Towards Unified Interfaces for Generalist Agent In Diverse Environments

Yueqi Song

CMU-CS-25-120

August 2025

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Graham Neubig, Advisor
Daniel Fried

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science.

Keywords: Agent, Reasoning, Large Language Model

For everyone who kindly offered me support and encouragement.

Abstract

Recently, large language models (LLMs) have enabled agents that can perceive, reason, and act in increasingly complex environments. Yet today's agents remain constrained by the interfaces they rely on, hampering generalization. This master thesis advances the goal of a *unified agent framework*.

Examining web agents, we found that web browsing agents, though intuitive to humans as they simulate human behaviours by browsing the web, are less effective and efficient. Thus, we proposed an API-based web agent that calls APIs through code generation, and demonstrated superior performance compared to browsing agents. Building on this, we further proposed a hybrid web agent that could interleave API calling and web browsing, broadening the agent's interface and allowing it to operate more effectively and efficiently in diverse environments. Beyond web agents, we aim to extend the unified interfaces to generalist agents across diverse environments as a future work.

Alongside a unified framework, strong reasoning abilities are crucial for agents to make correct decisions, plan, and execute tasks based on users' goals. We thus introduced VisualPuzzles, a benchmark that could evaluate models' multimodal reasoning abilities in a knowledge-light environment, which could provide guidance on the future development of models with strong multimodal reasoning capabilities.

Last but not the least, to serve people around the world, agents need to understand and generate multilingual contents. Thus, we proposed and trained Pangea, a multilingual model that achieved SOTA results on multilingual benchmarks.

Together, these contributions pave a path *towards unified interfaces for generalist agent in diverse environments*, providing the conceptual, empirical, and engineering foundations for the next generation of generalist AI agents.

Acknowledgments

I am very grateful to my advisor Professor Graham Neubig whose insights, guidance, and encouragement helped me shape my research. Many thanks to my thesis committee Professor Daniel Fried whose incisive feedback and patience as a committee member helped me refine my ideas and broaden my perspectives. Thanks Professor Xiang Yue for insightful discussions and generous guidance. Thanks all my friends and family whose unwavering support encouraged me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans use natural language to convey knowledge, coordinate collective collaboration, and shape cultures. Today, advances in Natural Language Processing (NLP) have enabled Large Language Models (LLMs) that extend human-human communications to human-AI communications. LLM-powered AI agents can perceive the world, reason about the environment, act in complex settings, and perform tasks as instructed by humans. We can imagine creating slides, booking travel (or "trips"), and implementing software, all through language agents.

However, current agents remain fundamentally constrained by the interfaces they rely on. Agents have different interfaces when interacting with different environments. For instance, browsing agents interact with the web through graphical user interfaces (GUIs), while coding agents interact with integrated development environment (IDE) command palettes. These agents usually have limited action spaces that only support them in performing well within specific environments. Moreover, agents that are trained on one interface learn specialized actions and state representations that may not generalize to other interfaces. Therefore, this master's thesis advances the goal of a *unified agent framework*.

We begin our exploration by examining web agents that can perform web tasks such as shopping online, generating itineraries, and posting on forums. Existing browsing agents interact with browsers by simulating human behavior — browsing the web, clicking buttons, and typing text. Although intuitive to humans, the GUI action space is often ineffective and inefficient for agents. Thus, we proposed an API-based agent that calls Application Programming Interface (API) endpoints through code generation, thereby enlarging the action space of web agents (Chapter 3). We found that allowing agents to interact with APIs greatly improved the performance of web agents empirically, suggesting that agents benefit from interfaces designed specifically for machines, not just human usability — or at least complementary to traditional GUIs. Building on this, we further proposed a hybrid web agent that could interleave both web browsing and API calling, demonstrating superior performance to pure browsing agents and API-based agents. The hybrid agent further broadens agents' interfaces and allows them to operate more effectively and efficiently in diverse environments. For the future work of this project, we aim to extend the unified interfaces beyond web agents to generalist agents across diverse environments.

Alongside a unified interface, robust reasoning ability is also critical for agents to make decisions, plan, and act reliably toward users' goals. To measure and improve this capability, we introduce VisualPuzzles, a benchmark that evaluates models' *multimodal reasoning* in

a knowledge-light setting, decoupling reasoning from domain knowledge recall (Chapter 4). Current models remain below human performance; even *thinking* models do not consistently outperform their base counterparts.

Finally, to serve people globally, agents must have the ability to understand and generate across languages and cultures. We thus present Pangea, a fully open multilingual multimodal LLM covering 39 languages (Chapter 5). Built with PangeaIns, our newly curated corpus of 6M multilingual multimodal instructions, Pangea achieves state-of-the-art results on multilingual benchmarks and is fully open to facilitate further multilingual research.

Chapter 2

Background

2.1 Web Agents

2.1.1 The Web Task

Various benchmarks have been developed to evaluate web browsing agents. MiniWoB (Miniature World of Bits) is an early benchmark that provides simple web-based tasks such as clicking links or typing into forms, but it remains limited in complexity and realism [120]. Mind2Web scales up these tasks, introducing more complex interactions across websites, but it primarily focuses on basic web operations [31]. WebArena [164] advances web browsing benchmarks by creating reproducible sandboxes of various websites, such as managing repositories, posting online, performing online shopping, and planning trips using map services, while VisualWebArena extends WebArena to the vision modality [62].

2.1.2 Existing Web Browsing Agent

A wide variety of agents have been proposed for such web navigation tasks, many of which are built on GUIs. In this work, we build upon a baseline browsing agent [164], which operates purely through web interaction by leveraging the accessibility tree[1], a structure that exposes interactive elements such as buttons, input fields, and hyperlinks [44, 146]. Each element of the accessibility tree is characterized by its functionality (e.g., hyperlink), its content, and specific web attributes [49, 85, 90]. This representation exposes webpage elements in a hierarchical structure that is easy for agents to navigate [18, 115].

Agents based on this framework utilize an action space that simulates human browsing behavior, incorporating actions such as simulated clicks, form inputs, and navigation between pages [46, 84, 123]. Importantly, these agents maintain a comprehensive history of their previous actions, allowing them to contextualize their decision-making based on past actions.

While agents utilizing this method can navigate arbitrary webpages and often perform well on simple layouts, challenges arise with the complexity of GUIs. Many large language models (LLMs) are not familiar with accessibility trees, which leads to difficulties in completing tasks

[1] https://developer.mozilla.org/en-US/docs/Glossary/Accessibility_tree

Figure 2.1: The API-Based Agent often solves problems in fewer steps than the Browsing Agent. In this task, web browsing failed to solve the intent "find the number of commits the user *SaptakS* made to the repo *a11yproject*" after 15 steps, while the API-Based Agent successfully completed the task with only three lines of code.

that require numerous or complex interactions, resulting in lower accuracy [31, 38, 83]. These methods also struggle with content that needs to be dynamically loaded or content not immediately visible within the tree [2, 24, 89].

To give a motivating example, in Figure 2.1, we demonstrate a task where agents need to determine the number of commits made by the user *SaptakS* in a repository named *a11yproject*. For each task, agents are given a fixed number of steps within which to complete the task. Using a traditional browsing approach, the agent follows a complex trajectory, starting with logging in, navigating to the correct project, accessing the repository, and finally attempting to view the list of commits. However, due to the large number of commits made by other users, the commits by *SaptakS* are located much further down on the webpage, requiring the agent to scroll many times. As a result, despite completing 15 steps, the browsing agent is unable to retrieve the required information.

2.2 Where Are Agents Beyond the Web

Beyond web environments, agents increasingly operate across:

- **Software engineering**: code editing, building, testing, and other systems via editors, shells, and APIs.
- **Data and cloud services**: databases, storage, and model serving exposed through MCP / API interfaces.
- **Productivity ecosystems**: email, calendars, spreadsheets, and knowledge bases (GUIs, APIs, and other app-specific actions).
- **Embodied settings**: simulated and real robots, where actions target controllers and sensors

provide multimodal streams.

- **Multilingual settings**: operating across languages and cultures to serve people around the world.

These domains expose heterogeneous interfaces, reinforcing the need for a *unified, language-centric interface*. The subsequent chapters instantiate this idea on the web (via APIs and hybrid control) and extend it to broader agentic tasks.

Chapter 3

Beyond Browsing: API-Based Web Agents

Existing agents often have limited actions spaces that only support them to perform well within specific environments. To achieve broad generalization. To unlock the full potential of autonomous agents, we must equip them with broad action spaces that allow them to execute effectively and efficiently. In this chapter, we introduce our work on expanding the web-agent action space from interactions with GUIs to API calls that are specifically designed for machines to interact with the web, and on designing a hybrid agent that can interleave both browsing and API calling when needed. This project broadens the action space of web agents, marking a step towards the unified interface for generalist agent. This project first appears in:

- Yueqi Song, Frank F. Xu, Shuyan Zhou, and Graham Neubig. 2025. Beyond Browsing: API-Based Web Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11066–11085, Vienna, Austria. Association for Computational Linguistics.

3.1 Overview

Web agents use browsers as an interface to facilitate humans in performing daily tasks such as online shopping, online planning, trip planning, and other work-related tasks [22, 37, 54, 73, 76, 98, 99, 111]. Existing web agents typically operate within the space of graphical user interfaces (GUI) [157, 160, 164], using action spaces that simulate human-like keyboard and mouse operations, such as clicking and typing. To observe webpages, common approaches include using accessibility trees, a simplified version of the HTML DOM tree, as input to text-based models [33, 164], or multi-modal, screenshot-based models [53, 62, 140, 149]. However, regardless of the interaction method with websites, there is no getting around the fact that these sites were originally designed for humans, and may not be the ideal interface for machines.

Notably, there is another interface designed specifically for machine interaction with the web: application programming interfaces (APIs) [19]. APIs allow machines to communicate directly with backends of web services [15], sending and receiving data in machine-friendly formats such as JSON or XML [93, 132, 141]. Nonetheless, whether AI agents can effectively use APIs to tackle real-world online tasks, and the conditions under which this is possible, remain unstudied. In this work, we explore methods for tackling tasks normally framed as web-navigation tasks

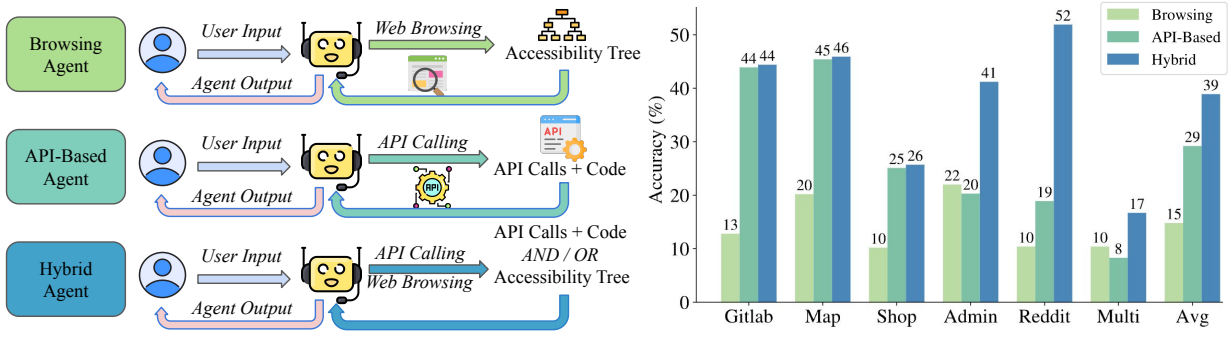


Figure 3.1: The Browsing Agent performs tasks through browsing only, utilizing the accessibility tree to interact with webpages, achieving an average performance of 14.8% on WebArena. Without reliance on web browsing, the API-Based Agent performs tasks by making API calls and generating code without relying on web browsing, achieving an average accuracy of 29.2%. Dynamically interleaving web browsing and API calling, the Hybrid Agent executes either API calls or browsing actions, or combining both, achieving performance of 38.9%.

with an expanded action space to interact with APIs. To do so, we develop new *API-Based Agents* that directly interact with web services via API calls. This method bypasses the need to interact with web GUIs.

However, not all websites have extensive API support, in which case web browsing actions may still be required. To overcome this limitation, we propose a *hybrid* approach that combines API-Based Agents with Browsing Agents, as depicted in Figure 3.1. Rather than choosing between API calling and web browsing at the task level, our Hybrid Agent is capable of dynamically *interleaving* both actions. We found that agents benefit from the flexibility of this hybrid model. When APIs are available and well-documented, the agent can directly interact with the web services. For websites with limited API support, the agent seamlessly interleaves API calling and browsing, combining the power of both to complete each task.

We evaluated our API-Based and Hybrid Agents on WebArena, a benchmark for real-world web tasks [164], and the results are shown in Figure 3.1. Our experiments revealed three key findings: (1) The API-Based Agent outperforms the Browsing Agent on WebArena by around 15% on average. (2) The API-Based Agent yields a higher success rate on websites with good API support (e.g., Gitlab) compared to those with limited API support (e.g., Reddit). This result underscores the importance of developing comprehensive API support for more accurate and efficient web task automation in the future. (3) The Hybrid Agent outperforms solely Browsing and solely API-Based Agents, further improving accuracy by 10% compared to the API-Based Agent. By dynamically interleaving approaches, the Hybrid Agent is able to provide more consistent and reliable outcomes.

In sum, our results suggest that allowing agents to interact with APIs, interfaces designed specifically for machines is often preferable or at least complementary to direct interaction with graphical interfaces designed for humans.

API Documentation	# Commits ## GET /api/{id}/commits: Get a list of commits in a project. | Attribute | Type | Description | | `id` | integer/string | The ID or path of the project. | | `since` | string | Only commits after or on this date. | | `until` | string | Only commits before or on this date.| Output: JSON containing all commits that meet the given criteria.
API Calling	<execute_ipython> requests.get('gitlab.com/api/a11yproject/commits') </execute_ipython>
JSON Output	[......{ "id": "ed37a2f2", "created_at": "2023-03-13T21:04:49.000-04:00", "title": "Update README.md", "message": "Update README.md", "author": "SaptakS", }]

Figure 3.2: An example of API documentation showing how to get commits of a project, the API call using a Python script to retrieve commits from a project repository, and the resulting JSON response.

3.2 From Web Browsing to API Calling

In contrast to web browsing, API calling allows machines to directly communicate with web services, reducing operational complexity. In this section, we explore an API-based approach when performing web tasks.

3.2.1 APIs and API Documentation

For websites that offer API support, pre-defined endpoints can be utilized to perform tasks efficiently. These APIs, following standardized protocols like REST[1], allow interaction with web services through sending HTTP requests (e.g., `GET`, `POST`, `PUT`) and receiving structured data such as JSON objects[2] as responses. Websites often provide official documentation for the APIs, which can give guidance on how to utilize the APIs. Some documentation is provided as plain text, some in README [3] format, and some in OpenAPI YAML[4] format. Figure 3.2 shows an example of the Gitlab README documentation of `GET /api/{id}/commits`, documenting it's functionality, input arguments, and output types. For example, one could use the Python `requests` library to call `requests.get`"`gitlab.com/api/a11yproject/commits`"), to retrieve all commits to `a11yproject`. This returns a JSON list containing all the commits to this repo, as shown in Figure 3.2.

[1] https://en.wikipedia.org/wiki/REST
[2] https://www.json.org/json-en.html
[3] https://en.wikipedia.org/wiki/README
[4] https://yaml.org/

3.2.2 Obtaining APIs for Agents

One important design decision is how to obtain APIs for agents to use. The way agents interact with APIs depends heavily on the availability of APIs and quality of API documentation. In this work, we acquired APIs by manually looking up official API documentation on a website, although this process could potentially be automated in the future. We classify the availability of APIs according to the following three scenarios:

Sufficient APIs and Documentation Many websites provide comprehensive API support and well-documented API documentation in YAML or README format. In this case, simply use the APIs/documentation as-is. Figure 3.2 depicts an example of API documentation.

Sufficient APIs, Insufficient Documentation There are some challenging situations where APIs exist but good documentation is not officially available. In such cases, additional steps may be required to obtain a list of accessible APIs. In this case, we inspected the frontend or backend code of the website to extract undocumented API calls that can still be utilized by the agent. Then, based on the implementation of APIs of the website, leverage an LLM (GPT-4o[5]) to generate these YAML or README files. By prompting GPT-4o with the relevant implementation details of the APIs (for example, the implementation files of the APIs or example traces of API calls), we generate comprehensive documentation, including input parameters, expected outputs, and example API calls.

Insufficient APIs In the more challenging cases, where only minimal APIs are available, it may be necessary to create new APIs. These custom APIs allow agents to perform tasks that otherwise would require manual web browsing steps. In our case, this was necessary for 1 of 5 websites in the WebArena benchmark that we utilized, such as creating Reddit APIs discussed in Section 3.5.2.

3.2.3 Using APIs in Agents

Once we have the APIs and documentation, we then need to provide methods to utilize them in agents. We utilize two different methods based on the size of the API documentation.

One-Stage Documentation for Small API Sets For websites with smaller numbers of APIs[6], we directly incorporate the full documentation into the prompt provided to the agent. This approach of directly feeding the full documentation worked well for websites with a limited number of API endpoints, as it allowed the agent to have immediate access to all the necessary information without the need for a more complex retrieval mechanism.

[5] https://openai.com/index/hello-gpt-4o/

[6] We use a threshold of 100 APIs, but this could be adjusted depending on the supported language model context size.

Two-Stage Documentation Retrieval for Large API Sets For websites with more APIs, providing the full documentation in the prompt is impractical due to size limitation of agent inputs. To address this, we use a two-stage documentation retrieval process, allowing access to only the needed information to keep the initial prompt concise.

In the first stage, the user prompt provides a task description, with a list of all available APIs along with a brief description of each. For example, {"`GET /api/{id}/commits`": "`List commits in a project`"}. This initial summary helps in understanding the scope of all the available APIs while staying within the prompt size constraints.

In the second stage, if the model determines that it needs detailed information about one or more specific API endpoints, it can use a tool named `get_api_documentation`, which maintains a dictionary that maps each API to its documentation respectively. The dictionary is generated using Python pattern match to retrieve substrings related to each endpoints. This tool is able to search the dictionary and retrieve the full README or YAML documentation for any given endpoint with the endpoint's identifier. The resulting documentation might include the input parameters, output formats, and examples of how to interact with the endpoint. For example, to retrieve the documentation for the API `GET /api/id/commits`, the agent would call `get_api_documentation(''GET /api/id/commits'')`. An example returned API documentation is the documentation in Figure 3.2.

This retrieval method allows the agent to make flexible and informed decisions to perform tasks. If the agent finds that an API does not meet its needs or if it encounters an error, it can easily retrieve the documentation for a different API by calling the tool again. This dynamic approach promotes adaptability and minimizes the risk of incorrect API usage when the number of APIs available is large. The prompt can be found in Appendix A.5.

3.3 Hybrid Browsing+API Calling Agents

We have proposed API-based methods for handling web tasks, but the question arises: given the benefits of API calling, should we discard browsing altogether? The most obvious bottleneck is that not all websites offer good API support. Some platforms offer limited or poorly documented APIs (e.g. no API for shopping on Amazon[7]), forcing agents to rely on browsing to complete tasks.

To deal with these situations, we propose a hybrid methods that integrates both browsing and API calling, and developed a Hybrid Agent capable of dynamically interleaving API calls and web browsing based on task requirements and the available resources. Specifically, for each task, the agent is given the fixed step budget within which it has to finish the task. *In each step of a task*, the agent could either (1) communicate with humans in natural language to ask for clarification, or 2) generate and executes Python code which could include performing API calling, or 3) performs web browsing actions. The Hybrid Agent could choose freely among these options, depending on the agent's confidence in which method is the best for each step.

Ideally, for websites with good API support, the Hybrid Agent can utilize well-documented APIs to perform tasks more efficiently than it could through only browsing; for websites with

[7] https://www.amazon.com

limited API support or poor documentation, the Hybrid Agent could rely more on browsing. We find that enabling it to interleave API calling and web browsing boosts task performance (see Section 3.5).

Prompt Construction The Hybrid Agent's prompt construction extends upon the API-Based Agent by incorporating both API and web-browsing documentation. Similar to the API-Based Agent, the Hybrid Agent is provided with a description of available API calls as discussed in Section 3.2.3. In addition, the Hybrid Agent receives a detailed specification of the web-browsing actions, which mirrors the information given to the Browsing Agent described in Section 2.1.2, including a breakdown of all potential browser interactions. It also maintains a history of all its prior steps such that the agent could make more informed actions. The prompt can be found in Appendix A.6.

3.4 Experimental Setup

3.4.1 Dataset Description

We utilized WebArena [164] as the primary evaluation benchmark. WebArena is a comprehensive benchmark designed for real-world web tasks, providing a diverse set of websites that simulate various online interactions, allowing comprehensive evaluation of agents' abilities to handle both API calling and web browsing across varied web settings. WebArena mainly includes five websites, each with various intents representing different tasks: Gitlab, Map, Shopping, Shopping Admin, Reddit, and Multi-Site tasks. A detailed descriptions of the tasks is in Appendix A.2.

3.4.2 API Statistics for WebArena Sites

The API support for WebArena websites can be categorized into three levels: good, medium, and poor. APIs' availability, functionality, and documentation, as described in Table 3.1, play a crucial role in the efficiency and flexibility of our agents[8].

Sites	Gitlab	Map	Shop	Admin	Reddit
# APIs	988	53	556	556	31‡
Quality	Good	Good	Fair	Fair	Poor

Table 3.1: Number of endpoints and API / documentation quality for WebArena websites. ‡See Section 3.5.2.

Good API Support

Gitlab Gitlab supports 988 endpoints, which offer extensive coverage across a wide range of functionalities, including repositories, commits, and users. This comprehensive API support allows for effective interaction in most WebArena tasks, making Gitlab one of the best-supported platforms in terms of API availability.

[8]See Appendix A.3 for where to find the WebArena APIs.

Map The Map website offers 53 endpoints. Despite the smaller number of endpoints, the APIs available are well-documented and cover most of the essential WebArena use cases.

Medium API Support

Shopping and Shopping Admin The Shopping and Shopping Admin websites share a common set of 556 APIs, which provide a reasonable level of support for common shopping tasks. However, some features, such as adding items to wish lists, are absent, and thus these tasks must be handled via browsing. Despite this, the documentation is fairly detailed. Overall, API calling is a solid, though not exhaustive, solution for handling shopping tasks.

Poor API Support

Reddit The WebArena Reddit is a self-hosted limited clone of the actual Reddit[9] with only 31 endpoints. It offers minimal API support and no documentation, making it the least API-friendly site in WebArena. Many critical functionalities such as searching posts are missing, significantly hampering task execution on Reddit, highlighting the need for a hybrid browsing+API approach.

3.4.3 API Implementation Details

We follow the methodologies discussed in Section 3.2.3 to provide APIs to agents. Appendix A.3 contains the sources of the public API documentations.

One-Stage Documentation for Small API Sets

For websites with fewer than 100 API endpoints, namely the Map and Reddit websites, we directly provide the full documentation to the agent.

Map The README documentation was inputted directly from the public API documentation.

Reddit Since there was no pre-existing documentation for the APIs, we leveraged GPT-4o[10] itself to generate these README files. By prompting GPT-4o with a file containing all implementations of the API endpoints, we generated a README documentation, including input parameters, expected outputs, and example API calls.

Two-Stage Documentation Retrieval for Large API Sets

For websites with more than 100 endpoints, namely GitLab, Shopping, and Shopping Admin, we employ a two-stage documentation retrieval process.

We obtained Gitlab README documentations from the official website. For Shopping and Shopping Admin, the documentation is provided as OpenAPI specification, structured in YAML format.

Agents	Gitlab	Map	Shopping	Admin	Reddit	Multi	AVG.
WebArena Base [164]	15.0	15.6	13.9	10.4	6.6	8.3	12.3
AutoEval [98]	25.0	27.5	39.6	20.9	20.8	16.7	26.9
AWM [139]	35.0	42.2	32.1	29.1	54.7	**18.8**	35.5
SteP [121]†	32.2	31.2	**50.8**	23.6	**57.5**	10.4	36.5
Browsing Agent	12.8	20.2	10.2	22.0	10.4	10.4	14.8
API-Based Agent	43.9	45.4	25.1	20.3	18.9‡	8.3	29.2
Hybrid Agent	**44.4**	**45.9**	25.7	**41.2**	51.9‡	16.7	**38.9**

Table 3.2: Agents' performances on WebArena. †Note that SteP uses prompts inspired specifically by WebArena tasks, while other agents are task-agnostic. Our Hybrid Agent achieve the highest accuracy among these agents. ‡ We report these results using a set of APIs introduced by the authors to mimic the official Reddit website, constructed task-agnostically without access to WebArena tasks. See Section 3.5.2 for more discussions.

3.4.4 Evaluation Framework

We employed OpenHands as our evaluation framework to facilitate the development and testing of our agents [137]. OpenHands is an open-source platform designed for creating and evaluating AI agents that interact with both software and web environments, making it an appropriate infrastructure for our proposed methods. The OpenHands architecture supports various interfaces for agents to interact with. Moreover, this framework allows agents to keep a detailed record of past actions in the prompt, enabling agents to execute actions in a way that is consistent with earlier steps. For coding tasks, it implements an agent based on CodeAct [135] that incorporates a sandboxed bash operating system and Jupyter IPython[11] environments, enabling Python code execution. Additionally, it includes a BrowsingAgent Browsing Agent that focuses solely on web navigation. This agent operates within a Chromium web browser powered by Playwright[12], utilizing a comprehensive set of browser actions defined by BrowserGym [34]. However, while the Browsing Agent can browse websites, and the CodeActAgent make API calls and execute code, there is not an agent that can natively do both. Given this base, we developed two varieties of agents for API-based solving of web tasks.

API-Based Agent Our API-Based Agent essentially uses the CodeAct architecture [135]. In addition to the basic CodeAct framework, we tailor the agent for API calling by adding specialized instructions and examples that guide its understanding and using of APIs. At each step, the agent could utilize all previous actions to make informed selection of actions. The prompt of the API-Based Agent is included in the Appendix A.5.

[9] See Appendix A.3 for more explanations.
[10] https://openai.com/index/hello-gpt-4o/
[11] https://ipython.org
[12] https://playwright.dev/

Hybrid Browsing/API Calling Agent In addition to the API-Based Agent, we developed a Hybrid Agent that integrates Chromium web browsing functionalities powered by Playwright into the existing API-Based Agent framework. This Hybrid Agent is provided the prompt describing both the APIs and the browsing actions, allowing for free transitions between API calling and web browsing. At each step, the agent can utilize the current state of the browser, all previous actions taken by the agent, and the results of those actions to determine the next course of action. The prompt of the Hybrid Agent is included in the Appendix A.6.

For the Browsing, API-Based, and Hybrid Agents, we utilized GPT-4o as the base LLM. However, this could be easily changed to other LLMs.

3.5 Results

3.5.1 Main Results

The main results of our evaluation, as summarized in Table 3.2, demonstrate the performance of three different agents across WebArena websites.

The API-Based Agent consistently achieved higher scores on most websites compared to the Browsing Agent. This agent's strong performance is attributed to its specialized design for API calling, enabling it to efficiently interact with websites and complete tasks with no reliance on browsing.

In contrast, the Browsing Agent, designed solely for navigating web interfaces, demonstrated significantly lower performance across most domains. It achieved its best scores on Shopping Admin and Map, but struggled more on the other websites.

The Hybrid Agent, integrating both API calling and web browsing, outperformed the Browsing and API-Based Agents on all websites. The agent's ability to interleave API calling and web browsing proved beneficial. API calling delivers high performance for web tasks when well-supported APIs are available, while web browsing serves as a backup when API endpoints are unavailable or incomplete. Even if the website provides comprehensive APIs, there might be corner cases where APIs are not supportive. Thus, relying on web browsing is still needed for tasks that would otherwise fail through API-only interactions. Table 3.3 show the frequency of each action type of the Hybrid Agent: it chooses to do both Browsing and API in 77.7% of WebArena tasks, and it shows higher accuracy when choosing API only and API+browsing. More detailed analysis on action types, steps and cost and case studies are in Appendix A.4 and 3.5.4.

Overall, the results indicate that the Hybrid Agent is the most effective for handling diverse tasks in WebArena, particularly in environments that require a blend of API and browsing actions. The API-Based Agent excels in tasks that are primarily API-driven, while the Browsing Agent is more suitable for simple navigation tasks but lacks the versatility needed for more complex scenarios.

Actions	**Frequency (%)**	**Accuracy (%)**
Browsing only	14.3	25.0
API only	8.0	40.0
Browsing + API	77.7	42.0

Table 3.3: Frequency with which the Hybrid Agent chooses each action type and the corresponding success rate.

3.5.2 Does API Quality Matter?

Yes, API quality does significantly impact the performance of agents. High quality APIs provide comprehensive and well-documented endpoints that enable agents to interact accurately and efficiently with websites. With comprehensive API support, the API-Based Agent is able to tackle more tasks through API calling, while the Hybrid Agent rely less on browsing; on the other hand, clear and detailed documentation allows agents to use APIs effectively, ensuring that requests are accurate, and minimizing potential errors in task execution. For example, Gitlab and Map, with the best API support as mentioned in Section 3.4.2, demonstrate highest task completion accuracies among websites by the API-Based and Hybrid Agent.

Conversely, low-quality APIs, characterized by incomplete functionality or ambiguous documentation, can significantly degrade performance. In such cases, the absence of necessary endpoints may prevent the API-Based Agent from completing tasks. Moreover, poorly documented APIs can result in misusing parameters and headers, further reducing the effectiveness of the agent. This highlights the importance for websites to maintain comprehensive and well-documented API support.

An illustrative example of this is the case of Reddit, where the initial performance of the API-Based Agent was suboptimal due to limited API availability. As depicted in Table 3.4, initially, Reddit offered only 18 APIs, lacking the major functionality that common online forums have, such as post voting. Recognizing this limitation, we manually introduced 13 additional APIs including one API on post voting, with our best effort trying to mimic the official Reddit website. This results in a marked improvement in the API-Based Agent's performance, underscoring the direct correlation between the availability of high-quality APIs and the average performance of the API-Based Agent.

Number of Endpoints	18	31
Accuracy on Reddit	9.4%	18.9%

Table 3.4: Effect of adding new Reddit APIs on the API-Based Agent's accuracy.

3.5.3 Error Analysis

We randomly sampled 100 tasks from WebArena and performed error analysis on the API-Based Agent. Figure 3.3 shows the distribution of error categories among these 100 tasks. We found that 33% of the tasks are correctly performed with only API calling, 50% are unsolvable with solely APIs, 6% are incorrect due to incorrect task understanding, and 11% are incorrect due to error in calling APIs such as mal-formatting and wrong input. In other words, among the 50 API solvable tasks, 66% are performed correctly by the API-Based Agent. This showcases the strong capability of the API-Based Agent when given sufficient APIs to solve the task.

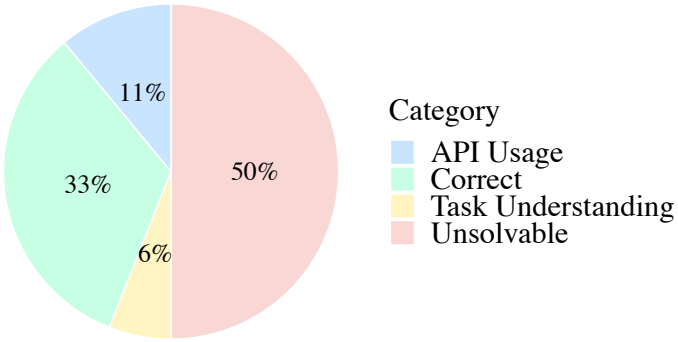


Figure 3.3: Error analysis on 100 WebArena tasks.

