Accelerating Innovation through AI-Powered Conceptual Abstraction and Interaction Design

HYEONSU BUTTWEILER KANG

CMU-HCII-24-105
August 2024

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aniket Kittur (Chair), Carnegie Mellon University
Sherry Tongshuang Wu, Carnegie Mellon University
Nikolas Martelaro, Carnegie Mellon University
Michael Terry, Google, Inc.

Submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation, the Office of Naval Research, the Allen Institute for Artificial Intelligence, and the Toyota Research Institute. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Human-AI Interaction, Machine Learning, Large Language Models (LLMs), Artificial Intelligence, Interaction Techniques, Reading Support Systems, Analogy Search Systems, Schematizing Unstructured Data, Analogy Engagement Support Systems, Sensemaking Support Systems, Creativity Support Systems, Ideation Support Systems, Innovation

Abstract

Isaac Newton famously said, “stand on the shoulders of giants,” to emphasize the importance of pre-existing synthesis for making new challenges tractable in a single human brain. Newton himself learned partial abstractions from Ptolemy, Copernicus, Kepler, and Galileo, as well as Descartes’ analytic paradigm, which he used as foundations for his calculus problem. However, rapidly accumulating knowledge makes it increasingly difficult to be aware of existing approaches and innovate upon them.

In this thesis, I argue that what we need are new tools to help people synthesize useful cross-cutting abstractions from knowledge, effectively organize knowledge with those abstractions, and use them to find novel cross-domain insights. I present four systems toward this goal, where I explore several kinds of abstractions to enable new interaction capabilities. These include ‘research threads’ for supercharging people’s reading experiences with AI to enable seamless interaction with thread-level abstractions while reading, the purpose-mechanism schema and how AI can help users find cross-domain analogies, and ‘active ingredients,’ a mechanism abstraction that helps designers engage with and transfer insights from biology to mobility design.

Through controlled laboratory studies, I demonstrate the value of these abstractions in elevating people’s focus during reading and exploration to a higher level (*e.g.,* from individual papers to how notable threads divide a research field; from individual species to active ingredients of mechanisms), thereby gaining efficiency and helping them broaden their pursuit of problem-solving strategies. The end result is more creative ideas.

In a world of abundant knowledge and large language models, the structuring and distilling of conceptual insights will be the defining characteristics of driving value in knowledge work. By putting powerful techniques that empower conceptual interaction with information into the hands of everyday people, I envision a future where innovators everywhere deeply engage with insights that overcome domain boundaries and develop novel ideas that address personal challenges they face to bring forth positive effects for the world.

Contents

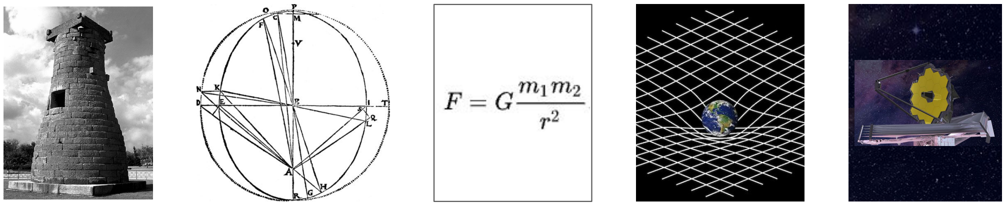
List of Figures

List of Tables

Chapter 1: Introduction

1.1 Is Innovators' Productivity Decreasing

Progress through Shifts in Abstraction



Cheom Seong Dae
c. 647

Regularities in Orbital Motion
Johannes Kepler
C. 1609

Universal gravitation
Issac Newton
C. 1687

Curved Spacetime
Einstein
C. 1915

James Webb Telescope
C. 2021

The scientific progress achieved over the centuries is truly awe-inspiring, expanding the scope of questions we can explore. Consider, for instance, the Cheom Seong Dae observatory (meaning "near the star place" in Korean), one of the oldest still-standing observatories, and juxtapose it with the cutting-edge James Webb Space Telescope (JWST), which orbits the second Lagrange point where no humans have reached to observe the earliest moments of the universe. Such advancements exemplify how far we have come in our quest for knowledge and understanding of the universe.

Tracing the history of this progress shows significant conceptual shifts driven by creation of new abstractions. For instance, in 1609, Johannes Kepler's work transitioned the prevailing model from Ptolemy's geocentrism to heliocentrism, and his insights laid the groundwork for Isaac Newton's universal gravitation theory, which in turn provided the foundation on which Einstein worked on his re-conceptualization of gravity. As such new conceptual frameworks can revolutionize our understanding and drive further innovation.

However, despite the seemingly remarkable progress, there is growing evidence that the rate of productivity growth among innovators is slowing down. A notable example is Moore's law, which describes the regular doubling of the number of transistors on an integrated circuit every two years, or equivalent to a 35% constant growth rate, that propelld advanced information technology and, along the way, remarkable economic growth over the past several decades. Yet, this masks the increasing input required to sustain such growth: with the number of Ph.D.s and funding in research growing more rapidly than 35% each year, the *effective innovator productivity is decreasing* (Fig. 1.1).

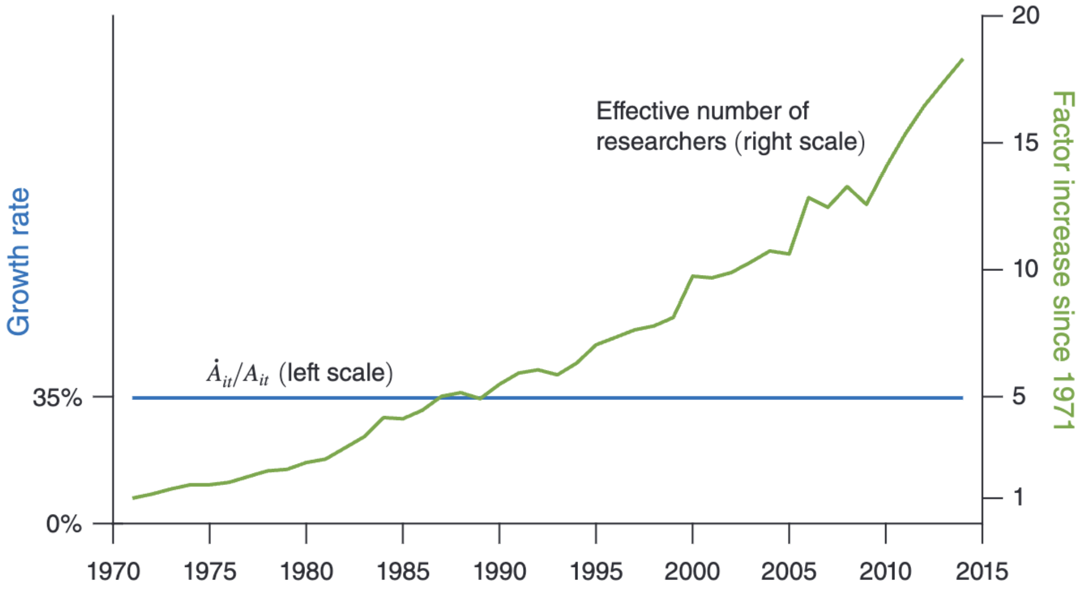


Figure 1.1: Trends in the growth rate in the effective number of researchers and the year-over-year growth rate specified by the Moore's law show a widening gap between the two, suggesting the productivity per researcher is falling over time (Figure adopted from [27]).

This slowdown in productivity is not limited to semiconductors but is observed across various key goods such as corn, cotton, and soy [27]. Furthermore, the quality of Nobel Prize-winning discoveries in physics show stagnant quality despite significantly increased investment [59], while patents and research findings are becoming more consolidatory rather than disruptive in nature [200]. In short, a unit of scientific and engineering innovation has become much more costly, suggesting decreasing innovators' productivity.

1.2 Harnessing the full potential of archived knowledge

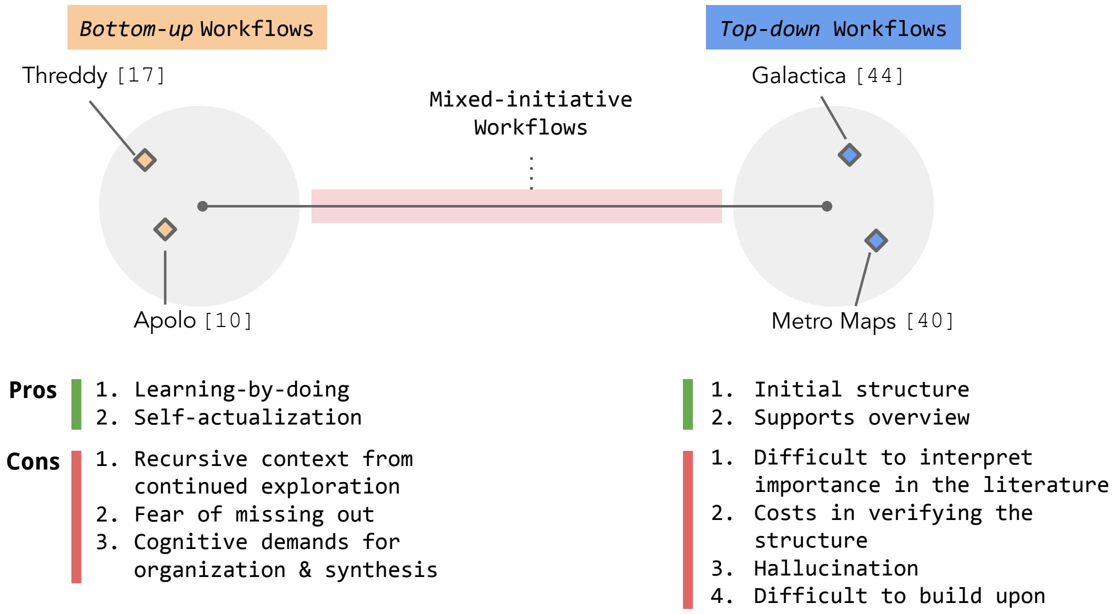
I believe one significant cause of the innovators' productivity decrease is the *increasing complexity and scale of knowledge production*, that leads to *increased burden of knowledge* and *challenging synthesis of existing information*. As the scale of accumulated knowledge grows, innovators must spend more time and resources to glean insights from it, and to consider many of the relevant existing works to contribute something new. An examination of conceptual diversity (calculated as the lexical diversity of title phrases) among research publications confirms this trend, as it has only grown linearly in recent years while the number of research publications has increased exponentially, widening the gap between the two [180]. Furthermore, Jones' analysis of the first time of great scientific achievements or invention show a trend in increasing amount of age before the first great achievements among the innovators [134].

Yet, Isaac Newton famously said, "standing on the shoulders of giants," to underscore the critical role pre-existing synthesis plays in making new challenges tractable in a single person's brain. Newton himself relied on the partial abstractions from the works of Ptolemy, Copernicus, Kepler, and Galileo, as well as the analytic paradigms of Descartes, to develop his solution to the calculus problem.

In this dissertation I argue that what we need for an antidote of the decreasing innovator productivity and realizing a bigger potential of archival knowledge is **new tools for helping people synthesize useful**

cross-cutting abstractions from knowledge, effectively organizing knowledge with those abstractions, and using them to find novel cross-domain insights.

1.3 Why aren't we getting better tools for building on existing knowledge?



Pros	1. Learning-by-doing 2. Self-actualization	1. Initial structure 2. Supports overview
Cons	1. Recursive context from continued exploration 2. Fear of missing out 3. Cognitive demands for organization & synthesis	1. Difficult to interpret importance in the literature 2. Costs in verifying the structure 3. Hallucination 4. Difficult to build upon

Figure 1.2: The initiative spectrum of workflow designs, with *bottom-up* and *top-down* workflows at the ends with corresponding pros and cons.

First, making sense of information – whether for scientists, designers, or lawyers – has inherent complexities [208] that involve many manual and interleaved steps before synthesizing useful *conceptual abstractions* that succinctly summarize the core aspects of information and their interrelationships. For example, consider a research scientist who wants to write a summary of notable research threads in a topic area she recently started exploring. To do so she opens up the PDF of a paper that she saved earlier for the topic. As she reads through the introduction and related work sections of the paper, she finds several sentences with a group of citations to prior work that describe interesting advances in the literature which she thinks are useful. She clips the snippets while taking notes on where they came from to review later, a cumbersome task that requires her to switch between a document editor, a PDF viewer, and a search engine. In the process, she opens up one of the new references found in a clipped snippet to examine further.

In the new paper, she once again discovers several citation contexts scattered around in the introduction and related work sections of the paper describing notable other approaches that have relevance. In order to track down the references, she recursively employs forward and backward citation chaining or footnote chasing [199], which quickly accumulates the layers in her search path. This prevents her from easily restoring the earlier context of research threads she started investigating (Fig. 1.2, left; Con 1). In addition, the possible search space from the collected references and citation contexts in the process increases exponentially, making it almost an impossible task to follow up on all of the discovery paths she might have, making her feel the fear of missing out on important research (Fig. 1.2, left; Con 2). Furthermore, as her collection of useful citation contexts and references expands, she finds it increasingly difficult to synthesize clips and references into a coherent outline describing multiple threads of research (Fig. 1.2, left; Con 3).

Systems that provide tailored support for each of the individual steps in the bottom-up process of sensemkaing do exist. However, existing tools often fall short of supporting the synthesis outcome users ultimately need to get at. While people are better at tracking down or organizing interesting references into folders during their sensemaking using such tools, they are left alone when it comes to extracting insights from the papers, discovering new papers and pinpointing and clipping the actual content that contribute to them, and organizing their structure together.

Switching to tools such as a new large language model (LLM)-powered and retrieval-augmented chat interface may represent the top-down end of the process well, and help users avoid some of the challenges by jumping over many steps in the iterative bottom-up process that people were previously engaged in. Instead now such a tool may provide a summary of several papers in response to a single user request. While it has potential to be remarkably useful, in such a scenario it is difficult for users to know how and why certain papers are included in the summary, and whether they are considered trustworthy and important works in the literature (Fig. 1.2 right, Con 1 & 2). It is also difficult to build upon the output (Fig. 1.2, right; Con 4), to expand by further including other related threads of research that divide the literature at a similar level of abstraction, or examine those sources by users themselves to deepen their understanding and consider alternative framing.

Once existing knowledge is mapped and relevant prior approaches are identified through the sensemaking process, the next goal is to build upon them to come up with novel solutions for a given problem. However, knowing what prior approaches have come before alone may not give rise to fresh new ideas. To this end, making analogies is an effective way of sourcing new inspirations. Aalogical inspirations have produced innovative breakthroughs throughout the history of science, engineering, and design. For example what used to be a 50-year-old challenge in aerospace engineering is “how do we transport a large number of solar arrays into the space for use by telescopes, satellites and others?”. One solution idea to this quest might be to make the arrays lighter, by optimizing the design of component parts such as the hinges, springs, etc. Another idea might be to build a bigger rocket, but the bigger the rocket that has to be built gets, the less feasible it gets given fuel requirements that significantly increase with the size of the rocket. When an astrophysicist working at NASA at the time encountered this problem, he happened to also be an Origami expert and recognized a deeper structural similarity between folding Origami papers and transporting and deploying large solar arrays into the space, despite the obvious surface level differences of one being typically involved with sheets of paper and the other with solar arrays. This kind of recognizing deeper structural similarities in spite of surface level differences is analogical processing. However, analogical innovations are rare because analogical transfer is challenging due to human memory retrieval sensitivity that favors analogs with surface-level similarity and the heavy cognitive load incurred during analogical processing.

Together, these represent the challenges that explain why building better tools for sensemaking of knowledge and ideation is challenging yet a valuable goal for exploration.

1.4 Systems for Helping People to Curate, Engage with, and Use Conceptual Abstraction during Sensemaking and Ideation

In this dissertation, I explore the design space of interacting with conceptual abstraction during sensemaking and ideation by designing and developing a series of novel interactive systems. These systems re-imagine existing tools by proposing new interaction techniques and computation to give users capabilties for curating, engaging with, and using conceptual abstraction during sensemaking and ideation. Within the design of these systems, I consider several questions such as 1) What use contexts do users need mechanisms of interaction with conceptual abstraction?; 2) How would users like to represent and interact with the abstractions created during their sensemaking and exploration?; and 3) What scaffolding can systems provide to help users engage with and iteratively build upon the artifacts they create?

The systems I develop explore a series of different user abilities for representing and interacting with conceptual abstraction. These include *extraction, externalization, and human-AI expansion of abstractions*, such as in the augmented reading experience where users can efficiently extract useful threads of research described in other scholars' existing work on the go (Threddy) and externalizing the extracted and saved threads as boundary objects that encapsulate user intent that can be used by AI seamless to expand in the neighboring citation graph (Synergi). I also examine *analogical inspirations based on a purpose-mechanism schema abstraction and user engagement with them*, such as in how scientists interact with an analogical search engine developed to retrieve analogical mechanism inspirations for their personal problem queries (Analogical Search Engine) and supporting their deeper engagement with the inspirations that may be unfamiliar at the outset but have potential to be valuable after transferring and further iterating on (BioSpark). The design of these systems were based on findings from workshops and pilot studies for needfinding where early prototype systems were built and used for interviewing users about their experience and problems they faced in their work before. After the development of each system, I conduct controlled lab experiments to deeply understand how the systems were used by people and compare to the status quo.

1.5 Thesis Contributions

In this thesis I focus on how to design novel sensemaking and ideation systems that scientists, engineers, and designers can use to curate valuable conceptual abstractions and engage with them to get inspired in their own problem domains. In order to build systems that help users to this end, I contribute research around the following three aspects of design across the systems and domains I explore.

1.5.1 Design for Interacting with Conceptual Abstraction by Research Threads

Design for Efficient Extraction and Organization of Threads

Scientists and inventors position their work in the context of most relevant prior work and draws conceptual similarities and contrasts with them. This positioning and synthesis of contribution provides a valuable resource for learning salient concepts in the field and knowing what language people in the field use to describe them. Such framing and positioning of contributions often appear in the introduction or related work sections of a paper, where text exposition is interleaved with reference notations and provided together. Therefore, this combination of exposition and references was the primary abstraction underlying the interaction design for Threddy and Synergi.

From discussions with other scholars and reflections on our own experience as researchers, we learned that people use a variety of processes for reading these sections and curating threads. Some described taking notes in the margins on the page while reading, others mentioned adding sticky notes near the passage they were referring to, yet others used different tools, like Google Docs, Zotero comments, or Figma boards for curating notes describing salient threads and associating them with the original papers. However, most people we talked to felt this resulted in a ungainly pile of notes, scattered across many papers, which made them not particularly conducive to the synthesis phase that followed. This also confirmed similar findings in the prior literature. We also found that in-text highlighting was commonly used among the note-takers and considered an intuitive input mechanism for this purpose, yet having an additional place for synthesis (*e.g.,* Google Docs or Spreadsheets) separate from the paper they were reading meant there was added context switching cost every time they moved between the two for viewing or updating content.

Based on these observations, the interaction design of Threddy and Synergi prioritized an integrated environment for both reading and synthesis. This materialized into a design choice that replaced the default Google Chrome browser PDF viewer with a native one that allowed users direct in-text highlighting while also having access to the highlighted content for backend computation that powered various information extraction and organization features. The content users highlighted for extraction typed in on their own as threads to save were added to the sidebar located on the right of the PDF viewer. In order to support efficient structuring, saved threads semantically relevant to the newly highlighted thread were ranked higher on the candidate list for associating at the time of saving.

Techniques for Human-AI Expansion of Research Threads

The thread objects in Threddy supported the discovery of recent papers by ranking them based on citation coverage related to the references included in a selected thread. This helped users decide which papers to read next, continuing their synthesis of research threads. However, users still struggled with the extensive effort required for manually engaging in the bottom-up process of moving between papers to identify and curate threads.

To address these challenges, we developed Synergi. Synergi expands on user-saved threads by leveraging them to automatically identify related papers, parse their full text to home in on specific content snippets with supporting citations, and process the most relevant snippets to produce a pre-digested structure which users could overview and incorporate into their synthesis thus far, seamlessly expanding its scope. By highlighting text that describes research threads, users can trigger Synergi to construct a graph of relevant papers and summarize their content hierarchically. This supports both detailed bottom-up and high-level overview processes, catering to different user needs.

A key insight of the Synergi thread-expansion algorithm is that research threads can act as boundary objects for AI, enabling the construction of a thread-based understanding that bridges individual papers. Synergi aids users in visualizing citation contexts and their interconnections, helping them develop comprehensive and reusable outlines. As such, Synergi helps to address the limitations of manual synthesis by providing a human-AI expansion of individual threads through discovery and summarization of research threads from related papers that have not been included in the synthesis thus far, but nonetheless related given its content. In this way, Synergi provides a mixed-initiative framework that supports both bottom-up and top-down workflows, facilitating a higher-level engagement with the literature and enabling users to effectively synthesize and organize research insights.

1.5.2 Empirical Understanding of Engagement Challenges in Purpose-Mechanism Abstraction-Based Analogical Transfer

Knowing what prior approaches exist in the literature is helpful for coming up with new ideas. However, to actually generate novel ideas, one must transform existing concepts into a new form. One way to achieve this is through analogical processing based on the purpose-mechanism schema abstraction, where diverse mechanisms with a similar purpose to the query purpose ca be discovered. An open question is how much the degree of purpose-match among the diverse mechanism ideas affects creative ideation during analogical processing, which is known to be challenging to human minds due to several cognitive constraints, including the human memory retrieval sensitivity that prefers analogs with surface-level similarity and the prohibitive cognitive load generated during analogical processing involving only a few relations at once.

In a controlled laboratory and case studies using an interactive analogy search engine operating on scientific papers, I found that partially matching purposes – similar at a high level but differing in low-level details – often led to more creative adaptation ideas than keyword-based searches. However, these partial mismatches can either facilitate creativity or render analogies useless, depending on their nature.

I identified two main challenges involved in this process. The first challenge was discovering salient structural elements. Participants noted difficulties in expressing their problems at the outset of search and recognized that important constraints are often identified only after encountering mismatched structures. This iterative process was essential for refining problem representations and steering research effectively.

The second challenge was supporting the right amount of mismatched elements to encourage useful variations. While exact purpose matches could result in missed opportunities, too much deviation made ideas infeasible. Balancing this mismatch was crucial for fostering creative adaptation.

In summary, the findings highlight the potential and challenges of using an analogy search engine to enhance creative scientific ideation. Purpose-match significantly mediates ideation outcomes, and understanding beneficial mismatches is essential. The iterative discovery of structural elements and balanced superpositions of mismatches are key factors in facilitating effective analogical transfer.

1.5.3 Techniques for Deepening User Engagement with Analogical Inspirations and Supporting Transfer

To mitigate the challenges involved in users engaging with analogical inspirations whose concepts may be unfamiliar, I explored LLM-powered strategies with BioSpark. BioSpark introduces several new interaction features designed to help users engage more deeply with analogical inspirations and boost the likelihood of successful transfer from these inspirations to target domains.

One of the core features of BioSpark is the ‘sparks’ button. This functionality allows users to generate new application ideas for a selected organism and its mechanism inspiration with each click. These ideas are contextually related to the user’s design problem and its constraints, enabling a continuous flow of relevant inspirations. This not only encourages playful exploration but also aids in the iterative process of refining and adapting these inspirations to their specific problem contexts. Additionally, BioSpark facilitates deeper engagement through a Q&A chat interface. This interface leverages an LLM-based agent to dynamically respond to user queries and provide contextual clarifications. Moreover, the tradeoff analysis button helps users to evaluate different aspects of the analogical inspirations, providing an analytical framework to assess the feasibility and potential impact of each idea. The combination of features enables users to interactively digest and integrate inspirations, which might initially seem alien or impractical.

To support the interactive capabilities, I explore a backend process for generating a repository of the active ingredients in analogical mechanisms found in nature that may be transferrable to target design domains. To do so it iteratively builds a hierarchy of organisms, known as the tree of life, to identify sparse branches on the hierarchy that indicates areas rich in potential for further expansion with the LLM. This back-end process ensures a growing and evolving repository of analogical inspirations aligned with user needs.

By combining these advanced interaction and backend features, BioSpark empowers users to navigate the unfamiliar terrain of analogical inspirations effectively. The system fosters a deep, iterative engagement, ensuring that even initially unfamiliar concepts can be explored and adapted to yield valuable innovations in the users' target domains. This LLM-powered approach facilitates not just the discovery of novel inspirations but also the practical transfer and integration of these inspirations, ultimately enhancing creative problem-solving and ideation processes.

1.6 Thesis Overview

- Chapter 2 situates this dissertation in the context of related research into sensemaking and analogical ideation, the theories I use to understand the sensemaking and analogical processing and transfer process, as well as new tools and techniques towards improving both the bottom-up and top-down ends of the initiative spectrum.
- Chapters 3 – 6 describe particular systems and studies that explore novel designs for sensemaking and ideation focused on the conceptual abstractions of research threads, purpose-mechanism schemas, and active ingredients of mechanisms. In these chapters, I report data from formative studies, provide descriptions of system specifications and implementation details, and results from evaluative studies.
- Chapter 3 describes Threddy, a system that introduces the research thread abstraction and develops an augmented scholarly reader interface that supports efficient extraction and organization of threads from research paper PDFs while reading. Then, I present the results of a controlled laboratory study evaluating Threddy with scientists.
- Chapter 4 examines human-AI collaborative expansion of research threads and introduces Synergi, that aims at augmenting users' abilities to gain thread-based understanding and discovery in the related literature by extracting notable threads from papers related to previously user saved threads, and pre-digesting papers along the saved threads to provide an initial structure that users can overview and build upon. I describe the Synergi retrieval algorithm as well as its hierarchical summarization process that uses an LLM. Then, I present the results from a laboratory study evaluating Synergi against Threddy.
- Chatper 5 examines how new ideas can be developed from understanding of prior works, and through analogical processing based on a purpose-mechanism schema abstraction. I examine the benefits and challenges involved with analogical processing with scientists using an interactive analogical search engine where users could type in descriptions of research problems they were personally interested in receiving inspirations for. I describe how the analogical search engine was developed, by first developing a sequence-to-sequence model trained on crowdsourced annotations of purpose and mechanism tokens from research paper abstracts, deploying this model to process papers at scale, and then by building a real-time search engine that searches for papers with a similar purpose and diverse mechanisms upon user queries. Then, I present the results from a controlled laboratory and case studies that describe how mismatching on certain purpose-level constraints can trigger generative ideation, yet how this process is hampered by participants' sensitivity to near

analogs, and how they cannot enumerate the important constraints of a purpose at the outset nor accurately assign the importance of matching on each constraint without seeing mismatching analogs.

- Chapter 6 describes BioSpark, a system for professional mobility designers seeking inspirations from organisms in nature for their concept designs. In order to bridge the challenges with engagement identified in Chapter 5, BioSpark introduces several new interaction features for users engaging with analogical inspirations, to deepen their engagement with inspirations and boost the likelihood of successful transfer from inspirations to target domains. I first describe the end-to-end system specification of BioSpark, which includes both a backend data generation process and the front-end interaction features. The backend dataset generation pipeline iteratively constructs the tree of life hierarchy using an LLM, by incorporating organisms represented in the dataset generated so far into the hierarchy. Then it identifies sparse branches on the hierarchy that represent potential opportunities for expansion which it uses to prompt an LLM for further generation. BioSpark interface supports LLM-based engagement features such as a 'sparks' button where users can generate new application ideas for a selected organism and its mechanism inspiration on each click of the button, a Q&A chat interface, and a trade-off analysis button, all of which are contextualized to the design problem and its constraints and additionally personalized to the interaction history when appropriate, to cater to specific user needs.
- Chapter 7 summarizes design lessons from the systems presented in the previous chapters and their evaluations and discusses their design implications with regards to what might need to change for broader adoption of conceptual abstraction-focused sensemaking and ideation systems. I conclude with Chapter 8 by reviewing the contributions of the dissertation.

Chapter 2: Related Work

Scholars have studied how people make sense of complex information for decades. Systems aimed at helping people with the sensemaking process have been developed for just as long. Pirolli and Card's Sensemaking loop modeled the process into two main phases, the foraging and the sensemaking phase, iteration through which is how people transform raw data into a coherent story for their inquiries [208]. In the foraging phase, people engage in activities such as seeking, searching, filtering, reading, and extracting relevant information from various sources for use later. In the latter phase, people develop schemas from the gathered and whittled down information, from which they develop hypotheses for their inquiries.

Klein et al. further investigated how the schema production process works in human psychology, and proposed a data/frame theory of sensemaking [152]. In it they use the term '*frame*', a cognitive structure in which entities are defined by their relations to others. The frame can take many different forms, such as a story – which explains the chronology of events following a causal relationship, a map – which shows feasible routes to destinations via geographical relations, a plan – which describes the sequence of intended actions, among others. Notably, in the data/frame theory of sensemaking the development of a frame follows an iterative process of how data is explored and perceived to discover relevant frames. And the relevant frames then guides the perception of further relevant spaces of data to explore. Hence, the two aspects – data and frames – are dual and have interleaving processes that cycle to move forward; a frame defines key elements of the situation, assigns significance, and describes their relations to each other. This then triggers subsequent actions, such as filtering irrelevant information and highlighting relevant ones, thereby constructing what counts as data. Information that passes through these actions then modifies the initial frame, and the cycle continues.

Many earlier works in the literature of sensemaking provide a foundation for the data/frame theory described above. Among them, Marvin Minsky formulated a similar process as fitting individual cases to a concise abstraction of a feature set and their parameters [181]. Piaget thought of schema as an internal representation that a person retains for reconstructing the persistent features or attributes of an entity [206]. For Schank and Abelson [221], it was understood as a pattern generalizing a recurring sequence of events or 'scripts'.

The interleaved processes of synthesizing a frame and using the synthesized frame to guide the data search and transformation process point to a broad design space for technological interventions. Specifically, I focus on three concrete forms of conceptual abstractions, or frames, that people produce when engaged in relevant sensemaking processes: *research threads* synthesized from reading scholarly texts, *purpose-mechanism schemas* that can enable the retrieval of novel analogical insights for problem-solving, and *active ingredients* of mechanisms that help designers transfer analogical insights into target problem domains. I study and develop novel interaction and computational techniques enabled by AI and LLMs that unlock new capabilities for users in exploring this design space. In the following sections, I review novel systems and techniques developed in the literature that can help people effectively make sense of information.

2.1 Systems that Support Bottom-up Synthesis

Scholars face numerous challenges when exploring and reviewing the large-scale and ever-changing literature [29, 131, 177, 193, 216, 245], such as information overload [179] and difficulty allocating attention effectively [229]. Finding and organizing relevant papers into multiple, evolving research threads, and updating these threads with recent literature, further complicates scholars' workflows. Consequently, literature review is often considered tedious, scattered, and reliant on chance [32, 161]. To mitigate these issues, various systems have been developed to aid scholars in understanding research articles.

Workflows at the *bottom-up* end of the spectrum (Fig. 1.2, left) involve practices like forward and backward citation chasing and footnote chasing. These are essential for scholars navigating citation graphs to discover significant papers related to a research problem [199]. However, these workflows often experience fragmented information environments and disjointed tools [278], adding complexity and distracting from synthesis. While related work or introduction sections of review papers may provide valuable syntheses of relevant domains [243], they are not ideal for iterative exploration of literature in creating one's own review with different focus, scope, or framing. Scholars must read multiple review papers for synthesis [198], which can significantly increase cognitive and interaction costs.

2.1.1 Systems that support information extraction while reading

To aid users in bottom-up sensemaking processes, Passages [106] introduced an approach to 'reify' [21] ephemeral user text selections into persistent objects shareable across multiple applications. These objects can be incorporated into various representations, such as a canvas or spreadsheet view, allowing users to maintain their reading context while extracting and storing relevant information. Likewise, ForSense [211] and Fuse [158] provided a sidebar view in webpage margins, enabling users to highlight content snippets and organize them into clusters while viewing. Crystalline [168] further supported the review-extract loop by automatically extracting software package-related criteria from developer websites using natural language parsing. Mesh [50] offered similar assistance by automatically pulling in Amazon product reviews from the API matching user-defined criteria about product categories.

2.1.2 Systems that augment reading interfaces to enhance comprehension

Another area of research related to bottom-up sensemaking focuses on augmenting reading interfaces. Reading individual papers can involve high cognitive costs due to unfamiliar terms, domain-specific jargon, nonce words [110], complex formulae [111], and required expertise for comprehension [20, 114, 190]. To address these costs, one thread of research has targeted cross-referencing within a single document to improve readability (e.g., [12, 77, 212]). A sub-thread of this focuses on enabling more efficient navigation between parts of a table and corresponding text [15, 146, 156]. Another sub-thread involves support for concept diagramming while reading (*cf.* [233]), which may help readers more deeply engaging with the material. In contrast, *NB* [282] enhances engagement and learning by adding contextually relevant discussions to the margins of a document students are reading. These systems demonstrate the importance of reducing cognitive costs associated with reading complex texts, like research papers, and suggest potential benefits of augmenting reading interfaces. This allows readers to build upon authors' pre-digested synthesis from multiple papers in a contextually relevant manner without disrupting their reading flow, promoting better focus and mental models of related research threads.

2.1.3 Systems that support paper-centric literature discovery and interaction

Significant research efforts have been devoted to developing interactive systems for scientists and professionals (*e.g.,* [188]) in various stages of literature discovery.

For example, PaperQuest [209] suggested relevant papers based on citation relationships using query papers. Apolo [53] allowed users to save papers or clips and expand with additional items via the Belief Propagation algorithm. Sturm [232] studied requirements for literature search systems and developed LitSonar where users could deploy nested queries to query over multiple sources of document streams. LitSense [235] included multiple citation relation visualizations and supported filtering and querying for homing in on specific references for further exploration. Papers101 [55] helped scholars search for additional relevant literature by generating unused keywords for query expansion. CiteSense [278] developed an information-rich environment which provides various features for searching, appraising, and managing the different tasks involved in a literature review. Lastly, Relatedly [198] recommended more specific content – relevant paragraphs in related work sections – than papers for exploration.

Except for Relatedly, most previous systems emphasize *documents* (i.e., research papers) for user interactions and discovery. Support for crosscutting abstractions, such as research threads spanning multiple papers and interaction techniques to build them over time, remains limited. Though the concept of 'reified' objects has been implemented before (*cf.* Passages [106]), it mainly focuses on extracting text selections and aggregating them into a single view, without capturing the rich context, such as cited references, semantic meaning, and the citation graphs. Moreover, the exploration of other synthesis-related objects, like expert author committees, is still lacking.

2.2 Systems that Support Top-down Synthesis

2.2.1 Systems that support overview of the information landscape

On the opposite end of the spectrum (Fig. 1.2, right), systems such as ConnectedPapers[1], Metro Maps of Science [224], and Wang et al.'s narrative visualization system [264] provide a *top-down* visual overview of the research landscape. These systems help scholars comprehend the structure of knowledge space and discover interesting areas within it. For a collection of documents, IntentStreams groups and visualizes documents relevant to a search query into streams [8], while Apolo [53] and IdeaHound [227] use a 2-D spatial arrangement to quickly overview similar document clusters. Although these representations are useful for entering a new knowledge domain, they often lack additional user interactions beyond the overview stage, limiting their utility for synthesizing knowledge scattered across multiple papers.

2.2.2 LLMs for generating an overview of a topic

Recent advances in Large Language Models (LLMs) like Galactica [242], ChatGPT[2], and Google Bard[3] showcase impressive capabilities in answering user questions using synthesized web knowledge. Tools such as Ask Your PDF[4] indicate promising future systems supporting personalization and specification based on user-curated documents. However, LLMs face challenges like hallucination and falsehood (*cf.* [18, 26, 248]), making their outputs uncertain and less trustworthy, requiring manual inspection

[1] https://www.connectedpapers.com/
[2] https://chat.openai.com/chat
[3] https://bard.google.com/
[4] https://askyourpdf.com/

and verification. Furthermore, their computation processes are obscured [9] and less interpretable to users [165, 279], limiting their ability to learn, iterate, and synthesize based on these outputs.

2.3 Systems for Supporting Analogical Inspirations

Many innovations in design, technology, and science have been driven by harnessing analogical inspirations from fields distant to one's own. Historical examples include Vetruvius' analogy of sound waves with water waves [61], the Wright brothers' wing control mechanism adapted from a bicycle inner tube box [133], and engineers using origami concepts to furl a solar array for space deployment [183, 204, 281]. These examples demonstrate that analogical innovation requires the inventor to map inspirations with deep structural similarities, often from seemingly unrelated domains [87, 89, 93].

However, supporting the entire cognitive process of analogical innovation in a single system has proven challenging [129, 140]. Most existing approaches focus on only one stage of the process and often rely on a small hand-coded set of inspirations [45, 66, 97], which limit their scalability and effectiveness. Past systems targeting analogy retrieval mainly focused on modeling analogical relations in non-scientific domains or within restricted scopes, such as structure-mapping [80, 81, 82, 87, 250], multiconstraint-based [72, 120, 127], connectionist [117], and rule-based reasoning [11, 39, 40, 259] systems. The high costs of developing structured representations hindered hand-crafted systems, such as DANE [127, 257]), from providing comprehensive topic coverage and real-world applicability.

Conversely, scalable computational approaches, including keyword or citation-based search engines, are limited by reliance on surface or domain similarity. These engines maximize similarity to queries which is helpful for identifying work within the target domain but less effective when seeking inspirations outside that domain (for example, for Salvador Luria's queries: "how do bacteria mutate?" or "why are bacterial mutation rates so inconsistent?", similarity maximizing search engines may have found Luria and Delbrück's earlier work on E.coli [171] but may have failed to recognize more distant sources of inspiration such as slot machines). Supporting users in finding relevant inspirations and aiding them in adapting these inspirations to their problems can lead to deeper engagement and potentially fruitful outcomes [10, 228, 260].

2.3.1 Bioinspired Design

One thread of research in design by analogy focuses on biological organisms and systems [129]. Projects such as AskNature [66] and DANE [97] rely on manual curation, which is labor-intensive. For example, redescribing a single biological organism in the Structure-Behavior-Function framework can take approximately 40-100 hours per model. Alternative approaches have used crowdsourcing to identify biomimetic inspirations in scientific articles (*e.g.,* [258, 280]), but high-quality annotations pose significant scalability challenges. Rule-based and data programming approaches also show promise but face issues of generalizability and scalability [54, 74].

2.3.2 LLMs for Ideation and Co-Creation

Recent advancements in LLMs offer potential for enhancing analogical innovation across all stages of the cognitive process [267]. LLMs can infer specific analogies, generate relevant ideas, and provide flexible natural language interfaces for interaction [1, 142, 178, 195].

However, studies have shown that improper incorporation of LLMs in the creative process can lead to fixation rather than increased creativity. Issues such as inaccurate inferences, hallucinations, and fixation on AI-generated ideas have been highlighted [147, 261]. Thus, a more nuanced approach to using LLMs in analogical innovation is necessary, aiming to augment human creativity without replacing it or causing undue fixation.

Chapter 3: Threddy

Supporting Personalized Thread-based Exploration and Organization of Scientific Literature

This work was previously published in ACM UIST 2022 ([136]) and has been adapted for this document.

Although numerous systems aim to assist users in their bottom-up sensemaking processes, many do not enable users to directly extract and curate synthesis outputs (*e.g.,* research threads) provided by others and develop them over time. Incorporating interaction features that align with users' existing workflows and support natural research thread discovery can substantially enhance sensemaking by reducing interaction, context-switching, and cognitive costs, while aiding knowledge relation and structure identification.

Furthermore, treating threads as first-class objects not only allows a broader exploration beyond the current paper but also enables effective interpretation of users' interests and intentions by AI during literature exploration. The combination of rich description and citations in threads can be leveraged to better target user interests.

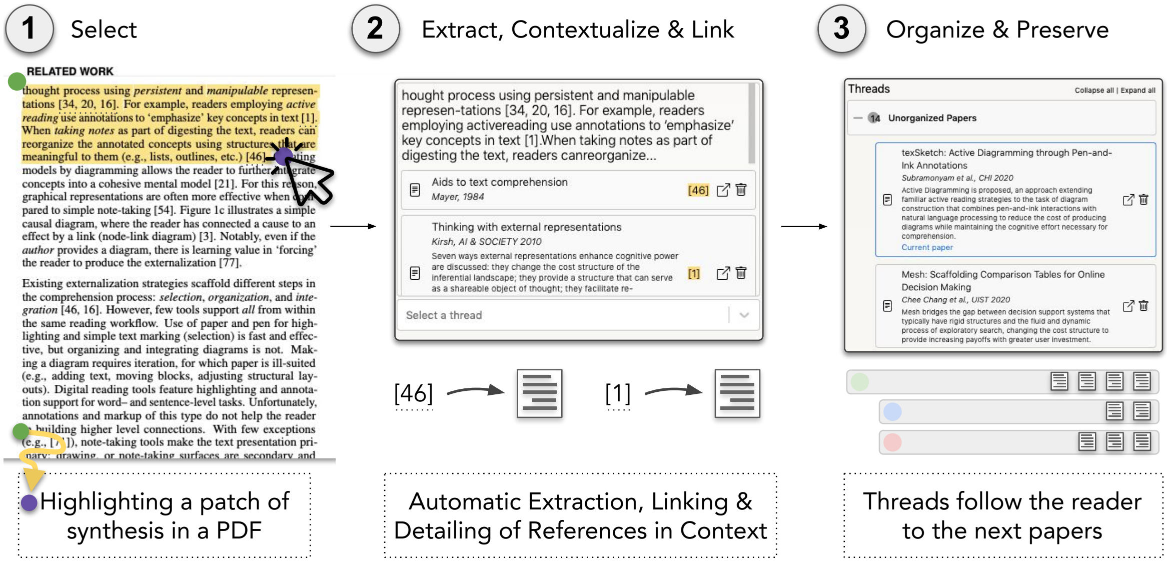
In this chapter, we explore a new paradigm for augmenting users' scientific paper reading by introducing Threddy, a platform that enhances a web-based PDF renderer with low-cost interaction features. Threddy enables efficient in-context extraction and organization of research threads and recommends additional relevant papers. It fosters users' understanding of their own work's contributions while organizing threads and supporting evidence during reading.

3.1 Introduction

Reviewing the literature to understand relevant threads of research is a critical part of scientific research and serves as a research facilitator and a vehicle for learning [33]. For example, a scholar trying to understand the history of tools that support scientific literature review might learn about research threads including overview visualizations based on citation networks; augmentative interfaces for active reading; collection tools that help scholars organize their papers; and so forth. Understanding prior threads of research is critical to building on past work, finding inspirations for new innovation, and positioning contributions in the appropriate research context.

However, as the scientific literature grows the challenges for users to find and make sense of the many different threads grow as well. Finding and keeping track of papers within a single thread can be challenging, requiring users to traverse references and citations, read through introductions and related work sections, and search across various keywords to avoid missing important work. Exacerbating the problem, scholars are often interested in multiple threads that are relevant to their work, with each often branching into multiple sub-threads as the example described above of literature review support tools demonstrates.

An effective strategy for 'shortcutting' the cumbersome process of assembling research threads is to harvest and build on the work that other scholars have already done in assembling them. This process, commonly used among scholars [33, 278], involves reading through papers (typically in the introductions and related work sections) to find how the authors have compressed and summarized the threads of research relevant to their papers in order to situate their own work's contributions. These predigested


1 Select
2 Extract, Contextualize & Link
3 Organize & Preserve
Highlighting a patch of synthesis in a PDF
Automatic Extraction, Linking & Detailing of References in Context
Threads follow the reader to the next papers
[46]
[1]
Select a thread
Threads
Collapse all | Expand all
14 Unorganized Papers
Current paper

RELATED WORK

thought process using *persistent* and *manipulable* representations [34, 20, 16]. For example, readers employing *active reading* use annotations to 'emphasize' key concepts in text [1]. When *taking notes* as part of digesting the text, readers can reorganize the annotated concepts using structures that are meaningful to them (e.g., lists, outlines, etc.) [46]. ating models by diagramming allows the reader to further integrate concepts into a cohesive mental model [21]. For this reason, graphical representations are often more effective when compared to simple note-taking [54]. Figure 1c illustrates a simple causal diagram, where the reader has connected a cause to an effect by a link (node-link diagram) [3]. Notably, even if the *author* provides a diagram, there is learning value in 'forcing' the reader to produce the externalization [77].

Existing externalization strategies scaffold different steps in the comprehension process: *selection*, *organization*, and *integration* [46, 16]. However, few tools support *all* from within the same reading workflow. Use of paper and pen for highlighting and simple text marking (selection) is fast and effective, but organizing and integrating diagrams is not. Making a diagram requires iteration, for which paper is ill-suited (e.g., adding text, moving blocks, adjusting structural layouts). Digital reading tools feature highlighting and annotation support for word– and sentence-level tasks. Unfortunately, annotations and markup of this type do not help the reader building higher level connections. With few exceptions (e.g., [ ]), note-taking tools make the text presentation primary: drawing, or note-taking surfaces are secondary and

hought process using persistent and manipulable represen-tations [34, 20, 16]. For example, readers employing activereading use annotations to 'emphasize' key concepts in text [1].When taking notes as part of digesting the text, readers canreorganize...

Aids to text comprehension
Mayer, 1984

Thinking with external representations
Kirsh, AI & SOCIETY 2010
Seven ways external representations enhance cognitive power are discussed: they change the cost structure of the inferential landscape; they provide a structure that can serve as a shareable object of thought; they facilitate re-

texSketch: Active Diagramming through Pen-and-Ink Annotations
Subramonyam et al., CHI 2020
Active Diagramming is proposed, an approach extending familiar active reading strategies to the task of diagram construction that combines pen-and-ink interactions with natural language processing to reduce the cost of producing diagrams while maintaining the cognitive effort necessary for comprehension.

Mesh: Scaffolding Comparison Tables for Online Decision Making
Chee Chang et al., UIST 2020
Mesh bridges the gap between decision support systems that typically have rigid structures and the fluid and dynamic process of exploratory search, changing the cost structure to provide increasing payoffs with greater user investment.

Figure 3.1: Thread creation and organization while reading on Threddy : ① The reader highlights a useful patch of text that interleaves references in a citation context. ② The system extracts the referenced papers from the highlighted text, link them to the citation context, and show the resulting data to the reader. ③ The reader creates a new thread together with the citation context and extracted references. Alternatively, the user can add the context only as a relevant clip to an existing thread, or add the extracted references to it. The threads and their context follow the reader as she continues on reading other papers.

threads provide scaffolding for users in assembling their own, both in terms of the references cited as well as the citing text describing those references. By following the most relevant references, finding more citing texts, and further chaining through papers, scholars can more quickly assemble an overview of the research threads in an area than by searching and collecting individual papers alone.

However, even this process of inferring threads is cumbersome. Consider the following scenario in which a scientist is learning from a new research paper. Quickly skimming the introduction, she may identify a useful patch of text in it which describes a research thread she would like to explore further. This patch of text (e.g., a sentence, a paragraph, or a section) often contains a number of citations pointing to related work and describes their relation which provides a helpful context. Deciding to save this context and follow up on one of the references requires her to first jump between the references section and the citation context in the body text to link the reference notation she wants to follow up with to the actual title and URL of the paper. Next she needs to locate the actual content of the paper, perhaps by querying its title on a search engine. Finally, she may make notes of the found paper and save it for future reference. She needs to repeat this multiple times for each patch of text she finds interesting, the cost of which compounds quickly. The reference notation used in papers may also differ, sometimes providing little context about what they are (e.g., numbers in a bracket such as '[1]'), and this lays an even more burdensome task of correctly mapping and linking papers, which (using one of our study participant's words) can be "*a real, damaging context break.*" In addition, she might after all end up finding that the actual content of the cited paper to be irrelevant or uninteresting, in which case she must resume her flow of reading by re-building the lost context and previous threads of thoughts.

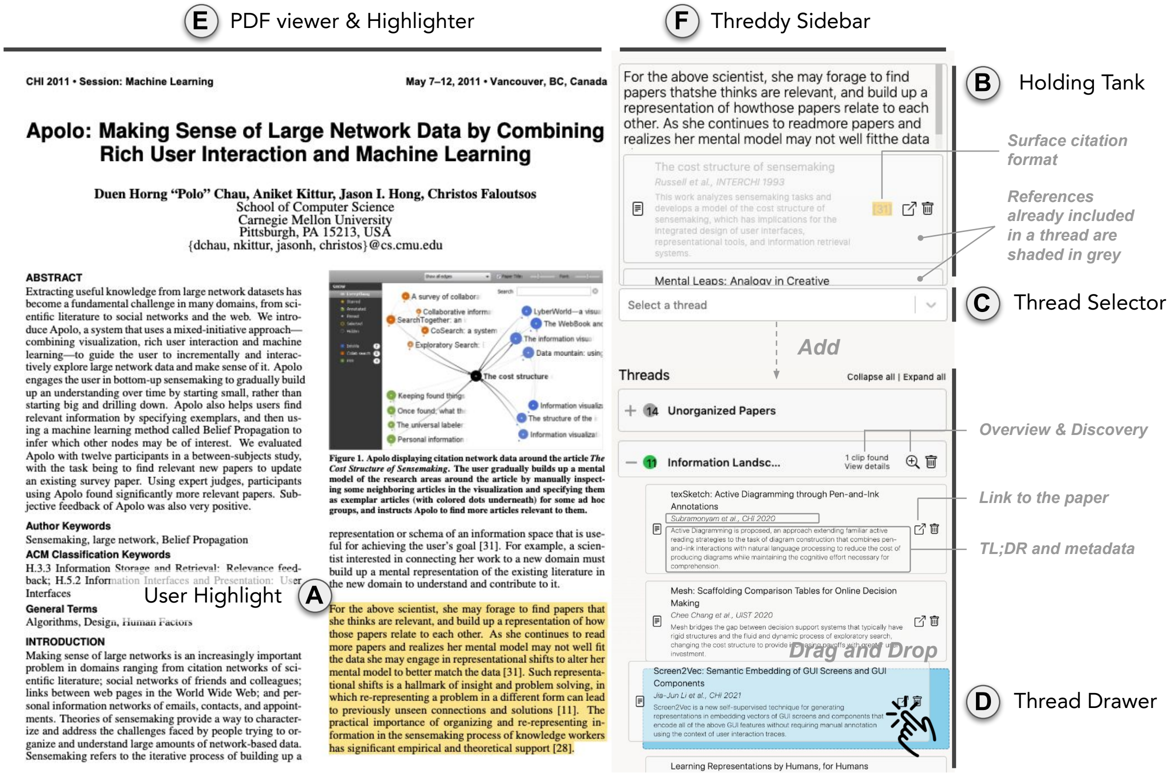
As she moves to other papers and collecting more patches of relevant information, it may become clear that they relate to each other along some threads she created, and this needs to be captured. In order to achieve this, she first needs to look through her own notes, threads, and references that she may have loosely organized, and this quickly becomes a sizable sub-task that once again requires her to break out of her flow of reading in order to complete it. Furthermore, it is easy to forget which references are already looked at and which are new that need to be processed, which may incur additional friction to the process. She may also have multiple threads she would like to follow up on at any given time, without having an easy way of keeping track of them while she is reading a paper. Using multiple tabs or groups of them might be an intuitive way to organize articles into related threads (albeit its potential for creating a tab overload [51, 52, 100]), but this does not help with maintaining patches of citation context that described the threads. Finally, once relevant patches of text are collected for each thread from source papers, there is no easy way to use this information to find additional relevant papers that she may use to further grow the threads.

While significant research has focused on supporting scientific literature search and collection, there is relatively little support for users building threads during reading. For example, *Papers101* [55] aims at supporting early-stage scholars' discovery of literature by recommending relevant but unused keywords for query. Alternatively, *Apolo* [53] adopts and applies the Belief Propagation [274] algorithm on the citation network in a novel way to support progressive retrieval of papers given a set of papers that the user has collected thus far. Once a list of recommendations are curated, systems such as *PaperQuest* [209] provide support for triaging what to read next, and *VisualBib* [63] aims at providing a more holistic support for managing the growing user-curated bibliographies. However, none of these systems provides a mechanism for leveraging the data scholars have collected and assembled into threads while reading research articles to recommend further relevant papers to continue growing the threads.

In this paper we aim to address this gap in the literature by developing Threddy , a system that supports users with collecting patches of text in research articles that contain pre-digested syntheses by other authors (i.e., a useful citation context along with automatically extracted references), and helps them assemble personal research threads using clippings of others' pre-digested threads. Different from the related prior work, Threddy does not create a new information environment nor an application context that requires users to context switch away from their natural flow of reading, but instead seamlessly integrates the support it provides into the user's in-situ context of reading. We evaluate Threddy in a controlled lab study with 9 scientists conducting literature review in personalized domains, and demonstrate how it increases users' effectiveness in leveraging pre-digested syntheses by other authors to enrich their own threads, decreases the cost of frequent context switching they would have experienced in a similar task without the tool, and heightens users' flow state while conducting a literature review.

3.2 Usage Scenario and System Design

Usage Scenario. We first illustrate how an end-user, Sam, would interact with Threddy to conduct a literature review. Sam is reading a paper when she encounters an interesting patch of text (see also figure 4.1). The two paragraphs in the related work section describe particularly relevant research threads that she wants to follow up on and save the references included in them for a deeper look. She quickly highlights the paragraph, which triggers Threddy to search for references included in it, automatically extract the metadata corresponding to each, and present them as interactive objects in the sidebar. Sam glances over the linked references and removes a couple of them which seemed less relevant based on their titles and TL;DR summaries, while keeping the rest. The extracted references and the context seem to form a good grouping for revisiting later, so she adds them as a new thread labeled '*Reifying ephemeral user interaction (e.g., text selections)*'. She continues reading the paragraphs and repeatedly extracts and saves additional context, references, and threads such as '*Systems for constructing narrative structures in*



CHI 2011 • Session: Machine Learning May 7–12, 2011 • Vancouver, BC, Canada

Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning

Duen Horng "Polo" Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dchau, nkittur, jasonh, christos}@cs.cmu.edu

ABSTRACT
Extracting useful knowledge from large network datasets has become a fundamental challenge in many domains, from scientific literature to social networks and the web. We introduce Apolo, a system that uses a mixed-initiative approach—combining visualization, rich user interaction and machine learning—to guide the user to incrementally and interactively explore large network data and make sense of it. Apolo engages the user in bottom-up sensemaking to gradually build up an understanding over time by starting small, rather than starting big and drilling down. Apolo also helps users find relevant information by specifying exemplars, and then using a machine learning method called Belief Propagation to infer which other nodes may be of interest. We evaluated Apolo with twelve participants in a between-subjects study, with the task being to find relevant new papers to update an existing survey paper. Using expert judges, participants using Apolo found significantly more relevant papers. Subjective feedback of Apolo was also very positive.

Author Keywords
Sensemaking, large network, Belief Propagation

ACM Classification Keywords
H.3.3 Information Storage and Retrieval: Relevance feedback; H.5.2 Information Interfaces and Presentation: User Interfaces

General Terms
Algorithms, Design, Human Factors

INTRODUCTION
Making sense of large networks is an increasingly important problem in domains ranging from citation networks of scientific literature; social networks of friends and colleagues; links between web pages in the World Wide Web; and personal information networks of emails, contacts, and appointments. Theories of sensemaking provide a way to characterize and address the challenges faced by people trying to organize and understand large amounts of network-based data. Sensemaking refers to the iterative process of building up a

Figure 1. Apolo displaying citation network data around the article *The Cost Structure of Sensemaking*. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles (with colored dots underneath) for some ad hoc groups, and instructs Apolo to find more articles relevant to them.

representation or schema of an information space that is useful for achieving the user's goal [31]. For example, a scientist interested in connecting her work to a new domain must build up a mental representation of the existing literature in the new domain to understand and contribute to it.

For the above scientist, she may forage to find papers that she thinks are relevant, and build up a representation of how those papers relate to each other. As she continues to read more papers and realizes her mental model may not well fit the data she may engage in representational shifts to alter her mental model to better match the data [31]. Such representational shifts is a hallmark of insight and problem solving, in which re-representing a problem in a different form can lead to previously unseen connections and solutions [11]. The practical importance of organizing and re-representing information in the sensemaking process of knowledge workers has significant empirical and theoretical support [28].

For the above scientist, she may forage to find papers thatshe thinks are relevant, and build up a representation of howthose papers relate to each other. As she continues to readmore papers and realizes her mental model may not well fitthe data

The cost structure of sensemaking
Russell et al., INTERCHI 1993
This work analyzes sensemaking tasks and develops a model of the cost structure of sensemaking, which has implications for the integrated design of user interfaces, representational tools, and information retrieval systems.
[31]

Mental Leaps: Analogy in Creative

Select a thread

Threads
Collapse all | Expand all

+ 14 Unorganized Papers

— 11 Information Landsc... 1 clip found View details

texSketch: Active Diagramming through Pen-and-Ink Annotations
Subramonyam et al., CHI 2020
Active Diagramming is proposed, an approach extending familiar active reading strategies to the task of diagram construction that combines pen-and-ink interactions with natural language processing to reduce the cost of producing diagrams while maintaining the cognitive effort necessary for comprehension.

Mesh: Scaffolding Comparison Tables for Online Decision Making
Chee Chang et al., UIST 2020
Mesh bridges the gap between decision support systems that typically have rigid structures and the fluid and dynamic process of exploratory search, changing the cost structure to provide increasing payoffs with greater user investment.

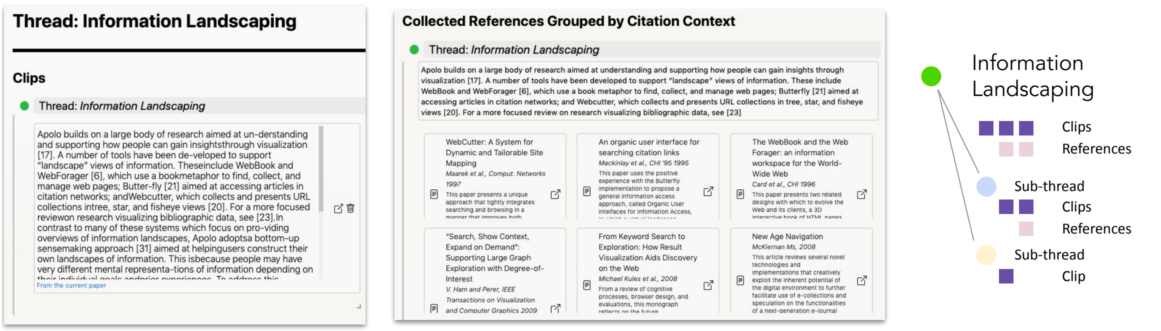
Screen2Vec: Semantic Embedding of GUI Screens and GUI Components
Jia-Jun Li et al., CHI 2021
Screen2Vec is a new self-supervised technique for generating representations in embedding vectors of GUI screens and components that encode all of the above GUI features without requiring manual annotation using the context of user interaction traces.

Learning Representations by Humans, for Humans

Figure 3.2: The design of Threddy consists of two primary panels, (E) PDF viewer & Highlighter and (F) Sidebar. When the user (A) highlights text in the PDF, its content and references are found and temporarily stored in the (B) holding tank. The user can review the content of the holding tank and clean up any errors in the automated extraction and linking of references. When the content looks good, the user can either type in the (C) thread selector to create a new thread, or choose an existing thread. Choosing an option selectively activates buttons for 1) creating a new thread with the references, 2) only adding the extracted references to the chosen thread, and 3) only adding the content of the holding tank as a clip to the chosen thread. Once the user chooses the intended operation, the (D) thread drawer's content is updated to reflect the change. The user can interact with the thread drawer to organize and re-organize its content. The changes are stored and persist across other paper PDFs.

sciences', and '*Systems that augment document margins*'. While these new groupings are helpful, she is not yet confident if the newly added threads and the hierarchy would make sense in light of additional references she would find later. She is also a little worried that the saved references may not provide a good coverage on the topic.

This leads her to click on one of the newly saved thread for a detailed look. In it, she finds a panel that shows additional references which cite several of the curated papers for the thread which seem relevant and useful. In particular, she finds three recent papers that cited a few of the papers she saved for the thread at the top of the panel. Seeing these papers leads her to create another thread with them. This also leads to a change in her mental model around the higher level research thread; she realizes there is a parent-level concept that aptly contains two of the threads she curated until now as its children. She creates it and nests the two threads as its children to better capture her updated mental model. She then collapses the parent thread to declutter the view and focus her attention to the other thread – '*Systems that augment document margins*' – that she has not yet looked at in detail. She finds one of the papers



Overview and Discovery Panel Structure

Thread: Information Landscaping

Clips

Thread: *Information Landscaping*

Apolo builds on a large body of research aimed at un-derstanding and supporting how people can gain insightsthrough visualization [17]. A number of tools have been de-veloped to support "landscape" views of information. Theseinclude WebBook and WebForager [6], which use a bookmetaphor to find, collect, and manage web pages; Butter-fly [21] aimed at accessing articles in citation networks; andWebcutter, which collects and presents URL collections intree, star, and fisheye views [20]. For a more focused reviewon research visualizing bibliographic data, see [23].In contrast to many of these systems which focus on pro-viding overviews of information landscapes, Apolo adoptsa bottom-up sensemaking approach [31] aimed at helpingusers construct their own landscapes of information. This isbecause people may have very different mental representa-tions of information depending on their individual goals andprior experiences. To address this

From the current paper

Collected References Grouped by Citation Context

Thread: *Information Landscaping*

Apolo builds on a large body of research aimed at understanding and supporting how people can gain insights through visualization [17]. A number of tools have been developed to support "landscape" views of information. These include WebBook and WebForager [6], which use a book metaphor to find, collect, and manage web pages; Butterfly [21] aimed at accessing articles in citation networks; and Webcutter, which collects and presents URL collections in tree, star, and fisheye views [20]. For a more focused review on research visualizing bibliographic data, see [23]

WebCutter: A System for Dynamic and Tailorable Site Mapping
Maarek et al., Comput. Networks 1997
This paper presents a unique approach that tightly integrates searching and browsing in a manner that improves both

An organic user interface for searching citation links
Mackinlay et al., CHI '95 1995
This paper uses the positive experience with the Butterfly implementation to propose a general information access approach, called Organic User Inteifaces for Infomiation Access, in which a virtual landscape

The WebBook and the Web Forager: an information workspace for the World-Wide Web
Card et al., CHI 1996
This paper presents two related designs with which to evolve the Web and its clients, a 3D interactive book of HTML pages

"Search, Show Context, Expand on Demand": Supporting Large Graph Exploration with Degree-of-Interest
V. Ham and Perer, IEEE Transactions on Visualization and Computer Graphics 2009

From Keyword Search to Exploration: How Result Visualization Aids Discovery on the Web
Michael Kules et al., 2008
From a review of cognitive processes, browser design, and evaluations, this monograph reflects on the future

New Age Navigation
McKiernan Ms, 2008
This article reviews several novel technologies and implementations that creatively exploit the inherent potential of the digital environment to further facilitate use of e-collections and speculation on the functionalities of a next-generation e-journal

(Top) Thread-specific Clips

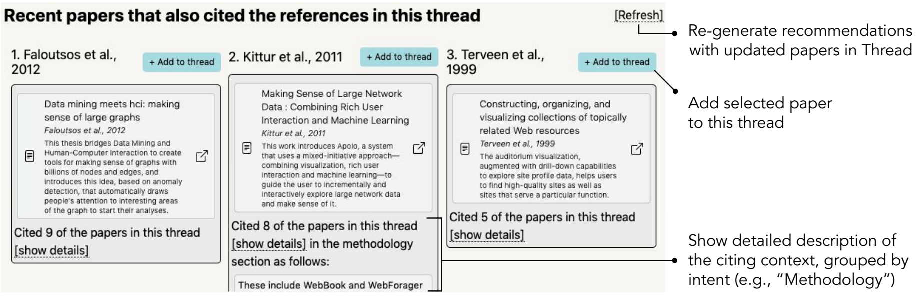
(Middle) References organized by citation context

Hierarchical Organization

Figure 3.3: The Overview and Discovery page consists of three components: At the top of the page is a section for the clips collected for the thread. In the middle of the page are references that belong to the thread grouped by their citation context, and at the bottom of the page (not shown here) is a recommendation panel that contain relevant papers. The overview shows all of the entry thread and its sub-threads' content in a hierarchical manner (indented tree). Readers can choose which thread they want to look at in more details from the sidebar view.

included in the thread particularly interesting, and clicks on it to switch to the PDF. Though her attention shifts towards reading the new paper, she nevertheless maintains her awareness of the current research thread she is interested in from the persisting content of the sidebar view on the right of the new PDF. At the top of the sidebar shows the most recent thread she made changes to, which shows the hierarchy of threads she has been organizing thus far. In the new paper's related work section, she finds interesting new examples of systems from the prior work and the corresponding threads describing how document margins might be augmented in different ways to improve learner discussion, engagement, or comprehension of technical documents. She adds these papers to '*Systems that augment...*' with the context and sub-threads corresponding to the specific ideas described in the paper.