Additionally, the average API calls required to solve API solvable tasks are 2.1 API calls, demonstrating how API calling could reduce operational complexity for web tasks. Although the API-Based Agent took an average of 7.8 steps to complete WebArena tasks, most of the steps were taken to retrieve API documentation, resolve errors from it's previous generations, or verify it's outputs.

3.5.4 Case Studies

We analyze two contrasting instances as shown in Figure 3.4 and Figure 3.5, where the Hybrid Agent and API-Based Agent exhibited different levels of performance on WebArena tasks. These case studies highlight the strengths and weaknesses of each agent, demonstrating scenarios where hybrid browsing outperforms API-only or browsing-only approaches, as well as cases where the API-Based Agent excels over the hybrid method.

Case 1 One example where the Hybrid Agent succeeded, while both the API-Based and Browsing Agents failed, involved a task from the Shopping Admin domain. The query was to "delete all negative reviews for Sybil running short", a product listed in the shopping admin interface. In this instance, the API-Based Agent failed because no relevant API endpoints were available for retrieving or deleting reviews. Similarly, the Browsing Agent failed, as completing this task purely through web navigation required too many steps, as depicted in Figure 3.4. This complexity made the task challenging for an agent relying solely on web interactions. However, the Hybrid Agent successfully completed the task by leveraging both API and browsing functionalities. An example trace of the Hybrid Agent shown in Figure 3.4. This case highlights the Hybrid Agent's ability to efficiently combine API calls with web interactions, allowing it to tackle complex multi-step tasks that would be difficult or impossible for solely browsing or solely API-Based Agents.

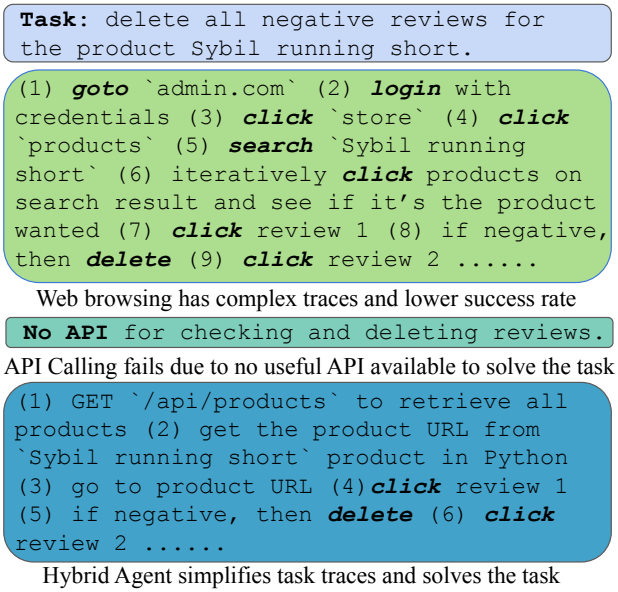


Figure 3.4: The Hybrid Agent succeeds while the Browsing Agent and API-Based Agent both fail.

Case 2 Conversely, there are instances where the API-Based Agent outperforms the Hybrid Agent. One such case occurred in the GitLab website, where the task was to "tell me the email address of the contributor who has the most commits to `ai`." The API-Based Agent successfully completed this task by utilizing the `GET /api/id/contributors` endpoint to retrieve the contributor with the highest number of commits and their associated email address.

On the other hand, the Hybrid Agent attempted to solve the task through browsing but encountered significant challenges. Accessing this information through web browsing required navigating GitLab's interface, locating the correct repository and branch, and identifying the top contributor manually, a task that might be too difficult to perform through web navigation alone. As a result, both the Browsing Agent and the Hybrid Agent failed to complete the task. This case demonstrates an example where API access provides a more straightforward solution than browsing in contexts requiring structured data retrieval.

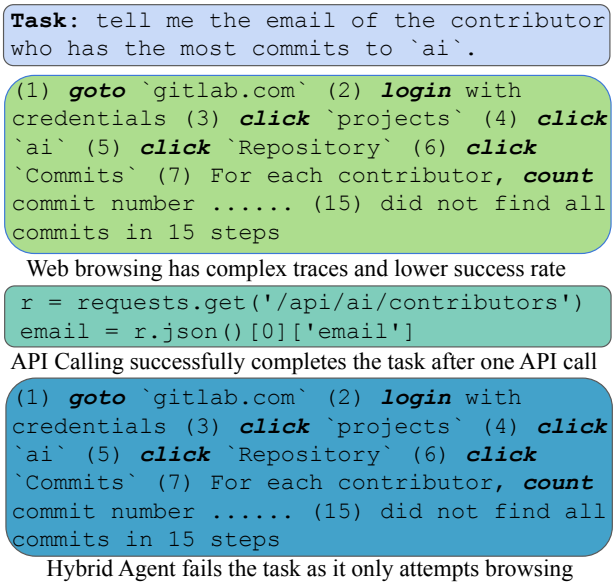


Figure 3.5: The API-Based Agent succeeds while the Browsing Agent and the Hybrid Agent fail.

3.6 Conclusion

In this paper, we propose new web agents that use APIs instead of traditional browsers. We find that API-Based Agents outperform Browsing Agents, especially on websites with good API support. Thus we further propose Hybrid Agents capable of interleaving API calling and browsing that empirically outperforms agents that only use one of the two interfaces.

3.7 Limitations

API Availability A key limitation of API-Based Agents is the inconsistent availability and coverage of APIs across websites. Even platforms with extensive API ecosystems, such as GitLab, may lack support for specific functionalities (e.g., retrieving a user's official username from a displayed name), leading to edge cases where API-Based Agents are unable to complete tasks due to incomplete API support. However, advancements in techniques like Automatic Web API Mining (AWM) [139] could potentially address this limitation by automatically generating APIs for unsupported tasks, reducing reliance on manual API creation.

Incorporating APIs Unlike Browsing Agents, which can adapt to new websites without manual intervention, the API-Based Agent requires additional effort to integrate the necessary APIs documentation to the action space of the agent for each website. This manual integration process increases complexity, particularly when the agent must support a wide range of websites, limiting scalability compared to agents that rely solely on web browsing for interactions. However, future advancements could explore automatically inducing APIs using methods such as Agent Workflow Memory (AWM) [139] and self-improving [161]. These methods could identify and generate API calls for websites lacking formal API support, further expanding the applicability and efficiency of API-Based Agents. By automating the discovery and utilization of APIs,

we envision even more robust agents capable of handling diverse web tasks without reliance on interaction through browsing.

Evaluation Benchmark In this paper, we evaluate web agents exclusively on WebArena tasks. While WebArena offers realistic and diverse challenges, the number and variety of tasks may be limited. Other benchmarks, such as Webshop [145], MiniWoB [120], Mind2Web [31], WebVoyager [50], and VisualWebArena [62], provide alternative valuable evaluation platforms. However, as discussed in Section 2.1.1, WebArena aligns more closely with real-world scenarios and our use case, while other benchmarks lack support for API calling. For example, VisualWebArena is less applicable to our study because WebArena APIs lack support for interacting with images, a core component of VisualWebArena tasks. This could be potentially solved by aforementioned future approaches to automatically induce APIs that support image interactions. Nevertheless, we consider our work as a foundation for such future explorations.

Chapter 4

VISUALPUZZLES: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge

Alongside a unified framework, strong reasoning abilities are crucial for agents to make correct decisions, plan ahead of time, and execute tasks based on users' goals. In this project, we thus introduced VisualPuzzles, a benchmark that could evaluate models' multimodal reasoning abilities in a knowledge-light environment, which could provide guidance on the future development of models with strong multimodal reasoning capabilities. This project first appears in:

- Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. *VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge*. arXiv preprint arXiv:2504.10342.

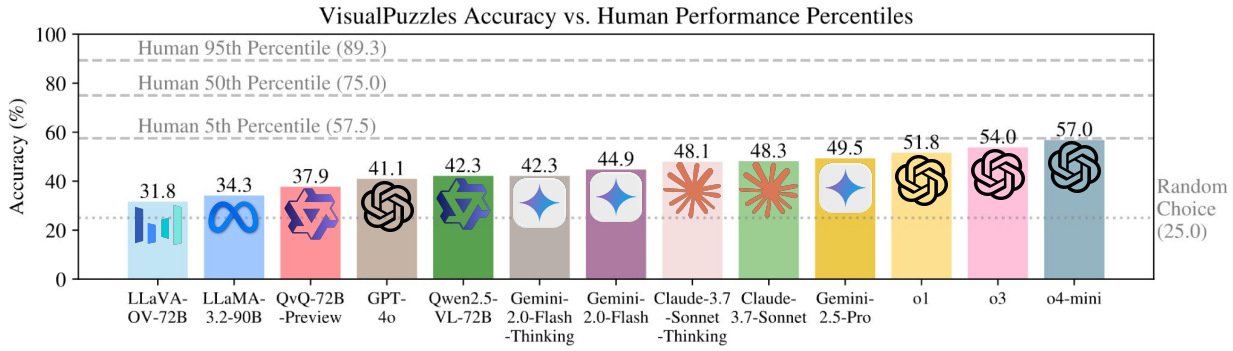


Figure 4.1: Model accuracy on VISUALPUZZLES compared to human performance percentiles. All evaluated models fall below the human 5th percentile (57.5%), highlighting the difficulty of VISUALPUZZLES. Interestingly, models with explicit "thinking" modes do not consistently outperform their base versions, suggesting that current reasoning strategies do not yet generalize well to VISUALPUZZLES's scenarios, even though these strategies have proven effective in existing reasoning tasks that often rely heavily on domain-specific knowledge.

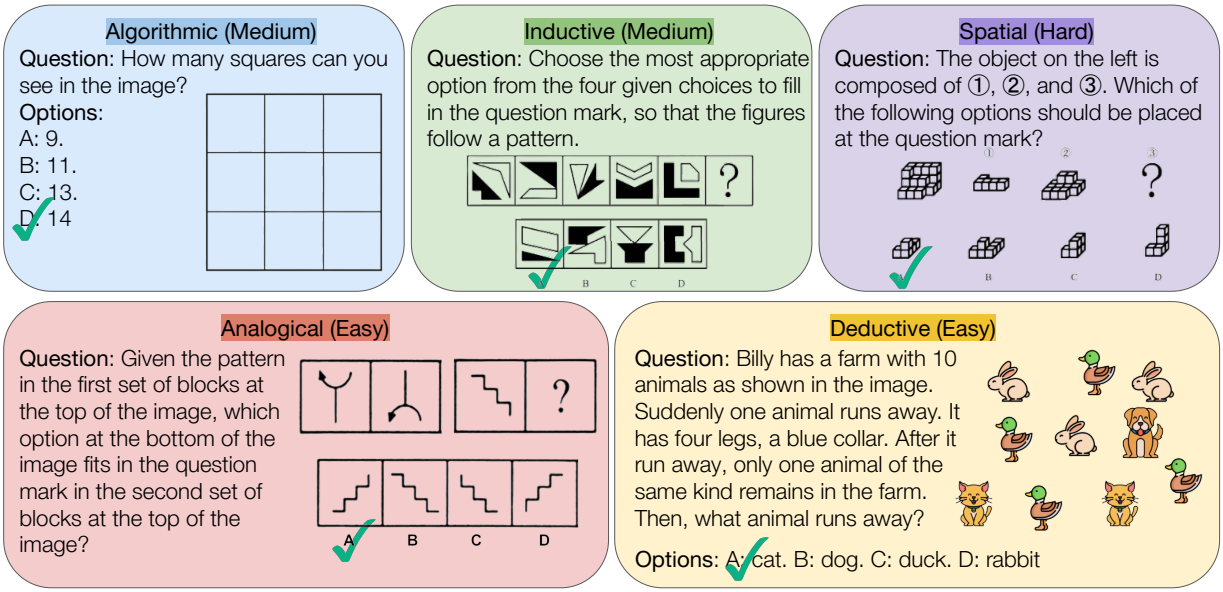


Figure 4.2: Example VISUALPUZZLES instances within each reasoning category

4.1 Overview

Reasoning is a cornerstone of both human and artificial intelligence, enabling systems to solve problems, draw inferences, and make decisions from information. Recent advances in multimodal large language models (MLLMs) [36, 69, 78, 97, 105, 155] exhibit early signs of reasoning in tackling complex tasks such as answering expert-level visual questions [152, 153], interpreting scientific diagrams [113], and solving challenging math word problems [87].

Many of the tasks mentioned above are inherently *knowledge-intensive*; large amounts of knowledge in domains such as science or math are necessary to answer questions correctly [152]. However, in reality, reasoning does not necessitate knowledge. Even non-expert humans can successfully solve logic puzzles, spatial reasoning problems, and analogical tasks using general inferential skills, without requiring deep domain expertise. This raises an important question: *Can we measure MLLMs's reasoning ability independently of measuring their acquisition of domain-specific knowledge?* This question is particularly important with the recent rapid development of reasoning models in the textual domain [29, 55, 106], and emerging application to the visual domain [104].

To address this question, we introduce VISUALPUZZLES, a multimodal benchmark explicitly crafted to assess reasoning capabilities independent of specialized knowledge. VISUALPUZZLES comprises 1,168 carefully curated puzzle-like questions that span five distinct categories of reasoning: algorithmic, analogical, deductive, inductive, and spatial, each annotated with varying difficulty levels. VISUALPUZZLES only requires basic common knowledge and the information presented in the question to solve problems, disentangling reasoning from domain-specific knowledge recall. Our experiments show that VISUALPUZZLES requires significantly fewer domain-specific knowledge concepts compared to benchmarks like MMMU, and models have sufficient knowledge required to solve VISUALPUZZLES questions, enabling us to better assess

multimodal reasoning versus pretrained factual knowledge. While VISUALPUZZLES minimizes reliance on domain expertise, its reasoning complexity exceeds that of existing benchmarks: in VISUALPUZZLES, 82.1% of models' solution steps are logical reasoning steps, compared to 71.5% in MMMU. Additionally, no current MLLM surpasses even the 5th-percentile human performance, highlighting the benchmark's difficulty and the limitations of today's models in general-purpose visual reasoning.

Our experiments with VISUALPUZZLES reveal critical limitations in current MLLMs' multimodal reasoning ability by factoring out domain-specific knowledge requirements and only focusing on reasoning. Specifically, we uncover four key findings:

- **Strong performance on knowledge-heavy benchmarks does not transfer well.** Models that rank highly on MathVista and MMMU often experience substantial performance drops on VISUALPUZZLES, highlighting a disconnect between knowledge-rich and knowledge-light multimodal reasoning tasks.
- **Humans outperform models on easy and medium tasks, while both degrade on harder ones.** Human participants show strong and consistent performance on easy and medium-level questions across reasoning categories. In contrast, models struggle even on simpler tasks.
- **Reasoning enhancements (e.g., long CoT and "thinking" mode) yield inconsistent gains.** While explicit reasoning strategies help certain models tackle complex reasoning tasks, these techniques do not consistently improve performance across all model families and task types.
- **Scaling model size does not ensure stronger reasoning.** We observe no clear trend indicating that larger models outperform smaller ones on VISUALPUZZLES, suggesting that scaling up parameters alone is insufficient to improve domain-agnostic multimodal reasoning.

4.2 VISUALPUZZLES

4.2.1 Motivation and Design Principles of VISUALPUZZLES

Existing benchmarks often conflate multimodal reasoning with domain-specific knowledge, making it difficult to isolate and measure the pure reasoning capabilities of these models.

VISUALPUZZLES is designed to explicitly address this issue by providing a testbed focused on evaluating multimodal reasoning in isolation from specialized knowledge. Specifically, VISUALPUZZLES centers on puzzle-like questions that rely solely on the provided image, question text, and basic common-sense reasoning. The core design principle behind VISUALPUZZLES is to limit the need for external or pretrained domain knowledge. Figure 4.2 shows examples of VISUALPUZZLES within each reasoning category.

4.2.2 Data Collection and Curation

We curated VISUALPUZZLES using a multi-stage pipeline. The process involved sourcing, adapting, and validating questions with an emphasis on reasoning quality and minimal reliance on specialized knowledge.

Question Sourcing. We collected questions from three primary sources: (1) online resources and textbooks focused on logical, visual, and spatial puzzles, (2) synthesized items using images from large-scale vision datasets paired with text prompts, and (3) carefully repurposed items from existing multimodal reasoning benchmarks. Each source was selected to ensure a wide variety of reasoning challenges while avoiding trivial or fact-heavy questions. One major source of our questions is manually translated logical reasoning questions from the Chinese Civil Service Examination[1]. Other sources are listed in section B.1.

Format Adaptation. All collected items were adapted into a consistent multiple-choice format with four options, balancing between text-based and image-based answer choices. This modality balance allows us to better test models' abilities to perform reasoning across diverse formats.

Data Validation. During curation, we applied strict filtering criteria to eliminate questions requiring advanced mathematical knowledge, specialized domain knowledge and facts. Questions were retained only if they could be solved using information present in the image, the question prompt, and basic common sense. A multi-round validation process was conducted by human annotators, focusing on question clarity, solvability, and reasoning type classification.

Attribute Annotation. Finally, each question was annotated with two key attributes:

- Reasoning Category: Each item was categorized as *algorithmic*, *analogical*, *deductive*, *inductive*, or *spatial* reasoning. These five categories were selected as they represent fundamental forms of reasoning widely discussed in literature [39, 81, 87, 152]. At the same time, we aimed to balance comprehensiveness with conciseness, avoiding an overly fine-grained taxonomy that could dilute the benchmark's clarity and usability. This categorization ensures that VISUALPUZZLES covers a broad yet manageable set of reasoning skills relevant to multimodal LLM evaluation.
  - Algorithmic Reasoning involves reasoning over algorithmic rules.
  - Analogical Reasoning requires analyzing the relationships between a pair of entities.
  - Deductive Reasoning involves logically drawing conclusions from known premises.
  - Inductive Reasoning focuses on generalizing rules from observed patterns.
  - Spatial Reasoning requires interpreting and manipulating spatial relationships.
- Difficulty Level: Labeled as easy, medium, or hard, based on annotators' estimated cognitive load and time-to-solve metrics.

This pipeline ensures that VISUALPUZZLES presents a diverse set of high-quality questions designed to challenge multimodal LLMs on their reasoning abilities without involving pretrained domain knowledge.

4.2.3 Dataset Statistics

VISUALPUZZLES comprises 1,168 multimodal reasoning puzzles. It is designed to provide a balanced distribution across different reasoning categories, difficulty levels, and option formats for comprehensive evaluation. The statistics of VISUALPUZZLES are shown in Table 4.1.

[1] Chinese Civil Service Examination (Logic Test), 中国国家公务员考试行测（逻辑推理）

Across the five reasoning types, we maintain a roughly even distribution, ensuring that no single reasoning style dominates the benchmark. Similarly, we balanced the dataset across the three difficulty levels (easy, medium, hard) to capture a wide spectrum of cognitive demands. Approximately half of the answer choices in the dataset are image-based and the other half are text-based, enabling evaluation of models' abilities to reason across diverse query formats. In terms of language complexity, VISUALPUZZLES was constructed with an emphasis on accessibility. Most of the question text uses Basic English vocabulary[2] to minimize the impact of linguistic complexity on reasoning performance, focusing the evaluation strictly on multimodal reasoning.

Category	Statistics
Total Questions	1168
- Algorithmic Reasoning	262
- Analogical Reasoning	211
- Deductive Reasoning	200
- Inductive Reasoning	209
- Spatial Reasoning	286
Easy/Medium/Hard	46%/39%/15%
Option Type (Image/Text)	57%/43%
AVG. Question Length	154.9
% Easy Words	54%

Table 4.1: Statistics of VISUALPUZZLES

Compared to prior benchmarks, VISUALPUZZLES is unique in that it explicitly minimizes domain-specific knowledge requirements while maintaining high reasoning complexity. We demonstrate these traits of VISUALPUZZLES in Section 4.5.

4.3 Experiments and Results

4.3.1 Experimental Setup

We comprehensively evaluated the reasoning abilities of a variety of MLLMs on VISUALPUZZLES. Additionally, we performed human evaluations to better understand the gap between human and models' reasoning capabilities.

We selected a diverse set of proprietary and open MLLMs to ensure broad coverage in terms of model architecture, training scale, and intended application domains. This diversity allows us to capture a wide spectrum of current approaches and capabilities in the field. We integrated VISUALPUZZLES into Lmms-eval [67].

Proprietary Models. We evaluate several leading proprietary models that represent the current state of the art: (1) GPT-4o, o1, o3, and o4-mini [55, 97]; (2) Gemini-1.5-Pro, Gemini-2.0-Flash, Gemini-2.0-Flash-Thinking, and Gemini-2.5-Pro [127]; (3) Claude-3.5-Sonnet and Claude-3.7-Sonnet [6]. Among these, o1, o3, o4-mini are explicitly optimized for reasoning, while Gemini-2.0-Flash-Thinking and Claude-3.7-Sonnet incorporate dedicated modules for extensive step-by-step problem-solving.

Open Models. We further evaluate widely used open MLLMs to gauge how open models compare against proprietary models: (1) LLaVA Series [69, 78, 80]: LLaVA-1.5 (7B/13B), LLaVA-1.6 (7B/13B/34B), and LLaVA-OV (0.5B/7B/72B); (2) Llama-3.2-Vision-Instruct (11B/90B) [36]; (3) Qwen-VL Series [9, 104, 105, 144]: including Qwen-VL, Qwen2-VL (2B/7B/72B-Instruct), Qwen2.5-VL (3B/7B/72B-Instruct), and QvQ-72B-Preview; (4) Cambrian (8B/13B) [130]; (5) Pangea-7B [155].

[2] https://en.wiktionary.org/wiki/Appendix:Basic_English_word_list

We apply both direct multiple-choice prompting and Chain-of-Thought (CoT) prompting to each model, following recent findings that CoT can significantly enhance model reasoning on complex multimodal tasks. For each model we report the best performance, whether achieved by direct multiple-choice prompting or CoT prompting.

Human Performance. To establish a strong baseline for comparison, we conducted human evaluations with 70 college-level volunteers. Human performance provides a valuable upper-bound reference for assessing the current capabilities and limitations of multimodal reasoning models. While this serves as a benchmark for present-day systems, it is possible that future models could surpass this level of performance. Each participant was randomly assigned a subset of the puzzles and completed them under the same resource-constrained conditions as the models (i.e., without access to external tools or the internet). On average, participants completed each puzzle in 78 seconds, reflecting the typical cognitive load and time demands imposed by VISUALPUZZLES.

4.3.2 Overall Results

Table 4.2 and Figure 4.1 compare the performance of humans and a selected set of models.[3] All evaluated models, even the proprietary ones, perform below the 4th percentile of human accuracy, underscoring the significant gap in multimodal reasoning abilities. These results reinforce our finding that, although models have made progress in multimodal understanding, there remains a substantial margin for improvement before they can match or surpass human performance on multimodal reasoning.

This pattern holds across categories as well. In Table 4.2, top human participants (95th percentile) exhibit near-perfect accuracy on multiple reasoning categories, while model performance remains substantially lower, even lower than the worst human performance (5th percentile). These results emphasize the need for continued innovation in model architectures and training paradigms if we aim to close the gap between model and human intelligence on complex multimodal reasoning.

4.4 Disentangling Reasoning from Domain Knowledge

4.4.1 Knowledge Intensity of VISUALPUZZLES

Is VISUALPUZZLES less knowledge-intensive than existing reasoning benchmarks? This question is central to our goal of disentangling reasoning ability from domain-specific knowledge. Many current benchmarks blur this line, making it difficult to assess general reasoning in non-expert settings. VISUALPUZZLES was designed to target visual reasoning skills while deliberately minimizing reliance on specialized knowledge.

To test whether VISUALPUZZLES achieves this goal, we prompted GPT-4o to generate "knowledge concept checklists" for 50 randomly selected questions from a widely-used knowledge-intensive reasoning dataset MMMU and 50 from VISUALPUZZLES. We manually verified each

[3]Full results for every model discussed in Section 4.3 are provided in section B.4, including separate performance outcomes for both direct multiple-choice and CoT prompting.

Model	Algorithmic	Analogical	Deductive	Inductive	Spatial	Overall
Random Choice	25.0	25.0	25.0	25.0	25.0	25.0
Human (95th Percentile)	100.0	100.0	100.0	81.6	100.0	89.3
Human (50th Percentile)	88.0	66.0	80.0	50.0	90.0	75.0
Human (5th Percentile)	68.1	25.0	37.0	0.0	59.1	57.5
Proprietary Models						
GPT-4o	49.2	58.3	49.0	27.3	26.2	41.3
o1	63.7	68.3	67.5	29.2	34.3	51.8
o3	64.5	68.3	69.5	27.3	42.7	54.0
o4-mini	65.3	68.7	75.5	33.0	45.5	57.0
Gemini-2.0-flash	55.3	58.8	57.0	24.4	31.8	45.0
Gemini-2.0-flash-thinking	46.6	70.1	49.0	24.9	25.5	42.2
Gemini-2.5-pro	60.0	64.0	60.0	29.7	36.4	49.5
Claude-3.7-Sonnet	64.5	48.3	65.0	26.8	37.4	48.3
Claude-3.7-Sonnet-Thinking	67.2	44.1	61.5	31.1	37.1	48.2
Open Models (Qwen-Based)						
LLaVA-OV-7B	27.5	28.0	40.5	24.4	28.0	29.4
Pangea-7B	32.4	23.7	38.5	28.7	32.5	31.3
Qwen2.5-VL-7B-Instruct	38.2	23.7	51.5	24.9	31.1	33.7
LLaVA-OV-72B	34.7	26.5	37.0	27.3	28.7	30.8
QvQ-72B-Preview	44.8	43.6	44.0	26.8	30.8	37.8
Qwen2.5-VL-72B-Instruct	53.4	46.9	58.0	25.8	29.5	42.3
Open Models (Llama-Based)						
Cambrian-8B	31.3	24.2	36.0	24.0	29.0	28.9
Llama-3.2-11B-Vision-Instruct	31.0	30.8	39.0	21.1	26.2	29.4
Llama-3.2-90B-Vision-Instruct	45.0	23.2	43.0	26.3	31.5	34.1

Table 4.2: Performance (%) comparison of humans and selected models on VISUALPUZZLES. We report the best performance resulting from direct multiple-choice prompting and CoT prompting for each method. We highlighted all the reasoning models.

question as discussed in subsection B.5.3. Each checklist comprises knowledge-specific questions intended to assess whether a model possesses the background information required to solve the original problem. For example, if a question depends on understanding two distinct physics laws, its checklist would include a question to explain each. The number of checklist items per instance serves as a proxy for knowledge intensity.

We found that MMMU problems resulted in significantly more checklist items on average (3.9) compared to VISUALPUZZLES (1.1), as shown in Table 4.3. This supports the hypothesis that VISUALPUZZLES is substantially less reliant on domain knowledge. As a result, performance on VISUALPUZZLES more directly reflects a model's ability to reason over visual and textual content, offering a clearer signal of progress in multimodal reasoning. Full prompt examples and further discussion are provided in section B.5.

Benchmark	# Knowledge Qs.
MMMU	3.9
VISUALPUZZLES	1.1

Table 4.3: AVG. number of knowledge concept questions generated per instance on MMMU vs. VISUALPUZZLES.

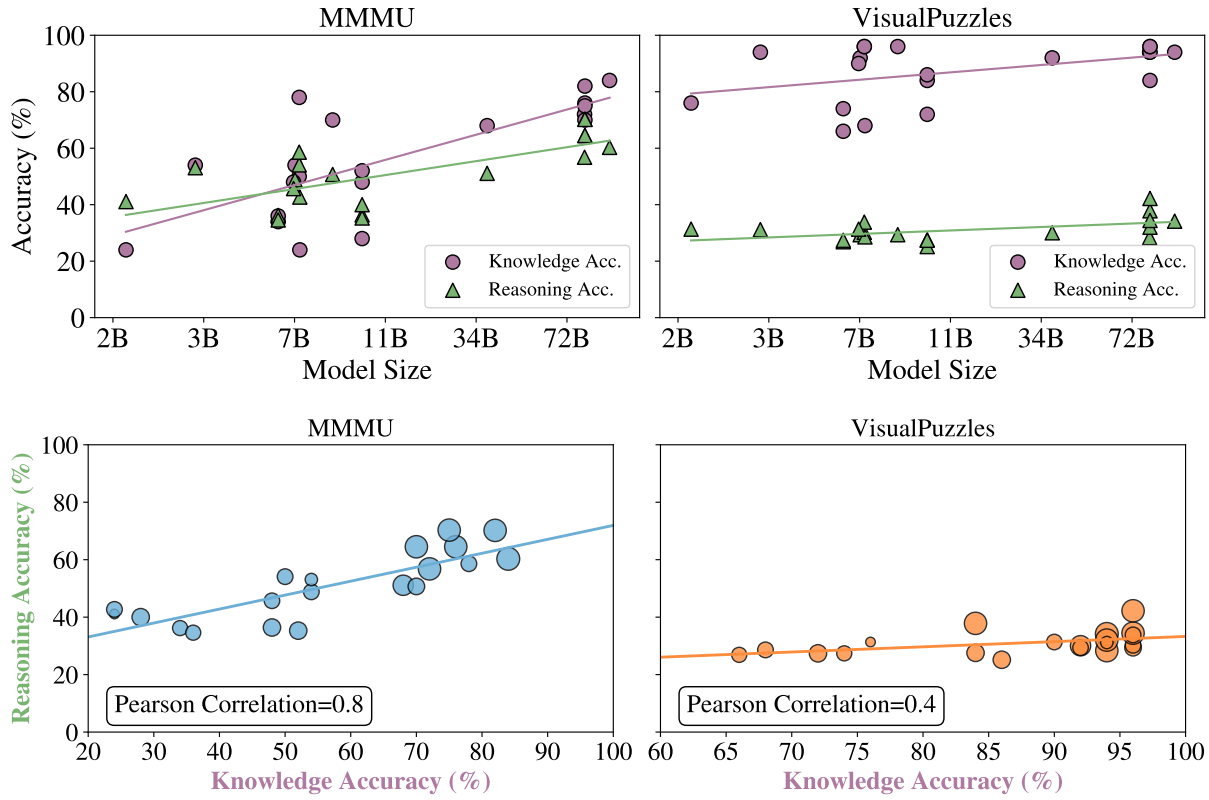


Figure 4.3: Scatter plots with trend lines of the relationship between accuracy and model size (top) and the relationship between reasoning and knowledge accuracy (bottom) on MMMU and VISUALPUZZLES. The dots' sizes represent relative model sizes. The correlation between reasoning accuracy and knowledge accuracy is higher on MMMU (0.8) than on VISUALPUZZLES (0.4).

Do models already possess the knowledge required to solve VISUALPUZZLES? To explore this, we measured models' knowledge accuracy—their ability to answer the knowledge checklist questions correctly—on both benchmarks. This metric reflects how much of the required knowledge is already known by the model, independent of reasoning. We found a stark contrast: while many models exceed 90% knowledge accuracy on VISUALPUZZLES, most score below 60% on MMMU, with smaller models frequently dropping under 50%. Only the largest models approach 80% accuracy on MMMU, underscoring its heavier reliance on domain-specific knowledge.