System Design Rationale. Literature review is a complex task that most likely spans long durations, may be interrupted by other tasks [185], and is often initiated and resumed across different device modalities [98]. Therefore, one of the core design rationale for THREDDY was how the context-switching cost may be reduced and the relevant task context such as research threads may be surfaced as the end-user moves from one paper to another. On the one end, active reading interfaces such as *LiquidText* [241] and *texSketch* [**?**] integrates an interactive canvas for note-taking and diagramming to an individual document to support in-depth reading. On the other end of the spectrum are systems for visualizing a collection of documents and their relevance between each other, as discussed in the related work section above. In contrast, THREDDY 's PDF renderer seamlessly replaces the end-user's default PDF reader in their browser, such that they can read papers as they would normally without any additional constraint of having to start reading papers using a new system, all the while collecting relevant papers and structuring them in a form that reflects the user's current mental model of the research space. **Highlighting and Selection.** THREDDY 's main PDF reader is divided into two areas (fig. 4.1): a PDF viewer and highlighter on the left Ⓔ, and a sidebar view holding the thread-related content on the right Ⓕ. The PDF viewer supports text (using mouse drag) and area highlights for images (drag-and-drop while pressing and holding the options/alt key), which trigger THREDDY to extract references included in the highlighted context (extraction is supported only for text highlights). Readers can view the extracted citation context and references in the Ⓑ holding tank in the side bar, and deselect any reference they do not want to include or to fix any extraction error. Readers can add the citation context together with the references as a new thread, or simply add them as a clip or papers to an existing thread using the Ⓒ thread selector. The selector uses the citation context and computes the *thread similarity* (Appendix A) to suggest which thread the highlighted and extracted content most likely belongs to. Threads and references are visually (e.g., different threads use colored dots with numbered counts of nested items on the left; papers use a 'document' icon in place of the colored dots) and organizationally differentiated (e.g., papers show a distinct title - metadata - TL;DR content structure within each card UI). Readers can edit the context or the label in the thread by clicking on the text. Citation context clips are visible only in the Overview and Discovery by default, to prevent clutter. **Organization.** Readers can (re-)organize the threads by drag-and-dropping a thread or papers *into* another thread as a nested thread or *out of* a thread to start a new one. Threads that most recently received an additive change (e.g., a new paper was added to it) are moved to the top. The rationale behind this design choice was that readers may have more organizational needs for the threads that last received content, and/or they are the ones readers most recently attended to and thus are likely to be re-visited. At the top of the thread drawer Ⓓ, the default 'Unorganized Papers' thread is created and each paper PDF opened in the viewer is initially added to it, such that readers can re-visit or organize it at a later time if they wish to (cf. 'deferred actions' interaction design [115]). The paper in the current reader is annotated with a 'current paper' message at the bottom of the corresponding paper card for awareness.



Recent papers that also cited the references in this thread

1. Faloutsos et al., 2012

Data mining meets hci: making sense of large graphs
Faloutsos et al., 2012
This thesis bridges Data Mining and Human-Computer Interaction to create tools for making sense of graphs with billions of nodes and edges, and introduces this idea, based on anomaly detection, that automatically draws people's attention to interesting areas of the graph to start their analyses.

Cited 9 of the papers in this thread
[show details]

2. Kittur et al., 2011

Making Sense of Large Network Data : Combining Rich User Interaction and Machine Learning
Kittur et al., 2011
This work introduces Apolo, a system that uses a mixed-initiative approach—combining visualization, rich user interaction and machine learning—to guide the user to incrementally and interactively explore large network data and make sense of it.

Cited 8 of the papers in this thread
[show details] in the methodology section as follows:

These include WebBook and WebForager

3. Terveen et al., 1999

Constructing, organizing, and visualizing collections of topically related Web resources
Terveen et al., 1999
The auditorium visualization, augmented with drill-down capabilities to explore site profile data, helps users to find high-quality sites as well as sites that serve a particular function.

Cited 5 of the papers in this thread
[show details]

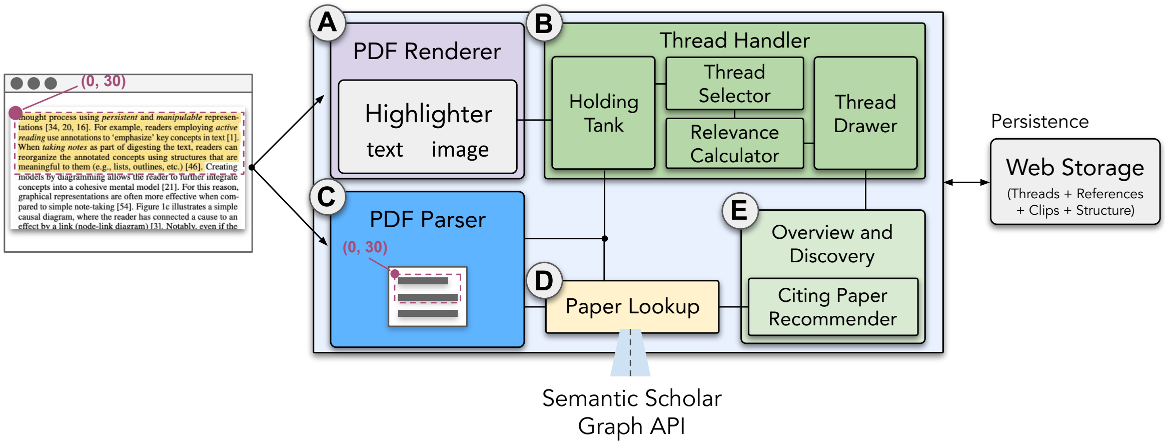
Re-generate recommendations with updated papers in Thread

Add selected paper to this thread

Show detailed description of the citing context, grouped by intent (e.g., "Methodology")

Figure 3.4: The discovery view shows recommendations with high citation coverage, recency, and semantic similarity in a grid. Users can examine the details of each recommendation and decide to add the recommended paper to the current thread. Once the new paper is added, the user can click [Refresh] to generate new recommendations using the updated thread.

Overview, Discovery and Persistence. Clicking on either the 'View Details' or 'Zoom' icon in each thread opens the Overview and Discovery panel for the selected thread (fig. 3.3). The clips that were minimized to prevent clutter in the sidebar view are now visible in full details, along with references grouped by the citation context they were collected from. The panel shows all of the nested threads and their content structured in a hierarchical manner, along with the content of the selected thread. At the bottom of the Overview and Discovery panel are new paper recommendations generated by searching for those that have most cited the papers curated for the selected thread (fig. 3.4). Each recommended paper conveys relevance by showing the number of papers that it cited from the curated, along with the citation context and intent (e.g., in the 'Methodology' section). These more recent papers help users discover newer development on the relevant research threads, akin to *forward chaining* commonly used by scholars conducting literature reviews.



thought process using *persistent* and *manipulable* representations [34, 20, 16]. For example, readers employing *active reading* use annotations to 'emphasize' key concepts in text [1]. When *taking notes* as part of digesting the text, readers can reorganize the annotated concepts using structures that are meaningful to them (e.g., lists, outlines, etc.) [46]. Creating models by diagramming allows the reader to further integrate concepts into a cohesive mental model [21]. For this reason, graphical representations are often more effective when compared to simple note-taking [54]. Figure 1c illustrates a simple causal diagram, where the reader has connected a cause to an effect by a link (node-link diagram) [3]. Notably, even if the

Figure 3.5: Threddy system architecture: (A) PDF Renderer captures and handles user highlight events, (B) Thread Handler fetches the references included in the highlighted text and visualizes them in the holding tank. (C) The parser provides the PDF parse to the thread handler for finding corresponding references in the highlighted text and fetching them via the (D) Paper Lookup module. It also supports (E) the citing paper recommender for user expanded threads. The data is stored in a web storage and persisted as the reader moves to other papers.

3.3 System Architecture

The front-end of Threddy handles user interaction with PDFs, rendering of threads, and the Overview and Discovery panel and is implemented as a Chrome browser extension. The back-end is implemented as a Flask server using GROBID [5] for parsing the PDF content such as the title, section headers, each in-line citation notation to the corresponding reference entry, and body text sentences and their coordinates within the document from its top left-hand side corner.

3.3.1 Automatic Extraction and Linking References to User Highlights

Backend PDF parsing, linking, and mapping user highlight locations. When the user opens a new paper PDF on the browser, Threddy sends the file data to the backend GROBID server for parsing. Our GROBID server uses a cascade of sequence labeling models including a fast linear chain CRF to parse a PDF (see [5] for more details). This server provides full text extraction and structuring of the received PDF, including the overall document segmentation (i.e., locating elements in pixel positions given the scale of the received PDF file) and structuring the text body into sentences, section titles, in-line citation notation to corresponding references (e.g., Whether and which reference '[1]' in a sentence represents), figures, tables, etc. This process is run once when a new paper is opened in Threddy (previously processed PDFs are cached) and takes up to a few 10s of seconds to complete. Additionally, we include the pixel coordinates for each 'sentence' parsed from GROBID, and run an additional sentence parsing using spaCy[1] to merge sentences that may have been erroneously broken. Using this parsed data, we search on the S2ORC [170] and Semantic Scholar APIs[2] to link the corresponding paper with its metadata including the URL, publication year, TL;DR [37], SPECTER embedding [58].

Using the parsed PDF, we first align the scale of the rendered PDF with the PDF used for parsing in

[1] `https://spacy.io/`
[2] `https://www.semanticscholar.org/product/api`

GROBID. When the end-user highlights a portion of the text in PDF, the scale-adjusted coordinates of the highlight location is used to search overlapping sentence coordinates of the parsed PDF. We also collect the surrounding context (i.e., pre- and post-sentences of the overlapping sentence) for clipping. References included in this contextualized selection are searched in the parsed PDF. For image highlights, we simply take a screenshot of the underlying content and convert it into a data url for storage and display.

3.3.2 User Highlights and Creation of Threads

Front-end PDF viewer, Highlighter, and Sidebar. Using the parsed PDF data, we replace the native Chrome PDF viewer with our custom viewer based on an existing highlighter[3] that provides convenient functionality for text and area selection which is a wrapper around the underlying rendering engine based on Mozilla's PDF.js[4]. We feed it with our parsed PDF data and align the rendering scale and user mouse coordinates in accordance with the parsed PDF's coordinate system.

The sidebar consists of multiple components (fig. 4.1). The *holding tank* view at the top visualizes the intermediate content based on user selection. This includes the user highlighted text and the references that are directly in or nearby the highlighted content or an image highlight from the PDF. The references are shown as a list of cards underneath the user highlighted content. Each reference card contains title and additional metadata about the paper, as well as the surface citation notation as shown in the PDF text for ease of mapping. The user may choose to select some but not all of the references automatically extracted to include (clicking on the trash icon discards the selected reference). The *toolbar* underneath the holding tank is selectively activated based on user selection of the thread and the input data. The text input box (fig. 4.1, Ⓒ) allows users to either type in new text (will create a new thread if no matching thread exists), or select one of the existing threads to add the data to. Thread suggestions are generated using an algorithm that first compares user highlighted text and to each vertical chain of the threads to find the most related top-level thread. Next it ranks the most closely related sub-thread within the best chain that the new content may be added to (see Appendix A for details).

Thread interaction. Threads are presented using an interactive nested structure which users can drag-and-drop to nest a thread under another or move it out of the parent thread, delete an existing thread, modify the label of the thread, or collapse/expand all its content (fig. 4.1, Ⓓ). At the top of the thread view is a single-level, unremovable thread titled 'Unorganized Papers' – this thread is automatically generated and adds any paper PDF that is opened by the user under it, allowing the them to defer the action of organizing the papers. This thread is not subject to the user interactions described above other than moving its member references to a different thread (or vice versa).

Clips and references included in each thread are visually differentiated. Clips are not shown to the users by default, but simplified as a simple counter message (e.g., '3 clips found. View details'), upon clicking which opens up an Overview and Discovery panel (Section. 3.3.3) to allow for the end-users to examine further details. In our user study, participants often clipped several textual and image content for each thread; showing all of them in the thread view will clutter it and make it hard to find the threads the user was building on. References, on the other hand, are directly shown as a list of separate cards underneath the thread that they belong to aid immediate access and further exploration. In addition to the content and metadata of the referenced paper, each paper card UI contains a URL icon, which automatically links the reference to its URL on Semantic Scholar. Clicking on the link directs end-users to the paper details page and reduces the amount of context switch they otherwise need in order to find the PDF.

[3] `https://github.com/agentcooper/react-pdf-highlighter`
[4] `https://github.com/mozilla/pdf.js`

3.3.3 Overview and Discovery of Additional Papers Related to Threads

Overview of threads. Once users have created multiple threads that may include sub-threads and relevant references, it can be challenging to review all of the collected and organized content, and to use all of its data to find additional papers in the related literature to further explore and grow the thread. To aid users with this overview and discovery experience, we designed a direct access to a separate panel (fig. 3.3) which opens up when the user clicks either the 'View details' text or the Zoom icon included in each thread card. The panel is expanded to the whole screen width when opened and shows the unrolled view of the selected thread and all of its sub-threads, along with the clips and references collected at each depth. We visualize this using an indented hierarchy to further differentiate threads at different levels (fig. 3.3, right). The panel has three main sections: At the top, user collected clips are shown in a grid with accompanying annotations of the source they were clipped from. Next, a list of references is grouped by their citation context and presented. Finally relevant papers are recommended at the bottom of the page.

Discovering new papers. Using the data collected and organized by each thread to find more relevant papers to further grow the thread could be a challenging task and may significantly interrupt with the end-user's flow of reading. In order to close this discovery loop, we automatically recommend new relevant papers when the Overview and Discovery panel is opened. End-users can then select any of the returned recommendations to add to the thread by clicking on the 'Add to thread' button (fig. 3.4, middle). This adds the paper to the thread in the sidebar as well and keeps it in sync. End-users may click 'Refresh' (fig. 3.4, top right) to re-generate the recommendations based on the updated list of references in the thread, which then invokes the recommender engine.

We use the Semantic Scholar API[5] to fetch necessary paper details for retrieval. Our recommendations use *citation coverage* as its primary source of relevance signal. The rationale behind this decision is that each selected thread contains user-curated relevant papers and the higher the number of thread references cited by a new paper, the higher the chance that it may be relevant to the thread. While we limit our search boundary up to 1,000 direct citations for each thread reference, future work may explore relevance via longer citation chains. For each of the citing paper, we simply count how many of the unique thread references were cited by the new paper. In our pilot tests, we found this to be a good proxy for relevance to the thread's content and return to this in our discussion. We sample a much smaller number (50) of top-ranked results from the top-ranked high-coverage citing papers and sort them by their publication recency. If the two citing papers have the same publication year, we further differentiate them by their semantic similarity computed as the cosine similarity between the centroid vector of the set of thread references included in the thread vs the new citing paper using SPECTER embeddings.

3.4 Evaluation

In evaluation our goal was to study how effectively Threddy supports scholars reading research papers to review the relevant literature in a new domain. To this end, we designed a short literature review task with the goal of producing an outline structure either for themselves in the future or someone else to build upon. We employed a within-subject study comparing Threddy to the commonly used Google Docs editor baseline (without any extensions for searching research papers installed) with research topics that scholars were personally interested in conducting a literature review of. Alternative choices for the baseline comparison may include combining a qualitative coding software such as NVivo with a reference

[5] https://www.semanticscholar.org/product/api

management tool Mendeley[6], or the recently announced Zotero 6[7] browser plug-in which allows users to annotate PDF documents opened in it and create notes that can be exported and imported into the desktop Zotero application. There are pros and cons of choosing each of these alternatives as a baseline for our comparison. However, in our study we decided to use Google Docs because: a) every participant currently uses it or has used it before to conduct literature review; b) it was directly accessible to everyone; and c) it was sufficiently versatile to support creation of research threads, clipping, and adding references to the thread. We return to the choice of the baseline condition in Discussion.

Participants and process. We recruited 9 participants (1 female) for the study. We employed a within-subjects study design, and counterbalanced the order of presentation using 4 Latin Square blocks and randomized rows. Due to the uneven number of recruited participants, the Google Docs-first presentation order was assigned one more time than the Threddy-first order. The mean age of participants was 29.3 (SD: 4.67) and all actively conducted research at the time of the study (1 Master's student, 2, Post-docs, 6 PhD students). Participants' fields of studies included: HCI (5), NLP (2), Material Sciences (2). Participants followed the following process in the study, which took place remotely using Google Meets: introduction and consent, installation of Threddy, two training tasks followed by the main tasks in an individualized order, and surveys. Participants were asked to share their screen during the study. We ended the study with a debrief interview with participants in which the interviewer asked follow-up questions on his observations. The study lasted around 1 hour 20 minutes and participants were compensated at a $30 USD per hour rate.

Training tasks. We used the following paper [137] and using one of its subsections in the Related Work (4 paragraphs) as the seed for practicing creation of an outline. The experimenter described the concepts used in the main task including: 'Threads', 'Clips', and 'References'; Threads are short descriptions of topics or concepts in the related domain and can form a hierarchy with other threads; Clips are supporting pieces of information related to a thread, which can range from a phrase to a paragraph-length text or images directly taken from the paper; References are papers relevant to the thread. Participants were shown a simple example outline and instructed that the outline needed a sufficient amount of details for comprehension and the accessibility of the source. In addition, participants were instructed to read at least one more paper that is relevant and create an outline that can incorporate multiple references in it, starting from the seed paper. Participants were shown a quick tour and core functionality demonstration (5 minutes) followed by a task to recreate the outline they created in a Google Doc using the same text in Threddy.

Timed main tasks. The main tasks used the two topically diverse papers that participants submitted as personally motivating sources for their own literature reviews as part of the sign up process. We randomly assigned each paper to a condition and instructed the participants to start from it as a seed for the task. The tasks were performed for 20 minutes each.

Surveys and interview. For demand (including physical and cognitive) and overall performance we adopt the validated 6-item NASA-TLX scale [108]. For technological compatibility with participants' existing literature review workflows and the easiness of learning we adapted the Technology Acceptance Model survey from [273] (5 items). For measuring the flow aspect [60] of participants' interaction with the system, we adopt Webster et al.'s research [268] uncovering multiple interrelated dimensions of flow in human-computer interaction and the corresponding questionnaire (11 items). Finally, we included 8 addi-

[6] https://tinyurl.com/yckre5md

[7] https://www.zotero.org/blog/zotero-6/

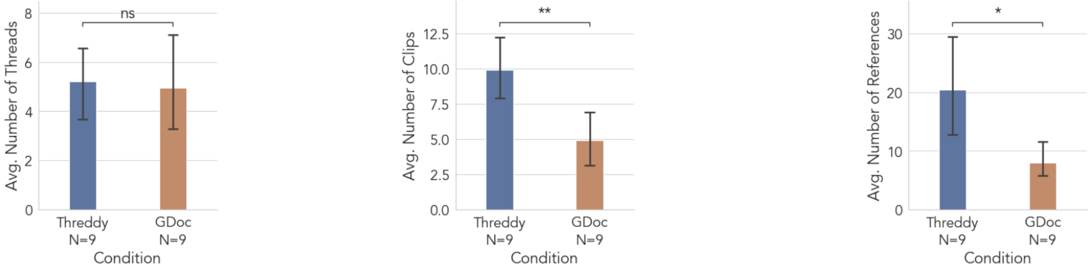


Figure 3.6: Evaluation Results: (a) The number of threads participants created in Threddy ($\mu = 5.2, \text{SD} = 2.28$) and Google Docs ($\mu = 4.9, \text{SD} = 3.03$) did not differ significantly (paired t-test, $t(14.87) = -0.22, p = 0.8$). (b) However, participants added significantly more clips to threads ($\mu = 9.9, \text{SD} = 3.48$) in Threddy vs. Google Docs ($\mu = 4.9, \text{SD} = 3.17$) (paired t-test, $t(15.86) = -3.19, p = 0.006$). (c) Participants also collected and placed a significantly higher number of references into threads ($\mu = 20.4, \text{SD} = 13.84$) in Threddy vs. Google Docs ($\mu = 7.9, \text{SD} = 4.91$) (paired t-test, $t(9.98) = -2.55, p = 0.03$).

tional questions asking participants about extraction of clips and references, as well as their organization into a thread structure (See Appendix B for details of the questionnaire).

Coding. For the baseline, two of the authors coded the first participant's outline together to count the number of threads created, as well as the number of clips and references collected. Then the coders independently coded the rest of the data. The ordinal Krippendorff's alphas were significant for all categories: 0.842, 0.962, 0.822 for threads, clips, and references, respectively. The main sources of disagreement included: whether to count unfinished notes and incomplete text as clips, repeated or slightly modified paper title text as threads, and so on. The final sets of counts for the baseline condition were therefore produced by resolving any disagreements by taking the average between the two coders.

3.5 Findings

3.5.1 Collection and Organization into Threads

Quantitative Results

Participants in both conditions created a similar number of threads ($\mu = 5.2, \text{SD} = 2.28$ in Threddy vs. $\mu = 4.9, \text{SD} = 3.03$ in Google Docs, paired t-test $p = 0.8$). However, the number of clips collected for threads was twice as high in the Threddy condition ($\mu = 9.9, \text{SD} = 3.48$) than in the Google Docs condition ($\mu = 4.9, \text{SD} = 3.17$, $t(15.86) = -3.19, p = 0.006$), demonstrating the utility of Threddy for supporting collection of clips for individual threads. In addition, the number of references collected and organized by the relevant threads was 2.6× higher ($\mu = 20.4, \text{SD} = 13.84$) in the Threddy condition than in the Google Docs condition ($\mu = 7.9, \text{SD} = 4.91, p = 0.03$), further demonstrating its support for efficient collection and organization of references by their relevant threads.

Qualitative Results

Saving time. All of the participants mentioned that automatic extraction of references from the highlighted citation context and linking them with metadata saved time. P1 mentioned that it "*saves a lot of time because I don't have to cross check between the context and the references section*" and similarly P3 said "*Collecting references is such a pain, such a context break... and when I go to the references section* *and finally connect the number to the actual paper, and it turns out that the paper itself doesn't even sound interesting, I need to go back to where I was, with the damage already being done in terms of breaking my reading flow.*" Compared to their experience with Threddy, participants described their typical workflow of conducting literature review as involving lots of "*scrolling back and forth*" and "*pointing and clicking*" (7/9 participants), and having to switch between different applications such as search engines, PDF viewers, and note-taking applications (3/9). Automatic linking to metadata such as the link to the paper details page on Semantic Scholar allowed participants to do without "*having to spend a lot of time to track down PDFs*" (P5, P7). Automatically binding references to their citation context "*reduced a few clicks that would have otherwise been necessary to organize and keep track of that way*" (P9). Uniformly formatting references in the card UI removed the subtask for "*formatting the references, for example it's always a pain to format the text I copied from a PDF... the text is either too large, colored differently, or have weird line breaks, and often times the URLs are way too long*" (P1) and made it easy as to not having to "*worry about all different (surface) citation forms*" (P5). This saved time was thought to be used for reading additional papers or switching to a different thread to explore more (P2).

Context awareness and flow. Participants also described the persistence of threads and pinning most recently worked on threads to the top as effective awareness mechanisms for continuing their thinking along those threads. P9 said: "*Persisting data is helfpul, I don't have to go back to the previous paper to be reminded of what I was thinking of.*" Context clipping was also considered helpful (3/9) and even better with its persistence across papers (P1). P9 thought being able to see the threads while reading any paper was "*a good forcing function to encourage myself to structure and organize as I go.*" Finally, P5 described the benefit of this awareness as follows:

> "Perpetualness of the information is nice because otherwise I have to do this kind of tasks in discrete chunks, just because I'll lose the context a lot in the process because I cannot see where things came from. Or just with a long list of citations and other things... I cannot remember exactly why something that I opened up later is relevant anymore. In comparison, here in the interface I have the continuous stream of cognition, like 'Past Me thought this was relevant to [production of materials] (thread created by P5)."

These results were also corroborated by participants' responses to survey questions. On a 7-point Likert scale (1: Strongly disagree, 7: Strongly agree), participants' agreement was significantly higher with the statement "*It was easy to collect relevant clips using the system.*" in the Threddy condition ($\mu = 6.2, \mathrm{SD} = 0.67$) than in the Google Docs condition ($\mu = 5.2, \mathrm{SD} = 1.09$, $t(13.23) = -3.46, p = 0.01$). Furthermore, participants felt like it was easier "*to keep track of relevant references*" using Threddy ($\mu = 5.9, \mathrm{SD} = 1.48$) than Google Docs ($\mu = 3.8, \mathrm{SD} = 1.27, t(15.63) = -2.80, p = 0.02$), and also easier "*to organize references into relevant threads*" using Threddy ($\mu = 5.6, \mathrm{SD} = 1.13$) than Google Docs ($\mu = 3.4, \mathrm{SD} = 1.33, t(15.58) = -3.59, p = 0.007$). Finally, participants also felt that they were in a heightened flow state while using Threddy based on the results from our survey. Participants' average composite responses to the flow questionnaire items was significantly higher in Threddy ($\mu = 51.0, \mathrm{SD} = 6.5$) than Google Docs ($\mu = 42.9, \mathrm{SD} = 6.77, t(15.97) = -2.94, p = 0.02$). Additionally, the overall demand required for accomplishing the tasks, measured as the sum of the scores to five NASA-TLX questionnaire items (excluding performance), showed no significant difference between the Google Docs ($\mu = 62.6$, SD=25.90) and Threddy conditions ($\mu = 55.3$, SD=20.34) ($t(15.15) = 0.76, p = .52$, Two-sided paired samples t-test; Table 1).

3.5.2 Discovering Papers Relevant to Threads

Participants mentioned that "*it was nice to see recommendations directly relevant to each thread*" (3/9) and directly accessible in the interface (8/9). Many recommendations "*seemed relevant*" (7/9) and especially in the first few rows of the recommendation section. Given their relevance, P6 felt that "*citation coverage as proxy for relevance seems to work well.*" Participants also agreed with the statement "*It was easy to find additional papers relevant to each thread using the system*" significantly more in the Threddy condition ($\mu = 6.1, \mathrm{SD} = 0.78$) than in the Google Docs condition ($\mu = 4.2, \mathrm{SD} = 2.28$, $t(9.86) = -2.35, p = 0.04$). They also felt featuring recommendations in terms of more recent papers that cited the references included in threads was helpful for "*getting a sense of how the field is progressing*" (P3, P5). Interestingly, one participant felt that she experienced "*no fear of missing out (FOMO)*" (P1) given the amount of recommendations available and their overall relevance, but another participant commented that "*I feel a little bit of FOMO because I built a thread which I thought was complete in a sense, but then seeing all these interesting articles made me think, what other things have I missed or simply not shown to me because I decided to organize this thread as such*" (P5).

3.5.3 Extracting Pre-digested Threads, Integrating Papers into Assembled Threads

One challenge with extracting pre-digested threads from other papers was *merged context*, which happened when the author of the paper combined multiple kinds and levels of research contribution into a single patch of text. P1 mentioned that "*(fuzzy) selection is nice, but it sometimes leads to too much... I have to still go through the list of extracted results and 'check off' something that I don't want. For example, these papers are general and maybe related in the loose sense at best... these on the other hand are more specific.*" Similarly P5 commented: "*My main frustration is that people put so many references into a single sentence, and they are not the same. Some of them are more specific and some of them are more general.*" Participants mentioned a need for specifying when they want an exact or pin-pointing selection mechanism that complements the current fuzzy extraction from the highlighted citation context ("*Sometimes I want to point at exactly one cite I want to add from the context.*" – P1), with additional support when a number of references were included in the citation context ("*And for other papers, (citation) styles like [12 – 15] are not great because they (the references) are in a batch, so how am I supposed to know which ones go to which and which one's interesting?"*). While automatically linking the metadata of the extracted references including the title and TL;DRs was helpful to the participants, sometimes they were unavailable due to missing data or provide insufficient context for comprehension specifically related to the citation context of interest (P1, P2, P5).

3.6 Discussion

3.6.1 Impact on the Workflow

Thread-first vs. paper-first exploration. Participants used Threddy to navigate through authors' pre-digested threads, collect ones that were interesting to them, and assembling their own threads using the collected threads. Compared to how they conducted literature review Threddy seemed to unlock a new capability for them to synthesize along the personally interesting threads, across multiple papers, without losing context in the process. P1 contrasted her experience on Threddy with how she currently conducts literature review on Google Docs and described that the former inverses the role between threads and papers in a sense:

> "On Google Docs, I can only focus on one paper at a time. When I find an interesting new paper, I'll skim it and write a short description of why it's interesting. If I find it's worth a more detailed read, I'll try to read it in full, taking notes... like creating my version of an annotated bibliography. In Threddy I can see my threads when I read a new paper, so I almost focus on those (threads) as opposed to individual papers." – P1

This phenomenon of *thread-first* exploration (in contrast to the *paper-first* exploration commonly used by our participants) was also observed in how our participants engaged with actions such as adding new papers to threads, moving references and sub-threads out of their parent thread once enough references with sufficiently different context were identified, and re-labeling and re-framing threads with different text to capture their evolving understanding of the research field. It seemed that Threddy did not create additional workload for users by encouraging them to take the thread-first perspective either; there was no significant difference in terms of the number of threads participants created within the duration of the timed study in each condition nor in terms of their perception of the demand required of them. It is possible, however, that the core of the literature review task that requires complex processing of information (e.g., identifying interesting research threads, summarizing and labeling threads, organizing them in a useful structure, updating the threads) is not helped directly by Threddy. Indeed, participants perceived the easiness of creating threads and subthreads to be roughly equal in both Google Docs and Threddy conditions (see Table 1). Instead, Threddy may help scholars with the task by freeing up their capacity and attention span that would otherwise be tied up by significant auxiliary tasks such as collecting supporting pieces of evidence or relevant context as clips, organizing them into threads, and growing the threads with additional references while keeping track of them. Indeed our participants responded that Threddy provided significant support for such tasks (Table 1).

On-the-go foraging and structuring. At a high level, Threddy users could continuously collect information while following the relevant threads of research. For example, improved context awareness and persistent threads across papers led participants to move between them and "*structure information on the go*" (P9) as opposed to reviewing papers individually and "in discrete chunks due to the frequently lost context" (P5). Moreover, in-thread recommendations integrated the stages for searching and reading, further reducing the context-switching cost. These reduced switching costs and support for externalizing working memory position Threddy to be especially useful in supporting the early sensemaking process of literature review as researchers forage for information, helping them create what Pirolli and Card term a "shoebox" of relevant information [208] easily, in a more organized way, and with affordances for helping them pull in even more relevant information. There are many other aspects of the process where other tools would remain useful. For example, active diagramming and concept mapping may help users externalize representations focused on relations of the concepts involved (e.g., understanding how the different components of a weather system fit together from a paper, rather than the threads of research on weather systems). Synthesis and summarization tools may also help further along the process; here it is possible that a thread-based approach could scaffold the creation of synthesized mental models by enabling users to work with pre-grouped sets of papers.

Checking assumptions via thread-centric recommendations. Participants generally appreciated the recommendations specific to each thread. In addition to the primary benefits of: 1) seeing what is out there, 2) getting a sense of how people are building off of the work curated for each thread, and 3) what may be relevant papers for further exploration along the specific threads, recommendations had secondary, unanticipated benefits as a "check for whether my thread makes sense (by looking at the returned recommendations)" (P3) and as a way to "think about how I might re-define or sub-divide my threads" (P7). Other participants felt additional mechanisms for specifying which context is personally more important for the recommendations, because they felt like "the list of recommendations is quite a spoonful (of papers), some of them are relevant but only at a high level, for example I don't want to see [this paper] just because most of the references in my thread have cited it in the background" (P1). A fruitful avenue for future work therefore may lie in designing alternative mechanisms for finding relevant papers to recommend for each thread that go beyond the simple citation coverage metric explored in this work.

3.6.2 Scaling over a Long Period of Time

An open question with Threddy is how the system would scale over time. With continued use, the number of threads and papers in them would grow significantly. Furthermore, a user's organization would require refactoring as they become more expert in an area; research areas grow and split; or their interests and mental models change. The design of Threddy was directly motivated by issues with scale faced by researchers using other collection tools such as Zotero or Mendeley, with the introduction of hierarchical threads developed in-situ aimed at providing a flexible and scalable way to keep track of diverse topics and subtopics during literature review. Though limited by the duration, our user study uncovered preliminary results speaking to the challenges of scale and time.

First, we found that users desired to focus primarily on a relatively small number of active threads that were most relevant to the context of their target paper. During the post-study interview participants noted that they are often limited in time by deadlines and focused on compiling the most relevant literature for their papers or grants, for example writing a related work section that might include 2-4 topic threads. As participants grew these threads, they noted that their mental models changed with new information, and could use Threddy to refactor the threads appropriately. This included pulling out a particularly dense topic into new subthreads, renaming threads as they learned more about what actually went in them, and using the hierarchy to nest new threads into existing ones and move between them.

However, participants also noted challenges such as not being able to see an overview of all their threads and easily reorganize them within a dedicated workspace. These suggest clear areas for future work including bringing in proven triage and workflow approaches such as tracking which papers have been read or are in different states of processing beyond the simple recency based mechanism introduced here. While these techniques were not core to testing the thread-based idea but they will become essential to a real-world system involving many threads, clips, and references. Another area includes support for working with threads over time, including more intelligent ways to split and merge threads or reorganizing them in the thread hierarchy, which would likely become more important as a user's library grows over time. At an even higher level, there are interesting questions around whether Threddy's hierarchical structure might be improved on by more flexible graph structures, and how such representations could be collaboratively aggregated and built on by others.

3.6.3 Beyond Citation Chaining

Chaining the references in forward and backward directions in time is a common practice used by scientists searching for high relevance papers in the literature and making sense of how the field has progressed over time [266]. However, one potential limitation of citation chaining-based approaches is that it may limit the discoverability of work outside frequently co-cited bodies of literature, and may lead to filter bubbles [189]. Certain domains of knowledge are less likely to interact with each other [56, 238] despite their potential for catalyzing significant scientific innovations [219]. Here, we believe that augmentative tools that help end-users discover articles directly in the context of their flow of reading have potential for helping scholars become more open to the literature that may exist outside the domains they are familiar with but are nonetheless relevant. Additionally, the reduced context switching with the aid of the tool may also help scholars more deeply engage with more distant articles.

In this vein, recent work on discovering analogical scientific literature [140] demonstrated an early evidence of the feasibility of computationally sourcing analogical papers that, although may be missed by conventional search engines, would inspire scientists to come up with novel conceptualization of their research problems (see also [122] for how similar mechanisms of sourcing computational analogies may spark inspirational ideas in a different task context). Complementary approaches may leverage the knowledge domains of papers that the scholar has recently read to automatically increase the frequency of cross-domain retrieval in subsequent recommendations (cf. [139]) or to design a user control for interactively tuning the retrieval domain diversity. At an even higher level, a broader design space for opportunities exists, for example how the system might source recommendations by taking into account the inferred 'social' relevance to frequently read or cited authors, or use such relevance for user engagement and prioritization of recommendations [138].

3.7 Conclusion

In this paper we developed Threddy , a system that supports users with collecting patches of text that contains pre-digested synthesis by other authors (i.e., useful citation context along with automatically extracted associated references), and helps them assemble threads they are personally interested in using clippings of other authors' pre-digested threads included in the introduction or related sections of papers. In contrast with prior work that sought to create separate information environments for similar objectives, Threddy seamlessly integrates into the user's in-situ context of reading, and aims at reducing the cost of context switching while harvesting, assembling, and synthesizing research threads. Further research is required to uncover additional design implications for in-situ reading support for collecting other's synthesis work and assembling them into their own threads.

A Algorithm for Ranking Relevant Existing Threads for Adding New Threads

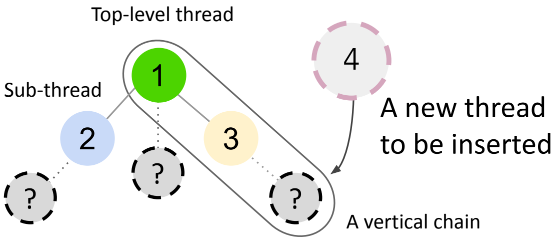


Figure 7: Closely related threads given a new target thread to be inserted are ranked in two stages. In stage one, the vertical chains of threads are grouped together, and a measure of fit that balances the group similarity (i.e., similarity to the centroid of the group) and the maximal member similarity is computed. In stage two, each of the member threads of a vertical chain is compared against to the target thread to further rank them based on similarity which helps the end-user quickly see which thread may be the most relevant for association. When no thread is matching the new thread's content the user can insert it as a new top-level thread.

With the continued use, scholars would likely accumulate a number of threads in the drawer of the interface, leading to a scanning cost that may increase linearly with it at the minimum. In the hope of scaling the usage of Threddy , an algorithm was developed to automatically sort closely related threads that the new thread most likely belongs to in a descending order of their relevance. The first step of this algorithm is finding closely related vertical 'chains' of threads to the to-be-added thread. The intuition here is that end-users most likely nested threads for their semantic relatedness, preserving of which may provide the system a valuable source of signal for discerning the similarity between the new thread and existing chains of threads. Therefore, we first group the members of each chained threads (traversed via the depth-first manner), and compute the similarity between the new target thread and the centroid of the chain. For computing the similarity and the chain centroid, the target thread and each member of the group are embedded into high-dimensional vectors that preserve their multifaceted semantic relatedness. We use the Microsoft's MiniLM model [263] fine-tuned by HuggingFace[8] with 1B+ training pairs, including 116M citation pairs from S2ORC. In our pilot test, this model provided a good trade-off between efficiency and performance for use in our real-time application setting. The chain centroid is computed using a simple average of the member thread embeddings.

However, optimizing only for the similarity to the group centroid runs the risk of finding a chain that although the members' centroid is close to the target thread's embedding, all of its members may be scattered far from one another (i.e., high dispersion, low cohesion). Therefore, we further measure the similarity between the target thread and the closest member thread in the chain and use it to deprioritize matching on such cases:

$$\text{Group Similarity}_{(\text{grp,T})} := \text{sim}\left(\sum_{n \in \text{grp}}^{N} (\overrightarrow{\text{emb(n)}})/N, \overrightarrow{\text{emb(T)}}\right)$$

$$\text{Cohesion}_{(\text{grp,T})} := \max_{n \in \text{grp}} \text{sim}\left(\overrightarrow{\text{emb(n)}}, \overrightarrow{\text{emb(T)}}\right)$$

For a given target thread T, our final rank objective is multiplicative:

$$\text{argmax}_{\text{grp}}\left(\text{Group Similarity}_{\text{grp},T} \times \text{Cohesion}_{\text{grp},T}\right)$$

[8] huggingface.co

to prioritize groups with coherent rather than lopsided similarity (e.g., a high score on only one of Group Similarity or Cohesion but low score on the other may result in an overall irrelevant thread to the user due to the potential situations as described above.). Once we have identified the best chained thread to insert the target thread into, we further rank its member threads in the order of its similarity to the target thread embedding. The resulting ranks of the threads are then presented to the user who may insert the target thread at a particular position of the chain.

B Additional Survey Results

Descriptions of additional questionnaire items and participants' responses grouped by condition are presented in Table 1. Two-sided paired samples t-tests were performed to compute the *p*-values between conditions. See Section 7.1 for a discussion of the results.

C Vignettes of Participants' Threads

The vignettes of threads were simplified by excluding the many clips and references added to each (sub-)thread, and loose yet-to-be organized papers. Note that the structure of threads is subject to change through participants' iteration.

Participant A's vignette of threads, simplified.

- Human-AI collaboration in healthcare
  - Barriers to AI adoption in healthcare
  - Human AI-onboarding
  - ML as second set of eyes
  - Clinical decision support systems
- Mental Model for Decision Making and Errors
- Explainable AI in healthcare

Participant B's vignette of threads, simplified.

- Interpretable model classes and explainability methods
- Usage of GAMs
  - GAMs are widely used to detect patterns of data
- Model interpretability (broadly)
- Explainable Boosting Machine
- GAM empirical studies and results

Participant C's vignette of threads, simplified.

- Table-based decision support tools
  - Sensemaking of collections of online information
- Review Summarization
- Aspect Extraction Methods
- Research on consumer product reviews

	Description	Google Docs	Thready	p-val.
Overall Workload	Sum of the participants' responses to the five NASA-TLX's [108] 21-point scale questionnaire items below.	62.6 (SD=25.90)	55.3 (SD=20.34)	p = .47
Mental Demand	"How mentally demanding was the task?"	10.4 (SD=6.91)	11.6 (SD=5.59)	p = .60
Physical Demand	"How physically demanding was the task?"	12.7 (SD=7.16)	9.3 (SD=4.18)	p = .14
Temporal Demand	"How hurried or rushed was the pace of the task?"	13.1 (SD=6.15)	12.3 (SD=4.47)	p = .64
Effort	"How hard did you have to work to accomplish your level of performance?"	16.2 (SD=3.90)	11.7 (SD=5.39)	p = .15
Frustration	"How insecure, discouraged, irritated, stressed, and annoyed were you?"	10.1 (SD=6.72)	10.4 (SD=5.94)	p = .92
Flow	Sum of the participants' responses to the 11 questionnaire items adopted from Webster et al. [268] measuring the flow aspect of participants' interaction with the system.	42.9 (SD=6.77)	51.0 (SD=6.50)	p = .02*
TAM	Sum of the participants' responses to the 5 questionnaire items adopted from [273] measuring the technological compatibility with participants' existing literature review workflows and the easiness of learning.	28.1 (SD=5.93)	24.7 (SD=5.12)	p = .18
Confidence	"Using the system increased my confidence in conducting literature review. (The response Likert scales for this question and below are 1: *Strongly disagree*, 7: *Strongly agree*)"	4.7 (SD=1.87)	4.8 (SD=1.56)	p = .88
Creating Threads	"It was easy to create different threads in the related literature using the system."	5.6 (SD=1.24)	5.9 (SD=0.93)	p = .54
Creating Subthreads	"It was easy to add sub-threads using the system."	5.4 (SD=1.51)	5.3 (SD=1.73)	p = .83
Collecting Clips	"It was easy to collect relevant clips using the system."	5.2 (SD=1.09)	6.2 (SD=0.67)	p = .001**
Organizing Clips	"It was easy to organize clips into relevant threads using the system."	5.8 (SD=0.97)	5.3 (SD=1.50)	p = .27
Keeping Track of References	"It was easy to keep track of relevant references using the system."	3.8 (SD=1.48)	5.9 (SD=1.27)	p = .02*
Growing Threads	"It was easy to find additional papers relevant to each thread using the system."	4.2 (SD=2.28)	6.1 (SD=0.78)	p = .04*
Organizing References	"It was easy to organize references into relevant threads using the system."	3.4 (SD=1.33)	5.6 (SD=1.13)	p = .007**

Table 1: Descriptions of additional questionnaire items and participants' responses grouped by condition. $p-$values are from two-sided paired samples t-tests. The results suggest that Threddy helps with collecting clips, growing threads by finding more references, and organizing them efficiently. However, Threddy did not seem to decrease the demand of the task or of creating and organizing threads. Furthermore, there was no significant difference in terms of technology compatibility/likelihood of adoption between Google Docs and Threddy, suggesting a familiarity bias favoring Google Docs.

Chapter 4: Synergi

A Mixed-Initiative System for Scholarly Synthesis and Sensemaking

This work was previously published in ACM UIST 2023 ([142]) and has been adapted for this document.

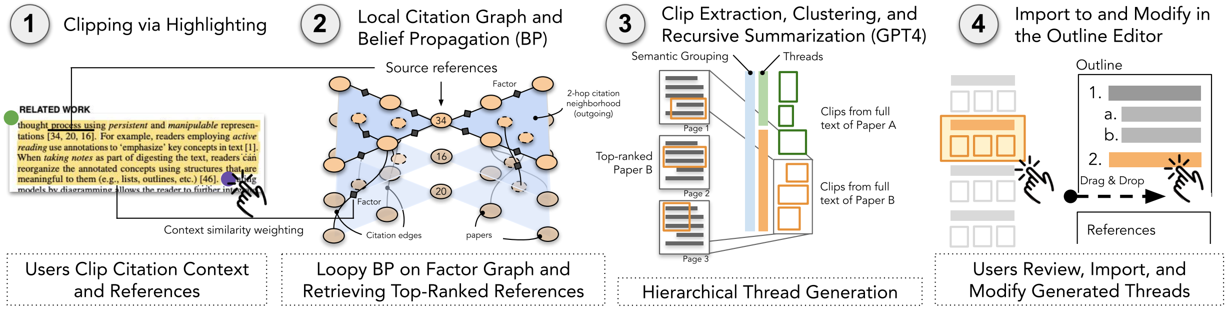
In this chapter, we build on Threddy to introduce Synergi which further utilizes user-curated threads as boundary objects for AI to home in on user interests and relevant other threads. To this end, we explore a mixed-initiative approach to directly support a crosscutting abstraction, *threads*, for scholarly synthesis and sensemaking. Synergi searches important papers relevant to user-curated citation contexts via Belief Propagation over a local citation graph, acquires their full text, hierarchically clusters relevant clips, and synthesizes abstractions using GPT-4. The structured abstractions and clips are then recommended to scholars as prominent threads of research, enabling iterative development and expansion. In our evaluation, we find that Synergi helps scholars efficiently make sense of relevant threads, broaden their perspectives, and increases curiosity. We discuss future design implications for thread-based, mixed-initiative scholarly synthesis support tools.

4.1 Introduction

Scientific and engineering innovations rely on synthesis of prior art: to know what approaches have been tried and identify most promising ideas for new problems; to unlock creative new ideas by combining existing ones; to reason about open challenges and unknown unknowns; and to contextualize one's research in a broader context of literature [154]. At the same time, scholarly synthesis is a cognitively difficult task because it involves many inter-related steps in the process such as discovering the relevant literature about a problem, reading and comprehending papers, collecting useful information and organizing it for further distillation, and recording and monitoring progress by developing an outline that summarizes current learning in the space [278]. Furthermore, scholarly synthesis becomes even more challenged by deepening specialization that makes the barrier of expertise for engaging with the literature higher [20, 110, 114], the accelerating rate of growth [29, 131], and its increasingly interdisciplinary nature [194, 251].

In order to synthesize knowledge scattered across multiple papers, scholars often employ an iterative workflow that involves multiple inter-related stages. This workflow can be characterized by its location on a spectrum of how much of the initiative is automated, between fully *bottom-up* and fully *top-down* workflows. Systems closer to the *bottom-up* end of the spectrum such as Apolo [53] or Threddy [136] allow users to explicitly save an interesting paper or clip, and expand to additional papers and clips. However, users are required to manually save individual clips and papers, making these systems fall short of helping scholars synthesize and distill knowledge after the early stages of discovery and foraging in sensemaking [208]. In contrast, systems near the *top-down* end of the spectrum such as ConnectedPapers[1] and Metro Maps of Science [224] provide users an initial visual overview of the research landscape to help scholars make sense of the structure of knowledge and discover interesting parts in it which can be especially useful for scholars new to a domain. In addition, recent Large Language Models (LLMs)-

[1] `https://www.connectedpapers.com/`


1
Clipping via Highlighting
2
Local Citation Graph and Belief Propagation (BP)
3
Clip Extraction, Clustering, and Recursive Summarization (GPT4)
4
Import to and Modify in the Outline Editor
RELATED WORK
Source references
Factor
2-hop citation neighborhood (outgoing)
34
16
20
Factor
Context similarity weighting
Citation edges
papers
Semantic Grouping
Threads
Clips from full text of Paper A
Top-ranked Paper B
Clips from full text of Paper B
Page 1
Page 2
Page 3
Outline
1.
a.
b.
2.
Drag & Drop
References
Users Clip Citation Context and References
Loopy BP on Factor Graph and Retrieving Top-Ranked References
Hierarchical Thread Generation
Users Review, Import, and Modify Generated Threads

thought process using *persistent* and *manipulable* representations [34, 20, 16]. For example, readers employing *active reading* use annotations to 'emphasize' key concepts in text [1]. When *taking notes* as part of digesting the text, readers can reorganize the annotated concepts using structures that are meaningful to them (e.g., lists, outlines, etc.) [46]. ...aking models by diagramming allows the reader to further int...

Figure 4.1: Main stages of Synergi. (A) A scholar highlights a patch of text in a paper PDF that describes an interesting research problem with references. (B) The system retrieves important papers specifically relevant to the highlighted context in terms of how they have been previously cited by other scholars, via Loopy Belief Propagation over a local 2-hop citation graph from the seed references (Section 4.3.1). (C) Relevant text snippets extracted from top-ranked papers are hierarchically structured and recursively summarized using GPT-4 in the chat interface (Section 4.3.2). (D) The outline of threads, supporting citation contexts, and references are presented to the scholar for importing, modifying, and refactoring in the editor (Section 4.3.3 and 4.3.4).

based systems such as Galactica [242], ChatGPT[2] and Google Bard[3] enable Q&A-based interactions with knowledge domains which users can iteratively query. However, the responses of such systems are similar to visual overviews described above in the sense that they are complete artifacts, rendering them less penetrable and useful for learning, iteration, and synthesis. Furthermore, despite the great potential for augmenting scholarly synthesis workflows, LLMs also suffer from issues of hallucination and falsehood (*cf.* [18, 26, 248]) that render their outputs uncertain, less trustworthy, and needing manual inspection and verification.

Here, we propose a novel mixed initiative workflow, Synergi, that augments scholars' existing synthesis workflows by providing them a structured outline view of research threads, which they can interactively review, curate, and modify. This outline can be iteratively generated to support scholars moving between the *bottom-up* and *top-down* workflows of scholarly synthesis, and help them combine the best of both worlds in the process. Synergi-generated research threads relate specifically to a query clip and seed references, that may match only on a specific citation context within a paper rather than its entirety, and can directly help scholars with making sense of existing threads of research in an area and understanding their relations. Synergi accomplishes this by automatically retrieving a set of important papers from a 2-hop neighborhood on the citation graph and summarizing them in a hierarchical manner with a synthesized label for each parent node that captures the core commonality among its children. In contrast to prior approaches that supported largely manual *bottom-up* synthesis workflows (*e.g.,* Threddy [136] and Apolo [53]), Synergi synthesizes threads from multiple papers and organizes them into a hierarchy that allows users to quickly discover most relevant threads and understand them through synthesis by other scholars, described in the citation contexts in their papers, that are provided together. Furthermore, in contrast to *top-down* LLM-based workflows that may generate difficult-to-inspect black-box outputs, Synergi-generated threads maintain rich provenance and context to help users relate and inspect them further by following up on the source papers and the specific parts in their body text.

Through case studies and a controlled laboratory experiment where domain experts compared the quality of user-generated outlines from Synergi against those of a baseline system based on Threddy and a GPT-

[2] https://chat.openai.com/chat
[3] https://bard.google.com/

4-based approach using the chat interface (henceforth referred to as Chat-GPT4) blind-to-condition, we found that Synergi resulted in the highest overall helpfulness ratings from expert judges. Our quantitative analysis showed that the overall helpfulness of outlines from Synergi was 1.6-point higher compared to Chat-GPT4-generated outlines and 2.6-point higher compared to Threddy-based outlines (on a 7-point Likert scale). In addition, experts judged that threads in the Synergi condition were better-supported with evidence from the literature compared to the Chat-GPT4 condition (+Δ3.3) and the Threddy condition (+Δ2.3; both on a 7-point Likert scale). Through quantitative and qualitative analyses of users' interaction logs, interviews, and responses to experience survey questions, we found that Synergi allowed users to think at a higher level of what existing salient threads of research are and how they divide the space, increased their curiosity in them, and boosted their confidence in conducting a literature review. In addition, we found that these benefits likely came from efficiency gains over a Threddy-based baseline, and also from gains in coverage of synthesis compared to a Chat-GPT4-based baseline. We discuss these results and conclude with design implications for future systems and workflow designs of AI-augmented scholarly synthesis.

In sum, the contributions of this paper include:

- Synergi, a novel mixed-initiative workflow consisting of retrieval and organizational algorithms and interaction features to support scholarly synthesis.
- The results of a controlled laboratory and case studies involving expert judges and detailed quantitative and qualitative analyses of user interaction logs, interviews, and surveys uncovering the benefits and challenges of the approach.
- Implications for future workflow designs and relevant research inquiries in this area.

4.2 Usage Scenario and Design Goals

Motivated by the challenges with existing tools and workflows described in the use scenario described in the introduction of this proposal (Section 1.1), our design goals are as follows:

[D1] When reading one research paper, allow scholars to clip passages and references of interests, and help them find important papers in the domain for synthesis, specific to query context and seed references.

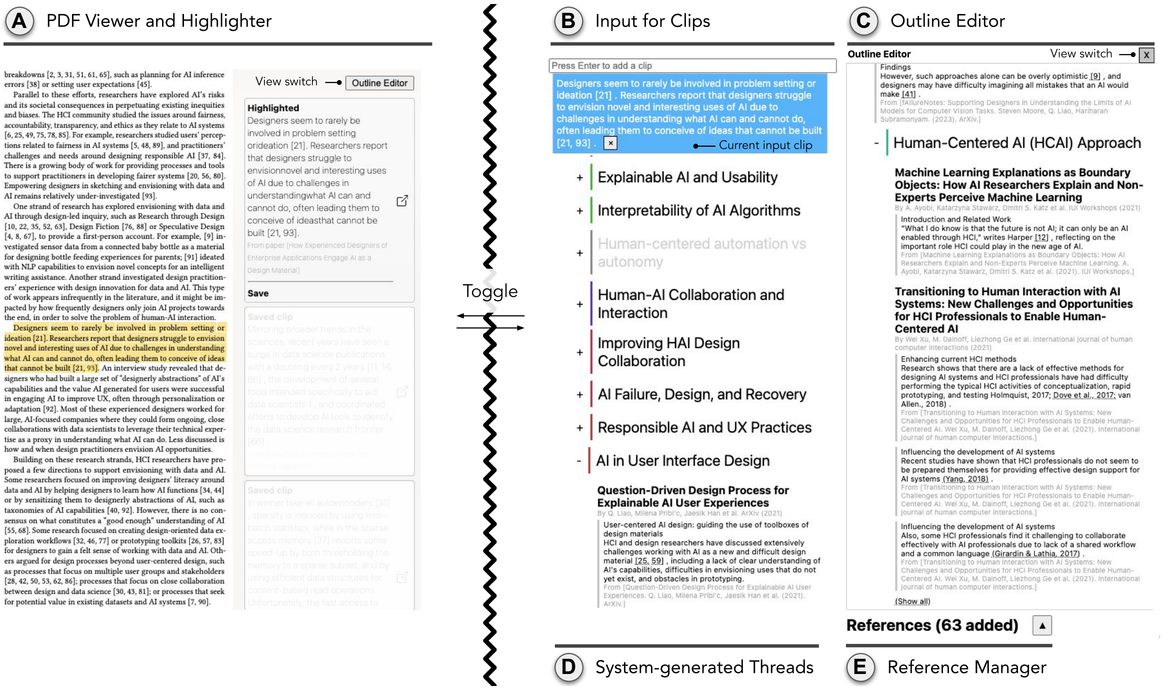
[D2] Based on clips and references collected by a scholar, the system should provide a structured outline of salient research threads to support their synthesis across multiple papers.

[D3] Help scholars understand the specific research contexts described in each thread in detail, and verify their sources.

[D4] Help scholars review the system-generated threads, curate ones that most interest them into their own outline, and iteratively build upon it.

4.3 System Architecture

The system consists of two primary backend algorithms and two sets of interface and interaction features corresponding to the design goals described above.



breakdowns [2, 3, 31, 51, 61, 65], such as planning for AI inference errors [38] or setting user expectations [45].

Parallel to these efforts, researchers have explored AI's risks and its societal consequences in perpetuating existing inequities and biases. The HCI community studied the issues around fairness, accountability, transparency, and ethics as they relate to AI systems [6, 25, 49, 75, 78, 85]. For example, researchers studied users' perceptions related to fairness in AI systems [5, 48, 89], and practitioners' challenges and needs around designing responsible AI [37, 84]. There is a growing body of work for providing processes and tools to support practitioners in developing fairer systems [20, 56, 80]. Empowering designers in sketching and envisioning with data and AI remains relatively under-investigated [93].

One strand of research has explored envisioning with data and AI through design-led inquiry, such as Research through Design [10, 22, 35, 52, 63], Design Fiction [76, 88] or Speculative Design [4, 8, 67], to provide a first-person account. For example, [9] investigated sensor data from a connected baby bottle as a material for designing bottle feeding experiences for parents; [91] ideated with NLP capabilities to envision novel concepts for an intelligent writing assistance. Another strand investigated design practitioners' experience with design innovation for data and AI. This type of work appears infrequently in the literature, and it might be impacted by how frequently designers only join AI projects towards the end, in order to solve the problem of human-AI interaction.

Designers seem to rarely be involved in problem setting or ideation [21]. Researchers report that designers struggle to envision novel and interesting uses of AI due to challenges in understanding what AI can and cannot do, often leading them to conceive of ideas that cannot be built [21, 93]. An interview study revealed that designers who had built a large set of "designerly abstractions" of AI's capabilities and the value AI generated for users were successful in engaging AI to improve UX, often through personalization or adaptation [92]. Most of these experienced designers worked for large, AI-focused companies where they could form ongoing, close collaborations with data scientists to leverage their technical expertise as a proxy in understanding what AI can do. Less discussed is how and when design practitioners envision AI opportunities.

Building on these research strands, HCI researchers have proposed a few directions to support envisioning with data and AI. Some researchers focused on improving designers' literacy around data and AI by helping designers to learn how AI functions [34, 44] or by sensitizing them to designerly abstractions of AI, such as taxonomies of AI capabilities [40, 92]. However, there is no consensus on what constitutes a "good enough" understanding of AI [55, 68]. Some research focused on creating design-oriented data exploration workflows [32, 46, 77] or prototyping toolkits [26, 57, 83] for designers to gain a felt sense of working with data and AI. Others argued for design processes beyond user-centered design, such as processes that focus on multiple user groups and stakeholders [28, 42, 50, 53, 62, 86]; processes that focus on close collaboration between design and data science [30, 43, 81]; or processes that seek for potential value in existing datasets and AI systems [7, 90].

Highlighted
Designers seem to rarely be involved in problem setting orideation [21]. Researchers report that designers struggle to envisionnovel and interesting uses of AI due to challenges in understandingwhat AI can and cannot do, often leading them to conceive of ideasthat cannot be built [21, 93].
From paper [How Experienced Designers of Enterprise Applications Engage AI as a Design Material]

Saved clip
Mirroring broader trends in the sciences, recent years have seen a surge in data science publications with a doubling every 2 years [11, 14, 66] , the development of several tools intended specifically to aid data scientists 1 , and coordinated efforts to develop AI tools to identify the data science research frontier [66] .
From [How Data Scientists Review the Scholarly Literature]

Saved clip
In winner take all autoencoders [31] , sparsity is induced by using mini-batch statistics, while in the sparse access memory [37] reports some speed-up by both thresholding the memory to a sparse subset, and by using efficient data structures for content-based read operations. Unfortunately, the fast access to

Press Enter to add a clip

Designers seem to rarely be involved in problem setting or ideation [21] . Researchers report that designers struggle to envision novel and interesting uses of AI due to challenges in understanding what AI can and cannot do, often leading them to conceive of ideas that cannot be built [21, 93] . ×

+ Explainable AI and Usability
+ Interpretability of AI Algorithms
+ Human-centered automation vs autonomy
+ Human-AI Collaboration and Interaction
+ Improving HAI Design Collaboration
+ AI Failure, Design, and Recovery
+ Responsible AI and UX Practices
- AI in User Interface Design

Question-Driven Design Process for Explainable AI User Experiences
By Q. Liao, Milena Pribi'c, Jaesik Han et al. ArXiv (2021)

User-centered AI design: guiding the use of toolboxes of design materials
HCI and design researchers have discussed extensively challenges working with AI as a new and difficult design material [25, 59] , including a lack of clear understanding of AI's capabilities, difficulties in envisioning uses that do not yet exist, and obstacles in prototyping.
From [Question-Driven Design Process for Explainable AI User Experiences. Q. Liao, Milena Pribi'c, Jaesik Han et al. (2021). ArXiv.]

Outline Editor

Findings
However, such approaches alone can be overly optimistic [9] , and designers may have difficulty imagining all mistakes that an AI would make [41] .
From [fAIlureNotes: Supporting Designers in Understanding the Limits of AI Models for Computer Vision Tasks. Steven Moore, Q. Liao, Hariharan Subramonyam. (2023). ArXiv.]

- Human-Centered AI (HCAI) Approach

Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning
By A. Ayobi, Katarzyna Stawarz, Dmitri S. Katz et al. IUI Workshops (2021)

Introduction and Related Work
"What I do know is that the future is not AI; it can only be an AI enabled through HCI," writes Harper [12] , reflecting on the important role HCI could play in the new age of AI.
From [Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning. A. Ayobi, Katarzyna Stawarz, Dmitri S. Katz et al. (2021). IUI Workshops.]

Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI
By Wei Xu, M. Dainoff, Liezhong Ge et al. International journal of human computer interactions (2021)

Enhancing current HCI methods
Research shows that there are a lack of effective methods for designing AI systems and HCI professionals have had difficulty performing the typical HCI activities of conceptualization, rapid prototyping, and testing Holmquist, 2017; Dove et al., 2017; van Allen., 2018) .
From [Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. Wei Xu, M. Dainoff, Liezhong Ge et al. (2021). International journal of human computer interactions.]

Influencing the development of AI systems
Recent studies have shown that HCI professionals do not seem to be prepared themselves for providing effective design support for AI systems (Yang, 2018) .
From [Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. Wei Xu, M. Dainoff, Liezhong Ge et al. (2021). International journal of human computer interactions.]

Influencing the development of AI systems
Also, some HCI professionals find it challenging to collaborate effectively with AI professionals due to lack of a shared workflow and a common language (Girardin & Lathia, 2017) .
From [Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. Wei Xu, M. Dainoff, Liezhong Ge et al. (2021). International journal of human computer interactions.]

(Show all)

References (63 added)

Figure 4.2: Two main interfaces of Synergi. (A) The PDF viewer and in-text highlighter is similar to that of Threddy [136], with a simplified stream of user-collected clips shown on the right. When the user clicks the 'Outline Editor' button, the view switches to the editor mode. (B) In the top lefthand side corner is an input for user-collected clips where keywords of clips can be typed in to trigger a dropdown menu (not shown). Users can also click on "Try these clips" button to see the most recently saved clips for convenience of their reference (not shown). When the user adds a clip in the input, Synergi kickstarts the pipeline to generate a 3-level hierarch of salient research threads in the literature specific to input clips. (C) Users can interact with the outline editor to curate interesting threads and citation contexts from the hierarchy (Section 4.3.4). (D) Synergi-generated threads and grouped citation contexts are made draggable for user curation into the editor (Section 4.3.3). (E) The reference manager automatically updates upon changes in the editor content (Section 4.3.3).

4.3.1 Retrieving important papers specific to user's query citation context (D1)

Background: Application of Loopy Belief Propagation for Sensemaking

Loopy Belief Propagation (LBP) [274] is a message-passing algorithm well-suited for iterative sensemaking over graphs that may contain cycles. LBP has previously been applied to sensemaking over citation graphs [53] due in part to its favorable qualities such as simultaneously being able to start from multiple entry points on a graph (e.g., multiple references in a user clipped paper passage), and supporting soft clustering (allowing each paper to belong to more than one research topics; see also Related Work in [53] for additional discussions of the algorithm's advantages over alternatives). While LBP on graphs with cycles may risk non-convergence, in practice the risk is extremely low on citation graphs due to the chronological ordering of citation edges leading to broken cycles and weak correlation [7].

Different from Apolo [53], in our workflow users start by specifying input that consists of *the initial set of seed references* as possible exemplars on the citation graph, along with the *citation context described in natural language in which they were referred to*. This setting does not assume user supervision is provided in an iterative manner throughout the process of discovery to prevent propagation of errors.

While previous use of Loopy BP over citation graphs only considered a set of user-provided seed papers to help discover additional papers [53], users in Synergi clips passages and references as they read a paper to discover relevant research threads and papers. To incorporate this additional context (i.e., text passages) into Loopy BP, we introduced a new multiplicative objective for context-sensitive message weighting (See Appendix A.1 for a detailed description), that goes beyond the constant message weighting scheme used in [53]. Intuitively, each component of the new multiplicative message weighting objective corresponds to the context similarity and reference overlap, respectively, optimization of which prioritizes papers that simultaneously meet the conditions of 1) that they are referred to in semantically related ways by other scholars in their literature reviews (typically appear in the introduction and related work sections of the paper) and 2) that they build upon related threads of research, represented by the overlapping set of references that they cited.

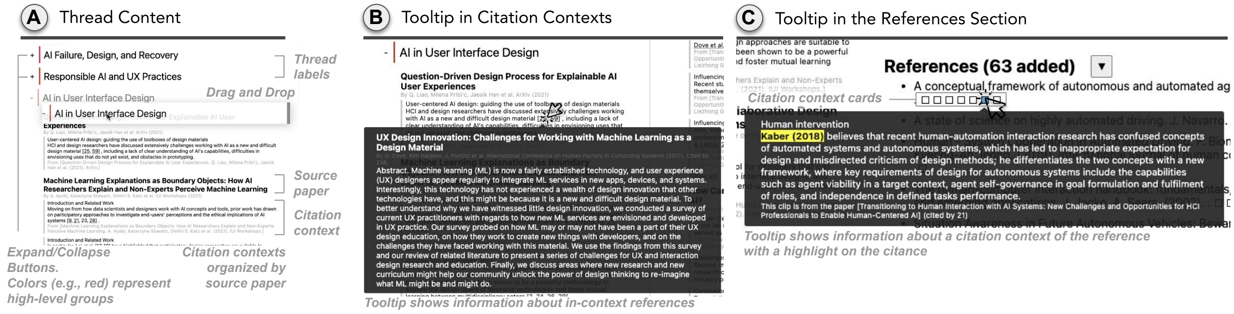
Construction of a factor graph using the 2-hop citation neighborhood

We run the LBP algorithm over the local citation graphs sourced starting from the seed references provided in the user clip. In order to construct a candidate set of papers for retrieval (Fig. 4.1, ➁), the system dynamically fetches the 2-hop citation neighborhood using each of the seed references in both directions (*i.e.,* incoming citations and references) using the Semantic Scholar APIs [149]. For each seed paper referenced in a clip, this allowed Synergi to fetch up to `50` most cited incoming or outgoing citations and `50` references for each hop, resulting a total of `50 * 50 * 2 = 5,000` candidate papers. Once the 2-hop citation neighborhood is retrieved for each seed reference, we construct our factor graph with each unique candidate paper as a variable and use the citation edges as factors connecting the variables. To more deeply consider how each candidate paper is semantically relevant to the user clips, we also retrieve from the APIs information about each candidate papers including the titles and citing contexts. These information were stored as annotations on each edge in the factor graph. Since a paper can be cited by the same paper multiple times in different contexts, each edge may end up with multiple citation context annotations. Furthermore, each variable can be connected to multiple papers that have citation connections with it, allowing Synergi to capture different ways a candidate paper had been characterized by other scholars.

Acquiring and parsing top-ranked paper PDFs

Prior work [136] showed that specific citation contexts and synthesis already provided by other scholars (often appear in the related work or introduction sections of a paper) are useful for scholars' sensemaking and literature review. In order to extract them, we developed a full-text PDF acquisition and parse pipeline. First, we ran the LBP algorithm described above until convergence to find `30` top-ranked papers to search for their full text PDFs. Then the pipeline initially searches the S2ORC corpus [170] to see whether a corresponding full text PDF URL is available for each paper. In cases where a PDF URL was not available in the S2ORC corpus, the pipeline uses the Google Custom Search API[4] to search for a matching paper title and its PDF URL using the "filetype:PDF" constraint. After obtaining a PDF file from the URL, the pipeline uses GROBID [5] to parse the PDF and extract the citation contexts along with metadata (*e.g.,* page number that the citation context appeared on; the header of the section containing the citation context, etc.) and the information of the references included in them to render in tooltips. Finally, if a candidate paper fails to fetch its PDF or be parsed, the pipeline defaulted to the paper title and abstract as its content.

[4] `https://developers.google.com/custom-search/`


A
Thread Content
+ AI Failure, Design, and Recovery
+ Responsible AI and UX Practices
- AI in User Interface Design
- AI in User Interface Design
Thread labels
Drag and Drop
Source paper
Citation context
Expand/Collapse Buttons.
Colors (e.g., red) represent high-level groups
Citation contexts organized by source paper
B
Tooltip in Citation Contexts
- AI in User Interface Design
Tooltip shows information about in-context references
C
Tooltip in the References Section
References (63 added) ▼
Citation context cards
□□□□□■□
Tooltip shows information about a citation context of the reference with a highlight on the citance

Question-Driven Design Process for Explainable AI User Experiences
By Q. Liao, Milena Pribi'c, Jaesik Han et al. ArXiv (2021)

User-centered AI design: guiding the use of toolboxes of design materials
HCI and design researchers have discussed extensively challenges working with AI as a new and difficult design material [25, 59] , including a lack of clear understanding of AI's capabilities, difficulties in envisioning uses that do not yet exist, and obstacles in prototyping.
From [Question-Driven Design Process for Explainable AI User Experiences. Q. Liao, Milena Pribi'c, Jaesik Han et al. (2021). ArXiv.]

Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning
By A. Ayobi, Katarzyna Stawarz, Dmitri S. Katz et al. IUI Workshops (2021)

Introduction and Related Work
Moving on from how data scientists and designers work with AI concepts and tools, prior work has drawn on participatory approaches to investigate end-users' perceptions and the ethical implications of AI systems [9, 21, 23, 28] .
From [Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning. A. Ayobi, Katarzyna Stawarz, Dmitri S. Katz et al. (2021). IUI Workshops.]

Introduction and Related Work

UX Design Innovation: Challenges for Working with Machine Learning as a Design Material
By G. Dove, Kim Halskov, J. Forlizzi et al. International Conference on Human Factors in Computing Systems (2017). Cited by 238.

Abstract. Machine learning (ML) is now a fairly established technology, and user experience (UX) designers appear regularly to integrate ML services in new apps, devices, and systems. Interestingly, this technology has not experienced a wealth of design innovation that other technologies have, and this might be because it is a new and difficult design material. To better understand why we have witnessed little design innovation, we conducted a survey of current UX practitioners with regards to how new ML services are envisioned and developed in UX practice. Our survey probed on how ML may or may not have been a part of their UX design education, on how they work to create new things with developers, and on the challenges they have faced working with this material. We use the findings from this survey and our review of related literature to present a series of challenges for UX and interaction design research and education. Finally, we discuss areas where new research and new curriculum might help our community unlock the power of design thinking to re-imagine what ML might be and might do.

- A conceptual framework of autonomous and automated ag

Human intervention
Kaber (2018) believes that recent human-automation interaction research has confused concepts of automated systems and autonomous systems, which has led to inappropriate expectation for design and misdirected criticism of design methods; he differentiates the two concepts with a new framework, where key requirements of design for autonomous systems include the capabilities such as agent viability in a target context, agent self-governance in goal formulation and fulfilment of roles, and independence in defined tasks performance.
This clip is from the paper [Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI] (cited by 21)

Figure 4.3: Interface features. (A) The thread outline view is organized using an indented tree visualization. Threads and clips are visually differentiated using colors (the latter always featured a grey bar) as well as information organization. The citation contexts for each thread were grouped by the source papers and presented as a list. By default 3 contexts were shown; clicking on a [show more] button at the end of the list expands the list (not shown). (B) Mouse over each reference in a citation context (dotted and underlined for feature visibility) showed a tooltip that contained information about the reference. (C) The reference section at the bottom of the outline editor was automatically updated with each reference featuring citation cards; mouse over on a card showed a tooltip that contained the citation context information.

4.3.2 Generating Salient Threads of Research (D2)

Using the top-ranked papers from previous steps, Synergi generates a structured summary of multiple relevant threads of research in the area (Fig. 4.1, ③). This consisted of steps to home in on specific citation contexts in the papers, structure them into a hierarchy, and summarize them to capture core commonalities among the lower-level components in the hierarchy.

Filtering citation contexts most relevant to seed clips

To synthesize relevant information scattered across the multiple top-ranked papers identified from the retrieval algorithm (Section 4.3.1) into a hierarchical structure using relevant text from them, Synergi embedded the extracted citation contexts using `text-davinci-003`, and filtered those that have a higher average cosine similarity to seed clips than `0.80`.[5]

Agglomerative clustering and tree-cutting

To present the most relevant topical clusters to the users, Synergi first uses the embeddings of the filtered citation contexts to measure how relevant they are to the user clip. For this, Synergi constructs a hierarchical structure from them using a unsupervised agglomerative clustering with the Ward linkage. We perform this using the `fastcluster` package [187]. Agglomerative clustering initializes citation contexts as singleton clusters and computes the ward distance of each pair to successively merge the most similar clusters. The result is a hierarchical binary tree (Fig. 11) where the height of the joint of branches represents the distance at which they were merged (the higher the height of the common ancestor of two leaf nodes on the hierarchy, the more distant they are as neighbors). The resulting binary tree is then converted into a 3-level hierarchy for the user to explore (see Appendix A.2 for a description of the rationale and the method).

[5] determined through a small scaled experiment with five example clips during development.

Recursively summarizing the children clusters

To help users explore the 3-level hierarchy, Synergi synthesizes labels for each parent thread that succinctly describes the underlying threads or citation contexts. In order to synthesize labels that are simultaneously coherent with the underlying children nodes' texts and are abstractions of them, we traverse the hierarchy in a bottom-up manner to recursively synthesize labels. We use Chat-GPT4 with a prompt (Fig. 12 in Appendix A.4) that instructs it to summarize the underlying text using 6 words or less. In each pass on a parent node, up to 25 text snippets from its children were provided during prompting. Therefore, in the first pass the 25 cluster citation contexts were added to the prompt and in successive runs, the text of the children clusters' synthesized labels were used. We also added a post-processing step to merge similar threads (see Appendix A.3 for a description of the rationale and the method) and assigned a unique color to each top-level thread such that the similarity among the children threads could be visually indicated later on the interface.

Finally, the 3-level tree structure with salient threads and their labels, along with the most relevant citation contexts attached to each, are returned to the front-end to render an overview of the relevant research landscape and salient threads in it.

4.3.3 Interface Features (D2 & D3)

Walk-through of the interface

Users on Synergi can highlight and clip relevant citation contexts directly from paper PDFs they are reading. Once they have one or more clips they are interested in investigating further, they switch to the editor view by clicking on the 'Outline Editor' button from the PDF viewer (Fig. 4.2 Ⓐ). In the Outline Editor view, the user can select one or more from the list of saved clips to generate structured research threads related to the citation context and seed references included in selected clips (Fig. 4.2 Ⓑ).

Once the system finishes processing, the structured thread recommendations appear under the clip input (Fig. 4.2 Ⓓ). The user can review the content by scrolling through the list and by expanding/collapsing individual threads which contain the detailed information about citation contexts related to the thread, grouped by source papers. The colored bars on the left also provide users with high-level research areas to quickly orient themselves among the surfaced research areas and help guiding their attention to interesting ones. When the user identifies interesting threads, they can curate them into the outline they are building by dragging and dropping the threads from the list on the lefthand side into the outline editor (Fig. 4.2 Ⓒ), into the appropriate location on the hierarchy. The reference section below automatically updates based on the content changes in the editor, providing the users an easy access to information about papers that have been most cited across multiple threads and citation contexts, which help them prioritize what to read next. The user can continue the cycle by opening up a new paper in the PDF viewer and switching between outline editor. The user data persists for iterative development and refinement.

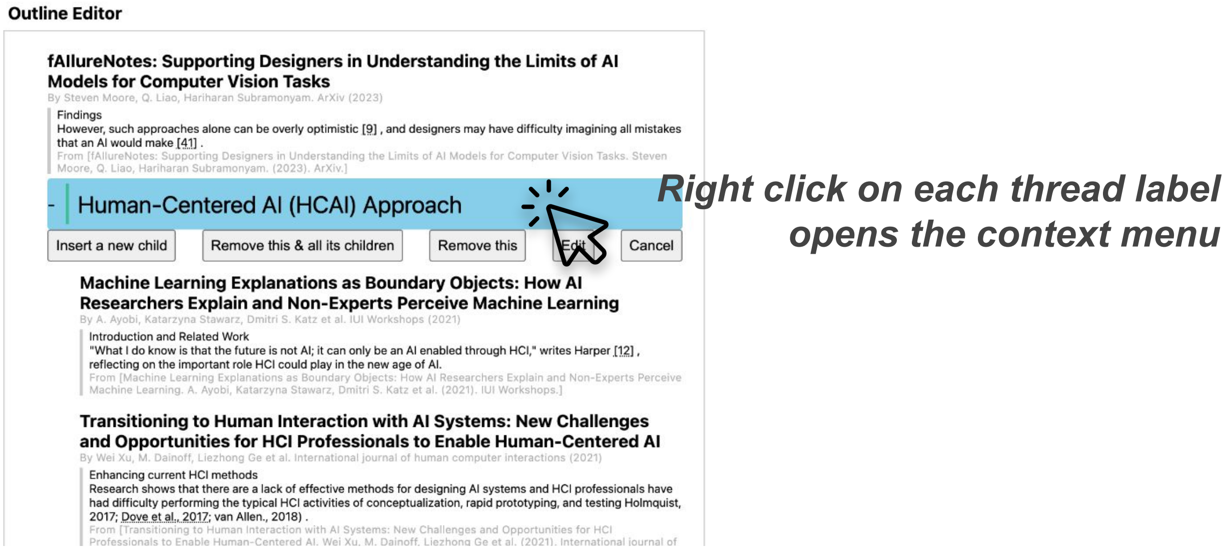
Tree-structured thread recommendations

The tree-structured thread recommendations can be expanded and collapsed to reveal the relevant citation contexts below, which are grouped by source papers (Fig. 4.3 Ⓐ), to provide users with easy access to the source materials and increase the verifiability. Each thread label also featured a color bar on the left to indicate semantically similar groupings among different threads. Each citation context included the specific context found from the paper, the section header that it appeared in, as well as other metadata about the source paper.

Citation context and reference tooltips

To help scholars quickly gain additional information about the cited references in each citation context, each citation notation (*e.g.,* '[4]') was rendered with a dotted underline (Fig. 4.3 Ⓑ), with an additional tooltip that reveals information about the reference such as its title, publication year and venue, number of citations, author names, and the abstract over a mouse-over. In the references section under the outline editor, each referenced paper was automatically updated when the content in the editor changes, and pulled in any citation contexts added in the editor that it was cited in. The grouped citation contexts were shown as squares next to the title (denoted as 'citation context cards' in Fig. 4.3 Ⓒ), which revealed a tooltip that contains information about the citation context with the corresponding reference notation highlighted in the yellow over a mouse-over.

4.3.4 Drag-and-Drop Outline Editor (D4)



Outline Editor

fAllureNotes: Supporting Designers in Understanding the Limits of AI Models for Computer Vision Tasks
By Steven Moore, Q. Liao, Hariharan Subramonyam. ArXiv (2023)

Findings
However, such approaches alone can be overly optimistic [9] , and designers may have difficulty imagining all mistakes that an AI would make [41] .
From [fAllureNotes: Supporting Designers in Understanding the Limits of AI Models for Computer Vision Tasks. Steven Moore, Q. Liao, Hariharan Subramonyam. (2023). ArXiv.]

- Human-Centered AI (HCAI) Approach

Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning
By A. Ayobi, Katarzyna Stawarz, Dmitri S. Katz et al. IUI Workshops (2021)

Introduction and Related Work
"What I do know is that the future is not AI; it can only be an AI enabled through HCI," writes Harper [12] , reflecting on the important role HCI could play in the new age of AI.
From [Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning. A. Ayobi, Katarzyna Stawarz, Dmitri S. Katz et al. (2021). IUI Workshops.]

Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI
By Wei Xu, M. Dainoff, Liezhong Ge et al. International journal of human computer interactions (2021)

Enhancing current HCI methods
Research shows that there are a lack of effective methods for designing AI systems and HCI professionals have had difficulty performing the typical HCI activities of conceptualization, rapid prototyping, and testing Holmquist, 2017; Dove et al., 2017; van Allen., 2018) .
From [Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. Wei Xu, M. Dainoff, Liezhong Ge et al. (2021). International journal of

Right click on each thread label opens the context menu

Figure 4.4: Users could edit the outline either by adding a new thread or citation context into it using drag-and-drop, or by right clicking on each node in the editor.

Threads or individual citation contexts were made draggable into the outline editor. Users could drop the dragged item into any thread node already in the editor or the default top-level thread ('Your Outline'). After the user drops an item to add to the editor, the references section below automatically updated to pull in any new references or new citation contexts for existing references (as shown in Fig. 4.3 Ⓒ). The added threads and citation contexts in the editor were interactive via right-clicking on them at which point the corresponding context menu was revealed. When a thread was right clicked, the following options were shown (Fig. 4.4):

Insert a new child: Add a new nested thread node.

Remove this & all its children: Completely remove the sub-tree rooted on this thread.

Remove this: Remove only the clicked thread and moves all its children one level up (equivalent to merging).

Edit: Edit the label of the thread.

Cancel: Close the menu.

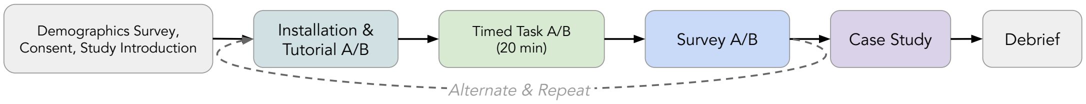

Demographics Survey, Consent, Study Introduction
Installation & Tutorial A/B
Timed Task A/B (20 min)
Survey A/B
Case Study
Debrief
Alternate & Repeat

Figure 4.5: The entire procedure of our study. The order of the middle section of the procedure was swapped based on the assignment (A/B). The order assignment was randomized and counterbalanced across participants (see text).

Right-clicks on citation contexts showed only the 'Remove this', 'Edit', 'Cancel' options in the menu.

4.4 Experimental Design

4.4.1 Objective & Research Questions

Based on a user query as the input, we aimed to study how Synergi-generated threads of research and supporting clips can benefit scholars conducting literature review to cover the broader areas of research. We designed the timed tasks in the experiment to mimic the practice and be coming up with a literature review outline for an assigned topic. This is because scholars often craft intermediary outlines before arriving at a fully written article to structure their thoughts, synthesis, and exploration of the literature in earlier stages. We chose two different topics of research based on the papers that our expert judges were lead authors on [123, 275]. To compare different conditions, we measure the quality of the outlines, the efficiency of constructing them, and the participants' perception of Synergi-generated threads and experience. We operationalized the quality of outlines as experts' judgment of the overall helpfulness, and thread-specific relevance, familiarity, and the goodness of the supporting citation context, on a Likert scale from 1 (Strongly disagree) to 7 (Strongly agree). We operationalized efficiency as the number of threads, clips, and references saved in the outline in a fixed amount of time, as well as the number of user actions taken to construct the outline. Our research questions were:

- RQ1. Does Synergi improve the quality of scholars' literature review outlines over the baselines?
- RQ2. Does Synergi improve the efficiency of outline construction over the baseline?
- RQ3. What are perceived benefits and limitations of Synergi-augmented workflows?

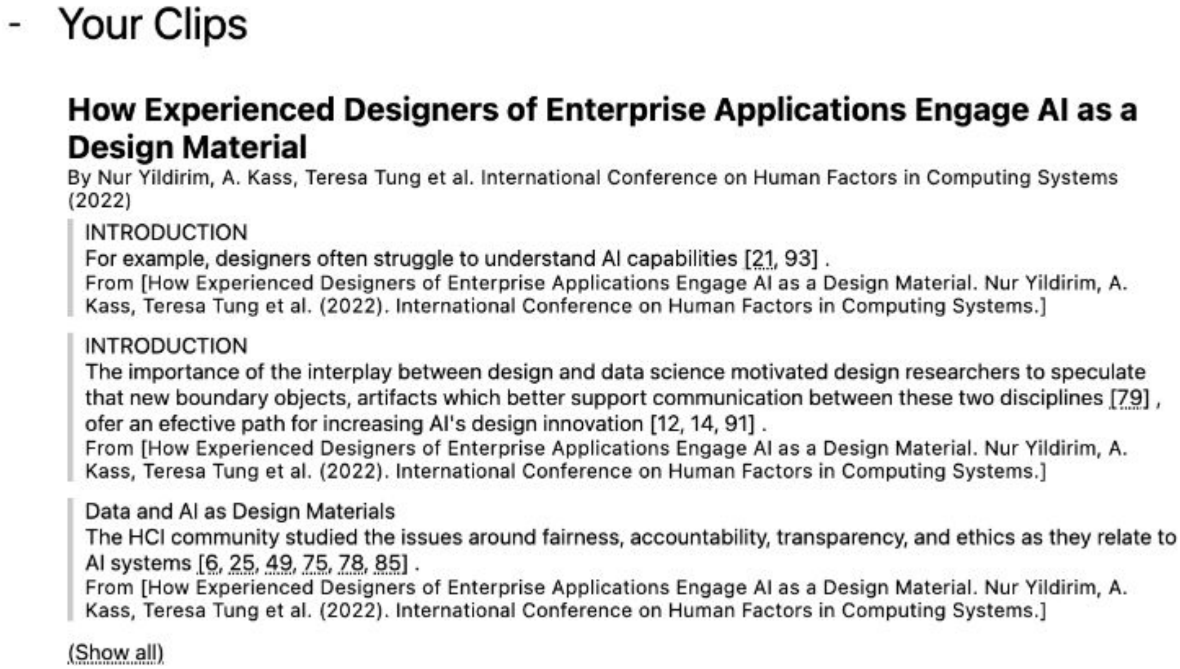
4.4.2 Participants

We recruited 12 participants (2 female) for the study. The mean age of participants was 26.4 (SD: 2.11) and all actively conducted research at the time of the study (9 Ph.D. students and 3 Pre-doctoral Investigators). Participants' fields of studies included (multiple choices): HCI (10), NLP (4), Information Retrieval (1), Cognitive Science (1). We also recruited two experts (both female) to review participants' outlines. Both experts judges were 5th-year Ph.D. students with multiple first-authored and peer-reviewed publications in HCI venues. Their domains of research were 'cross-functional AI teams in envisioning AI products and experiences' and 'designing and building novel tools to help developers better annotate and share their learning materials'. The expert judges spent 1.5 hours to review participants' outlines and were compensated $60 USD. The study lasted for 80 minutes and participants were compensated $40 USD.

4.4.3 Baseline Implementation

Baseline based on Threddy

The baseline system, based on prior work Threddy [136], supported users in manually curating citation contexts via direct in-text highlighting in the PDF, with persisting clips across papers for increased context awareness, and featured a list of user curated clips on the lefthand side of the editor view that users could drag-and-drop into the editor easily. The user-curated clips replaced the system-generated outline provided in the treatment condition. All other interaction features were kept the same.



- Your Clips

How Experienced Designers of Enterprise Applications Engage AI as a Design Material

By Nur Yildirim, A. Kass, Teresa Tung et al. International Conference on Human Factors in Computing Systems (2022)

> INTRODUCTION
> For example, designers often struggle to understand AI capabilities [21, 93] .
> From [How Experienced Designers of Enterprise Applications Engage AI as a Design Material. Nur Yildirim, A. Kass, Teresa Tung et al. (2022). International Conference on Human Factors in Computing Systems.]

> INTRODUCTION
> The importance of the interplay between design and data science motivated design researchers to speculate that new boundary objects, artifacts which better support communication between these two disciplines [79] , ofer an efective path for increasing AI's design innovation [12, 14, 91] .
> From [How Experienced Designers of Enterprise Applications Engage AI as a Design Material. Nur Yildirim, A. Kass, Teresa Tung et al. (2022). International Conference on Human Factors in Computing Systems.]

> Data and AI as Design Materials
> The HCI community studied the issues around fairness, accountability, transparency, and ethics as they relate to AI systems [6, 25, 49, 75, 78, 85] .
> From [How Experienced Designers of Enterprise Applications Engage AI as a Design Material. Nur Yildirim, A. Kass, Teresa Tung et al. (2022). International Conference on Human Factors in Computing Systems.]

(Show all)

Figure 4.6: The baseline system was based on Threddy [136] which supported clipping, persistence of clips across multiple papers, and an easy access to the outline editor where users could organize their own outlines using the self-curated clips.

Baseline based on Chat-GPT4

We also generated two literature review outlines for each paper using Chat-GPT4 on the OpenAI Playground interface[6]. We instructed it to complete a literature review that the user has started and provided the same citation context clip used in the treatment condition for generating an outline, along with an additional label of one thread. We also replaced the citation notations with the actual titles of the cited references, with a clear demarcation to provide further context about the research topic (see Fig. 13 in Appendix B.2). The temperature for Chat-GPT4 was set to 1 with the maximum token length for generation as `2,048`. We repeatedly sampled two outlines for each of the two source paper-clips. These outlines were then manually formatted/blinded (*e.g.,* removing auxiliary characters demarcating headers, reference notations, and unifying the style) for expert review.

[6] `https://platform.openai.com/playground`

4.4.4 Procedure

Structure

We employed a within-subjects study to compare Synergi to a baseline system based on a prior system [136]. We chose two different research areas and topics for timed literature review tasks, and let individual participants choose personally interesting topic/paper for case studies in the end (three tasks in total per participant). We randomly assigned systems to the topics for the timed tasks. We counterbalanced the order of presentation using 6 Latin Square blocks and randomized rows. Participants followed the following procedure in the study, which took place remotely using Zoom: Introduction, Consent, Demographics survey; Installation and Tutorial (detailed in Appendix G.1) of the first system; Main task for the first system; Survey for the first system; Alternate and repeat for the second system; Case Study based on a personally interesting topic; Debrief. Participants were asked to share their screen during the timed tasks and think-aloud during the case studies.

Timed Literature Review Tasks on Pre-defined Topics (20 mins each)

In each of the two timed tasks, participants were instructed to perform a literature review on a randomly assigned topic.The interviewer provided the initial URL to the paper and pointed the participants to the exact location of the clip in each paper that contained the target problem statement. The scenario given to the participants was motivated as 'conducting a review of the relevant literature on behalf of their colleague, who is studying a related research question'.

Post-task Surveys

After each task, participants were administered a survey containing questions on their subjective feelings about the experience. Demand (both physical and cognitive) and overall performance were measured using the validated 6-item NASA-TLX scale [108], where a more compact 7-point scale, mapped to the original 21-point scale, was instrumented [217]. In order to probe the compatibility and adoptibility of the technology with participants' existing literature review workflows, we included a modified Technology Acceptance Model survey from [273] (4 items). Furthermore, 8 types of benefits around discovery, sensemaking, outlining, curiosity, confidence, fear of missing out, and organization of clips and references were measured for each system (See Appendix D for details of the questionnaire).

Data Collection

We collected participant-generated literature review outlines at the end of each timed task. The outlines were then transformed into a spreadsheet while preserving the indentation of the original tree structure with additional columns on the left for experts' judgement. Each tree was traversed to tally the number of threads, clips, and references for each participant for analysis. During the experiment, participant's interaction traces (*i.e.,* timestamped action details during timed tasks) on each system were logged. The details of each timestamped action included a unique user ID, time of the action, the type of the action (*i.e.,* clip, import, create, move, edit, remove, merge), and corresponding details. Participants' think-alouds during the case study and debrief were recorded and transcribed.

Experts' Evaluation

The participant-generated literature review outlines were anonymized and blended with two randomly sampled outlines from Chat GPT-4 for each paper (See Appendix B.2 for the details of the prompts used).

Therefore outlines were generated from three conditions in total, *Baseline* – the Threddy-based baseline system described in Section 4.4.3, *Treatment*, and the *Chat-GPT4*-based baseline (Section 4.4.3). Experts reviewed each outline independently and blind-to-condition, and evaluated on the basis of the following 7-point Likert-scale (1: Strongly disagree, 7: Strongly agree) questions:

- (Overall Outline Helpfulness) "*I found the outline with supporting context helpful for reviewing the relevant literature.*"
- (Thread Familiarity) "*I found the thread of research familiar.*"
- (Thread Relevance) "*I found the thread of research relevant.*"
- (Thread is Well-Supported by Citation Context) "*I found the thread to be well-supported by the specific citation context(s).*"

The overall helpfulness question was evaluated once per participant resulting in 12 data points in *Baseline* and *Treatment* conditions and 4 data points in the *Chat-GPT4* condition; the three thread-level questions were evaluated once per thread per participant, leading to 108 data points (*i.e.,* 31 in *Baseline*; 10 in *Chat-GPT4*; and 67 in *Treatment*) in total.

Case Studies

At the end of the timed tasks, the interviewer asked participants to find and open the PDF of a paper that they were personally interested in that was also in their domain of research using the treatment system. Each participant highlighted and clipped a patch of text (one sentence or longer) that described a particular research problem that also included at least one citation in it, then generated a list of threads using it in the same way as earlier in the timed task. Once the result has returned, the participants were asked to review the generated list of threads, their semantic grouping, the clips, and the references that the clips had originated from. The interviewer then asked questions around their quality, benefits, and limitations.

Data Analysis

The mappings between the research questions and analyses of collected data are as follows.

- RQ1. We analyzed the quality measures of the outlines, which were on a 7-point Likert scale, using non-parametric tests. For expert-evaluated overall helpfulness of outlines, the Wilcoxon's signed rank test was performed for the paired-samples data (*i.e.,* the Baseline vs. Treatment comparison) and the Mann-Whitney U test was performed for the independent data (*i.e.,* the Chat-GPT4 baseline vs. Treatment comparison). For independent data such as thread-level familiarity and relevance, the Mann-Whitney U test was used.
- RQ2. We analyzed the efficiency measures (*e.g.,* the average number of saved threads/clips/references in 20 minutes and the number of user actions taken to construct the outline) between the conditions using paired Student's t-test.
- RQ3. The Likert-scale and Likert-item responses in the survey data were analyzed using the non-parametric paired-samples Wilcoxon's signed rank test. Participants' comments during the case studies were transcribed and qualitatively analyzed using open coding. Participants' interaction logs were visualized as time graphs and used for triangulating relevant survey responses and qualitative data.

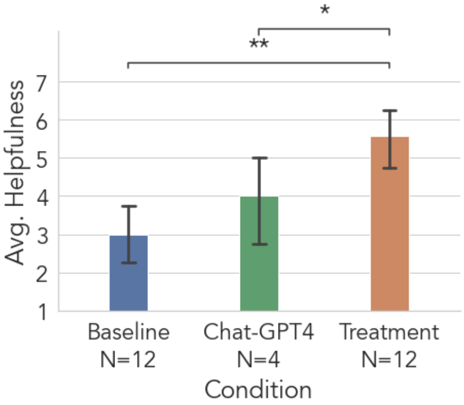


Figure 4.7: The overall helpfulness judged by experts was the highest in Treatment (M=5.6), followed by Chat-GPT4 (M=4.0) and Baseline (M=3.0) conditions. The pairwise differences between Treatment and others were significant (see text).

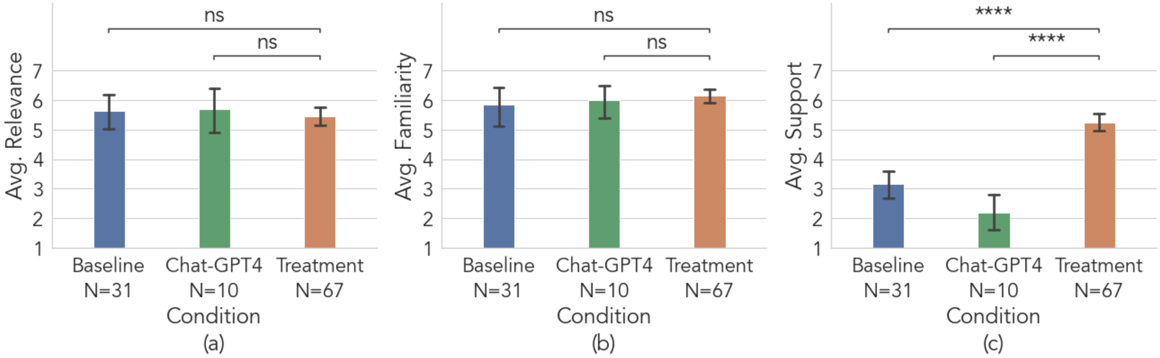


Figure 4.8: Neither (a) average thread relevance nor (b) familiarity significantly differed between the conditions. (c) However, the average goodness of support from relevant citation context differed significantly, as it was judged higher in the Treatment condition (M=5.5, SD=1.25) than in the Chat-GPT4 (M=2.2, SD=1.03) or the Baseline (M=3.2, SD=1.27) conditions.

4.5 Findings

4.5.1 RQ1. Quality of Outlines

Higher quality outlines.

While using SYNERGI , participants were able to generate literature review outlines that were rated as higher quality. The average expert judges' ratings on the overall helpfulness of literature review outlines in the Treatment condition was M=5.6 (SD=1.38), followed by the Chat-GPT4 condition (M=4.0, SD=1.41) and the baseline condition (M=3.0, SD=1.41) (Fig. 4.7). Both differences between the Treatment and the Chat-GPT4 conditions (two-sided Mann-Whitney U=7, p=0.036) and between the Treatment and the Baseline conditions (Wilcoxon W=4, p=0.003) were significant. The experts were blind to the conditions that each of the outlines were generated under.

Improved support while maintaining relevance and familiarity.

We further examined the overall outline helpfulness by comparing between the conditions their component threads' relevance, familiarity, and how well each thread was supported by relevant citation contexts found in the literature (Fig. 4.8). The results showed that the average thread relevance did not differ between the Treatment (M=5.4, SD=1.32) and the Chat-GPT4 (M=5.7, SD=1.34) conditions, nor between the Treatment and the Baseline (M=5.6, SD=1.74) conditions. Similarly, the average thread familiarity between the Treatment (M=6.1, SD=1.03) and the Chat-GPT4 (M=6.0, SD=0.94) conditions did not differ significantly, nor did the difference between the Treatment and the Baseline (M=5.8, SD=1.86) conditions. This suggests that while Synergi considered a large set of 2-hop references and citations (more than 5,000 candidate papers), it is able to maintain high relevance to the user query when presenting related research topics.

Further, the average support each thread received from relevant citation contexts differed significantly. Experts' judgement on the goodness of supporting citation contexts was the highest in the Treatment condition (M=5.5, SD=1.25) and positive (between 'slight' (5) and 'moderate' (6) levels), whereas in the Chat-GPT4 (M=2.2, SD=1.03; two-sided two-sided Mann-Whitney $U = 26, p < .0001$) and the Baseline (M=3.2, SD=1.27; two-sided Mann-Whitney $U = 255, p < .0001$) conditions, it was negative and significantly lower. The goodness of support from relevant citation contexts also seemed to be a differentiating factor of the overall helpfulness of outlines among the conditions; while the relevance and familiarity measures for each thread were highly correlated (Kendall's $\tau = .78, p < .0001$ for Baseline; $\tau = .45, p < .0001$ for Treatment; $\tau = .88, p = .002$ for Chat-GPT4 threads), the only other significant correlations between the support and other measures showed a weak relation (*i.e.,* between relevance and support, $\tau = 0.21, p = 0.04$).

It is notable that despite the lack of supporting citation contexts, both the relevance and familiarity of an average thread generated by Chat-GPT4 tied with those of human-generated threads in the Baseline and Treatment conditions. However, our expert judges noted significant qualitative differences between the Chat-GPT4-generated threads from others, despite not knowing the sources of each outline during the evaluation. The judges proactively offered descriptions of how they differed qualitatively:

> "[One of the Chat-GPT4-generated outlines was] *Probably the most coherent/thoughtful summarization and distillation of the source paper, but most of the stuff seems like something you could just get from reading only that paper and less of a literature review... no citations in any of the points... although the points are reasonable and feel like informed either by my work or other relevant source.*" – E2

> "[After correctly pointing out the two Chat-GPT4-generated outlines] *They seem like maybe someone read over some of the citations in my paper and pulled some points from that, but synthesis is generic. Overall, they are both not great as they don't include citations for the points outlined... Numbered lists in both outlines feel as if they were AI-generated, basically too generic to be useful without citations.*" – E1

4.5.2 RQ2. Outline Construction Process

Synergi showed significant efficiency gains in the outline construction process

The number of research threads, clips, and references saved in the duration of the experiment were all significantly higher in the Treatment than the Baseline condition (Fig. 4.9a – c). For threads, the average

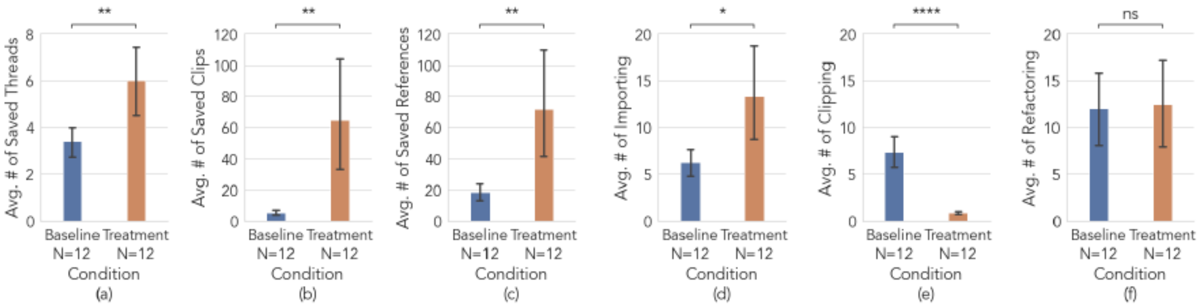


Figure 4.9: The average (a) number of threads, (b) clips, and (c) references saved during the experiment (fixed length) were significantly higher in the Treatment condition than in the Baseline condition. (d) The differences in the saved numbers could be explained by how much more efficiently users in the Treatment condition imported system-generated outputs, rather than (e) spending time in manually clipping the relevant citation contexts, while (f) performing an overall similar amount of refactoring after adding new items to the outline editor.

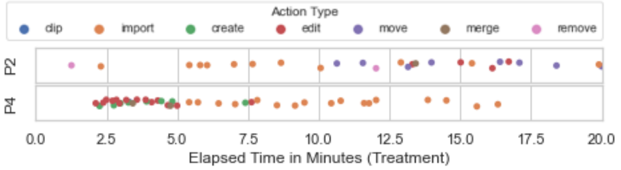


Figure 4.10: (Top) A prototypical time-graph of user actions demonstrating a bottom-up approach of constructing the outline. (Bottom) Same for a top-down construction approach.

number saved was 6.0 (SD=2.76) in the Treatment condition vs. 3.4 (SD=1.16) in the Baseline condition (t_{paired}(14.79)=-2.98, p=0.01). The average number of saved clips was 64.3 (SD=66.27) in the Treatment condition vs. 5.5 (SD=2.81) in the Baseline condition (t_{paired}(11.04)=-3.12, p=0.010). The average number of saved references was also significantly higher in the Treatment (M=71.5, SD=63.40) vs. Baseline (M=18.4, SD=9.62) conditions (t_{paired}(11.51)=2.98, p=0.01).

The higher numbers of saved items in the treatment condition could be explained by the overall higher frequency of 'import' actions that users in the treatment condition performed (Fig. 4.9d) compared to the baseline condition, instead of manually clipping (Fig. 4.9e). On average, the users in the treatment condition performed 13.3 (SD=9.06) imports vs. 6.3 (SD=2.80) in the baseline (t_{paired}(13.08)=-2.75, p=0.019; Fig. 4.9d) and 0.9 clipping (SD=0.29, Treatment) vs. 7.3 (SD=3.20, Baseline; Fig. 4.9e) (t_{paired}(11.18)=7.0, p=0.00002). The overall number of refactoring operations (*i.e.,* moving nodes in the outline editor, editing their labels, merging different thread nodes, removing nodes, creating a new parent thread) did not differ significantly between the two conditions (M=12.4, SD=8.44 in Treatment vs. M=12.0, SD=7.75 in Baseline; Fig. 4.9f, t_{paired}(21.84)=0.17, p=0.87), further suggesting that the efficiency gains originated from replacing the manual clipping of data with examining and importing the system-generated threads and clips in the treatment condition.

Synergi supported both top-down and bottom-up workflows

Interestingly, the users in the Treatment condition exhibited diverging patterns of constructing the outlines. Specifically, some users showed a pattern of top-down construction where they first carefully read through the problem statement and the rest of the source paper to come up with most salient threads of research in their mind before moving on to importing clips that fit those threads, and updating them when a new thread that expands or modifies the initial threads ideated by themselves. Fig. 4.10 (bottom) demonstrates a prototypical action time-graph which shows a densely populated area of refactoring in the beginning (*e.g.,* in the first 5 minutes in the graph) followed by successive importing. In contrast, Fig. 4.10 (top) demonstrates a prototypical time-graph for a bottom-up construction approach. In this case, the participant (P2) first imports a number of system-generated threads and clips onto the editor on the right, then moves on to refactor them (*e.g.,* past the 10 minute mark) to work towards a personally interesting outline.

4.5.3 RQ3. Perceived Benefits and Challenges with Synergi-augmented Workflows

Our quantitative analysis of survey results and qualitative analysis of interviews uncovered different types of benefits from Synergi, such as encouraging participants to gain a higher-level perspective about the literature, think about relations among the threads, and increasing their curiosity. They also uncovered limitations of Synergi-augmented workflows such as additional refinement need related to identifying concepts at a similar level on the conceptual hierarchy, support for probing the relations among threads, and the desire to see explanatory relevance signals for user trust and acceptance.

Reviewing Synergi-generated threads encouraged broader perspectives, sensemaking, and curiosity

Participants commented on how having a list of automatically generated threads of research pushed them to think more broadly about the research space. P1 mentioned that the threads "*help you visualize the literature review outline in your head*" and "*provide better and more context, especially useful for a new topic*" (P1). Relatedly, P4 commented that:

> "*This is giving me a super-power to even begin to think at the level of 'how are different threads of research dividing the space?', which would've been impossible for me to do otherwise.*" – P4

Compared to how they typically conduct a literature in a new domain, they described feeling like saving a lot of time and cognitive effort ("*I usually have to scroll back and forth so many times*" – P2; "*Overhead is significantly reduced... I can now just read, copy-paste, and re-organize stuff*" – P3) that would have otherwise interfered with forming higher-level perspectives. Participants' responses to the survey question: "The system helped me discover relevant threads of research in the literature." also significantly favored the treatment condition (M=6.3, SD=0.75) over the baseline condition (M=3.3, SD=2.14; Wilcoxon W=0, p=0.009). Participants also felt as though the "*colors denoted good groupings of threads, for example this brown (color) shows a group about 'Evaluation of toxicity' which was the core question in our research project.*" (P7) and that "*the thread titles are pretty informative. I could easily tell what I should be paying attention to.*" (P8). Interestingly, P1 commented on how "*it's refreshing to find threads on definitions and studies of 'social capital' that may differ in non-western and global south's regional context of use*" (P1) because manually chasing the citations alone tend to get you "*sucked into*" the "*West-dominant*" perspectives in the literature, since "*asymmetry in the citation behaviors exists between the western and non-western bodies of literature*" – P1.

Furthermore, participants' responses to survey questions: "The system helped me make sense of relevant threads of research in the literature." (M=5.3, SD=1.66 in Treatment vs. M=4.3, SD=2.00 in Baseline, Wilcoxon W=10.5, p=0.088) and "The system helped me outline a review of the literature." (M=6.1, SD=0.67 in Treatment vs. M=5.1, SD=2.02 in Baseline, Wilcoxon W=2, p=0.089) showed marginal significance between the two conditions at $\alpha = .10$.

Participants commented that the list of papers included in the references section of the outline, automatically extracted from the imported clips, was particularly relevant and contained "*inspiring papers to read in this area*" (P7) and one that some participant wanted to take home ("*Can I get a copy of the list on the left?*" – P11). P10 also described how the list "*Matches the threads and references that I curated for my own on-going literature review of the domain, which is good*" (P10). Participants' responses to the survey questions also showed significant preference for the treatment condition over the baseline condition in terms of boosting their curiosity around different threads of research (M=6.0, SD=0.74 in Treatment vs. M=3.9, SD=1.73 in Baseline; Wilcoxon W=2, p=0.01), confidence in conducting the literature review (M=5.8, SD=0.94 in Treatment vs. M=4.0, SD=1.71 in Baseline; Wilcoxon W=2, p=0.01), and in reducing the fear of missing out on important research (M=5.2, SD=1.22 in Treatment vs. M=3.2, SD=1.64 in Baseline; Wilcoxon W=5, p=0.01) (See Appendix D for the details of survey questions).

Trade-offs between Completeness vs. Information Overload

While participants reacted positively towards the initial utility of Synergi in the context of the timed literature review outlining task ("*This is a great starting point for a literature review*" – P10), they also commented on limitations that point to future directions of research in the area. One of the common concerns for longer-term use of Synergi raised by participants related to how to make sense of the quantity of threads presented to them. On the one hand, "having this many, around 20 or so threads would overwhelm me easily" (P10) and especially "seeing similar threads, even though I like how they are grouped together using the same color, could really overwhelm me" (P4). On the other end of the spectrum, seeing a widely varying number of threads returned for queries made P8 wonder if "the result here is complete in this area because I only got 5 threads for this query. Or am I missing something important?" (P8).

Additional Support for Refining and Relating Threads

Participants also commented on how in some cases the variations among the threads within the same high-level color group may be insignificant yet repeated, leading to visual clutter and information overload: "*[Newcomer Integration in OSS Projects] and [Newcomer barriers] are too similar, they can be merged*" (P10); "*[Prompt engineering in NLP models] and [Prompting in Natural Langugage Processing] feel really similar*" (P7). On the other hand, participants also pointed out threads that were seemingly too narrow in scope for them to be at the same level as other threads that seemed to synthesize across multiple papers: "*The [Skip-thought] thread is kind of weird to have be its own category because it's the name of a specific technique from a single paper.*" (P6); "*[Numeric and logical reasoning] is focused on a very specific aspect of the papers in it, which I appreciate but feels too specific to be included in my review.*" (P7).

P4 described how the threads of research helped him 'lift' his perspective going into the literature review task which was beneficial. However, he also described how he was trying to interpret the relations and the order among different threads within each group and between differently colored high-level groups, and how he wished to "*also be able to reason about what the overlapping spaces are between the threads, for example in a 'Venn diagram' of the research space... which is hard to do with a list of threads.*" (P4).

An interesting sub-thread emerged in this topic when participants examined some of the 'and' conjugated threads and found examples where the phrase before and after the 'and' were at different levels of conceptual abstraction. Often the problematic cases featured one concept that felt too broad to be meaningful in relation to the other concept in the thread. For P10, a thread titled '[Augmenting scientific reading] and [machine learning]' was a clear demonstration of how the 'and'-conjugated concepts could appear at different levels of abstraction, with the second concept in this specific example (*i.e.,* machine learning) being too high-level to be useful. Similarly P6 pointed out two examples, '[Text classification] and [feature weighting]' where the first concept was too broad to be meaningful, and '[Image Captioning] and [Computer Vision]' where the second concept "*did not feel like adding useful information*" (P6).

Scaffolding explanatory relevance information for trust and confidence in recommendations

Last but not least, participants wished to see additional information to understand how each thread was generated, and efficient at-a-glance information around which specific aspect in the query each clip is relevant to, in order to boost their confidence and trust in the recommendations. P10 said that:

> "*Understanding the sourcing mechanisms would help me gauge how much trust I should be lending to the system and stay vigilant for potential failure modes, because there are so many different kinds of relations that could be surfaced, for example 'is it (relation) by authors? venues? publication years? topical similarity?' which makes me want to understand more.*"
> – P10

For some, being able to group threads by a given paper was desired for helping orient their sensemaking process. P11 commented that "*In my process I move between papers when conducting a literature review... Here, some of the clips look similar to one another and I can see how the same paper is touching on different threads and I appreciate that the system has added clips from the same paper across multiple relevant papers... but it would be nice to be able to see which other threads that this paper has been added to so that I can quickly decide whether to read that paper in more details.*" (P11). P12 commented that "*It would be helpful if I could see the connections between a thread and each clip in the thread because there are a lot of clips in this thread... and I want to quickly go through them, discarding the ones that look tangentially related.*" (P12).

4.6 Discussion

4.6.1 Summary of Contribution

In this work, we designed and developed SYNERGI as a mixed-initiative system for scholarly synthesis and sensemaking and studied the benefits and challenges of augmenting scholars' workflows for synthesizing knowledge from many papers. In doing so, we built on relevant threads of prior work in Human-Computer Interaction and Natural Language Processing to explore novel algorithmic and interaction designs.

In contrast to prior approaches that were limited to fully manual or fully automated synthesis, our approach is aimed at generating the structure of research threads relevant to specific a query context and seed references that scholars can iteratively review, curate, and build upon. To enable this, SYNERGI first searches important papers by simultaneously considering the specific user query context and seed references over the citation graph through Loopy Belief Propgation, automatically clips useful citation context described by other scholars from their full text, and synthesizes a hierarchical structure using agglomerative clustering and GPT-4. The constrained use and curated input data to GPT-4 limits the risk of hallucination frequent in many *top-down* workflow systems. Furthermore, preservation of rich provenance and contextual information allows scholars to easily examine the structure and its details and check the veracity when in doubt.

Our evaluation task focused on generating a literature review outline which is a common practice in scholarly research and can serve as a useful practice for organizing one's learning about the literature reviewed thus far. Comparing to a baseline system based on prior work [136], and a Chat-GPT4-based prompting approach, we found that SYNERGI improves the expert-judged overall helpfulness of outlines and the participants' efficiency in constructing them. Participants in the study commented how being able to see the SYNERGI-generated threads of research broadened their perspectives and freed their cognitive bandwidth to focus more on higher-level thinking about salient threads of research and their relations. Our results also found implications for future AI-augmented scholarly synthesis workflows, which we discuss further in Section 4.6.3 and 4.6.4.

4.6.2 Broader Scope of Synthesis over Chat-GPT4; Efficiency Gains over THREDDY

Our evaluation results showed that the overall helpfulness ratings of the outlines generated in the treatment condition was significantly higher than that of the baseline or Chat-GPT4 conditions. In comparing to the outlines from the Chat-GPT4 baseline, the expert judges found (blind-to-condition) that Chat-GPT4-generated outlines were surprisingly well-synthesized, distilled key points about the target problem statement, and were "thoughtful". However, they also thought the helpfulness of outlines was significantly limited due to the small scope of its content which felt to have focused only on the source paper alone, and because the threads generated did not have any supporting citation context from other related papers in the literature, unlike the outlines generated in other conditions. These results highlighted the role and value of supporting evidence in scholarly synthesis.

Surprisingly, the average expert-judged familiarity and relevance of threads did not differ between the three conditions, which included both human-generated and Chat-GPT4 threads. This suggested that both human and GPT4-generated threads felt on-topic and exhibited a similar level of linguistic fluency when it comes to a limited scope of literature review, perhaps involving a single source paper. However, the outlines generated from each condition showed significant differences in the scope of synthesis among them, measured by the number of supporting references and citation contexts included for each thread. A much higher level of comprehensiveness was seen in the treatment condition compared to the other two conditions. By examining the outline construction process, we found that participants in the treatment condition gained significantly in their efficiency of foraging and making sense of the space of related research, compared to the baseline condition, which could have allowed them to broaden their scope of synthesis and to incorporate more relevant papers and supporting citation contexts into their outline. Taken together, these findings suggest that while LLMs such as GPT-4 made remarkable advances in appropriately condensing scholarly text, being able to synthesize across multiple papers from the broader literature remains a uniquely human capability today, albeit human scholars may be challenged by limited cognitive bandwidth while performing the cognitive taxing tasks of literature review and synthesis.

4.6.3 Workflows, Cost Structures, User Reliance

Our examination of user interaction logs also revealed two salient behavioral patterns during synthesis around how and when they incorporated the SYNERGI-generated threads into their own outlines which we labeled as *top-down* and *bottom-up* synthesis workflows (Section 4.5.2). In the top-down workflow, users often started by processing the problem statement in more depth compared to the bottom-up workflow, and involved reading broader surrounding contexts in the source paper to distill their understanding into an initial outline of their own. In our evaluation participants using this workflow tended to have more prior knowledge in the broader research area that they could draw upon in creating the initial structure. Once appropriate empty threads in their initial structure were identified, they subsequently imported relevant system-generated threads into them.

In contrast, in the bottom-up process participants often started off by iteratively importing system-generated threads into their editor on an individual thread basis, and creating ad-hoc parent threads when they find commonalities among existing threads. Though lacking initial outline structures, this workflow was popular among the participants most of whom were new to the subject domains used in the experiment. Furthermore, given their access to a readily available list of threads, the cost structure of sensemaking [150] may have shifted such that their reliance on system recommendations was increased to make economic decisions [157]. The contrast in synthesis workflows among the scholars therefore points to interesting future inquiries around their relationships with the cognitive cost structures of scholarly synthesis. Future empirical research in this area has potential to elucidate the factors that can effect changes on behavioral outcomes such as scholars' acceptance and reliance on AI-generated recommendations. Studies may be designed to measure changes in synthesis workflows and outcomes while manipulating participants' prior knowledge in task subjects and introduce different modalities of AI recommendations as well as interaction features that alter the participants' efficiency in verifying or modifying the recommendations in the process. In this context, recent work such as [254] provides helpful exemplar study designs for this domain.

4.6.4 Implications for Future Thread-focused Mixed-Initiative Workflow Designs

Our evaluation also points to design implications for future workflows that are collaborative and AI-augmented. Our expert evaluation showed that fully AI-generated synthesis was competitive against outlines synthesized by human users in a manual or an AI-augmented workflow in coherence and distillation. While fully automated AI-synthesis was limited in its scope – where it seemed to have focused on a single source of paper – which led the judges to believe its utility is significantly reduced, future, more improved LLMs with a sufficiently larger context window may overcome this issue via new capabilities in processing many papers at once.

However, even with such an improved AI, a fully automated workflow may not be the best design for future systems aimed at supporting scholarly synthesis. An important relevant observation here is that 'putting in the work' during the literature review may be critical for scholars' learning and building up the necessary repository of knowledge to successfully perform subsequent synthesis of the domain. Rather than adopting a design that may disincentivize self learning and self-actualization [174], successful mixed-initiative systems therefore would need to consider tasks that AI augmentation can be most beneficial *without* interfering with the core cognitive tasks and human learning processes. This may be hinged on selectively delegating tasks involved in the synthesis based on their high vs. low importance or the core vs. periphery division. For example, scholars may specify a subset of research threads deemed peripheral to be further reviewed and summarized by an AI agent, such that they can efficiently make decisions with respect to whether newly identified threads from the summary merits further exploration and attention from the user, without sacrificing attention and cognitive bandwidth in case they turn out to irrelevant or uninteresting.

4.6.5 Limitations

Though our evaluations uncovered new insights into scholarly synthesis workflows and implications for future mixed initiative synthesis support tools, our experiments focused only on end-to-end evaluations of the entire pipeline. Further ablation studies may be conducted to tease apart contributions from each component in the pipeline (*e.g.,* the retrieval algorithm based on the modified Loopy Belief Propagation algorithm; the algorithm for formation of a thread-based hierarchy; and the recursive summarization method using GPT4). In addition, future evaluation against a baseline that has an expanded prompt context (*e.g.,* using multiple paper body text as input) relative to the prompt we used on GPT-4 in this work would tell us whether GPT-4's synthesis capabilities generalize multiple papers. Furthermore, while our PDF acquisition and parsing modules were performant in the case studies that involved generating outlines for personalized queries, scaling our approach to real-world scenarios with many users may require a significant investment into engineering. A notable example here is how our system aimed to acquire and parse the full text PDFs for important papers, but it relied on best effort (by involving use of commercial APIs such as Google's Custom Search; Section 4.3.1), without a guarantee of coverage. While significant combined research and engineering efforts such as the S2ORC corpus [170] is notable in greatly increasing access to a large paper index with full text PDFs, we note that a significant portion of human knowledge is still locked in non-accessible PDFs, and concerted legal and institutional efforts may be required to make a significant step forward in this domain.

Finally, we believe that future empirical evaluations that go beyond the short duration for studies reported here, and in a more ecologically valid use context (*e.g.,* in a field deployment study rather than a laboratory study) may uncover exciting new opportunities and important challenges in this space.

4.7 Conclusion

In this paper we develop Synergi, a mixed-initiative system that supports scholarly synthesis and sensemaking of the scientific literature. In contrast to prior approaches that cater to either ends of the initiative spectrum (*i.e., bottom-up* or *top-down* workflows), here we develop a a novel approach to help scholars iteratively review the structure of literature related to a specific query context, curate important threads and references, and outline a useful review. Our evaluation that involved 12 participants and domain experts found that Synergi allowed users to create a higher-quality outline of a literature review, compared to a baseline based on the prior system, Threddy [136] and Chat-GPT4. We also found that Synergi achieves this through efficiency gains over the Threddy baseline. Moreover, we show that Synergi increased the coverage of synthesis while also enabling effective curation of supporting evidence from multiple papers over Chat-GPT4. Participants of the user studies found Synergi to be useful in broadening their perspectives about the literature, increasing curiosity while decreasing the fear of missing out on important research in the area. Finally, we conclude with implications for future mixed-initiative workflow designs for scholarly synthesis and interesting inquiries for research in the space. We believe more work is needed in this area to uncover new workflow models on the initiative spectrum and envision improved interactive scholarly systems that would help accelerate scientific innovation for all.

A Detailed System Descriptions

A.1 Loopy Belief Propagation Algorithm in Synergi

Background

The use of LBP in prior work [53] was limited to a scalar conversion weighting of the probability (`0.58`) when messages are exchanged between connected nodes in the graph. In other words, when the user assigns a category to a paper, the papers connected to that via citations would receive messages to increase their marginal probabilities of also being assigned the same category, regardless of the specific citation context. Furthermore, while this simple message weighting is a suitable configuration for interaction scenarios where the user provides iterative supervision over graph nodes (*i.e.,* user assigns a category $c \in C$ for each node n; each node state $s(n) \in \{c, \neg c, \text{not-seen}\}$), which can be used to correct subsequently propagating errors due to insensitivity to diverse citation relations, it is not suitable for our problem setting where no iterative supervision from the user can be supplied during the initial outline generation phase.

In contrast, in our problem setting the user input consists only of *the initial set of seed references* as possible exemplars on the citation graph, along with the *citation context described in natural language in which they were referred to*, without iterative supervision.

Running LBP with context-specific message scaling

In order to prioritize papers that globally optimizes relevance and importance to the user input, we developed a multiplicative message weighting scheme which we assign to each factor in the factor graph to change the marginal probability after each local message passing between the two papers v_i and v_j:

$$\frac{\left(\sum_{s\in S,k\in K} \text{sim}\left(\text{emb}(a_{i,j,k}), \text{emb}(c_s)\right)\right)}{|S \times K|} \times \frac{1}{1 + e^{-\left|\text{ref}(v_i)\cap\text{ref}(v_j)\right|}}$$

where $\{\forall s \in S : c_s\}$ is the set of seed clips, $\{\forall k \in K : a_k\}$ are the annotation texts stored on each edge between paper variable v_i and v_j (*i.e.,* note that $k \geq 1$ because the candidate paper's title text is always available even when no citation context text was found), $\text{sim}(\cdot, \cdot)$ represents the cosine similarity function that takes two embedding vectors as its input, $\text{emb}(\cdot)$ represents a text embedding using the Open AI's `text-davinci-003` model, and $\text{ref}(\cdot)$ represents a function that takes a paper v_i as its input to return the IDs of its referenced papers.

Intuitively, the first component of the multiplication corresponds to the average semantic similarity of possible pairings between the citation contexts in seed clips provided by the user and the citation contexts of the two papers. This is relevant because we are concerned with prioritizing papers with *similarity specific to the query aspect*, rather than the entire paper's topical or thematic similarity to another paper.