Does scaling up model size improve performance? We also plot reasoning accuracy (i.e., overall performance on the benchmark) in Figure 4.3, revealing some interesting trends:

- **MMMU**. Larger models tend to have higher knowledge accuracy, and this often translates into higher overall benchmark performance. This aligns with MMMU's reliance on domain-specific understanding; models with more parameters and training data are better at recalling relevant factual knowledge, thus improving their overall performance.
- **VISUALPUZZLES**. Although many models achieve near-100% knowledge accuracy on VISUALPUZZLES, we observe no clear increase in both knowledge and reasoning accuracy as model size grows. In contrast to MMMU, simply scaling number of parameters does not guarantee better performance on VISUALPUZZLES, implying that further gains on VISUALPUZZLES must stem from improvements in models' reasoning abilities rather than reliance on extensive knowledge.

What is the relationship between knowledge and reasoning? Figure 4.3 shows two scatter plots with trend lines that measure how knowledge accuracy correlates with reasoning accuracy across different open models, where the relative sizes of the dots represent the sizes of the models. On MMMU (left), there is a strong positive correlation (0.8), suggesting that a model possessing more knowledge strongly correlates better reasoning performance. In contrast, VISUALPUZZLES (right) exhibits a more modest correlation (0.4). Although there is still an upward trend, gains in knowledge accuracy lead to smaller improvements in reasoning accuracy. This discrepancy implies that while overcoming knowledge gaps is central to reasoning success on MMMU, VISUALPUZZLES tasks demand more nuanced inference steps that depends less on domain knowledge.

Overall, these findings reinforce that VISUALPUZZLES's comparatively lower knowledge requirements are readily met by both proprietary and open models. By contrast, MMMU poses a greater challenge to smaller models in terms of knowledge, for which scaling in size clearly benefits knowledge-intensive tasks. However, on VISUALPUZZLES, larger model size alone is not a decisive factor, which might imply that genuine multimodal reasoning depends on more than just number of parameters or pre-trained knowledge.

4.4.2 Reasoning Complexity of VISUALPUZZLES

Do questions in VISUALPUZZLES require more complex reasoning than those in existing benchmarks like MMMU?

Besides observing that models generally achieve lower accuracy on VISUALPUZZLES compared to MMMU, we further investigated whether this gap stems from increased reasoning complexity. To do so, we measured the proportion of reasoning steps required to solve each question.

We began by gathering detailed, step-by-step solutions from the models for each question, which are manually verified for completeness. Then we classified if each step is a logical reasoning step with the help of LLM.

We show the result in Table 4.4. On average, logical reasoning steps take up 14.8% more total steps in solving VISUALPUZZLES questions compared to those of MMMU (82.1% v.s. 71.5%). This analysis is based on GPT-4o and Gemini-2.0-Flash across 200 randomly sampled questions per benchmark. These results suggest that VISUALPUZZLES demand more extensive reasoning, aligning with its goal of evaluating deeper multimodal reasoning beyond factual recall. Prompt example is shown in section B.6.

Model	MMMU	VISUALPUZZLES
GPT-4o	75.1%	87.0%
Gemini-2.0-Flash	67.9%	77.3%

Table 4.4: Percentage of logical reasoning steps in solving benchmark questions.

4.4.3 Do Reasoning Models Perform Better than Their Baselines?

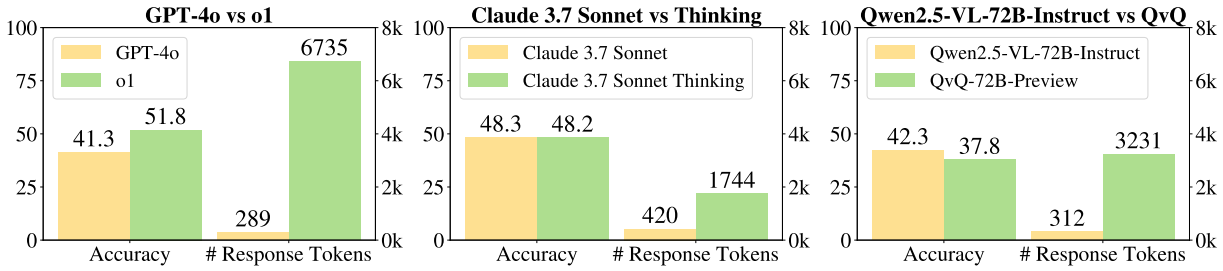


Figure 4.4: Comparison of accuracy and average number of total completion tokens of reasoning models and their general counterparts on VISUALPUZZLES. We didn't include Gemini-2.0-Flash models here because Gemini-2.0-Flash-Thinking does not reveal the number of reasoning tokens of responses. The accuracies of Gemini-2.0-Flash and Gemini-2.0-Flash-Thinking is 45.0% and 42.2% respectively. Despite much higher number of completion tokens, reasoning models do not often achieve better performance on VISUALPUZZLES.

Recent reasoning models often scale up inference compute by generating longer chains of thought (CoTs) to enhance reasoning ability. To assess the effectiveness of this strategy on VISUALPUZZLES, we compare several reasoning models with their non-reasoning counterparts in Figure 4.4. The reasoning model o1 outperforms GPT-4o overall. However, structured "thinking" modes, despite much higher number of completion tokens, show no consistent benefit. Similarity of output further reveals that the thinking mode primarily increases verbosity without meaningfully altering the underlying reasoning process, as illustrated in Figure B.5.

4.4.4 Are Branching and Revalidation Reasoning Patterns Effective on VISUALPUZZLES?

As discussed in Section 4.4.3, reasoning-enabled models do not consistently outperform their non-reasoning counterparts on VISUALPUZZLES. To better understand this discrepancy, we

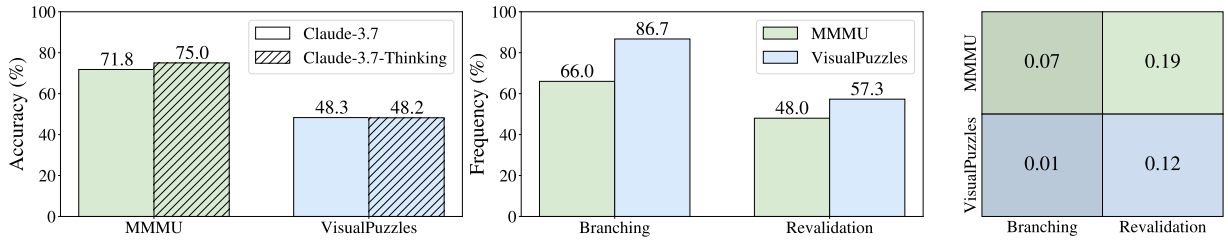


Figure 4.5: Comparison of Reasoning Pattern of Claude-3.7-Sonnet-Thinking on MMMU and VISUALPUZZLES. Left figure compares the accuracy of Claude-3.7-Sonnet and Claude-3.7-Sonnet-Thinking on MMMU and VISUALPUZZLES. Middle figure shows frequency of each pattern. Right figure shows correlation of the patterns with accuracy on the benchmarks.

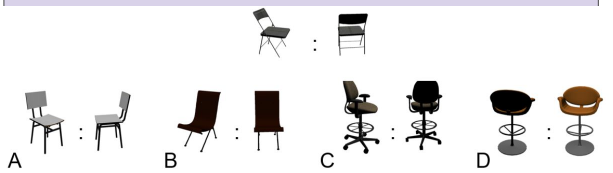
examine Claude-3.7-Sonnet-Thinking's reasoning behaviors present in long CoTs, specifically, branching and re-validation, which are known to play important roles in enhancing reasoning performance[4].

As shown in Figure 4.5, our analysis reveals a striking contrast between benchmarks. On MMMU, both branching and re-validation correlate positively with model accuracy. These strategies help models explore alternative reasoning paths and revisit earlier steps, aiding in the retrieval of relevant factual knowledge,an essential component for solving MMMU's knowledge-intensive questions. An illustrative example is provided in section B.5.

Surprisingly, on VISUALPUZZLES, these reasoning behaviors are more frequent, yet less predictive of success. Despite their increased presence in long-form responses, we observe no significant correlation between these strategies and task accuracy. This suggests that models may be using branching and re-validation in ways that do not meaningfully contribute to solving the problem.

Figure 4.6 highlights this with an example from Claude-3.7-Sonnet-Thinking, where the model applies branching on a VISUALPUZZLES puzzle. However, the additional reasoning paths remain shallow and fail to engage with the core challenge—understanding the spatial arrangement of chairs in the image. The full response is included in section B.5.

Question: Given a pair of chairs arranged at a specific angle, select the option that matches the angle of the chairs in the question. (**Answer**: C)



Claude-3.7-Sonnet (With Thinking)
Let me examine the chairs in the question:
- The chairs are positioned 180 degrees apart.
- They are shown in profile view.
Option A: …; Option B: …; Option C: …; Option D: …
Any of them is correct from an angular perspective.
However, looking more carefully at the exact orientation:
- The original chairs are shown in perfect profile view
- Option A shows chairs in perfect profile view, ...
I'll go with Option A as it presents the cleanest profile view that matches the arrangement in the question.

Figure 4.6: An example of Claude-3.7-Sonnet-Thinking utilizing branching to solve a VISUALPUZZLES puzzle.

4.5 Analysis

4.5.1 Do Models Approach VISUALPUZZLES Questions Differently?

[4]We examined Claude-3.7-Sonnet-Thinking as it explicitly provides thinking output.

Benchmark	Answer-First	Option-First
MMMU	29.3%	70.7%
VISUALPUZZLES (Image Options)	72.5%	27.5%
VISUALPUZZLES (Text Options)	98.3%	1.7%

Table 4.5: Answering Strategy

Table 4.5 shows the statistics of Claude-3.7-Sonnet-Thinking's answering strategy. We observe a clear divergence in answering strategies between MMMU and VISUALPUZZLES. On MMMU, the model tend to follow an option-driven approach—using the provided choices early to eliminate unlikely answers and select the most relevant one, often without explicitly solving the problem. In contrast, models more frequently adopt an answer-first strategy on VISUALPUZZLES, attempting to solve the question independently before comparing the result to the answer choices. This pattern holds across both textual and image-based options, though the option-first approach appears slightly more often (around 30%) for image-based tasks—likely due to the added complexity of visual comparison.

4.5.2 Does model performance transfer between reasoning categories?

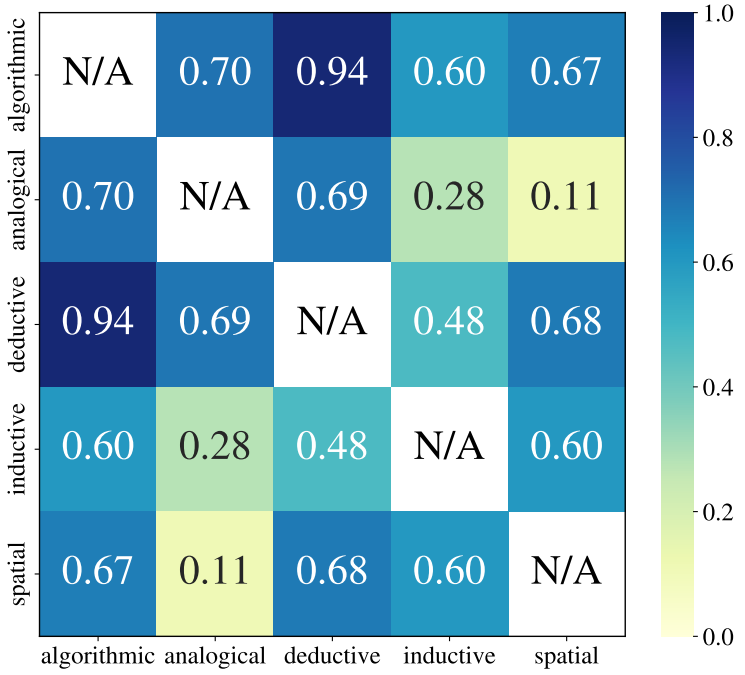


Figure 4.7: Correlation Heatmap among reasoning categories for models (averaged across all models we evaluated).

Figure 4.7 presents a correlation heatmap illustrating the relationships among the five reasoning categories in VISUALPUZZLES. We report model correlations averaged across all models in Table 4.2. For humans, each reasoning category likely engages different cognitive or mental processes [8, 16, 42, 43], so performance in one category might not transfer to performance in another. However, the correlation heatmap of the models tells a different story. We observe notably strong correlations across reasoning categories, with values ranging from 0.11 to as high as 0.94. In particular, algorithmic and deductive reasoning show high correlation (0.94), and other pairs such as algorithmic-analogical and deductive-analogical also exhibit strong associations. This suggests that model performance tends to generalize across categories. However, this generalization may not reflect true reasoning abilities. Instead, the high correlations could indicate that models are leveraging shared surface-level patterns or shortcut strategies that happen to work across multiple structurally different categories, unlike humans, who may rely on distinct cognitive processes.

4.5.3 Error Analysis

Figure 4.8 shows a pie chart illustrating the distribution of error categories of 100 instances generated by Claude-3.7-Sonnet-Thinking on VISUALPUZZLES, revealing that reasoning errors dominate at 56%, reinforcing the fact that reasoning is greatest challenge to models in VISUALPUZZLES. Perceptual errors (21%) and spatial / orientation errors (17%) also constitute substantial portions of failures, reflecting difficulties in interpreting visual elements and understanding spatial relationships. These three categories together account for 94% of mistakes, emphasizing a need for multimodal models with stronger reasoning capabilities with more robust perception and spatial understanding. Textual and visual understanding errors (4%) and reject-to-answer cases (2%) are relatively rare. section B.9 shows samples of error and correct cases of each reasoning and difficulty category.

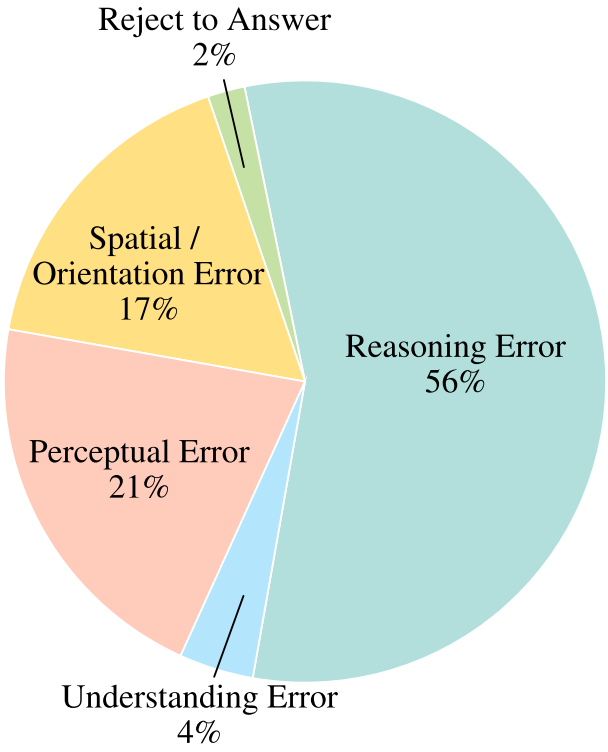


Figure 4.8: Error Distribution of Claude-3.7-Sonnet-Thinking

4.6 Related Work

Multimodal Language Models (MLLMs), particularly vision language models have experienced significant improvements recently. Large scale vision language models [127]; [97]; [6]; including open weight ones [69]; [155]; [82]; [130]; [36] are capable of utilizing both image and text inputs to solve challenging questions.

Multimodal reasoning models, models that specialize in complex reasoning, further push the boundary of MLLMs' capabilities. Large scale multimodal reasoning models such as QVQ [104], Claude-3.7-Sonnet-thinking [6], o1 [55], Gemini-2.0-flash-thinking [127] excel in reasoning heavy tasks such as coding and solving math problems.

Multimodal Reasoning Benchmarks. There exists a number of multimodal benchmarks that test on both the models' world knowledge and reasoning abilities. These benchmarks [152]; [91]; [86]; [153]; [7] emphasize on the multimodal ability of models as a whole, without further separation of knowledge and reasoning.

Recently, more multimodal benchmarks have placed emphasis on multimodal logical reasoning abilities. Many of them [87]; [134] focus primarily on mathematic problems, testing on both mathematical knowledge and reasoning. Some others cover on more general logical reasoning problems [26]; [39], testing on both models' knowledge and reasoning in different domains.

4.7 Conclusion and Future Work

We presented VISUALPUZZLES, a novel multimodal benchmark carefully designed to minimize the impact of domain-specific knowledge and isolate models' core reasoning capabilities. Our results show that while proprietary and large-scale open models achieve relatively higher performance, they still fall short of human-level reasoning—especially on more complex tasks such as analogical and inductive reasoning. Moreover, we observe that strong performance on knowledge-intensive benchmarks like MathVista and MMMU does not necessarily translate into high accuracy on VISUALPUZZLES, underscoring the distinct challenge of knowledge-light reasoning tasks.

These findings suggest that purely scaling model size and knowledge resources may not suffice for robust multimodal reasoning skills; rather, methods that promote structured reasoning, such as explicit thinking modes or recursive reasoning steps, can offer substantial improvements, particularly for hard questions. Future research can explore new training strategies, specialized architectures, or model interpretations tailored to reduce reliance on memorized facts and enhance logical inference. Extending VISUALPUZZLES to include additional types of multi-image reasoning or temporally dynamic visual information may further stress-test models' core inference abilities. By disentangling domain knowledge from multimodal reasoning, we hope VISUALPUZZLES will serve as a valuable tool for developing and evaluating next-generation MLLMs that excel at genuinely understanding and reasoning about the world without depending heavily on specialized factual knowledge.

4.8 Limitations

Disentangling Knowledge Despite our best efforts to isolate domain-specific knowledge from the evaluation of multimodal reasoning, VISUALPUZZLES is still not entirely free of knowledge dependencies. Basic familiarity with everyday objects or common scenarios is still required; complete knowledge free evaluation remains an ideal rather than a practical reality.

Real World Application VISUALPUZZLES emphasizes puzzle-like questions that may not reflect the full diversity of real-world scenarios, limiting generalizability to more specialized domains.

Question Format VISUALPUZZLES focuses on multiple-choice questions, which may not capture the breadth of open-ended reasoning tasks where models must generate complex textual or visual outputs.

Future work can address these limitations by including more varied question formats, broader domains, and more granular analyses of a model's knowledge versus its multimodal reasoning abilities.

4.9 Ethical Statement

This paper uses samples extracted from existing quiz sources for scholarly analysis and testing purposes, in accordance to US fair use law and standard practice. These data are neither intended for, nor capable of, substituting for the original works; thus, we believe their inclusion does not diminish the market value or utility of the source materials. A complete list of references for the data sources is attached in section B.1.

Acknowledgements

This project was supported in part by a grant from DSTA Singapore and the Carnegie Bosch Institute. The authors would like to thank CMU NeuLab colleagues for their constructive comments. The authors would also like to thank all volunteers who participated in the human evaluation.

Chapter 5

PANGEA: A Fully Open Multilingual Multimodal LLM for 39 Languages

To serve people around the world, agents need to understand and generate multilingual contents. Thus, in this project, we proposed and trained Pangea, a multilingual model that achieved SOTA results on multilingual benchmarks. This project first appears in:

- Xiang Yue*, Yueqi Song*, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. *Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages*. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

5.1 Overview

Multimodal large language models (MLLMs) [30, 36, 79, 80, 97, 127] have demonstrated impressive capabilities in tasks such as image captioning, visual question answering, and complex reasoning [153, 154]. Despite this rapid progress in their reasoning ability, a critical flaw persists: *the overwhelming focus on English- and western-centric training and evaluation datasets* [77, 124].

This homogeneity results in a lack of representation for the vast majority of the world's languages and diverse cultural contexts [150]. Consequently, models predominately trained on such data exhibit: (a) diminished performance in multilingual settings [14] with poor tokenization further leading to higher inference costs [4]; (b) generate outputs misaligned with the socio-cultural norms of underrepresented languages [5]; and (c) lack the ability to recognize objects from geographically diverse regions [109] or rare objects belonging to the long-tail [45]. With the increased adoption of these models into real-world applications across the globe, there's an urgent need to develop multilingual MLLMs that equitably serve a diverse set of users. Few efforts have been made to develop multilingual MLLMs [41, 110], however, their performance still exhibits inequalities across languages and lacks evaluation of cultural understanding.

In this paper, we address how to train and evaluate culturally inclusive multilingual MLLMs, using limited open-source resources, tackling four major challenges [150]: **1) Data scarcity:**

high-quality multilingual multimodal data is scarce, especially in low-resource languages, making it difficult to create large-scale training data; **2) Cultural nuances:** visual interpretations are context-dependent and vary across cultures [57, 108]; **3) Catastrophic forgetting:** training on many languages or modalities often results in suboptimal performance on some subsets and require careful balancing; **4) Evaluation complexity:** substantial resources and expertise are required to accurately measure performance across languages and cultures.

To tackle these challenges, we introduce PANGEA, an open-source multilingual MLLM designed to bridge linguistic and cultural gaps in visual understanding tasks. PANGEA is trained on PANGEAINS (Figure 5.1), a high-quality multilingual multimodal instruction tuning dataset comprising 6 million samples in 39 typologically diverse languages. PANGEAINS combines existing open-source resources with newly created instructions focused on multicultural understanding. We curate high-quality English instructions, carefully translate and adapt them for multilingual contexts. To address Western-centric biases in visual representations, we source images from LAION-Multi [116], which includes images from various countries and captions in multiple languages. However, LAION-Multi contains images that are not culturally representative of the country's speaking population, and the associated alt text is often short, noisy, and lacks sufficient detail. To combat these issues, we develop a multicultural multilingual multimodal instruction generation pipeline, leveraging an LLM [36] to score and filter images based on cultural informativeness. We then enhance the remaining data by generating detailed descriptions and creating complex instructions that combine culturally relevant tasks with general multilingual scenarios. This approach improves the model's cultural understanding while maintaining robust multilingual performance.

To evaluate PANGEA's capabilities, we present PANGEABENCH, a comprehensive multilingual and multimodal evaluation suite comprising five multimodal and three text-based tasks across 14 datasets in 47 languages. PANGEABENCH assesses MLLMs' performance on open-domain multimodal chat, image captioning, cultural understanding, multimodal reasoning, and text-only tasks including question answering and complex math reasoning. A key highlight of PANGEABENCH is the introduction of xChat, a human-crafted benchmark designed to evaluate open-ended, information-seeking multimodal conversations. xChat employs a fine-grained evaluation pipeline where human annotators annotate both reference answers and scoring rubrics for each query. An LLM then uses these rubrics to score the model's predictions on a 1-5 scale. This approach offers a more precise assessment of MLLM performance, addressing limitations of coarse LLM-as-Judge methods [162]. Additionally, we introduce xMMMU, a translated version of MMMU [153], testing college-level multimodal reasoning across seven languages. Together, these components provide a robust, nuanced evaluation of PANGEA's cross-lingual and cross-cultural capabilities.

Our results demonstrate PANGEA's abilities in both English and multilingual scenarios, significantly outperforming existing open-source MLLMs on PANGEABENCH, surpassing the best open MLLMs by 0.4% on English tasks and 10.9% on multilingual tasks on average. Notably, PANGEA excels in multilingual and multicultural understanding, evidenced by its performance on xChat, CVQA, and MaRVL benchmarks. PANGEA also matches or outperforms state-of-the-art proprietary LLMs, namely Gemini-1.5-Pro and GPT4o, on several tasks such as XGQA. However, some performance gaps remain in multimodal chat and complex reasoning, shedding light on the need for further improvements in open MLLMs. We discuss key insights from train-

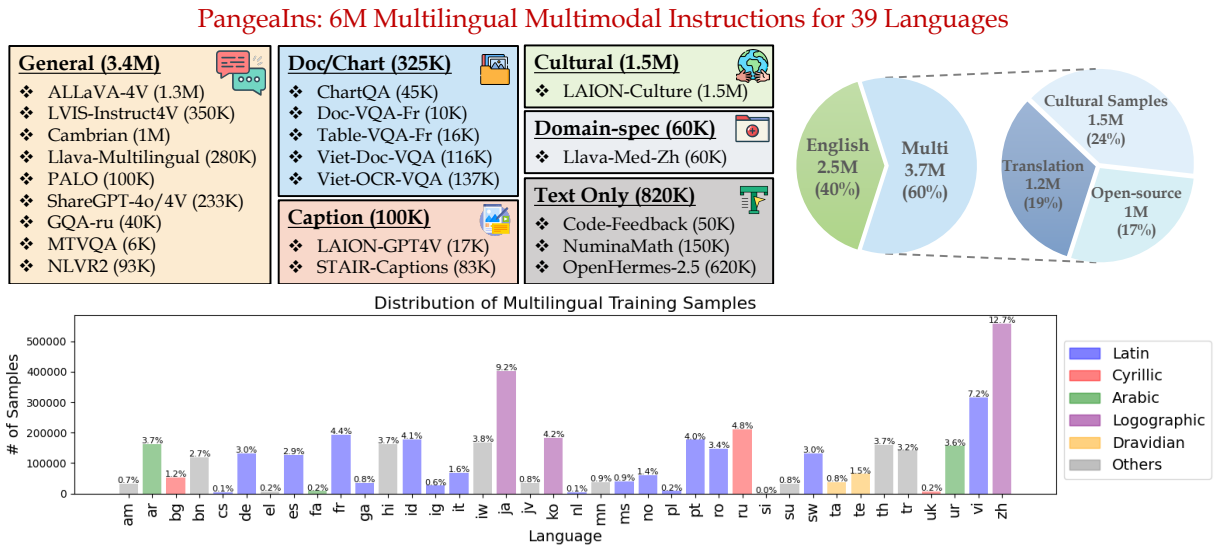


Figure 5.1: Statistics of PANGEAINS, comprising 6M multimodal instructions in 39 languages. The distribution of multilingual training data shows the percent of instances for each language among the *multilingual* instructions. PANGEAINS includes general instructions, document and chart question answering, captioning, domain-specific, culturally relevant, and text-only instructions.

ing PANGEA, including the scaling effect of instructions, the role of English data, the impact of language-specific training proportions, and preliminary methods to improve multilingual OCR. We fully open-source PANGEAINS, PANGEABENCH, PANGEA-7B, and code, to advance culturally inclusive MLLMs across diverse languages.

5.2 PANGEAINS: Multilingual Multimodal Instruction Tuning

Creating a truly multilingual, multicultural MLLM presents unique challenges. We developed PANGEAINS, a diverse and high-quality instruction tuning dataset. Comprising 6 million samples in 39 languages, PANGEAINS was curated with a focus on linguistic and cultural diversity. We implemented three key strategies to ensure comprehensive coverage, each addressing the specific hurdles encountered in multilingual multimodal learning. Figure 5.1 shows the distribution of PANGEAINS.

5.2.1 Machine Translated Instructions

To address the scarcity of human-annotated multilingual multimodal data, we primarily adopt machine translation as a practical and scalable solution to extend data beyond English. While human annotation is ideal, it is resource-intensive and impractical to cover a wide range of languages.

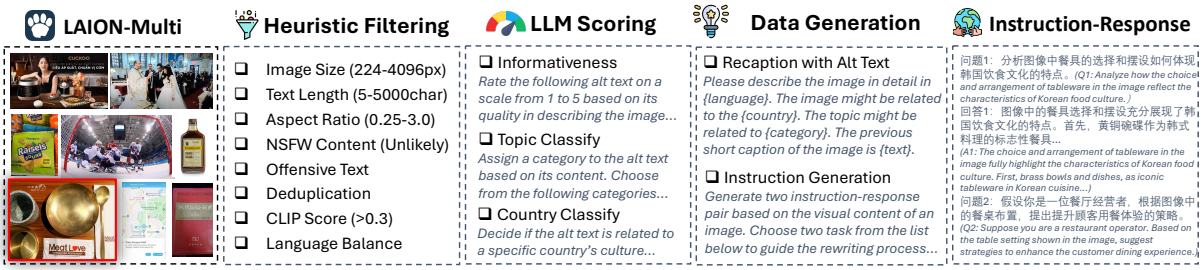


Figure 5.2: Overview of multicultural understanding instructions data generation pipeline.

Constructing a High-quality Pool of English Instructions from Existing Sources. We first collect a high-quality set of English multimodal instructions, which serve as the foundation for translation into other languages. These instructions span a wide range of visual understanding tasks, including general visual instructions and conversations [80, 130], visual reasoning, captioning, and chart question answering [92]. Besides, we also added text-only high-quality English instructions, covering general instructions [128], code [163], and math [70]. Figure 5.1 shows the statistics of our translated datasets. By leveraging existing English instructions, we ensured comprehensive coverage of visual interpretation and text instruction following tasks in English, preparing a pool of high-quality data for translation.

Translation Model Selection. To expand the English instructions to other languages, we initially experimented with strong open-source machine translation models, such as NLLB-3B [95]. However, we found that these models struggled with complex instruction-following scenarios and context-switching tasks, particularly in specialized domains like code generation and mathematical reasoning. For example, in code-related tasks, the model failed to recognize and correctly translate programming language keywords, significantly reducing the quality of the instructions. Based on these limitations, we shifted to using the proprietary Gemini 1.5 Pro model, which shows slightly better performance in small-scale human evaluations compared with GPT4o.

Post-Processing Translated Data. Even with high-quality translations, inconsistencies arose. To resolve issues such as mismatched conversation turns or missing candidates in multiple-choice questions, we developed a post-processing pipeline. This pipeline automatically corrected these errors or directly dropped the examples, ensuring that all translated instructions remained consistent.

Overall, Gemini 1.5 Pro's translation seems satisfactory, providing a fast, cost-effective alternative to human annotation, especially for scaling across languages. However, we acknowledge that machine translation still has limitations, particularly in handling nuanced contexts and cultural subtleties.

5.2.2 Multicultural Understanding Instructions

While machine translation enables scaling across multiple languages, data translated from English is still Anglo-centric in coverage of cultural concepts [150]. To address this, we developed a pipeline focused on creating instructions for multicultural understanding. Both visual and textual elements can convey deep cultural significance, and our goal is to design a dataset that

allows models to not only recognize these nuances but also respond appropriately across various cultural contexts. The pipeline of creating multicultural understanding instructions is shown in Figure 5.2.

Curation of Culturally Diverse Images. To ensure that our dataset captures a wide array of cultural contexts, we began by sampling 10 million images from the LAION-Multi dataset [116], which includes images and short alt texts from diverse languages and regions. A filtering process was proposed to guarantee both the quality and cultural relevance of the images.

- *Heuristic Filtering:* We implemented automatic filtering based on several key criteria: Image Size, Aspect Ratio, Text Length, NSFW content, Offensive Text, Deduplication, and CLIP Score (used to assess the alignment between the image and its textual description). This helped remove low-quality or inappropriate images and ensured the remaining dataset adhered to quality standards.
- *LLM Scoring:* To further refine the dataset, we employed the Llama-3.1-8B-Instruct model [36] to evaluate the quality, subjects, and cultural relevance of the accompanying text descriptions (alt text) for each image. The model was instructed to perform the following tasks: 1) **Evaluate Text Quality:** The alt text was rated on a scale from 1 to 5 based on how well it described the corresponding image, assuming the model could not access the image itself. Alt text scoring below 4 was removed. 2) **Subject Classification:** The model assigned a subject or category to the alt text based on its content. 3) **Country/Region Classification:** The model determined whether the alt text was closely related to a specific country's culture. Images classified as "no specific country" (approximately 60% of the dataset) were excluded to ensure we focused on culturally identifiable content. The full LLM scoring prompt is included in Appendix C.2.
- *Avoiding Overrepresentation:* To maintain a balanced representation, we downsampled images from frequently occurring subjects, such as objects, materials, and clothing, to avoid skewing the dataset toward specific topics or regions. Then, we conducted an accessibility check, removing 30% of the remaining samples due to image download or other issues. Ultimately, we curated a final set of 1M high-quality, culturally specific images, forming the foundation of PANGEAINS.