The second term of the multiplication corresponds to the degree of overlapping references between the two papers. Intuitively, the higher the number of overlapping references between the two papers, the more likely they would be building on similar threads of research, which can be a useful signal. Similar mechanism of triadic closure has been shown to be capable of surfacing missing friends [6, 225], relevant paper recommendations [138], and author recommendations [141]. However, the effect of a small increase of the count of the overlapping references early on (*e.g.,* consider the effect from a step change $0 \mapsto 1$, in terms of the number of overlapping references between two papers; because there are many more papers that do not share any references, this step change may contain more discriminative information for classification than any other subsequent increases) may exhibit a steeper effect than the same difference at a higher base count of overlapping references. As such, we model the diminishing returns of this signal using the sigmoid function. Finally, the LBP is run until conversion[7].

A.2 From Binary Tree to a 3-level Hierarchy

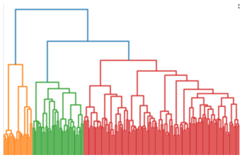


Figure 11: Example hierarchy from agglomerative clustering.

The resulting binary tree from the agglomerative clustering step in the algorithm (Section 4.3.2) may contain within it the high-level hierarchy that resembles the structure that emerges from bottom-up coding of clips via this clustering process. However, in practice each thread in a literature review outline may have more than just two children citation contexts supporting it. For example, in the example binary tree outputted in Fig. 11, the tri-colored branches may correspond well to three distinctive research areas and thus need to be grouped into three semantic categories. Therefore, we condense and re-structure the binary tree in a way that hides the unnecessary complexity arising from the particular clustering method, while preserving the high-level semantic groupings converted, into an 3-level N-ary tree by cutting it at 3 different heights and pruning the branches that form elongated chains.

A.3 Merging Similar Threads

After piloting the synthesized labels of threads (Section 4.3.2), we realized that the conversion of the full binary tree from agglomerative clustering into a 3-level hierarchy may have resulted in sub-groups that have similar citation contexts, that may be better described as a single larger high-level group. Therefore, we introduced a post-processing step that greedily merges parent threads that are highly similar in content from one another, thus reducing redundant sub-groups. We achieved this by using the pairwise cosine similarity of 0.92 as threshold, which was determined from pilot testing.

A.4 Chat-GPT4 Prompt for Label Synthesis

The input prompt to Chat-GPT4 consisted of a system message and a user message (Fig. 12). The outputs were generated using the OpenAI Playground interface[8] in the chat mode using the GPT-4 model. The temperature was set to 0. The content of the user message was infilled with up to 25 citation context text snippets in each cluster.

B Details of the study

B.1 Tutorials

Before participants start with each of the two main task with different conditions, they were given a tutorial of the assigned systems via screen sharing. The interviewer demonstrated a step-by-step installation process and the main features of each system using a prepared script that took around 10 minutes in each condition. In the baseline condition, participants were instructed to clip citances using in-text highlighter

[7] We did not encounter a non-converging case in the user studies.

[8] https://platform.openai.com/playground

```
[System Message]
You are an agent that summarizes scientific articles.
- Follow the user’s requirements carefully & to the letter.
```

```
[User Message]
What is the topic commonly described in the following text snippets?
Summarize the topic succinctly (i.e., 6 words or less).
Reply with "Common topic: " followed by your response.
---
{input documents}
---
```

Figure 12: The prompt used to synthesize labels for each cluster using cluster members ({`input documents`}).

directly in the PDF, and switch between the editor and PDF viewer to organize saved clips into an outline. Participants could search for the PDFs of relevant papers on the Web using any popular search engines and continuously collect relevant clips from them. Participants in the treatment condition were instructed to start by reviewing the Synergi-generated threads and recommended clips to construct an outline.

B.2 Chat-GPT4 Prompt for Literature Review

For the prompt in Fig. 13, the temperature was set to 1 for repeated random sampling. The content of the user message was infilled using the content of each clip used in timed tasks, augmented by the titles of the references included in the clip.

C Detailed User Interaction Logs

A time-graph of user actions in each condition is shown in Fig. 14.

D Full Survey Results

Descriptions of survey items and participants’ responses grouped by condition are presented in Table 1. Two-sided Wilcoxon’s signed rank tests were performed to compute the *p*-values between conditions. See Section 4.5.3 for discussions of the results.

```
[System Message]
You are an assistant to a scientist who's conducting a literature review.
- Follow the user's requirements carefully & to the letter.
```

```
[User Message]
Complete the following survey paper:

Title: Using Annotations for Sensemaking about Code - A Survey

### Code comments are not commonly used for keeping track of facts learned or open
↪ questions

Code comments are commonly utilized for keeping track of open tasks [START_REF]The
↪ emergent structure of development tasks.[END_REF][START_REF]Work Item Tagging:
↪ Communicating Concerns in Collaborative Software Development.[END_REF] and can be
↪ used as navigational aids [START_REF]How Software Developers Use Tagging to
↪ Support Reminding and Refinding.[END_REF][START_REF]Work Item Tagging:
↪ Communicating Concerns in Collaborative Software Development.[END_REF], but are
↪ not commonly used for keeping track of the other previously mentioned information
↪ needs developers have such as facts learned or open questions. This may be
↪ partially because the cost of externalizing this information, especially when the
↪ information may be incorrect, is too high [START_REF]Resumption strategies for
↪ interrupted programming tasks.[END_REF], and these code comments must then be
↪ cleaned up [START_REF]TODO or to bug.[END_REF].

###
```

Figure 13: The prompt used to generate outlines for expert review (showing content for one of the two papers used in timed tasks of the experiment). (Top) The system message component of the prompt. (Bottom) The user message component of the prompt. The temperature was set to 1. The prompt for the first paper in the timed task was similarly constructed, using the clipped citation context with demarcated (*e.g.,* enclosed within each [START_REF]...[END_REF] pair) reference titles.

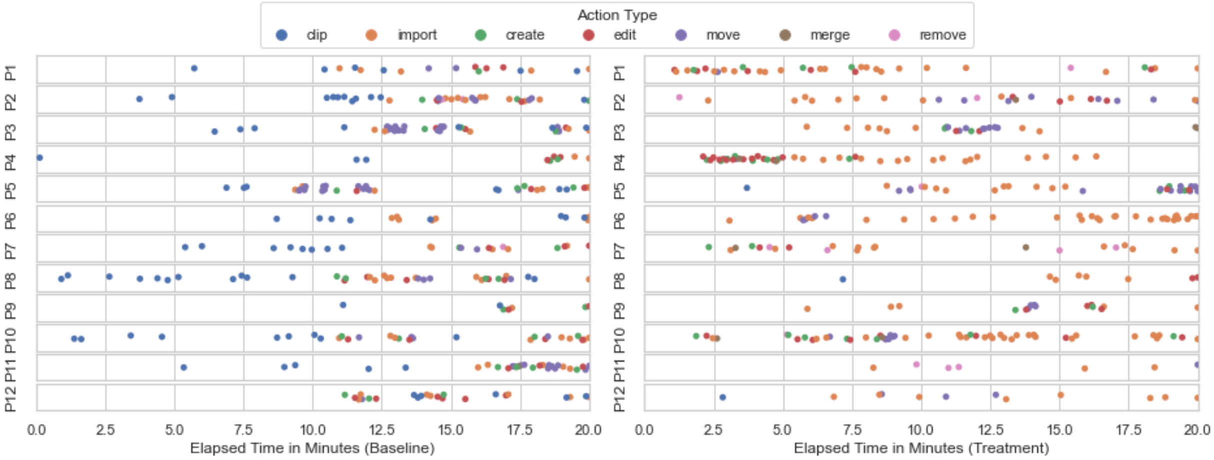


Figure 14: User interaction logs on each system showing the timestamps of seven types of actions.

	Description	Baseline	Synergi	*p*-val.
1. NASA-TLX	Sum of the participants' responses to the five NASA-TLX's [108] Likert-scale questionnaire items below. The original 21-point scale was mapped to a 7-point scale, similarly with [217].	22.3 (SD=6.00)	17.9 (SD=4.19)	.08
1a. Mental	"How mentally demanding was the task?"	4.8 (SD=1.36)	4.3 (SD=1.42)	.34
1b. Physical	"How physically demanding was the task?"	4.6 (SD=1.62)	3.8 (SD=1.47)	.32
1c. Temporal	"How hurried or rushed was the pace of the task?"	5.0 (SD=1.21)	3.5 (SD=1.31)	.003**
1d. Effort	"How hard did you have to work to accomplish your level of performance?"	4.4 (SD=1.44)	4.3 (SD=0.98)	.93
1e. Frustration	"How insecure, discouraged, irritated, stressed, and annoyed were you?"	3.5 (SD=2.11)	2.0 (SD=1.21)	.08
2. TAM	Sum of the participants' responses to the 4 questionnaire items below adopted from [273] measuring the technological compatibility with participants' existing scholarly discovery workflows and the easiness of learning.	19.1 (SD=4.48)	21.0 (SD=5.00)	.06
2a. Compatibility	"*Using the system is compatible with most aspects of how I search for scholars and their papers.*" (The response Likert scales for this question and below are 1: *Strongly disagree*, 7: *Strongly agree*)	4.1 (SD=1.51)	4.8 (SD=1.70)	.33
2b. Fit	"*The system fits well with the way I like to search for scholars and their papers.*"	4.7 (SD=1.83)	4.6 (SD=1.73)	.89
2c. Easy-to-Learn	"*I think learning to use the system is easy.*"	5.8 (SD=1.05)	6.2 (SD=1.02)	.48
2d. Adoption	"*Given that I had access to the system, I predict that I would use it.*"	4.5 (SD=188)	5.4 (SD=1.73)	.15
3. Discovery	"*The system helped me discover relevant threads of research in the literature.*"	3.3 (SD=2.14)	6.3 (SD=0.75)	.009**
4. Sensemaking	"*The system helped me make sense of relevant threads of research in the literature.*"	4.3 (SD=2.00)	5.3 (SD=1.66)	.09
5. Outlining	"*The system helped me outline a review of the literature.*"	5.1 (SD=2.02)	6.1 (SD=0.67)	.09
6. Curiosity	"*The system made me curious about different threads of research in the literature.*"	3.9 (SD=1.73)	6.0 (SD=0.74)	.01*
7. Confidence	"*The system increased my confidence in reviewing the literature.*"	4.0 (SD=1.71)	5.8 (SD=0.94)	.01*
8. Fear of Missing Out	"*The system reduced my fear of missing out on important research.*"	3.2 (SD=1.64)	5.2 (SD=1.22)	.01*
9. Organizing Clips	"*The system helped me organize the clips I found.*"	5.7 (SD=1.15)	5.5 (SD=1.73)	.79
10. Organizing References	"*The system helped me organize the references I found.*"	5.2 (SD=1.59)	5.8 (SD=1.66)	.34

Table 1: Descriptions of full questionnaire items and responses grouped by condition. $p-$values are from two-sided paired samples Wilcoxon's signed rank tests.

Chapter 5: Analogical Search Engine

Augmenting Scientific Creativity with an Analogical Search Engine

This work was previously published in ACM Transactions on Computer-Human Interaction (Volume 29, Issue 6) ([140]) and has been adapted for this document.

While knowing what notable prior approaches exist is useful, the knowledge needs to be further transformed to produce new problem-solving insights. To this end, professionals often need to engage in many cycles of deliberate ideation.

One way to generate novel problem-solving insights is by way of generating analogies, which has been central to creative problem-solving throughout the history of science and technology. As the number of scientific papers continues to increase exponentially, there is a growing opportunity for finding diverse solutions to existing problems. However, realizing this potential requires the development of a means for searching through a large corpus that goes beyond surface matches and simple keywords. In this chapter, we describe the first end-to-end system for analogical search on scientific papers and evaluate its effectiveness with scientists' own problems. Furthermore, we uncover interaction challenges and new design spaces around building an interactive analogy search engine for professional uses.

5.1 Introduction

Analogical reasoning has been central to creative problem solving throughout the history of science and technology [71, 86, 101, 112, 119, 197]. Many important scientific discoveries were driven by analogies: the Greek philosopher Chrysippus made a connection between observable water waves and sound waves; an analogy between bacteria and slot machines helped Salvador Luria advance the theory of bacterial mutation; a pioneering chemist Joseph Priestly suggested charges attract or repel each other with an inverse square force by an analogy to gravity.

Today the potential for finding analogies to accelerate innovation in science and engineering is greater than ever before. As of 2009 fifty million scientific papers had been published, and the number continues to grow at an exceedingly fast rate [29, 65, 131, 192]. These papers represent a potential treasure trove for finding inspirations from distant domains and generating creative solutions to challenging problems.

However, searching analogical inspirations in a large corpus of papers remains a longstanding challenge [76, 88, 176, 231]. Previous systems for retrieving analogies have largely focused on modeling analogical relations in non-scientific domains and/or in limited scopes (e.g., structure-mapping [80, 81, 82, 87, 250], multiconstraint-based [72, 120, 127], connectionist [117], rule-based reasoning [11, 39, 40, 259] systems), and the prohibitive costs of creating highly structured representations prevented hand-crafted systems (e.g., DANE [127, 257]) from having a broad coverage of topics and being deployed for realistic use. Conversely, scalable computational approaches such as keyword or citation based search engines have been limited by a dependence on surface or domain similarity. Such search engines aim to maximize similarity to the query which is useful when trying to know what has been done on the problem in the target domain but less useful when trying to find inspiration outside that domain (for example, for Salvador Luria's queries: "how do bacteria mutate?" or "why are bacterial mutation rates so inconsistent?", similarity maximizing search engines may have found Luria and Delbrück's earlier work on E.coli [171] but may have failed to recognize more distant sources of inspiration such as slot machines as relevant).

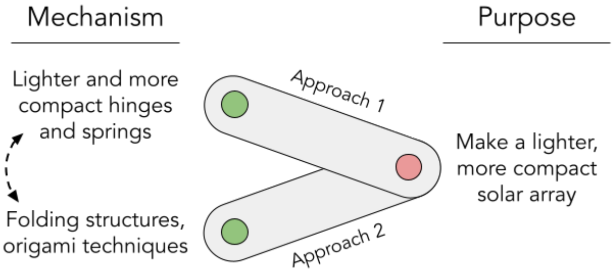

Mechanism
Purpose
Lighter and more compact hinges and springs
Folding structures, origami techniques
Approach 1
Approach 2
Make a lighter, more compact solar array

Figure 5.1: A diagram of two different yet analogical approaches (dashed arrow) for building lighter and more compact solar arrays, and their representations in purposes and mechanisms.

Recently a novel idea for analogical search was introduced [121]. In this idea what would otherwise be a complex analogical relation between products is pared down to just two components: purpose (*what problem does it solve?*) and mechanism (*how does it solve that problem?*). Once many such purpose and mechanism pairs are identified, products that solve a similar problem to the query but using diverse mechanisms are searched to help broaden the searcher's perspective on the problem and boost their creativity for coming up with novel mechanism ideas. Anecdotal evidence suggests that this approach may also be applicable to the domain of scientific research. For example, while building lighter and more compact solar panel arrays has been a longstanding challenge for NASA scientists, recognizing how the ancient art form of origami may be applied to create folding structures led to an innovation to use compliant mechanisms to build not just compact but also self-deployable solar arrays [64, 204, 281] (diagrammatically shown in fig. 5.1). The first remaining challenge of analogical search in the scholarly domain is how we might represent scientific articles as purpose and mechanism pairs at scale and search for those that solve similar purposes using different mechanisms. Recent advances in natural language processing have demonstrated that neural networks that use pre-trained embeddings to encode input text can offer a promising technique to address it. Pre-trained embeddings are real-valued vectors that represent tokens (*Tokenization* means breaking a piece of text into smaller units; *Tokens* can be words, characters, sub-words, or n-grams.), in a high-dimensional space (e.g., typically dimensions of a few dozens to a few thousands) and are shown to capture rich, multi-faceted semantic relations between words [23, 236]. Leveraging them, neural networks may be trained to identify purposes and mechanisms from text [121, 122] to enable search-by-analogy (i.e. different mechanisms used for similar purposes). Once candidate papers are retrieved, searchers may use them to come up with novel classes of mechanisms or apply them directly to their own research problems to improve upon the current state. Prior studies in product ideation showed that users of analogical search systems could engage with the results to engender more novel and relevant ideas [49, 96, 151]. Here, we study the remaining open questions as to whether such findings also generalize to the scientific domains of innovation and how they may differ.

In this paper we present a functioning prototype of an analogical search engine for scientific articles at scale and investigate how such a system can help users explore and adapt distant inspirations. In doing so our system moves beyond manually curated approaches that have limited data (e.g., crowdsourced annotations in [49] with ~2000 papers) and machine learning approaches that have been limited to simple product descriptions [96, 121, 122]. Using the prototypical system, we explore how it enables scientists to interactively search for inspirations for their personalized research problems in a large (~1.7M) paper corpus. We investigate whether scientists can recognize mapping of analogical relations between the results returned from our analogical search engine and their query problems, and use them to come up with novel ideas. The scale of our corpus allows us to probe realistic issues including noise, error, and scale as well as how scientists react to a search engine that does not aim to provide only the most similar results to their query.

In order to accomplish these goals we describe how we address several technical issues in the design of an interactive-speed analogical search engine, ranging from developing a machine learning model for extracting purposes and mechanisms in scientific text at a token level granularity, the pipeline for constructing a *similarity space* of purpose embeddings, and enabling these embeddings to be queried at interactive speeds by end users through a search interface. We construct the similarity space by putting semantically related purpose embeddings in close indices from each other such that related purposes can be searched at scale.

In addition to the technical challenges there are several important questions around the design of analogical search engines that we explore here. A core conceptual difference that distinguishes analogical search engines from other kinds is that the analogs they find for a search query need to maintain some kind of distance from the query, rather than simply maximizing the similarity with it. However, only certain kinds of distance may support generative ideation while others have a detrimental effect. Another question remains as to how much distance is appropriate when it comes to finding analogical inspirations in other domains. While landmark studies of analogical innovation suggest that highly distant domains can provide particularly novel or transformative innovations [93, 95, 113], recent work suggests the question may be more nuanced and that intermediate levels of distance may be fruitful for finding ideas that are close enough to be relevant but sufficiently distant to be unfamiliar and spur creative adaptation [47, 83, 99]. Using a concrete example from one of our participants who studied ways to facilitate heat transfer in semiconductors, a keyword search engine might find commonly used mechanisms appropriate for direct application (e.g., tweaking the composition of the material) while an analogical search engine might find similar problems in more distant domains which suggest mechanisms that inspire creative adaptation (e.g., nanoscale fins that absorb heat and convert it to mechanical energy). Though more distant conceptual combinations may not always lead to immediately feasible or useful ideas, they may result in outsized value after being iterated on [24, 46, 153].

In the following sections we explore the technical and design challenges for an analogical search engine and how users interact with such a system. First, we describe the development of a human-in-the-loop search engine prototype, in which most elements of the system are functional but human screeners are used to remove obvious noise from the end results in order to maximize our ability to probe how users interact with potentially useful analogical inspirations. Using this prototype we characterize how researchers searching for inspirations for their own problems gain the most benefit from papers that partially match their problem (i.e., match at a high level purpose but mismatch at a lower level specifications of the purpose), and that the benefits are driven not by direct application of the ideas in the paper but by creative adaptation of those ideas to their target domain. Subsequently we describe improvements to the system to enable a fully automated, interactive-speed prototype and case studies with researchers using the system in a realistic way involving reformulation of their queries and self-driven attention to the results. We synthesize the findings of the two studies into design implications for next-generation analogical search engines.

Through extensive in-depth evaluations using an ideation think-aloud protocol [79, 253] with PhD-level researchers working on their own problems, we evaluate the degree to which inspirations spark creative adaptation ideas in a realistic way on scientists' own research problems. Unlike previous work which has often used undergraduate students in the classroom or lab [257], and often evaluated systems on pre-determined problems [84], this study design provides our evaluation with a high degree of external validity and allows us to deeply understand the ways in which encountering our results can engender new ideas. Our final, automated search engine demonstrates how the human-in-the-loop filtering can be removed

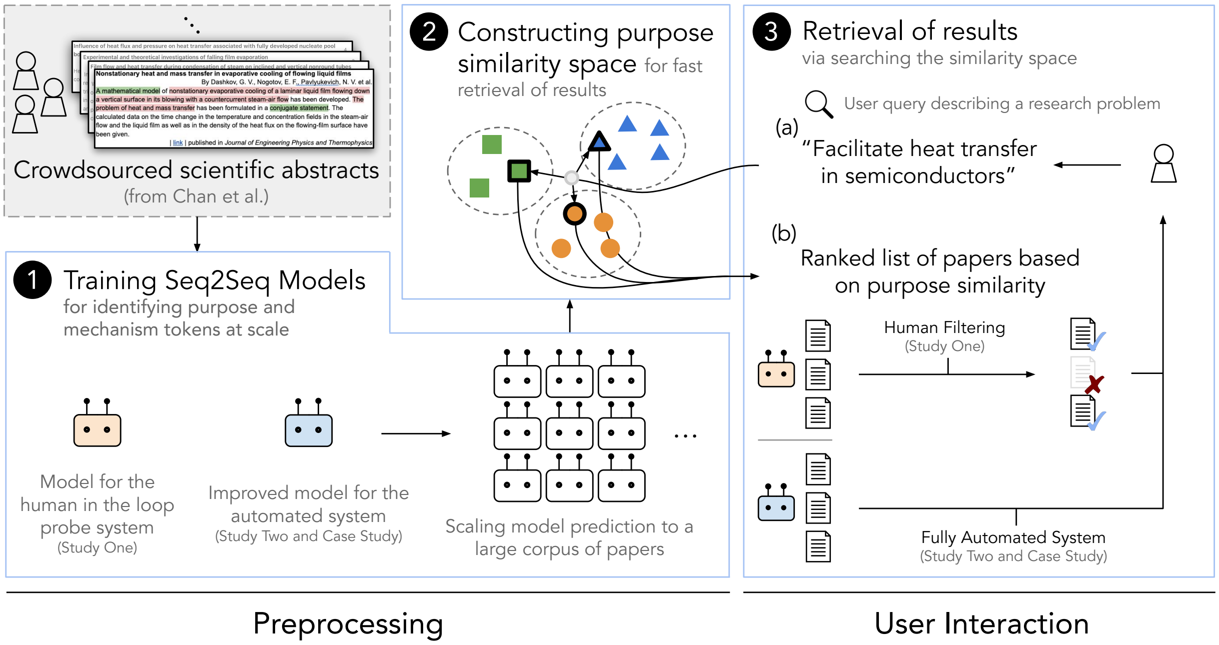

Influence of heat flux and pressure on heat transfer associated with fully developed nucleate pool
Experimental and theoretical investigations of falling film evaporation
Film flow and heat transfer during condensation of steam on inclined and vertical nonround tubes
Nonstationary heat and mass transfer in evaporative cooling of flowing liquid films
By Dashkov, G. V., Nogotov, E. F., Pavlyukevich, N. V. et al.
A mathematical model of nonstationary evaporative cooling of a laminar liquid film flowing down a vertical surface in its blowing with a countercurrent steam-air flow has been developed. The problem of heat and mass transfer has been formulated in a conjugate statement. The calculated data on the time change in the temperature and concentration fields in the steam-air flow and the liquid film as well as in the density of the heat flux on the flowing-film surface have been given.
| link | published in Journal of Engineering Physics and Thermophysics
Crowdsourced scientific abstracts
(from Chan et al.)
1 Training Seq2Seq Models
for identifying purpose and mechanism tokens at scale
Model for the human in the loop probe system
(Study One)
Improved model for the automated system
(Study Two and Case Study)
Scaling model prediction to a large corpus of papers
...
2 Constructing purpose similarity space for fast retrieval of results
3 Retrieval of results
via searching the similarity space
User query describing a research problem
(a)
"Facilitate heat transfer in semiconductors"
(b)
Ranked list of papers based on purpose similarity
Human Filtering
(Study One)
Fully Automated System
(Study Two and Case Study)
Preprocessing
User Interaction

Figure 5.2: Components of our system design that address the three core challenges. ① Purpose and mechanism tokens are extracted from paper abstracts at scale. We develop sequence-to-sequence classifiers to classify tokens into purpose, mechanism, or neither, going beyond previous approaches that worked on sentences or relied on crowds. ② We embed the extracted purpose texts using a pre-trained language model (Google's Universal Sentence Encoder (USE) [44]) and train a tree-based index of vectors to place high semantic similarity vectors in close neighborhoods for efficient lookup. ③ When the user query arrives at the system, it is first embedded with USE. This query embedding is then used to lookup the pre-computed tree indices for high similarity purpose embeddings. Paper abstracts for the corresponding purpose embeddings are retrieved from Google Datastore. In the first system, additional human filtering is performed to remove obviously irrelevant results that may have been included due to model errors. Finally, a set of papers with similar purposes to the query but different mechanisms are returned to the users for ideation.

while achieving a similar accuracy. We conclude with the benefits, design challenges, and opportunities for future analogical search engines from case studies with several researchers. To encourage innovation in this domain, we release our corpus of purpose and mechanism embeddings[1].

5.2 System Design

The design of our analogical search engine for scientific papers involves three main system requirements. First, a computational pipeline for automatically identifying purposes (*what problems does it solve?*) and mechanisms (*how does it solve those problems*) at scale (e.g., millions of papers), in a token-level granularity from scientific abstracts. Second, an efficient retrieval algorithm for incorporating the identified purpose and mechanism texts into the system to enable search-by-analogy (i.e. paper abstracts that contain similar purposes to a query problem but different mechanisms). Third, end-user interactivity for querying problems of interest (e.g., "transfer heat in semiconductors," "grow plants using nanoparticle fertilizers").

[1] `https://github.com/hyeonsuukang/augmenting_tochi22`

Kind (# of papers)	Avg. length	# of PP	# of MN
Train (2021)	196	65 261	120 586
Validation (50)	170	1510	1988

Table 5.1: Summary statistics of the training and validation datasets: the number of purpose (PP) and mechanism (MN) tokens, the number and avg. token length of paper abstracts.

Domain	CS	Eng	BioMed	B & Eng	**Total**
Count	675K	568K	336K	145K	1.7M

Table 5.2: Corpus used in the deployed search engine and its topical distribution: Computer Science (CS), Engineering (Eng), Biomedicine (BioMed), and Business and Engineering (B & Eng).

We describe the system design in detail in the following subsections.

5.2.1 Stage One. Training Seq2Seq models for identifying purpose and mechanism tokens

Overview of Modeling

In the first stage of the system, purpose and mechanism tokens are identified from paper abstracts (fig. 5.2, ➀). Research paper abstracts often include descriptions of the most important purpose or *the core problem addressed in a paper* and the proposed mechanism or *the approach taken to address the problem*, making them good candidates for identification and extraction of tokens corresponding to them. For example, for a similar problem of facilitating heat transfer, Paper A may propose an approach that modifies the structure of the material used at the interface between crystalline silicon (semiconductor material) and the substrate, while Paper B may propose a more distant mechanism (due to the mismatch on scale) of fin-based heat sinks commonly used for electronic devices. The goal of this first stage is to automatically identify and extract tokens that correspond to the similar purpose (e.g., ‘facilitate heat transfer’) as well as the mechanisms (e.g., ‘modifying the structure of the material used at the interface between crystalline silicon’ vs. ‘fin-based heat sinks’) from the abstract A and B.

One relevant automated approach for identifying purposes and mechanisms from scientific abstracts is DISA [124], which formulates the task as supervised sentence classification. However, we found that many key sentences in abstracts include both purpose and mechanism, breaking the assumptions of a sentence-level classifier (e.g., “In this paper, [*a wavelet transforms based method*] for [*filtering noise from images*] is presented.”). To overcome this limitation we follow [122] and frame purpose and mechanism identification as a sequence-to-sequence (Seq2Seq) learning task [17, 237] and develop deep neural networks with inductive biases capable of learning token-level patterns in the training dataset. Our dataset consists of crowdsourced annotations from Chan et al. (the dataset is constructed via application of [49] to a larger corpus of around 2000 paper abstracts largely in computer science domains) (Table 5.1). We train the models to classify input features (tokens or spans of tokens) as either purpose (PP), mechanism (MN) or neither.

We train two deep neural networks (Model 1 and 2), achieving increasing accuracy of classification. The first model is based on a Bi-directional LSTM (BiLSTM) architecture for sequence tagging [116, 126], in which the forward (the beginning of the sequence to the end) and the backward passes condition each token position in the text with its left and right context, respectively. A main source of improvement of Model 2 over Model 1 is the ability to more selectively attend to informative tokens in a sentence rather than treating each token in a sequence as independent of each other (as a hypothetical example, an extremely effective model based on this approach may assign more weights to the tokens 'selectively attend to informative tokens', as they represent the core mechanism described in the previous sentence) and to leverage the regularities of co-occurrence with surrounding words through the self-attention mechanism [256].

Seq2Seq Model Implementation Details

We implement the BiLSTM architecture of Model 1 in PyTorch [201]. We use pre-trained GloVe [203] word embeddings with 300 dimensions, consistent with prior work [28, 160, 203] to represent each token in the sequence as 300-dimensional input vectors for the model. We train the model with a cross entropy loss objective for per-token classification in the three (PP, MN, Neither) token classes.

For Model 2, we adapt the SpanRel [130] architecture and implement it on AllenNLP [85]. We implement a self attention mechanism that tunes weights for the core word in each span as well as the boundary words that distinguish the context of use, consistent with [162]. We use the pre-trained ELMo 5.5B [205] embeddings for token representation following the near state-of-the-art performance reported in [130] on the scientific Wet Lab Protocol dataset. We train the model using a similar procedure as Model 1. We leave detailed training parameters for Model 1 and 2 to the Appendix.

Introducing Human-in-the-loop Filtering for Model 1

The final classification performance (F1-scores) of Model 1 on the validation set is 0.509 (Purpose), 0.497 (Mechanism), and 0.801 (neither). We found that the limited accuracy contributed to how the system retrieves irrelevant search results. Because reactions to obviously irrelevant results are not useful, we added a human-in-the-loop [70] filtering stage. The filtering proceeded as follows: members from the research team inputted problem queries received from study participants into the system. Once the model produced matches, they went over from the top of the sorted list and removed only those that are irrelevant to the problem context. They continued filtering until at least 30 papers with reasonable purpose similarity were collected. After Winsorizing at top and bottom 10% [272], the human filterers reviewed 45 papers per query (SD: 27.6, min: 6, max: 138) for 5 queries (SD: 2.4, min:2, max: 9) to collect 33 (SD: 3.5, min: 30, max: 40) purpose-similar papers (about 12/45 = 26% error rate). In Study 1 we show that the limited retrieval accuracy of Model 1 is sufficient for use as a probe with this additional human-in-the-loop filtering. In Study 2 and case studies, we demonstrate how this filtering can be removed with Model 2 while achieving a similar accuracy.

Scaling Model Inference

In order to have sufficient coverage to return diverse results, we collected an initial corpus of 2.8 million research papers from Springer Nature[2]. After deduplication (based on Digital Object Identifier using BigQuery[3]) and filtering only papers with at least 50 characters in the abstract we were left with 1.7 million papers in four subjects (Table 5.2). We stored the resulting corpus in Google Cloud storage buckets[4]. To scale the classification of the Seq2Seq models we used the Apache Beam API[5] on Google

[2] `https://dev.springernature.com/`
[3] `https://cloud.google.com/bigquery`
[4] `https://cloud.google.com/storage`
[5] `https://beam.apache.org/`

Cloud Dataflow[6] to parallelize the operation.

5.2.2 Stage Two. Constructing a purpose similarity space

Overview

In the second stage, the identified purpose texts are incorporated into the system to enable search-by-analogy of papers that solve similar problems using different mechanisms, at an interactive speed (fig. 5.2, ②). Relevant previous approaches include Hope et al. [121] which first clusters similar purposes (through *k*-means with pruning) and subsequently samples within each cluster of similar purposes to maximize the diversity of mechanisms (via a GMM approximation algorithm [213]), or [122] which employs similarity metrics to balance the *similarity* to a purpose query and the *distance* to a mechanism query (and vice versa). In contrast, from pilot tests in our corpus we discovered that even close purpose matches of scientific papers already had high variance in terms of the mechanisms they propose. We hypothesize that this may be the case due to the enormous span of possible research topics and the relative sparseness of their coverage in our corpus, and/or due to the emphasis on novelty in scientific research that discourages future papers which might contribute relatively small variations to an existing mechanism. We leave exploration of these hypotheses for future work and simplify our sampling of the scientific papers to the one based solely on the similarity of purpose, sufficient for ensuring diversity.

In order to support fast retrieval (e.g., sub-second response time) of papers with similar purposes at scale (e.g., millions of papers), we pre-train Spotify's Annoy[7] indices of nearest neighboring purposes. Annoy trains a neural network to assign an embedding vector corresponding to a purpose an index in the high-dimensional space that brings it close to other indices of purpose vectors that have similar meaning (see §5.2.2 for details of the metric used for the similarity of meaning). Annoy uses random projection and tree-building (see [3, 4]) to create read-only, file-based indices. Because it decouples creation of the static index files from lookup, it enables efficient and flexible search by utilizing many parallel processes to quickly load and map indices into memory.

Interactive Speed

Additionally Annoy minimizes its memory footprint in the process. This efficiency, critical for real-time applications such as ours, was further validated during our test of the end-to-end latency on the Web, with the average response taking 2.4s (SD = 0.56s)[8]. The level of latency we observed was sufficiently low to enable interactive search by end users (both human-in-the-loop filterers in Study One and researcher participants in case studies).

Implementation Details

To construct the similarity space, we first encode the purpose texts into high-dimensional embedding vectors which then can be used to compute pairwise semantic similarity. Here, the choice of an encoding algorithm depends on three main constraints. First, the pairwise similarity, when computed, should correlate well with the human-judged semantic similarity between the purposes. Second, similarity calculation between varying lengths of texts should be possible because extracted purposes can differ in length. Third, computationally efficient methods are preferred for scaling. To meet these requirements, we chose

[6] `https://cloud.google.com/dataflow/`

[7] `https://github.com/spotify/annoy`

[8] We tested with 20 topically varied search queries that have not previously been entered to the engine to test the latency end-users experience and to exclude the effect of caching from it.

Universal Sentence Encoder (USE)[9] to encode purposes into fixed 512-dimensional vectors. Universal Sentence Encoder trains a transformer architecture [256] on a large corpus of both unsupervised (e.g., Wikipedia) and supervised (e.g., Stanford Natural Language Inference dataset [30]) data to produce a neural network that can encode text into vectors that meaningfully correlate with human judgment (e.g., evaluated on the semantic textual similarity benchmark [43]). USE can handle texts of varying lengths (e.g., from short phrases to sentences to paragraphs), and with high efficiency [44], thereby making it suitable for our system.

We pre-compute pairwise similarity of the purpose embeddings and store the indices in neighborhoods of high similarity for fast retrieval of similar purposes. As mentioned before, we train the Annoy indices on Google Cloud AI Platform[10]. We use 1 - the Euclidean distance of normalized vectors (i.e., given two vectors $\mathbf{u}$ and $\mathbf{v}$, distance($\mathbf{u}, \mathbf{v}$) = $\sqrt{(2\,(1 - \cos{(\mathbf{u}, \mathbf{v})}))}$) as a similarity metric (using a Euclidean distance based metric for nearest neighbor clustering shows good performance, see [14] for a related discussion on the impact of the distance metric on the retrieval performance). We set the hyper-parameter k specifying the number of trees in the forest to 100 (larger k's result in more accurate results but also decreases performance; see [3] for further details). Empirically, 100 seemed to strike a good balance between the precision-performance trade-off, thus we did not experiment with this parameter further.

5.2.3 Stage Three. Retrieving the results

In the last stage, the front-end interface interacts with end users and receives problem queries. These queries are then relayed to the back-end for retrieval of papers that solve similar problems using different mechanisms. The retrieved papers are presented on the front-end for users to review (fig. 5.2, ③). When a user query is received, the back-end first encodes it using the same encoding algorithm used as the construction method of the purpose similarity space (i.e. Universal Sentence Encoder). Using this query embedding, the back-end searches the pre-trained similarity space for papers with similar purposes. The papers with high purpose similarity are then returned to and displayed on the front-end. We describe the actual interfaces used in the studies in the corresponding design sections (§5.3.2, §5.3.2).

Together the design of our system enabled what is to our knowledge the first functioning prototype of an interactive analogical search engine for scientific papers at scale. In the following sections we report on how such a search engine can help researchers find analogical papers that facilitate creative ideation.

5.3 Study 1: Creative Adaptation with a Human-in-the-loop Analogical Search Engine

In Study 1 we set out to establish the viability of an analogical search engine using a human-in-the-loop probe in the domain of scholarly recommendations. We investigate whether analogical search returns a distinct and novel set of papers compared to keyword search results, and capture participants' reaction to each result in a randomized order, blind to condition. To deeply understand the process of ideation using analogical papers we ask participants to come up with new ideas for their own research projects after reviewing each paper. Using this data we code ideation outcomes in depth to explore the various ways in which analogical distance can shape ideation outcomes, such as inspiring direct transfer of solutions, or sparking adaptation of ideas into novel combinations.

[9] https://tfhub.dev/google/universal-sentence-encoder-large/5

[10] https://cloud.google.com/ai-platform

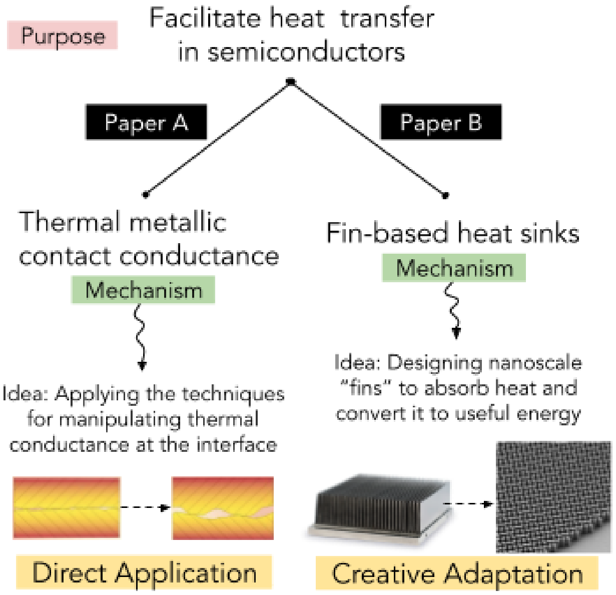

Purpose
Facilitate heat transfer in semiconductors
Paper A
Paper B
Thermal metallic contact conductance
Mechanism
Idea: Applying the techniques for manipulating thermal conductance at the interface
Direct Application
Fin-based heat sinks
Mechanism
Idea: Designing nanoscale "fins" to absorb heat and convert it to useful energy
Creative Adaptation

Figure 5.3: Example papers for the purpose of facilitating heat transfer heat in semiconductors. (Top) A Direct Application paper involves directly applicable ideas and techniques for manipulating the interface material and structure to control thermal conductance. (Bottom) A Creative Adaptation example involves transferring a distant idea (fin-based design for heat sinks) and creatively adapting it into the target problem context (designing nano-scale fins that could absorb heat and convert it to useful energy). Figure credits: contact configurations and interface resistance from [276], fin-based heat sink from [247], nanofins from [220].

5.3.1 Coding ideation outcomes

We are interested in studying whether an analogical search engine provides distinctive and complementary value to other commonly used search approaches that rely on surface similarity. In particular, our focus is on the inspirational value rather than the immediate relevance of search results or the direct usefulness of solutions. The highest value of creative inspiration often comes from creatively adapting ideas to reformulate a problem and recognizing new bridges to previously unknown domains that open up entirely new spaces of ideas. For example, recognizing a connection from the ancient art form of origami to fold intricate structures with paper and building a sufficiently compact, deployable solar panel arrays and radiation shields led NASA to hire origami experts [64, 204, 281].

Our approach to measuring ideation outcome is through the use of a quaternary variable categorizing the types of ideation. To capture the inspirational value of analogical search and move beyond the measurements focused on the immediate relevance or the direct usefulness we distinguish the Creative Adaptation and Direct Application types of ideation. In our studies these two types corresponded to think-alouds that resulted in novel ideas whereas the rest (Background and None) corresponded to think-alouds in which no new ideas were produced.

- **Creative Adaptation:** Novel mechanism ideas that involve substantial adaptation of the information provided in the paper. These ideas are typically associated with a higher uncertainty of success due to the less familiar nature of the domains involved.

- **Direct Application:** More directly applicable ideas that involve less adaptation than Creative Adaptation. These ideas are typically associated with a lower uncertainty of success because researchers are more familiar with the domains.

- **Background:** The information provided in the paper is good for background reading (e.g., to learn about other domains).

- **None:** Did not result in new ideas nor was useful for background reading.

Creative Adaptation ideas generally involved a substantial amount of adaptation, while Direct Application ideas were closer to the source domain and more directly applicable. For example, using the data from one of our participants, applying the techniques for manipulating thermal conductance at solid-solid interfaces was considered a direct application idea for P1 (fig. 5.3, left) because he was familiar with the concept of controlling the interfacial thermal conductivity given the relevant approaches he developed in his current and past research projects. Thus the connections to the source problem were directly recognizable. On the other hand, creating a fin-based wall structure for heat transfer was an example of creative adaptation idea (fig. 5.3, right) because of its novelty and the participant's unfamiliarity in relevant domains. The unfamiliarity and uncertainty was generally more associated with analogs for creative adaptation than direct application. On the other hand, the unfamiliarity also sometimes acted as a barrier to participants' openness and subsequent ideation. Though challenging, in order to recognize novel connections to the source problem the participants may need to suspend their early rejection of a seemingly foreign idea and its surface-level mismatches and engage in deeper processing which could lead to re-imagination and re-formulation of the research problem at hand. To code the Creative Adaptation and Direct Application types of ideation outcomes, the coders took into consideration different linguistic and contextual aspects of the descriptions of the ideas and their think-aloud process (details in §5.3.2).

5.3.2 Design of the study

Participants

We recruited eight graduate (four women) researchers in the fields of sciences and engineering via email advertisement at a private R1 U.S. institution. Four were senior PhD students (3rd year or above and one recently defended their thesis) and the rest was 2nd year or below. Disciplinary backgrounds of the participants included: Mechanical (3), Biomedical (2), Environmental (1), Civil (1), and Chemical Engineering (1). Once a participant signed up for the study, we asked them to describe their research problems and send the research team search queries they use to look for inspirations on popular search engines such as Google Scholar[11]. Members of the research team screened papers with relevant purposes using these queries on the filtering interface (fig. 5.4, left). Despite our efforts to collect papers over diverse topical areas, the search engine did not contain enough papers for two of the participants who work on relatively novel fields (e.g., "machine learning methods of 3D bioprinting"). These participants were interviewed on their current practices for reviewing prior works and coming up with new ideas for research and were not included in the subsequent analyses.

Study Procedure and Keyword-search Control

The rest of the participants were then invited to in-person interviews. To ensure that participants would be exposed to a sufficiently diverse set of analogical mechanisms and to maximize our power to observe the ideation process, we generated a list of top 30 results from the analogical search engine using the search queries provided by the study participants. As a control condition we also included top 15 results from a keyword-based search engine using the standard Okapi BM25 algorithm [172] ($k_1 = 1.2, b = 0.75$) using the same search queries as the analogical search engine. The order of results in the list was randomized and participants were blind to condition. To account for the difference in the quantity of exposure in the analysis, we normalized the ideation outcomes by the number of results returned in each condition. Using this list we employed a think-aloud protocol [164, 252] in which participants were presented with the title, abstract, and other metadata of papers and asked to think aloud as they read through them with the goal of generating ideas useful for their research using our Web-based interface (fig. 5.4, right). Although time consuming, this approach allowed us to capture rich data on participants' thought process and how those processes changed and evolved as participants considered how a paper might relate to their own research problems. In addition, we asked the participants to make a judgment on the novelty of each paper on a 3-point Likert-scale. After participants finished reviewing the 45 papers, we interviewed them about their overall thoughts on the results' relevance and novelty and whether there were any surprising or unique results. Each interview lasted about one and a half hours and the participants were compensated $15/hr for their participation.

Data and Coding

In total, our data consisted of 267 paper recommendations for six participants and their Likert-scale questionnaire responses measuring the content novelty, after removing 3 within-condition duplicates (these papers included cosmetic changes such as different capitalization in the title or abstract). One participant ran out of time towards the end of the interview and only provided novelty measures for the last 17 paper recommendations in the randomized list. Thus, 250 transcripts of participants' think-aloud ideation after reading each paper were used for analyzing ideation outcomes. To code the distance between the Creative Adaptation and Direct Application types of ideation outcomes, the coders took into consideration (1) the

[11] https://scholar.google.com/

verbs used to describe the ideas (e.g., ‘design’, ‘develop’, or ‘invent’ were generally associated more with distant ideas compared to ‘apply’, ‘use’, ‘adopt’; see Table. 5.3); (2) the context of ideas such as participants’ expression of unfamiliarity or uncertainty of the domain involved (e.g., “I’m not really sure” vs. “I’m familiar with this domain”); and (3) participants’ perceived immediacy of the idea’s applicability (i.e., ideas perceived by participants as more immediately applicable were associated with direct application but not creative adaptation ideas). Two of the authors coded a fraction of the data together (13/250, 5.2%) and then independently coded the rest blind-to-condition, using the four ideation outcomes types described in §5.3.1 and with the following protocol: The coders first judged the existence of an idea. If there was, then its type was further distinguished between Creative Adaptation and Direct Application using the linguistic and contextual descriptions described above (e.g., Creative Adaptation ideas were more frequently associated with the ‘design’ words, higher unfamiliarity and uncertainty of the domains, and less immediate applicability, compared to Direct Application ideas). In case there was no concrete idea in the data, coders further distinguished between the Background vs None cases.

The agreement between coders was significant, with Cohen’s $\kappa = 0.89$ (near perfect agreement) for the four categories of ideation outcome. Given the high level of agreement between the coders, any disagreements were resolved via discussion on a case-by-case basis.

Apparatus 1: the human-in-the-loop filtering interface

In Study 1, members of the research team first received search queries from study participants and reviewed the model-produced purpose matches to filter irrelevant papers using a filtering interface (fig. 5.4, left). This additional step was introduced to ensure that papers with obviously dissimilar purposes are not returned to study participants. Reviewers determined whether each paper contained a clearly irrelevant purpose in which case it was removed by clicking the *Dissimilar* button at the bottom of the paper. On the other hand when the *Similar* button was clicked it turned the background of the paper green in the interface and increased the number of the papers collected so far. Reviewers continued the screening process until at least 30 papers with reasonable purpose similarity were collected.

Apparatus 2: the ideation task interface

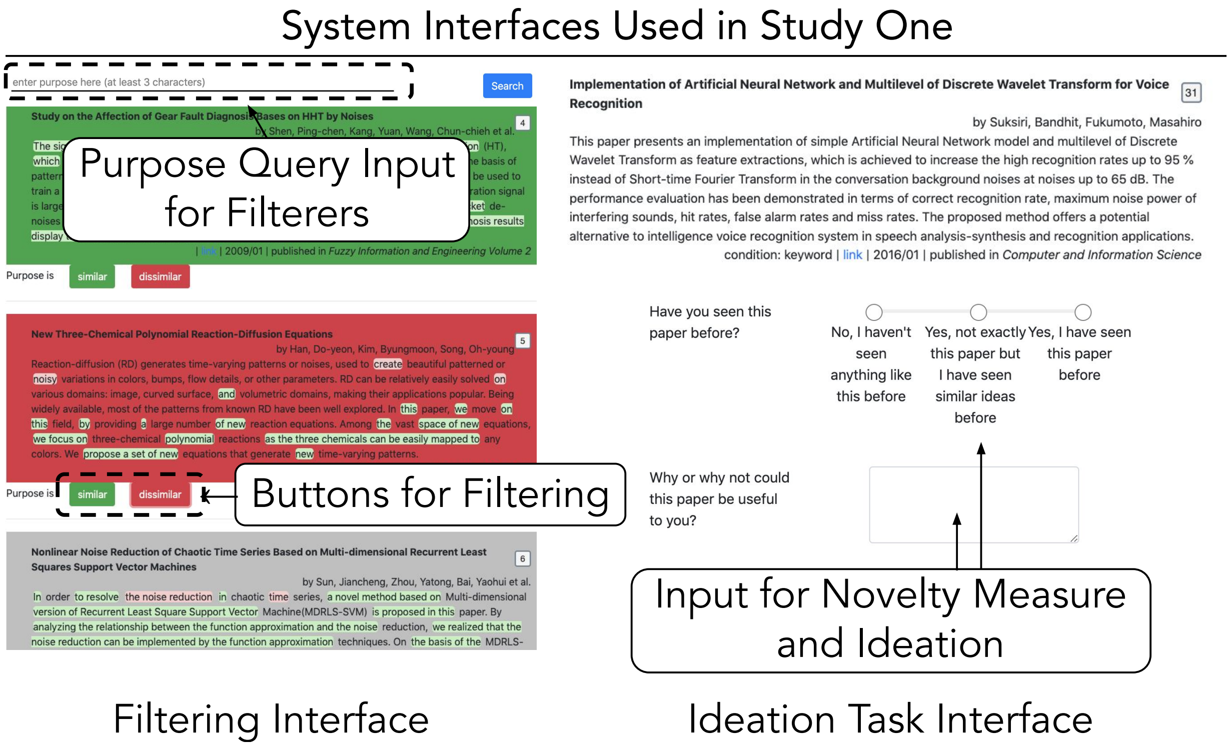
The filtered papers were then displayed as a randomized list of papers to study participants (fig. 5.4, right). In addition to the content and metadata of papers (e.g., authors, publication date, venue, etc.), each paper was presented with a Likert-scale question for measuring content novelty and a text input for ideation.

Limitations

To reduce potential biases, our coders were blind to experimental conditions and relied on participants’ statements of ideas’ novelty and usefulness (e.g., “I’ve never seen something like this before,” “this is not a domain I would’ve searched if I used Google Scholar”), and achieved a high inter-rater reliability. We believe coders had a reasonable understanding of how participants arrived at specific ideas from descriptions of their current and past research topics, think-alouds, and end-of-experiment discussions. Despite this, we also acknowledge the limitations of this approach and discuss how future research may improve upon it (see §5.7.2).

On reporting the results

We report the result of our studies below. To denote statistical significance we use the following notations: $^{*}(\alpha = 0.05)$, $^{**}(\alpha = 0.01)$, $^{***}(\alpha = 0.001)$, $^{****}(\alpha = 0.0001)$. Alpha levels were adjusted when appropriate


System Interfaces Used in Study One
Purpose Query Input for Filterers
Buttons for Filtering
Input for Novelty Measure and Ideation
Filtering Interface
Ideation Task Interface

enter purpose here (at least 3 characters) Search

Study on the Affection of Gear Fault Diagnosis Bases on HHT by Noises 4

by Shen, Ping-chen, Kang, Yuan, Wang, Chun-chieh et al.

The sig … on (HT), which … he basis of pattern … be used to train a … ration signal is large … ket de- noises … nosis results display …

| link | 2009/01 | published in *Fuzzy Information and Engineering Volume 2*

Purpose is similar dissimilar

New Three-Chemical Polynomial Reaction-Diffusion Equations 5

by Han, Do-yeon, Kim, Byungmoon, Song, Oh-young

Reaction-diffusion (RD) generates time-varying patterns or noises, used to create beautiful patterned or noisy variations in colors, bumps, flow details, or other parameters. RD can be relatively easily solved on various domains: image, curved surface, and volumetric domains, making their applications popular. Being widely available, most of the patterns from known RD have been well explored. In this paper, we move on this field, by providing a large number of new reaction equations. Among the vast space of new equations, we focus on three-chemical polynomial reactions as the three chemicals can be easily mapped to any colors. We propose a set of new equations that generate new time-varying patterns.

Purpose is similar dissimilar

Nonlinear Noise Reduction of Chaotic Time Series Based on Multi-dimensional Recurrent Least Squares Support Vector Machines 6

by Sun, Jiancheng, Zhou, Yatong, Bai, Yaohui et al.

In order to resolve the noise reduction in chaotic time series, a novel method based on Multi-dimensional version of Recurrent Least Square Support Vector Machine(MDRLS-SVM) is proposed in this paper. By analyzing the relationship between the function approximation and the noise reduction, we realized that the noise reduction can be implemented by the function approximation techniques. On the basis of the MDRLS-

Implementation of Artificial Neural Network and Multilevel of Discrete Wavelet Transform for Voice Recognition 31

by Suksiri, Bandhit, Fukumoto, Masahiro

This paper presents an implementation of simple Artificial Neural Network model and multilevel of Discrete Wavelet Transform as feature extractions, which is achieved to increase the high recognition rates up to 95 % instead of Short-time Fourier Transform in the conversation background noises at noises up to 65 dB. The performance evaluation has been demonstrated in terms of correct recognition rate, maximum noise power of interfering sounds, hit rates, false alarm rates and miss rates. The proposed method offers a potential alternative to intelligence voice recognition system in speech analysis-synthesis and recognition applications.

condition: keyword | link | 2016/01 | published in *Computer and Information Science*

Have you seen this paper before?

○ No, I haven't seen anything like this before
○ Yes, not exactly this paper but I have seen similar ideas before
○ Yes, I have seen this paper before

Why or why not could this paper be useful to you?

Figure 5.4: The front-end interfaces. (Left) Human reviewers used this filtering interface to input search queries received from the participants and remove papers with obviously irrelevant purposes. To assist the reviewers' filtering process, model predicted purpose (e.g., *the noise reduction* and *time*, highlighted in red at the bottom of the filtering interface) and mechanism (highlighted in green) tokens were also provided along with the title and the abstract text. The background color turned green when the "Similar" button is clicked and red when the "Dissimilar" button is clicked. (Right) The ideation task interface was populated with a list of human filtered papers for review by the participants in Study 1 (the order of papers was randomized).

in post-hoc analyses using Bonferroni correction.

5.3.3 Result

Finding novel papers for creative ideas. Our key measure of success is how paper recommendations from the analogy search engine (hereinafter *analogy papers*) help scientists generate creative ideas for their own research problems. To this end, we investigate a) whether analogy papers are novel and complementary to the papers found from the keyword-search baseline (hereinafter *keyword papers*) and b) whether analogy papers resulted in more creative adaptation ideas than direct application ideas in ideation.

Analogy papers differed from keyword papers and were judged more novel

The viability of our approach is based on the assumption that the analogy search pipeline returns a different distribution of results than a keyword-based baseline. This assumption appeared to hold true: the keyword-search and analogy-search conditions resulted in almost completely disjoint sets of paper recommendations. Out of the total 267 papers, the overlap between analogy and keyword papers was only one.

Analogy papers appeared to represent a complementary set of results users would be unlikely to encounter through keyword-based search.

To further examine this assumption we had participants rate the novelty of the results by asking them "*have you seen this paper before?*" on a 3-point Likert scale response options of 1: "*Yes, I have seen this paper before*", 2: "*Yes, not exactly this paper but I have seen similar ideas before*", and 3: "*No, I have not seen anything like this before*". Participants found papers recommended in the analogy condition to contain significantly more novel ideas (2.7, SD: 0.48) compared to the keyword condition (2.3, SD: 0.55) (Welch's two-tailed t-test, $t = -5.53, p = 1.33 \times 10^{-7}$) (fig. 5.5, left). Participants thought the "variance in results is much higher than using other search engines" (P5) and "there're a lot of bordering domains... which can be useful if I want to get ideas in them" (P4).

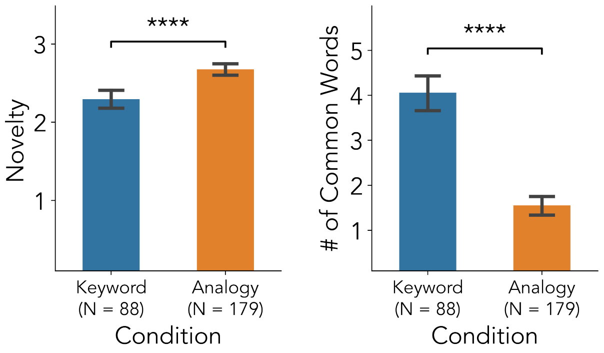


Figure 5.5: (Left) Participants judged analogy papers significantly more novel. The mean response to the question *"Have you seen this paper before?"* was significantly higher in Analogy: 2.7 (SD: 0.48) than in Keyword: 2.3 (SD: 0.55). (Right) There were significantly more overlapping words between search query terms provided by participants and the title and abstract text of papers: Keyword: 4.1 (SD: 1.74) vs. Analogy: 1.6 (SD: 1.42).

This difference was also reflected in the content of papers, with keyword papers having significantly more overlapping terms with participant-provided query terms (4.1, SD: 1.74) than analogy papers (1.6, SD: 1.42) (Welch's two-tailed t-test, $t(145.27) = 11.70, p = 1.10 \times 10^{-22}$) (fig. 5.5, right)[12]. More occurrences of familiar query terms in keyword papers' titles and abstracts may have led participants to perceive them as more familiar.

Analogy papers resulted in more creative adaptation ideas than direct application ideas

We found that the distribution of ideation outcome types differed significantly between analogy and keyword papers ($\chi^2(3) = 52.12, p < 1.0 \times 10^{-10}$). Participants came up with more creative adaptation ideas (N = 53; 32% of total) over direct application ideas (N = 3; 2%) using analogy papers. In contrast, keyword papers resulted in more direct application ideas (N = 16; 19%) than creative adaptation ideas (N = 10; 12%) (fig. 5.6). The difference between creative adaptation and direct application was significant ($\chi^2(1) = 28.41, p = 9.84 \times 10^{-8}$).

To illustrate more concretely the divergent patterns of ideation leading to Creative Adaptation and Direct Application ideas, we describe vignettes from three participants (table 5.3). While Direct Application ideas represented close-knit techniques and mechanisms directly useful for the source problem (described with verbs such as *apply* and *adopt*), Creative Adaptation type ideas were more distant from the source problem and could be characterized with the use of different verbs associated with significant adaptation (*design* and *invent*). For example, P1's research focused on the methods for improving nanoscale heat transfer in semiconductor materials. Previously he developed mechanisms for manipulating the thermal conductivity at solid-solid interfaces, specifically by adjusting the semiconductor wall structures. Thus,

[12] We measured the term overlap between participants' queries and the content of papers (title and abstract). To preprocess text, we used NLTK [25] to tokenize papers' content, remove stopwords, digits, and symbols, and lemmatize adjectives, verbs, and adverbs. Finally, using the processed tokens we constructed a set of unique terms for each paper and the query which was then compared to find overlapping terms.

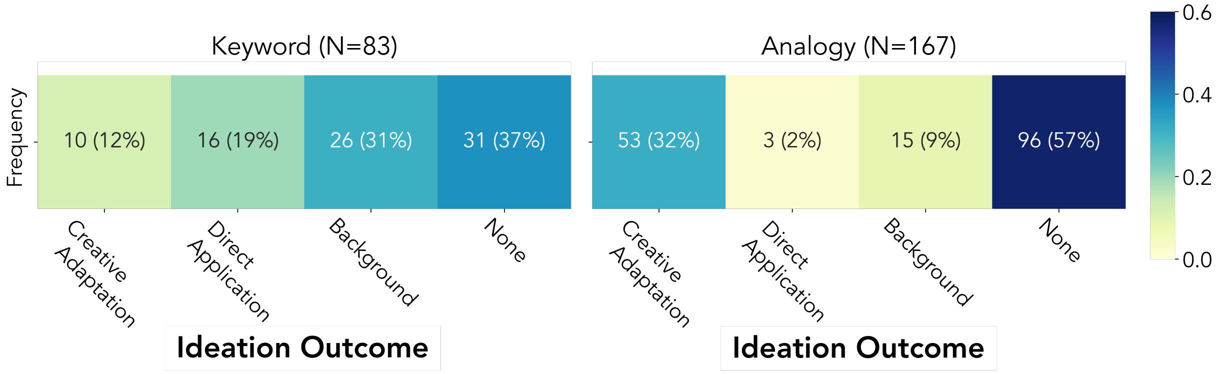


Figure 5.6: Frequency of the ideation outcome types by condition. Darker colors represent higher rates. Creative adaptation is 5.3 times more frequent among analogy papers (53 in Analogy vs. 10 in Keyword), while most of direct application is from keyword papers (3 in Analogy vs. 16 in Keyword). The distributions differed significantly (chi-squared test, $\chi^2(3) = 52.12, p < 1.0 \times 10^{-10}$ overall and $\chi^2(1) = 28.41, p = 9.84 \times 10^{-8}$ for the contrast between the rates of creative adaptation and direct application ideas).

a paper reporting experimental results of manipulating thermal conductance on planar metallic contact points was deemed a directly useful paper that might contain helpful techniques. On the other hand an analogy paper which dealt with the heat transfer phenomenon at a macroscale, using fin-based heat sink designs for electronic devices, gave him a new inspiration: to adapt fins for nanoscale heat transfer in semiconductors to not only transfer heat but also convert it into a useful form of mechanical energy. Despite the mismatch on scale ([macroscale] ↮ [microscale]), challenging the assumption of the typical size of a fin-based design engendered an idea to creatively adapt it to convert heat into energy through an array of tiny fins, rather than merely dissipating it into space as in the original formulation of the problem. P1 also found another analogy paper focused on thermal resistance at a liquid-solid interface useful for future ideation because despite its surface dissimilarities, there was a potential mapping that may open up a new space of ideas (e.g., [liquid] ↮ [polymer substrate], [solid] ↮ [germanium], yet the pairwise relation [liquid:solid] ↔ [polymer substrate:germanium] may be analogous and interesting): "This is liquid... but it's about liquid-solid interface which can be useful... because for the substrate that sits on top of silicon or germanium you use polymers which have liquid-like properties" (P1).

In the case of P2, a paper focused on computational methods for toxicity prediction was deemed directly helpful because "if certain nanomaterials are toxic to certain microorganisms that eat plants or kill them but safe for the plant, we can target these organisms using the nanomaterials as pesticide. Another way this can be helpful is in predicting the chance of toxicity of the nanoparticles in our fertilizers" (P2). Whereas an analogy paper that uses image analysis for plant identification reminded her of "hyperspectral imaging in plants, like a CT scan for plants. So making a hyperspectral 3D model using something like this... to optically sense and trace plant cells (such that the entry of fertilizer nanoparticles into plant cells can be monitored, a sub-problem of P2's research problem) would be pretty cool."

As a third example, P6's research focused on recording and simulating electrical activity using microelectrode arrays. To him, an analogy paper about printing sensors for electrocardiogram (ECG) recording seemed to present an interesting idea despite its mismatch in terms of scale ([nanoscale] ↮ [macroscale]) and manufacturing mechanism ([fabrication] ↮ [printing]), because the pairwise relation between [nanoscale:fabrication] ↔ [macroscale:printing] engendered a reflection on the relative advantages of different methods and future

PID	Research Problem	Type	**Paper Title → New Idea** (paraphrased)
1	Improve nanoscale heat transfer in semiconductor material	Direct Application	*Experimental investigation of thermal contact conductance for nominally flat metallic contact* → Apply the techniques in the paper to manipulate thermal conductance at the solid-solid interface
		Creative Adaptation	*Investigation on periodically developed heat transfer in a specially enhanced channel* → Design nanoscale "fins" to absorb heat and convert it to mechanical energy
2	Grow plants better by optimizing entry of nanoparticle fertilizers into the plant	Direct Application	*Nanoinformatics: Predicting Toxicity Using Computational Modeling* → Apply the computational modeling from the paper for predicting toxicity of candidate nanoparticles
		Creative Adaptation	*Identification of Plant Using Leaf Image Analysis* → Invent a hyperspectral 3D imaging mechanism for plants that optically senses, traces, and images plant cells in 3-dimensional structures
3	Enhance the evaporation efficiency of thin liquid films in heat pipes and thermosyphones	Direct Application	*Thin film evaporation effect on heat transport capability in a grooved heat pipe* → Adopt the techniques in the paper for manipulating the solid interface's surface properties to balance the film thickness and disjoining pressure
		Creative Adaptation	*Alkaline treatment kinetics of calcium phosphate by piezoelectric quartz crystal impedance* → Design novel liquid film materials for manipulating hydrophobicity to change disjoining pressure

Table 5.3: Examples of Direct Application and Creative Adaptation types for three participants (PID). Each participant's research problem is described in the Problem column. While the topics of research problems vary, Creative Adaptation ideas are more distant in terms of content compared to the source problem than Direct Application ideas are, and may be characterized by the use of different sets of verbs ({*design*, *invent*} in Creative Adaptation ideas versus {*apply*, *adopt*} in Direct Application ideas).

research directions): "Interesting idea! Instead of nanoscale fabrication, printing can be a good alternative for example for rapid prototyping. But I think the resolution won't be enough (for use) in nanoscale... works for this particular paper's goal, but an idea for future research is whether we can leverage the benefit of both worlds – rapid printing and precision of nanoscale fabrication" (P6).

The level of purpose-match had different effects on the ideation outcome

Suggested in these examples is a certain kind of distance the ideas in analogy papers maintain in order to spur creative adaptation. We hypothesize that some amount of difference in purpose facilitates creative adaptation. This process may involve a curvilinear relationship between the degree of purpose mismatch and the resulting ideation outcome, with too much or too little deviation leading to a little-to-no benefit or even an adverse effect on the ideation outcome, a pattern that is consistent with the findings in the literature of creativity and learning outcomes (e.g., Csikszentmihalyi's optimal difficulty [60]). For this analysis, we coded each paper based on three levels of purpose-match to the source problem:

- **Full:** Both high- and low-level purposes match
- **Part:** Only the high-level abstract purpose matches. Explicit descriptions of the high-level purpose exist in either title and abstract of the paper. At the same time, certain low-level aspects of the participant's

Purpose-Match	PID	Participant Comment
Full	2	"It's a little bit old (from 2010) but I have read papers from that era. I love this... because the paper mentions everything else and especially one word which is 'disjoining pressure' – if I were to publish my current project that's going to be the core topic."
Part	1	"Though I'm not familiar with GFRP-GFRP... but I can see that they're referring to glass fiber reinforced plastic, so this is something not crystalized material... learning about this kind of materials is interesting."
None	3	"I don't know what a lot of words mean. I don't typically work with animals cells."

Table 5.4: Examples of different purpose-match types. Purpose-Match shows the level of purpose-match between a recommended paper and each participant's research problem (see table 5.3 for descriptions of research problems). Fully matching purposes are those that match at both high- (more abstract) and low-levels (specific details). Partial matches only match at the high-level abstraction and differ in details. The Participant Comment column shows relevant excerpts from the participant.

research problem are mismatched as evidenced by relevant comments from the participant

- **None:** Neither high- nor low-level purposes match

Examples of these types of purpose-match are provided in Table 5.4. High-level match can be considered as a first-order criterion of purpose match and low-level match as a second-order criterion: If the paper does not have overlapping terms in terms of its purpose with the user query cast at a high level (e.g., transfer heat, grow plants) then the low-level match does not matter, but if the paper's purpose matches at the high level, its low-level alignment (e.g., specific aspects of the purpose, such as its scale or materialistic phase) will additionally determine full (i.e., aligned in both high- and low-level aspects of the purpose) vs partial match (i.e., aligned only in the high-level but not low-level aspects of the purpose). Therefore, the coding procedure was symmetrical to the procedure described for coding four types of ideation outcome, with the high-level purpose match deciding between {Full, Part} and None match types, while the low-level purpose further distinguishing between Full vs. Partial match. Following this procedure, two independent coders achieved an inter-rater reliability Cohen's $\kappa = 0.72$ (substantial agreement) and disagreements were resolved with case-by-case discussion.

We used the MEDIATION package[13] [249] to conduct a mediation analysis between the condition, the kind of purpose-match, and the binary Creative Adaptation ideation outcome. The analysis showed that the effect of condition (Keyword vs. Analogy) on the binary outcome of creative adaptation was mediated by the degree of purpose-match, but not by the novelty of content, suggesting that the difference between full vs. partial matching on purpose is much more significant than the variance in the content novelty. We come back to this result in the discussion (§5.7.2). Table 5.5 presents the result of the mediation analyses. The regression coefficient between creative adaptation and condition was significant as was the regression coefficient between the degree of purpose match and creative adaptation. The indirect effect was $(-.42) \times (.21) = -.09$. We tested the significance of this indirect effect using a bootstrapping procedure [210] ($p < 2 \times 10^{-16}$), by computing the unstandardized indirect effects for each of 1000

[13] https://cran.r-project.org/web/packages/mediation/index.html

Mediator	Effect of Condition on Mediator (*a*)	Unique Effect of Mediator (*b*)	Indirect Effect ($a \times b$)	CI 95% Lower	CI 95% Upper
Purpose-match	-0.42^{****} (.08)	0.21^{****} (.05)	-0.09^{****}	−0.14	−0.05
Novelty	0.40^{****} (.07)	−0.06 (.05)	−0.02	−0.07	0.02
Pid	−0.02 (.22)	0.03^{*} (.02)	−0.001	−0.02	0.02

Table 5.5: Regression table of three mediation analyses using *Purpose-match*, *Novelty* and *Pid* (Participant ID) as mediators between Condition and the binary Creative Adaptation outcome variable. Purpose-match was the only significant mediator between Condition and Creative Adaptation (indirect effect=-.09, significant using a bootstrapping method [210] with 1000 iterations, $p < 2 \times 10^{-16}$).

boostrapped samples as well as the 95% confidence interval (CI)[14].

Partial purpose matches in both keyword and analogy papers led to creative adaptation, but the rate was significantly higher with analogy papers. As expected, the ratio of direct application decreased from the keyword papers that fully match in purpose (Keyword Full, 68%) to the keyword papers that partially match in purpose (Keyword Part, 6%) (fig. 5.8). At the same time, the rate of creative adaptation increased from the keyword papers that fully match in purpose (Keyword Full, 0%) to the keyword papers that partially match in purpose (Keyword Part, 21%). However, the rate of creative adaptation differed significantly between the keyword and analogy papers, with the rate more than doubling among the analogy papers over keyword papers (Analogy Part 47% vs. Keyword Part 21%). Homing in on the partial matches, these papers led to creative adaptation ideas significantly more often in analogy search (47%) than keyword search (21%) (Welch's two-tailed t-test, $t(112.22) = -3.40, p = 9.0 \times 10^{-4}$, fig. 5.7, left). While the partial purpose mismatch was highly associated with creative adaptation ideas, it could be a double-edged sword. Among the analogy papers, 38% of the partial mismatches resulted in no useful ideation outcome as opposed to the 47% that resulted in creative adaptation (fig. 5.8, Analogy Part). Therefore, **knowing what mismatches are beneficial to creative adaptation** has important implications for facilitating generative misalignment for ideation.

5.4 Study 2: Enabling a Fully Automated Analogical Search Engine

5.4.1 Motivation and structure of the study

The findings of Study 1 suggest potential benefits of an analogical search engine for scientific research, but a core limitation of interactivity due to the human-in-the-loop system design prevented its use as a more realistic probe for understanding researchers' natural interaction with analogical results. Specifically, the results of Study 1 are limited by the lack of participants' ability to reformulate search queries and the study design that involved returning only a fixed number of papers that blended both keyword and analogy papers in a randomized order. These factors significantly deviate from realistic usage scenarios of a deployed analogical search engine and prevent us from observing the full scope of user interaction.

[14] Alternatively, it is possible that the mediating effect of the degree of purpose-match on the likelihood of creative adaptation outcome is moderated by novelty. However, the result of our analysis showed that this was unlikely: The effect was insignificant using the boostrapping method -.04, ($p = 0.12$, 95% CI = $[-.09, .01]$).

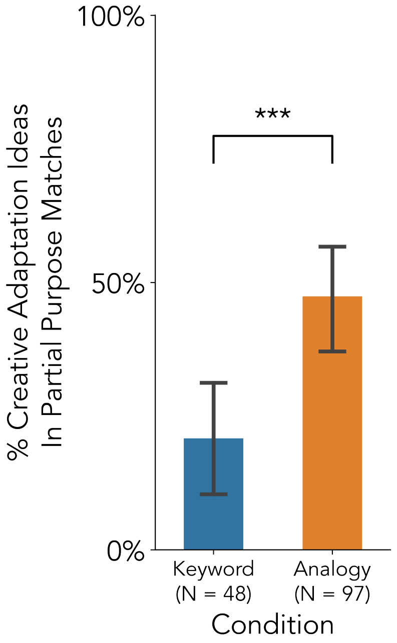


Figure 5.7: Proportion of creative adaptation ideas among the partial purpose-match papers. Creative Adaptation was significantly more frequent among the analogy papers (47%) than keyword papers (21%) (Welch's two-tailed t-test, $p = 9.0 \times 10^{-4}$.

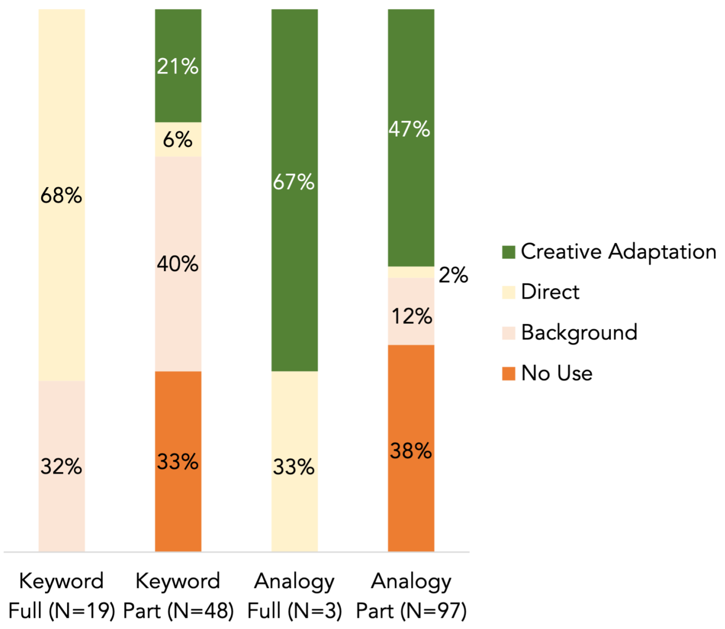


Figure 5.8: The rate of ideation outcome types in full and partial purpose matches. Among the keyword papers as the purpose mismatch increases, the rate of creative adaptation also increases from 0% to 21% (middle). However, this rate is significantly higher among the analogy papers (47%) than the keyword papers (21%). Note that while purpose mismatches led to more creative adaptation among analogy papers, a large portion of them also resulted in no ideation outcome (38%).

In order to move beyond these limitations, first we need a fully automated pipeline that removes the need for human-in-the-loop filtering, thus allowing us to enable query reformulation and interaction with corresponding search results. To achieve this, we improved the model accuracy on extracting purposes and mechanisms from paper abstracts by training a more sophisticated neural network that leverages more nuanced linguistic patterns. Specifically, we implemented an attention mechanism within a span-based sequence-to-sequence model (Model 2) such that it could learn words that frequently co-occur to describe coherent purposes or mechanisms in paper abstracts, and as a result, learning more informative words for our purpose (see Appendix for details of implementation). Through evaluating the system backed by this improved pipeline, we demonstrate how it can remove the human-in-the-loop while maintaining similar levels of accuracy. In the following sections, we report the evaluation results that show 1) an improved token-level prediction accuracy using the span-based Model 2; 2) rankings of the results aligning well with human-judgment of purpose-match from Study 1; and 3) top ranked results of the system maintaining a similar rate of partial purpose matches relative to that of the human-in-the-loop system from Study 1.

The interactivity enabled by the automated analogical search pipeline further allows us to observe its use in more realistic scenarios. To probe how researchers would interact with an analogical search engine and what challenges they might face in the process, we ran case studies with six researchers (§5.5). From these studies, we uncover potential challenges (§5.5) and synthesize design implications for future analogical search engines (§5.6).

5.4.2 Result

Model	Embedding (finetuned)	All	PP	MN
1. Model 2 [130]	ELMo (N)	**0.65**	0.65	0.64
2. BiLSTM	ELMo (N)	0.63	0.67	0.59
3. BiLSTM	SciBERT (N)	0.62	0.69	0.55
4. BiLSTM-CRF [205]	ELMo (N)	0.58	0.59	0.57
5. BiLSTM	GloVe (Y)	0.55	0.56	0.53
6. Model 1	GloVe (N)	0.50	0.51	0.50

Table 5.6: F1 scores of different models, sorted by the overall F1 score of Purpose (PP) and Mechanism (MN) detection. The span-based Model 2 gave the best Overall F1 score (blue). In comparison, the average agreement (%) between two experts' and crowdworkers' annotations was 0.68 (PP) and 0.72 (MN) [49]. We used AllenNLP [85] to implement the baseline models 1 – 5.

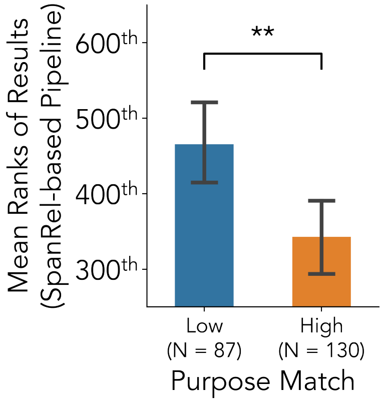


Figure 5.9: Mean ranks of human-judged high and low purpose match papers from the span-based pipeline. Low matches were ranked significantly lower (the rank number was higher), on average at 465th (SD: 261.92) than high matches at 343th (SD: 279.48).

Improved token-level prediction of a span-based model

First we compared the span-based Model 2 with five other baselines to evaluate the token-level classification performance (Table 7). Model 2's overall F1 score was the highest at 0.65 (Purpose; PP: 0.65, Mechanism; MN: 0.64, an 0.14- and 0.14-absolute-point increase from Model 1, respectively) on the validation set which represents an overall 0.15-absolute-point increase from Model 1 used for the initial human-in-the-loop analogical search engine.

Pipeline with a span-based model reflected human judgment for ranking the results

The improved token-level prediction performance materialized as an increase in the pipeline's ability to judge the degree of purpose match. For this evaluation, we first recorded every query provided by Study 1 participants that human-in-the-loop filterers used to search and filter the relevant papers. Then, we simulated the search condition of the filterers for the automated pipeline by providing it input as the exact queries they used. We capped the number of top search results sufficiently large at 1000 for each query. From these top 1000 results, we selected papers that also appeared in the human-in-the-loop system and collected the corresponding human-vetted judgments of high or low purpose-match. For each of these papers, we also collected its corresponding rank positions on the new (automated) pipeline's list of results.

We compared the mean ranks of papers that are judged by human filterers as high purpose match to those of low purpose matches. The result showed that the new pipeline indeed was able to distinguish between the two groups of papers; low purpose matches (i.e. papers that were deemed not relevant and subsequently filtered by the judges in Study 1) were placed at significantly lower positions on the list than high purpose matches (i.e. unfiltered papers in Study 1). The mean rank for low purpose matches was 465 while for

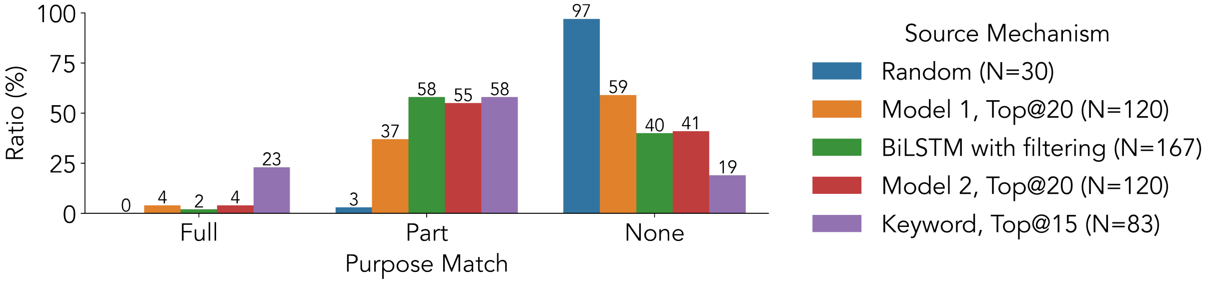


Figure 5.10: Distribution of Full, Part, and None purpose matches among the five sourcing mechanisms: *BiLSTM with filtering* represents the human-in-the-loop system (Study 1); *Model 1* represents a system based on the BiLSTM model alone, without human-in-the-loop filtering; *Model 2* represents the fully automated system; *Random* represents randomly sampled papers; *Keyword* represents keyword-based search (Control in Study 1). *Model 2* and *BiLSTM with filtering* showed a similar distribution of purpose matches, and more partial purpose matches than *BiLSTM* alone. Random showed mostly no matches. The *Keyword* condition resulted in the highest number of fully matched papers and the lowest number of no matches, suggesting that keyword-based search may be an effective mechanism for direct search tasks, but potentially less effective for inspirational/exploratory search tasks.

high purpose matches it was 343 (fig. 5.9). This difference was significant ($t(192.49) = 3.29$, $p = 0.0012$. Welch's two-tailed t-test.).

Different model performance on finding papers that fully or partially match on purpose

Data and coding. In addition to the overall rankings reflecting human-vetted judgments we also found that the proportion of partial purpose matches was significant among the top-ranked results. We sourced top 20 results for each participant's research problem with the automated system (Model 2) using the exact queries and order used by the human-in-the-loop filterers in Study 1. We compared this to four other approaches: 1) the human-in-the-loop system in Study 1 (*BiLSTM with filtering*), 2) a BiLSTM-based system excluding the human-in-the-loop from 1 (*BiLSTM*), 3) randomly sampled papers (Random), and 4) a keyword-based search results, which was used as control in Study 1 (*Keyword*). There were no overlapping papers between Model 2 and other conditions except for the Keyword condition which had 1 overlapping paper. To code the degree of purpose match, we blended the results of Model 2, biLSTM, and Random conditions. Two of the authors coded a fraction of the data together blind-to-condition (7.4%, $N = 20/270$) following the same procedure used in Study 1. Then they independently coded the rest blind-to-condition achieving an inter-rater agreement of $\kappa = 0.80$ (substantial agreement). We resolved any disagreement through discussion on an individual case basis.

Result. We found that the Model 2-based system achieved a parity with the human-in-the-loop system (Study 1) for finding purpose matches (fig. 5.10), with more than half of the system's top 20 results judged to be partial purpose matches. In contrast, when human-in-the-loop filtering was removed from the BiLSTM-based system, the frequency of partial purpose matches decreased from 58% to 37% while the frequency of no matches increased from 40% to 59%. Random sampling resulted in mostly irrelevant results, with no alignment on purpose with the source problem. An interesting point of comparison is between the keyword-based and Model 2-based search results. Keyword search mostly outperformed Model 2-based system by finding full purpose matches at a much higher rate (23% in keyword search vs. 4% in the Model 2-based system), with similar rates of partial purpose matches (58% vs. 55%),

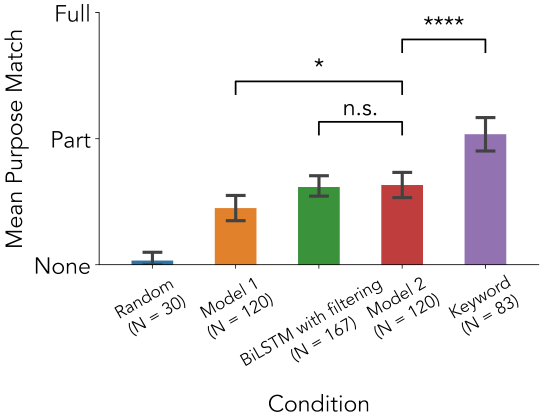
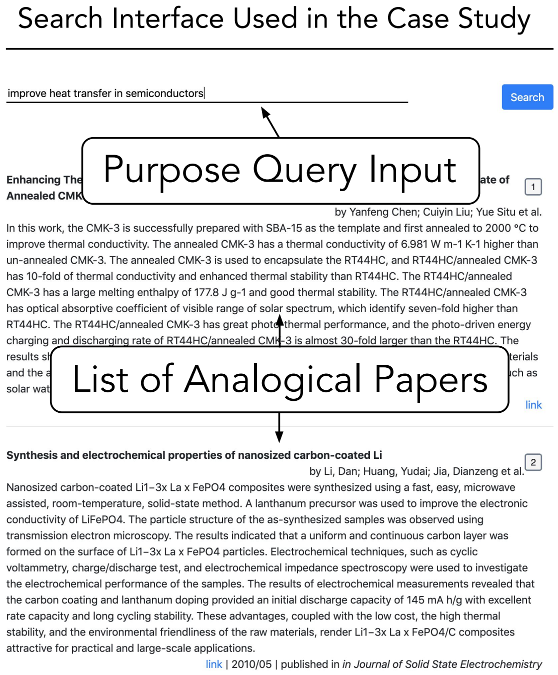


Figure 5.11: The distribution of mean purpose match scores over different conditions (mappings: None ↦ 0, Part ↦ 1, and Full ↦ 2). The mean purpose-match score of the system backed by Model 2 (0.63, SD: 0.56) is significantly higher than that of the system used in Study 1 without the human-in-the-loop (BiLSTM, $\mu = 0.45$, SD: 0.58) (Welch's two-tailed t-test, $t(237.87) = 2.49, p = 0.0135$), similar to that of the system with the human-in-the-loop (BiLSTM with filtering, $\mu = 0.62$, SD: 0.52) ($t(244.65) = 0.25, p = 0.80$), and significantly lower than that of the keyword-based search (Keyword, $\mu = 1.04$, SD:0.65) ($t(159.38) = -4.57, p = 0$).


Search Interface Used in the Case Study
Search
Purpose Query Input
List of Analogical Papers
1
link
2

improve heat transfer in semiconductors

Enhancing The … **ate of Annealed CMK**

by Yanfeng Chen; Cuiyin Liu; Yue Situ et al.

In this work, the CMK-3 is successfully prepared with SBA-15 as the template and first annealed to 2000 °C to improve thermal conductivity. The annealed CMK-3 has a thermal conductivity of 6.981 W m-1 K-1 higher than un-annealed CMK-3. The annealed CMK-3 is used to encapsulate the RT44HC, and RT44HC/annealed CMK-3 has 10-fold of thermal conductivity and enhanced thermal stability than RT44HC. The RT44HC/annealed CMK-3 has a large melting enthalpy of 177.8 J g-1 and good thermal stability. The RT44HC/annealed CMK-3 has optical absorptive coefficient of visible range of solar spectrum, which identify seven-fold higher than RT44HC. The RT44HC/annealed CMK-3 has great photo-thermal performance, and the photo-driven energy charging and discharging rate of RT44HC/annealed CMK-3 is almost 30-fold larger than the RT44HC. The results sh … terials and the a … ch as solar wat

Synthesis and electrochemical properties of nanosized carbon-coated Li

by Li, Dan; Huang, Yudai; Jia, Dianzeng et al.

Nanosized carbon-coated Li1−3x La x FePO4 composites were synthesized using a fast, easy, microwave assisted, room-temperature, solid-state method. A lanthanum precursor was used to improve the electronic conductivity of LiFePO4. The particle structure of the as-synthesized samples was observed using transmission electron microscopy. The results indicated that a uniform and continuous carbon layer was formed on the surface of Li1−3x La x FePO4 particles. Electrochemical techniques, such as cyclic voltammetry, charge/discharge test, and electrochemical impedance spectroscopy were used to investigate the electrochemical performance of the samples. The results of electrochemical measurements revealed that the carbon coating and lanthanum doping provided an initial discharge capacity of 145 mA h/g with excellent rate capacity and long cycling stability. These advantages, coupled with the low cost, the high thermal stability, and the environmental friendliness of the raw materials, render Li1−3x La x FePO4/C composites attractive for practical and large-scale applications.

link | 2010/05 | published in *in Journal of Solid State Electrochemistry*

Figure 5.12: The search interface used for case studies featured an input for query reformulation which participants used to iteratively reformulate their queries.

and significantly less no purpose matches (19% vs. 41%). On average the purpose match score was the highest in keyword-search followed by the Model 2-based and the human-in-the-loop systems (fig. 5.11). Combined with the results of Study 1, this suggests the complementary value of analogical search: The higher rate of full-matches in keyword-search may be good when searchers know what they are looking for, such as in direct search tasks and foraging from familiar sources of ideas. Nonetheless, because analogy papers were both deemed significantly more novel by the scientists and had little-to-no overlap with keyword-search papers, they augmented keyword-based search results with a complementary set of papers that introduce useful mistmatches in their purposes. This set of papers may open up new domains of ideas that scientists may not have been aware of, and encourage creative adaptation.

5.5 Case Studies with Researchers

To further understand what potential interaction challenges prevent future analogical search engines from reaching their full potential, we ran case studies with 6 participants. To this end, we developed a frontend interface that includes a text input for reformulating purpose queries (fig. 5.12, right). This frontend interfaced with our automated, Model 2-based backend to display a ranked list of analogical results for a given purpose query. Leveraging the fully automated search engine, we also removed the structure of Study 1

PID	Participants’ Description of Research Problem
1	Improve heat pipe evaporation
2	Computer simulations for fluids in nanoscale and uncovering their heat-transfer properties
3	Developing a model to identify complex steps in Nuclear Power Plant (NPP) operation, and understanding what task features and structures cause the complexity and how this influences the operators’ performance
4	Designing simulators for training bridge inspectors
5	Developing algorithms and extensible frameworks for detecting personal protective equipment (PPE) in construction sites to improve the safety of construction workers
6	Convergence rates of optimization algorithms under multiple initial starting positions

Table 5.7: Case study participants’ descriptions of own research problems

that asked participants to engage with each result they encountered, thus allowing us to observe which results researchers more naturally attend to and engage with. In sum, the design of our case studies differ from Study 1 in three aspects: 1) participants interacted with only the analogical search results ranked in the order of purpose similarity, without blended keyword-based search results; 2) participants reviewed search results returned for their queries and reformulated the queries when needed; and 3) participants looked for papers that interest them and may serve as sources of inspiration for their research problems at their own pace, without being explicitly asked to engage with each result they encounter.

The primary goal of our case studies was to identify generalizable challenges that analogical search engines may face in interactive use, thus providing us insights on how future engines may be designed and improved. Specifically, we were interested in the challenges related to 1) how researchers recognize relevance of analogical search results and 2) how researchers formulate and reformulate purpose search queries while interacting with analogical search results.

5.5.1 Participants and Design

Participants were asked to formulate purpose queries for their own research problems and interact with the results to find interesting papers. If a paper gave them a new idea relevant to their research project, they were asked to write a short project proposal in a shared Google Doc and explain how the paper helped them to come up with the idea. Interviews were conducted via Zoom and lasted for roughly an hour. Participants were paid $20 in compensation. One participant was an assistant professor in mechanical engineering at a public R1 U.S. university and five were PhD researchers in the fields of sciences and engineering at a private R1 U.S. university. Two were senior PhD students (3rd year or above) and the rest were 2nd year or below. Disciplinary backgrounds of the participants included Chemical (2), Civil (3), and Mechanical Engineering (1). We note that one participant previously took part in Study 1, whose research focus was the same in terms of its general domain. However, the participant’s ideas and the specific papers of interest that led to them did not have overlap between the two studies. Table 5.7 describes participants’ research problems.

Apparatus: Search interface. The improved performance of Model 2 backed the fully automated pipeline without human filtering. The search interface interacting with this back-end included a text input for reformulating purpose search queries as well as a list view of search results that showed a sorted list of papers with similar purposes (fig. 5.12).

5.5.2 Result

Overall impressions

Overall participants described their experience with the analogical search engine in a positive light (e.g., "helps me think at a broad topic or a big picture level" – P2; "find some very interesting and useful ideas, the design is also very simple, good when focusing on key areas of research" – P5; and "very interested now what the future of this engine would look like" – P3), but a deeper look suggested that the success of ideation depended on how well searchers were able to engage with analogical results that deviate from their expectations: "It's surprising that the engine recommends examples like these" – P3; "If I input the same search queries on Google Scholar it'd not normally return these things... this search engine works in a different way" – P1.

"Not the kind of paper I'd look for but...": The challenge of early rejections

Unlike similarity-maximizing search engines, the diversity in analogical search results can lead to premature rejection of alternative mechanism ideas. One of the factors contributing to premature rejection of alternatives may be the tendency for adherence to a set of existing ideas or concepts, as studied in the literature of design fixation (e.g., [128]). In our study, the participants found the variety of domains featured in search results confusing, and it sometimes prevented them from engaging with the ideas therein. For example, P3, whose research studies ways to manage or reduce task complexity for nuclear power plant operators, expected to see results similar to Google Scholar which are typically in the domains of operational and managerial sciences, but was surprised by unfamiliar domains represented in search results: "These (*distributed networked systems design* or *path planning for automated robots*) are not the kinds of fields that I normally read in, if I found them elsewhere I would've probably thought they're irrelevant and skipped" (P3). Ranging from unfamiliar terms (P1, P4, P5) to unfamiliar categories of approaches (e.g., "Not sure what 'Gauss-Newton approach for solving constrained optimization' is" – P6), or high-level research directions (e.g., "this is different from my research direction, people who work on this direction might find it interesting, though" – P1), participants saw the diversity of results as a challenge for engagement. P1 pointed out a perceived gap between the expectation of least effort and the cognitive processing required when engaging with analogical ideas and adapting them:

> *"As I understand it, I think this search engine is trying to present papers from related but different fields to let people make connections. But people expect less friction. (The result is) something interesting but I can't directly write it into a project proposal... I think it would be challenging to make people get interested in investing time to read the papers in depth to come up with connections. I wonder what would happen if this was hosted just as an online website (instead of the study context)"* – P1

On the other hand, analogs that did get examined more deeply could ultimately lead to creative adaptation. For example, P3 mapped task scheduling among computer processes to task assignment among the nuclear power plant operators, and came up with an idea to adapt algorithmic scheduling used in real-time distributed systems to a scheduling mechanism that could be useful in her research context. Represented symbolically this process was akin to ideating what might best fill in the '?' in the relational structure [scheduling algorithm:processes in distributed systems] ↔ [?:nuclear power plant operators]: "I think the algorithms proposed in this paper could be useful for calculating the operator task execution time, the power plant system's response time, and the time margin between the execution time and the system response time... so that the next task assignment can factor in these margins and things related to workers'

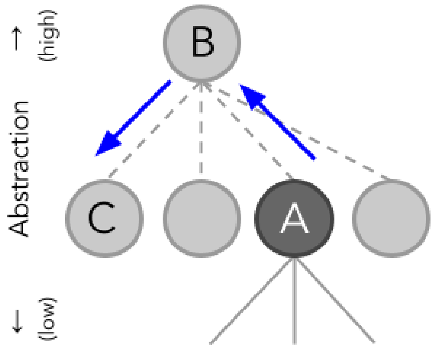


Figure 5.13: Diagram showing different abstraction levels of purposes and their relations. Node Ⓐ corresponds to a more specific query than its higher-level representation, denoted as Ⓑ. Similarly, node Ⓒ represents a more specific purpose representation of Ⓐ, accessible via the Ⓐ $\underset{\text{abstraction}}{\rightarrow}$ Ⓑ $\underset{\text{specification}}{\rightarrow}$ Ⓒ path.

well-being like rest and time required between switching tasks" (P3).

Participants seemed to recognize a small number of core relations as kernel for creative adaptation. In the example of P3, *scheduling processes* in the distributed systems paper piqued her interest and led her to connect them with similar concepts in the literature she was already familiar with: "You need to make that connection... I saw parallels between (distributed systems domain) concepts like [scheduling] and [tasks] and [scheduling tasks for the operators]" (P3). Similarly, P5 recognized a similarity between [monitoring people's performance] in fitness training and [monitoring whether construction workers are wearing personal protective equipment] in construction sites. He then adapted the idea of tracking heat emission in the fitness context to his own: "I like the idea of [monitoring heat emissions] in fitness training... maybe I can also detect heat emissions from construction workers to see if they are wearing the safety vests or masks while also monitoring the site conditions and worker efficiency. It also gives me an idea to monitor the CO_2 emissions from workers so as to improve the robustness of detection" (P5). In this case, *monitoring* and the *physical nature* of the activities involved helped P5 see the connection useful for creatively adapting the source idea.

"I don't know what to type in": The challenge of query (re-)formulation

Another challenge participants faced was that they were not used to formulating their search queries in terms of high level purposes of their research. On average participants entered 5.2 queries (Min: 1, Max: 18, SD: 5.87), 87% (27) of which were in the form of a single noun phrase (e.g., "heat pipe evaporation," – P1, "task complexity" – P3, "theoretical optimization convergence for non-convex functions" – P6) or a comma-separated set of multiple noun phrases (e.g., "heat transfer, nanoscale, fluid" – P2) that represented specific aspects related to research purposes rather than the core purposes themselves. For example, the purpose of 'heat pipe evaporation' may be to transfer heat, and the purpose of searching for 'theoretical optimization convergence for...' may be to detect when optimization converges or diverges, or to effectively sample unknown (non-convex) distributions.

One of the reasons why participants formulated search queries in this way may be wrongly assuming that the search engine used keyword matching to find results. For example, extensive prior experience with search engines that highlight matching keywords in abstracts (e.g., Google Scholar) in response to users' search queries can reinforce such assumptions among the users. In addition, participants' domain knowledge useful for judging which of the returned papers are relevant may have led them to notice a set of keywords the inclusion of which strongly signifies the relevance of a paper. In contrast, the analogical search results often seemed to not feature such directly similar terms and this contributed to the difficulty of judging whether a result is relevant and how: "I find these papers not very related to my search query at first. It'd be better if you can use some graph or some pictures to indicate how these papers can relate to my keywords" (P5); "I'd not consider... (because) they are totally different, right? They look irrelevant... until I think about it I can realize that it's useful. But if you give me the paper, at first I don't realize that" (P3).

While it may not feel as compelling or natural to participants, formulating and abstracting queries at a high level may lead to searching more distant results that are analogous at a higher level. For example, by querying "detect personal protective equipment" instead of "personal protective equipment construction," P5 found novel mechanisms of detection, such as general image segmentation algorithms or an approach to monitoring heat in the context of fitness training not specific to construction sites and personal protective equipment but nonetheless useful for creative adaptation. Querying "scheduling tasks" instead of "task complexity" for P3 resulted in finding scheduling algorithms in distributed computer systems that otherwise P3 would not have encountered, while "assigning tasks" led to novel auction mechanisms which made her think about a system in which each power plant operator can bid for a task as opposed to being assigned one. Schematically, fig. 5.13 shows how formulating queries at a higher level of abstraction than specifying the problem context in full details (Ⓐ → Ⓑ) may lead to discovering novel mechanisms that are relevant at the high level of abstraction, and in more distant ways from the original problem formulation (Ⓑ → Ⓒ).

5.6 Design Implications

From both the case studies' and Study 1's participants' reflection on the challenges of interacting with analogical search results, common themes emerged. Here we present three design implications for future analogical search systems synthesized from these results. We use subscripts to denote which study participants participated in when appropriate.

5.6.1 Support purpose representation at different levels of abstraction

Analogical search engines should support users to formulate their purpose queries at different levels of abstraction. Additionally the search engine may prompt users to consider abstracting or specifying their purpose queries in the first place, and explain how it might help bring new insights into their problems. As seen in the case studies (Section 5.5.2), scientists recognized their purpose queries may be represented at multiple levels, but prior experiences with similarity maximizing search engines may also have anchored them around pre-existing rigid formulation of purposes. Prompting users to consider their research problems at multiple levels may work against this rigidity, and providing candidate suggestions at varying levels may further reduce the cognitive demand. Moving up on the hierarchy to abstract purpose queries may be possible through removing parts of the query words that correspond to specific constraints, or by replacing them with more general descriptions. For example two participants of Study 1 had an identical purpose representation at a high level ("facilitate heat transfer") despite the differences in materialistic phases targeted in each purpose: solid material and semiconductors for P1$_{\textit{Study 1}}$ and liquid thin films for P3$_{\textit{Study 1}}$.

Furthermore, we also observed that looking for only the exact match of a purpose can lead to missed opportunities. For example, although "fins represent a different idea for transferring the heat" and "they (fins) don't match in terms of the scale – macro, not nano," it nevertheless made P1$_{Study\ 1}$ wonder "what if we could design nanoscale wall structures that act like fins that convert heat to mechanical energy?". A corollary to this observation is that sometimes the superpositions of misalignment with just the right amount can lead to interesting results. For P4$_{Study\ 1}$, a paper presenting experimental techniques for piezoelectric properties was interesting despite its misalignment such as [*simulation*-based] (source) ↮ [*experimental*] (analog) and [*dielectric properties*] (source) ↮ [*piezoelectric properties*] (analog): "Though it's an experimental study, it's very close in terms of the material and phenomenon so likely to be helpful. Because we might be able to pick up some trends like, if we increased the temperature, the dielectric response gets stronger or weaker, inferred from the experimental piezoelectric responses, which can then be used to corroborate simulation results or help configure its parameters" (P4$_{Study\ 1}$). However, too much deviation seemed detrimental to its potential for inspiration: "[Molecular dynamic simulation] is the same tool, but (this paper studies) [thermal] (not [dielectric]) properties on [polymer composites]... [polymer composites] are harder to model" (P4$_{Study\ 1}$). In sum, analogical search engines should support not only the capability to 'narrow it down' with specific constraints, but also ways to relax them to broaden the search space when suitable, thus making feasible the sweet spot between too little (i.e. similarity maximization and trivial matches) and too much deviation (i.e. critical misalignment and unusable analogs).

5.6.2 Support iterative steering from critical misalignment and towards generative misalignment

Analogical search engines should recognize that important constraints may be discovered by users only after seeing misaligned analogs, and support this discovery process by presenting effective examples of misalignment to users. Analogs that deviate on some aspects of the source problem but preserve important relations may be particularly conducive to analogical inspiration that opens up not just individual solutions, but entirely new domains of solutions. However at the same time scientists also found it challenging to know how to come up with effective search queries because combinations of misalignment can sometimes lead to an unintended intersection of domains: "I feel like I'm tricking the machine because [thin film] is often used with [solids], and the term [pressure] also appears a lot in [manufacturing]... so combining them gives a subset of papers concerned with heat transfer in solid materials and in manufacturing" (P3$_{Study\ 1}$); "on Google Scholar also, I get a lot of polymer strings and get (irrelevant) results like *we use an [electric] device to study [vibration and stress] of [polymers]...* the machine is picking up [electric] and [properties] such as vibration and stress in the context of studying polymers but what I really want is [electric properties] of [polymers] *not* [electronic devices] to study the [mechanical properties] of [polymers]" (P4$_{Study\ 1}$). Nonetheless, seeing misaligned analogs can be an effective way of reasoning about salient constraints and reflecting on hidden assumptions. For example, while evaluating papers about designing microelectrode arrays, P6$_{Study\ 1}$ said: *"Now I think about this (result), I assumed a lot of things when typing that search query... though impedance and topology are my main focus in microelectrode arrays, the coating, size, interface between a cell membrane and electrodes/sensors, biocompatibility, softness of electrodes, fabrication process, material of the platform: silicon or polymer or graphene, form factor: attaching electrodes to a shank-like structure or a broom-like structure, degree of invasiveness, are all part of the possible areas of research and it makes sense that they showed up – there is no way the machine would have known that from my query."* This excerpt illustrates how knowing what the necessary specifications are and which constraints need to be abstracted to cast a wide-enough net to catch interesting ideas appeared to be a difficult task for scientists, especially when they had to recall important attributes rather than simply recognize them from examples of misalignment. Prior work in cognitive sciences also show how dissimilarity associated with various factors in analogical mappings [90] can pressure working memory [262], increase cognitive load [240], and increases response time taken to produce correct mappings for analogy problems [144]. Therefore, analogical search engines should help to reduce the cognitive effort required in the process, for example by proactively retrieving results that are 'usefully' misaligned such that searchers can better recognize (as opposed to having to recall) salient constraints and refine their problem representation. This process is deeply exploratory [218, 269, 278] in nature, and suggest the importance of both providing end-users a sense of progress over time [244] as well as adequate feedback mechanisms for the machine to adjust according to the changing end-user search intent [145, 222, 223]. For example, while the machine may 'correctly' recognize a significant anaogical relevance at a higher level of purpose representation and recommend *macro*-scale mechanisms to a scientist who studies *nano*-scale phenomena (P1$_{\textit{Study 1}}$) or solid and semiconductor-based cooling mechanisms to a scientist in liquid and evaporative cooling systems (P3$_{\textit{Study 1}}$), these analogs may be critically misaligned on the specific constraints of the problem (i.e. the scale or materialistic phase) and thus considered by end-users as useless and even harmful.

5.6.3 Support reflection and explanation of analogical relevance

Throughout the process of analogical search, human-AI coordination is critical for success, and an important aspect is how deeply the human users can reflect on the retrieved analogs [107] and recognize how different notions of relevance may exist for their own problem context, despite potential dissimilarity on the surface. Looking at previous examples of the tools and techniques developed for targeted reflection support may be useful to this end. For example, ImageCascade [155] provides intelligent support such as automatically generated mood-boards and semantic labels for groups of images to help designers communicate their design intent to others. Another system, Card Mapper, visualizes relative co-occurrences of design concepts using proximity in the design space [62]. Similarly representing the space of analogical ideas using spatial encoding of similarity between two analogs, or designing information that supports getting a sense of the space of search results — e.g., semantic category labels similar to ImageCascade's or the distribution of the domains that analogs are pulled from — may be an avenue for fruitful future research. The explanation of relevance is also important especially when there is a risk of early rejection (§5.5.2). Using examples from the case studies, one approach to explaining relevance might be to surface a small number of core common features between an analog and a problem query. Such common features were considered useful by scientists for making analogical connections, and they could creatively adapt them for their own research problem context. When common features are not directly retrieved, generation of more elaborate explanations may be required. We refer to [19, 34, 230**?**] for those interested in future design considerations of automatically generated recommendation explanation. Further complementing the direct explanation of relevance approach, techniques such as prompting or reminding the searchers of previously rejected or overlooked ideas may also trigger deeper reflection and delay premature rejection of the ideas based solely on their surface dissimilarity. Participants from both studies commented that the critical first step towards analogical inspiration may be raising similarly enough attention and interest above the initial 'hump' of cognitive demand. Gentle reminders (e.g., "Ask me later if this would be interesting and also provide a list of items" – P1$_{\textit{Case Studies}}$) or resurfacing previously rejected papers in light of new information (P1$_{\textit{Case Studies}}$, P3$_{\textit{Case Studies}}$) may help with users cross this barrier.

5.7 Discussion

5.7.1 Summary of contribution

With the exponential growth of research output and the deepening specialization within different fields, encouraging analogical inspiration for scientific innovation that connects distant domains becomes ever more challenging. Our human-in-the-loop and fully automated analogical search engines represent an approach for supporting such analogical inspirations for challenging scientific problems. We have demonstrated in Study 1 that our human-in-the-loop system finds novel results that participants would be unlikely to encounter from keyword-based search, and that these results lead to high levels of creative adaptation. Through a mediation analysis we also showed that this success was driven by the analogical search engine's ability to find *partial* purpose matches (e.g., matching at the high-level purpose but differs at the low-level details). We saw the nuanced effects of partial purpose alignment on the results' goodness as analogs for inspiration. Through qualitative observations, we described how certain attributes of analogical mapping were perceived as more salient by participants, and that mismatches on them can have either a positive (i.e. generative insights) or a negative impact (i.e. critical misalignment) on creative adaptation. In contrast, keyword-based search resulted in more *full* purpose matches and a higher level of direct application. The value of keyword-based search and analogy-based search thus may complement each other, while keyword-based search can help researchers find 'exactly that', analogy-based search can help researchers to switch from a preservative mode (i.e. aiming to find results with maximal similarity to the query) to a generative mode (i.e. aiming to find analogs that are relevant despite the surface dissimilarity) of searching, and ultimately lead them to recognize unusual relations and come up with ways to creatively adapt existing ideas for novel domains.

We also demonstrated how improving the sequence-to-sequence purpose and mechanism identification model can remove the human-in-the-loop but maintain a similar level of accuracy on purpose-match by human judges. This improvement enabled us to develop a fully automated analogical search system to use as a probe to study searchers' more natural interaction with analogical results. Through a series of evaluation we first show that our automated analogical search pipeline can emulate human judgment of purpose match and that it finds partial purpose matches in top ranked results with a similar rate compared to the human-in-the-loop system used in Study 1. Then through case studies we find generalizable challenges that future analogical search engines may face, such as early rejection of alternative mechanism ideas and the difficulty of abstracting and representing purposes at the right level. From our studies we synthesize design implications for future analogical search engines, such as supporting purpose representations at different levels of abstraction, supporting the iterative process of steering away from critically misaligned analogs and towards a fertile land of generative misalignment, and providing explanations on why certain analogical search results may be relevant. We envision that future studies will shed light on deeper cognitive sources of the challenges identified here. A fruitful avenue of research may be studying how the dual processing theory [135, 265] underlies or interacts with analogical search interaction. Studying also how simplification heuristics [182] may negatively bias interaction with analogical results and how it may be reduced for expert user populations may be an interesting future direction [41, 159].

5.7.2 Limitations and future work

Experimental design and improving its validity

Our findings have several limitations. First the design of our studies may be improved to increase the experimental validity. We believe that our coders of the ideation outcomes had a reasonable understanding of participants' research context from descriptions of current and past research topics, think-alouds with 45 papers, and end-of-experiment discussions, and that the procedure of coding reduced potential biases (e.g., the coders were blind to experimental conditions, relied on participants' statements of novelty and distance). Despite this, it is possible that they judged ideas differently from domain experts, for example coding more or fewer ideas as creative adaptation, or pre-filtering useful ideas in the human-in-the-loop stage. In addition, other quality dimensions such as potential for impact or domain-expert-judged idea quality are largely inaccessible within the studies presented here. Future research may improve on these limitations by iterating on the experimental design, collecting data for triangulating the results and capturing other quality dimension of the generated ideas.

Additionally, future work may add ablation studies to quantify the effects of human filtering in Study 1 on the ideation outcome as well as sensitivity studies to relate how much the increased token-level classification performance of trained models may reduce the burden of human filtering. Furthermore, additional experiments with baselines other than keyword-based search using the whole abstract will help pinpoint the potential advantages of representing and matching papers using extracted purposes and mechanisms. For example, Chan et al. [49] found that embedding all words from an abstract (using GloVe embeddings) resulted in retrieval of fewer analogical items than when extracted purposes and mechanisms were used. Replicating this result with additional approaches such as contextualized word embeddings and pre-trained language models (e.g., ELMo [205], BERT [67], and SciBERT [22]) will be valuable.

Potential sampling bias

The sampling strategy in Study 1 was purposefully unbalanced, where analogical papers were sampled twice as much as keyword papers to ensure participants' exposure to sufficiently diverse results. This was crucial for uncovering potential benefits and challenges of our analogical search engine and investigating its viability. This ratio was chosen purposefully, to balance the statistical power for detecting potentially significant differences between the conditions, while also limiting the number of papers that each participant had to review. Given the cognitive burden of reviewing a paper while thinking aloud, we decided on 45 in total with the 2:1 ratio to fit the practical time limits of interviews. However, this may have led to unanticipated effects on ideation outcomes despite having accounted for the difference in sample sizes by measuring the outcomes in ratios. For example, when the results were combined into a single blinded list, the over-representation of analogical results over more purpose-aligned keyword results may have shifted the users' overall perceived value of the list to be more or less positive. Users' perception of diverse results may have been further affected by their relative over-representation. For example, increased cognitive load for processing analogical mapping [103, 104, 240] may suggest that results that fully match on the purpose search query may have been perceived even more favorably than analogical results, due to a negative spill over effect from the rest of the papers in the list, which were less likely matched on the purpose. Investigating whether such factors led to compounding effects beyond our ratio-based measures of usefulness remains an open question for future work.

Controlling the diversity of search results

Our work is also limited by the lack of controllability in sampling the search results beyond purpose similarity. As described in §5.2.2, from pilot tests in our corpus we discovered that even close purpose matches of scientific papers already had high variance in terms of the mechanisms they proposed which allowed us to focus our approach to sampling based solely on purpose similarity. The simplicity of this approach also means fewer hyper parameters in the sampling mechanism compared to other approaches [121, 122]. However, all the approaches including this work thus far lacked a mechanism for explicitly controlling the diversity in retrieved search results which remains a fruitful avenue for future work. For example, prior research has uncovered the nuanced effects of distance (e.g., near vs. far sources of inspiration [48, 226]), suggesting the benefit of targeting analogs at different distance from the source problem for the right context. Future research may also uncover further complexities in the relationship between novelty and purpose-match. The result of our mediation analysis (Table 5.5) showed that the novelty of content among the search results in Study 1 was not a significant factor to the same extent that the three levels of purpose match was. However, the relationship between novelty and purpose match may be more complex than the levels of manipulation presented in this work. For example, [68] suggested a greater importance of novelty than usefulness for predicting creativity scores. Future work may design mechanisms to manipulate the variance in content novelty and alignment in the purpose-mechanism schema to uncover dynamics between the two that go beyond the results from mediation analyses presented here (§5.3.3). Furthermore, challenges with the abstraction of purposes remain open, for example how core versus peripheral attributes of research purposes may be identified, and how they may be selectively matched at a specific level of the conceptual hierarchy. Finally, not all query formulations are created equal in terms of their suitability for analogical search. We observed in the case studies that participants wanted to express different query intent via reformulation (§5.5.2). While participants could reformulate their search queries and examine the returned results from our analogical search engine in real-time, it was unclear whether and how specific query formulations may lead to more or less diverse results, and how subsequent queries may be updated after reviewing them. As such, systems that assist users in the potentially tedious process of query reformulation [270] (for example, by way of automatic query expansion [42]) in the context of analogical search will be important.

Studying the effect of larger context of scientific innovation on analogical innovation

Due to our focus on ideation outcomes, our results do not explain how these ideas may be integrated, developed, and shared across the research communities. Studying the lifetime of ideas that goes beyond their inception will deepen our understanding of the factors that currently make analogical innovation such a rare event in sciences (for example, Hofstra et al. suggested that more semantically distant conceptual combinations receive far less uptake [118]). Through interviewing our study participants and other colleagues in academia we found emerging structures related to this challenge. Our interviews informed us that in general the context in which a scientist exists – such as the scientist's role in a project, the maturity of a project, and the broader academic culture – can ultimately change how they interact with and seek analogical inspirations. For example a third-year PhD student studying chemical engineering commented "In the current stage of my project it's more about parameter-tuning – running multiple experiments at once and comparing which configuration works the best... If I were a first year PhD student maybe I would be in a broader field and exploration." In contrast, a PhD in biology who recently defended noted that "analogical inspirations would perhaps be more useful if you're looking for a postdoc or a faculty position."

In addition, the underlying career incentive structures in academia may also affect researchers' perception of and openness to analogical inspirations. One of the study participants commented "since I'm already a third year PhD student and my project is further along and more firmed up, I'm not really looking for really far inspirations... first we push the specific way we have in mind with many iterations on the experiments until, say, publication." In addition to the career-wise incentives there are other types of competitive resourcefulness (e.g., social resources such as the advisors' and colleagues' expertise that participants can easily tap into; physical and other forms resources such as tangible artifacts like previously developed code packages or experimental processes and setups). These factors can influence scientists' perception of their advantage and lead them to interpret analogical inspirations as more or less useful, feasible, and directly applicable to their research. This observation is further suggested by survey results that asked our participants: "*Could this paper be useful to you?*," their ratings were significantly higher for keyword papers than analogy papers despite them having come up with creative adaptation ideas more often with analogy papers. Therefore, future work that studies incentive structures, the quality of ideation outcome, their feasibility, the differences in research context e.g., frames of research contribution such as discovery-oriented vs. novel system development-oriented, and taking a longitudinal observation of the variation in such factors will add a significant depth to our understanding.

5.8 Conclusion

In this paper we present our novel human-in-the-loop and fully automated analogical search engines for scientific articles. Through a series of evaluations we found that analogous papers that our systems retrieved were novel and useful for sparking creative adaptation ideas. However, significant work is needed to continue this trajectory, including additional understanding of the context and incentives of scientists as well as advances in the data pipeline and interaction methods beyond those described here for a system to maximize its real-world impact.

We imagine a future in which scholars and designers could find inspirations based on deep analogical similarity that can drive innovation across fields. We hope this work will encourage scientists, designers, and system builders to collaborate across disciplinary boundaries to accelerate the rate of scientific innovation.

A Reproducibility

Training and validation datasets. The original annotation dataset from [49] also includes Background and Findings annotations which we exclude due to their relatively high confusion rates among the annotators with the Purpose and Mechanism classes and to balance the number of available training examples per annotation class.

Model parameter selection. We experimented with changing the model capacity relative to the signal present in the training dataset by tuning the number of hidden layers and the nodes used in each model architecture. For Model 1 we found a hidden layer of 100 nodes was sufficient. We optimized this model using the cross-entropy loss and the Adam optimizer [148] with a 0.0001 learning rate. For Model 2, we found three hidden layers with 256 nodes led to an improved accuracy on the validation set. We trained this model with an L2 regularizer ($\alpha = 0.01$), dropouts with the rate of 0.3, and the Adam optimizer with a 0.001 learning rate.

Span-based model architecture. We adapt SpanRel [130] as architecture for the span-based Model 2. SpanRel combines the boundary representation (BiLSTM) and the content representation with a self-attention mechanism for finding the core words. More specifically, given a sentence $\boldsymbol{x} = [e_1, e_2, \cdots, e_n]$, of n token embeddings, a span $s_i = [\omega_{s_i}, \omega_{s_i+1}, \cdots, \omega_{f_i}]$ is a concatenation of the *content representation* $\boldsymbol{z_i}^c$ (weighted average across all token embeddings in the span; SelfAttn) and the *boundary representation* $\boldsymbol{z_i}^b$ of the start (s_i) and end positions (f_i) of the span:

$$\begin{aligned}
\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_n} &= \text{BiLSTM}(\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_n}) \\
\boldsymbol{z_i^c} &= \text{SelfAttn}(\boldsymbol{e_{s_i}}, \boldsymbol{e_{s_i+1}}, \cdots, \boldsymbol{e_{f_i}}) \\
\boldsymbol{z_i^b} &= [\boldsymbol{u_{s_i}}; \boldsymbol{u_{f_i}}] \\
\boldsymbol{z_i} &= [\boldsymbol{z_i^c}; \boldsymbol{z_i^b}]
\end{aligned}$$

We use the contextualized ELMo 5.5B embeddings[15] for token representation, following the near state-of-the-art performance reported on the named entity recognition task on the Wet Lab Protocol dataset in [130]. We refer to [130, 162] for further details.

Other parameters. We use GloVe vectors for input feature representation for Model 1 with 300 dimensions, consistent with the prior work [28, 160, 203]. For Model 2, we use the contextualized ELMo 5.5B embeddings as described above which have pre-determined 1024 dimensions. We use Universal Sentence Encoder (USE) [44] for encoding purposes. A USE embedding vector has pre-determined 512 dimensions.

[15] `https://allennlp.org/elmo`

Chapter 6: BioSpark

An LLM-based End-to-End System for Facilitating Analogical Design Sparks and Deepening Engagement

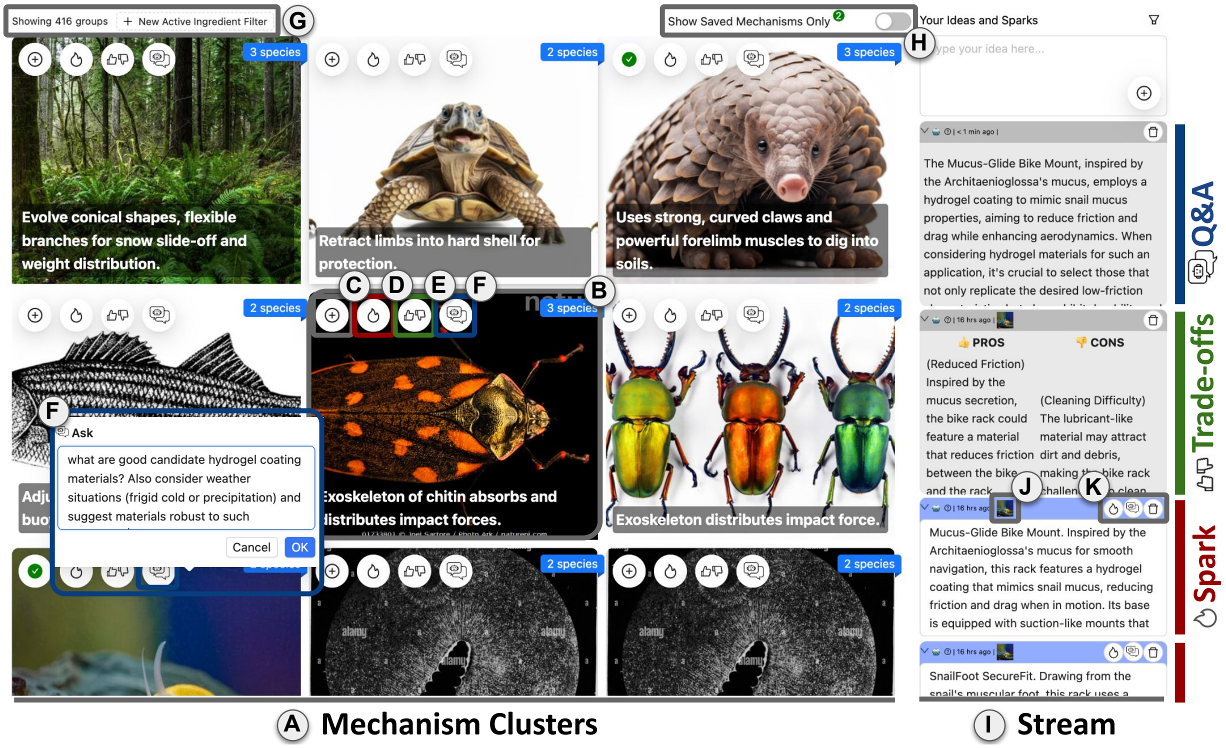
An earlier version of this work was previously published as an Extended Abstract in ACM CHI 2024 ([143]) and NeurIPS 2023 Creativity Workshop, and has been adapted for this document.

This paper presents BioSpark, a system for analogical innovation designed to act as a creativity partner in reducing the cognitive effort in finding, mapping, and creatively adapting inspirations from distant fields. While previous approaches have largely focused on the initial stages of identifying inspirations, often limited to a narrow set of hand coded data, BioSpark uses LLMs embedded in a familiar, visual, Pinterest-like interface to support users in more deeply engaging with inspirations across multiple stages of analogical innovation while avoiding fixation and over-reliance on AI-generated ideas. To do so we introduce several novel features, including a tree-of-life enabled approach for generating relevant and diverse inspirations; 'sparks' that connect inspirations to the source problem domain; tradeoff cards that scaffold user consideration of key constraints; and a free-form chat interface grounded in the context of the design problem and the inspiration to help users more deeply explore adapting inspirations. We designed and evaluated the effectiveness of BioSpark through a workshop with professional designers, a pilot study with a functional prototype, and a controlled user study. Our results suggest that participants found value in BioSpark's potential to embed AI support into interfaces seamlessly, promote deeper engagement with inspirations, and augment human creativity, while mitigating the risks of hallucinations, loss of ownership, and idea fixation.

6.1 Introduction

Many innovations in design, technology, and science have been driven by people finding and adapting inspirations from fields distant to their own. Whether Vetruvius explaining how sound waves work through analogy with water waves [61], the Wright brothers designing a lightweight wing control mechanism based on a bicycle inner tube box [133], or engineers partnering with an origami expert to furl a solar array into a narrow rocket [183, 204, 281], such innovations have required their inventors to engage in a complex cognitive process of finding and creatively adapting inspirations that had limited surface similarities but deep structural similarities.

While such analogies may sometimes seem like 'lightning strikes' of serendipity, researchers have identified that analogical innovation involves several cognitive stages, all of which can require significant mental effort. Finding inspirations in distant domains is difficult because of the challenge of going beyond surface keywords and visual similarity to finding out-of-domain mechanisms that share a deeper underlying problem structure. Once encountered, determining which inspirations are relevant to solving the problem requires the inventor to map the inspiration to the source domain to understand how its mechanisms could be instantiated [87, 89, 93]. Often, the inspiration may not itself be used directly; instead it may identify a profitable design space that the designer might creatively adapt to the problem [140, 167]. For example, an inspiration of a simple paper crane might identify origami as a profitable design space but not be directly used in the complex design of a folding solar array. Finally, the particular way in which the design mech-



Evolve conical shapes, flexible branches for snow slide-off and weight distribution.

Retract limbs into hard shell for protection.

Uses strong, curved claws and powerful forelimb muscles to dig into soils.

Exoskeleton of chitin absorbs and distributes impact forces.