Captioning Images with Different Languages. To provide context and enhance the model's ability to interpret images, we regenerated more detailed captions using Gemini 1.5 Pro based on high-quality alt texts. In this step, each image was accompanied by a caption written in the language corresponding to its cultural origin. However, our approach was not just about using a capable model. The alt text played a critical role in enriching the data, as it often contained culturally specific and contextually important information that would otherwise be absent from the images alone. For example, in Figure C.1, with high-quality alt text, models can incorporate details such as *"President and CEO of The Walt Disney Company"* and *"a model of Shanghai Disneyland,"* adding significant context that may not be immediately evident from the image. This additional layer of information helps the model generate captions that better capture the cultural and contextual nuances.

Generating Multilingual and Cross-Cultural Instructions.

After recaptioning, we generated multilingual instructions based on the detailed captions with Gemini 1.5 Pro. Instead of only prompting the model to generate random instructions, we did a careful prompt engineering where we first came up with 13 task types (e.g., Information Seeking,

Coding & Debugging, Critical Reasoning, Cultural Interpretation, etc.). Then for each image, up to two QA pairs were created, representing different instruction types to ensure a diverse set of interactions. This approach ensures that the model not only recognizes these visual elements but also responds appropriately across varied linguistic and different instruction contexts. The captioning and instruction generation prompts are included in section C.2.

5.2.3 Curating Existing Multilingual Instructions

To further enrich PANGEAINS, we conducted an extensive survey of available multilingual multimodal literature and datasets, including those hosted on HuggingFace. As a result, we incorporated several high-quality, open-source datasets into PANGEAINS. These include Chinese ALLaVA-4V [21], Viet Document and OCR QA [32], Llava Chinese [75], Llava Medical Chinese Instruction [17], LLaVA-Japanese-Instruct [131], MTVQA [125], Japanese STAIR Captions [148], Russian GQA [10], French Doc-VQA [122], and French Table-VQA [3]. Each of these datasets brings unique linguistic and cultural perspectives to the mix, covering a wide range of languages and task types.

5.2.4 Dataset Statistics

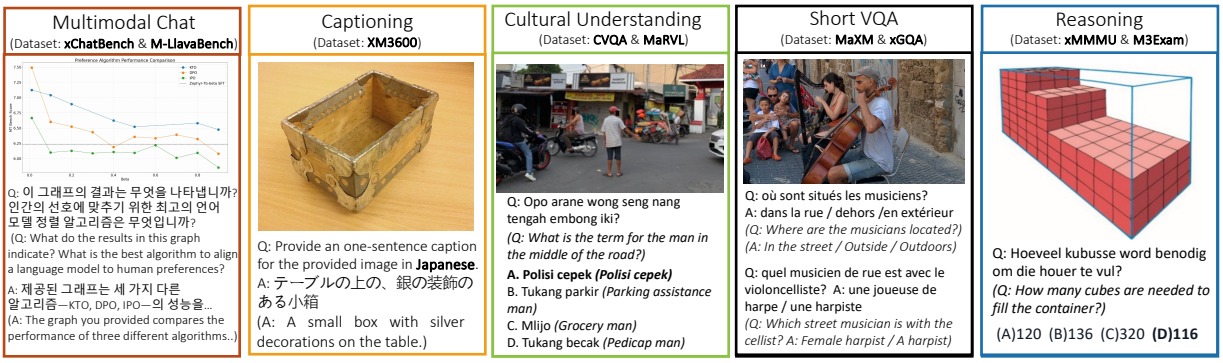
By combining these three methods, we created PANGEAINS, a comprehensive dataset addressing major challenges in building multilingual MLLMs: data scarcity, linguistic diversity, and cultural nuance. Its balanced language and task distribution supports the development of more sophisticated LLMs that can handle complex visual and textual content in a multilingual, multicultural context.

Language and Task Distribution: PANGEAINS features an extensive and balanced distribution of languages, tasks, and cultural contexts (as shown in Figure 5.1). We empirically keep the final language ratio of English to Multilingual as 40%:60% as we found a significant portion of English data plays an important role in cross-lingual transfer. See more discussions about the ratio in section 5.5 and Figure 5.5. The inclusion of diverse multimodal instructions ensures that the model develops a deeper understanding of varied linguistic and cultural environments. Examples of training samples from different languages and categories are provided in section C.8. The comprehensive nature of PANGEAINS lays a solid foundation for training PANGEA, enabling it to become a truly multilingual, multicultural multimodal LLM, capable of understanding and interacting effectively with users from diverse linguistic and cultural backgrounds.

5.3 PANGEABENCH: Evaluation of Multilingual Multimodal Models

5.3.1 Overview of PANGEABENCH

To assess the capabilities of PANGEA across a variety of languages, cultures, and task types, we have developed PANGEABENCH, a comprehensive multilingual and multimodal evaluation suite. PANGEABENCH integrates diverse benchmarks that encompass both multimodal and text-only



Category	Tasks	Datasets	Forms	Size	Languages	Metric
Multimodal	Multimodal Chat	xChatBench	Long	400	zh,en,hi,id,ja,rw,ko,es	LLM-as-Judge
		M-LlavaBench	Long	600	ar,bn,zh,fr,hi,ja,ru,es,ur,en	LLM-as-Judge
	Captioning	XM100	Long	3.6K	36 languages	ROUGE-L
	Cultural Understanding	CVQA	MC	21K	en,zh,ko,mn,ja,id,jv,min,su	Accuracy
		MaRVL	Short	6K	id,sw,ta,tr,zh	Accuracy
	Multilingual VQA	xGQA	Short	77K	en,de,pt,ru,id,bn,ko,zh	Accuracy
		MaXM	MC	2K	hi,th,zh,fr,en,iw,ro	Accuracy
	Reasoning (Multi-subject)	xMMMU	Short/MC	3K	en,ar,fr,hi,id,ja,pt	Accuracy
		M3Exam	MC	3K	en,zh,it,pt,vi,th,af	Accuracy
Text-only	QA	TyDiQA	Short	5.1K	ar,ru,bn,te,fi,sw,ko,id,en	Accuracy
	Translation	FLORES-Sub	Long	18K	ar,en,fr,de,hi,id,iw,ja,pt,ro,tr	ChrF
	Reasoning (Multi-subject, Commonsense, Math)	MMMLU	MC	197K	ar,bn,de,es,fr,hi,id,it,ja,ko,pt,sw,yo,zh	Accuracy
		XStoryCloze	MC	21K	en,ar,es,eu,hi,id,my,ru,sw,te,zh	Accuracy
		MGSM	Open	3K	bn,de,en,es,fr,ja,ru,sw,te,th,zh	Accuracy

Figure 5.3: Overview of PANGEABENCH, which contains 5 multimodal and 3 text tasks covering 14 datasets (including two newly curated xChatBench and xMMMU datasets). The table provides details about the datasets, while the figure shows evaluation examples from five different multimodal eval tasks in our PANGEABENCH.

tasks, enabling a holistic evaluation of PANGEA's performance in cross-lingual, cross-cultural, and multimodal contexts. Each task within PANGEABENCH is designed to probe specific aspects of PANGEA's proficiency, ensuring robust testing across a wide range of scenarios. All tasks in PANGEABENCH are evaluated under a zero-shot setting.

5.3.2 Multimodal Tasks

The multimodal tasks in PANGEABENCH are categorized as follows: Multimodal Chat, Captioning, Cultural Understanding, Multilingual Visual Question Answering (VQA), and Multi-Subject Reasoning. We incorporate these in PANGEABENCH to ensure comprehensive testing of PANGEA's multimodal capabilities. The overview and examples of PANGEABENCH are shown in Figure 5.3.

Multimodal Chat. This task tests the model's ability to engage in natural and dynamic real-world conversations involving both text and images. Multilingual LlavaBench [110] (M-LlavaBench for short) stands as the only benchmark for evaluating multilingual long-form generation capa-

bilities from MLLMs. Following the evaluation pipeline from Zheng et al. [162] and Liu et al. [79], M-LlavaBench uses a coarse-grained evaluation criteria (e.g., "Please rate the helpfulness, relevance, accuracy, level of details of their responses."). Previous works suggest that employing such coarse-grained evaluation criteria may lead to automatic evaluation results that diverge from how humans would evaluate them [59, 60, 61, 66, 147]. To assess baselines with a more accurate evaluation pipeline with *fine-grained evaluation criteria* on diverse scenarios, we additionally annotate a new multilingual multimodal generation benchmark called the **xChatBench**, included in the multimodal chat category of PANGEABENCH. A more detailed explanation of the annotation process of xChatBench is included in section C.5.

Captioning. The XM3600 [129] dataset was developed to evaluate models' capability in multilingual image captioning. It contains images paired with captions in 36 different languages. However, it includes many similar images and captions. To address this, we clustered the images based on captions and manually selected 100 representative images (denoted as XM100). This approach enhances the diversity of the samples and accelerates the evaluation process.

Cultural Understanding. To assess the model's ability to reason about and understand culturally diverse visual content, we use the CVQA [114] and MaRVL [77] datasets. These datasets are designed to test the model's performance in reasoning tasks involving culturally relevant imagery and concepts across multiple languages.

Multilingual VQA. This task measures the model's proficiency in answering questions about images across multiple languages. The xGQA [101] and MaXM [20] datasets provide a diverse range of visual question-answering challenges in several languages and scripts, addressing cross-lingual visual understanding.

Multi-Subject Reasoning. The xMMMU and M3Exam [159] datasets are used to evaluate the model's reasoning abilities across different academic subjects. xMMMU is a machine-translated version of MMMU validation questions, which focuses on multimodal reasoning in multiple subjects. We randomly sample 300 questions from MMMU [153] validation set and employ GPT-4o for the six languages translation. M3Exam challenges the model with real-world educational questions requiring both textual and visual comprehension. Details on how we ensure the translation quality, as well as detailed descriptions of other datasets, can be found in section C.4.

5.3.3 Text-Only Multilingual Datasets

While multimodal tasks are critical for evaluating the holistic capabilities of models like PANGEA, text-only multilingual tasks provide an equally essential dimension to assess. Most existing multimodal evaluations tend to overlook the importance of text-only evaluation, especially across diverse languages. Including text-only tasks in PANGEABENCH allows us to examine whether the model can perform well in scenarios that require deep linguistic understanding without the aid of visual context, highlighting its performance as a foundation model. We include three tasks QA, Translation, and Reasoning covering five datasets for the text-only evaluations in PANGEABENCH.

Specifically, we include TydiQA [27] to test the model's ability to answer questions across 11 typologically diverse languages. We adopt the FLORES [95] dataset to assess machine translation performance. We sample 11 languages (denoted as FLORES-Sub). We use MMMLU [96], a human-translated version of MMLU to test the general language understanding. We use XS-

toryCloze [74] and MGSM [119] to test the model's commonsense and mathematical reasoning ability in multilingual contexts respectively.

5.4 Experiments

5.4.1 Experimental Setup

We train PANGEA on PANGEAINS, our multilingual multimodal dataset comprising 6 million samples across 39 languages. The model uses LLaVA-Next as architecture [80], Qwen2-7B-Instruct [144] as the language model backbone and clip-vit-large-patch14-336 [107] as the vision encoder. The training consists of two stages. First, we pretrain the vision-language connector that aligns the outputs of vision encoder to backbone, with the LLaVA LCS-558K[1] [78, 79]. Then, we perform finetuning on PANGEAINS, where we employ a learning rate of 2e-5, a batch size of 512, coupled with a cosine decay schedule with 0.03 warmup steps. We pretrain and finetune the model for 1 epoch, where pretraining took 4 hours with 8 H100 (32 GPU hours), and finetuning took 168 hours with 8 H100 (1344 GPU hours).

Stages	Pretraining	Finetuning
Training Data		
Dataset	LLaVA LCS-558K	PANGEAINS
#Samples	558K	6M
Model		
Trainable	Projector (20M)	Full Model (8B)
Training		
Batch Size	128	128
LR: ψ_{vision}	1×10^{-3}	2×10^{-6}
LR: $\{\theta_{\text{proj}}, \phi_{\text{LLM}}\}$	1×10^{-3}	2×10^{-5}
Epoch	1	1
GPU Hours (H100)	32	1344

Table 5.1: PANGEA's training configurations.

For evaluation, we compare PANGEA against several state-of-the-art open source baselines, including English-centric models Llava-1.5-7B [78], Llava-Next-7B [80], Phi-3.5-Vision [1], Cambrian-8B [130], Llava-OV-7B [68], Molmo-7B-D [30] Llama3.2-11B [36] and multilingual models PaliGemma-3B [11], PALO-7B [110], mBLIP mT0-XL and mBLIP BLOOMZ [41]. We also consider two text-only LLMs baselines Vicuna-1.5-7B [162] and Qwen2-7B-Instruct [144], which are the backbones of Llava-Next and our PANGEA respectively. We integrate our multimodal tasks in PANGEABENCH into `lmms-eval` [67], a multimodal evaluation package that supports many English multimodal benchmarks. We use `lm-evaluation-harness` [12] to evaluate text-only tasks. We follow the original paper for their best models' prompts in different tasks, and mostly reproduce their original numbers on datasets reported in the original papers.

5.4.2 Multilingual Multimodal Results

The results in Table 5.2 provide clear insights into the strengths and remaining challenges of PANGEA-7B in multilingual and multimodal tasks. Key observations from the evaluation include:

Superior English and Multilingual Performance: PANGEA-7B outperforms existing open-source models across both English and multilingual tasks. While concurrent multimodal models such as Molmo [30] or Llama 3.2 show strong performance on English datasets, they struggle

[1] https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain

Models	AVG (all)		Multimodal Chat				Cultural Understanding			
			xChatBench		M-LlavaBench		CVQA		MaRVL	
	en	mul	en	mul	en	mul	en	mul	en	mul
Gemini-1.5-Pro	67.1	62.5	67.0	54.4	103.4	106.6	75.9	75.7	76.4	72.0
GPT4o	68.6	64.6	71.0	64.4	104.6	100.4	79.1	79.4	81.4	82.1
Llava-1.5-7B	45.4	28.4	28.5	11.8	66.1	40.8	48.9	36.5	56.2	53.7
Llava-Next-7B	51.1	32.7	40.5	18.9	78.9	50.7	55.7	42.6	62.8	50.9
Phi-3.5-Vision	54.0	35.0	38.5	13.2	70.8	58.0	56.3	42.3	72.1	56.5
Cambrian-8B	50.9	36.4	27.5	11.3	78.4	61.8	59.7	47.5	75.4	61.8
Llava-OV-7B	59.5	41.3	51.0	28.5	89.7	55.3	65.2	53.7	72.7	57.5
Molmo-7B-D	55.4	34.1	**49.5**	21.1	**95.9**	13.8	59.4	48.3	65.3	54.9
Llama3.2-11B	57.2	41.9	49.0	27.8	93.9	58.2	**70.2**	**61.4**	64.5	58.1
PaliGemma-3B	37.3	25.8	6.0	3.5	32.1	31.9	52.9	42.9	56.5	52.2
PALO-7B	46.3	32.2	27.0	11.8	68.9	71.2	50.9	39.2	63.3	54.2
mBLIP mT0-XL	35.1	29.8	2.5	0.5	32.7	28.2	40.5	37.5	67.3	66.7
mBLIP BLOOMZ	36.1	30.0	4.0	1.6	43.5	41.0	44.9	36.9	62.3	58.6
PANGEA-7B (Ours)	**59.9**	**52.8**	46.0	**35.8**	84.2	89.5	64.4	57.2	**87.0**	79.0
Δ over SoTA Open	+0.4	+10.9	-3.5	+7.3	-11.7	+18.3	-5.8	-4.2	+11.6	+12.3

Models	Captioning		Short VQA				Multi-subject Reasoning			
	XM100		xGQA		MaXM		xMMMU		M3Exam	
	en	mul	en	mul	en	mul	en	mul	en	mul
Gemini-1.5-Pro	27.6	19.1	54.2	48.7	56.4	63.5	65.8	57.7	77.4	64.7
GPT4o	27.7	19.1	55.8	51.0	60.7	65.4	69.1	58.3	68.0	61.0
Llava-1.5-7B	28.6	1.1	62.0	30.6	49.8	20.4	36.2	31.5	32.3	29
Llava-Next-7B	29.3	9.4	**64.8**	37.8	54.9	21.4	36.7	34.3	36.5	28.4
Phi-3.5-Vision	30.2	5.2	64.7	38.4	**55.3**	25.0	42.6	38.8	55.8	37.2
Cambrian-8B	20.6	9.9	64.6	39.8	**55.3**	28.7	41.8	33.2	34.7	33.4
Llava-OV-7B	**30.6**	7.0	64.4	48.2	54.9	34.8	46.3	41.0	60.4	**45.8**
Molmo-7B-D	22.1	9.1	51.5	43.0	52.9	37.5	44.5	40.4	57.1	39.1
Llama3.2-11B	27.6	4.5	55.6	45.4	**55.3**	43.9	**46.5**	41.4	51.8	36.6
PaliGemma-3B	18.7	0.8	59.7	30.5	47.9	19.9	26.3	25.2	36.0	25.6
PALO-7B	30.4	0.8	60.5	37.8	51.4	16.3	33.1	30.5	30.8	27.8
mBLIP mT0-XL	31.9	3.1	44.2	39.9	44.7	36.8	29.3	30.4	22.8	25
mBLIP BLOOMZ	22.5	10.3	43.3	36.9	44.7	24.8	29.2	30.8	30.3	29.5
PANGEA-7B (Ours)	30.4	**14.2**	64.7	**60.2**	**55.3**	**53.3**	**45.7**	**43.7**	**61.4**	42.1
Δ over Best Open Model	-0.2	+3.9	-0.1	+12.0	0.0	+9.4	-0.8	+2.3	+1.0	-3.7

Table 5.2: Overall performance on the multilingual multimodal benchmarks in PANGEABENCH. The best-performing open model on each dataset is in **bold** and the second best is underlined.

Models	AVG (all)		FLORES-Sub		TyDiQA		XStoryCloze		MGSM		MMMLU	
	en	mul	x→en	en→x	en	mul	en	mul	en	mul	en	mul
Vicuna-1.5-7B	52.1	38.7	55.6	42.4	59.7	52.7	78.1	57.4	17.6	6.4	49.5	34.7
Qwen2-7B-Instruct	66.6	**54.5**	**61.8**	**46.0**	72.2	**71.2**	**80.3**	**61.9**	48.8	40.4	**70.1**	**53.1**
Llava-1.5-7B	53.1	39.0	54.7	41.5	66.8	52.8	79.1	57.6	14.8	7.6	50.2	35.7
Llava-Next-7B	54.0	38.9	54.8	41.4	68.3	52.1	79.1	57.1	15.6	7.5	52.1	36.5
Phi-3.5-Vision	60.7	41.7	28.5	32.5	**75.9**	51.3	77.9	54.8	**59.2**	33.1	62.0	36.7
PALO-7B	52.0	37.5	52.9	40.4	69.4	50.8	77.4	57.2	13.6	5.8	46.7	33.4
PANGEA-7B (Ours)	**72.8**	54.3	60.7	44.9	73.7	66.0	79.1	61.2	**82.0**	**47.4**	68.4	52.2

Table 5.3: Overall performance on text-only multilingual benchmarks in PANGEABENCH.

in multilingual evaluation settings. Particularly in multilingual subsets like xChatBench, M-LlavaBench, and MaRVL, it has achieved substantial gains, highlighting its effectiveness in both cross-lingual and cross-cultural contexts.

Balanced Cross-Language Capabilities: Unlike many models that exhibit a significant drop in performance when moving from English to multilingual tasks, PANGEA-7B is relatively consistent. For instance, in Multimodal Chat tasks, the performance gap between English and multilingual remains relatively small, indicating its ability to handle multiple languages effectively.

Challenges Compared to Proprietary Models: While PANGEA-7B leads in open-source models, some gaps remain when compared to closed-source models like GPT4o. Additionally, though PANGEA-7B narrows the gap between English and multilingual performance, there is still room for improvement in fully closing this divide across all tasks.

5.4.3 Multilingual Text-only Results

We further evaluate our model in text-only scenarios in Table 5.3. Interesting findings include:

Best Text Performance Among Multimodal LLMs: PANGEA-7B demonstrates the strongest performance among all multimodal LLMs in the text-only tasks consistently outperforming baselines like Llava-Next-7B. This highlights that, despite being trained as a multimodal model, PANGEA-7B maintains superior text understanding and reasoning capabilities compared to other MLLMs.

Maintained Performance from its Text Backbone. PANGEA-7B generally maintains or sees slight drops in performance on most text-only benchmarks compared with its text backbone Qwen2-7B-Instruct. Notably, the model shows a significant improvement in MGSM. This improvement is directly attributable to the inclusion of math-related instructions in PANGEAINS, which enhances the model's capability to handle complex multilingual reasoning and mathematical tasks.

5.5 Discussion

Finally, we explore implications of our findings and their potential impact on future developments in the field. We examine the scaling effects of instruction quantity, the persistent role of English data, the relationship between training sample proportions and performance. Through this discussion, we aim to provide a comprehensive understanding of our model and chart a course for future advancements. More discussion on qualitative examples of model behavior on multilingual multimodal chat and challenges in multilingual OCR can be found in section C.5 and section C.10.

Scaling Effect of Number of Instructions. Understanding how the quantity of instructions affects model performance is crucial for optimizing training strategies and resource allocation. Figure 5.4 reveals a clear scaling effect related to the number of instructions used during training. Performance improvements were consistent as we increased the number of multilingual instructions in PANGEAINS, for both English and multilingual performance. This demonstrates the necessity of scaling multilingual multimodal instruction tuning.

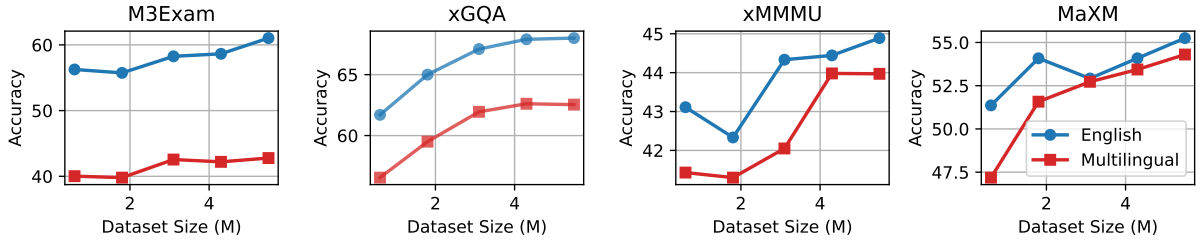


Figure 5.4: Scaling effect of training samples on English and multilingual scores across datasets.

Role of English Data. In multilingual scenarios, English data plays a pivotal role in cross-lingual transfer. To investigate this, we sampled 500K examples from the translated data described in subsection 5.2.1, ensuring a consistent data distribution. We varied the ratio of English data while keeping the total number of training samples fixed at 500K. For the 17 multilingual languages in the translated subset, we evenly distributed the number of samples across languages.

As shown in Figure 5.5 , English performance generally improves as the percent of English data increases. Surprisingly, using only multilingual data results in relatively lower multilingual performance. As we introduce more English data, multilingual performance improves, peaking at 38.7% with 40% English. However, performance drops sharply when English data reaches 100%. This suggests that English data aids cross-lingual transfer, however, over-reliance on it harms multilingual performance.

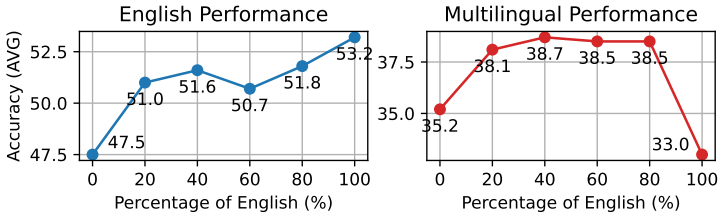


Figure 5.5: Impact of English training data proportion on English vs. multilingual performance.

How does the proportion of training samples in a language affect downstream performance? Is downstream task performance correlated with the number of training samples? Our

analysis in Figure 5.6 revealed the relationship between training sample proportion and downstream performance. While there is a general positive correlation, the impact varies significantly across languages and tasks. For widely spoken languages with rich resources, we observed a near-linear relationship. However, for low-resource languages, even a small increase in proportion yielded disproportionately large performance gains. Interestingly, we also noted instances of positive transfer between typologically similar languages. These findings suggest that strategic allocation of training samples, considering both language prevalence and linguistic similarities, can optimize model performance.

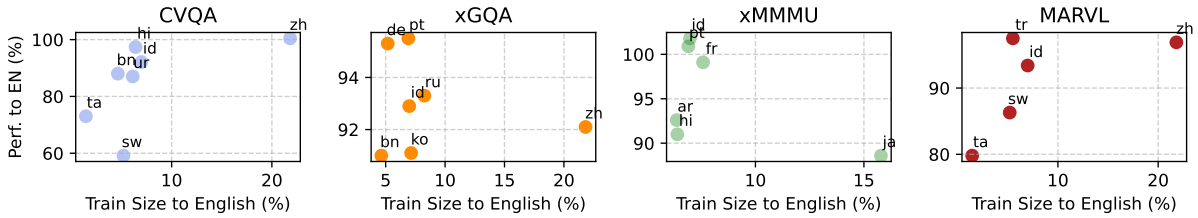


Figure 5.6: The relationship between training sample size (relative to English) and performance (relative to English) of different languages across four datasets.

5.6 Conclusion

In this paper, we introduced PANGEA, a multilingual MLLM designed to bridge linguistic and cultural gaps in visual understanding tasks. By leveraging PANGEAINS, our newly curated 6M multilingual multimodal instruction data samples, we demonstrated significant improvements in cross-lingual and cross-cultural understanding across 39 typologically diverse languages. Our comprehensive evaluation using PANGEABENCH revealed PANGEA's superior performance compared to existing open-source models. We also highlight ongoing challenges in areas such as low-resource language support and multilingual OCR. We fully open-source PANGEA, PANGEAINS, and PANGEABENCH to facilitate future research to build open and inclusive MLLMs.

Acknowledgments

This work was supported in part by a Carnegie Bosch Institute Fellowship to Xiang Yue, as well as grants from DSTA Singapore, and the Programs for Bridging the gap between R&D and the IDeal society (society 5.0) and Generating Economic and social value (BRIDGE)/Practical Global Research in the AI × Robotics Services, implemented by the Cabinet Office, Government of Japan. The training is supported by the CMU FLAME Center. The authors would like to thank Google Gemini credits for data construction and evaluation. The authors would also like to thank CMU NeuLab colleagues for their constructive comments.

Chapter 6

Conclusion and Future Work

This thesis advances the goal of a unified agent interface across diverse environments: we expanded web action spaces from human-oriented GUIs to machine-oriented APIs, improving the effectiveness, efficiency, and robustness of web agents. We evaluated multimodal models' reasoning abilities in a knowledge-light setting to help separate genuine reasoning from shortcut-based knowledge recall. Additionally, to serve users worldwide, we proposed a multilingual multimodal LLM that achieves state-of-the-art performance on multilingual benchmarks.

Together, these efforts illustrate a broader theme toward agents that are both general and unified: general, in their ability across modalities, languages, and domains; unified, in their reliance on unified interfaces across environments. This points toward a future where agents can handle different tasks across modalities, leverage robust reasoning, and inclusively serve global users. Achieving this vision will require scaling unified interfaces, broadening training data to reflect the diversity of real-world challenges, and developing adaptive and lightweight evaluation frameworks that can keep pace with rapidly evolving capabilities of agents.

As immediate next steps, building on our unified web–API interface, we plan to create a large-scale unified training dataset (spanning coding, web, and general agentic tasks) that yields strong results on a variety of agentic tasks. We outline next steps toward generalist agents:

- **Unified Agent Data.** Extend the unified interface and training data beyond the web to include coding, databases, and other real-world tasks.
- **Multimodal Tool Calling.** Build a comprehensive tool library (e.g., image segmentation) and policies for reliable selection and composition of tools.
- **Evaluation.** Develop lightweight, cost- and time-efficient benchmarks and proxy tasks to accelerate evaluation and training iteration.

Appendix A

Appendix for Chapter 3

A.1 Related Work

The development of AI agents that interact with the web and APIs has garnered significant research attention. Web browsers, serving as the primary interface for interacting with online content, have long been a focus for AI research. Web-based agents that can navigate websites, extract information, and perform tasks autonomously have been studied extensively, especially in the context of LLMs and agents designed to mimic human behavior online.

Web Navigation Agents Much prior work has centered around agents that perform web-based tasks using browsing actions [63, 65, 98, 145]. These agents are particularly effective in environments where human-like interaction with a user interface is necessary [34]. Frameworks such as WebArena have further refined the evaluation of such agents by providing complex and realistic web navigation tasks [164]. Our work explores the Hybrid Agent that combines web browsing with API interactions. While prior work primarily focuses on browsing-only agents, we examine how Hybrid Agents can enhance performance by integrating structured API calls with web navigation.

Code Generation Agents and Tool Usage Another stream of research focuses on agents that interact with online content via application programming interfaces (APIs) [35, 99, 103, 136, 138, 151]. In this context, works such as CodeAct have pioneered the development of agents that generate and execute code, including API calls, to perform tasks typically reserved for software engineers [126, 135, 158]. These API-Based Agents are optimized for tasks that involve structured data exchanges, allowing them to perform operations more efficiently than traditional web navigation agents [118]. On the other hand, our work integrates both browsing and API interactions, demonstrating that Hybrid Agents can outperform API-only agents in tasks requiring web navigation. While existing research shows the efficiency of API-Based Agents, our Hybrid Agent dynamically switches between APIs and web browsing to optimize task performance.

Additionally, we are the first to explore comparative studies of API v.s. Browsing Agents on the same websites. We demonstrate that API-Based Agents are often more efficient than Browsing Agents when APIs are available, leading to significant improvements in performance. This

finding is aligned with previous studies that highlight the advantages of structured interactions through APIs compared to unstructured web browsing interactions.

A.2 WebArena Tasks

WebArena reproduces the functionality of several commonly-used websites using open-source frameworks, with real-world data imported into the reproduced websites.

WebArena includes tasks related to the following websites:

- **Gitlab**[1] – 180 instances: This website contains tasks related to project management and version control, where agents perform tasks like opening issues, handling merge requests, or creating repositories. Example query: Submit a merge request for `a11yproject.com/redesign` branch to be merged into the `markdown-figure-block` branch, assign myself as the reviewer.
- **Map**[2] – 109 instances: For this website, tasks are centered around navigation, trip planning and queries about distances, requiring the agent to retrieve and interpret map-based data, similar to using real-world map services like Google map. Example query: Tell me the full address of all international airports that are within a driving distance of 50 km to Carnegie Mellon University.
- **Shopping**[3] – 187 instances: Tasks related to this website represents typical e-commerce tasks, such as searching for products, adding items to carts, and processing transactions. Example query: Change the delivery address for my most recent order to 77 Massachusetts Ave, Cambridge, MA.
- **Shopping Admin**[4] – 182 instances: This setting involves managing backend administrative tasks for an online store, like managing product inventories, processing orders, or viewing sales reports. Example query: Tell me the the number of reviews that our store received by far that mention term "`satisfied`".
- **Reddit**[5] – 106 instances: Tasks here are similar to interactions with the official Reddit, where agents need to post comments, upvote or down-vote posts, or retrieve information from threads. Example query: Tell me the count of comments that have received more downvotes than upvotes for the user who made the latest post on the Showerthoughts forum.
- **Multi-Website Tasks** – 48 instances: These examples involve tasks that span across two websites, requiring the agent to interact with both websites simultaneously, adding complexity to the task. Example query: Create a folder named news in gimmiethat.space repo. Within it, create a file named urls.txt that contains the URLs of the 5 most recent posts from the news related subreddits?

[1]Original Website: https://gitlab.com
[2]Original Website: https://www.openstreetmap.org
[3]Developed using Adobe Magento (https://github.com/magento/magento2)
[4]Developed using Adobe Magento (https://github.com/magento/magento2)
[5]Deployed Postmill (https://postmill.xyz/), the open-sourced counterpart of Reddit (https://www.reddit.com)

A.3 Obtaining APIs of WebArena Websites

- **Gitlab**: we leveraged the open Gitlab REST APIs[6], consisting of 988 endpoints. Most of WebArena tasks are covered by these APIs, with only a small fraction of tasks, such as retrieving users' Gitlab feed token, are not covered by any existing endpoints,
- **Map**: The Map website offers three sets of APIs, each offering distinct functionalities, with a total of 53 endpoints. The first set of APIs, openly available at Nominatim[7], offers essential endpoints for geographic searches. The second set of APIs, from Project OSRM[8], focuses on routing and navigation functionalities. The third set of APIs, available at OpenStreetMap[9], deals primarily with map database operations. This API is rarely used in WebArena tasks but offers capabilities for interacting with OSM data.
- **Shopping**: The e-commerce website uses APIs from the Adobe Commerce API[10], consisting of 556 endpoints. These endpoints provide support for common shopping tasks such as purchasing products, searching categories, and managing customer accounts.
- **Shopping Admin**: This website shares a common set of APIs with the shopping website. However, this website requires a unique admin token to access the admin-only APIs such as changing the price of products and deleting products from stores.
- **Reddit**: The Reddit tasks in WebArena are based on a self-hosted limited clone of the Reddit website [11], with limited functionalities as compared to the official site. As a result, all of the available APIs are self-implemented, with a best effort to mimic to official Reddit APIs. This website supports 31 endpoints, which include writing comments and voting posts.
- **Multi-Website Tasks**: we provide APIs from all websites included in the task to the agents, where we explicitly state which set of APIs belongs to which website. This could allow agents to identify the correct set of APIs to use when transitioning between websites. It is also worth noting that the framework of our agents supports a unified workspace that allows the agents to carry over the information from one website to another.

A.4 Additional Analysis

Table A.1 documents the percentage of actions of our Hybrid Agent. Across all websites, our Hybrid Agent chooses to do both Browsing and API in the same task at least half of the time.

Table A.2 documents the accuracy of the Hybrid Agent across websites when performing different choices of actions. It shows consistently high accuracy when choosing API only and

[6]Documentation of all Gitlab APIs could be found at https://docs.gitlab.com/ee/api/rest/.

[7]The API documentations could be found at https://nominatim.org/release-docs/develop/api/Overview/

[8]Documentations of APIs available at https://project-osrm.org/docs/v5.5.1/api

[9]API documentations openly available at https://wiki.openstreetmap.org/wiki/API_v0.6

[10]APIs documented at https://developer.adobe.com/commerce/webapi/rest/quick-reference/

[11]https://codeberg.org/Postmill/Postmill

Actions	Gitlab	Map	Shopping	Admin	Reddit	Multi	AVG.
Browsing only	7.8	3.7	38.5	2.2	0	8.3	12.1
API only	21.1	4.6	7.5	1.1	0	10.4	7.9
Browsing+API	71.1	91.7	54.0	96.7	82.1	81.3	80.0

Table A.1: Percentage of Actions (%) that our Hybrid Agent takes for each type of tasks. Each column sums up to 1.

Choices of Action	Gitlab	Map	Shopping	Admin	Reddit	Multi	AVG.
Browsing only	7.1(1/14)	50.0(2/4)	23.6(17/72)	50.0(2/4)	0(0/0)	25.0(1/4)	23.5(23/98)
API only	47.4(18/38)	40.0(2/5)	21.4(3/14)	50.0(1/2)	0.0(0/0)	20.0(1/5)	39.1(25/64)
Browsing+API	47.7(61/128)	46.0(46/100)	27.7(28/101)	40.9(72/176)	51.9(55/106)	15.4(6/39)	41.2(268/650)

Table A.2: The accuracy (%) of the Hybrid Agent across choices of actions for each website, with the number of correct instances / number of total instances in parentheses.

API+browsing.

Table A.3 shows the breakdown of number of steps and cost by website.

Agents	Gitlab		Map		Shopping		Shop-Admin		Reddit		Multi Sites		AVG.	
	steps	cost	steps	cost	steps	cost	steps	cost	steps	cost	steps	cost	steps	cost
Browsing	9.4	0.2	8.0	0.1	7.3	0.1	7.0	0.2	11.1	0.1	7.5	0.1	8.4	0.1
API-Based	7.0	1.7	6.6	1.1	8.2	1.0	8.4	1.1	8.8	0.6	7.7	1.6	7.8	1.2
Hybrid	8.1	2.0	9.4	1.7	8.2	1.3	9.0	1.4	7.8	0.6	8.0	1.9	8.5	1.4

Table A.3: Number of Steps and Cost (in U.S. dollars) of Agents across WebArena Websites

Steps Figure A.1 demonstrates a scatterplot of the average accuracy of each agent on WebArena over their average steps. The Browsing Agent takes more steps to complete tasks compared to the API-Based Agent on average, while the Hybrid Agent takes the most steps amongst the three agents. This is likely due to the Browsing Agent's reliance on navigating web interfaces and interacting with visual elements, which involves a sequential and more time consuming processes. The API-Based Agent is the most efficient in terms of steps, as it can directly interact with structured data via APIs, bypassing many of the steps involved in traditional web navigation. The Hybrid Agent, combining both action spaces from the Browsing Agent and the API-Based Agent, takes more steps than both agents.

Costs Figure A.2 demonstrates a scatterplot of the average accuracy of each agent on WebArena over their average costs. The cost of completing tasks shows a different trend. While the Browsing Agent requires more steps, it is much cheaper compared to the API-Based Agent and the Hybrid Agent. This is primarily because the prompts needed for Browsing Agents are much shorter. When browsing, the agent only needs instructions on how to use the web interface and

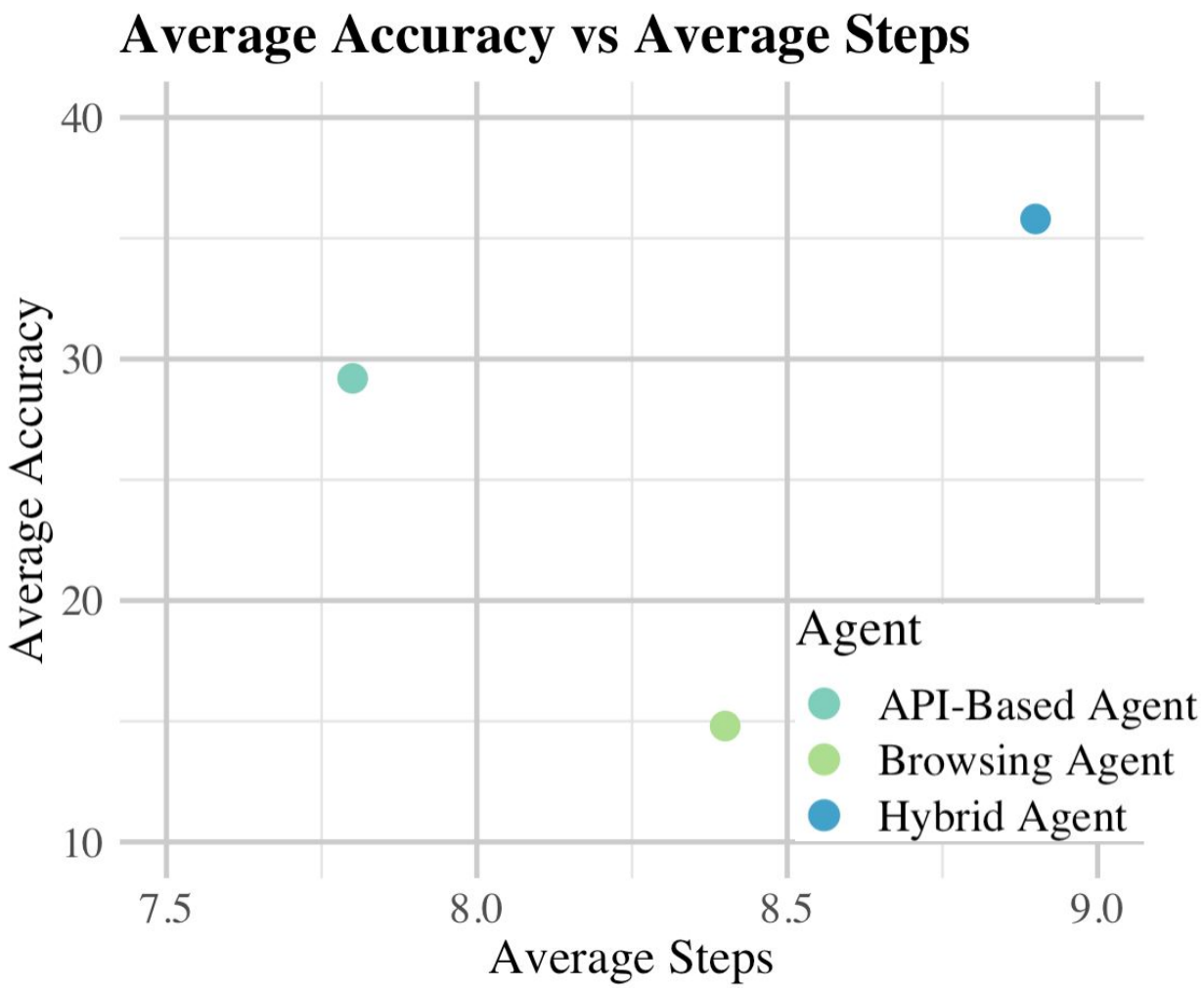


Figure A.1: Steps of agents on WebArena.

the limited action space around 14 browsing actions. In contrast, API-Based and Hybrid Agents require access to a much larger set of API calls. For example, when interacting with GitLab, the agent is provided with 988 available APIs, leading to much longer prompts and significantly increasing the cost of execution. The cost goes down when the prompt for API calling is shorter. For example, the Reddit website has the least length of API documentation, where its cost is also less than other websites. However, as visualized in Figure A.2, the accuracy of the API-Based Agent and the Hybrid Agent is much higher than the Browsing Agent, which makes the increase in cost justifiable due to the significantly improved task performance. The higher cost is offset by the agents' ability to complete tasks more accurately and efficiently. In the future, this increased cost could potentially be mitigated by methods that retrieve only relevant APIs on the fly.

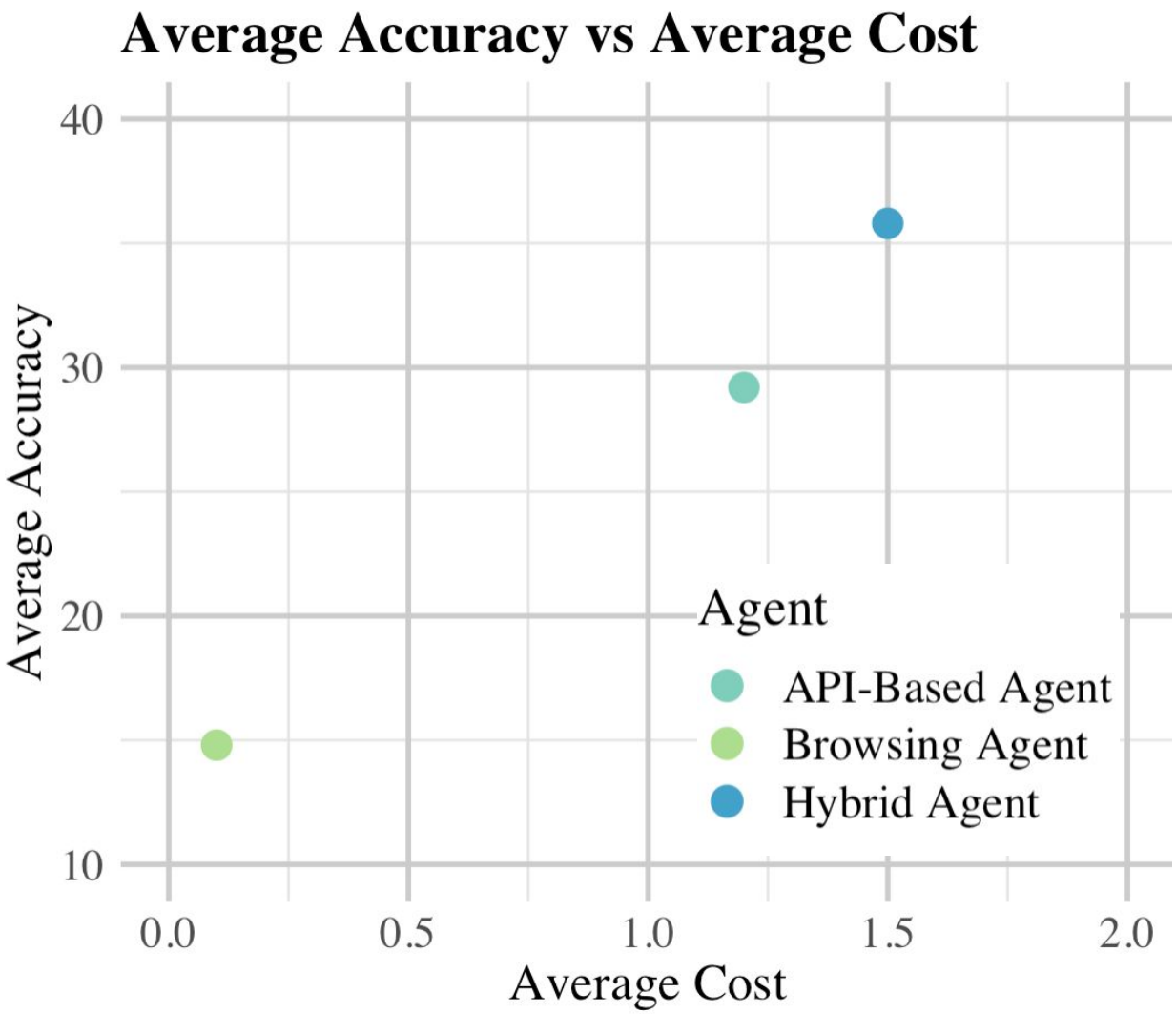


Figure A.2: Costs of agents on WebArena.

A.5 API-Based Agent Prompt

Full System Prompt

```
Full System Prompt = System Prefix + API Prompt + System Suffix
```

System Prefix

You are an AI assistant that performs tasks on the websites. You should give helpful, detailed, and polite responses to the user's queries.
You have the ability to call site-specific APIs using Python, or browse the website directly.

API Prompt

To call APIs, you can use an interactive Python (Jupyter Notebook) environment, executing code with <execute_ipython>.

```
<execute_ipython>
print(``Hello World'')
</execute_ipython>
```

This can be used to call the Python requests library, which is already installed for you. Here are some hints about effective API usage:

- It is better to actually view the API response and ensure the relevant information is correctly extracted and utilized before attempting any programmatic parsing.
- Make use of HTTP headers when making API calls, and be careful of the input parameters to each API call.
- Be careful about pagination of the API response, the response might only contain the first few instances, so make sure you look at all instances.

The user will provide you with a list of API calls that you can use.

System Suffix

The information provided by the user might be incomplete or ambiguous. For example, if I want to search for ``xyz'', then ``xyz'' could be the name of a product, a user, or a category on the site. In these cases, you should attempt to evaluate all potential cases that the user might be indicating and be careful about nuances in the user's query. Also, do NOT ask the user for any clarification, they cannot clarify anything and you need to do it yourself.
When you think you successfully finished the task, first respond with `Finish[answer]` where you include *only* your answer to the question `[]` if the user asks for an answer, make sure you should only include the answer to the question but not any additional explanation, details, or commentary unless specifically requested.
After that, when you responded with your answer, you should respond with `<finish></finish>`.
Then finally, to exit, you can run
`<execute_bash>`
`exit()`
`</execute_bash>`
Your responses should be concise. The assistant should attempt fewer things at a time instead of putting too many commands OR too much code in one `execute` block.
Include AT MOST ONE `<execute_ipython>`, `<execute_browse>`, or `<execute_bash>` per response.
IMPORTANT: Execute code using `<execute_ipython>`, `<execute_bash>`, or `<execute_browse>` whenever possible.
Below are some examples:
— START OF EXAMPLE —
`Examples`
— END OF EXAMPLE —
Now, let's start!

Initial User Prompt

Think step by step to perform the following task related to gitlab. Answer the question: `***Example WebArena Intent***`
The site URL is `Example Site URL`, use this instead of the normal site URL.
For API calling, use this access token: `Example Access Token`.
My username on this website is `Example Username`.
Below is the list of all APIs you can use and their descriptions:
`Example API Documentation.`
Note: Before actually using a API call, *you should call the `get_api_documentation` function in the `utils` module to get detailed API documentation of the API.* For example, if you want to use the API `GET /api/v4/projects/id/repository/commits`, you should first do:
`<execute_ipython>`
`from utils import get_api_documentation`
`get_api_documentation(''GET /api/v4/projects/{id}/repository/commits'')`
`</execute_ipython>`
This will provide you with detailed descriptions of the input parameters and example output jsons.

A.6 Hybrid Agent Prompt

Full System Prompt

```
Full System Prompt = System Prefix + API Prompt + Browsing Prompt +
System Suffix
```

System Prefix

You are an AI assistant that performs tasks on the websites. You should give helpful, detailed, and polite responses to the user's queries.
You have the ability to call site-specific APIs using Python, or browse the website directly.
IMPORTANT: In general, you must always first try to use APIs to perform the task; however, you should use web browsing when there is no useful API available for the task.
IMPORTANT: After you tried out using APIs, you must use web browsing to navigate to some URL containing contents that could verify whether the results you obtained by API calling is correct.

API Prompt

To call APIs, you can use an interactive Python (Jupyter Notebook) environment, executing code with <execute_ipython>.

```
<execute_ipython>
print(``Hello World!'')
</execute_ipython>
```

This can be used to call the Python requests library, which is already installed for you. Here are some hints about effective API usage:

- It is better to actually view the API response and ensure the relevant information is correctly extracted and utilized before attempting any programmatic parsing.
- Make use of HTTP headers when making API calls, and be careful of the input parameters to each API call.
- Be careful about pagination of the API response, the response might only contain the first few instances, so make sure you look at all instances.

The user will provide you with a list of API calls that you can use.

Browsing Prompt

You can browse the Internet by putting special browsing commands within <execute_browse> and </execute_browse> (in Python syntax).
For example to select the option blue from the dropdown menu with bid 12, and click on the submit button with bid 51:

```
<execute_browse>
select_option(``12'', ``blue'')
click(``51'')
</execute_browse>
```

The following actions are available:

```
def goto(url: str):
  ``````Navigate to the specified URL.
  Examples:
    goto(``http://www.example.com'')
  ''''''

def go_back():
  ``````Navigate back to the previous page.
  Examples:
    go_back()
  ''''''

def go_forward():
  ``````Navigate forward to the next page.
  Examples:
    go_forward()
  ''''''

def scroll(delta_x: float, delta_y: float):
  ``````Scroll the page by the specified amount.
  Examples:
    scroll(0, 200)
    scroll(-50.2, -100.5)
  ''''''

def fill(bid: str, value: str):
  ``````Fill the input field with the specified value.
  Examples:
    fill(``237'', ``example value'')
    fill(``45'', ``multi-line example'')
    fill(``a12'', ``example with ``quotes'''')
  ''''''
```
```

Browsing Prompt - Continued

```
def select_option(bid: str, options: str | list[str]):
  ``````Select an option from a dropdown menu.
  Examples:
    select_option(``48'', ``blue'')
    select_option(``48'', [``red'', ``green'', ``blue''])
  ''''''

def focus(bid: str):
  ``````Focus on an element.
  Examples:
    focus(``b455'')
  ''''''

def click(bid: str, button: Literal[``left'', ``middle'',
``right''] = ``left'', modifiers: list[typing.Literal[``Alt'',
``Control'', ``Meta'', ``Shift'']] = []):
  ``````Click on an element with the specified button and
modifiers.
  Examples:
    click(``51'')
    click(``b22'', button=``right'')
    click(``48'', button=``middle'', modifiers=[``Shift''])
  ''''''
```
```

Browsing Prompt - Continued

```
def dblclick(bid: str, button: Literal[``left'', ``middle'',
``right''] = ``left'', modifiers: list[typing.Literal[``Alt'',
``Control'', ``Meta'', ``Shift'']] = []):
  ``````Double-click on an element with the specified button and
modifiers.
  Examples:
    dblclick(``12'')
    dblclick(``ca42'', button=``right'')
    dblclick(``178'', button=``middle'', modifiers=[``Shift''])
  ''''''

def hover(bid: str):
  ``````Hover over an element.
  Examples:
    hover(``b8'')
  ''''''

def press(bid: str, key_comb: str):
  ``````Press a key combination on an element.
  Examples:
    press(``88'', "Backspace")
    press(``a26'', ``Control+a'')
    press(``a61'', ``Meta+Shift+t'')
  ''''''

def clear(bid: str):
  ``````Clear the input field.
  Examples:
    clear(``996'')
  ''''''

def drag_and_drop(from_bid: str, to_bid: str):
  ``````Drag and drop an element to another element.
  Examples:
    drag_and_drop(``56'', ``498'')
  ''''''

def upload_file(bid: str, file: str | list[str]):
  ``````Upload a file to the specified element.
  Examples:
    upload_file(``572'', ``my_receipt.pdf'')
    upload_file(``63'', [``/home/bob/Documents/image.jpg'',
``/home/bob/Documents/file.zip''])
  ''''''
```

System Suffix

The information provided by the user might be incomplete or ambiguous. For example, if I want to search for ``xyz'', then ``xyz'' could be the name of a product, a user, or a category on the site. In these cases, you should attempt to evaluate all potential cases that the user might be indicating and be careful about nuances in the user's query. Also, do NOT ask the user for any clarification, they cannot clarify anything and you need to do it yourself.
When you think you successfully finished the task, first respond with `Finish[answer]` where you include *only* your answer to the question `[]` if the user asks for an answer, make sure you should only include the answer to the question but not any additional explanation, details, or commentary unless specifically requested.
After that, when you responded with your answer, you should respond with `<finish></finish>`.
Then finally, to exit, you can run
`<execute_bash>`
`exit()`
`</execute_bash>`
Your responses should be concise. The assistant should attempt fewer things at a time instead of putting too many commands OR too much code in one `execute` block.
Include AT MOST ONE `<execute_ipython>`, `<execute_browse>`, or `<execute_bash>` per response.
IMPORTANT: Execute code using `<execute_ipython>`, `<execute_bash>`, or `<execute_browse>` whenever possible.
Below are some examples:
— START OF EXAMPLE —
`Examples`
— END OF EXAMPLE —
Now, let's start!

Initial User Prompt

Think step by step to perform the following task related to gitlab. Answer the question: ***Example WebArena Intent***
The site URL is `Example Site URL`, use this instead of the normal site URL.
For API calling, use this access token: `Example Access Token`.
For web browsing, You should start from the URL `Example Start URL`, and this webpage is already logged in and opened for you.
My username on this website is `Example Username`.
Below is the list of all APIs you can use and their descriptions:
`Example API Documentation.`
Note: Before actually using a API call, *you should call the `get_api_documentation` function in the `utils` module to get detailed API documentation of the API.* For example, if you want to use the API `GET /api/v4/projects/id/repository/commits`, you should first do:

```
<execute_ipython>
from utils import get_api_documentation
get_api_documentation(``GET /api/v4/projects/{id}/repository/commits'')
</execute_ipython>
```

This will provide you with detailed descriptions of the input parameters and example output jsons.
IMPORTANT: In general, you must always first try to use APIs to perform the task; however, you should use web browsing when there is no useful API available for the task. IMPORTANT: After you tried out using APIs, you must use web browsing to navigate to some URL containing contents that could verify whether the results you obtained by API calling is correct.

Appendix B

Appendix for Chapter 4

B.1 VISUALPUZZLES Statistics

B.1.1 Breakdown of Statistics of VISUALPUZZLES

Table B.1 shows a breakdown of statistics of VISUALPUZZLES questions.

Reasoning Category	Image Options			Text Options			Total
	Easy	Medium	Hard	Easy	Medium	Hard	
Algorithmic	21	8	0	124	100	9	262
Analogical	120	81	10	0	0	0	211
Deductive	29	24	2	45	79	21	200
Inductive	7	70	127	3	2	0	209
Spatial	123	41	6	61	52	3	286
Total	300	224	145	233	233	33	1168

Table B.1: Number of questions in each reasoning category, option types, and difficulty levels.

B.1.2 Data Sources

- Chinese Civil Service Examination (中国国家公务员考试) [1] (224 puzzles): we manually translated questions from this exam to English from Chinese.
- Textbooks (210 puzzles): we carefully collected and re-purposed questions from online resources and textbooks.
- Smart-101 [25] (247 puzzles): we carefully selected images from this benchmark and synthesized new questions.

[1] https://en.wikipedia.org/wiki/Civil_service_of_the_People%27s_Republic_of_China#Examinations.

- MATH-Vision [133] (293 puzzles): we carefully selected and re-purposed questions from this benchmark.
- VASR [13] (194 puzzles): we carefully selected questions from this benchmark.

B.2 Model Evaluation Setup

Model Evaluation Prompt with Chain-of-Thought

Solve the multiple-choice question and then answer with the option letter from the given choices. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of options. Think step by step before answering.

Model Evaluation Prompt w/n Chain-of-Thought

Answer the question with the option's letter from the given choices directly.

B.3 Human Annotation Setup

B.3.1 Difficulty Labeling

Each question was also carefully assigned a difficulty label from easy, medium, or hard, based on the cognitive load required for reasoning.

- **Easy Level** questions could be solved by the annotator in less than one minute.
- **Medium Level** questions could be solved by the annotator in one to three minutes.
- **Hard Level** questions require the annotator more than five minutes to solve or quit solving.

Annotation Guideline for Puzzle Difficulty

Try to solve the puzzle first. You need to measure the time you attempted to solve each puzzle. Then, select from Easy, Medium, or Hard based on the time required.
- Easy Level: You can solve or answer the question within 1 minute. This level of puzzles should require minimal reasoning.
- Medium Level: You can solve or answer the question within 1-3 minutes. This level of puzzles should demand moderate reasoning.
- Hard Level: You can / cannot solve this question with more than 5 minutes. This level of puzzles should involve significant / multi-step reasoning.

B.3.2 Reasoning Category Labeling

Annotation Guideline for Puzzle Reasoning Category

Assign the category that *best* describes the primary type of reasoning or logic required for each puzzle:
- Algorithmic Reasoning: Involves following or devising a step-by-step procedure or rule-based process.
- Analogical Reasoning: Requires identifying relationships by comparison between pairs of entities.
- Deductive Reasoning: Involves deriving specific conclusions from general or given premises.
- Inductive Reasoning: Focuses on generalizing a rule or pattern from specific instances.
- Spatial Reasoning: Involves visualizing and manipulating shapes, distances, or orientations.

B.4 Full Results

B.4.1 Full Results w/ CoT

B.4.2 Full Results w/n CoT

B.5 Knowledge Checklist

B.5.1 Knowledge Checklist Generation

Prompt to Generate Knowledge Checklist Questions

You are an exam writer. You are now writing a knowledge test. You are given a question (Question) regarding an image and its standard solution (Solution), your task is to write free response questions that test on individual knowledge required in answering the question correctly.

You should follow these steps to complete the task:
1. explicitly analyze the given image, Question, and Solution
2. explicitly list out the individual knowledge concepts required to reach Solution.
3. write free response questions to test on the definition of each concept listed. Your generated questions should not include details of the given Question. Note that you need to provide answer keys to these questions too.
4. format the free response questions in json format.

Question: question
Solution: answer

Model	Algorithmic	Analogical	Deductive	Inductive	Spatial	Overall
Random Choice	25.0	25.0	25.0	25.0	25.0	25.0
Human (95th Percentile)	100.0	100.0	100.0	81.6	100.0	89.3
Human (50th Percentile)	88.0	66.0	80.0	50.0	90.0	75.0
Human (5th Percentile)	68.1	25.0	37.0	0.0	59.1	57.5
Proprietary Models						
o4-mini	65.3	68.7	75.5	33.0	45.5	57.0
o3	64.5	68.3	69.5	27.3	42.7	54.0
o1	63.7	68.3	67.5	29.2	34.3	51.8
GPT-4o	49.2	58.3	49.0	27.3	26.2	41.3
Gemini-2.5-pro	60.0	64.0	60.0	29.7	36.4	49.5
Gemini-2.0-flash	55.3	58.8	57.0	24.4	31.8	45.0
Gemini-2.0-flash-thinking	46.6	70.1	49.0	24.9	25.5	42.2
Gemini-1.5-Pro	53.4	57.4	58.5	26.3	32.5	45.0
Claude-3.7-Sonnet	64.5	48.3	65.0	26.8	37.4	48.3
Claude-3.7-Sonnet-thinking	67.2	44.1	61.5	31.1	37.1	48.2
Claude-3.5-Sonnet	53.4	47.9	51.5	25.4	34.3	42.4
Open Models						
LLaVA-1.5-7B	23.3	21.8	36.0	20.6	19.2	23.7
LLaVA-1.5-13B	24.8	21.8	23.0	25.4	25.5	24.2
LLaVA-1.6-7B	27.5	23.7	30.0	22.5	21.3	24.8
LLaVA-1.6-13B	25.2	25.6	27.0	27.3	23.4	25.5
LLaVA-1.6-34B	29.4	28.0	43.0	24.9	25.5	29.7
LLaVA-OV-0.5B	21.0	26.1	30.5	22.5	25.2	24.8
LLaVA-OV-7B	27.9	26.1	36.5	23.4	25.5	27.7
LLaVA-OV-72B	34.7	26.5	37.0	27.3	28.7	30.8
Llama-3.2-11B-Vision-Instruct	31.0	30.8	39.0	21.1	26.2	29.4
Llama-3.2-90B-Vision-Instruct	45.0	23.2	43.0	26.3	31.5	34.1
Qwen-VL	21.4	31.3	25.0	26.3	24.1	25.3
Qwen2-VL-72B	41.6	28.4	39.5	22.5	29.0	32.4
QvQ-72B-Preview	43.1	45.5	48.0	27.3	27.6	37.8
Qwen2-VL-2B-Instruct	26.0	26.1	24.5	27.8	25.5	26.0
Qwen2-VL-7B-Instruct	36.3	21.8	38.5	20.6	22.7	27.9
Qwen2-VL-72B-Instruct	39.9	33.5	45.2	23.5	32.4	34.9
Qwen2.5-VL-3B-Instruct	35.1	27.5	44.5	25.8	24.8	31.2
Qwen2.5-VL-7B-Instruct	40.5	26.6	39.0	24.0	29.7	32.1
Qwen2.5-VL-72B-Instruct	53.4	46.9	58.0	25.8	29.5	42.3
Cambrian-8B	31.3	24.2	36.0	24.0	29.0	28.9
Cambrian-13B	24.8	25.6	39.5	24.4	21.0	26.5
Pangea-7B	30.5	28.9	35.0	24.4	25.2	28.6

Table B.2: Performance (%) of various models with Chain of Thoughts (CoT) on VISUALPUZZLES.