Exoskeleton distributes impact force.

what are good candidate hydrogel coating materials? Also consider weather situations (frigid cold or precipitation) and suggest materials robust to such

| < 1 min ago |

The Mucus-Glide Bike Mount, inspired by the Architaenioglossa's mucus, employs a hydrogel coating to mimic snail mucus properties, aiming to reduce friction and drag while enhancing aerodynamics. When considering hydrogel materials for such an application, it's crucial to select those that not only replicate the desired low-friction

| 16 hrs ago |

👍 PROS	👎 CONS
(Reduced Friction) Inspired by the mucus secretion, the bike rack could feature a material that reduces friction between the bike and the rack	(Cleaning Difficulty) The lubricant-like material may attract dirt and debris, making the bike rack challen clean

| 16 hrs ago |

Mucus-Glide Bike Mount. Inspired by the Architaenioglossa's mucus for smooth navigation, this rack features a hydrogel coating that mimics snail mucus, reducing friction and drag when in motion. Its base is equipped with suction-like mounts that

| 16 hrs ago |

SnailFoot SecureFit. Drawing from the snail's muscular foot, this rack uses a

Figure 6.1: The BioSpark interface is organized into the mechanism clusters panel (Ⓐ) and the stream (Ⓘ). Each cluster card (Ⓑ) consists of an image of the first (or user-selected in the cluster modal view) species in the cluster, its active ingredient description, action buttons, and a ribbon indicating the size of the cluster. When the card is clicked a modal view shows up, revealing more details about each mechanism in the cluster. The action buttons include: 'save mechanism' (Ⓒ), which updates the count in the badge for the saved mechanisms toggle (Ⓗ), 'spark' (Ⓓ) that generates new sparks of inspiration that build on the clicked mechanism and are diversified by previously generated sparks, 'trade-off' (Ⓔ) that generates a run-down of potential design trade-offs of using the clicked mechanism in the context of the design problem, and 'Ask AI' (Ⓕ) that opens up a pop up window with a text area for typing any requests (*e.g.*, follow-up questions about the mechanism). The stream panel (Ⓘ) includes system- and user-generated outputs such as sparks, trade-offs, and responses to user questions. Each spark card also includes helpful features such as a caret for expanding/collapsing the card, the timestamp of creation, a clickable thumbnail (Ⓙ) showing the source mechanism, which expands the modal view upon clicking it, and control buttons (Ⓚ) for further generating new sparks of the spark content, Q&A, and deletion. The content of each spark is directly editable.

anism is instantiated and used needs to be thought through, with limitations and trade-offs considered and mitigated [10, 228, 260]; for example, while the Wright brothers used the twisting of the cardboard box for inspiration, they had to find materials and use mechanisms that would support a manned plane in flight without tearing or being too heavy.

Supporting these complex needs in a single system has been challenging, with most approaches focusing on one or two stages of the process and largely limited to a small, hand-coded set of inspirations [45, 66, 97, 129]. Approaches to collecting inspirations at a larger scale have begun to appear [73, 121, 125, 140], but have mostly been limited to helping with the finding stage of analogical innovation. Relying on users to do the hard work of determining which inspirations could be relevant and how they could be adapted can lead to them not noticing or putting in the effort to go beyond surface similarities and try to understand how an inspiration could be used; as noted in Kang et al. 2022, "the critical first step towards analogical inspiration may be raising... enough attention and interest above the initial 'hump' of cognitive demand" [140].

In BioSpark we explore the idea of a LLM-powered computing system acting as a creativity partner to proactively help with the intellectual work of not only finding analogies but also transferring and adapting those ideas to the target domain. By doing so we aim to help free up the cognitive effort of users to engage in the creative process of exploring new design spaces and considering more ideas more deeply than they would be able to otherwise. A key goal of our approach is to augment human creativity rather than replacing it with AI-generated ideas or resulting in fixation on those ideas.

To achieve this, BioSpark explores several new design patterns for partnering AI with human analogical ideation, including:

- A tree-of-life enabled approach for generating new and relevant biological inspirations from a small set of 'gold standard' inspirations taken from AskNature;
- An analogical ideation interface leveraging familiar interaction concepts from designers' practice of browsing Pinterest and curating moodboards;
- Proactively generating 'sparks' that help users understand the mapping between inspirations and their design problem;
- Providing pro/con tradeoffs to scaffold users in considering key aspects of the design problem;
- Supporting a free-form chat interface grounded in the inspiration and the design problem context to help users more deeply consider inspiration mechanisms.

We instantiated BioSpark in a prototype system and evaluated and iterated on it through a formative study with 4 participants with design and engineering backgrounds as well as a user study with 12 participants of varied backgrounds. Our results suggest ways in which AI support can be embedded into interfaces to support and augment human creativity through deeper engagement with AI inspirations while addressing potentially negative effects such as hallucinations, lack of ownership and control, and fixation on AI prompts and generated ideas.

6.2 Related Work

6.2.1 Design by analogy

Throughout history, analogies have often driven breakthroughs in science, engineering, and design (*e.g.,* [61, 133, 183]). Yet, analogical innovation in human minds has proven rare due to the cognitive challenges involved with the underlying analogical processing. One challenge is the high sensitivity to surface-level similarity during retrieval from memory that favors analogs with shared visual or keyword similarities over the ones that share a deeper underlying structure [89]. In addition, the heavy cognitive load incurred during analogical processing, even with just a few relations, significantly burdens working memory and leads to performance degradation [91, 94, 105]. To support people with analogical processing, researchers have designed various systems for analogy retrieval. One thread of research here focuses on modeling analogical relations, albeit in limited scopes. This includes system based on the structure-mapping theory [80, 81, 87], multiconstraints theories (*e.g.,* [120], connectionist designs [117, 127], and rule-based approaches [11, 39, 40]). Many methods involve labor-intensive processes, such as the WordTree methodology [167]. Additionally, numerous systems depend on hand-coded and meticulously structured data, the curation of which is often resource-intensive (*e.g.,* [97, 260]).

Recent work in computational methods for finding analogical inspirations at scale have shown promising results using a significantly simplified schema (*e.g.,* the purpose and mechanism schema in [121, 140]) with just a fraction of data (*e.g.,* [49, 121, 140]). However these systems primarily focus on facilitating the discovery of potential analogies and do not extend support to the subsequent, intricate stages of design that follow. This involves navigating potential limitations or trade-offs, which are essential for the successful transfer of these analogies in real-world scenarios [10, 228, 260].

6.2.2 Bioinspired design

One particularly relevant thread of research in design by analogy focuses on finding inspirations in biological organisms and systems [129]. However, prior approaches have been limited due to their reliance on costly manual curation (*e.g.,* AskNature [66] or DANE [97]; the researchers of DANE found that redescribing a single biological organism in the Structure-Behavior-Function framework can take approximately ~40-100 hours per model). Alternative approaches demonstrated the feasibility of using crowdsourcing to power supervised learning for identifying scientific articles with biomemetic inspirations (*e.g.,* [258, 280]), but the cost of curating high-quality annotations presented a significant bottleneck for scalability. Yet another line of research has explored rule-based (*e.g.,* [54]) or data programming [74] approaches, and showed promising results, albeit potential concerns of their generalizability and scalability.

Our iterative tree-of-life-based algorithm for expanding the mechanism dataset builds on these threads of research, while also leveraging recent advances in AI, such as Large Language Models (LLMs), that present promising new opportunities for designing scalable approaches for bio-analogy generation. However, naively prompting LLMs in a zero-shot manner may still result in limited diversity on abstract concepts [57]. One promising avenue of research here is exploring knowledge-augmented or knowledge-guided prompting for the purpose of increasing conceptual diversity in generation output. Previous work in this area (*e.g.,* [16]) has explored this idea in the domain of factual Q&A, and has shown increased factuality in responses to questions with simple answers (*e.g.,* "*Where did Alex Chilton die?*") when a relevant knowledge graph was traversed first to retrieve relevant facts to contextualize LLM prompts.

6.2.3 LLMs for ideation and co-creation

Recent advances in LLMs also suggest the potential for scalably augmenting analogical innovation for users throughout the entire cognitive process, from finding potential analogical inspirations to mapping them to the problem domain to helping users more deeply engage with their mechanisms and trade-offs. LLMs have shown the capability to infer specific analogies and to generate ideas relevant to a design goal (*cf.* [267]). They also can serve as more flexible natural language processing components in an interface, allowing for powerful interface augmentation approaches (*e.g.,* [13, 78, 142, 169] as well as direct interfaces using chat-based dialog (*e.g.,* [1, 178, 195]).

However, studies examining the use of LLMs and generative AI in the creative process have shown that improperly incorporating LLMs into the creative process can end up doing more harm than good. Using generative AI systems such as image generation (*e.g.,* Midjourney) or text generation (*e.g.,* ChatGPT) has been shown to lead users to become more fixated rather than more creative [261]. Several core properties to LLMs have been identified as potentially problematic, including tendencies for inaccurate inferences and hallucinations, user fixation on the initial prompts they enter, and overly accepting the results of AI-generated ideas rather than adapting them or using them to further explore the design space [147, 261]. These results suggest a more nuanced approach to incorporating LLMs and AI into the analogical innovation process may be needed.

6.3 Formative Studies

6.3.1 Workshop with Professional Designers

To better understand how our system could engage with the needs and constraints of designers, we conducted a day-long workshop with professional automotive designers. During the workshop we conducted a design probe with several worked examples of analogical inspirations that had the potential to help them in various phases of their design process. We also discussed their existing workflow and how a tool could be usefully integrated into that flow.

Several findings from this workshop informed the design of BioSpark. First, designers' frequent, often daily, practice of scrolling through inspirations on online sites such as Pinterest suggested an opportunity for a familiar interface and behavior that our system leverage. Relatedly, the designers expressed a desire for visual representations of inspirations (consistent with previous findings such as [166]) which reinforced the value of a visually dense, scrollable interface and suggested that we needed to support ways of creating visual representations for the textual inspirations the system would provide. Finally, designers stressed the importance of their design brief and the time constraints of generating and iterating on concepts to address it, which informed the overall flow of our prototypes in first setting a design problem to be addressed and reducing the cognitive effort and time in helping designers adapting inspirations to address those problems.

6.3.2 Functional Prototype Pilot

Based on the workshop findings we developed an initial functional prototype to test with participants. The prototype included a component to generate biological inspirations that could address a design problem, for example given the need to 'design a secure bike rack for sedans' the system provides inspirations such as shapeshifting algae and parasitic copepods that attach to their hosts using friction-based mechanisms (details on the method used for inspiration generation are provided in the full system description). Each inspiration was represented visually with an image retrieved from online search queries (Appendix A.3).

To assist designers in interacting with these inspirations, the prototype (fig. 21) included features for explanation, comparison, combination, and critique of mechanisms using GPT4-generated content within the interface (Appendix A.4). These features allow users to delve into individual mechanisms, directly compare mechanisms, synthesize new ideas by combining mechanisms, and critique their own design ideas.

For the pilot we recruited four participants (all male, avg. 29.0 age: SD: 7.39) with backgrounds in:

mechanical engineering, CAD (2), visual, communication & UX design (1), and creative coding and visual design (1). The study took place virtually on Zoom and lasted about 45 minutes.

Each participant was given one of two randomly selected design briefs (1) '*design a secure bike rack for sedans*' or 2) '*design improvements for sedans on slippery road conditions*') and asked to use the prototype to find as many inspiring mechanisms to come up with new design ideas as they could in 20 minutes. Participants shared their screen and thought aloud during the task, and were encouraged to use all of the system features to come up with at least two new design ideas for the brief. After the design tasks participants were interviewed about their experience with the prototype.

Overall, participants found the prototype valuable, with all participants finding several mechanisms that inspired new design ideas. Some examples included: the coiling of octopus tentacles and lizard tails inspiring bike rack components that could expand and contract with turbulence; and scale and fur arrangements in rodents inspiring groove patterns on tires that would create more downforce on slippery roads.

However, all participants noted a similar theme in terms of desire for additional support, which was greater help in **engaging with the inspirations to understand and adapt their 'active ingredients'**, the core abstraction underpinning how each mechanism actually works. For example, while mechanisms commonly included whole body images of organisms, participants commented that showing active ingredients such as *"feathers and feather constellation patterns rather than the whole body image of birds"* (P1); and *"curvature of the tail and claws in geckos"* (P3) would be better for transferring a concept.

Relatedly, participants also wanted additional support for **envisioning how mechanism inspirations transferred into their target design domain**. Participants also commented that incorporation of mechanism ideas into focused areas of target domains may facilitate engagement with more distant ideas that have potential for high impact when iterated on. P2 said: *"I think more targeted treatment, focused in scope would be good. Maybe we (users) can apply scales to only parts of the vehicle, such as wheel rims or front grilles or spoilers"*.

Finally, participants noted the desire for **deeper exploration of mechanism ideas**, such as from P1: *"'Bike rack' and 'slime' are somewhat contradictory but it (slime mechanism) makes me think about the attachment aspects of the design... maybe new ideas around loading and unloading of bikes that have dynamically adjusting surface friction... I'm going to click "explain" on slime... (after the detailed explanation loads) I wish I could know more about the lubrication mechanism aspect of slimes"*.

6.4 BioSpark

6.4.1 Design Goals

Together, the design workshop and pilot study with an initial functional prototype informed the development of BioSpark's design goals for a system aimed at helping augment engagement with analogical inspirations with the potential of being incorporated into designers' existing practices. Specially, we aimed to help users: 1) Find analogical inspirations; 2) transfer them into their target domain; 3) understand the active ingredients of their mechanisms; and 4) explore those mechanisms more deeply.

We instantiated these design goals in BioSpark, with the high level intention of acting as a creative partner in the analogical design and innovation process beyond simply finding inspirations. The design of BioSpark is premised on reducing the cognitive workload of users by proactively helping the user see multiple connections between inspirations and their own problem domain, potentially sparking new ideas in the design spaces thus unlocked; and/or by reducing the effort for the user to engage with inspirations more deeply by considering their tradeoffs and design constraints or by asking for more information about their details or characteristics.

Implicit but nonetheless critical in these design goals is the need for the system as a creative partner to not take over the human element of creativity and ownership of the resulting ideas. To address this we frame AI-generated content as intermediate products that are ephemeral and editable and minimize explicit context switching by embedding AI actions in the flow of the system. Specifically, the AI provides its suggestions on mappings between inspirations and the problem as idea 'spark' cards that are added to a sidebar whenever the user saves an inspiration; 'tradeoff' cards that contextualize the pros and cons of an inspiration's mechanism within the problem domain; and Q&A cards that allow the user to submit freeform queries to the LLM which are automatically contextualized with the problem and inspiration contexts. As the user engages with the system a typical flow involves them perusing and saving inspirations, and engaging with the cards in the sidebar to more deeply consider particular mechanisms or the design spaces they represent.

In the following sections we describe in more detail the design of the system, starting with groundwork and infrastructure for generating inspirations and their active ingredients; how we find visual representations for them; and then a detailed discussion of the system interface features including a scenario walkthrough.

6.4.2 Biological Mechanisms and Active Ingredients Dataset Generation

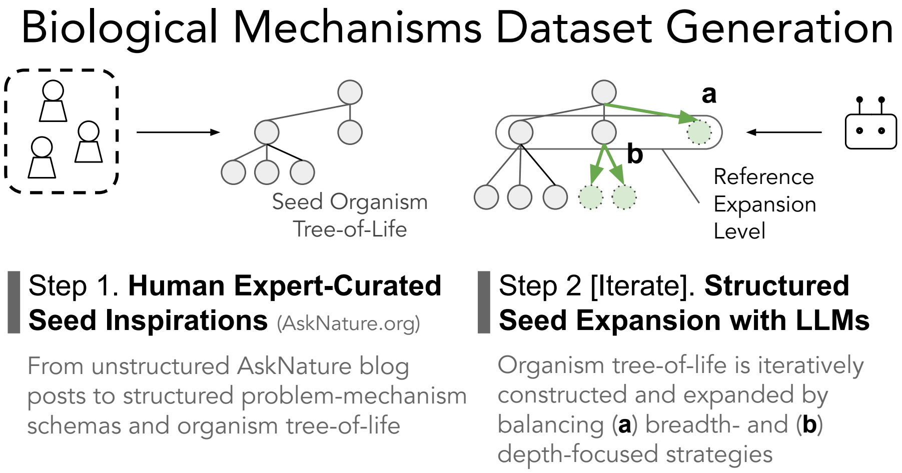
Iterative Tree-of-life Construction and Expansion

The first stage of the BioSpark dataset pipeline is running an iterative algorithm for generating a diverse set of mechanism inspirations, starting from a small set of expert-curated AskNature seed inspirations. Prior approaches in this research area can be described as either directly retrieving information from the Web (through various means such as crowdsourcing [258, 280], rule-based programs (*e.g.,* [54]), data programming [74]) or generating information from LLMs using prompt augmentation, such as by adding directly relevant facts retrieved from a knowledge graph [16]. In comparison, while our approach also explores prompt augmentation for LLMs, it also differs with prior approaches in new ways. First, our approach conceptually follows the hierarchy-based expansion method such as the WordTree method [167] that demonstrated how the up-then-down traversal on the abstraction hierarchy in structured brainstorming settings could lead to novel insights. Here, we design a similar approach for structurally expanding a seed dataset, but unlike the focus of the prior work on designing a collective process that involves human ideators and the word abstraction hierarchy, our approach applies it to LLMs and the Tree-of-Life[1] hierarchy to design a scalable algorithm.

Second, our approach also differs in terms of its use of proxy data for expansion. Unlike prior approaches that aimed at retrieving directly relevant facts or scientific articles, we anchor our expansion algorithm on species of nature as a mediator for new spaces of mechanisms, as species often adapt to changing natural environments by evolving with new mechanisms.

In order to generate a diverse set of biological mechanisms from curated blog posts on AskNature, we design a two-stage process. In the first stage, we start by structuring the natural text in AskNature blog posts into problem-mechanism-organism schemas through HTML parsing of the blog text and GPT4-based extraction (Details in Appendix A.1). For each organism in an extracted schema, we then construct a 7-level tree-of-life hierarchy consisting of the {`domain,kingdom,phylum,class,order,family,g`

[1] `https://en.wikipedia.org/wiki/Tree_of_life`


Biological Mechanisms Dataset Generation
Seed Organism Tree-of-Life
a
b
Reference Expansion Level

Step 1. **Human Expert-Curated Seed Inspirations** (AskNature.org)

From unstructured AskNature blog posts to structured problem-mechanism schemas and organism tree-of-life

Step 2 [Iterate]. **Structured Seed Expansion with LLMs**

Organism tree-of-life is iteratively constructed and expanded by balancing (**a**) breadth- and (**b**) depth-focused strategies

Figure 6.2: We design a two-stage pipeline for generating a diverse set of species and their mechanisms. In Step 1, we start from an initial set of seed mechanisms and species extracted from AskNature. In Step 2, we use LLMs to iteratively construct the tree-of-life hierarchy using species generated up to that point. We sort the hierarchy to identify sparse branches that have maximal diversification opportunities, and traverse them in a depth- or breadth-focused manner to generate further species and their mechanisms.

`enus,species`} levels. To do so in our initial investigation we explored available resources such as the Global Biodiversity Information Facility API[2], Catalogue of Life [36], or the Encyclopedia of Life [75], using canonical species names retrieved from the Darwin Core List of Terms[3] for corresponding organisms in our problem-mechanism-organism schemas. However, the limited coverage, data consistency, and API availability of these tools prevented their adoption in our pipeline. We also investigated Wikipedia as a source of ground-truth taxonomic information given the name of a species, which exist for some of the organism articles we tested in the form of the 'biota' information box that appears on the right-hand side of the corresponding organism Wiki article, but this data was not readily available at scale.

Instead, we wondered if given the scale of biology articles GPT4 has processed during pre-training if the model could reliably generate the 7-level taxonomy when prompted using only the name of the species. To evaluate the feasibility of this idea, we constructed 90 'ground-truth' taxonomies using Wikipedia's 'biota' scientific classification information box. To our surprise, this evaluation showed satisfactory accuracy levels for use that were ranging between 94.4% – 100% for each of the 7 levels on the taxonomy (Table 2, Appendix C).

In the second stage, using the constructed initial tree-of-life hierarchy, we identify sparsely populated branches as diverse mechanism generation opportunities. Indeed recent data shows that there are significant branching opportunities on the tree-of-life hierarchy with its exponential growth (for example, the `Genus` level has an estimated number of 310K members [214], while the number in the subsequent level, `families`, is estimated at 8K [184]. The number of direct children in each node also varies significantly, with the majority of genera within non-avian reptiles hosting a single species each. In contrast, insect genera, for example *Lasioglossum* and *Andrena* boast over 1,000 species each, and the of flowering plant known as *Astragalus* includes more than 3,000 known species [271]). In order to exploit these opportunities, we adopt both breadth- and depth-first strategies for finding new mechanisms.

[2] `https://www.gbif.org/developer/species`
[3] `https://dwc.tdwg.org/list/#dwc_Organism`

In depth-focused diversification, we traverse the tree-of-life hierarchy of organisms (that include all of the organisms in the dataset up to that point) and filter its nodes at a given depth (*e.g.,* 'order'). We then sort the filtered nodes in the ascending order based on the number of its children, such that the first node in the sorted list has the least number of children. This node represents the highest opportunity for vertical (*i.e.,* generating its children species) exploration, due to its sparsity. Using the first five nodes in the list as candidates for depth-focused expansion, we design a prompt (fig. 12 in Appendix A.2) that requests GPT4 to generate children nodes and their mechanisms that are applicable to the problem in the schema. In breadth-focused expansion, we design a prompt (fig. 13 in Appendix A.2) that requests GPT4 to generate sibling nodes of a given node, excluding the previously generated nodes to avoid duplicate generation. The output of the prompts contains natural text descriptions of the step-by-step execution which we then feed into a simple prompt for extracting and structuring the data into a JSON array.

Active Ingredient Extraction

Informed by our design goals and formative study findings, one purpose for BioSpark design and development was enhancing its data focus around *active ingredients*, or transferable concepts in mechanism descriptions, to streamline downstream user interaction along the active ingredients. To this end, we further developed the BioSpark backend dataset pipeline to process the mechanisms in the dataset to extract active ingredients and organize them in semantically meaningful clusters.

To extract active ingredient descriptions from mechanisms, we designed a prompt (Appendix F.1) for a GPT4 model (`gpt-4-turbo-preview`) that consists of a system message and a user message. Using the system message we instruct the model with three criteria for identifying active ingredients. Through pilot testing we found that active ingredients that are short (*i.e.,* 15 words or less) are easier to skim and increase the cluster separation by excluding secondary features of commonality among the species. We also found that an active ingredient description with a verb or verb phrase is easier to parse, as it often presented the information in the form of 'what acts upon what'; thus we explicitly instructed the model to focus on this information in the system prompt. Finally, we provide some examples of concrete active ingredients and instruct the model to also focus on those elements in extraction. The prompt also takes in the description of a mechanism as its user message to apply these rules.

Recursive Clustering of Active Ingredients

In order to organize the active ingredients in semantically meaningful groups, we create a recursive clustering algorithm. In our pilot testing we found that directly applying the off-the-shelf clustering algorithm such as the density-based algorithm `DBSCAN` resulted in two challenges: 1) there remained a large (often ~20 - 40% of the total) cluster of seemingly diverse mechanisms, and 2) that cluster appeared at the beginning of clusters, contributing to potentially mis-orienting users as to what each cluster represented.

To address these challenges, our algorithm iteratively targets the "could not cluster" cluster, which conventionally is denoted as the "-1" cluster in many off-the-shelf clustering algorithm output including `DBSCAN`, to re-cluster among its members. To this end, the algorithm successively re-clusters using members of the -1 cluster generated from the previous run of clustering, with a gradual relaxation of the minimum distance parameter (denoted as ϵ in `DBSCAN`), that decides the global sensitivity to cluster separation. We set this parameter to start off at .3, and gradually increases with a decreasing slope, meaning the step increase added to the parameter at each run decreases over time *i.e.,* by a factor of 1.1, until no improvement in clustering results could be made even after the epsilon has increased. Intuitively this works by taking out the relatively straightforward clusters (*i.e.,* groups of mechanisms with very similar surface text forms, resulting in very low distances among them in the embedding space) early on, and by sufficiently lowering the sensitivity threshold subsequently in order to identify less obvious yet coherent clusters (*i.e.,* groups of mechanisms that look different in the surface text form yet are semantically related, resulting in relatively higher distances among them) in the distance terrain of the remaining mechanisms. After the final run of the algorithm, if there remains any -1 cluster mechanisms, they are broken down into a series of singleton clusters, and appended to the end of the list to orient users towards clusters of multiple members for high information density clusters.

6.4.3 Mechanism Image Retrieval

Searching on Google+Adobe Stock Images

In order to visually represent each mechanism, we consider retrieval- and generation-based approaches. In our formative study, we found that generated mechanism images needed more improvements such as identifying effective time and vantage points as well as effective zoom or scales that are specific to each mechanism to be effective. We also found that participants liked seeing the portrait of a species as the first step before engaging more deeply with its mechanisms. Furthermore, we similarly observed from the design of expertly curated AskNature.org webpages that the use of a close-up and centered portrait of a species often creates a striking visual that also invokes curiosity.

Therefore, we decided to use a retrieval-based approach to visually represent the mechanism, and specifically to focus on finding effective animal portrait images. We use Google Search and Adobe Stock Images for this purpose, each of which had strengths and weaknesses. We found that directly searching on Google using its API[4] with an animal name or mechanism description query often resulted in images such as book covers or graphs in relevant research paper, which were ineffective (this can also be seen in our formative study system interface, fig. 21). On the other hand, using Adobe Stock Images[5] with animal species names as queries led to top-ranked results that were often high-quality photos, but also with other potential visual representativeness concerns, as sometimes the results included photos of nature with the query animal shown in the distance, rather than as portraits. In addition, we had coverage concerns with Adobe Stock Images. Thus, we combined the top-5 results from Google Search and Adobe Stock Images to create a set of images to rank and choose the best candidate from.

Visual Representativeness Ranking

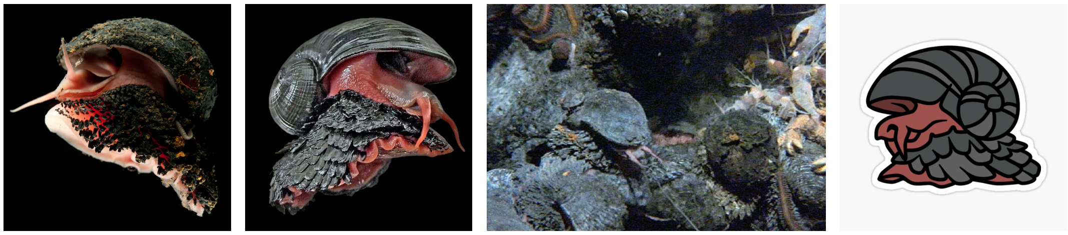


Figure 6.3: The top-2 and bottom-2 species images ranked by GPT-4V (higher scores to the left) for ‘Scaly-foot snail’.

Using up to 10 image candidates for each species, we ranked them using GPT-4V (`gpt-4-vision-preview`) by prompting the model with instructions for focusing on specific types of photos (*i.e.,* “focus on one individual species, preferably in the wild, over cartoons, drawings, or the species photographed in the distance”) with potential value for mechanism understanding (*i.e.,* “contains visual details that help

[4] `https://programmablesearchengine.google.com/about/`

[5] `https://stock.adobe.com`

viewers understand the following biological mechanisms...") In addition, we instructed the model to score all 10 candidate images from 0 to 100 in terms of goodness based on the criteria described above, and provide rationale for each. The prompt used can be found in Appendix F.2.

For example, fig. 6.3 shows the top-2 and bottom-2 scoring images (from left-to-right) for the species 'Scaly-foot snail' using this approach. The score and rationale of each image was as follows.

```
Top-1 | Score: 95 | Rationale: This image provides an excellent close-up view of the
↪  scaly-foot snail, highlighting intricate details of the iron-infused sclerites and
↪  the shell's texture which are indicative of its biological mechanisms.

Top-2 | Score: 92 | Rationale: This image offers a clear and detailed close-up of the
↪  scaly-foot snail, clearly showcasing the iron-infused sclerites and the unique
↪  appearance of its iron-mineralized shell, which are valuable for understanding its
↪  biological mechanisms.

Bottom-2 | Score: 30 | Rationale: Although this underwater image shows the scaly-foot
↪  snail in its natural environment, the details necessary to comprehend the
↪  biological mechanisms such as the sclerites and shell are obscured and not clearly
↪  visible.

Bottom-1 | Score: 10 | Rationale: The image is a stylized, cartoon-like representation
↪  of a scaly-foot snail, lacking detailed visual information about the species'
↪  biological mechanisms such as the iron-infused sclerites and iron-mineralized shell
↪  .
```

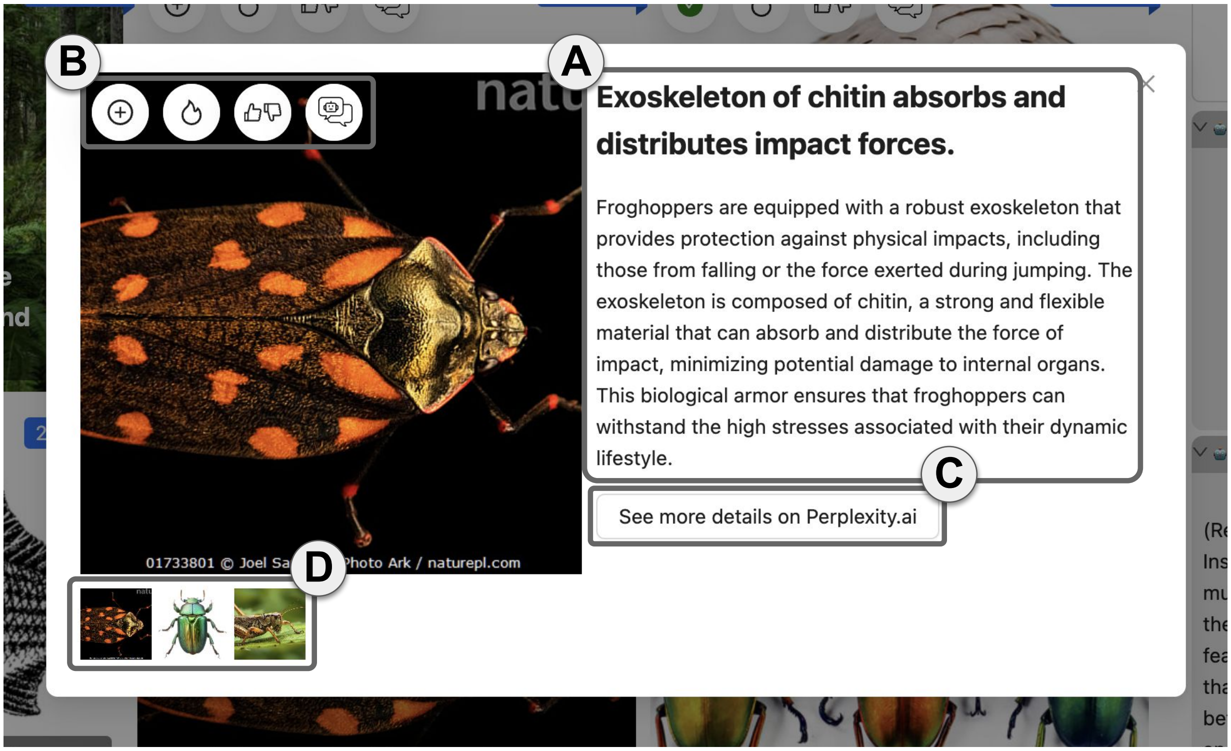
The first author reviewed ranked images, their scores, and rationale for 10 species, and found a consistent pattern of reasonable visual processing and instruction-following to be useful.

6.4.4 Interface

Walk-through

Consider an automotive designer, Sarah, looking for inspirations that could spark new ideas for novel bike rack design, similarly as the persona we described for the formative study. When she arrives at the BioSpark interface, she first scrolls through the board UI on the left of the screen to review different clusters of mechanisms. She is initially drawn to the 'exoskeleton' cluster, showing an image of a froghopper, as the exoskeleton structure may have insights into the skeletal design of new bike racks. She clicks the cluster card (fig. 6.1, Ⓐ) to examine its details further. The mechanism description in the modal that expands out upon her click highlights a particular material, 'chitin', as strong and flexible that can absorb and distribute the force of impact. She clicks on the 'See more details on Perplexity.ai' button to explore this material further (fig. 6.4). She finds a few related scientific research providing additional details of the exoskeleton composition, such as how pleural arches of the froghopper exoskeleton contain a composite structure of both rigid chitin and the elastic protein resilin that allows the exoskeleton to store energy and then release it quickly to power the froghopper's powerful jumps (fig. 6.6). She takes a quick note on the research and returns to BioSpark.

She then finds another mechanism that seems counter-intuitive yet interesting, the mucus and muscular foot of 'Architaenioglossa', that includes different species of snail, as potentially interesting mechanisms for the problem. She clicks on the 'spark' button to receive inspirations for new ideas that may use this mechanism in new ways (fig. 6.1, Ⓓ). She receives two sparks in response; the first, titled ‘Mucus-Glide Bike Mount’, describes an idea that uses hydrogel coating to reduce friction in motion. Intrigued by the idea, but concerned with the durability of hydrogel in various weather conditions, she asks BioSpark using the ‘Ask AI’ button (fig. 6.1, Ⓕ): “*what are good candidate hydrogel coating materials? Also consider weather situations (frigid cold or precipitation) and suggest materials robust to such conditions.*”. BioSpark returns an information card that provides alternative material choices, such as Polyacrylamide Hydrogels, described as capable of maintaining their mechanical strength and elasticity in a wide range of temperatures and as resistant to degradation in wet conditions, or Polyvinyle Alcohol (PVA) Hydrogels, notable for excellent mechanical properties and withstanding repeated freeze-thaw cycles while maintaining a low-friction surface even when wet, which makes them an appealing case for use in cold weather conditions (the Q&A card in the top of the stream, fig. 6.1, Ⓘ). She writes down these materials as potential leads to pass on to the engineering research team later, and clicks on the ‘Trade-off’ button (fig. 6.1, Ⓔ) to learn more about the potential disadvantages of a design that incorporates a lubricant-like material directly on the surface of the rack where bike wheels are loaded on to. The returned trade-offs card raises cleaning difficulty as a potential concern, which she uses to ideate related usage scenarios and constraints involved to develop the idea further.



Exoskeleton of chitin absorbs and distributes impact forces.

Froghoppers are equipped with a robust exoskeleton that provides protection against physical impacts, including those from falling or the force exerted during jumping. The exoskeleton is composed of chitin, a strong and flexible material that can absorb and distribute the force of impact, minimizing potential damage to internal organs. This biological armor ensures that froghoppers can withstand the high stresses associated with their dynamic lifestyle.

Figure 6.4: The modal view of a clicked mechanism cluster shows additional mechanism and active ingredient details (Ⓐ). The same action buttons featured on the main page of the interface (Ⓑ) are shown, as well as the ‘See more details on Perplexity.ai’ for finding additional details and related scientific researech (Ⓒ), and a carousel displaying other species that belong to the cluster which can be viewed by clicking on any of the images (Ⓓ).

Sparks

We expected that BioSpark users may engage in an explore-exploit trade-off while interacting with the system. Many prior studies have investigated patterns involved with this trade-off, for example in the

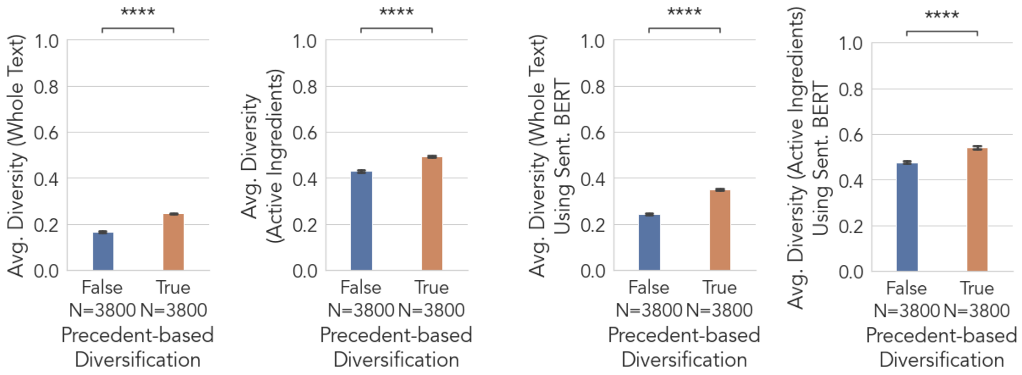


Figure 6.5: (First & Second) Bar graphs show that semantic diversity increased when using the precedent-based diversification approach, both at the whole spark and active ingredient levels; (Third & Fourth) Repeat analyses show the robustness of these results against the choice difference of the encoder model, when the Sentence-bert model [215] is used instead of OpenAI's text-embedding-3-large.

context of organizational learning (*cf.* [173]) and information foraging (*cf.* [207]), producing valuable implications for system design. In BioSpark, this trade-off may manifest in the form of users scrolling and browsing various clusters in the interface, and wanting to efficiently exploit an interesting design space that surrounds a particular mechanism and its active ingredient. In order to support efficient exploitation of an interesting design space that surrounds a mechanism inspiration, we design a one-click feature for new related spark generation.

We generate two new sparks using GPT4 (`gpt-4-turbo-preview`) each time the user clicks on the 'spark' button (fig. 6.1, Ⓓ) on a mechanism. We design a spark-generation prompt (Appendix F.3) to request the generation. In the prompt we contextualize the user-selected mechanism inspiration with the design problem description and the constraints provided with the problem. We instruct GPT4 to be succinct when generating sparks (*i.e.,* under 500 characters) and provide a descriptive title for each. However during pilot testing, we noticed that directly generating multiple sparks for the same mechanism inspiration led to highly similar generations, despite the explicit instruction included in the prompt that requested diversification in generation.

To address this, we add the most recently generated 20 sparks as part of the prompt, and deliberately request that the new generation be novel, and not redundant with them. We term this approach 'precedent-based diversification'. We test whether precedent-based diversification leads to a significant improvement in terms of semantic diversity compared to generation without such diversification. To this end, we repeatedly generate 20 sparks for each of the 10 randomly selected seed mechanism inspirations, and for each of the two design problems (the same design problems that our participants saw in the user study).

We investigate semantic diversity at two levels, the whole text and the active ingredient of a spark. To get the active ingredient, we process the generated spark using the same process as before for extracting active ingredients from mechanisms (§6.4.2). We then encode each spark or active ingredient text into an embedding using the OpenAI's `text-embedding-3-large` model. We construct pairs of spark or active ingredient embeddings using the 20 sparks generated for each seed mechanism for each of the two design problems, which amounted to 3,800 pairs, and calculate the average cosine distance among the pairs. This average represents the semantic diversity measure, which has been used in similar context in prior studies and was shown to be a viable measure of semantic diversity of natural language texts (*cf.* [92, 109, 246]).

In order to ensure robustness of our results against the choice difference of the encoder model, we repeat the analysis using another popular encoder – the `Sentence-bert` model for embedding the text [215]. We find that, at the whole spark text level, the semantic diversity was significantly higher when precedent-based diversification was used (M=.24, SD=.073) than not (M=.17, SD=.090) ($t_{\text{two-tailed}}$(7291.87)=-42.41, p=0.0). The result is consistent for the active ingredient level (M=.49, .43, p = 0.0), and robust against the choice of encoder models (fig. 6.5).

Finally, in order to further facilitate users' engagement and exploration in a potentially interesting design space, we add two sparks whenever the user saves a mechanism inspiration as well.

Trade-off Analysis

We generate a new trade-off analysis card using GPT4 (`gpt-4-turbo-preview`) each time the user clicks on the trade-off button (fig. 6.1, Ⓔ) on a mechanism. We design a trade-off analysis prompt (detailed in Appendix F.4) to request the generation. In the prompt we contextualize the user-selected mechanism inspiration using the design problem description and the constraints provided with the problem. We instruct GPT4 to return the 'pros' and 'cons' of the mechanism inspiration in the context of the design problem using a markdown table format that places each pro-and-con pair in a new row, and give each item in the table a succinct (3 words or less) label. In the view, we display the analysis in each trade-off card in the stream (fig. 6.1, Ⓘ) and implement a scrollable and formatted table view using `React-Markdown`[6] and `remark-gfm`[7].

Two-Stage User Request Triaging and Handling Q&A

In order to flexibly respond to various requests that users could type in the Q&A text area (fig. 6.1, Ⓕ) and generate appropriate responses, we design a two-stage process for handling user Q&A. In the first `action-triage` stage we prompt GPT4 (`gpt-4-turbo-preview`) with user-typed text to act as an agent that reads the content and triage it to any of the following five action choices:

```
[Action 1]. Generate **two** related but highly different ideas based on the user-
↪  selected mechanism.
[Action 2]. Perform an analysis of anticipated pros-and-cons design tradeoffs of
↪  applying the user-selected mechanism.
[Action 3]. Answer the user's follow-up question or respond to their comment related to
↪   the user-selected mechanism
[Action 4]. None of the above actions are appropriate for the user comment; take no
↪  action.
```

We instruct GPT4 to pick appropriate actions and for each choice, return an 'appropriateness score' and supporting 'rationale' for the choice. We pass the latter information to the interface to feature it on a tooltip next to the timestamp of the returned card (fig. 6.1, the ?⃝ icon left to each card in the stream Ⓘ). The first action indicates that the user intent inferred from the text by GPT4 is in seeking new ideas. The second action indicates that the user intent is in seeking a design trade-off analysis. The third action indicates that the user intent is in additional details about a mechanism. The final action indicates that none of the user intent above was deemed appropriate. We sort by the appropriateness score in the descending order and send corresponding requests for performing each action, as described in the previous sections.

[6] `https://github.com/remarkjs/react-markdown`
[7] `https://github.com/remarkjs/remark-gfm`

In the second stage, the BioSpark backend performs the action and returns the result to the interface.

Stream Organization & Supporting Efficient Exploitation of an Interesting Design Space

In order to support user engagement with freshly produced sparks and other system-generated information, we organize the stream (fig. 6.1, Ⓘ) by recency, placing the most recently generated items to the top. In addition, to support efficient exploitation of an interesting design space, we support action buttons directly in each spark (fig. 6.1, Ⓚ). Users can use these buttons to build off of any of the existing sparks in the stream, for example by clicking on the spark generation or the Q&A button in the card. We leverage the same machinery for generating sparks as before, but contextualize the generation using the selected spark in the stream, to anchor generation in the design space being exploited by the user.

Furthermore, the stream contains helpful organizational feature, such as quick filtering of different types of information (*e.g.,* sparks, trade-offs, or Q&A only, fig. 6.1, top of the stream Ⓘ), as well as deleted items with additional support for restoration.

Drill-down on related research

In order to support users with drilling down on related scientific research for each mechanism inspiration on demand, we designed a designated button (The 'See more details on Perplexity.ai[8]' button, fig. 6.4, Ⓒ). Through interface pilots, we anticipated that the most common user workflow for drilling down on related research to be taking place after the user decides on a particularly interesting cluster for further consideration. When designing the button, we initially considered its placement on each of the cluster cards in the main interface, but decided to move it to the cluster modal view in order to prevent clutter and support effective exploration of diverse design space on the main interface. In addition, to support the streamlined exploration – decision – further research workflow, we specifically placed the button at the end of the extended mechanism description featured in the modal fig. 6.4, Ⓒ).

We implemented the button's functionality as opening a new browser tab that contains search results of relevant research on the Perplexity.ai website. The search query was pre-populated using the following template:

```
Give me relevant details about "[active ingredient]" commonly found in [species]
```

An example of the Perplexity.ai page result is shown in fig. 6.6. This functionality design was a compromise following our technical investigation that showed the difficulty of implementing Perplexity.ai's search page results inside a native `React.js` application interface[9] and the lack of API[10] support for evidence generation[11].

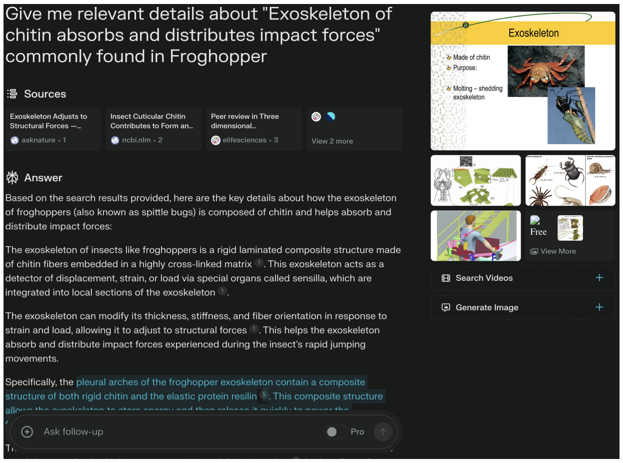
BioSpark was implemented using `React.js` for the interface and the `Flask` server in `Python3.11` for the backend components.

[8] `https://www.perplexity.ai/`

[9] Perplexity.ai prohibits user requests that attempt to render its search results natively.

[10] `https://docs.perplexity.ai/`

[11] Last tested on March 17th, 2024.



Give me relevant details about "Exoskeleton of chitin absorbs and distributes impact forces" commonly found in Froghopper

Sources

Exoskeleton Adjusts to Structural Forces —... asknature • 1

Insect Cuticular Chitin Contributes to Form an... ncbi.nlm • 2

Peer review in Three dimensional... elifesciences • 3

View 2 more

Answer

Based on the search results provided, here are the key details about how the exoskeleton of froghoppers (also known as spittle bugs) is composed of chitin and helps absorb and distribute impact forces:

The exoskeleton of insects like froghoppers is a rigid laminated composite structure made of chitin fibers embedded in a highly cross-linked matrix [1]. This exoskeleton acts as a detector of displacement, strain, or load via special organs called sensilla, which are integrated into local sections of the exoskeleton [1].

The exoskeleton can modify its thickness, stiffness, and fiber orientation in response to strain and load, allowing it to adjust to structural forces [1]. This helps the exoskeleton absorb and distribute impact forces experienced during the insect's rapid jumping movements.

Specifically, the pleural arches of the froghopper exoskeleton contain a composite structure of both rigid chitin and the elastic protein resilin [5]. This composite structure allows the exoskeleton to store energy and then release it quickly to power the

Figure 6.6: An example results page on Perplexity.ai that opens up in a new browser tab when the user clicks on the ‘See more details on Perplexity.ai’ button on the mechanism modal view. The page describes how the froghopper exoskeleton contains a composite structure of both rigid chitin and the elastic protein resilin that allows the exoskeleton to store energy and then release it quickly to power the froghopper’s powerful jumps, and its supporting research, that may provide valuable details as described in our scenario (§6.4.4).

6.5 User Study

We conducted a within-subjects laboratory study to investigate whether BioSpark was more effective in helping people engage with inspirations for ideating new solutions to design problems compared to a baseline condition that involved using both the expert-curated AskNature.org and ChatGPT. The rationale for using AskNature+ChatGPT as a baseline condition is that AskNature is a ”gold standard” for highly curated bioinspired design inspirations, and in combination with ChatGPT participants could in theory perform all of the same features as enabled by BioSpark, though with higher friction. Thus this baseline provides a strong test for the system: while we might not necessarily expect the quality of its inspirations to match that of AskNature, we anticipated them to be sufficiently useful to test the value of the system’s workflow support as a whole.

6.5.1 Methodology

Research Questions

Our research questions included:

[RQ1] How do users engage with inspirations and how does the depth and type of engagement differ between conditions?

[RQ2] How do users explore the design space during ideation and how do the quantity and diversity of their ideas differ between conditions?

Structure

We employed a within-subjects study design to compare BioSpark with a baseline system for inspiration and a shared Google Spreadsheet participants accessed to write down their own ideas. We chose two design problems for user ideation, including how to design wheelchairs that allow users to go up the stairs easily and how to design an innovative bike rack for sedans. These problems were chosen because they involve multiple, potentially competing constraints (e.g., lightweight but durable) and were pilot tested for being able to be completed within the timed ideation task.

```
(The 'Wheelchair' problem) Design advanced wheelchairs that can also allow users to go
↪ up the stairs easily.

Constraint 1 (Lightweight yet Durable Construction): The wheelchair should be
↪ lightweight and be able to withstand a heavy load without structural failure.

Constraint 2 (Compact and Foldable Design): The wheelchair must be foldable to a 1/4 of
↪  the volume within 30 seconds without the use of tools.
```

```
(The 'Bike rack' problem) Design innovative bike racks for sedans.

Constraint 1 (Integration without Interfering with Aerodynamics): The bike rack's
↪ design must not significantly reduce the vehicle's fuel efficiency when installed
↪ and with bikes mounted.

Constraint 2 (Versatility in Accommodating Different Bike Types): The rack must be able
↪  to securely hold at least three different bike frame sizes (e.g., 16", 20", and
↪ 26") without the need for additional adapters.
```

We randomly assigned problems to conditions for the main timed tasks (20 minutes each), counterbalancing the order of presentation using 3 2x2 Latin Square blocks. Participants followed a fixed procedure in the study, which took place remotely using Zoom: Introduction, Consent, Demographics survey; Tutorial (detailed in Appendix G.1) of the first system via screensharing; Main task for the first system (20 min); Rating task for the first system (only in the BioSpark condition) Survey for the first system; alternating and repeating for the second system; followed by a debrief. Participants were asked to share their screen during the timed tasks and think-aloud. To probe how participants felt about the utility of different information generated using various AI-based system features, after the BioSpark's main task, participants were also presented with a rating interface that showed a list of saved Spark, Q&A, and Trade-off cards along with a 5-point Likert-scale for them to rate its usefulness in their process.

Participants

We recruited 12 researchers (7 women, 5 men) through advertisement on Upwork[12] and email lists at a State Arts College. Participants' background included professional UX design experience (6), professional illustration and graphic design (1), PhD in Psychology (3), and a current undergraduate student in Arts and Design (1) and a master's student in AI and Data Science (1). Participants' average age was 36.1 (SD=9.91).

Baseline

The baseline system used AskNature.org+ChatGPT. Participants were given 5 URLs, each of which pointing to a functional category equivalent to those that were used for the BioSpark backend dataset

[12] https://www.upwork.com/

pipeline: Manage Impact[13], Manage Tension[14], Manage Compression[15], Manage Turbulence[16], and Modify Speed[17]. Before the baseline task began, participants organized their screen by opening up all 5 tabs in their browser on the left-hand side of the screen and sign-in and open the ChatGPT[18] interface on the right-hand side of the screen. They were instructed to freely use the platforms to help themselves understand and ideate with mechanism inspirations found on AskNature for the design problems. Each participant was also instructed to write down the ideas they come up with in the process in a prepared Google spreadsheet, with a brief description of the species that inspired each idea.

Qualitative coding of the types of participants' engagement with mechanism inspiration

The research team met to discuss coding of interview and think-aloud data from the study. One salient feature of the data was that participants seemed to engage with mechanism inspirations differently in depth, for example with or without follow-up actions that related to attempting to deepen their understanding of the inspirations, of their relevance to the design problem and of trade-offs regarding different design constraints, and attempting to come up with new ideas that could adapt the inspirations to a design problem in new ways. In order to capture this, the first two authors came up with the following four codes that describe participants' different engagement patterns:

[**S1**: "*Interesting!*"]: Comments on a mechanism inspiration, but directly followed by moving on to a different mechanism that was visible to the participant.

[**S2**: "*I'm not sure how this might be relevant*"]: Comments on a mechanism inspiration, but similarly followed by moving on to a different mechanism that was visible to the participant.

[**D1**: *Engaging with relevance understanding and constraints consideration*]: Engages with AI to understand a mechanism inspiration's relevance to the design problem, for example by asking the following types of questions "*tell me examples of...*" or "*how might this be used/applied...*"

[**D2**: *Actively coming up with new ideas*]: Explores the design space and actively generates new ideas "*it made me think of...*"

One author transcribed the interview and think-aloud data from the study, and incorporated descriptions of participants' actions with each platform (*e.g.,* what the participant is typing in the ChatGPT interface or what the participant is clicking in BioSpark), that participants' think-aloud did not describe but were relevant to understanding their engagement process and intent. This amounted to 266 transcripts across 12 participants. Coders coded a set of randomly selected 16 transcripts together blind-to-condition and arrived at an agreed-upon set of codes through a discussion. Then the two coders coded 30 additional randomly selected transcripts independently. The inter-rater agreement of codes for this set showed a moderate to strong level of agreement $\kappa = 0.76$. Thus, the first author coded the remaining 218 transcripts alone.

Extraction of unique design constraints described in each idea

To analyze the user engagement patterns involving consideration of design constraints (§6.5.2), we first extract the unique design constraints described by participants in each idea. We use GPT4 (`gpt-4-tu`

[13] (Manage Impact) `https://rb.gy/rvz17u`

[14] (Manage Tension) `https://rb.gy/t3se2z`

[15] (Manage Compression) `https://rb.gy/xvogjb`

[16] (Manage Turbulence) `https://rb.gy/9apgoq`

[17] (Modify Speed) `https://rb.gy/r7o2c8`

[18] `https://chat.openai.com/`

`rbo-preview`) with a prompt (Appendix H.1) to perform the extraction. The first author reviewed the extracted constraints in terms of their coherence and uniqueness for a random set of `20` ideas and found that the extraction was satisfactory in terms of both the uniqueness of extracted constraints and their coherence.

Extraction of the species' names that participants described as inspiring their ideas

To analyze the diversity of the species that participants were inspired from for their own ideas (§6.5.2), we use GPT4 (`gpt-4-turbo-preview`) with a prompt (Appendix H.2) to extract the species name from each participant idea and normalize it. The first author then reviewed the extracted species' names in a random sample of `20` ideas and found that the extraction accuracy was satisfactory.

Length-constrained Summarization and Diversity Calculation of Participants' Ideas

In order to accurately analyze the semantic diversity of participants' ideas without the length of idea description as a confound (§6.5.2), we first summarize each participant idea into 10 words or less using GPT4 (`gpt-4-turbo-preview`) with a prompt (Appendix H.3). In the prompt we instructed GPT4 to succinctly summarize each idea in 10 words or less, and provided four examples of summarization. The first author then reviewed summarized ideas in terms of accuracy for a random set of `20` ideas and found that the summarization performance was satisfactory.

Each summarized idea was then encoded into an embedding using the OpenAI's `text-embedding-3-large` model. Using these embeddings, we construct pairs in each individual participant-condition combination and calculate the average per condition for semantic diversity analysis between the conditions, similarly as before in our evaluation of precedent-based diversification (§6.4.4).

Measures

In relation to the research questions described in §6.5.1, we collect the following measures for analysis:

[RQ1] To analyze the differences in participants' engagement patterns we measure: the frequency of each code for each participant in each condition; the character length of each idea; the number of unique design constraints mentioned in each idea.

[RQ2] To analyze the diversity in design space exploration we measure per condition: the aggregate pairwise cosine distance for each participant; the number of unique species each of the participants were inspired by; the number of ideas participants generated.

Furthermore, to analyze how participants' thought about the two systems' usefulness and their various AI features, we collected participants' subjective ratings to a modified Technology Acceptance Model survey questionnaire items focused around task performance and easiness of learning from [273] (4 items). In addition, we employed questions focused on serendipity and exploration adapted from [175, 191] (9 items) and the questions on the value of AI assistance and the quality of inspirations found in the system.

For the BioSpark condition, additional 6 questions were included in the survey asking participants about the usefulness of 6 different features of the system.

Analysis

[RQ1] To analyze the potential differences in engagement patterns with mechanism inspirations, we perform a χ^2 test followed by pairwise paired-samples t-tests (two-tailed) with corrections for multiple tests (Bonferroni) when appropriate. Furthermore, we perform independent-samples t-tests (two-tailed) for idea-level analyses, *e.g.,* for their length and the number of design constraints described.

[RQ2] To analyze the diversity in design space exploration, we perform paired-samples t-tests (two-tailed), *e.g.,* for each participant's aggregate pairwise cosine distance between conditions, and the number of ideas participants generated.

For analyses of how participants' thought about the two systems' usefulness and their various AI features, we perform Wilcoxon's signed rank test on the survey data, which was participants' responses on a 7-point Likert scale (1: *strongly disagree*, 7: *strongly agree*), using the non-parametric paired-samples and two-tailed Wilcoxon's signed rank test. We also perform a thematic analysis [31] on the transcripts. Our analysis focused on themes around: how users interacted with the integrated interaction features in BioSpark, and how that contrasted with their usage patterns in the baseline condition, and the potential challenges related to that.

6.5.2 User Study Findings

We structure our user study findings around the two research questions and report on how participants felt that the BioSpark's integrated AI support compared to ChatGPT for helpfulness in engaging with mechanism inspirations and exploring the diverse design space.

RQ1: How did participants engage with inspirations and how does the engagement differ in depth between the conditions?

We consider three factors with significance to how designers engage with analogical inspirations to arrive at insights. The first factor is the depth in which they engage with an inspiration, which is particularly relevant to analogical inspirations that may require challenging, deep cognitive processing of analogs [89, 94] for valuable ideas of transfer to emerge. The second factor is the 'first impression' of an analogical inspiration, where the negative first impression could be particularly damaging by snipping the bud of potentially valuable insight despite surface-level irrelevance. The last factor is how much designers consider the feasibility constraints of an inspiration or idea that may have significant implications for its practical impact.

To analyze these factors, we first look at a subset of the transcribed participants' think aloud along with their behavior descriptions and find distinctive engagement patterns. Table 6.1 shows representative cases and how they map to the four codes of engagement patterns (§6.5.1). Across the cases that involved participants' deep engagement (D1 & D2), we see a consistent pattern of participants following up on their thoughts with exploration of additional information (*e.g.,* such as from using the Ask feature in BioSpark; D1), or going into the details of what the inspiration made them think of (D2). In contrast, for shallow types of engagement (S1 & S2), we find that participants stop short of the kinds of extensive follow-up we see with deep engagement.

Following this observation, we analyze the data and hold the following hypotheses:

[H1] Participants engage more deeply with mechanism inspirations when they can access integrated 'deep' engagement features (*e.g.,* requesting 'sparks' of mechanism-to-problem mappings for a user-selected mechanism that are different from previously generated sparks or user ideas; requesting a run-down of anticipated design trade-offs for a user-selected mechanism; and being able to ask a follow-up question about a mechanism and receive a problem-context-specific answer).

[H1a] (*Depth*) Participants in BioSpark show a higher frequency of 'deep' engagement than the baseline.

[H1b] (*First Impression*) Participants in BioSpark see interesting connections or be curious about a mechanism inspiration significantly more often than the baseline.

[H1c] (*Feasibility Consideration*) Participants in BioSpark elaborate on design constraints in their ideas more extensively than the baseline.

Qualitative log analysis of engagement types. In order to examine how participants' engagement depth differed between the conditions, we start our analysis by categorizing the first two engagement codes as 'shallow' (*i.e.,* S1 – participants commenting "*Interesting!*" or S2 – "*I'm not sure how this might be relevant*" without any follow-up actions), and the remaining two codes as 'deep' (*i.e.,* D1 – participants' comments while interacting with ChatGPT or BioSpark's deep engagement features related to understanding the relevance of a mechanism inspiration, *e.g.,* "*tell me examples of...*" or "*how might this be used/applied...*", and D2 – exploring the design space and actively generating new ideas *e.g.,* "*it made me think of...*"). We first perform a χ^2 test to find if there is a significant distributional difference between the frequencies of the two types of codes and conditions. We find a significant distributional difference ($\chi^2(1)$=12.93, p=.0003). Following this result, we perform paired two-tailed t-tests to identify the between-condition difference for each engagement type. Pairwise comparisons after Bonferroni correction for multiple (2) testing for each type of engagement shows a significantly higher frequency of deep engagement in BioSpark (M=3.3, SD=2.67) over the baseline condition (M=1.3, SD=1.87) ($t_{\text{paired}}(21.16)$=-3.12, p==.01, fig. 6.7, left). Interestingly, we find no significant differences in the frequency of shallow engagement between conditions (Baseline: M=6.3, SD=3.28; BioSpark: M=5.8, SD=3.60, $t_{\text{paired}}(21.81)$=.34, p==.74, fig. 6.7, right).

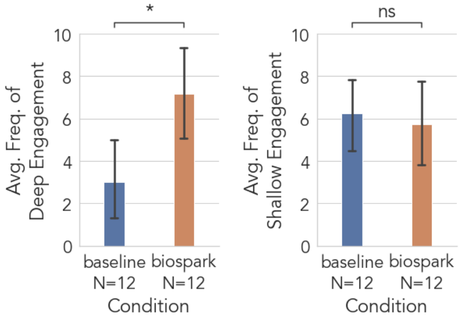


Figure 6.7: (Left) The bar graph shows that the average number of deep engagement was significantly higher in BioSpark; (Right) The bar graph shows that there were equally many shallow engagement types in both conditions.

These results suggest that participants were more efficient when engaging with diverse mechanism inspirations in BioSpark, leading to more frequent 'deep' engagement in spite of similar levels of frequency in otherwise 'shallow' engagement. Indeed, we see that participants' follow-up actions for deeply engaging with source inspirations differed between conditions which involved different time and cognitive demands required for performing the actions. For example, when asking follow-up questions about a mechanism in the baseline condition using ChatGPT, participants had to manually provide relevant mechanism context from the AskNature webpage and iterate on their prompts. In contrast, BioSpark provided dedicated one-click buttons for requesting a design trade-off analysis, Q&A interaction, and generating ideation "sparks" based on a user-selected mechanism that provided descriptions of relevant mappings to the problem. Participants' behavioral log data show that they used these system features throughout the experiment (fig. 6.8). Participants also commented in support of the difference in efficiency of follow-up actions, as P1 described: *"I just felt like it was able to produce things without me having to like prompt*

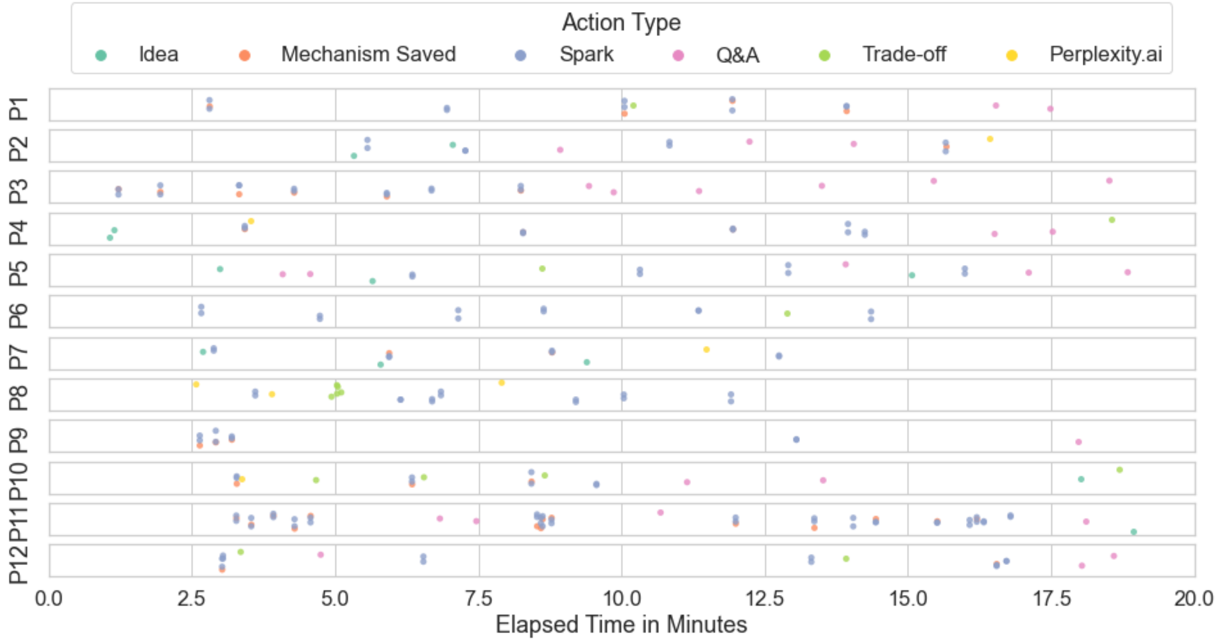


Figure 6.8: This user log visualization shows that participants used BioSpark features throughout the span of experiment to generate new sparks of inspiration based on user-selected or saved mechanisms (Spark), learn more about design trade-offs (Trade-off), write down their own idea (Idea), ask follow-up questions about the mechanism and design constraints (Q&A), or drill down on related research (Perplexity.ai).

it. And I think that allowed me to spend more time, maybe thinking about specific connections between the mechanisms and the design features it was suggesting. Whereas... with [ChatGPT] I felt like I had to spend more time like doing to allow it to actually help, but with [BioSpark] I felt like the AI system already knew what I needed, so it saved that step.". Together these results suggest that the availability and design of system features in BioSpark may have freed up participants' cognitive bandwidth, allowing them to focus more on deeper engagement with mechanism inspirations. **Thus we confirm H1a.**

Deep engagement can only take place if participants recognized the potential relevance or be curious about a mechanism inspiration without initially realizing the relevance. In either case, the first impression of a potential inspiration has a significant implication for how that inspiration may be taken up or pursued for deeper engagement down the road, and especially important for analogical inspirations whose value may be nontrivial at the beginning, but may have an outsized value after iterations [24, 46, 153]. To examine how BioSpark changed the saliency of participants' first impressions of mechanism inspirations compared to the baseline system, we analyze the frequency of the two positive and negative forms of shallow engagement. Specifically, we analyze the balance of the frequency of the positive-to-negative first impressions, by mapping the codes $S1 \rightarrow +1$ and $S2 \rightarrow -1$. We find that the balance is significantly higher in the BioSpark condition (M=3.1, SD=3.92) than the baseline condition (M-.3, SD=4.16) (t_{paired}(21.92)=-2.25, p==.046). **Thus we confirm H1b.**

How participants engaged with design constraints related to the ideation task. In addition to the analysis of user *behaviors* of engagement, we also examine the *output* of user behaviors for signatures of deep engagement. One such signature is the number of different design constraints participants described in each idea. Anticipating and engaging with different design constraints is important as design

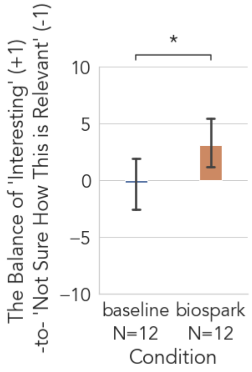


Figure 6.9: The graph shows that the frequency balance of recognizing or not recognizing connections between a mechanism and the design problem at hand (*i.e.,* ‘interesting’ – S1 and ‘not sure how this is relevant’ – S2, mapped to `+1` and `-1`, resp.) is roughly at parity in the baseline condition, while there were significantly more positive first impressions in BioSpark.

constraints are important for the feasibility and practical impact of ideas. In order to analyze the level of design constraint elaboration we group user ideas per condition and first examine the average idea length. We perform a two-tailed, independent samples t-test over ideas grouped by condition and find that participants’ ideas were significantly longer in the BioSpark condition (M=375.5, SD=96.15) over the baseline condition (M=141.9, SD=108.93, $t_{\text{two-tailed}}$(119.25)=-14.65, p=¡.0001). However, the length alone may not necessarily represent how deeply participants elaborate on design constraints without being corroborated by its content focused on related design constraints.