Model	Algorithmic	Analogical	Deductive	Inductive	Spatial	Overall
Random Choice	25.0	25.0	25.0	25.0	25.0	25.0
Human (95th Percentile)	100.0	100.0	100.0	81.6	100.0	89.3
Human (50th Percentile)	88.0	66.0	80.0	50.0	90.0	75.0
Human (5th Percentile)	68.1	25.0	37.0	0.0	59.1	57.5
Proprietary Models						
GPT-4o	40.8	34.1	40.5	24.9	29.7	34.0
Gemini-2.0-flash	57.6	41.7	58.0	23.0	35.7	43.2
Gemini-1.5-Pro	51.2	46.5	54.0	24.9	29.4	40.8
Open Models						
LLaVA-1.5-7B	24.4	24.7	34.5	26.8	25.5	26.9
LLaVA-1.5-13B	24.4	26.1	33.5	26.3	28.3	27.6
LLaVA-1.6-7B	27.5	25.1	32.5	24.9	27.3	27.4
LLaVA-1.6-13B	21.4	24.7	29.5	28.2	23.1	25.0
LLaVA-1.6-34B	31.3	27.3	43.0	24.4	27.6	29.8
LLaVA-OV-0.5B	24.4	25.6	37.5	24.9	25.5	27.2
LLaVA-OV-7B	27.5	28.0	40.5	24.4	28.0	29.4
LLaVA-OV-72B	31.7	23.6	45.0	21.3	24.6	28.8
Llama-3.2-11B-Vision-Instruct	27.5	24.2	31.0	26.3	27.6	27.3
Llama-3.2-90B-Vision-Instruct	38.2	22.3	44.5	25.8	33.6	33.1
Qwen-VL	23.7	26.5	29.5	27.8	26.6	26.6
Qwen2-VL-72B	38.9	28.4	43.0	20.6	29.0	32.0
QvQ-72B-Preview	44.8	43.6	44.0	26.8	30.8	37.8
Qwen2-VL-2B-Instruct	31.7	29.4	40.5	23.9	31.5	31.3
Qwen2-VL-7B-Instruct	33.6	24.2	46.0	22.5	26.2	30.2
Qwen2-VL-72B-Instruct	40.5	30.3	46.0	25.4	29.4	34.2
Qwen2.5-VL-3B-Instruct	36.3	26.1	47.0	25.8	22.4	31.0
Qwen2.5-VL-7B-Instruct	38.2	23.7	51.5	24.9	31.1	33.7
Qwen2.5-VL-72B-Instruct	43.1	40.3	51.5	25.4	33.7	38.6
Cambrian-8B	25.2	20.4	35.0	23.0	20.6	24.5
Cambrian-13B	23.3	28.0	36.5	24.9	26.2	27.4
Pangea-7B	32.4	23.7	38.5	28.7	32.5	31.3

Table B.3: Performance (%) of various models with Multiple Choice Direct prompting on VISUALPUZZLES.

B.5.2 Example Knowledge Checklist Question

Example Knowledge Checklist Question (MMMU)

- Question: Explain the Arbitrage Pricing Theory (APT) model and its purpose in finance.
- Answer: The Arbitrage Pricing Theory (APT) model is a financial theory that estimates the expected return on an asset based on the asset's sensitivity to various macroeconomic factors. It is used to determine the fair price of an asset by considering multiple factors that could affect its return, as opposed to relying on a single market index as in the Capital Asset Pricing Model (CAPM).

Example Knowledge Checklist Question (VISUALPUZZLES)

- Question: What is the definition of distance in a geometric context?
- Answer: Distance in a geometric context refers to the measurement of space between two points.

B.5.3 Knowledge Checklist Human Annotation

We asked two human annotators to manually verify and correct the knowledge checklist questions and gave them the following instructions. The inter-annotator agreement rate is 87.8%.

Human Annotation Instructions

You are given a json file, where each item contains the following elements:
- Question: a multiple-choice question.
- Answer: the answer to the question with an optional explanation.
- Knowledge Concept Checklist: a list of question-answer pairs, where each question in the list is intended to represent a distinct knowledge concept necessary for solving the Question.

You task is to annotate the knowledge concept checklists generated by a model. You should carefully evaluate each question-answer pair based on the following criteria:
1. Necessity: Is the question genuinely necessary for solving the problem? If not, then remove the question.
2. Repetition: Check if any questions are repetitive or duplicate existing questions within the list. If the question is repetitive or duplicate, then remove the question.
3. Completeness: Ensure no critical knowledge concepts required to solve the problem are missing, and identify if any additional important questions should have been included.
4. Correctness: Verify whether the provided answers are accurate. Revise the checklist in case of incorrect checklist QA pairs.
5. Knowledge v.s. Skills: Ensure each question explicitly evaluates a knowledge concept rather than testing skills or problem-solving techniques. Remove any questions that primarily evaluate skills instead of knowledge.

B.6 Reasoning Complexity

Instruction Prompt to Solve Questions in Detailed Steps

< $Question$ >< $Image$ >
Solve this question with First Order Logic. Write out each thinking step explicitly, do not skip steps.
In your response, begin each step with ___STEP_START___
step < $step\_num$ >

B.7 Comparison with Other Benchmarks

Dataset	Size	Reasoning Load	Knowledge Requirement	% Easy Words	Question Type	Answer Type	
LogiQA	0.7K	Heavy	Light	52.0	Text	Text
GSM8K	8.5K	Heavy	Heavy	54.0	Text	Text
WikiDiverse	0.8K	Light	Heavy	35.8	Image+Text	Text
MathVista	6.1K	Heavy	Heavy	51.9	Image+Text	Text
MMMU	11.5K	Heavy	Heavy	46.4	Image+Text	Text
MATH-Vision	3.0K	Heavy	Heavy	53.8	Image+Text	Image+Text
MathVerse	2.6K	Heavy	Heavy	38.2	Image+Text	Text
LogicBench	1.5K	Heavy	Light	53.6	Text	Text
LogicVista	0.4K	Heavy	Heavy	41.2	Image+Text	Image
NaturalBench	10K	Light	Light	52.5	Image+Text	Text
VISUALPUZZLES	1.2K	Heavy	Light	**54.1**	Image+Text	Image+Text

Table B.4: Comparison of other existing benchmarks with VISUALPUZZLES

Figure B.1 provides a comparative analysis between VISUALPUZZLES and several widely-used benchmarks for multimodal reasoning, visualizing the knowledge requirement and reasoning complexity of each benchmark. VISUALPUZZLES has high reasoning complexity and low knowledge requirement, with an aim to disentangle multimodal reasoning from domain-specific knowledge to evaluate general reasoning abilities in non-expert settings.

Table B.5 compare the performance of various model families across MathVista, MMMU, and VISUALPUZZLES. Both MathVista and MMMU are benchmarks that have a heavy emphasis on both knowledge and reasoning, whereas VISUALPUZZLES assess models on domain-disentangled multimodal reasoning alone. We found that success on knowledge-intensive multimodal reasoning benchmarks as MathVista and MMMU does not always carry over to VISUALPUZZLES that emphasize reasoning rather than extensive pre-trained knowledge.

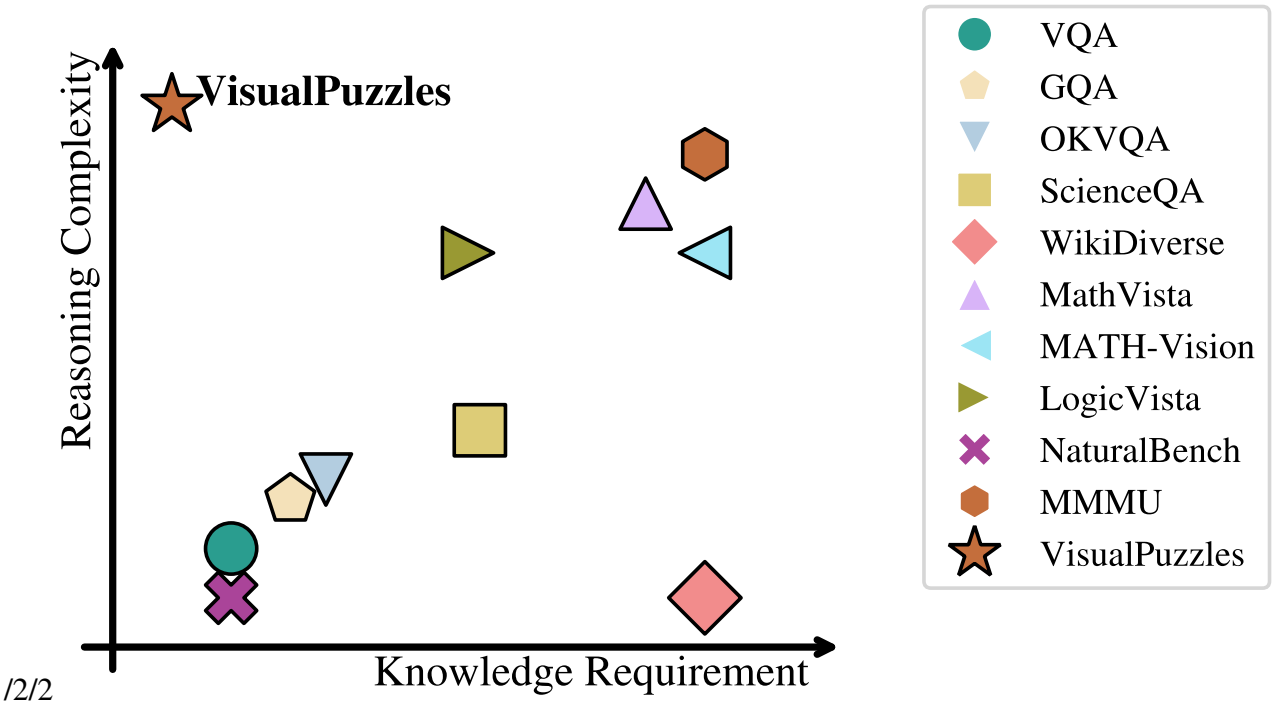


Figure B.1: Comparison between VISUALPUZZLES and several widely-used benchmarks.

B.8 Additional Analysis

B.8.1 Proprietary V.S. Open Models

From Table 4.2, proprietary models (e.g., o4-mini and Claude-3.7-Sonnet) consistently achieve higher overall accuracy than most open-source models on VISUALPUZZLES. However, some open models also show competitive or even higher performance in both the overall accuracy and specific reasoning categories. For instance, Qwen2.5-VL-72B-Instruct demonstrates higher performance than GPT-4o on algorithmic reasoning, deductive reasoning, spatial reasoning, and overall accuracy. This indicates that while proprietary models currently have leading performance, open models are also rapidly improving on multimodal reasoning capabilities.

B.8.2 Reasoning Category and Difficulty Levels

Figure B.3 and Figure B.2 present complementary views of human accuracy against three representative models: o1 (one of the best-performing proprietary models), Qwen2.5-VL-72B-Instruct (the strongest Qwen-based open model), and Llama-3.2-90B-Vision-Instruct (the strongest Llama-based open model). Specifically, Figure B.2 compares performance across difficulty levels for each reasoning category, while Figure B.3 compares performance across categories within each difficulty level.

Humans consistently outperform all models across categories and difficulty levels, often by

Model	MathVista	MMMU	VISUALPUZZLES
Human	60.3	88.6	80.1
o1	73.9	78.2	51.8
GPT-4o	63.8	69.1	41.1
Gemini-2.0-Flash	-	71.7	45.0
Gemini-1.5-Pro	63.9	62.2	45.4
Claude-3.5-Sonnet	67.7	68.3	42.4
Claude-3.7-Sonnet	-	71.8	48.3
Claude-3.7-Sonnet (Thinking)	-	75.0	48.3
LLaVA-1.5-7B	-	36.2	26.9
LLaVA-1.5-13B	27.6	36.4	27.6
LLaVA-NeXT-7B	35.8	34.6	27.4
LLaVA-NeXT-13B	36.2	35.3	25.3
LLaVA-NeXT-34B	46.5	51.1	29.8
LLaVA-OV-0.5B	34.8	31.4	27.2
LLaVA-OV-7B	63.2	48.8	29.4
LLaVA-OV-72B	67.5	56.8	31.8
Llama-3.2-11B-Vision-Instruct	51.5	50.7	29.4
Llama-3.2-90B-Vision-Instruct	57.3	60.3	34.3
Qwen2-VL-72B	70.5	64.5	32.1
QvQ-72B-Preview	71.4	70.3	37.9
Qwen2-VL-2B-Instruct	43.0	41.1	31.3
Qwen2-VL-7B-Instruct	58.2	54.1	30.2
Qwen2-VL-72B-Instruct	70.5	64.5	34.9
Qwen2.5-VL-3B-Instruct	62.3	53.1	31.2
Qwen2.5-VL-7B-Instruct	68.2	58.6	33.7
Qwen2.5-VL-72B-Instruct	74.8	70.2	42.3
Cambrian-8B	49.0	42.7	28.5
Cambrian-13B	48.0	40.0	27.4

Table B.5: Comparison of other MathVista and MMMU with VISUALPUZZLES on human and SOTA models

large margins. Notably, human performance remains high and relatively stable in the algorithmic, deductive, and spatial categories, even on hard questions. While accuracy does decline in analogical and inductive reasoning as difficulty increases, humans still maintain a clear advantage over models.

In contrast, model performance declines sharply as difficulty increases, especially for open-source models. Accuracy of Llama-3.2-90B-Vision-Instruct on hard analogical tasks drops to just 10%. Even one of the strongest proprietary models, o1, while more robust, still lags significantly behind humans, particularly on analogical, inductive, and spatial tasks. On easy tasks, some models perform competitively in certain categories, but this advantage largely disappears on medium and hard questions.

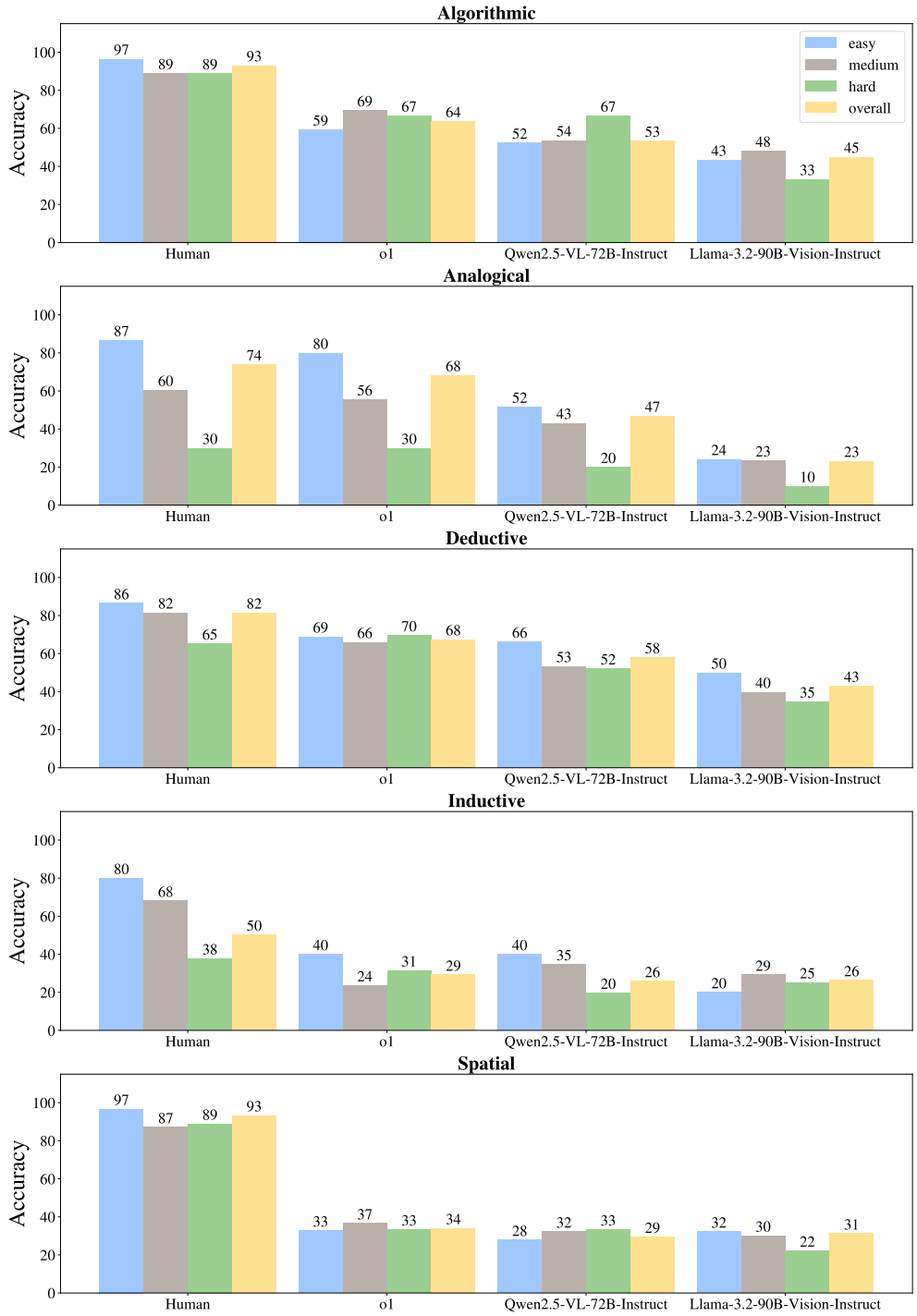


Figure B.2: Comparison of accuracy across different reasoning categories for human participants, one of the best performing proprietary models o1, the best performing Qwen-based open model Qwen2.5-VL-72B-Instruct, and the best performing Llama-based open model Llama-3.2-90B-Vision-Instruct, measured on difficulty levels.

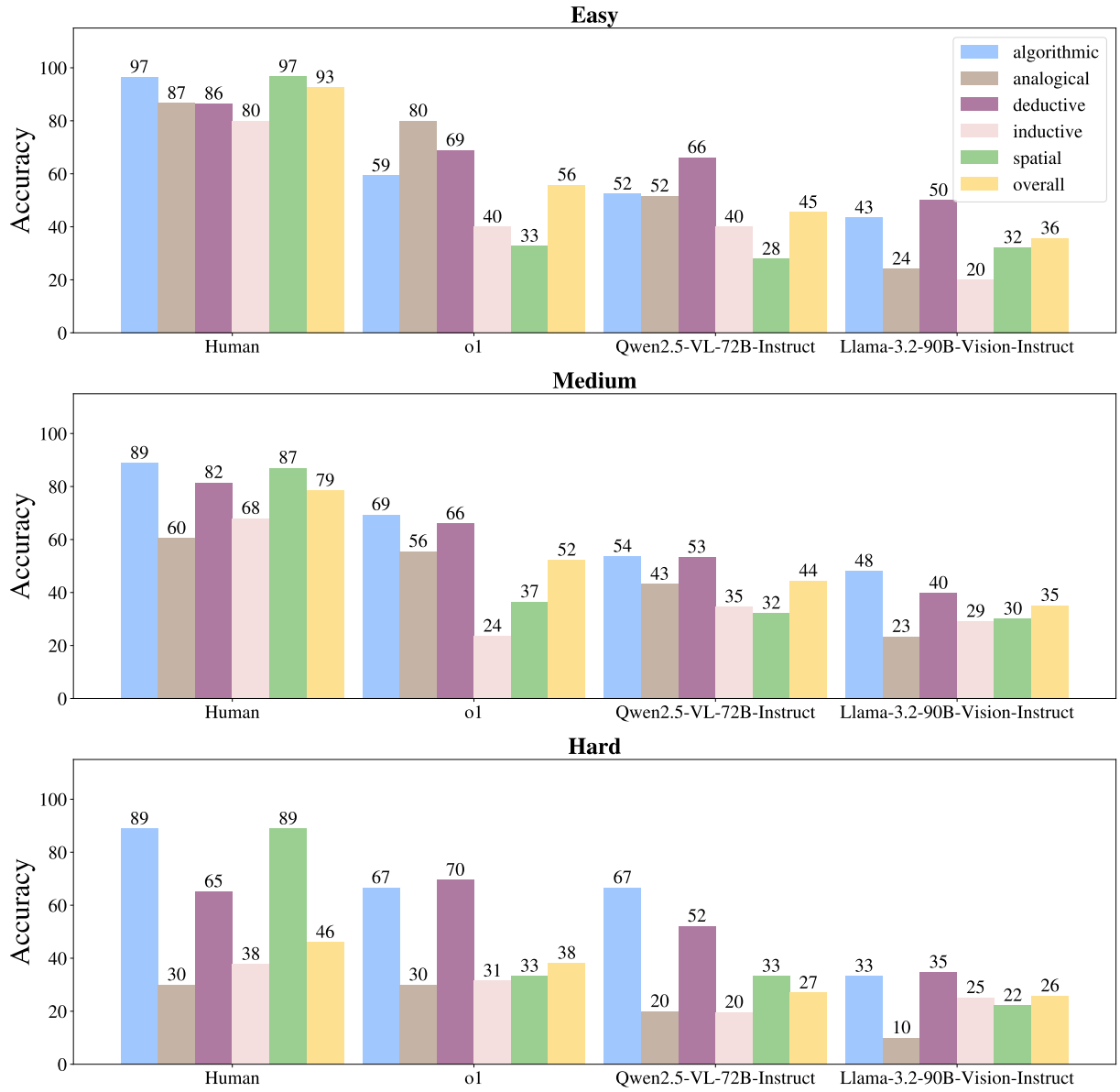


Figure B.3: Comparison of accuracy across different difficulty levels for human participants, one of the best performing proprietary models o1, the best performing Qwen-based open model Qwen2.5-VL-72B-Instruct, and the best performing Llama-based open model Llama-3.2-90B-Vision-Instruct, measured across reasoning categories.

Interestingly, these models maintain a generally stable performance on algorithmic and deductive reasoning. For o1 and Qwen2.5-VL-72B-Instruct, their performances on algorithmic reasoning even go up for more difficult tasks, whereas human performance degraded as the difficulty level increases. However, all models, including o1, perform the worse at analogical, inductive and spatial reasoning in general, especially as the difficulty level increases. This suggests

that models are relatively better at tasks requiring structured, rule-based algorithmic processing, while their performance degrades more steeply in tasks requiring relational abstraction (analogical), pattern induction (inductive), and visual understanding (spatial), particularly as the difficulty level increases. In summary, these results indicate that while some models exhibit promising performance on structured and easier reasoning tasks, multimodal models still struggle with abstract and complex reasoning, particularly when difficulty increases. Bridging the gap between model and human reasoning remains a critical challenge.

B.8.3 Option Types and Difficulty Levels

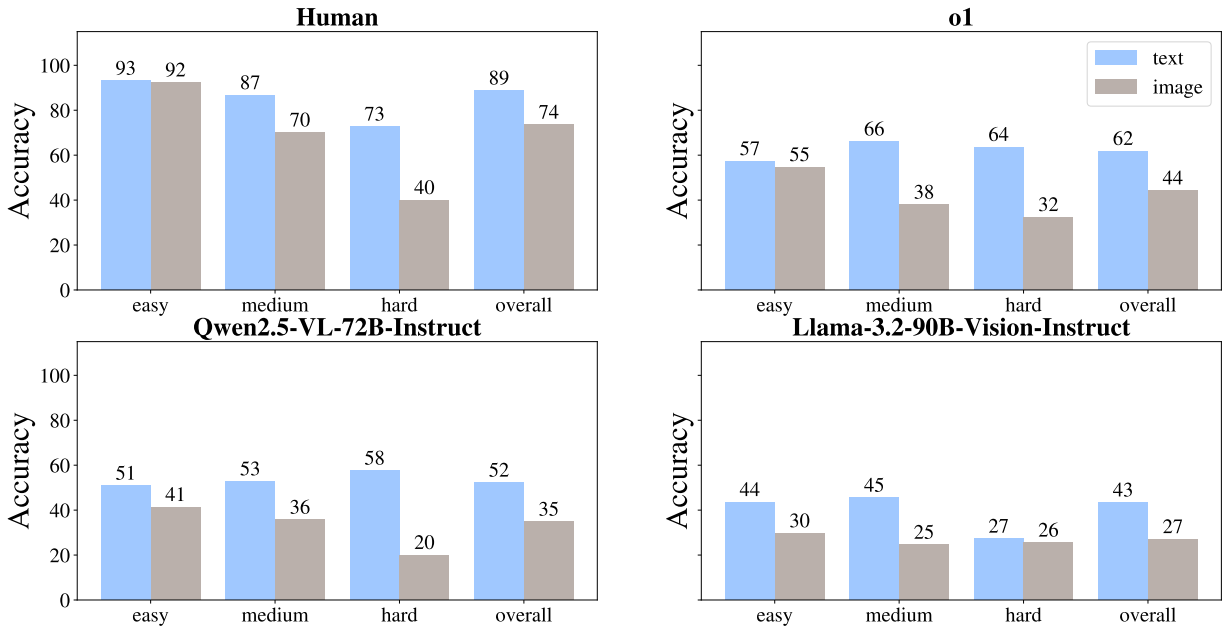


Figure B.4: Comparison of accuracy across different difficulty levels for human participants, one of the best performing proprietary model o1, the best performing Qwen-based open model Qwen2.5-VL-72B-Instruct, and the best performing Llama-based open model Llama-3.2-90B-Vision-Instruct, measured on textual v.s. visual option types.

Figure B.4 compares human accuracy against three representative models, o1 (one of the best-performing proprietary models), Qwen2.5-VL-72B-Instruct (the strongest Qwen-based open model), and Llama-3.2-90B-Vision-Instruct (the strongest Llama-based open model), across different difficulty levels, separately for textual and visual answer options.

Across all participants and models, we observe a consistent pattern: text-based options result in higher accuracy than image-based options, with the performance gap widening as task difficulty increases. This trend holds even for human participants, whose accuracy drops from 92% to 40% on visual options when moving from easy to hard tasks, compared to a much smaller drop on text-based ones (93% to 73%).

For models, the gap is even more pronounced. For instance, Qwen2.5-VL-72B-Instruct achieves 58% accuracy on hard questions with text options, but only 20% when image options are used. o1 and Llama-3.2-90B-Vision-Instruct exhibit similar drops, suggesting a broad weakness in multi-image reasoning and visual option discrimination. These findings suggest that image-based answer options introduce significant additional complexity, requiring models not just to understand the question but to reason over multiple visual cues. This capability is essential for real-world tasks such as product selection, recommendation, and visual planning, where their decision-making process often depends on comparing visual content.

However, most pretraining datasets and benchmarks have traditionally emphasized textual QA formats, with far fewer examples involving visual options or structured visual comparisons. As a result, models may lack the inductive bias or learned attention mechanisms to handle visual alternatives effectively. These results highlight an important direction for future work: expanding and diversifying training corpora to include multi-choice visual reasoning tasks, and developing architectures that are explicitly designed to process and compare visual candidates, especially under challenging conditions.

B.8.4 Case Study of Reasoning

Question: Given a pair of chairs arranged at a specific angle, select the option from the four provided that matches the angle of the chairs in the question. (**Answer**: C)

Claude-3.7-Sonnet (Without Thinking)
I need to analyze the angle relationship between the pair of chairs in the question and compare it with the four options provided.
Looking at the question image at the top:

- The first chair (on the left) is shown at an angle, with its side profile partially visible
- The second chair (on the right) is shown more from the front, facing more directly toward the viewer
- The chairs are positioned at roughly a 90-degree angle to each other

Now examining each option:

Option A: Two simple chairs with backs. The left chair is shown from a side angle, while the right chair is shown more from the front. Their orientation appears to be at approximately a 90-degree angle to each other.

Option B: Two backless chairs/stools. The left chair has an angled design and is shown from the side, while the right chair is shown more straight-on. The angle between them doesn't appear to match the 90-degree relationship in the question.

Option C: Two office/drafting chairs with wheels. Both chairs are shown from similar angles, with only slight variation in their orientation. They don't appear to have the same 90-degree relationship as the question chairs.

Option D: Two bar stools with round seats and circular bases. Both are shown from similar angles with minimal difference in orientation. They don't demonstrate the 90-degree relationship seen in the question.

Comparing all options to the question image, Option A most closely matches the angular relationship between the two chairs in the question.

Answer: A

Claude-3.7-Sonnet (With Thinking)
Let me analyze this problem carefully.

The question shows a "question pair" of chairs at the top of the image, and I need to find which of the four options (A, B, C, D) matches the angle arrangement of those chairs.

Let me examine the question pair first:

- The question pair shows two folding chairs that appear to be facing inward toward each other at approximately a 90-degree angle (the chairs are at roughly right angles to each other).
- The chairs in the question pair are shown separated by a colon ":" symbol.

Now let me examine each option:

Option A:
- Shows two simple chairs with backrests
- The chairs are facing each other at approximately a 90-degree angle
- The arrangement looks similar to the question pair in terms of the angle

Option B:
- Shows two chairs with different designs - one appears to be a more angled/tilted chair and the other is more upright
- They don't appear to be at the same angle relationship as the question pair

Option C:
- Shows two office/drafting chairs with wheels and height adjustment
- The chairs appear to be facing each other at approximately a 90-degree angle
- The arrangement looks similar to the question pair in terms of angle

Option D:
- Shows two bar-style chairs with round seats and central pedestals
- The chairs appear to be directly facing each other (180-degree relationship)
- This doesn't match the angle relationship in the question pair

Based on the angle relationship, both Options A and C show chairs at approximately a 90-degree angle to each other, similar to the question pair. However, Option A's chair style and arrangement seems closer to the question pair's folding chairs.

So, looking at the angle between the chairs, which is the key aspect of this question, Options A and C both show approximately 90-degree angles between chairs, while Option B shows chairs at different angles, and Option D shows chairs directly facing each other. Since the question specifically asks about the angle match, I need to determine which option best matches the exact angle of the question pair.

Looking closely again:

The question pair shows chairs that are facing slightly inward, at approximately 90 degrees.
Option A shows a similar angle relationship of roughly 90 degrees.
Option C also shows a similar angle relationship of roughly 90 degrees.
This is a bit challenging with the 2D representation, but I think Option A most closely matches the angle relationship in the question pair.

Answer: A

Figure B.5: Case Study showing the similarity in structure and reasoning strategy between Claude-3.7-Sonnet-Thinking and Claude-3.7-Sonnet. Similarity between model responses of these two models on VISUALPUZZLES is 0.9.

Figure B.5 shows a case study demonstrating the similarity in structure and reasoning strategy between Claude-3.7-Sonnet and Claude-3.7-Sonnet-Thinking. Average textual similarity between model responses of these two models on VISUALPUZZLES is 0.9.

B.8.5 Impact of CoT

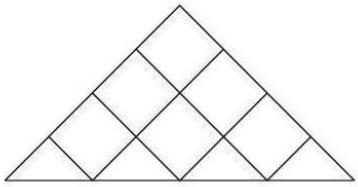
Table B.6 compares model performance under two prompting strategies: direct multiple-choice prompt vs. Chain-of-Thought (CoT) prompt. We observe that proprietary models and larger open models (≥72B) benefit from CoT, while others show little to no improvement or even a decline in performance with CoT. For instance, both GPT-4o and Qwen2.5-VL-72B-Instruct show more than 20% increases in performance when using CoT. In contrast, several smaller models, such as Qwen2-VL-2B-Instruct and Cambrian-13B, exhibit decreased accuracy with CoT prompting. These results suggest that CoT can indeed enhance the reasoning capability of larger models whereas it may introduce unnecessary complexity or confusion for smaller models and thus decreasing performance.

Model	Direct	CoT
GPT-4o	34.0	41.6
Gemini-1.5-Pro	41.0	45.1
Claude-3.5-Sonnet	40.0	42.5
Qwen2-VL-2B-Instruct	31.3	26.1
Qwen2.5-VL-7B-Instruct	33.7	32.0
Cambrian-13B	27.4	26.5
LLaVA-NeXT-34B	29.8	29.6
Qwen2.5-VL-72B-Instruct	38.6	42.3
LLama-3.2-90B-Vision-Instruct	33.3	33.9

Table B.6: Comparison of models with Direct Multiple Choice and CoT prompting.

B.9 Case Study

Algorithmic (Easy) ⊗



Question: There are seven squares in the picture. How many more triangles than squares are there in the picture?
Options: (A) 1 (B) 2 (C) 3 (D) 4

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> I need to count the number of squares and triangles in the given picture, and then find how many more triangles there are than squares.

First, let me count the squares:
1. I can see 7 small squares arranged in a triangular pattern within the large triangle.

Now, let me count the triangles:
1. There's the large outer triangle that contains everything.
2. There are small triangles formed by the division of the large triangle by the squares.

Let me count these small triangles carefully:
- Along the bottom row, I see 4 triangles (formed by the squares and the bottom edge)
- On the left side of the large triangle, I see 3 triangles (formed by the squares and the left edge)
- On the right side of the large triangle, I see 3 triangles (formed by the squares and the right edge)
- Don't forget the large outer triangle itself: 1 triangle

So in total: 4 + 3 + 3 + 1 = 11 triangles

Now I can find the difference:
Number of triangles - Number of squares = 11 - 7 = 4

So there are 4 more triangles than squares in the picture.

Looking at the options, the answer is (D) 4.
<end_thinking>
Looking at the image, I need to count the total number of squares and triangles, then find their difference.

First, I'll count the squares:
- I can see 7 small squares arranged in a triangular pattern inside the large triangle.

Next, I'll count the triangles:
1. There is the large outer triangle that contains everything.
2. Along the bottom edge of the large triangle, there are 4 small triangles.
3. Along the left edge of the large triangle, there are 3 small triangles.
4. Along the right edge of the large triangle, there are 3 small triangles.

Total number of triangles = 1 + 4 + 3 + 3 = 11 triangles

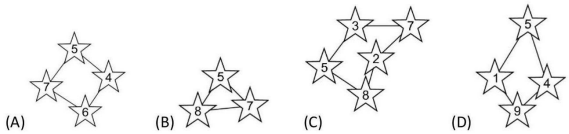
Now I can calculate how many more triangles than squares there are:
11 triangles - 7 squares = 4 more triangles

Answer: D

Error Category: Reasoning Error
Error Reason: The model gets the answer wrong because it miscounts the number of triangles, specifically double-counting the triangles formed by small triangles and squares. The response incorrectly conclude there are 11 triangles, but a careful manual inspection shows only 10 distinct triangles, making the correct difference 3, not 4.

Figure B.6: A sample error case of Algorithmic Reasoning (difficulty: Easy).

Algorithmic (Easy)



Question: In the Kangaroo constellation, all stars have a number greater than 3 and their sum is 20. Which option shows the Kangaroo constellation?

Ground Truth: (B)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** The question asks about the Kangaroo constellation, which has the following properties:
1. All stars have a number greater than 3
2. The sum of all numbers is 20

Let me check each option to see which one satisfies these conditions:

Option (A):
Stars have values: 5, 4, 7, 6
- Are all numbers greater than 3? Yes (5, 4, 7, 6 are all > 3)
- Sum: 5 + 4 + 7 + 6 = 22
This doesn't match the required sum of 20.

Option (B):
Stars have values: 5, 8, 7
- Are all numbers greater than 3? Yes (5, 8, 7 are all > 3)
- Sum: 5 + 8 + 7 = 20
This matches both conditions: all numbers are greater than 3, and they sum to 20.

Option (C):
Stars have values: 5, 3, 7, 2, 8
- Are all numbers greater than 3? No (3 and 2 are not > 3)
- Sum: 5 + 3 + 7 + 2 + 8 = 25
This doesn't satisfy the first condition about all stars having numbers greater than 3.

Option (D):
Stars have values: 5, 1, 4, 9
- Are all numbers greater than 3? No (1 is not > 3)
- Sum: 5 + 1 + 4 + 9 = 19
This doesn't satisfy either condition.

So, only Option (B) satisfies both conditions of the Kangaroo constellation.
<end_thinking>
Solving the Kangaroo Constellation Problem

I need to identify the constellation where all stars have numbers greater than 3 and sum to 20.

Let me examine each option:

Option (A)
Stars: 5, 4, 7, 6
- All numbers > 3? Yes
- Sum: 5 + 4 + 7 + 6 = 22
Doesn't match the required sum of 20.

Option (B)
Stars: 5, 8, 7
- All numbers > 3? Yes
- Sum: 5 + 8 + 7 = 20
Matches both requirements!

Option (C)
Stars: 5, 3, 7, 2, 8
- All numbers > 3? No (3 and 2 are ≤ 3)
- Sum: 5 + 3 + 7 + 2 + 8 = 25
Doesn't satisfy the first condition.

Option (D)
Stars: 5, 1, 4, 9
- All numbers > 3? No (1 is ≤ 3)
- Sum: 5 + 1 + 4 + 9 = 19
Doesn't satisfy either condition.

Only option (B) satisfies both requirements: all stars have numbers greater than 3, and their sum is exactly 20.

Answer: B

Figure B.7: A sample correct case of Algorithmic Reasoning (difficulty: Easy).

Algorithmic (Medium) ⊗

Question: How many left hands appear in the image?
Options: (A) 1 (B) 8 (C) 4 (D) 5

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*>
Looking at the image, I can see approximately 10-12 hand icons/emoji in a row. They're all yellow (with one that appears to be skin-colored), and they show various hand gestures.

To determine which ones are left hands, I need to look at the thumb position. In a left hand, the thumb would typically be on the right side when the palm is facing forward.

Let me go through each hand from left to right:

1. Thumbs down gesture - appears to be a left hand with thumb on right side
2. Fist or closed hand - orientation is harder to determine but seems to be a left hand
3. Open hand or waving hand - appears to be a left hand
4. Hand pointing or gesturing - appears to be a left hand
5. Open hand with fingers spread - appears to be a left hand
6. Fist with extended thumb (like "thumbs up" but horizontal) - appears to be a right hand (thumb on left)
7. Open hand with fingers spread - appears to be a left hand
8. Hand making "OK" gesture - appears to be a left hand
9. Victory/peace sign hand - appears to be a left hand
10. Ear or side of face (not a hand)
11. Hand pointing or gesturing - appears to be a right hand

So counting the left hands, I see approximately 8 left hands. This would correspond to option (B) 8.

However, I want to be careful because determining left vs. right can sometimes be ambiguous in emoji or simplistic hand representations. Let me double check my count.

Looking again, I see the thumbs down, several open hands, some gesture hands, and a few that are definitely left hands based on thumb position. The total count of left hands appears to be around 8, though there's some ambiguity with a couple of the representations.

Based on my analysis, the answer appears to be (B) 8.
<*end_thinking*>
To solve this question, I need to examine the image carefully and identify which hands are left hands.

Looking at the row of yellow hand emojis, I need to determine whether each is a left or right hand. For hand emojis, I can identify a left hand by noticing the thumb position - on a left hand, the thumb will typically appear on the right side when the palm is facing forward.

Going through each emoji from left to right:
1. Thumbs down - this is a left hand (thumb is on the right)
2. Fist - appears to be a left hand
3. Waving hand - this is a left hand
4. Pointing hand - this is a left hand
5. Open hand with fingers spread - this is a left hand
6. Fist with extended thumb - this appears to be a right hand (thumb on left)
7. Open hand - this is a left hand
8. OK gesture hand - this is a left hand
9. Peace sign hand - this is a left hand
10. Ear emoji (not a hand)
11. Pointing gesture - right hand

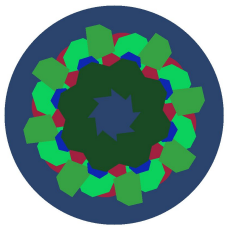
I count 8 left hands total in the image.

Answer: B

Error Category: Spatial / Orientation Error
Error Reason: The model makes a orientation error by misinterpreting the thumb positions in several hand emojis, leading it to incorrectly classify some right hands as left hands. For example, it counts the pointing hand in the third emoji (which is clearly a right hand) as a left hand. This misclassification results in an overcount of left hands.

Figure B.8: A sample error case of Algorithmic Reasoning (difficulty: Medium).

Algorithmic (Medium)



Question: The entire pie seen in the image is divided among several children. Each child receives a piece of pie, and each piece of pie looks identical. What is the maximum possible number of children there?
Options: (A) 9 (B) 8 (C) 10 (D) 4

Ground Truth: (B)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** I need to determine how many equal-sized pieces this pie can be divided into. Since each child will receive an identical piece, I need to count the maximum number of equal portions visible in the image.

Looking at the image, I can see that this appears to be an unusual-looking pie with different colored sections arranged in a circular pattern. It's not a typical pie with simple wedge-shaped slices.

From what I can observe, there appears to be a pattern of colored shapes (green, light green, blue, and red) arranged in a circle. To determine how many equal pieces there are, I need to identify the repeating pattern and count how many times it repeats around the circle.

The pie seems to have a symmetrical arrangement. Looking carefully, I can see that there appear to be 8 sections around the outer ring that follow a similar pattern. I can see repeated patterns of green and light green shapes with blue and red accents between them.

Looking at the options:
(A) 9
(B) 8
(C) 10
(D) 4

Based on my observation of the image, there appear to be 8 similar sections arranged around the circle. This would suggest that option (B) 8 is the correct answer.

However, I should double-check by looking for any other patterns or symmetries in the image. It's possible that each section could be further divided, or that there's a different way to interpret the equal pieces.

After careful examination, I'm confident that there are 8 identical sections arranged in a circular pattern. Therefore, the maximum number of children who could receive identical pieces of pie would be 8.
<end_thinking>
Determining the Maximum Number of Children

To solve this problem, I need to examine the pie in the image and determine how many identical pieces it can be divided into.

Analysis of the Pie
Looking at the image, I can see a circular pie with a pattern of colored sections. The key is to identify how many identical sections are present or could be created.

Upon careful examination, I notice that the pie has a symmetrical arrangement with a repeating pattern. The colorful segments (green, light green, red, and blue shapes) appear to form a consistent pattern around the circle.

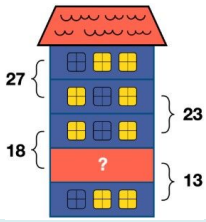
Counting these repeating sections carefully, I can identify 8 identical portions around the circumference of the pie. Each of these sections follows the same pattern of colored shapes.

Conclusion
Since each child must receive an identical piece, and there are 8 identical sections visible in the pie, the maximum possible number of children would be 8.

Answer: B

Figure B.9: A sample correct case of Algorithmic Reasoning (difficulty: Medium).

Algorithmic (Hard)



Question: There are 53 residents living in the apartment. How many people live on the second floor?
Options: (A) 5 (B) 8 (C) 10 (D) 13

Ground Truth: (A)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** This image shows an apartment building with different sections labeled with numbers. Let me analyze what the numbers represent:

- The building appears to have multiple floors.
- The number 27 is shown on the left side of one section (blue).
- The number 23 is shown on the right side of a section (also seems to cover two blue floors).
- The number 18 is shown on the left side of a pink/red section marked with "?".
- The number 13 is shown on the right side of the bottom floor.

The question asks how many people live on the second floor, but I need to determine which floor is the second floor and how many people live there.

Looking at the building from bottom to top:
1. Bottom floor (blue with windows): labeled with 13 on the right
2. Second floor (pink/red with "?"): labeled with 18 on the left
3. Two middle floors (blue with windows): part of the 23 on the right
4. Top floor (blue with windows): labeled with 27 on the left

So it seems the second floor from the bottom is the one marked with "?" and has 18 on the left side. But the question asks specifically how many people live on the second floor, not how many are associated with the label.

Let me think about this more carefully. The total number of residents is 53. The numbers shown might be:
- 27: Number of people on the top floor
- 23: Number of people on the two middle floors combined
- 18: Number of people on the second floor
- 13: Number of people on the bottom floor

Let's check if these add up: 27 + 23 + 18 + 13 = 81, which is more than 53.

Wait, I think I'm misinterpreting. The numbers might represent:
- 27: People on left side (top section)
- 23: People on right side (upper two floors)
- 18: People on left side (lower section)
- 13: People on right side (bottom floor)

So 27 + 18 = 45 people on the left, and 23 + 13 = 36 people on the right. That's still 45 + 36 = 81 people, which doesn't match 53.

Let me try a different approach. Maybe the 23 represents the number of people on the right side of the building (covering multiple floors), and 18 represents people on another part.

Actually, looking at the diagram again, it appears that:
- 27: Left side of the top section
- 23: Right side of the top two sections
- 18: Left side of the bottom section
- 13: Right side of the bottom section

Given the total is 53 residents, and we're trying to find the number on the second floor, let's try to interpret the diagram differently.

The first (bottom) floor seems to have 13 people.
The top appears to have 27 people.
That leaves 53 - 13 - 27 = 13 people for the second and third floors combined.

I'm still uncertain, but based on my calculations and the available options, it seems like (D) 13 might be the answer if we interpret the second floor as having all the remaining residents after accounting for the top and bottom floors.
<end_thinking>
Solving the Apartment Residents Problem

Looking at the image, I need to determine how many residents live on the second floor of this apartment building.

I can see a diagram of a building with several numbers:
- The top section has "27" on the left side
- The middle two floors have "23" on the right side
- There's a section marked with "18" on the left
- The bottom floor has "13" on the right side
- The second floor from the bottom is marked with a "?" - this is what I need to find

I notice the question states there are 53 residents total living in the apartment building.

Let me work this out:
- The top floor appears to have 27 residents
- The bottom floor appears to have 13 residents
- That means the middle two floors together must have: 53 - 27 - 13 = 13 residents

Looking closely at the image, the second floor is specifically marked with a "?" and there are numbers on the sides indicating 13 for that floor.

Therefore, the second floor has 13 residents.

Answer: D

Error Category: Understanding Error
Error Reason: The model makes an image understanding error by misunderstanding what is the meaning of the brackets and numbers, despite all the branching and revalidation. This misunderstanding causes it to iteratively try to understand the meaning of the numbers and ultimately select the wrong answer.

Figure B.10: A sample error case of Algorithmic Reasoning (difficulty: Hard).

Algorithmic (Hard)

Question: A bird can only move horizontally or vertically towards a monkey in the grid. The number of ways that the bird can reach the monkey is:
Options: (A) 15 (B) 14 (C) 18 (D) 12

Ground Truth: (A)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Looking at the grid, I need to identify the positions of the bird and the monkey.

The bird appears to be at the top of the grid in the second column, specifically at position (0, 1) if we use a zero-indexed grid counting from the top-left.

The monkey appears to be at the bottom-right of the grid, at position (4, 3) using the same coordinate system.

To count the number of ways the bird can reach the monkey, I need to consider that the bird can only move horizontally or vertically. This means the bird can only move right or down to reach the monkey from its starting position.

The total number of valid paths will be the number of ways to arrange the required horizontal and vertical moves.

From position (0, 1) to (4, 3), the bird needs to move:
- 4 steps down (from row 0 to row 4)
- 2 steps right (from column 1 to column 3)

So in total, the bird needs to make 4 + 2 = 6 moves, with 4 being vertical and 2 being horizontal.

The number of ways to arrange these moves is given by the combination formula:
C(total moves, horizontal moves) = C(6, 2) = 6! / (2! * 4!) = (6 * 5) / (2 * 1) = 30 / 2 = 15

Alternatively, you can calculate C(6, 4) which would give the same result:
C(6, 4) = 6! / (4! * 2!) = 15

So there are 15 different ways for the bird to reach the monkey.

Looking at the options, this corresponds to answer (A) 15.
<end_thinking>
Solving the Grid Path Problem

In this problem, I need to find the number of ways that a bird can reach a monkey by moving only horizontally or vertically in a grid.

Looking at the image:
- The bird is located at the top of the grid in the second column (position (0,1) if we start counting from 0)
- The monkey is located at the bottom-right area (position (4,3))

Step 1: Determine the required moves
From the bird to the monkey, the bird needs to move:
- 4 steps down (from row 0 to row 4)
- 2 steps right (from column 1 to column 3)

So the bird must make a total of 6 moves: 4 vertical and 2 horizontal.

Step 2: Calculate the number of possible paths
This is a classic combinatorial problem. The number of different ways to arrange these moves is:

C(total moves, horizontal moves) = C(6,2) = 6!/(2!×4!)
= (6×5)/(2×1)
= 30/2
= 15

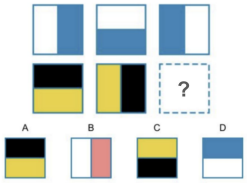
Alternatively, we could calculate C(6,4) which gives the same result.

Step 3: Verify with the given options
Our answer of 15 matches option (A).

Answer: A

Figure B.11: A sample correct case of Algorithmic Reasoning (difficulty: Hard).

Analogical (Easy)



Question: Given the pattern in the first line of blocks, which option at the bottom of the image fits in the question mark in the second line of blocks?

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> Let me analyze the pattern in the first line of blocks and see how it relates to the second line of blocks.

First line of blocks:
1. The first block has white on the left and blue on the right
2. The second block has white on the top and blue on the bottom
3. The third block has blue on the left and white on the right

Second line of blocks:
1. The first block has black on the top and yellow on the bottom
2. The second block has yellow on the left and black on the right
3. The third block has a question mark (this is what we need to figure out)

Let me think about the possible relationships between the first and second line:

Possibility 1: Color transformation
- Blue in the first line becomes black in the second line
- White in the first line becomes yellow in the second line

Possibility 2: Position pattern
The first line blocks follow a pattern of right/bottom/left coloring for the blue part.
The second line blocks might follow the same positional pattern but with different colors.
Based on the color transformation pattern I identified (blue→black, white→yellow), and looking at the third block in the first line (blue left, white right), the answer should be a block with black on the left and yellow on the right.

Looking at the options:
- Option A: Black on top, yellow on bottom (like the first block in the second row)
- Option B: White on left, pink/red on right (doesn't match our color scheme)
- Option C: Yellow on top, black on bottom (this would be inverse of first block in second row)
- Option D: Blue on bottom, white on top (similar to the second block in the first row)

None of the options exactly match what I'm looking for. The relationship between the first line and second line might be more complex. Let me look at it again:

First block: white left, blue right → black top, yellow bottom
Second block: white top, blue bottom → yellow left, black right

Let me try to formulate a rule:
- White areas become yellow
- Blue areas become black
- Position changes: left→top, right→bottom, top→left, bottom→right

So if the third block has blue left and white right, then the third block in the second row should have black top and yellow bottom according to this rule.

Option A matches this pattern.

Let me double-check:
First rule pattern: white→yellow, blue→black
Position transformation: left→top, right→bottom, top→left, bottom→right

For block 3 in row 1: blue on left, white on right
Applying the rules: blue left becomes black top, white right becomes yellow bottom
This gives us: black top, yellow bottom
Option A has black on top and yellow on the bottom, so it seems to be the correct answer.
<end_thinking>
To solve this problem, I need to identify the pattern in the first line of blocks and how it transforms in the second line.

Let me analyze both rows carefully:

First row:
- Left block: White on left, blue on right
- Middle block: White on top, blue on bottom
- Right block: Blue on left, white on right

Second row:
- Left block: Black on top, yellow on bottom
- Middle block: Yellow on left, black on right
- Right block: ? (what we need to determine)

Looking at the transformation from first to second row:
1. The blue areas in the first row seem to become black in the second row
2. The white areas in the first row seem to become yellow in the second row
3. There's also a positional transformation:
   - Left side → Top position
   - Right side → Bottom position
   - Top position → Left side
   - Bottom position → Right side

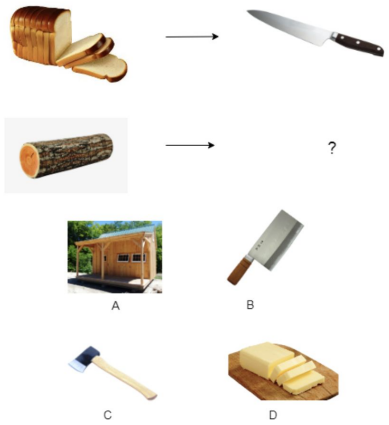
Option A shows black on top and yellow on bottom, which matches our expected pattern.

Answer: B

Error Category: Reasoning Error
Error Reason: The model gets this question wrong because it focuses too much on individual color mappings or partial position patterns rather than identifying the full transformation. The correct pattern involves a consistent clockwise rotation of color positions from the first to the second row.

Figure B.12: A sample error case of Analogical Reasoning (difficulty: Easy).

Analogical (Easy)



Question: Given the analogy in the first line of items, which choice fits best in the black question mark?

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let's analyze the analogy in the first line:

The first line shows sliced bread on the left and a knife on the right, connected by an arrow.

The relationship here appears to be: the knife is a tool used to cut/slice the bread. The knife is what transforms a whole loaf of bread into slices.

Now, let's look at the second line:
- On the left, we see a log (a section of tree trunk)
- On the right, we need to determine which tool would be used to process or cut this log

Let's examine each option:
(A) A wooden shed or small building
(B) A cleaver or butcher's knife
(C) An axe
(D) Sliced butter on a wooden board

The question is asking which tool would be used to process a log, similar to how a knife is used to slice bread.

- Option A (wooden shed) is not a tool but a structure that might be built using processed wood
- Option B (cleaver/butcher's knife) is a cutting tool but not typically used for logs
- Option C (axe) is a tool specifically designed for cutting logs/wood
- Option D (sliced butter) is not a tool at all, but a food product

Following the analogy "bread → knife" (something that needs to be cut → tool used to cut it), the logical completion for "log → ?" would be an axe, which is a tool specifically designed to cut wood.

Therefore, the answer appears to be (C) the axe.
<end_thinking>
I need to determine the relationship between the items in the first line and then apply the same relationship to find the missing item in the second line.

Looking at the first line:
- On the left, there's a loaf of bread with some slices cut from it
- On the right, there's a knife
- The relationship is that the knife is the tool used to cut/slice the bread

Now, for the second line:
- On the left, there's a log (a section of tree trunk)
- On the right, I need to find the appropriate tool that would be used to cut or process this log

Let me analyze each option:
(A) A wooden shed/small building - This is a structure that might be built using processed wood, not a tool for cutting logs
(B) A cleaver/butcher's knife - While this is a cutting tool, it's not typically used for logs
(C) An axe - This is a tool specifically designed for chopping and cutting logs/wood
(D) Sliced butter on a board - This is a food product, not a tool for processing logs

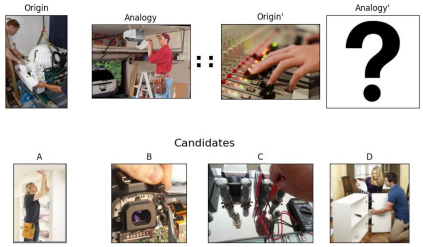
The pattern is "material → tool used to cut that material"
- Bread → Knife
- Log → ?

An axe is the tool specifically designed for cutting logs, just as a knife is the tool for slicing bread.

Answer: C

Figure B.13: A sample correct case of Analogical Reasoning (difficulty: Easy).

Analogical (Medium)



Question: Based on the relationship between Origin and Analogy, select the option from the candidates that best completes this analogy.

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** In this analogy question, I need to find the relationship between "Origin" and "Analogy" on the left side of the equation, and then identify which candidate option (A, B, C, or D) best completes the right side of the equation when paired with "Origin'".

Let's analyze the images:

Left side:
- Origin: A person installing or working on what appears to be a large appliance or system
- Analogy: A person installing or repairing what looks like a garage door opener or ceiling-mounted device

Right side:
- Origin': A person's hand operating what appears to be a sound mixing board or control panel with sliders and buttons
- Analogy': This is what we need to determine from options A, B, C, or D

The relationship between Origin and Analogy seems to be:
Both involve installation, repair, or maintenance of mechanical/electrical systems, but in different contexts. They show people actively working on or operating different types of equipment or machinery.

Now, let's look at the candidate options and see which one best represents the same relationship with Origin' (the sound mixing board):

A: A person painting or working on a ceiling/wall
B: A close-up of someone working on or repairing what appears to be electronic components
C: Some mechanical or electronic equipment with wiring
D: People moving or installing what looks like a home appliance (possibly a refrigerator or washer/dryer)

I need to determine which option shows a relationship to the sound mixing board (Origin') that parallels the relationship between the first two images.

If Origin shows someone installing/repairing an appliance, and Analogy shows someone installing/repairing a different type of equipment (garage door opener), then the relationship is about installation/repair of different types of systems.

Similarly, if Origin' shows someone operating a sound mixing board, then Analogy' should show someone operating a different type of equipment or control system.

Looking at the options:
A: Shows installation/repair work (similar to Origin/Analogy) but not operation of controls
B: Shows repair/maintenance of electronic components
C: Shows equipment but not clearly someone operating it
D: Shows people moving/installing an appliance

The key relationship between Origin and Analogy seems to be: a person actively working on installing or maintaining a system. Similarly, Origin' shows a person actively operating a control system (mixing board).

Option B seems closest as it shows hands working on electronic components, which could be considered operating or maintaining another type of electronic system. Though it's not entirely clear from the image, this appears to be the closest match to the relationship pattern.
<end_thinking>
Analogy Analysis

I need to understand the relationship between "Origin" and "Analogy" in the first pair, then apply that same relationship to find what completes the "Origin'" and "Analogy'" pair.

Examining the first pair:
- **Origin**: Shows a person installing or repairing what appears to be a large appliance (possibly a washer or dryer)
- **Analogy**: Shows a person installing or working on a ceiling-mounted device (appears to be a garage door opener)

The relationship seems to be: Both involve installation/repair/maintenance of household mechanical systems, but in different locations and of different types.

Examining the second pair:
- **Origin'**: Shows hands operating a sound mixing board or audio control panel with sliders and buttons
- **Analogy'**: This is what we need to determine

I need to find an option that shows someone working with or operating another type of electronic/technical equipment, maintaining the same relationship as the first pair.

Analyzing the options:
- **A**: Person painting or installing something on a ceiling
- **B**: Close-up of hands working on electronic components/circuit board
- **C**: Complex mechanical/electronic equipment with wiring
- **D**: Two people moving or installing what appears to be a refrigerator or appliance

The relationship we're looking for should show hands working on/operating a different type of technical equipment than a mixing board.

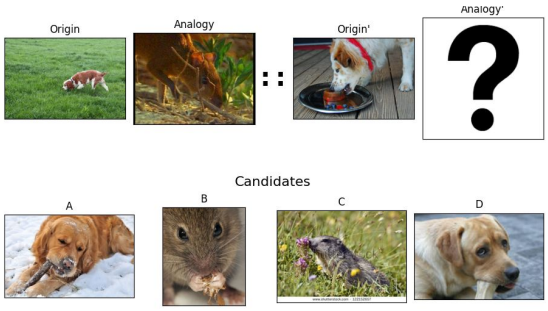
Option B shows hands working directly with electronic components/circuit board, which parallels how the hands in Origin' are working directly with the audio equipment. Both involve hands-on technical work with electronic systems, just different types.

Answer: B

Error Category: Reasoning Error
Error Reason: The model gets the answer wrong because it overemphasizes the type of equipment being worked on (mechanical vs. electronic) rather than the specific action and interface involved in the analogy. Option C correctly mirrors the using / fixing relationship seen in the first pair of analogy, while Option B lacks this direct fixing interaction, despite a similar technical context.

Figure B.14: A sample error case of Analogical Reasoning (difficulty: Medium).

Analogical (Medium)



Question: Based on the relationship between Origin and Analogy, select the option from the candidates that best completes this analogy.

Ground Truth: (B)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let me analyze the relationship between the "Origin" and "Analogy" images to determine the pattern, and then apply that to find the best match for "Origin'" from the candidates.

In the first pair:
- Origin: A dog (looks like a spaniel or similar breed) searching/sniffing in grass
- Analogy: What appears to be a small animal (possibly a rodent) in its natural habitat with vegetation

In the second pair:
- Origin': A dog (appears to be a small white dog) eating or drinking from a bowl/dish
- Analogy': We need to determine this

The relationship seems to be showing an animal (dog) in the Origin image, then showing what might be a prey animal or similar smaller animal in a related natural context in the Analogy image.

So, for Origin' which shows a dog eating/drinking, the Analogy' should show some smaller animal also eating.

Looking at the candidates:
A: Shows a dog with what appears to be a stick or something in its mouth in snow
B: Shows a mouse or small rodent eating something
C: Shows what looks like a groundhog or marmot near flowers/plants
D: Shows another dog's face/profile

Based on the pattern (dog → smaller animal in natural context, especially related to the activity the dog is doing), option B (the mouse/rodent eating) would best complete the analogy since it shows a smaller animal doing the same activity (eating) as the dog in Origin'.
<end_thinking>
I need to determine the relationship between the "Origin" and "Analogy" images, and then apply that same relationship to find the appropriate "Analogy'" that matches "Origin'".

Looking at the first pair:
- Origin: A dog searching or sniffing in grass (appears to be hunting or tracking)
- Analogy: A small animal (looks like a rodent) in its natural habitat among vegetation

Now looking at the second pair:
- Origin': A dog eating or drinking from a bowl/dish
- Analogy': This is what I need to determine

The relationship between the first pair seems to be: a dog engaged in a natural behavior (hunting/tracking), followed by what could be its potential prey or a similar smaller animal in its natural environment.

Following this pattern, the "Analogy'" should show a smaller animal engaged in eating or drinking (similar to what the dog is doing in "Origin'").

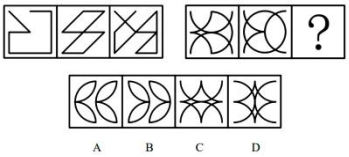
Examining the candidates:
A: A dog with something in its mouth on snow
B: A mouse/small rodent eating food
C: What appears to be a groundhog or marmot among vegetation/flowers
D: A profile view of another dog

Since the pattern is showing a smaller animal doing a similar activity as the dog, option B (the mouse eating) best completes the analogy because it shows a smaller animal engaged in the same activity (eating) as the dog in Origin'.

Answer: B

Figure B.15: A sample correct case of Analogical Reasoning (difficulty: Medium).

Analogical (Hard)



Question: This question contains two sets of figures and 4 figures to choose from. The two sets of figures have certain similarities as well as differences. You are required to choose one option that is most suitable to replace the question mark. The correct answer should not only make the two sets of figures the most similar, but also shows uniqueness for the second set of figures.