To examine the constraint-focused-content, we first extract unique chunks from each idea description that each corresponds to consideration of a single coherent design constraint. We use the `gpt-4-turbo-preview` model with a prompt for extraction, which showed a satisfactory performance (details in §6.5.1). We then counted the number of extracted constraints for each idea. Using this data, we find that the length of each idea was indeed significantly correlated with the number of design constraints described in it (ρ=.58, p ¡ .0001). We also find that BioSpark users mention a significantly higher number of design constraints in their ideas (M=2.7, SD=1.01) than the baseline condition (M=1.6, SD=.78; $t_{\text{two-tailed}}$(163.27)=-8.45, p= ¡ .0001). Participants’ comments also supported these results. For example, P4 described how the default behavior of BioSpark providing information about relevant design constraints to consider nudged him: *“In the [baseline system], I was asking just random questions to ChatGPT, whereas in [*BioSpark*] I was asking related questions, you know, like, first would ask, how to manufacture this, but then I can follow up with like... I didn’t have to even ask ‘lightweight’ or ‘durable’ material. But when the system suggested them and provided alternative choices like carbon fiber, griffin, polymer... I thought this was very helpful. And then I looked at the pros and cons of using different kinds of materials, and manufacturing cost... which gave me ideas.”* **Thus we confirm H1c.**

Taken together, we confirm H1: Users engage more deeply with mechanism inspirations when they can access integrated ‘deep’ engagement features.

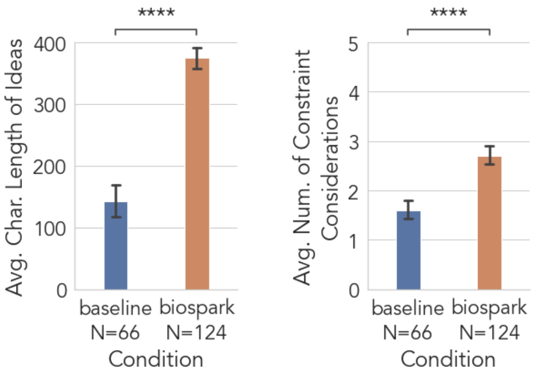


Figure 6.10: (Left) The bar graph shows that user ideas were significantly longer in BioSpark than in the baseline, and (Right) a similar trend was observed for the number of design constraints described in each idea.

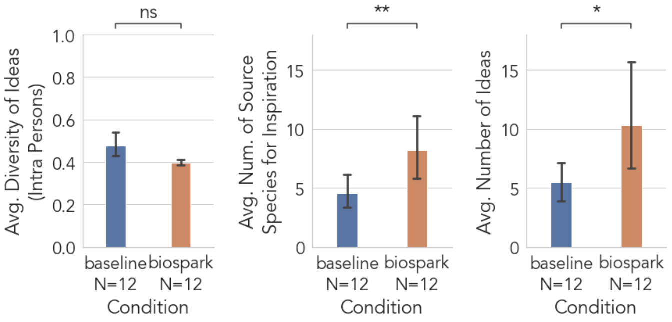


Figure 6.11: (Left) The bar graph shows a marginally higher diversity of ideas in the baseline condition. Diversity is measured as the average cosine distance of idea embeddings with length-constrained summarization, for each participant-condition combination; (Middle) The number of unique species in nature that participants were inspired by in their ideas was significant higher in BioSpark than the baseline; (Right) The number of ideas participants came up with during the experiment was significantly higher in BioSpark than the baseline.

RQ2: How much design space did users explore during ideation and how does its diversity differ between the conditions?

To answer the research question we look at the aggregate semantic diversity of ideas, the number of individual species represented in the ideas, and the number of ideas per participant, and hold the following hypotheses:

[H2] Participants explore a more diverse design space using BioSpark than the baseline system.

[H2a] Participants' idea diversity is significantly higher in BioSpark than the baseline.

[H2b] Participants engage with significantly more inspiring species in BioSpark than the baseline.

[H2c] Participants generate significantly more ideas in BioSpark than the baseline.

To analyze the semantic diversity of ideas, we calculate the aggregate pairwise cosine distance using text embeddings as described in §6.5.1, adopting an approach similarly used in prior studies and was shown to be a viable measure of semantic diversity of natural language texts (*cf.* [92, 109, 246]). However, directly applying the pairwise cosine distance measure without accounting for the significant difference in the character length and the number of design constraints mentioned in the ideas may lead to an inaccurate result because conceptually different ideas could result in a lower level of diversity in the BioSpark condition just as a side effect of mentioning similar constraints. To mitigate the potential confounding from length (and descriptions of design constraints rather than actual design ideas), we first perform a length-constrained summarization of each idea in 10 words or less using `gpt-4-turbo-preview`, and embed the summarized idea text to calculate the average pairwise cosine distance in each individual participant-condition combination for analysis (details in §6.5.1).

We find that diversity of ideas was .48 (SD=.109) in the baseline condition and .40 (SD=.025) in the BioSpark condition, but the difference is marginally significant (t_{paired}(12.17)=2.20, p==.05) (fig 6.11, left). **Thus we cannot confirm H2a: Participants' idea diversity is significantly higher in BioSpark than the baseline.**

Another measure of diversity in design space exploration is how many different species in nature participants are engaging with for ideation, which has potential for not only inspiring ideas based on a specific mechanism of the particular species but also opening up a new space of design that encompasses other mechanisms of the species or its related species. To this end, we analyze the number of unique species that participants describe as inspirations for their ideas. We first extract the inspiring species' names using `gpt-4-turbo-preview` with a prompt (details in §6.5.1) and lowercase the extracted names to construct a set of unique species per participant. We calculate the size of each set and average them across the participants in each condition. We perform a paired two-tailed t-test to identify the between-condition difference. We find that the number of unique species for inspiration is significantly higher in the BioSpark condition (M=8.2, SD=4.97) than the baseline condition (M=4.6, SD=2.71) (t_{paired}(17.02)=-3.30, p==.007) (fig. 6.11, middle). **Thus we confirm H2b: Participants engage with significantly more inspiring species in BioSpark than the baseline.**

Complementary to the types of content diversity described above, the number of unique ideas that participants came up with itself also has a direct implication to how much design space participants explored. To analyze this, we perform a paired two-tailed t-test of the number of participant-generated ideas. We find that participants in the BioSpark condition generated significantly more ideas (M=10.3, SD=8.46) than the baseline condition (M=5.5, SD=2.91) (t_{paired}(13.56)=-2.35, p==.04) (fig 6.11, right). **Thus we confirm H2c: participants generate significantly more ideas in BioSpark than the baseline.**

Taken together, we find that while the semantic diversity score of ideas shows a ∼17% (0.48 → 0.40) decrease from the baseline to the BioSpark condition, the average number of ideas (5.5 → 10.3) and the number of different species participants engaged with (4.6 → 8.2) both show significantly higher levels of increase (~105% and ~78%, respectively). Thus, **to the extent of considering the aggregate effect of the factors and their relative proportions, we conclude that exploration of design space is effectively broadened in BioSpark over the baseline system (H2).**

Furthermore, the interview, survey, and observation data allow us to more deeply understand what aspects of the BioSpark design and interactive features most contributed to broadening participants' exploration of design spaces.

Participants felt that the AI features in BioSpark helped them explore more design spaces and get creative. On the usefulness of the 'spark' generation feature (the second highest-scoring feature for usefulness, M=5.9, SD=1.08), P1 commented that:

> *"The "Sparks" button provided me with a lot of interesting insights and got me thinking in directions I may not have thought of on my own. It spurred my creative thought. There were some instances where I felt that the Sparks button produced insights that were a bit technical and somewhat excessive, so I*

streamlined the text to enhance accessibility for myself."

On the usefulness of the 'Ask AI Anything' feature (the third highest-scoring feature for usefulness, M=5.8, SD=1.19), participants mentioned that it *"helped clarify the details of the general overview of the idea"* (P3), and that *"I could ask some very specific questions about very specific mechanisms such as finding a fabricated material that is comparable to chitin, and get a useful reply"* (P4).

However, participants also felt in some occasions the AI's response was re-coursing to make connections to the original problem constraint, which seemed forced and adding less value *"it's still trying to navigate the conversation towards the original topic and doens't seem to be "progressive" enough to talk more in depth with specific areas"* (P12). Overall, participants agreement with the statement '*I was able to examine a variety of inspirations*' was significantly higher in the BioSpark condition (M=6.6, SD=.67) than the baseline condition (M=5.4, SD=1.62) (Wilcoxon W=0.0, p=.03), as well as for the statement '*I could easily explore many inspirations without getting lost*' (M=6.3, SD=.89 in BioSpark, M=5.2, SD=1.80 in baseline, Wilcoxon W=2.0, p=.05).

Participants' strategies for prompting and integrating ChatGPT with AskNature.org varied, without clear guidelines

Furthermore, BioSpark's tailored AI support helped participants find a relevant design space and ideate within it, and this was made clear when contrasting with how participants interacted with AI support in the baseline condition. Consider the three participants P12 and P9, and P4 who navigated AskNature and used AI in different ways in the baseline condition. P12 was an undergraduate student at an arts college who have only casually used prompt-based AI tools such as ChatGPT while P9 worked in a tech company as a UX designer and have extensively used prompting on a daily basis in his work. P9 even volunteered a few of his system prompts saved on ChatGPT that detailed the persona, task instructions, and performance guidelines which he described perfecting over time. P4 was a PhD student in educational psychology who have occasionally used ChatGPT in the past.

While P12 initially struggled to make relevance connections from the AskNature article pages to the problem context, subsequent interaction on ChatGPT was also time-consuming and ofttimes required multiple back-and-forth's to communicate his intent to AI, for example by asking AI to define an unfamiliar concept (*e.g., "what is [large pelagic cruisers]?"* – P12) and rephrasing and re-asking his question to focus its response to explanation of relevance to a problem domain (*e.g., "Well it's giving me specific animals (rather than general and transferrable concepts)... So I'm going to pause the generation... let me see... (types on ChatGPT) 'what shape?'... oops that didn't work, let me try again 'what shape that could be useful for industrial design?"'* – P12).

In contrast, and unlike other participants, P9 spent the first 6-70% of time in the study configuring the first prompt to ChatGPT, in which he provided the task details, the AskNature.org website details, and even including screen shots of the first few mechanisms he saw on AskNature. At the end of this context he added a description of his request to ChatGPT to then filter the mechanisms provided in 5 functional categories[19] in terms of relevance to the design problem at hand. In subsequent prompts he selected a few of the mechanisms ChatGPT generated in response, and prompted AI to generate more ideas based on those. While P9 felt this approach was effective in generating ideas, it certainly cost him a significant amount of time for crafting a prompt that had enough context and he deliberately chose to completely bypass the diverse mechanisms existing on AskNature and to rely fully on ChatGPT-filtered mechanisms.

[19] That were provided as part of the task instruction to participants in the baseline condition

Finally, P4 felt that AskNature was *"like a generic platform"*, but that he felt like meandering while interacting with AI to build off of the material, as *"I felt like I was asking just random questions to ChatGPT, while in the second one I was asking related questions"*. While he thought the functional category-based organization in AskNature helped his navigation, it also somewhat fixed the broader design space he explored in, as he felt like *"got stuck somewhere a little bit because I came up with this sideways top-mount bike rack idea early on from the 'manage turbulence' concept"*. P4 also thought it was easier to see the relevance of mechanisms in BioSpark and follow-up with more exploration.

In contrast, BioSpark's integrated design helped participants map out the design space

Participants commented on how the design and presentation of information in BioSpark streamlined their exploration and helped them accomplish the task. P10 described it as: *"I like that it's integrated into one space. I can press a button to get to the particular need that I had."*. While some of the participants thought clustering of different mechanisms was helpful for navigation, others either did not notice the 'clusters' or engage with them.

Furthermore, the stream organization of sparks was generally thought as helpful, and was the highest-scoring system design feature in BioSpark (M=6.3, SD=.98). Relatedly, P1 said:

> *"I liked these because they kept my thoughts and all the information very organized. It allowed me to focus on the actual text vs focusing on the organization of everything. It would have been even more helpful though if there was a way to enable bullet points or formatting tools within these. I would have used bold or italics for example."*

Some participants wanted additional support for *"comparing and contrasting"* (P7) the ideas, to 'easily highlight the strengths of each design and extract useful design features from it for integration into a new idea for mitigating anticipated challenges' (P5). Overall, participants' agreement with the statement '*Using this system would improve my task performance*' was significantly higher in the BioSpark condition (M=6.5, SD=.80) than in the baseline condition (M=5.3, SD=1.07) (Wilcoxon *W*=2.0, *p*=.02). For the statement '*The system enabled me to make connections between different inspirations*' the difference was marginal (M=6.5, SD=.80 in BioSpark, M=5.6, SD=1.56 in baseline, Wilcoxon *W*=1.5, *p*=.07).

6.6 Discussion

We introduced BioSpark, a system exploring the idea of acting as a creativity partner in analogical innovation. BioSpark builds on insights from a design workshop and formative pilot study to support not only finding inspirations but also transferring inspirations into the target design domain and more deeply engaging with them during ideation. We found in a user study that the LLM-enabled features we explored in BioSpark – generating and clustering inspirations, introducing sparks to help map the inspiration to the design problem, tradeoffs to help users consider design constraints, and free from chat to explore inspirations more deeply – resulted in participants generating more ideas and exploring more different species without a significant decrease in diversity compared to a 'gold standard' condition using AskNature inspirations and chatGPT. Furthermore, BioSpark appeared to keep users in the flow of ideation, reduce the cognitive effort in transferring and adapting ideas, and help people engage more deeply in considering how they could use inspirations and the design spaces they unlocked.

One significant concern we had was that the features that were aimed at deeper engagement, such as sparks, might counterproductively decrease engagement and increase fixation because of how fleshed out the connections were in terms of articulating an entire, detailed design idea embodying the inspiration's mechanism in the target domain (*e.g.,* using spider silk for lifting a wheelchair or creating a ramp). The higher the fidelity of an inspiration the more it may incur fixation and direct use rather than creative adaptation [255]. In our case we did not see evidence of this happening; instead, it led to deeper engagement and further exploration of the design space suggested by the idea.

Why did this occur? We believe there are several factors at play. First, the cognitive load of mapping the idea to the design space was reduced by the AI, but the decision to do the mapping in the first place was driven by the user by saving an inspiration. Thus before seeing the AI mapping they needed to notice something interesting or relevant about the inspiration, even if they didn't fully make the connection to the problem domain themselves. This self-driven curiosity and agency could play a role in their deeper engagement with the sparks and tradeoff cards. Future systems might explore what user actions and agency are necessary for them to feel ownership and spur initial engagement with inspirations.

Another factor that might have driven engagement was perceived ownership of inspirations. Previous work has identified that ownership and attribution are key elements of human-AI collaboration [202]. In our study we noticed participants making attribution statements about the sparks, such as "That's not really my idea, [it's] ChatGPT's idea but okay". A common theme among participants was discussing how they modified the sparks to make ideas more their own and to avoid "plagiarism", even though they were told they could use the sparks as they wish. It's unclear why designers in other studies finding fixation when using LLM and generative AI did not similarly adapt and riff on ideas in order to build ownership, but a possibility that might be explored is that an integrated system that frames AI-generated ideas as intermediate products, as we do in BioSpark, might be more effective at promoting deeper engagement than an unstructured system or one where they claim prominence as more final products.

Another concern we had in our system design was whether contextualizing the system interface features in the design problem would lead to narrowing of the design ideas users would explore. Sparks, tradeoff cards, and the freeform chat interface were all contextualized with the source design problem, with the goal of reducing the cognitive effort needed for users to engage with the details of the inspirations relevant to their goals. This largely appeared to hold true, with users finding the contextualization useful, and even sometimes "magical". While they were technically able to do this with the baseline system and sometimes did ("It feels like I got the seed like the very starting point idea from AskNature and then generating actual ideas from it was done by ChatGPT, like translating the seed into actual ideas"), the efficiency of the built-in contextualization was frequently mentioned as useful as a jumping off point rather than replacing cognitive work (*e.g.,* "it was able to produce things without me having to like prompt it. And I think that allowed me to spend more time, maybe thinking about specific connections between the mechanisms and the design features it was suggesting").

Overall, our results suggest a more nuanced consideration of context and fixation than previously considered, in which helping users reduce cognitive load throughout the analogical innovation process while avoiding fixation by keeping AI suggestions as intermediate products in the system flow could be a profitable paradigm to explore.

6.6.1 Limitations

Our findings and analysis have several limitations. Our access to professional designers working in large organizations doing ongoing work was limited to the design workshop, and our formative study and user study involved heterogeneous pools of participants that included freelance designers recruited from UpWork and design and PhD students recruited from an arts college. These study participants may not be representative of all design professionals.

There may be alternative interpretations of our data based on the operalization of the measures we used. Measuring creativity and ideation has been the subject of a large stream of research across multiple disciplines, and the particular measures we used are grounded in a specific subset of that stream we perceived as most relevant to the goals of measuring engagement with inspirations and the resulting quantity and diversity of ideas generated. Diversity in particular has been measured in many different ways, and our approach to it may be biased by choices we made in our computation pipeline, such as reframing ideas to control for word length or the particular LLM used.

While we developed our study protocol based on input about professional designers' practice, our user studies are also limited in terms of introducing artificial scenarios and are time-limited in a way that may not be representative of the constraints designers have in their jobs. Future studies deploying similar systems into designers' actual practice, or making such systems publicly available for volunteer usage could result in important learnings about the benefits and remaining challenges for supporting analogical innovation to have real world impact.

6.7 Conclusion

In this work we present BioSpark, an end-to-end system for generating a biological-analogical mechanisms dataset and an interactive interface that facilitates learning new biological mechanisms for design challenges and synthesizing new solution ideas inspired by analogical mechanisms. We imagine a future in which engineers and designers could find inspirations based on deep analogical similarity between mechanisms found in nature to problems common to engineering and design challenges to drive innovation across fields. Future work remains in this area to provide improvements on the pipeline for generating analogical inspirations from nature, and in supporting users' recognition and synthesis from them for materializing downstream innovations.

PID	Active Ingredient Inspiration	Species	Code	Participants’ Think-aloud & Related Behavior Descriptions
P1	Shape allows air to flow over and around (like in a sail)	Caryophyllales (seed)	**S1**	*“Okay. WindSail carrier.* ***Oh that’s pretty cool!*** *Inspired by the aerodynamically shaped seeds of Caryophyllales, this bike rack utilizes a lightweight, sail-like structure that harnesses airflow to reduce drag...”*
			D1	*“The sails are adjustable to snugly fit bikes, mimicking the efficiency of seed dispersal by wind... What does it mean by a sail”* (**Asks BioSpark a question to explain how ‘sail’ would work in the idea**) *“That’s a lot of information... Aerodynamic Shape. The sail-like structure of the WindSail Carrier is not just for aesthetic appeal; it serves a functional purpose by mimicking the shape of aerodynamically efficient seeds.* ***The shape allows air to flow over and around the bike rack... Oh okay now we’re getting something.*** *”*
P4	Appendages retract into an empty space	Turtle	**D2**	*“Okay, so, the tortoise shell* ***made me think about*** *how things can be folded into empty space. That was the thing I got from the tortoise. There’s empty space inside the shell and it can fold like it can take its feet into the shell. But it doesn’t break the feet into simple pieces, or fold it like, roll it like, or anything like that. Just takes the feed into empty shell. So that’s what I came up with, and then I started thinking about like, oh what does it mean to have a slot inside, and then I thought, airplane wheels, and the Alaska airline door incident, which made me think about the pin-release mechanism (that was supposed to hold the door).”*
P6	Small, powerful thrusts that allow for quick, upward propulsion and quick directional changes	Lepidoptera (butterflies and moths)	**S1**	*“Okay* ***that’s pretty interesting****, the propulsion (mechanism) and changing directions... that could be relevant to changing directions on wide stairs.”*
	N/A	N/A	**S2**	*“okay so since I don’t know these concepts from nature,* ***I need help in understanding whether*** *I can use that technology in this? So* ***it’s hard to understand the relevance.*** *”*
P11	Sliding and collapsing (like in a telescope)	Armadillo	**D2**	*“Okay so this shell that can collapse is an interesting mechanism. Like this* ***makes me think of a telescope****, like a* ***telescoping mechanism for sliding and collapsing****... so that could be a really interesting design space.”*

Table 6.1: Transcripts of participants’ think aloud and behavioral records in the BioSpark condition (‘Participants’ Think-aloud & Related Behavior Descriptions’) and how they were coded into different types of engagement patterns (‘Code’). The bold-faced text in each row highlights the important signatures of the assigned code. Each row also contains the following: the participant ID (‘PID’), the active ingredient description that participants found interesting / relevant (‘Active Ingredient Inspiration’), and the associated species (‘Species’). Exhibits in the baseline condition were similar, with the exception of tools participants interacted with.

A System Implementation Details

A.1 Structuring AskNature blog posts into seed problem-mechanism-organism schemas

To source a set of diverse, high-quality biological mechanisms for a given problem, BioSpark starts from a seed set of expert-curated biological mechanisms on AskNature (fig. 6.2, Step 1). AskNature.org provides a curated list of organisms with detailed descriptions of their unique strategies to functional problems (*e.g.,* 'Manage Impact', 'Modify Speed'). The organisms and strategies can be grouped by function and viewed as a list. To curate a seed set of high-quality mechanisms, we first choose a functional problem p predicted to be highly relevant to automobile designers, excluding irrelevant functions such as 'Adapt Behaviors', 'Adapt Genotype', 'Coevolve', 'Maintain Community' We access the sub-list of organisms $o \in O$ and strategies posted to p on AskNature's group-by-function page by parsing the `HTML` code using the `BeautifulSoup` package on `Python`. We then access the blog post for each organism-strategy page using the parsed URL and parse the returned `HTML` page to get the title, description, and references (if available).

At this stage, the returned unstructured text is yet to contain a succinct mechanism description. Furthermore, we found that some blog posts do not contain any body text despite having a title and are accessible via the URL. Some of these missing blog posts indicated that they are in maintenance and/or planned to be updated. To structure the raw blog post text AskNature$_{(o,p)}$, we prompt GPT4 [196] to succinctly describe (*i.e.,* using 12 words or less) the core mechanism (*i.e.,* excluding the qualities or effects, and focusing on mechanisms with engineering design implications), given (o, p) (if blog post text is missing) or $(o, p, \text{AskNature}_{(o,p)})$. The returned mechanism description m along with the function description makes up the problem-mechanism schema for each organism: $\{o \in O | (p, m, o)\}$.

A.2 Iteratively expanding mechanisms dataset by traversing constructed taxonomic trees

Using each schema as a seed, we iteratively prompt GPT4 to find relevant mechanisms for the given mechanism and problem, using an even mixture of breadth- and depth-focused expansion strategies (fig. 6.2, Step 2). To enable structured diversification of organisms and their mechanisms beyond prior work that relied on token-level manipulation or naïvely prompting LLMs, we guide LLMs *how* and *where* to expand by leveraging organism taxonomic hierarchies. At each iteration of expansion (fig. 6.2, Step 2), we aggregate the organisms represented in found mechanisms up to that point, and construct a taxonomic tree featuring seven levels of hierarchy on Tree of Life: {`domain, kingdom, phylum, class, order, family, genus, species`}, where `domain` representing the highest level and `species` representing the lowest level on the hierarchy.

Given this tree, we aim to identify sparsely populated branches for expansion. We cut the tree at a given reference expansion level (*e.g.,* `class`), and sort the taxonomic ranks (nodes) on that level by the number of its immediate children nodes[20], in an increasing order. For performance, we batch 10 prompts to send to GPT4 for expansion. For half of the prompts, we instruct **breadth-first expansion** which asks GPT4 to first identify *sibling* nodes at the given reference taxon level and existing nodes (up to 50 most populated nodes).

For example, the prompt asks "come up with a few biological `classes` not in {`...names of excluded`

[20] Alternatively, the entire size of the subtree, rather than immediate children, could be used for sorting

classes...}”. The breadth-first expansion prompt then instructs GPT4 to repeat the following: 1) Choose one taxon from the list it came up with; 2) Succinctly describe (*i.e.,* using 14 words or less) new mechanisms *m* related to a problem *p*. For the rest of the prompts, we instruct **depth-first expansion** which asks GPT4 to first identify a new *children* node at the given reference taxon level and existing children nodes (up to 50 randomly sampled children). For example, the prompt asks “come up with a few biological families in order araneae that are not any of {araneidae, ...}”. The depth-first expansion prompt then instructs GPT4 to repeat a similar procedure as breath-first expansion. The prompt details are provided in fig. 12 (the depth-focused expansion prompt) and fig. 13. In the prototype system, we run 10 batches for expansion to construct dataset of mechanisms for each problem.

The returned list of mechanisms and organisms text are then fed into the second GPT4 prompt for structuring them into a list of {mechanism, organism} dictionaries. Finally, using each organism name, we prompt GPT3.5-turbo to retrieve the seven-level taxonomic hierarchy, based on our model evaluation result showing its high accuracy (Appendix C).

```
[System Message]
You are an expert biologist who knows species and their taxonomic hierarchy in detail
↪ .
You can also come up with diverse problem-solving strategies found in nature relevant
↪  to engineering design problems.
Do the following step-by-step.
```

```
[User Message]
1. Come up with a few biological {lower-taxon-plural} **IN** the {taxon} "{term}" AND
↪  **NOT** {exclude-user-prompt}
2. Select one {lower-taxon_singular} from the list you came up with.
3. Come up with short descriptions (up to 14 words or less) of new mechanisms found
↪ in the selected {lower-taxon-singular} that are applicable to the challenge of "{
↪ prob}".
4. Repeat step 2 and 3 for each selected {lower-taxon-singular} and think step-by-
↪ step. Number each step in your thinking and make it as short as possible.
```

Figure 12: The prompt used for depth-focused expansion of the mechanism dataset. The “lower-taxon-singular” or “lower-taxon-plural” is the singular and plural name of the subsequent level on the tree-of-life hierarchy, of the level “taxon”, respectively. The “term” is the name of the selected taxon. The “exclude-user-prompt” includes previously generated “taxon” names which are used to instruct the LLM to avoid duplicate generation. The “prob” and “src-mech” contain the problem and mechanism schemas to constrain generation.

A.3 Representative Mechanism Image Curation

To aid designers’ visual understanding of and pique curiosity for biological-analogical mechanisms, we retrieve representative images for corresponding textual mechanism descriptions. We use Google Custom Search[21] with queries as “[organism name]:[mechanism description]” and the file type set to images and the safe search mode enabled. We choose the first place result of Custom Search as the visual representation of each mechanism.

[21] https://developers.google.com/custom-search/v1/overview

```
[System Message]
{same as in the depth-focused expansion prompt}}
```

```
[User Message]
1. Come up with a few biological {taxon-plural} **NOT IN** the excluded {taxon-plural
↪ } below:
{exclude-user-prompt}
2. {same as in the depth-focused expansion prompt}
3. {same as in the depth-focused expansion prompt}
4. {same as in the depth-focused expansion prompt}
```

Figure 13: The prompt used for breadth-focused expansion of the mechanism dataset. See the depth-focused expansion prompt (fig 12) for parameters descriptions.

A.4 Interacting with Mechanism Inspirations: Explain, Compare, Combine, and Critique

To facilitate designers' understanding and synthesis of mechanism inspirations, we develop several interaction features available on the interface (fig. 6.1). The **Explain** button is located in tooltips that pop up when the user places the mouse over on a mechanism card in the board UI (fig. 6.1, first panel). When the user clicks on the button, BioSpark sends a prompt to GPT4 requesting elaboration of the interacted mechanism and the organism in the context of the chosen engineering design problem. The **Compare** tab is located in the control bar of the sidebar of the interface. To use this, users need to first click on (at least) two mechanism cards from the left, saving them to the 'saved inspirations' panel at the top of the sidebar. There, users can check any two of the saved mechanisms they wish to compare. BioSpark sends a prompt to GPT4 when the user clicks on the tab, requesting comparison of pros and cons between the two mechanisms in the context of the chosen engineering problem. The result is formatted into a markdown table, with each mechanism as the header followed by pros and cons rows detailing each point. The **Combine** tab is also located in the control bar of the sidebar in the interface. Similarly with Compare, users can check two saved mechanisms they wish to see combined. BioSpark sends a prompt to GPT4 then requesting elaboration of a mechanism that combines the two selected mechanisms, and explain its potential advantages in the context of the chosen engineering problem. The result is also formatted into a markdown page using section title and headers for demarcating the content. Finally, the **Critique** button is located inside the Ideate tab. Upon clicking the Ideate tab, users can type in their own idea in the rich text editor in the opened tab, and optionally click on the button below to receive constructive feedback on it. BioSpark sends a prompt to GPT4 with the content of the text editor describing the idea, and requests additional revision that may improve the quality, such as anticipated failure modes and potential improvements.

A.5 Extending BioSpark to support any problem queries by incorporating rich problem-mechanism relations

One of BioSpark's limitations is its fixed problem queries. Though the five pre-generated problem queries provide a useful entry to mechanism organisms that may be applicable to a diverse set of design challenges, it comes at the cost of an inability to query biological mechanisms for *any* engineering design problems described in natural language text. As designers and engineers progress in interacting with the system, they may naturally come up with follow-up queries that may differ from the source queries, that could emphasize important constraints around the design problem, or specify low-level details newly understood to be important to consider. Adaptation to such evolving user query intent requires further personalization and scaffolding in the workflow. In future work, mixed-initiative workflows may leverage user interaction traces as input to LLM operations (*cf.* [142, 169]) to augment query input and automatically search data to retrieve analogical results.

In order to enable search by free-text problem queries, the underlying data model needs to be extended to contain multiple problem-mechanism relations beyond the single problem present in the schema $\{\forall i | (p_i, m_i, o_i)\}$, and into an enriched dataset with mappings between problems $p_1, p_2, \cdots, p_n \in P$ and an applicable mechanism m_k, as commonly the case in engineering (*e.g.,* 'spider silk' can be used for multiple engineering challenges such as replacing steel bars in concrete or wound suture and prosthesis [102]).

One way to expand the rich problem-mechanism relations in a scalable manner is to prompt LLMs to come up possible engineering design problems that a given biological mechanism could be applied to. Here, naïve prompting may suffer from conceptual redundancy, analogous to the challenge of curating diverse mechanisms, that limits the diversity in mechanism-to-problem mappings. Another approach may be to intelligently use the existing dataset $\{\forall i | (p_i, m_i, o_i)\}$ to identify similar mechanisms m_i in (p_i, m_i, o_i) and m_j in (p_j, m_j, o_j) that can be mapped onto disparate problems: $m_i \sim m_j \rightarrow p_i, p_j$. This approach however assumes the presennce of many correlated such mechanisms with disparate problem pairs in the dataset, which need empirical examination for support. Once the enriched dataset $\{\forall i | m_i \rightarrow S(m_i)\}$ (Here, $S(m_i)$ denotes the set of engineering design problems that m_i is applicable to) is made available, one simple approach for allowing querying on any problem text is to construct a similarity search index (*e.g.,* the HNSW index of FAISS [132]) using a chosen text embedding approach.

B Synthetic Mechanism Visualization

We explored two complementary approaches for visualization.

B.1 Direct Prompting Approach

In the direct prompting approach for synthetic mechanism visualization, we prompt Dall-E3 using a combination of the description of the mechanism and the species of the organism using the template in fig. 17. We set the image resolution for generation to `1024x1024` in size and `standard` quality resolution. We set the '`style`' parameter to `vivid`, which guides the model to "lean towards generating hyper-real and dramatic images"[22]. We repeatedly sample 3 images to incorporate visual diversity among the images.

B.2 Aspect- and Vantage-Point-Focused Prompting Approach

We noticed that the simple repeated sampling approach incorporated variations among the generated images focused on visual aesthetics. However, we also noticed that the variations did not sufficiently incorporate diversification based on different aspects involved with each mechanism. For example, though differing in the specific color palette or orientation used for depicting the organism (fig. 16, left), the images commonly featured an anatomical drawing of the whole body, using a top-down perspective, similar to what might be expected in displays in natural science museums. We also noted that generated images rarely focused on a single aspect or view point of the mechanism, and instead frequently incorporated many confusing visual details such as organism anatomy into one image, potentially misguiding viewers'

[22] `https://platform.openai.com/docs/api-reference/images/create`

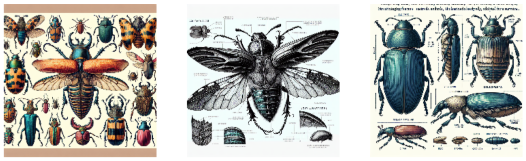


Figure 14: Direct prompting

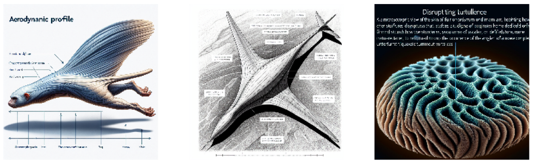


Figure 15: Aspect-focused

Figure 16: Synthetic mechanism visualizations using two approaches. (Left) Images generated via direct prompting on Dall-E3 for ‘Body shape streamlining found in coleoptera’. (Right) Images generated via aspect-focused prompting for ‘Body shape streamlining found in Dermoptera’. The three aspects identified by GPT4 for this mechanism were (from the leftmost image) ‘Aerodynamic Profile View’, ‘Top-Down Aerodynamics View’, ‘Microscopic Surface Texture View’, each highlighting different aspects of the mechanism (see Appendix B.2 for the full generated content).

```
{Mchanism Description: Body shape streamlining} found in {Species: coleoptera}
```

Figure 17: The prompt used in the direct prompting approach for synthetic mechanism visualization.

attention to less important aspects of the organism and the mechanism. Motivated by this, in the second approach we aimed to control for visual differentiation by focusing generation along one relevant aspect of the mechanism at a time. To this end, we developed a two-step generation pipeline: In the first stage, we prompt GPT4 to generate up to four unique aspects about a mechanism using an aspect- and vantage-point-elaboration prompt with both the mechanism description and the organism name. The prompt asks GPT4 to describe how the mechanistic aspect described could be best visually displayed using a particular vantage point for the species and to format the aspect descriptions into a list of JSON objects. In the second stage, we iteratively prompt Dall-E3 using each JSON object and the same text-to-image generation prompt (Appendix B.1) as before to generate mechanism visualizations along a single salient aspect. Pilot experimentation of the outcome of this pipeline showed a clearer emphasis on individual mechanism aspects and the use of a fixed vantage point for the corresponding aspect that might visualize the aspect effectively.

C BioSpark Dataset Pipeline Evaluation: Accuracy of LLM-based Taxonomy Construction

The main process in our diversification strategy is iterative construction of taxonomic trees at each stage of expansion with a set of problem-mechanism schemas and corresponding organisms $\{o \in O | (p, m, o)\}$ curated (in case of AskNature seeds) or generated up to that point. To construct the trees, the taxonomic hierarchy of each organism needs to be known. Here, we restrict our tree construction to seven levels of

Model	Domain	Kingdom	Phylum	Class	Order	Family	Genus
GPT4	100% (90/90)	100% (90/90)	100% (90/90)	100% (90/90)	96.7% (87/90)	94.4% (85/90)	98.9% (89/90)
GPT3.5-turbo	100% (90/90)	100% (90/90)	100% (90/90)	100% (90/90)	95.6% (86/90)	95.6% (86/90)	93.3% (84/90)

Table 2: The accuracy of zero-shot taxonomy generation using only the organism name.

depth, ranging from the highest to lowest levels: domain, kingdom, phylum, class, order, family, genus, species. These levels provide considerable branch-switching opportunities for diversification, through significant changes in the number of members between levels and within each level of the hierarchy. For example, while the highest level domain consists of three members, Bacteria, Archaea, and Eukarya, there are estimated 8.7M species in the world [239]. The next level on the hierarchy, Genus, has an estimated number of 310K members [214], while the number in the subsequent level, families, is estimated at 8K [184] in 2011. The number of known species for each node on the hierarchy also changes considerably, further contributing to the diversification opportunities. For example while most non-avian reptile genera have only 1 species each, insect genera such as *Lasioglossum* and *Andrena* have over 1,000 species each, while the flowering plant genus, *Astragalus*, contains over 3,000 known species [271].

Our initial exploration of suitable approaches to retrieve organism taxonomies involved using available resources such as the Global Biodiversity Information Facility API[23], Catalogue of Life [36], or the Encyclopedia of Life [75], where canonical species names were retrieved from the Darwin Core List of Terms[24] for corresponding organisms in problem-mechanism schemas. However, the limited coverage, data consistency, and API availability of these tools prevented their adoption. On the other hand, Wikipedia provides scientific classification for some of the organism articles (for example in the Pomelo article[25], taxonomic names for Kingdom, Clade, Order, Family, Genus, and Species are available in the 'biota' information box that appears on the right-hand side of the page). However, this data was not readily available for scalable generation.

C.1 Procedure

LLMs may provide an alternative solution to the limitations of existing approaches for retrieving the taxonomic hierarchy for a given organism name. To test this idea, we curated 90 gold taxonomies using Wikipedia that have complete information in the 'biota' scientific classification info box (the complete list of 90 organism names can be found in Appendix C.5). For each organism, we prompted LLMs with each organism name zero-shot using the chat completions API endpoint[26] using each model key. The prompt used for taxonomy generation for LLMs can be found in fig. 19. Once the hierarchy data is generated, we lower-cased the rank names for consistency.

C.2 GPT4's Accuracy

We find that GPT4's zero-shot taxonomy generation accuracy to range between 94.4% and 100% (Table 2). The lowest accuracy was observed in the family taxonomy, followed by order (96.7%) and genus (98.9%).

[23] https://www.gbif.org/developer/species
[24] https://dwc.tdwg.org/list/#dwc_Organism
[25] https://en.wikipedia.org/wiki/Pomelo
[26] https://api.openai.com/v1/chat/completions

C.3 Error analysis

We find that some error cases in taxonomy generation could be attributed to recent changes in classification in the literature. For example, both GPT4 and GPT3.5-turbo models classified naked mole-rats as then literature-accepted ‘Bathyergidae’ for their family, same as other African mole-rats. However, more recently naked mole-rats were placed in a separate family, Heterocephalidae [2].

Among the error cases overlapping between the two models, we found cases that either the GPT3.5-turbo or the GPT4 model wins over the other (*e.g.,* for ‘hummingbird’, GPT3.5-turbo generated ‘archilochus’ as its genus whereas GPT4 generated ‘various’; for ‘boxer crab’, GPT3.5-turbo generated ‘hymenoptera’ which is an order of insects, whereas GPT4 generated ‘decapoda’, which is the correct order). In other cases, both models outputted similarly incorrect answers, for example for ‘sea snail’, GPT3.5-turbo generated ‘neogastropoda’ whereas GPT4 generated ‘archaeogastropoda’ (the Wikipedia gold answer was ‘lepetellida’).

C.4 System Optimization: GPT3.5-turbo's Accuracy

We find that GPT3.5-turbo has comparable accuracy levels with GPT4 in zero-shot taxonomy generation. The highest misaglignment occurred in `genus`, with a 6.67% error rate (equivalent to 6 out of 90). Appendix C.3 provides a further qualitative error analysis of models' comparative performance. Based on these results, we opted for the more efficient GPT3.5-turbo model in our pipeline. We leave further exploration of the capabilities of smaller, fine-tuned base LLMs, with implications for LLM cascade[27], to future work.

C.5 Complete List of Organisms Used for Taxonomy Generation

```
{‘spidermonkey’,‘prairiedog’,‘gardentigermoth’,‘africansacredibis’,‘argiopeargen
tata’,‘ostrich’,‘groundhog’,‘daniorerio’,‘gannet’,‘deer’,‘cattle’,‘glyptodon’,‘a
lligatorsnappingturtle’,‘leopard’,‘arcticgroundsquirrel’,‘cormorantsandshags’,‘b
ears’,‘squirrels’,‘herons’,‘europeanbadger’,‘goldensilkorb-weaver’,‘aardvark’,‘s
eahorses’,‘banner-tailedkangaroorat’,‘hyenas’,‘pinkfairyarmadillo’,‘giantotter’,
‘bighornsheep’,‘hippopotamus’,‘californiagroundsquirrel’,‘europeanbee-eater’,‘be
echmarten’,‘leopardgecko’,‘tailorbird’,‘testudinidae’,‘emperorpenguin’,‘northern
pike’,‘giantclam’,‘stoat’,‘horse’,‘nutria’,‘tree-kangaroo’,‘giraffe’,‘guineababo
on’,‘ferret’,‘bonytailchub’,‘bayaweaver’,‘brooktrout’,‘pelican’,‘mallard’,‘rosea
tespoonbill’,‘mountainweasel’,‘pocketgophers’,‘lybiaedmondsoni’,‘giantanteater’,
‘commonraccoondog’,‘dewdropspiders’,‘armadillogirdledlizard’,‘arcticfox’,‘bison’
,‘swordfish’,‘baldeagle’,‘chimpanzee’,‘asbolusverrucosus’,‘spermwhale’,‘abalone’
,‘goldenjackal’,‘hornet’,‘zebra’,‘orangutans’,‘peregrinefalcon’,‘atlanticcod’,‘b
urrowingowl’,‘africanwilddog’,‘manedwolf’,‘honeybee’,‘nakedmole-rat’,‘echidnas’,
‘bowerbirds’,‘rhinoceros’,‘beaver’,‘bombyxmori’,‘commonboxturtle’,‘hummingbird’,
‘domesticsheep’,‘wolverine’,‘raccoon’,‘evergreenbagworm’,‘pig’,‘muskrat’}
```

[27] LLM cascade refers to a system design approach that adaptively chooses optimal LLM APIs for a given query. Smaller, task-specific LLMs are regarded as optimal when they exhibit higher or similar levels of performance compared to models that are orders of magnitude larger [69], with all else being equal.

D BioSpark Dataset Pipeline Evaluation: Increase in Organism Diversity

In order to evaluate the effectiveness of diversification through our expansion strategies from iteratively constructed taxonomic trees, we investigated how organism diversity changes upon a series of mechanism generation.

D.1 Procedure

We generate mechanisms and corresponding species for five problems closely related to automobile design: ‘managing impact’, ‘managing tension’, ‘managing compression’, ‘managing turbulence’, ‘modifying speed’. We index the species at the time of its appearance in the corresponding mechanism generation. Hence, the index corresponds to when a new mechanism was generated via our pipeline. At each generation index, we count the unique number of names that are generated up to that point, for each taxonomic rank, and average the numbers across the five problems.

D.2 The pattern of increasing organism diversity

Qualitatively we observe that the number of species generated are monotonically increasing (fig. 20), albeit at a decreasing rate. The ratio between the number of unique species and the generation index is 2:1 at index=200, and approaches 4:1 near index=800. This suggests more mechanisms are generated and become concentrated on individual species (*e.g.,* for grasshoppers, there might be several distinct mechanisms relevant to the problem of ‘managing turbulence’ such as their foldable wing structures, lightweight exoskeleton designs, or joints in their legs enabling repeated high jumps) on average as generation continues. In our future work, we will explore whether and how the mechanisms generated for the same species semantically differ from one another. In addition, we plan to examine how our generation approach compares to alternative curation (*e.g.,* [74], which explored a data programming approach for mining concise biological problem-solving inspirations on Wikipedia) or generation approaches in terms of the efficiency of generating mechanisms across diverse organisms, by measuring the slope of organism diversity over generation index.

E Formative Study System Interface

Based on the design workshop findings (§6.3.1), we designed an initial, functional protoype interface to test with participants. The protoype included a selector for generating biological inspirations that could address a design problem such as ‘design a secure bike rack for sedans’. The interface also included interactive features for explanation, comparison, combination, and critique of mechanisms that used GPT4 to generate corresponding content. Fig. 21 shows the interface design.

F BioSpark

F.1 Active Ingredient Extraction

The prompt used is detailed in fig. 22.

F.2 Ranking Candidate Species Images

The prompt used for ranking candidate species images is detailed in fig. 23.

F.3 Spark Generation

The prompt used for generating sparks is detailed in fig. 24.

F.4 Trade-off Analysis Generation

The prompt used for generating a trade-off analysis is detailed in fig. 25.

G Details of the User Study

G.1 Tutorials

Before participants start with each of the two main task in each condition, they were given a tutorial of the assigned systems via screen sharing. The interviewer demonstrated a step-by-step process and the main features of each system using a prepared script that took around 8 minutes for the BioSpark condition which had more features, and around 5 minutes for the baseline condition. In the baseline condition, participants were instructed to open up 5 different URLs each pointing to a pre-curated list of mechanisms for a functional category. The 5 functional categories used in the study were the same as those that were used for the BioSpark backend dataset pipeline, and they were: Manage Impact, Manage Tension, Manage Compression, Manage Turbulence, and Modify Speed. In addition, participants in the baseline condition were instructed to sign in and open ChatGPT, and freely use it for understanding and ideating for the design problems using the information found from AskNature. When participants came up with each new idea during the task, they were told to write it down in a prepared Google spreadsheet that was shared in the beginning of the task. In the BioSpark condition, participants were told to keep the stream space as a holder for their ideas, and thus delete any ideas they did not like or edit the text directly.

H Prompts Used for Pre-processing Ideas in the User Study

H.1 Prompt used for extracting coherent and unique design constraints

The prompt used is detailed in fig. 26.

H.2 Prompt used for extracting the species' name that inspired each participant's idea

The prompt used is detailed in fig. 27.

H.3 Prompt used for summarizing the participants' ideas

The prompt used is detailed in fig. 28.

```
[System Message]
Identify up to three different vantage points that illuminate the mechanism described
↪ .
The vantage points for the mechanism should consider the problem context of designing
↪  a secure bike rack for sedans, where management of turbulence is critical.
The vantage points should differ in terms of the scale and the angle of focus, such
↪ as micro versus macro or portrait, side, or top views and so on.
For each vantage point, describe in detail how the mechanism aspect according to the
↪ specific vantage point can be visually drawn.
Output your response in a list of JSON dictionaries with "vantage_point" as keys and
↪ their corresponding values.
For example, [{ "vantage_point": "...", "desc": "..."}, {"vantage_point": "...", "
↪ desc": "..."}, ...].
Do not output any text other than the JSON.
```

```
[User Message]
{Body shape streamlining} found in the species {galliformes}
```

```
[Example Generated Content for the mechanism 'Body shape streamlining found in
↪ Dermoptera' (in section "Synthetic Mechanism Visualization")]
[{
    "vantage_point": "Aerodynamic Profile View",
    "desc": "From a side view, the streamlined body shape of the gliding primate or
↪ rodent can be depicted with smooth, elongated contours that taper towards the rear
↪ . This profile would highlight the gradual reduction in cross-sectional area from
↪ the head to the tail, illustrating how the body shape is optimized for reducing
↪ drag. The limbs would be shown close to the body or extended in a way that aligns
↪ with the airflow, further reducing turbulence."
},
{
    "vantage_point": "Top-Down Aerodynamics View",
    "desc": "From a top-down perspective, the organism's body can be visualized as an
↪  elongated oval with a slight taper towards the tail end, indicating a teardrop
↪ shape that is known for its aerodynamic efficiency. The limbs, if extended, would
↪ be seen as narrow lines that either align with the body or are positioned to
↪ maintain the streamlined form. This view would emphasize the symmetry and
↪ smoothness of the body, which are critical for minimizing air resistance and
↪ managing turbulence."
},
{
    "vantage_point": "Microscopic Surface Texture View",
    "desc": "At a microscopic scale, the surface texture of the organism's skin or
↪ fur can be illustrated to show how it may contribute to reducing turbulence. The
↪ visual would depict any specialized structures, such as scales or fur patterns,
↪ that are designed to disrupt the laminar flow just enough to reduce the formation
↪ of larger, more chaotic turbulent vortices. This close-up view would focus on the
↪ intricacies of the surface that interact with the air at a very small scale, which
↪  collectively contribute to the overall aerodynamic performance."
}]
```

Figure 18: The prompt used in the aspect- and vantage-point-focused prompting approach for synthetic mechanism visualization and a sample generated output. The content in curly brackets ({...}) in the user message portion were replaced with user-specific content.

```
[System Message]
You are an expert biologist who knows species and their taxonomic hierarchy very well
↪ . Follow the instructions to the letter.
- Return the scientific term for each taxonomic rank the species belongs to.
- Enclose keys and values using double quotes ("...") and format them into a Python
↪ dictionary.
- Use the taxonomic ranks as keys and corresponding scientific terms as their values.
- Do not add any other text.
```

```
[User Message]
What {"domain", "kingdom", "phylum", "class", "order", "family", "genus"} does "{
↪ organism}" belong to? Format your reply into a Python dictionary.
```

Figure 19: The prompt used to generate the taxonomy of each organism.

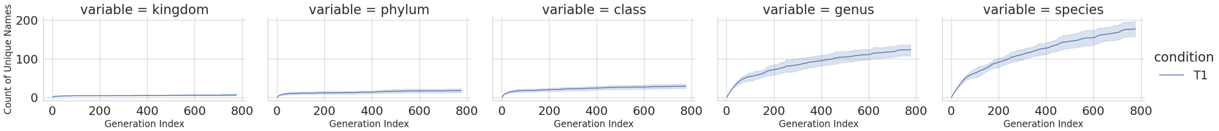
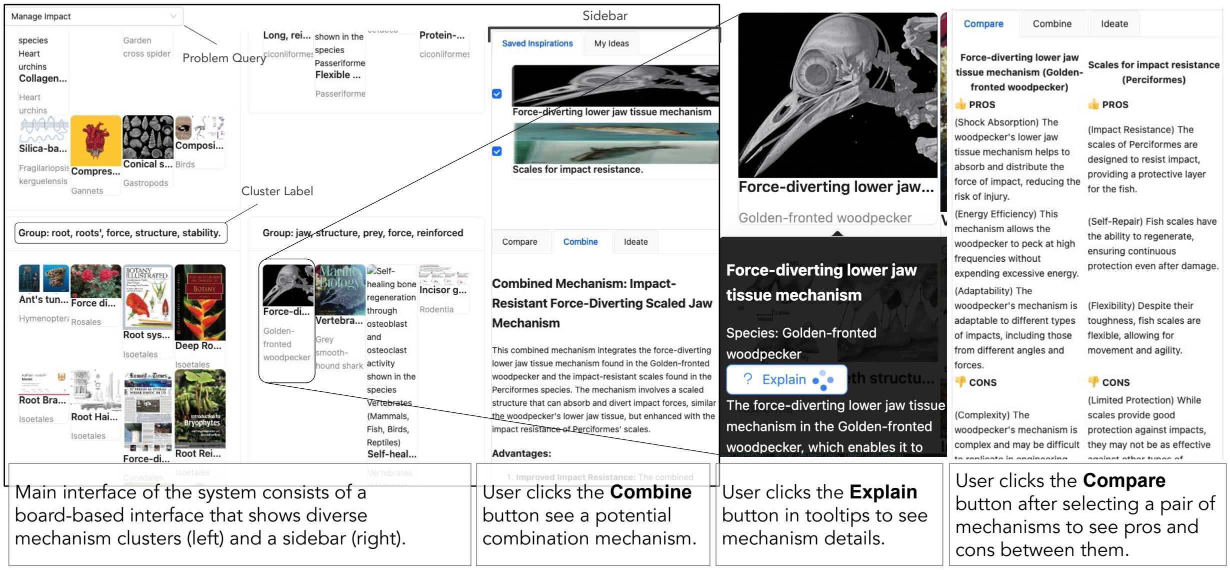


Figure 20: Organism diversity, measured by the number of unique names among the generation, increases monotonically as generation continues, while the number of mechanisms generated per organism also increases, as evidenced by the decreasing slope. This suggests mechanism concentration over each organism increases over time.



Combined Mechanism: Impact-Resistant Force-Diverting Scaled Jaw Mechanism

This combined mechanism integrates the force-diverting lower jaw tissue mechanism found in the Golden-fronted woodpecker and the impact-resistant scales found in the Perciformes species. The mechanism involves a scaled structure that can absorb and divert impact forces, similar the woodpecker's lower jaw tissue, but enhanced with the impact resistance of Perciformes' scales.

Advantages:

1. Improved Impact Resistance: The combined

Force-diverting lower jaw tissue mechanism

Species: Golden-fronted woodpecker

The force-diverting lower jaw tissue mechanism in the Golden-fronted woodpecker, which enables it to

Force-diverting lower jaw tissue mechanism (Golden-fronted woodpecker)	Scales for impact resistance (Perciformes)
👍 PROS	👍 PROS
(Shock Absorption) The woodpecker's lower jaw tissue mechanism helps to absorb and distribute the force of impact, reducing the risk of injury.	(Impact Resistance) The scales of Perciformes are designed to resist impact, providing a protective layer for the fish.
(Energy Efficiency) This mechanism allows the woodpecker to peck at high frequencies without expending excessive energy.	(Self-Repair) Fish scales have the ability to regenerate, ensuring continuous protection even after damage.
(Adaptability) The woodpecker's mechanism is adaptable to different types of impacts, including those from different angles and forces.	(Flexibility) Despite their toughness, fish scales are flexible, allowing for movement and agility.
👎 CONS	👎 CONS
(Complexity) The woodpecker's mechanism is complex and may be difficult to replicate in engineering	(Limited Protection) While scales provide good protection against impacts, they may not be as effective against other types of

Main interface of the system consists of a board-based interface that shows diverse mechanism clusters (left) and a sidebar (right).

User clicks the **Combine** button see a potential combination mechanism.

User clicks the **Explain** button in tooltips to see mechanism details.

User clicks the **Compare** button after selecting a pair of mechanisms to see pros and cons between them.

Figure 21: Formative study prototype interface and a subset of available interaction features (excluding the 'critique' button that was also available to study participants). The interface consists of a left-hand side panel that shows clusters of semantically similar mechanisms that was scrollable, and a right-hand side panel that included a holding tank for user-saved mechanisms. When the user checks two of the saved mechanisms in the holding tank and clicks one of the tabs underneath, the system generated the corresponding content, such as the comparison of two mechanisms in a pros-and-cons table, a new idea that combines the two mechanisms, and the 'Ideate' button that provided critique on the 'Combine' idea.

```
[System Message]
Reply with a succinct (i.e., 15 words or less) description of the following
↪ biological mechanism's active ingredient. Follow the instructions.
[Instructions]
- The active ingredient should describe how the species "act" upon its challenges to
↪ mitigate them, and include verb or verb phrasees.
- Active ingredient descriptions should also focus on the integral ingredients such
↪ as its bodily parts, liquids, or evolutionary tactic that are concrete and
↪ distinctive.
- Structure your output in the following format (do not output any characters other
↪ than the actual json-formatted dictionary):
{"desc": "..."}
```

```
[User Message]
{mechanism}
```

Figure 22: The prompt used to extract the active ingredient from a mechanism. The mechanism is description to extract from is provided as part of the user message to GPT4.

```
[User Message]
Judge each image based on how clearly it shows the real species (i.e., photos
↪ focusing on one instance of the species in the wild is better than cartoons,
↪ drawings, or species photographed in the distance)
{species} and contains visual details that help viewers understand the following
↪ biological mechanisms:
===
{formatted_mechanisms}
===
For each image given, reply with a number between 0 and 100 as its "score", where a
↪ higher number represents a higher quality of the picture,
and also provide rationale for your decision in "rationale".
Output a list, with the following format. Exclude any other character than the comma
↪ between dictionaries in the list:
[{{"score": "50", "rationale": "..."}}, ...]
```

Figure 23: The prompt used to extract the active ingredient from a mechanism. The mechanism is description to extract from is provided as part of the user message to GPT4.

```
[System Message]
Generate **2** highly different ideas that could solve the design problem: "{
↪ design_prob}".
The design problem has constraints that the ideas must satisfy: {design_constraints}
Generated ideas must be at least broadly related to the user-selected inspiration
↪ found in nature.
Generated ideas should be novel and not redundant with the following ideas generated
↪ in the past: {prev_sparks}
Describe each idea succinctly (i.e., max 500 characters), but ensure to provide
↪ sufficient details to help the user visualize the idea.
Start each idea description with a short, eye-catching name that captures the gist.
Output exactly in the following format, WITHOUT ANY OTHER TEXT:
[{{"name": "IDEA 1 NAME", "desc": "IDEA 1 DESCRIPTION"}}, ...]
```

```
[User Message]
User-selected inspiration from nature to base your generation on:
===
{user_selected_mechanism_inspiration}
===
```

Figure 24: The prompt used to generate new sparks. We contextualize the prompt using the design problem description and the constraints provided with the problem, as well as 20 previously generated sparks for precedent-based diversification. We explicitly instruct the model to generate non-redundant sparks based on the history of precedents, and be succinct (*i.e.,* under 500 characters), with a descriptive title. Finally the user-selected mechanism inspiration is provided as part of the user message to GPT4.

```
[System Message]
Generate up to **3** anticipated pros and cons for applying the user-selected
↪ mechanism to the design problem: "{design_prob}".
The design problem has constraints that the ideas must satisfy: {design_constraints}
Format the 'pros' and 'cons' into each column in a markdown table.
Place the header row at the top of the table: "|        **PROS** |        **CONS**
↪ |".
After the header row, place each 'pro'-'con' row; start each 'pro' or 'con' text with
↪  a succinct label (3 words or less), enclosed in parantheses.
```

```
[User Message]
User-selected inspiration from nature to base your generation on:
===
{user_selected_mechanism_inspiration}
===
```

Figure 25: The prompt used to generate a new potential design trade-off analysis. We contextualize the prompt using the design problem description and the constraints provided with the problem. We instruct the model to return the ‘pros’ and ‘cons’ of the mechanism inspiration in the context of the design problem using a markdown table format that places each pro-and-con pair in a new row, and give each item in the table a succinct (3 words or less) label. Finally the user-selected mechanism inspiration is provided as part of the user message to GPT4.

```
[System Message]
We're preparing each idea description for measuring the degree to which various
↪ design constraints were considered.

Chunk the idea description into unique segments each of which describe consideration(
↪ s) for a single coherent design constraint.
Output exactly in the following format (use double quotation marks to encapsulate any
↪  string), without any other text:
{"constraint_considerations": [["idea 1 constraint 1", "idea 1 constraint 2", ...],
↪ ... }
```

```
[User Message]
{list_of_participant_ideas}
```

Figure 26: The prompt used to extract coherent chunks of text that relates to unique design constraints. Participants' ideas are stringified and provided as part of the user message to GPT4.

```
[System Message]
We're extracting the source species' name that inspired each idea from the idea
↪ description.

Output exactly in the following format (use double quotation marks to encapsulate any
↪  string), without any other text:
{"source_species": ['species name for idea 1', 'species name for idea 2', ...] }
```

```
[User Message]
{list_of_participant_ideas}
```

Figure 27: The prompt used to extract the species' names that inspired participants' ideas. Participants' ideas are stringified and provided as part of the user message to GPT4.

```
[System Message]
We're preparing each idea description for measuring the diversity.

One confounder for the diversity measure is the length of each idea and the descriptions of constraints that are not core to
↪ the idea itself.

To address this, your job is to summarize each idea using exactly 10 words (or less) to capture the core active ingredient.
Focus on how the bike rack is constructed, or what it looks like, rather than what it was inspired by.
Follow the examples:
===
Example 1:
{idea: ”Scots Pine − Hanging bike racks by mimicing the scots pine structure since it is very good at handling the weight
↪ stress of the bikes, the flexible wood is an important feature to consider when dealing with wide range weighted bike like
↪ light weight to heavy weight. Also the climate witholding is an exceptional add−on to have.”, 'probId': 1}

Should be summarized to:
'Hanging bike racks mimicking scots pine'

Example 2:
{idea: ”Tortoise Shell Geometry − For the bike racks stability & durability, I would use layers of different properties to add
↪ on or improve the strength mimicing the totroise shell geometry. Also I would adpat this feature since it prodvides very good
↪  longitivity of the product which enhances user satisfaction & trust.”, 'probId': 1}

Should be summarized to:
'Use layers like tortoise shell geometry'

Example 3:
{idea: ”Include articulated joint systems at key points such as the seat−to−frame and frame−to−wheel connections; these would
↪ allow for controlled movement in response to sudden impacts; engineers could use elastomeric materials for flexibility or
↪ incorporate spring−loaded mechanisms for controlled movement”, 'probId': 2}

Should be summarized to:
'Articulated joint systems at key points'

Example 4:
{idea: ”BeetleGlide Stair Wheelchair. Drawing from the ironclad beetle's exoskeleton, this wheelchair uses a bio−inspired shell
↪  design for strength and lightness. It incorporates a unique, foldable track system that extends around stairs like a
↪ caterpillar, enabling safe ascent and descent. The tracks are made from a durable, lightweight material that grips stairs
↪ securely, while the wheelchair's body folds by collapsing the seat and retracting the tracks into the shell, ensuring it
↪ meets the compactness requirement.”, 'probId': 2}

Should be summarized to:
'Retractable shell−like track system for gliding'
===

Note that the problem ID 2 corresponds to a ”bike rack” design problem and 1 corresponds to a  ”wheelchair” design problem.
Ensure that your the summarized description include ”bike rack” or ”wheelchair” in its description.

Ensure that you are not adding any additional content to the description other than the design domain as described above.

Output exactly in the following format, without any other text:
{”summarized_ideas”: ['idea 1', 'idea 2', ...]}
```

```
[User Message]
{list_of_participant_ideas}
```

Figure 28: The prompt used to summarize the participants' ideas. Participants' ideas are stringified and provided as part of the user message to GPT4.

Chapter 7: Discussion

The four systems described in this thesis demonstrate the possibilities for human-AI interaction design that centers on conceptual abstraction to increase human cognitive efficiency and creativity in problem-solving. From the need-finding studies, as well as the observational and interview data of the people interacting with the systems developed as part of this thesis, I synthesize design implications for future systems that focus on interaction with conceptual abstractions for knowledge discovery and insight generation.

7.1 Supporting Exploration of Many Abstractions

When multiple abstraction artifacts are created that lay diverging exploratory paths, how should design support this exploration of abstractions? In order to make good progress during the exploration, one needs to understand and evaluate the goodness of each abstraction and make decisions around what kinds of data are relevant and needed to further refine the abstraction and continue the exploration [152].

To illustrate this process, consider the following scenario of an intelligence officer gathering counterintelligence information, abstracting a pattern, and assessing its threat level. In this example, an intelligence officer is tasked with figuring out the nature of overflight activity around nuclear power plants[1] that significantly increased in frequency in reports following 9/11. The first possible explanatory pattern might be that this is a threat from terrorists who suddenly increased their surveillance activities. Alternatively, the officer might pursue a different explanation – that this is a series of random coincidences. Exploring each of these patterns further guides the decision around what kinds of data are relevant. For example, the officer may gather a candidate terrorist organization's tactics manual to see whether such a tactic would align with the organizational goal. To determine whether the reports indeed contained a series of random incidents, the officer could gather information about the pilots and the routes the planes were taking. These data become 'critical constraints' of the pattern, or 'anchors' in the terminology used in [152]. The manual the officer acquired clearly instructs members of the organization not to bring attention to themselves. Moreover, after analyzing the flight paths, the officer finds that 90% of the flights were heading from the east and flying at low altitudes, about 1,500 feet. Both types of data are misaligned with the first pattern. The regularity in the second data refutes the second pattern. Based on this information, the officer considers another pattern – that these flights perhaps were by less informed, training pilots who were following standard instructions in the pilot manual. Finally, the officer finds instructions that describe 'visual flight rules' about how pilots, when lost, should look for nuclear power plants. Because they are so easily sighted and recognizable as landmarks, they are useful for getting one's bearings. The instructions further specify that students should fly east to west and low. The officer reinforces this explanatory pattern by interviewing a flight instructor who described this as common knowledge for a generation of pilots. One remaining mystery is why the frequency of sightings only increased after 9/11. The officer finds an explanation for this: critical infrastructure, including nuclear power plants, was designated as "temporary restricted flight" zones following 9/11, meaning the reporting sensitivity likely increased, suggesting that such flight patterns could have existed for a longer period but were only being picked up following the

[1] The unmodified version of the example can be found in [152])

designation.

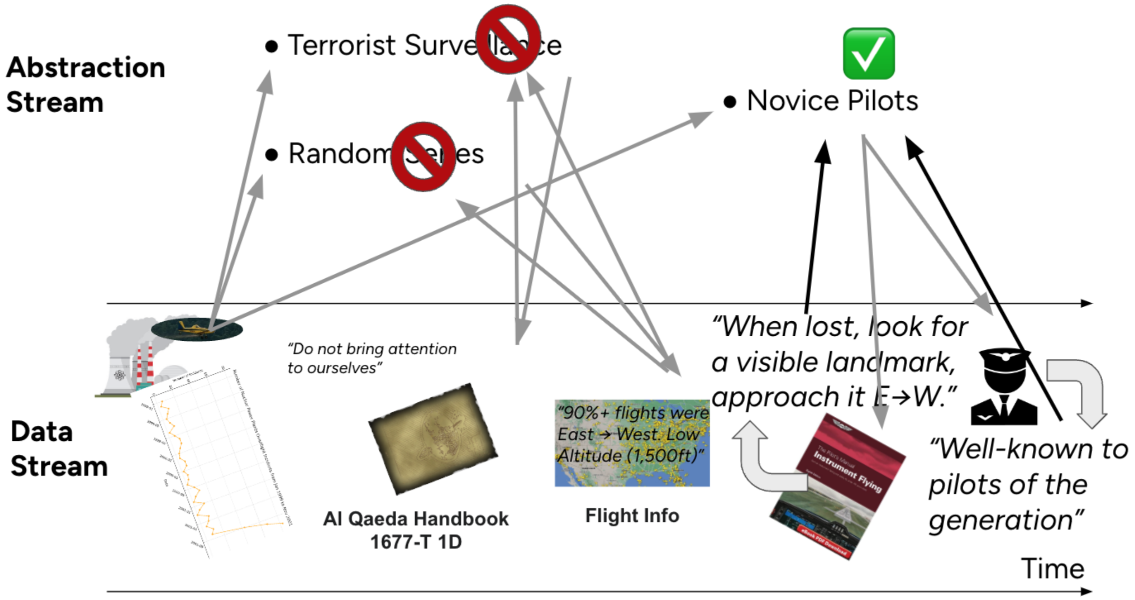

Abstraction Stream
• Terrorist Surveillance
• Random Series
✅
• Novice Pilots
Data Stream
"Do not bring attention to ourselves"
Al Qaeda Handbook 1677-T 1D
"90%+ flights were East → West. Low Altitude (1,500ft)"
Flight Info
"When lost, look for a visible landmark, approach it E→W."
"Well-known to pilots of the generation"
Instrument Flying
Time

Figure 7.1: Visual representation of the scenario found in [152]: 'intelligence officer tasked with figuring out the nature of overflight activity around nuclear power plants'.

This example, as visually shown in fig. 7.1, illustrates the notion of *critical (anchor) constraints* and how their iterative discovery guides subsequent abstractions' formation by confirming or refuting them. In this sense *the goodness of abstraction* rests on *how well its representation aligns with existing critical constraints*. Because of the iterative nature of discovery, designing a layer for *superimposed structures that can overlay abstraction representations on top of source materials* without altering them permanently and linking them for efficient access (similar to how discussion artifacts were designed over the raw discussion threads [277]) may be a helpful design principle. Furthermore, as shown in the example, people could explore multiple alternative abstractions at once and may need to branch back to earlier abstractions as new critical constraints emerge, acquiring new data as needed, and propagating changes to other data as they go.

Large design spaces remain open as to how AI may assist in this process of data acquisition, and generation and propagation of changes as users explore multiple abstractions. For example, *AI could assist people in defining data relevance criteria, and retrieving appropriate sources from the Web*. This direction may be cautiously explored with development of new AI agents that can operate tools such as web interface manipulation or web APIs. For dynamic generation and propagation of appropriate updates, approaches that extend initial works such as PaperWeaver [163] may be explored to automatically generate textual update descriptions anchored on a previous state, and by combining it with an additionally developed module that translates the text-based description into appropriate update lists. Luminate is also a notable work here, where a similar pipeline for automatically generating possible dimensions of information with an LLM first, then generating information for each combination of dimensions [234] could be useful for systematically generating and maintaining change lists across many target dimensions.

7.2 Supporting Usefully Misaligned Abstractions

From observations of users interacting with the analogy search engine in laboratory and case studies, I found that they were motivated to iteratively and flexibly re-represent their purpose queries at different levels of abstraction. Such re-representation allowed them to uncover novel abstractions that might be missed with exact matches, and prompted new ways to adapt ideas even if they were initially misaligned.

For example, users frequently removed certain constraints from a purpose representation in response to the returned search results, thereby moving it to a higher level of abstraction and casting a wider net for search results. They also added further constraint descriptions to the query to filter results misaligned with critical constraints. This observation was echoed in how users interacted with Synergi-generated thread structures, where they commented on how some threads were too narrow in their conceptual focus within a broader structure to be represented at the same level in the hierarchy as other threads, while other times a thread was thought to be too broad-scoped and would benefit from further delineation of differences among its members.

From these insights, I present the design implication of supporting usefully misaligned abstractions that facilitate the discovery of important constraints in the problem representation. For the misalignment to be useful, one should be able to engage with its examples to reveal latent assumptions, identify core constraints, and learn how they might formulate more effective queries. Given the cognitive load involved in the process of mapping relations between abstractions, systems should reduce the effort required by retrieving results that are usefully misaligned, helping users to recognize rather than recall critical constraints.

In addition, because users often struggle to come up with queries that suitably represent the core abstraction while also avoiding unintended intersections of domains that lead to irrelevant results, supporting efficient exploration of abstractions is key. For example, users may engage in dropping or altering certain constraints included in the representation and resume exploration with the revised representation. Various such configurations could be explored in parallel, and their results incorporated back to guide subsequent explorations.

7.3 Abstraction and Exploration History as Boundary Object

Users interacting with Synergi-generated thread structures noted that sometimes the generated threads seemed mismatched at the level of abstraction. For example, the delineation between two different threads “Newcomer Integration in OSS Projects” and “Newcomer Barriers,” despite each containing a coherent and disjoint set of clips, felt too similar in level for a user to be separated. On the other end of the spectrum, a thread on “Numeric and Logical Reasoning” for a user whose research interests included techniques for sub-sentential representations for LLMs, felt it focused on too specific an aspect of the literature that does not rise to the level of abstraction the rest of the structure was focusing on. This suggests a potential limitation of using an individually user-curated thread as a boundary object that AI uses to expand user synthesis. While users could engage with the AI-generated structure and benefited from it to visualize a literature review outline that synthesizes across multiple papers rather than from a single paper, the tail parts of the structure skewed too high on the abstraction ladder and too generic or too low in level and too specific to be really useful.

This means AI would need further context surrounding the boundary object to unpack the intricacies in user intent. In BioSpark, the LLM handling user requests to generate new ‘sparks,’ which were application ideas that showed ways in which the source inspiration could be applied to the target design domain to help designers bridge the transfer gap, was provided with the previous history of generation to guide its generation towards more conceptually diverse ideas. Even with such a simple concatenation of generation history, the LLM showed improvement in the diversity of generation by avoiding the generation of immediately similar ideas that the user has already seen. Using the history more judiciously by focusing on how users interacted with prior information at the level of conceptual abstraction, future work may improve on the diversity metric even further.

To scaffold user learning, the system may curate conceptually diverse abstractions and their misaligned examples, presenting them to the user in a manner that supports maximum gains in both user learning and AI understanding of user intent. In the literature, two relevant approaches have been explored for a similar purpose. In the context of a problem concerned with detecting traffic lights in noisy images from the perspective of a car on a road, Mozannar et al. used an iterative algorithm for finding examples congruent with one another to belong to a region given an optimization object of human learning gains for forming a region [186]. Approaches that generalize this line of work to identify regions based on the user's changing mental models of conceptual abstraction and multiple possible ways that each data point can be conceptualized could effectively augment human onboarding processes. Furthermore, human domain experts' insights may be crowdsourced to provide conceptual abstractions for examples to learn 'concept activation vectors', similar to Cai et al.'s work [38], to bootstrap learning in a high-stakes domain.

7.4 Supporting Curation of Parts from Generated Abstraction Outputs

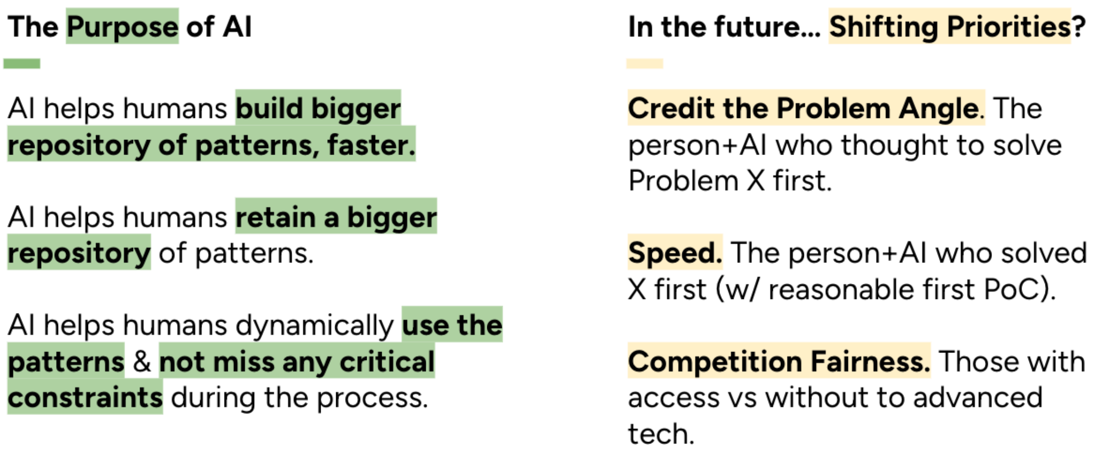
In addition, we found that users often engaged in taking only parts from the system generated abstractions and incorporated them into their own output. For example, in BioSpark, users also engaged in selecting the best features from various inspirations to form a new idea (*e.g.,* combining the idea of 'sliding scale' from armadillos and 'lubricating surface' from mucus snails to produce an idea around an easily expandable telescoping rack for bikes). In Synergi, sub-selection and curation were supported via dragging and dropping the useful threads into the user's own editor.

Extrapolating from these examples, for going beyond a common chat-based interface for LLMs, it would be important to support ways to easily select and curate only a subset of abstractions from the broader structure to improve their usefulness. Supporting user intent for sub-selecting from the system-generated abstraction outputs while learning from the signals embedded in the rest of the structure could meaningfully improve the model of user intent for the system. For example, the threads users selected before moving on to generate other system outputs could indicate that the rest of the threads in the structure that were not selected constitute negative examples that did not meet the level of abstraction the selected threads fit well. Learning from the comparisons between these cases, therefore, could provide valuable nuances that the system missed before.

7.5 Abstraction as Lens

Zooming out a little bit, abstraction can be understood as a lens for users' interpretation and authoring processes. Streamlining users' flow of data interpretation – by, for instance, making important parts of the new information salient through highlighting its conceptual relations to prior knowledge – and personalizing the interpretation of data using adaptive levels of abstraction can significantly aid comprehension. Moreover, structuring outlines for authoring tasks can bring further clarity and focus. This implies that focusing interaction design on abstraction in future systems opens up new spaces of design. These systems would support users in easily moving between different levels of abstraction, matching system presentation levels accordingly, and anticipating users' abstraction intents. By acting on these intents, the system can provide timely and relevant insights, thereby assisting users more effectively in their various tasks. Ultimately, an emphasis on abstraction could make systems more intuitive, responsive, and aligned with the cognitive processes of their users.

7.6 Broader Implications



The Purpose of AI

AI helps humans **build bigger repository of patterns, faster.**

AI helps humans **retain a bigger repository** of patterns.

AI helps humans dynamically **use the patterns** & **not miss any critical constraints** during the process.

In the future... Shifting Priorities?

Credit the Problem Angle. The person+AI who thought to solve Problem X first.

Speed. The person+AI who solved X first (w/ reasonable first PoC).

Competition Fairness. Those with access vs without to advanced tech.

Figure 7.2: Broader Implications for Design

Envisioning future systems that focus on end-user interaction with conceptual abstractions, I consider the following broader implications for design. The purpose of AI systems is to help humans build a bigger repository of patterns faster, retain that repository of patterns, and dynamically apply them with high accuracy and yield in complex problem scenarios. As such, the effectiveness of designs could be measured according to these goals. Projecting into the future, some factors currently important may only become more critical as human-AI team capabilities increase. As the execution cost drops, the importance of understanding how to approach a problem and the types of problems to solve will be amplified. The speed at which a reasonable proof of concept can be developed will remain important, and concerns regarding systemic fairness will become even more pronounced as access to these tools becomes increasingly tied to costly resources.

Chapter 8: Conclusion

In 1945, Vaneveer Bush proposed a vision for an associative trails machine that never forgets information and assists users in their day-to-day inquiry and knowledge work [35]. In the years since, research into related areas has provided valuable models for how humans make sense of complex information, and how the data and the conceptual abstraction generated during information exploration guide one another. This dissertation argues that the challenges we face today in the innovation economy, such as information overload and the difficulties associated with synthesizing valuable insights, can be addressed by focusing user interaction and AI design on how we synthesize conceptual abstractions from prior knowledge and use them to discover new knowledge and insights.

To this end, this thesis presents four sensemaking and ideation systems that demonstrate how end users, such as scientists, engineers, and designers, could be supported in curating, engaging with, and utilizing conceptual abstractions.

Key contributions include the development of Threddy and Synergi, systems that demonstrate efficient extraction and human-AI collaborative expansion of research threads, providing a foundation for more effective literature synthesis. Additionally, the Analogical Search Engine and BioSpark systems showcase the potential for LLM-powered analogical processing to generate innovative ideas, emphasizing the importance of purpose-mechanism schema abstractions.

The empirical studies and controlled lab experiments conducted with these systems reveal important design principles and interaction techniques that can enhance user engagement and support the practical transfer of inspirations. These findings and design lessons point to future systems aimed at addressing the challenges posed by the increasing complexity and scale of knowledge production, ultimately aiming to improve innovator productivity.

In conclusion, this dissertation advocates for the development of mixed-initiative systems that combine user input with advanced AI capabilities. By empowering users to better synthesize and utilize conceptual abstractions, these systems hold the promise of unlocking new avenues for innovation and effectively leveraging the vast repository of archived knowledge. The insights and design principles derived from this research have far-reaching implications, potentially transforming how individuals across various domains engage with complex information and generate novel solutions.

Bibliography

[1] Ask your pdf. `https://askyourpdf.com/`. Accessed: 2023-04-05. 2.3.2, 6.2.3

[2] Naked mole-rat. `https://en.wikipedia.org/wiki/Naked_mole-rat`. Accessed: 10-02-2023. C.3

[3] Annoy: How it works. URL `https://github.com/spotify/annoy#how-does-it-work`. [Online; accessed 23-Jan-2022]. 5.2.2

[4] Random projection in locality-sensitive hashing. URL `https://en.wikipedia.org/wiki/Locality-sensitive_hashing#Random_projection`. [Online; accessed 23-Jan-2022]. 5.2.2

[5] Grobid. `https://github.com/kermitt2/grobid`, 2008–2021. 3.3, 3.3.1, 4.3.1

[6] Rediet Abebe, Nicole Immorlica, Jon Kleinberg, Brendan Lucier, and Ali Shirali. On the effect of triadic closure on network segregation. *Economic Review*, 97(3):890–915, 2022. A.1

[7] Yuan An, Jeannette Janssen, and Evangelos E Milios. Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6:664–678, 2004. 4.3.1

[8] Salvatore Andolina, Khalil Klouche, Jaakko Peltonen, Mohammad Hoque, Tuukka Ruotsalo, Diogo Cabral, Arto Klami, Dorota Głowacka, Patrik Floréen, and Giulio Jacucci. Intentstreams: smart parallel search streams for branching exploratory search. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 300–305, 2015. 2.2.1

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 2.2.2

[10] Michael F Ashby and Kara Johnson. *Materials and design: the art and science of material selection in product design*. Butterworth-Heinemann, 2013. 2.3, 6.1, 6.2.1

[11] Kevin D Ashley. Reasoning with cases and hypotheticals in hypo. *International journal of man-machine studies*, 34(6):753–796, 1991. 2.3, 5.1, 6.2.1

[12] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing, 2022. URL `https://arxiv.org/abs/2203.00130`. 2.1.2

[13] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*, 30(5):1–38, 2023. 6.2.3

[14] Yoram Bachrach, Yehuda Finkelstein, Ran Gilad-Bachrach, Liran Katzir, Noam Koenigstein, Nir Nice, and Ulrich Paquet. Speeding up the xbox recommender system using a euclidean transformation for inner-product spaces. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 257–264, New York, NY, USA, 2014. Association for

Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645741. URL https://doi.org/10.1145/2645710.2645741. 5.2.2

[15] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. Elastic documents: Coupling text and tables through contextual visualizations for enhanced document reading. *IEEE Transactions on Visualization and Computer Graphics*, 25:661–671, 2019. 2.1.2

[16] Jinheon Baek, Alham Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, pages 70–98, Toronto, ON, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.matching-1.7. URL https://aclanthology.org/2023.matching-1.7. 6.2.2, 6.4.2

[17] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016. 5.2.1

[18] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 2.2.2, 4.1

[19] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021. 5.6.3

[20] Charles Bazerman. Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written communication*, 2(1):3–23, 1985. 2.1.2, 4.1

[21] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E. Mackay. Generative theories of interaction. *ACM Trans. Comput.-Hum. Interact.*, 28(6), nov 2021. ISSN 1073-0516. doi: 10.1145/3468505. URL https://doi.org/10.1145/3468505. 2.1.1

[22] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371. 5.7.2

[23] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003. 5.1

[24] Justin M. Berg. The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes*, 125(1):1–17, 2014. ISSN 07495978. doi: 10.1016/j.obhdp.2014.06.001. 5.1, 6.5.2

[25] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P04-3031. 12

[26] Michael J. Black. Michael j. black on twitter, November 2022. URL https://twitter.com/Michael_J_Black/status/1593133722316189696. Accessed: 2023-03-28. 2.2.2, 4.1

[27] Nicholas Bloom, Charles I Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020. (document), 1.1

[28] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://www.aclweb.org/anthology/Q17-1010. 5.2.1, A

[29] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015. 2.1, 4.1, 5.1

[30] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075. 5.2.2

[31] Virginia Braun and Victoria Clarke. *Thematic analysis.* American Psychological Association, 2012. 6.5.1

[32] Corinna Breitinger, Patrick Wortner, Bela Gipp, and Harald Reiterer. 'too late to collaborate': Challenges to the discovery of in-progress research. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 134–137. IEEE, 2019. 2.1

[33] Christine Susan Bruce. Research students' early experiences of the dissertation literature review. *Studies in Higher Education*, 19(2):217–229, 1994. 3.1

[34] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. 5.6.3

[35] Vannevar Bush et al. As we may think. *The atlantic monthly*, 176(1):101–108, 1945. 8

[36] O. Bánki, Y. Roskov, M. Döring, G. Ower, D. R. Hernández Robles, C. A. Plata Corredor, T. Stjernegaard Jeppesen, A. Örn, L. Vandepitte, D. Hobern, P. Schalk, R. E. DeWalt, K. Ma, J. Miller, T. Orrell, R. Aalbu, J. Abbott, R. Adlard, E. M. Adriaenssens, and et al. Catalogue of life checklist. *Catalogue of Life*, September 2023. doi: https://doi.org/10.48580/ddz4x. URL https://doi.org/10.48580/ddz4x. 6.4.2, C

[37] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.428. URL https://aclanthology.org/2020.findings-emnlp.428. 3.3.1

[38] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019. 7.3

[39] Jaime G Carbonell. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine learning*, pages 137–161. Springer, 1983. 2.3, 5.1, 6.2.1

[40] Jaime Guillermo Carbonell. Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1985. 2.3, 5.1, 6.2.1

[41] E Moulton Carol-anne, Glenn Regehr, Maria Mylopoulos, and Helen M MacRae. Slowing down

when you should: a new model of expert judgment. *Academic Medicine*, 82(10):S109–S116, 2007. 5.7.1

[42] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)*, 44(1):1–50, 2012. 5.7.2

[43] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://www.aclweb.org/anthology/S17-2001`. 5.2.2

[44] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018. (document), 5.2, 5.2.2, A

[45] Amaresh Chakrabarti, Prabir Sarkar, B Leelavathamma, and BS Nataraju. A functional representation for aiding biomimetic and artificial inspiration of new ideas. *Ai Edam*, 19(2):113–132, 2005. 2.3, 6.1

[46] Joel Chan and Christian D. Schunn. The importance of iteration in creative conceptual combination. *Cognition*, 145:104–115, December 2015. ISSN 0010-0277. doi: 10.1016/j.cognition.2015.08.008. URL `http://www.sciencedirect.com/science/article/pii/S0010027715300524`. 5.1, 6.5.2

[47] Joel Chan, Steven P. Dow, and Christian D. Schunn. Do The Best Design Ideas (Really) Come From Conceptually Distant Sources Of Inspiration? *Design Studies*, 36:31–58, 2015. doi: 10.1016/j.destud.2014.08.001. 5.1

[48] Joel Chan, Pao Siangliulue, Denisa Qori McDonald, Ruixue Liu, Reza Moradinezhad, Safa Aman, Erin T Solovey, Krzysztof Z Gajos, and Steven P Dow. Semantically far inspirations considered harmful? accounting for cognitive states in collaborative ideation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, pages 93–105, 2017. 5.7.2

[49] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. Solvent: A mixed initiative system for finding analogies between research papers. *Proc. ACM Hum.-Comput. Interact.*, 2 (CSCW), November 2018. doi: 10.1145/3274300. URL `https://doi.org/10.1145/3274300`. (document), 5.1, 5.2.1, 5.6, 5.7.2, A, 6.2.1

[50] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. Mesh: Scaffolding comparison tables for online decision making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 391–405, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375146. doi: 10.1145/3379337.3415865. URL `https://doi.org/10.1145/3379337.3415865`. 2.1.1

[51] Joseph Chee Chang, Nathan Hahn, Yongsung Kim, Julina Coupland, Bradley Breneisen, Hannah S Kim, John Hwong, and Aniket Kittur. When the tab comes due:challenges in the cost structure of browser tab usage. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. 3.1

[52] Joseph Chee Chang, Yongsung Kim, Victor Miller, Michael Xieyang Liu, Brad A Myers, and Aniket Kittur. Tabs. do: Task-centric browser tab management. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 663–676, 2021. 3.1

[53] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. Apolo: Interactive

large graph sensemaking by combining machine learning and visualization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 739–742, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020524. URL `https://doi.org/10.1145/2020408.2020524`. 2.1.3, 2.2.1, 3.1, 4.1, 4.3.1, A.1

[54] Hyunmin Cheong and L. H. Shu. Retrieving causally related functions from natural-language text for biomimetic design. *Journal of Mechanical Design*, 136(8):081008–081008–10, Jun 2014. ISSN 1050-0472. doi: 10.1115/1.4027494. URL `http://dx.doi.org/10.1115/1.4027494`. 2.3.1, 6.2.2, 6.4.2

[55] Kiroong Choe, Seokweon Jung, Seokhyeon Park, Hwajung Hong, and Jinwook Seo. Papers101: Supporting the discovery process in the literature review workflow for novice researchers. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pages 176–180. IEEE, 2021. 2.1.3, 3.1

[56] Johan SG Chu and James A Evans. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118, 2021. 3.6.3

[57] John Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.34. URL `https://aclanthology.org/2023.acl-long.34`. 6.2.2

[58] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.207. URL `https://aclanthology.org/2020.acl-main.207`. 3.3.1

[59] Patrick Collison and Michael Nielsen. Science is getting less bang for its buck. *The Atlantic*, 16, 2018. 1.1

[60] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. *Flow: The psychology of optimal experience*, volume 1990. Harper & Row New York, 1990. 3.4, 5.3.3

[61] Olivier Darrigol. The analogy between light and sound in the history of optics from the ancient greeks to isaac newton. part 1. *Centaurus*, 52(2):117–155, 2010. 2.3, 6.1, 6.2.1

[62] Dimitrios Darzentas, Raphael Velt, Richard Wetzel, Peter J Craigon, Hanne G Wagner, Lachlan D Urquhart, and Steve Benford. Card mapper: Enabling data-driven reflections on ideation cards. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019. 5.6.3

[63] Antonina Dattolo and Marco Corbatto. Assisting researchers in bibliographic tasks: A new usable, real-time tool for analyzing bibliographies. *Journal of the Association for Information Science and Technology*, 2021. 3.1

[64] Nicola Davis. Nasa needs you: space agency to crowdsource origami designs for shield, 2017. URL `https://www.theguardian.com/science/2017/jul/20/nasa-needs-you-space-agency-to-crowdsource-origami-designs-for-shield`. 5.1, 5.3.1

[65] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965. ISSN 0036-8075. doi: 10.1126/science.149.3683.510. URL `https://science.sciencemag.org/`

content/149/3683/510. 5.1

[66] Jon-Michael Deldin and Megan Schuknecht. The asknature database: enabling solutions in biomimetic design. In *Biologically inspired design: computational methods and tools*, pages 17–27. Springer, 2013. 2.3, 2.3.1, 6.1, 6.2.2

[67] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423. 5.7.2

[68] Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C Neubauer. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1):35, 2015. 5.7.2

[69] David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-Dickstein, et al. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022. 27

[70] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. Wizard of oz support throughout an iterative design process. *IEEE Pervasive Computing*, 4(4):18–26, 2005. 5.2.1

[71] K. N. Dunbar. How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, and J. Vaid, editors, *Creative thought: An investigation of conceptual structures and processes*, pages 461–493. Washington D.C., 1997. 5.1

[72] Chris Eliasmith and Paul Thagard. Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, 25(2):245–286, 2001. 2.3, 5.1

[73] Hen Emuna, Nadav Borenstein, Xin Qian, Hyeonsu Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. Imitation of life: A search engine for biologically inspired design. *arXiv preprint arXiv:2312.12681*, 2023. 6.1

[74] Hen Emuna, Nadav Borenstein, Xin Qian, Hyeonsu Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. Imitation of life: A search engine for biologically inspired design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 503–511, 2024. 2.3.1, 6.2.2, 6.4.2, D.2

[75] Encyclopedia. Encyclopedia of life (eol). https://eol.org/docs/what-is-eol. Accessed: 10-02-2023. 6.4.2, C

[76] Maryam Fazel-Zarandi and Eric Yu. Ontology-based expertise finding. In Takahira Yamaguchi, editor, *Practical Aspects of Knowledge Management*, pages 232–243, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-89447-6. 5.1

[77] Raymond Fok, Andrew Head, Jonathan Bragg, Kyle Lo, Marti A. Hearst, and Daniel S. Weld. Scim: Intelligent faceted highlights for interactive, multi-pass skimming of scientific papers, 2022. URL https://arxiv.org/abs/2205.04561. 2.1.2

[78] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. *arXiv preprint arXiv:2310.07581*, 2023. 6.2.3

[79] Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. A description of think aloud method

and protocol analysis. *Qualitative health research*, 3(4):430–441, 1993. 5.1

[80] Kenneth Forbus. *Exploring analogy in the large*. MIT Press, 2001. 2.3, 5.1, 6.2.1

[81] Kenneth D Forbus, Ronald W Ferguson, and Dedre Gentner. Incremental structure-mapping. In *Proceedings of the sixteenth annual conference of the Cognitive Science Society*, pages 313–318, 1994. 2.3, 5.1, 6.2.1

[82] Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, 41(5):1152–1201, 2017. 2.3, 5.1

[83] Katherine Fu, Joel Chan, Jonathan Cagan, Kenneth Kotovsky, Christian Schunn, and Kristin Wood. The meaning of "near" and "far": the impact of structuring design databases and the effect of distance of analogy on design output. *Journal of Mechanical Design*, 135(2):021007, 2013. 5.1

[84] Katherine Fu, Joel Chan, Christian Schunn, Jonathan Cagan, and Kenneth Kotovsky. Expert representation of design repository space: A comparison to and validation of algorithmic output. *Design Studies*, 34(6):729 – 762, 2013. ISSN 0142-694X. doi: https://doi.org/10.1016/j.destud.2013.06.002. URL http://www.sciencedirect.com/science/article/pii/S0142694X13000495. 5.1

[85] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2501. URL https://www.aclweb.org/anthology/W18-2501. (document), 5.2.1, 5.6

[86] D. Gentner, S. Brem, R. W. Ferguson, P. Wolff, A. B. Markman, and K. D. Forbus. Analogy and Creativity in the Works of Johannes Kepler. In T. B. Ward, J. Vaid, and S. M. Smith, editors, *Creative thought: An investigation of conceptual structures and processes*, pages 403–459. American Psychological Association, Washington D.C., 1997. 5.1

[87] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170, 1983. 2.3, 5.1, 6.1, 6.2.1

[88] Dedre Gentner and Russell Landers. Analogical reminding: A good match is hard to find. In *Unknown Host Publication Title*, pages 607–613. IEEE, December 1985. 5.1

[89] Dedre Gentner and Russell Landers. Analogical reminding: A good match is hard to find. In *Unknown Host Publication Title*, pages 607–613. IEEE, 1985. 2.3, 6.1, 6.2.1, 6.5.2

[90] Dedre Gentner and Linsey Smith. Analogical reasoning. *Encyclopedia of human behavior*, 2: 130–136, 2012. 5.6.2

[91] Dedre Gentner, Mary Jo Rattermann, and Kenneth D Forbus. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive psychology*, 25(4):524–575, 1993. 6.2.1

[92] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019, 2022. 6.4.4, 6.5.2

[93] Mary L Gick and Keith J Holyoak. Analogical problem solving. *Cognitive psychology*, 12(3): 306–355, 1980. 2.3, 5.1, 6.1

[94] Mary L Gick and Keith J Holyoak. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38, 1983. 6.2.1, 6.5.2

[95] Mary L. Gick and Keith J. Holyoak. Schema induction and analogical transfer. *Cognitive Psychology*, 15(1):1 – 38, 1983. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(83)90002-6. URL `http://www.sciencedirect.com/science/article/pii/0010028583900026`. 5.1

[96] Karni Gilon, Joel Chan, Felicia Y. Ng, Hila Liifshitz-Assaf, Aniket Kittur, and Dafna Shahaf. Analogy mining for specific design needs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 121:1–121:11, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173695. URL `http://doi.acm.org/10.1145/3173574.3173695`. 5.1

[97] Ashok K Goel, Swaroop Vattam, Bryan Wiltgen, and Michael Helms. Cognitive, collaborative, conceptual and creative—four characteristics of the next generation of knowledge-based cad systems: A study in biologically inspired design. *Computer-Aided Design*, 44(10):879–900, 2012. 2.3, 2.3.1, 6.1, 6.2.1, 6.2.2

[98] Sebastian Gomes, Miriam Boon, and Orland Hoeber. A study of cross-session cross-device search within an academic digital library. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 384–394, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531929. URL `https://doi.org/10.1145/3477495.3531929`. 3.2

[99] Milene Gonçalves, Carlos Cardoso, and Petra Badke-Schaub. Inspiration peak: exploring the semantic distance between design problem and textual inspirational stimuli. *International Journal of Design Creativity and Innovation*, 1(4):215–232, 2013. 5.1

[100] Miriam Greis, Emre Avci, Albrecht Schmidt, and Tonja Machulla. Increasing users' confidence in uncertain data by aggregating data from multiple sources. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 828–840, 2017. 3.1

[101] Howard E. Gruber and Paul H. Barrett. *Darwin on man: A psychological study of scientific creativity*. Darwin on man: A psychological study of scientific creativity. E. P. Dutton, New York, NY, England, 1974. ISBN 978-0-525-08877-6. Pages: xxv, 495. 5.1

[102] Yunqing Gu, Lingzhi Yu, Jiegang Mou, Denghao Wu, Peijian Zhou, and Maosen Xu. Mechanical properties and application analysis of spider silk bionic material. *e-Polymers*, 20(1):443–457, 2020. A.5

[103] Graeme S Halford. Analogical reasoning and conceptual complexity in cognitive development. *Human Development*, 35(4):193–217, 1992. 5.7.2

[104] Graeme S Halford, William H Wilson, and Steven Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences*, 21(6):803–831, 1998. 5.7.2

[105] Graeme S Halford, Rosemary Baker, Julie E McCredden, and John D Bain. How many variables can humans process? *Psychological science*, 16(1):70–76, 2005. 6.2.1

[106] Han L Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E Mackay, and Michel Beaudouin-Lafon. Passages: Interacting with text across documents. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022. 2.1.1, 2.1.3

[107] Ning Hao, Yixuan Ku, Meigui Liu, Yi Hu, Mark Bodner, Roland H Grabner, and Andreas Fink. Reflection enhances creativity: Beneficial effects of idea evaluation on idea generation. *Brain and cognition*, 103:30–37, 2016. 5.6.3

[108] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of

empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988. 3.4, **??**, 4.4.4, **??**

[109] Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*, 2023. 6.4.4, 6.5.2

[110] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021. 2.1.2, 4.1

[111] Andrew Head, Amber Xie, and Marti A Hearst. Math augmentation: How authors enhance the readability of formulas using novel visual design practices. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022. 2.1.2

[112] M. Hesse. *Models and analogies in science*. Notre Dame, IN, 1966. 5.1

[113] Mary B Hesse. Models and analogies in science. 1966. 5.1

[114] Terje Hillesund. Digital reading spaces: How expert readers handle books, the web and electronic paper. 2010. 2.1.2, 4.1

[115] Ken Hinckley, Xiaojun Bi, Michel Pahud, and Bill Buxton. Informal information gathering techniques for active reading. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1893–1896, 2012. 3.2

[116] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. 5.2.1

[117] Douglas R Hofstadter, Melanie Mitchell, et al. The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory*, 2:205–267, 1995. 2.3, 5.1, 6.2.1

[118] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020. 5.7.2

[119] K. J. Holyoak and P. Thagard. The analogical scientist. In K. J. Holyoak and P. Thagard, editors, *Mental Leaps: Analogy in Creative Thought*, pages 185–209. Cambridge, MA, 1996. 5.1

[120] Keith J Holyoak and Paul Thagard. Analogical mapping by constraint satisfaction. *Cognitive science*, 13(3):295–355, 1989. 2.3, 5.1, 6.2.1

[121] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 235–243, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098038. URL `http://doi.acm.org/10.1145/3097983.3098038`. 5.1, 5.2.2, 5.7.2, 6.1, 6.2.1

[122] Tom Hope, Ronen Tamari, Hyeonsu Kang, Daniel Hershcovich, Joel Chan, Aniket Kittur, and Dafna Shahaf. Scaling creative inspiration with fine-grained functional facets of product ideas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3491102.3517434. URL `https://doi.org/10.1145/3491102.3517434`. 3.6.2, 5.1, 5.2.1, 5.2.2, 5.7.2

[123] Amber Horvath, Brad Myers, Andrew Macvean, and Imtiaz Rahman. Using annotations for sense-

making about code. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393201. doi: 10.1145/3526113.3545667. URL `https://doi.org/10.1145/3526113.3545667`. 4.4.1

[124] Hen-Hsen Huang and Hsin-Hsi Chen. Disa: A scientific writing advisor with deep information structure analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 5229–5231, 2017. doi: 10.24963/ijcai.2017/773. URL `https://doi.org/10.24963/ijcai.2017/773`. 5.2.1

[125] Ting-Hao'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C Lee Giles. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. *arXiv preprint arXiv:2005.02367*, 2020. 6.1

[126] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. 5.2.1

[127] John E Hummel and Keith J Holyoak. A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, 110(2):220, 2003. 2.3, 5.1, 6.2.1

[128] David G Jansson and Steven M Smith. Design fixation. *Design studies*, 12(1):3–11, 1991. 5.5.2

[129] Shuo Jiang, Jie Hu, Kristin L Wood, and Jianxi Luo. Data-driven design-by-analogy: state-of-the-art and future directions. *Journal of Mechanical Design*, 144(2):020801, 2022. 2.3, 2.3.1, 6.1, 6.2.2

[130] Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. Generalizing natural language analysis through span-relation representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2120–2133, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.192. URL `https://www.aclweb.org/anthology/2020.acl-main.192`. 5.2.1, 5.4.2, A

[131] Arif E Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010. 2.1, 4.1, 5.1

[132] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. A.5

[133] Philip N Johnson-Laird. Flying bicycles: How the wright brothers invented the airplane. *Mind & Society*, 4:27–48, 2005. 2.3, 6.1, 6.2.1

[134] Benjamin F Jones. Age and great invention. *The Review of Economics and Statistics*, 92(1):1–14, 2010. 1.2

[135] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011. 5.7.1

[136] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. Threddy: An interactive system for personalized thread-based exploration and organization of scientific literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393201. doi: 10.1145/3526113.3545660. URL `https://doi.org/10.1145/3526113.3545660`. (document), 3, 4.1, 4.2, 4.3.1, 4.4.3, 4.6, 4.4.4, 4.6.1, 4.7

[137] Hyeonsu B Kang, Gabriel Amoako, Neil Sengupta, and Steven P Dow. Paragon: An online gallery for enhancing design feedback with visual examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018. 3.4

[138] Hyeonsu B Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S Weld, Doug Downey, and Jonathan Bragg. From who you know to what you read: Augmenting scientific recommendations with implicit social networks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517470. URL `https://doi.org/10.1145/3491102.3517470`. 3.6.2, A.1

[139] Hyeonsu B Kang, Sheshera Mysore, Kevin J Huang, Haw-Shiuan Chang, Thorben Prein, Andrew McCallum, Aniket Kittur, and Elsa Olivetti. Augmenting scientific creativity with retrieval across knowledge domains. In *Second Workshop on Bridging Human-Computer Interaction and Natural Language Processing at NAACL 2022*. arXiv, 2022. doi: 10.48550/ARXIV.2206.01328. URL `https://arxiv.org/abs/2206.01328`. 3.6.2

[140] Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. Augmenting scientific creativity with an analogical search engine. *ACM Trans. Comput.-Hum. Interact.*, mar 2022. ISSN 1073-0516. doi: 10.1145/3530013. URL `https://doi.org/10.1145/3530013`. Just Accepted. 2.3, 3.6.2, 5, 6.1, 6.2.1

[141] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. Comlittee: Literature discovery with personal elected author committees. *arXiv preprint arXiv:2302.06780*, 2023. A.1

[142] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. Synergi: A mixed-initiative system for scholarly synthesis and sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606759. URL `https://doi.org/10.1145/3586183.3606759`. 2.3.2, 4, 6.2.3, A.5

[143] Hyeonsu B Kang, David Chuan-En Lin, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K. Hong. Biospark: An end-to-end generative system for biological-analogical inspirations and ideation. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3651035. URL `https://doi.org/10.1145/3613905.3651035`. 6

[144] Mark T Keane, Tim Ledgeway, and Stuart Duff. Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, 18(3):387–438, 1994. 5.6.2

[145] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, volume 37, pages 18–28. ACM New York, NY, USA, 2003. 5.6.2

[146] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. Facilitating document reading by linking text and tables. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018. 2.1.2

[147] Jingoog Kim and Mary Lou Maher. The effect of ai-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation*, 11(2):81–98, 2023. 2.3.2, 6.2.3

[148] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. A

[149] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023. 4.3.1

[150] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, Trupti Telang, and Michael R. Bove. Costs

and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2989–2998, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481415. URL `https://doi.org/10.1145/2470654.2481415`. 4.6.3

[151] Aniket Kittur, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E Kraut, and Dafna Shahaf. Scaling up analogical innovation with crowds and ai. *Proceedings of the National Academy of Sciences*, 116(6):1870–1877, 2019. 5.1

[152] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. A data–frame theory of sensemaking. In *Expertise out of context*, pages 118–160. Psychology Press, 2007. (document), 2, 7.1, 1, 7.1

[153] Madeline K Kneeland, Melissa A Schilling, and Barak S Aharonson. Exploring uncharted territory: Knowledge search processes in the origination of outlier innovation. *Organization Science*, 31(3): 535–557, 2020. 5.1, 6.5.2

[154] Jeffrey W Knopf. Doing a literature review. *PS: Political Science & Politics*, 39(1):127–132, 2006. 4.1

[155] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E MacKay. Imagesense: An intelligent collaborative ideation tool to support diverse human-computer partnerships. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 2020. 5.6.3

[156] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. Extracting references between text and charts via crowdsourcing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014. 2.1.2

[157] Wouter Kool and Matthew Botvinick. Mental labour. *Nature human behaviour*, 2(12):899–908, 2018. 4.6.3

[158] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. Fuse: In-situ sensemaking support in the browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393201. doi: 10.1145/3526113.3545693. URL `https://doi.org/10.1145/3526113.3545693`. 2.1.1

[159] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety*, 25(10): 808–820, 2016. 5.7.1

[160] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997. 5.2.1, A

[161] Esther Landhuis. Scientific literature: Information overload. *Nature*, 535(7612):457–458, 2016. 2.1

[162] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL `https://www.aclweb.org/anthology/D17-1018`. 5.2.1, A

[163] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and

Pao Siangliulue. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642196. URL https://doi.org/10.1145/3613904.3642196. 7.1

[164] Clayton Lewis. *Using the" thinking-aloud" method in cognitive interface design.* IBM TJ Watson Research Center Yorktown Heights, NY, 1982. 5.3.2

[165] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020. 2.2.2

[166] Julie S Linsey, Emily F Clauss, Tolga Kurtoglu, Jeremy T Murphy, Kristin L Wood, and Arthur B Markman. An experimental study of group idea generation techniques: understanding the roles of idea representation and viewing methods. 2011. 6.3.1

[167] Julie S Linsey, Arthur B Markman, and Kristin Lee Wood. Design by analogy: A study of the wordtree method for problem re-representation. 2012. 6.1, 6.2.1, 6.4.2

[168] Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. Crystalline: Lowering the cost for developers to collect and organize information for decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501968. URL https://doi.org/10.1145/3491102.3501968. 2.1.1

[169] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. Selenite: Scaffolding decision making with comprehensive overviews elicited from large language models. *arXiv preprint arXiv:2310.02161*, 2023. 6.2.3, A.5

[170] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*, 2020. URL https://arxiv.org/abs/1911.02782. 3.3.1, 4.3.1, 4.6.5

[171] Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943. 2.3, 5.1

[172] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008. ISBN 0521865719. 5.3.2

[173] James G March. Exploration and exploitation in organizational learning. *Organization science*, 2 (1):71–87, 1991. 6.4.4

[174] Abraham Maslow. Self-actualization and beyond. 1965. 4.6.4

[175] Lori McCay-Peet and Elaine Toms. Measuring the dimensions of serendipity in digital environments. *Information Research: An International Electronic Journal*, 16(3):n3, 2011. 6.5.1

[176] David W. McDonald and Mark S. Ackerman. Expertise recommender: A flexible recommendation system and architecture. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, page 231–240, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132220. doi: 10.1145/358916.358994. URL https://doi.org/10.1145/358916.358994. 5.1

[177] Anamika Megwalu. Academic social networking: a case study on users' information behavior. In *Current Issues in Libraries, Information Science and Related Fields*. Emerald Group Publishing Limited, 2015. 2.1

[178] Microsoft Corporation. Microsoft Copilot: Your everyday ai companion. https://copilot.microsoft.com, 2023. Accessed: 2023-04-05. 2.3.2, 6.2.3

[179] James G Miller. Information input overload and psychopathology. *American journal of psychiatry*, 116(8):695–704, 1960. 2.1

[180] Staša Milojević. Quantifying the cognitive extent of science. *Journal of Informetrics*, 9(4):962–973, 2015. 1.2

[181] Marvin Minsky et al. A framework for representing knowledge, 1974. 2

[182] Henry Mintzberg, Duru Raisinghani, and Andre Theoret. The structure of" unstructured" decision processes. *Administrative science quarterly*, pages 246–275, 1976. 5.7.1

[183] Meg Monk. BYU engineers use origami to make more space in space. https://universe.byu.edu/2013/12/12/byu-engineers-use-origami-to-make-more-space-in-space/, December 2013. Accessed: YYYY-MM-DD. 2.3, 6.1, 6.2.1

[184] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011. 6.4.2, C

[185] Dan Morris, Meredith Ringel Morris, and Gina Venolia. Searchbar: A search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1207–1216, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi: 10.1145/1357054.1357242. URL https://doi.org/10.1145/1357054.1357242. 3.2

[186] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Effective human-ai teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems*, 36, 2024. 7.3

[187] Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, 53:1–18, 2013. 4.3.2

[188] Sheshera Mysore, Mahmood Jasim, Haoru Song, Sarah Akbar, Andre Kenneth Chase Randall, and Narges Mahyar. How data scientists review the scholarly literature. *arXiv preprint arXiv:2301.03774*, 2023. 2.1.3

[189] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 677–686, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2568012. URL https://doi.org/10.1145/2566486.2568012. 3.6.3

[190] David Nicholas, Peter Williams, Ian Rowlands, and Hamid R Jamali. Researchers'e-journal use and information seeking behaviour. *Journal of Information Science*, 36(4):494–516, 2010. 2.1.2

[191] Xi Niu, Fakhri Abbas, Mary Lou Maher, and Kazjon Grace. Surprise me if you can: Serendipity in health information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018. 6.5.1

[192] Richar Van Noorden. Global scientific output doubles every nine years, May 2014. URL http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html. 5.1

[193] Richar Van Noorden. Global scientific output doubles every nine years, May 2014. URL http://blogs.nature.com/news/2014/05/

global-scientific-output-doubles-every-nine-years.html. 2.1

[194] Keisuke Okamura. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1), 2019. 4.1

[195] OpenAI. Chatgpt - your friendly ai chatbot. https://chat.openai.com/, 2023. Accessed: 2023-04-28. 2.3.2, 6.2.3

[196] OpenAI. Gpt-4 technical report, 2023. A.1

[197] R. Oppenheimer. Analogy in science. *American Psychologist*, 11(3):127–135, 1956. ISSN 0003-066X. 5.1

[198] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. Relatedly: Scaffolding literature reviews with existing related work sections. *arXiv preprint arXiv:2302.06754*, 2023. 2.1, 2.1.3

[199] Carole L Palmer, Lauren C Teffeau, and Carrie M Pirmann. Scholarly information practices in the online environment. *Report commissioned by OCLC Research. Published online at: www. oclc. org/programs/publications/reports/2009-02. pdf*, 2009. 1.3, 2.1

[200] Michael Park, Erin Leahey, and Russell J Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144, 2023. 1.1

[201] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 5.2.1

[202] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. Exploring the effects of technological writing assistance for support providers in online mental health community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020. 6.6

[203] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP '14)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162. 5.2.1, A

[204] Edwin A Peraza-Hernandez, Darren J Hartl, Richard J Malak Jr, and Dimitris C Lagoudas. Origami-inspired active structures: a synthesis and review. *Smart Materials and Structures*, 23(9):094001, 2014. 2.3, 5.1, 5.3.1, 6.1

[205] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202. 5.2.1, 5.4.2, 5.7.2

[206] Jean Piaget. *The construction of reality in the child.* Routledge, 2013. 2

[207] Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999. 6.4.4

[208] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005. 1.3, 2, 3.6.1, 4.1

[209] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. Paperquest: A visualization tool

to support literature review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2264–2271, 2016. 2.1.3, 3.1

[210] Kristopher J. Preacher and Andrew F. Hayes. Spss and sas procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36(4):717–731, Nov 2004. ISSN 1532-5970. doi: 10.3758/BF03206553. URL https://doi.org/10.3758/BF03206553. (document), 5.3.3, 5.5

[211] Napol Rachatasumrit, Gonzalo Ramos, Jina Suh, Rachel Ng, and Christopher Meek. Forsense: Accelerating online research through sensemaking integration and machine research support. In *26th International Conference on Intelligent User Interfaces*, pages 608–618, 2021. 2.1.1

[212] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. Citeread: Integrating localized citation contexts into scientific paper reading. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 707–719, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391443. doi: 10.1145/3490099.3511162. URL https://doi.org/10.1145/3490099.3511162. 2.1.2

[213] Sekharipuram S Ravi, Daniel J Rosenkrantz, and Giri Kumar Tayi. Heuristic and special case algorithms for dispersion problems. *Operations Research*, 42(2):299–310, 1994. 5.2.2

[214] Tony Rees, Leen Vandepitte, Bart Vanhoorne, and Wim Decock. All genera of the world: an overview and estimates based on the march 2020 release of the interim register of marine and nonmarine genera (irmng). *Megataxa*, 1(2):123–140, 2020. 6.4.2, C

[215] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. (document), 6.5, 6.4.4

[216] Stephen Rowland. Overcoming fragmentation in professional life: The challenge for academic development. *Higher education quarterly*, 56(1):52–64, 2002. 2.1

[217] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4), jan 2018. 4.4.4, **??**

[218] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276. ACM, 1993. 5.6.2

[219] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574, 2015. 3.6.3

[220] Neil Savage. Nanofins make a better hologram, 2016. URL https://spectrum.ieee.org/tech-talk/semiconductors/optoelectronics/nanofins-make-a-better-hologram. (document), 5.3

[221] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press, 2013. 2

[222] Tobias Schnabel, Paul N Bennett, and Thorsten Joachims. Shaping feedback data in recommender systems with interventions based on information foraging theory. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 546–554, 2019. 5.6.2

[223] Tobias Schnabel, Gonzalo Ramos, and Saleema Amershi. "who doesn't like dinosaurs?" finding and eliciting richer preferences for recommendation. In *Fourteenth ACM Conference on Recom-*

mender Systems, pages 398–407, 2020. 5.6.2

[224] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1122–1130, 2012. 2.2.1, 4.1

[225] Amit Sharma and Dan Cosley. Do social explanations work? studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1133–1144, 2013. A.1

[226] Pao Siangliulue, Joel Chan, Krzysztof Z Gajos, and Steven P Dow. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pages 83–92, 2015. 5.7.2

[227] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. Ideahound: Improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, page 609–624, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341899. doi: 10.1145/2984511.2984578. URL `https://doi.org/10.1145/2984511.2984578`. 2.2.1

[228] L Siddharth and Amaresh Chakrabarti. Evaluating the impact of idea-inspire 4.0 on analogical transfer of concepts. *Ai Edam*, 32(4):431–448, 2018. 2.3, 6.1, 6.2.1

[229] Herbert A Simon. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, 70:187–202, 1996. 2.1

[230] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 5.6.3

[231] L. Streeter and K. Lochbaum. An expert/expert-locating system based on automatic representation of semantic structure. In *Proceedings. The Fourth Conference on Artificial Intelligence Applications*, pages 345,346,347,348,349,350, Los Alamitos, CA, USA, mar 1988. IEEE Computer Society. doi: 10.1109/CAIA.1988.196129. URL `https://doi.ieeecomputersociety.org/10.1109/CAIA.1988.196129`. 5.1

[232] Benjamin Sturm and Ali Sunyaev. Design principles for systematic search systems: a holistic synthesis of a rigorous multi-cycle design science research journey. *Business & Information Systems Engineering*, 61(1):91–111, 2019. 2.1.3

[233] Hariharan Subramonyam, Colleen Seifert, Priti Shah, and Eytan Adar. Texsketch: Active diagramming through pen-and-ink annotations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 2.1.2

[234] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642400. URL `https://doi.org/10.1145/3613904.3642400`. 7.1

[235] Nicole Sultanum, Christine Murad, and Daniel Wigdor. Understanding and supporting academic literature review workflows with litsense. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–5, 2020. 2.1.3

[236] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with

neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf. 5.1

[237] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press. 5.2.1

[238] Don R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986. doi: 10.1086/601720. URL https://doi.org/10.1086/601720. 3.6.3

[239] L. Sweetlove. Number of species on earth tagged at 8.7 million. *Nature*, August 2011. doi: https://doi.org/10.1038/news.2011.498. URL https://doi.org/10.1038/news.2011.498. C

[240] John Sweller, Paul Chandler, Paul Tierney, and Martin Cooper. Cognitive load as a factor in the structuring of technical material. *Journal of experimental psychology: general*, 119(2):176, 1990. 5.6.2, 5.7.2

[241] Craig S. Tashman and W. Keith Edwards. Liquidtext: A flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 3285–3294, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979430. URL https://doi.org/10.1145/1978942.1979430. 3.2

[242] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 2.2.2, 4.1

[243] Jaime Teevan. A formula for academic papers: Related work, November 2014. URL http://slowsearching.blogspot.com/2014/11/a-formula-for-academic-papers-related.html. 2.1

[244] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, page 415–422, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581137028. doi: 10.1145/985692.985745. URL https://doi.org/10.1145/985692.985745. 5.6.2

[245] Carol Tenopir, Donald W King, Sheri Edwards, and Lei Wu. Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings*. Emerald Group Publishing Limited, 2009. 2.1

[246] Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. *arXiv preprint arXiv:2004.02990*, 2020. 6.4.4, 6.5.2

[247] ThermoCool. Skived fin heat sinks, 2021. URL https://thermocoolcorp.com/project/skived-fins/. (document), 5.3

[248] H Holden Thorp. Chatgpt is fun, but not an author, 2023. 2.2.2, 4.1

[249] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. Mediation: R package for causal mediation analysis. 2014. 5.3.3

[250] Peter D Turney. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655, 2008. 2.3, 5.1

[251] Richard Van Noorden et al. Interdisciplinary research by the numbers. *Nature*, 525(7569):306–307, 2015. 4.1

[252] MW Van Someren, YF Barnard, and JAC Sandberg. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress*, 1994. 5.3.2

[253] MW Van Someren, YF Barnard, and JAC Sandberg. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress*, 1994. 5.1

[254] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *arXiv preprint arXiv:2212.06823*, 2022. 4.6.3

[255] Luis A Vasconcelos and Nathan Crilly. Inspiration and fixation: Questions, methods, findings, and challenges. *Design Studies*, 42:1–32, 2016. 6.6

[256] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. 5.2.1, 5.2.2

[257] Swaroop Vattam, Bryan Wiltgen, Michael Helms, Ashok K Goel, and Jeannette Yen. Dane: fostering creativity in and through biologically inspired design. In *Design Creativity 2010*, pages 115–122. Springer, 2011. 2.3, 5.1, 5.1

[258] Swaroop S. Vattam and Ashok K. Goel. Semantically Annotating Research Articles for Interdisciplinary Design. In *Proceedings of the Sixth International Conference on Knowledge Capture*, K-CAP '11, pages 165–166, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0396-5. doi: 10.1145/1999676.1999707. URL `http://doi.acm.org/10.1145/1999676.1999707`. 2.3.1, 6.2.2, 6.4.2

[259] Manuela M Veloso and Jaime G Carbonell. Derivational analogy in prodigy: Automating case acquisition, storage, and utilization. In *Case-Based Learning*, pages 55–84. Springer, 1993. 2.3, 5.1

[260] Julian FV Vincent and Darrell L Mann. Systematic technology transfer from biology to engineering. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 360(1791):159–173, 2002. 2.3, 6.1, 6.2.1

[261] Samangi Wadinambiarachchi, Ryan M Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. The effects of generative ai on design fixation and divergent thinking. *arXiv preprint arXiv:2403.11164*, 2024. 2.3.2, 6.2.3

[262] James A Waltz, Albert Lau, Sara K Grewal, and Keith J Holyoak. The role of working memory in analogical mapping. *Memory & Cognition*, 28(7):1205–1212, 2000. 5.6.2

[263] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. A

[264] Yun Wang, Dongyu Liu, Huamin Qu, Qiong Luo, and Xiaojuan Ma. A guided tour of literature review: Facilitating academic paper reading with narrative visualization. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, pages 17–24, 2016. 2.2.1

[265] Peter C Wason and J St BT Evans. Dual processes in reasoning? *Cognition*, 3(2):141–154, 1974.

5.7.1

[266] Anthony Watkinson, David Nicholas, Clare Thornley, Eti Herman, Hamid R Jamali, Rachel Volentine, Suzie Allard, Kenneth Levine, and Carol Tenopir. Changes in the digital scholarly environment and issues of trust: An exploratory, qualitative analysis. *Information processing & management*, 52 (3):446–458, 2016. 3.6.3

[267] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023. 2.3.2, 6.2.3

[268] Jane Webster, Linda Klebe Trevino, and Lisa Ryan. The dimensionality and correlates of flow in human-computer interactions. *Computers in human behavior*, 9(4):411–426, 1993. 3.4, **??**

[269] Ryen W White and Resa A Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–98, 2009. 5.6.2

[270] Ryen W White, Paul N Bennett, and Susan T Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1009–1018, 2010. 5.7.2

[271] Wikipedia. Genus. `https://en.wikipedia.org/wiki/Genus#cite_note-10`. Accessed: 10-02-2023. 6.4.2, C

[272] Rand R Wilcox and HJ Keselman. Modern robust data analysis methods: measures of central tendency. *Psychological methods*, 8(3):254, 2003. 5.2.1

[273] Jen-Her Wu and Shu-Ching Wang. What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & management*, 42(5):719–729, 2005. 3.4, **??**, 4.4.4, **??**, 6.5.1

[274] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. *Understanding Belief Propagation and Its Generalizations*, page 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003. ISBN 1558608117. 3.1, 4.3.1

[275] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. How experienced designers of enterprise applications engage ai as a design material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517491. URL `https://doi.org/10.1145/3491102.3517491`. 4.4.1

[276] Michael Yovanovich, Richard Culham, and Peter Teertstra. Calculating interface resistance, 2004. URL `http://www.thermalengineer.com/library/calculating_interface_resistance.htm`. (document), 5.3

[277] Amy Xian Zhang. *Systems for collective human curation of online discussion*. PhD thesis, Massachusetts Institute of Technology, 2019. 7.1

[278] Xiaolong Zhang, Yan Qu, C. Lee Giles, and Piyou Song. Citesense: Supporting sensemaking of research literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 677–680, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi: 10.1145/1357054.1357161. URL `https://doi.org/10.1145/1357054.1357161`. 2.1, 2.1.3, 3.1, 4.1, 5.6.2

[279] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. 2.2.2

[280] Yuanshuo Zhao, Ioana Baldini, Prasanna Sattigeri, Inkit Padhi, Yoong Keok Lee, and Ethan Smith. Data driven techniques for organizing scientific articles relevant to biomimicry. In *ACM/AAAI Artificial Intelligence, Ethics and Society (AIES) conference*, 2018. 2.3.1, 6.2.2, 6.4.2

[281] Shannon A Zirbel, Mary E Wilson, Spencer P Magleby, and Larry L Howell. An origami-inspired self-deployable array. In *ASME 2013 Conference on Smart Materials, Adaptive Structures and Intelligent Systems*. American Society of Mechanical Engineers Digital Collection, 2013. 2.3, 5.1, 5.3.1, 6.1

[282] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. Successful classroom deployment of a social document annotation system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1883–1892, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi: 10.1145/2207676.2208326. URL `https://doi.org/10.1145/2207676.2208326`. 2.1.2