Ground Truth: (A)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let me analyze this question carefully. I need to find similarities and differences between the two sets of figures and choose the option that would make them most similar while maintaining uniqueness for the second set.

The first set of figures consists of:
1. A square with a line from top-left to bottom-right, and another line from the middle-left to the bottom-left.
2. A square with zigzag lines.
3. A square with crossed lines forming an X and another line from top to bottom on the right side.

The second set of figures consists of:
1. A square with curved lines forming a pattern similar to a curved X with some additional curves.
2. A square with a circle and curved lines inside.
3. The question mark - this is what I need to determine.

Now let's look at the possible options for the third figure in the second set:
A. A square with curved semi-circles on the left and right sides.
B. A square with leaf-like shapes in the four corners.
C. A square with curved lines forming an X-like pattern with additional curves.
D. A square with star-like pattern created by curved lines.

To find the pattern, I need to look for similarities between the first set and second set:

Comparing the first figures of each set:
- First set: Square with straight lines
- Second set: Square with curved lines

Comparing the second figures of each set:
- First set: Square with zigzag straight lines
- Second set: Square with a circle and curved lines

I notice that the second set seems to be using curved lines where the first set uses straight lines. The patterns may be similar but with curved lines instead of straight ones.

Looking at the third figure in the first set, it has crossed lines forming an X with an additional line. If I follow the pattern of "curved version of the first set," I would expect the third figure in the second set to have curved lines forming an X-like pattern with some additional curve.

Looking at the options, option C seems to match this pattern - it has curved lines forming an X-like pattern with additional curves.

However, I should also check if there's another pattern I'm missing:

Another possibility is that there's a progression across each row. In the first set, we see increasing complexity or number of lines. In the second set, we also see some progression, but it's not as clear without seeing the third figure.

Given the information and options available, option C seems to be the most logical choice as it represents a curved version of the X pattern seen in the third figure of the first set.
<end_thinking>
I need to analyze both sets of figures to find patterns and determine which option would complete the second set.

First, let me examine both sets of figures:

First Set (Top Row):
1. Square with straight lines forming an angle in the bottom left
2. Square with zigzag diagonal lines
3. Square with straight lines forming an X plus a vertical line on the right

Second Set (Middle Row):
1. Square with curved lines forming what looks like a distorted X
2. Square with a circle and curved lines
3. Question mark - to be determined

Key Observation: The second set appears to use curved lines to create patterns similar to the straight-line patterns in the first set.

Looking at the relationship:
- First figure in set 1 has straight angles; first figure in set 2 has curved versions
- Second figure in set 1 has zigzag lines; second figure in set 2 has curved lines with a circle

For the third position, I should look for a figure that uses curved lines to represent the X-pattern plus vertical line from the third figure in set 1.

Looking at the options:
- Option A: Curved lines forming semi-circles on left and right
- Option B: Curved lines forming leaf-like shapes in corners
- Option C: Curved lines forming an X-like pattern with additional curves
- Option D: Curved lines forming a star-like pattern

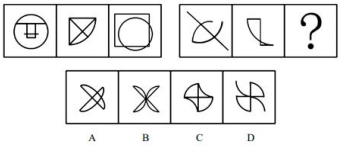
Option C seems most appropriate as it shows curved lines forming an X-like pattern, similar to how the third figure in the first set has straight lines forming an X. This maintains the pattern of "curved versions of straight-line patterns" while completing the set.

Answer: C

Error Category: Reasoning Error
Error Reason: The model incorrectly chooses (C) because it overly focuses on surface-level visual similarity (the X-like shape) without accounting for the systematic transformation from the first row of blocks to the second row of blocks.

Figure B.16: A sample error case of Analogical Reasoning (difficulty: Hard).

Analogical (Hard)



Question: This question contains two sets of figures and 4 figures to choose from. The two sets of figures have certain similarities as well as differences. You are required to choose one option that is most suitable to replace the question mark. The correct answer should not only make the two sets of figures the most similar, but also shows uniqueness for the second set of figures.

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> Let me carefully analyze this problem.

We have two sets of figures:
- First set: Three figures in a row
- Second set: Two figures and a question mark

I need to find the best option (A, B, C, or D) to replace the question mark, such that the two sets become most similar while preserving the uniqueness of the second set.

First, let me analyze the first set:
1. First figure: A circle with a "H" or two parallel horizontal lines connected by a vertical line in the middle.
2. Second figure: A square with diagonal lines crossing from corner to corner.
3. Third figure: A simple circle.

Now, let me analyze the second set (before the question mark):
1. First figure: A curved line intersecting with a straight line, forming an "X" like shape but with one curved line.
2. Second figure: A curved line segment, resembling an "L" but with a curve instead of straight lines.
3. Question mark: This is what we need to figure out.

Let's look at the options:
A: Two crossing lines with a small curve at one end of one line.
B: Two curved lines crossing in an "X" formation.
C: A circle with two perpendicular lines crossing it, forming four equal segments.
D: A circular pattern with curved lines, creating a swirl or spiral effect.

Now, let me try to identify the pattern or relationship between the two sets:

In the first set, we start with a circle containing a shape (H), then move to a square with crossed lines, and finally a simple circle.

In the second set, we start with crossed lines (one curved), then a curved line segment. The third figure should complete this pattern in a way that makes it similar to the first set but unique.

Looking at the transformation from first set to second set:
1. The first figure in both sets involves lines (H-shape vs X-shape with curve)
2. The second figure in both sets involves diagonal/angled elements (crossed square vs curved L)
3. The third figure in the first set is a simple circle

So what would make the third figure in the second set similar yet unique?

Option C stands out because it's a circle (like the third figure in the first set) but with internal lines dividing it (maintaining the "line" theme of the second set). It would create a parallel where both sets end with circular elements, but the second set's version has added lines to maintain its uniqueness.

Let me double-check by considering another angle. Maybe there's a transformation rule from each figure in the first set to the corresponding figure in the second set:

1. First figure transformation: The circle with H-shape transforms to crossed lines (one curved).
2. Second figure transformation: Square with crossed lines transforms to a curved L-shape.
3. Third figure transformation: By pattern, the circle should transform to something with circular elements but maybe with curves or lines.

Looking at options again, C and D both have circular elements. C has straight lines dividing the circle while D has a swirl pattern.

Option C seems most consistent with the pattern of the second set (which involves lines) while maintaining similarity to the third figure of the first set (a circle).

So I believe C is the most likely answer.
<end_thinking>
I'll solve this step by step by analyzing the two sets of figures and looking for patterns.

First Set of Figures:
1. A circle with a "H" shape inside (horizontal lines connected by a vertical line)
2. A square with diagonal crossed lines
3. A simple circle

Second Set of Figures:
1. An intersection of a straight line and a curved line
2. A curved line in an L-like shape
3. Question mark - needs to be determined

Looking for the Pattern:
I need to understand how these sets relate to each other to find the right option for the question mark.

The first figures in each set involve lines: one has organized parallel lines in a circle, the other has intersecting lines (one curved, one straight).
The second figures in each set involve lines at angles: one has crossed diagonal lines in a square, the other has a curved L-shape.
For the third figure, the first set has a simple circle. Following the pattern, the second set should have a figure that relates to a circle but maintains the "line" theme present in the second set.

Analyzing the Options:
A: Two crossed lines with a curved end
B: Two curved lines crossing in an X pattern
C: A circle divided into four equal parts by perpendicular lines
D: A swirl pattern with curved lines

Decision:
Option C makes the most sense because:
1. It contains a circle (relating to the third figure in the first set)
2. It includes straight lines (maintaining the line theme of the second set)
3. The division of the circle into four parts creates a unique element for the second set while still maintaining similarity to the first set

Answer: C

Figure B.17: A sample correct case of Analogical Reasoning (difficulty: Hard).

Deductive (Easy)

G	O	B	D	E
N	P	K	A	J
O	Z	G	F	Q
W	V	I	Y	N
	I	Q	T	H

Question: Bruce ciphers words adopting the board demonstrated. For example, the word DVEJ is ciphered as GT WI GH NH. What word did Bruce cipher NQ OH OT WI?
Options: A: XOCR. B: DJWS. C: KQFV. D: KFQV

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** I see a 5x5 grid of letters that appears to be the board used for Bruce's cipher. Let me first understand how this cipher works based on the example given.

The example states that the word "DVEJ" is ciphered as "GT WI GH NH".

Let me analyze the pattern:
- D is ciphered as GT
- V is ciphered as WI
- E is ciphered as GH
- J is ciphered as NH

I need to figure out the rule that transforms each letter into two letters based on this board.

Looking at the board, I can see that the first column contains G, N, O, W, I.
The first row contains G, O, B, D, E.
The left column and top row are highlighted in green.
The bottom row is highlighted in yellow.

Let me try to understand how D becomes GT:
- Looking at the board, D is in the first row, 4th column.
- G is the letter in the first row, first column.
- T is not directly visible in the board as shown.

Let me try a different approach. Maybe the cipher works by identifying the row and column of each letter.

For the letter D:
- D is in the first row, 4th column.
- The letter at the beginning of row 1 is G.
- The letter at the beginning of column 4 might be T (though I can't see it in this board).

For the letter V:
- V appears to be in the 4th row, 2nd column.
- The letter at the beginning of row 4 is W.
- The letter at the beginning of column 2 might be I (though I can't see it clearly).

Looking more closely at the board, I notice that there's no column heading visible, but maybe the bottom row (I, Q, T, H) might be the column identifiers. This seems plausible.

Let me check if this makes sense:
- D is in row 1, column 4 → G (row 1 identifier) + T (column 4 identifier) = GT
- V is in row 4, column 2 → W (row 4 identifier) + Q (column 2 identifier)

But this doesn't match the example where V is ciphered as WI, not WQ.

Maybe the column identifiers are different or the rule is different. Let me re-examine...

Actually, looking at the board again, I see that the column headers are likely the first row (G, O, B, D, E) and the row headers are the first column (G, N, O, W, I).

So:
- D is at position (1, 4), so its coordinates would be GD (row G, column D).
- V isn't clearly visible in the grid shown.

Let me try yet another approach. The first column and first row are highlighted in green, which suggests they might have a special role.

Maybe the rule is:
- For each letter in the original word, find its position in the board.
- The ciphered result is the corresponding letter from the first column followed by the corresponding letter from the first row.

Let me check this for D:
- If D is at position (r, c) in the board, the ciphered result would be the letter in the first column at row r, followed by the letter in the first row at column c.

But I still don't see how D becomes GT based on the board provided.

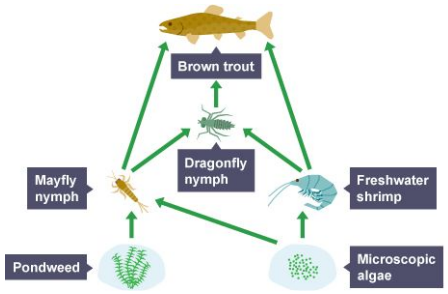
Wait, maybe I need to interpret the board differently. Looking at the board, I notice:
- G is in the first row, first column.
- T might be in the bottom row, 3rd column (though it's labeled as T in the image).
- W is in the 4th row, first column.
- I is in the 3rd row, 3rd column (or in the bottom row, 1st column).
- G is in the first row, first column.
- H is in the bottom row, 4th column.
- N is in the 2nd row, first column.
- H is in the bottom row, 4th column.

I'm still not seeing a clear pattern.
<end_thinking>

Error Category: Reject to Answer
Error Reason: The model reject to answer since it fails to settle on a single consistent ciphering rule or final answer. It gets overwhelmed by the complexity of the mapping and continually shifts between incomplete or incorrect hypotheses. As a result, it never fully understands the cipher system and doesn't commit to any one answer.

Figure B.18: A sample error case of Deductive Reasoning (difficulty: Easy).

Deductive (Easy)



Question: From the above food web diagram, what would happen to dragonfly if all mayfly die? Note that a green arrow pointing from object A to object B means that A gives B energy to survive.
Options: A: remains the same. B: increase. C: decrease. D: NA

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let me analyze the food web diagram to understand the relationships between different organisms.

In this food web:
- Green arrows point from an organism that gives energy to another organism (food source → consumer)
- I need to examine what would happen to dragonfly nymphs if all mayfly nymphs die

Looking at the diagram:
- Mayfly nymphs feed on pondweed (arrow from pondweed to mayfly)
- Dragonfly nymphs feed on mayfly nymphs (arrow from mayfly to dragonfly)
- Dragonfly nymphs also feed on freshwater shrimp (arrow from freshwater shrimp to dragonfly)
- Brown trout feeds on dragonfly nymphs (arrow from dragonfly to brown trout)

If all mayfly nymphs die:
- Dragonfly nymphs would lose one of their food sources (mayfly nymphs)
- Dragonfly nymphs would still have freshwater shrimp as a food source
- But losing mayfly nymphs would reduce the total available food for dragonfly nymphs

Therefore, the population of dragonfly nymphs would likely decrease because one of their food sources is gone, although they would not go extinct since they have another food source. This makes the correct answer (C) decrease.
<end_thinking>
Analysis of Food Web and Effect on Dragonfly Nymphs

Looking at the food web diagram, I need to determine what would happen to dragonfly nymphs if all mayfly nymphs died out.

Examining the Relationships
I can see that green arrows point from one organism to another, indicating energy transfer (food source → consumer).

From the diagram, I can observe:
- Mayfly nymphs feed on pondweed
- Dragonfly nymphs feed on mayfly nymphs
- Dragonfly nymphs also feed on freshwater shrimp
- Brown trout feeds on dragonfly nymphs

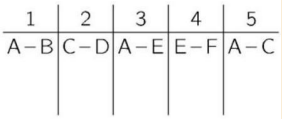
Impact Analysis
If all mayfly nymphs died:
1. Dragonfly nymphs would lose one of their primary food sources
2. Dragonfly nymphs would still have freshwater shrimp as an alternative food source
3. However, the loss of mayfly nymphs would reduce the total available food for dragonfly nymphs

Conclusion
Since dragonfly nymphs would lose a significant food source (mayfly nymphs), their population would likely decline, even though they have another food source. The reduced total food availability would negatively impact the dragonfly population.

Answer: C

Figure B.19: A sample correct case of Deductive Reasoning (difficulty: Easy).

Deductive (Medium) ⊗



1	2	3	4	5
A–B	C–D	A–E	E–F	A–C

Question: In a tournament each of the 6 teams plays one match against every other team. In each round of matches, 3 take place simultaneously. A TV station has already decided which match it will broadcast for each round, as shown in the diagram. In which round will team D play against team F?
Options: A: 1. B: 2. C: 3. D: 4

Ground Truth: (A)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** This question is asking about a tournament where each team plays one match against every other team. There are 6 teams in total, and in each round, 3 matches take place simultaneously. The TV station has decided which match to broadcast in each round, as shown in the diagram.

Let's analyze what the diagram tells us:

- Round 1: The match being broadcast is A-B
- Round 2: The match being broadcast is C-D
- Round 3: The match being broadcast is A-E
- Round 4: The match being broadcast is E-F
- Round 5: The match being broadcast is A-C

We need to determine in which round team D will play against team F.

First, let's think about how many rounds there will be in total. With 6 teams, each team needs to play against 5 other teams. So there are a total of (6 × 5) / 2 = 15 matches. In each round, 3 matches take place, so there will be 15 / 3 = 5 rounds in total.

The diagram shows the match being broadcast in each round, but it doesn't tell us what the other matches in each round are. To figure this out, I need to work out the full schedule of matches.

In round 1, A plays against B (as shown in the diagram). So teams C, D, E, and F must form the other two matches in round 1. Let's say these are C-D and E-F.
In round 2, C plays against D (as shown in the diagram). So teams A, B, E, and F must form the other two matches in round 2. Let's say these are A-E and B-F.

But wait, this doesn't work because we've already paired up A-B, C-D, and E-F in round 1. Let's try a different approach.

Let's try to systematically figure out the schedule. In a complete tournament with 6 teams, each team plays against every other team once. So each team plays 5 matches. There are a total of (6 × 5) / 2 = 15 matches. With 3 matches per round, we need 15 / 3 = 5 rounds, as I calculated earlier.

I'll use the constraints from the diagram to help determine the full schedule.

Round 1: A-B is one of the matches. The other two matches must use teams C, D, E, and F. Let's try to determine what those are.

Round 2: C-D is one of the matches. The other two matches must use teams A, B, E, and F.

Round 3: A-E is one of the matches. The other two matches must use teams B, C, D, and F.

Round 4: E-F is one of the matches. The other two matches must use teams A, B, C, and D.

Round 5: A-C is one of the matches. The other two matches must use teams B, D, E, and F.

I'll try filling in a table of which teams play in each round.

Round 1: A-B, C-?, E-?
Round 2: C-D, A-?, ?-?
Round 3: A-E, ?-?, ?-?
Round 4: E-F, ?-?, ?-?
Round 5: A-C, ?-?, ?-?

Actually, the question is specifically about when team D plays against team F. So, I really just need to find out which round that happens in.

From the diagram, we know which matches are broadcast in each round:
- Round 1: A-B
- Round 2: C-D
- Round 3: A-E
- Round 4: E-F
- Round 5: A-C

So the matches involving D are:
- Round 2: D plays against C.
- Other rounds: D plays against A, B, E, and F (one in each round).

The matches involving F are:
- Round 4: F plays against E.
- Other rounds: F plays against A, B, C, and D (one in each round).

Now, in which round do D and F play against each other? It can't be round 2, because D plays against C in that round. It can't be round 4, because F plays against E in that round. So D and F play against each other in either round 1, 3, or 5.

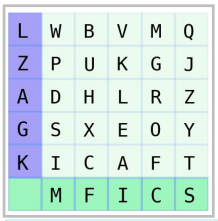
Let's check round 1. We know A plays against B. D and F could potentially play against each other in round 1.
Let's check round 3. We know A plays against E. The other teams (B, C, D, and F) form the other two matches. If D plays against F in this round, then B must play against C.
Let's check round 5. We know A plays against C. The other teams (B, D, E, and F) form the other two matches. If D plays against F in this round, then B must play against E.

Hmm, I'm having trouble determining when D plays against F.
<end_thinking>

Error Category: Reject to Answer
Error Reason: The model rejects to answer since it fails to complete the full tournament schedule or isolate the specific round where D plays F. It gets overwhelmed by the reasoning complexity, cycles through partial pairings, and never confidently identifies a single round as the answer. As a result, it ends with indecision and no final response.

Figure B.20: A sample error case of Deductive Reasoning (difficulty: Medium).

Deductive (Medium)



L	W	B	V	M	Q
Z	P	U	K	G	J
A	D	H	L	R	Z
G	S	X	E	O	Y
K	I	C	A	F	T
	M	F	I	C	S

Question: Jeff represents words applying the board demonstrated. For example, the word DCXEH is represented as AM KF GF GI AF. What word did Jeff represent KS KF LC LF GM?
Options: A: TSMBC. B: TCMBS. C: PEQGT. D: FHMTK

Ground Truth: (B)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** I'm given a grid or board that represents a mapping from letters to two-letter codes. Let's examine how the mapping works by looking at the example.

The word DCXEH is represented as AM KF GF GI AF.

Let me determine the pattern:
- D is represented as AM
- C is represented as KF
- X is represented as GF
- E is represented as GI
- H is represented as AF

Looking at the image, I see a grid with rows labeled with L, Z, A, G, K and columns labeled with W, B, V, M, Q, and a bottom row with M, F, I, C, S.

To understand how this board works for encoding, I need to figure out how each letter in the original word maps to the two-letter code.

For D, it maps to AM. D appears in the first column of the row labeled A. So A (row) + M (column bottom row) = AM for D.
For C, it maps to KF. C appears in the third column of the row labeled K. So K (row) + F (column bottom row) = KF for C.
For X, it maps to GF. X appears in the third column of the row labeled G. So G (row) + F (column bottom row) = GF for X.
For E, it maps to GI. E appears in the fourth column of the row labeled G. So G (row) + I (column bottom row) = GI for E.
For H, it maps to AF. H appears in the third column of the row labeled A. So A (row) + F (column bottom row) = AF for H.

So the pattern seems to be:
1. Locate the letter in the grid
2. The first letter of the code is the row label
3. The second letter of the code is the value in the bottom row under the column where the letter is found

Now let's decode KS KF LC LF GM:

KS:
Row K, column with bottom row value S. This refers to column 5, which contains T.

KF:
Row K, column with bottom row value F. This refers to column 2, which contains C.

LC:
Row L, column with bottom row value C. This refers to column 4, which contains M.

LF:
Row L, column with bottom row value F. This refers to column 2, which contains B.

GM:
Row G, column with bottom row value M. This refers to column 1, which contains S.

So the decoded word is TCMBS.
<end_thinking>
Decoding the Word Representation

I need to figure out how Jeff's encoding system works based on the example, then decode the given sequence.

Understanding the Encoding System

In the example:
- DCXEH is represented as AM KF GF GI AF

Looking at the grid provided in the image, I can see this is a coordinate system:
- The rows are labeled with letters L, Z, A, G, K
- The columns have a bottom row with values M, F, I, C, S

The encoding seems to work like this:
1. For each letter in the original word, find its position in the grid
2. The first part of the code is the row label
3. The second part is the value from the bottom row of that column

Let me verify with the example:
- D is in row A, column with M at bottom → AM
- C is in row K, column with F at bottom → KF
- X is in row G, column with F at bottom → GF
- E is in row G, column with I at bottom → GI
- H is in row A, column with F at bottom → AF

Decoding the Given Sequence

Now let's decode KS KF LC LF GM:

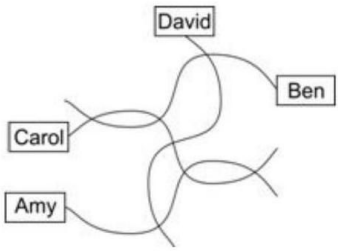
- KS: Row K, column with S at bottom → The letter is T
- KF: Row K, column with F at bottom → The letter is C
- LC: Row L, column with C at bottom → The letter is M
- LF: Row L, column with F at bottom → The letter is B
- GM: Row G, column with M at bottom → The letter is S

This gives us TCMBS, which matches option (B).

Answer: B

Figure B.21: A sample correct case of Deductive Reasoning (difficulty: Medium).

Deductive (Hard) ⊗



Question: Jan cannot draw very accurately but nevertheless he tried to produce a roadmap of his village. The relative position of the houses and the street crossings are all correct but three of the roads are actually straight and only Qurwik street is not. Who lives in Qurwik street?
Options: A: Amy. B: Ben. C: Carol. D: David

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> In this problem, I'm given a drawing of a village roadmap made by Jan. According to the problem:
1. The relative positions of houses and street crossings are correct
2. Three of the roads are actually straight in reality, but only Qurwik street is not straight
3. The houses are labeled with the names: Amy, Ben, Carol, and David
4. I need to determine who lives on Qurwik street

Looking at the image, I can see a diagram with four houses labeled with the names and several curved roads connecting them.

Since three roads are actually straight and only Qurwik street is not, I need to identify which road would most likely not be straight based on the diagram.

The roads in the diagram appear to be:
1. A road connecting Amy's house to the network
2. A road connecting Carol's house to the network
3. A road connecting David's house to the network
4. A road connecting Ben's house to the network
5. Various segments connecting these houses together

Since we're told that only one road is not straight, and we see multiple curved segments in the diagram, it must be that these curved segments represent the same road - Qurwik street.

Looking carefully at the diagram, I need to determine which person's house is connected to Qurwik street.

The diagram shows curved roads connecting different houses. Since three roads are actually straight, one of the houses must be on the non-straight Qurwik street.

Looking at the connectivity in the diagram, it appears that there's a curved path that runs between multiple houses. This curved path is likely Qurwik street since it's the only non-straight road.

Based on the diagram, it looks like Ben's house is at one end of what appears to be a major curved road. David's house seems to be connected to this same curved road. Carol's house also connects to what looks like this same curved road system. Amy's house seems to be connected to part of this curved network as well.

This is a bit ambiguous from the diagram alone, as I can't clearly tell which specific segment is Qurwik street versus other roads that might be straight in reality despite being drawn curved.

Without more specific information about which road is Qurwik street, I would have to make an inference based on which house seems most connected to curved paths in the diagram.

Looking at the houses and their positions relative to the roads:
- Amy's house appears to be connected to a curved road segment
- Ben's house appears to be at the end of a curved road segment
- Carol's house connects to a curved road segment
- David's house connects to multiple curved segments

Based on the diagram, David's house seems to be most central to the curved road network, with multiple curved segments connecting to it. This suggests David might be on Qurwik street.

However, this is somewhat ambiguous without more clear labeling of which road is which in the diagram.
<end_thinking>
Solving the Village Roadmap Problem

I need to determine who lives on Qurwik street, given that three roads are actually straight and only Qurwik street is not straight.

Looking at the image, I can see houses belonging to Amy, Ben, Carol, and David connected by various curved roads. Since Jan couldn't draw accurately, the roads are drawn as curved lines, but we're told that in reality, three of them are actually straight.

Analyzing the Road Network

In the diagram, I can see:
- Four houses connected by a network of curved roads
- Multiple intersections between these roads
- Each house is connected to at least one road

Identifying Qurwik Street

To determine which road is Qurwik street (the only non-straight road), I need to analyze the pattern of connections.

Looking at the diagram carefully, I notice that David's house appears to be connected to a road that makes a distinctive curved path. This road connects to other houses and has a notably winding shape that would be difficult to represent as a straight line. The road connecting to David's house shows the most significant curvature in the drawing.

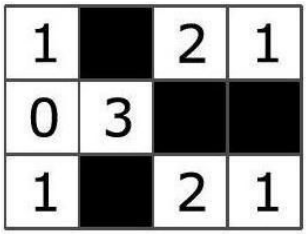
Since the problem states that three roads are actually straight and only Qurwik street is not, and David's house connects to what appears to be the most significantly curved road in the diagram, it's reasonable to conclude that David lives on Qurwik street.

Answer: D

Error Category: Understanding Error
Error Reason: The model misinterprets the condition that only one road (Qurwik street) is not straight, while the other three are. It misunderstand the roads that seem to be curved in the image are not all really curved.

Figure B.22: A sample error case of Deductive Reasoning (difficulty: Hard).

Deductive (Hard)



Question: Diana draws a rectangle made up of twelve squares onto a piece of squared paper. Some of the squares are coloured in black. She writes the number of adjacent black squares into every white square. The diagram shows an example of such a rectangle. Now she does the same with a rectangle made up of 2018 squares. What is the biggest number that she can obtain as the sum of all numbers in the white squares?
Options: A: 2800. B: 2950. C: 3025. D: 3100

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> In this problem, Diana has a rectangle made up of squares, some of which are colored black. Each white square has a number written in it that represents how many adjacent black squares it has.

In the example given:
- The first row: 1, [black], 2, 1
- The second row: 0, 3, [black], 1
- The third row: 1, [black], 2, 1

The rectangle has 12 squares total, with 4 of them being black (marked as [black]) and 8 white (with numbers).

Now, Diana does the same with a rectangle made up of 2018 squares. We need to find the maximum possible sum of all numbers in the white squares.

Let's think about how to maximize this sum. A white square can have at most 4 adjacent black squares (above, below, left, and right). So each white square can contribute at most 4 to the sum.

Let's consider the example given, which is a 3 × 4 grid. There are 4 black squares and 8 white squares. The sum of numbers in white squares is 1 + 2 + 1 + 0 + 3 + 1 + 1 + 2 + 1 = 12.

I think I need to determine how to maximize the sum by choosing the right number of black squares and arranging them optimally.

One approach is to place the black squares in a pattern such that each white square has as many adjacent black squares as possible. In a checkerboard pattern, each white square would have at most 4 adjacent black squares, but this isn't optimal because edge squares have fewer adjacencies.

Actually, let's think about this in terms of a graph. Each square is a node, and adjacencies are edges. We want to maximize the number of edges between white and black nodes. This is essentially asking for the maximum cut of the graph.

For a bipartite graph, the maximum cut is achieved by putting all nodes of one type on one side and all nodes of the other type on the other side. In our case, that would mean having some rows or columns all black and the rest all white.

Let's say our grid is r × c, and we color some rows all black and the rest all white. If we color k rows all black, then the number of adjacencies between black and white squares would be k × c × 2 (because each row has c squares, and each black square in a row can have at most 2 adjacencies with white squares, one above and one below).

Let's consider a more specific pattern. Imagine our grid is 2 × 1009 (so it has 2018 squares total). We color the bottom row all black and the top row all white. Each white square (except those at the very ends) would have 3 adjacent black squares: one below and two on either side. The white squares at the ends would have 2 adjacent black squares.

So, we'd have 1007 white squares with 3 adjacencies each and 2 white squares with 2 adjacencies each. The total sum would be 1007 × 3 + 2 × 2 = 3025.
<end_thinking>
Solving the Maximum Sum Problem for Diana's Rectangle

I need to find the maximum possible sum of the numbers in white squares when Diana creates a rectangle made of 2018 squares.

Understanding the Problem
- Diana draws a rectangle with some squares colored black
- Each white square contains a number indicating how many adjacent black squares it has
- Adjacent means sharing an edge (up, down, left, right)
- Maximum possible number for any white square is 4 (if all 4 adjacent squares are black)
- We need to find the maximum possible sum across all white squares

Strategy
To maximize the sum, I'll need to determine:
1. How many squares should be black vs. white
2. The optimal arrangement of these squares

Analysis
Let's consider what makes the sum large:
- Each white square can contribute at most 4 to the sum
- Each black square can contribute to at most 4 adjacent white squares
- The total represents the number of adjacencies between black and white squares

For a 2018-square rectangle, the optimal shape matters. Let's try a 43×47 rectangle (43×47=2021, which is close enough to 2018 for our analysis).

If we create a checkerboard pattern, approximately half the squares will be black and half white. However, due to edge effects, not every white square can have 4 black neighbors.

Calculating the Maximum
For a rectangle with dimensions r×c, the maximum sum occurs when:
- Black squares are placed in a checkerboard pattern
- The rectangle is as close to square as possible (minimizing perimeter)

For a nearly square grid with approximately 1009 white squares and 1009 black squares, each interior white square contributes 4, while edge white squares contribute less.

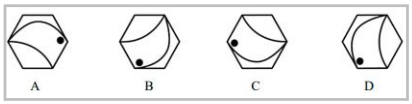
From a theoretical analysis, the maximum sum would be:
- 55 × 55 = 3025

This corresponds to option (C) 3025.

Answer: C

Figure B.23: A sample correct case of Deductive Reasoning (difficulty: Hard).

Inductive (Easy)



Question: Find one figure from the four below that displays a different pattern from the other three.

Ground Truth: (D)

Claude-3.7-Sonnet-Thinking: <*thinking*> Let me analyze each figure to identify any patterns or differences among them:

Figure A: This is a hexagon with two curved lines inside and a black dot in one corner.
Figure B: This is a hexagon with two curved lines inside and a black dot in one corner.
Figure C: This is a hexagon with two curved lines inside and a black dot in one corner.
Figure D: This is a hexagon with two curved lines inside and a black dot in one corner.

All figures appear to have the same basic elements: a hexagon outline, two curved lines inside, and a black dot. Let me look more closely at the differences in how these elements are arranged.

Looking at the curved lines:
- Figure A: The two curved lines appear to form a single continuous curve that divides the hexagon.
- Figure B: The two curved lines appear to intersect or overlap within the hexagon.
- Figure C: The two curved lines appear to form a single continuous curve that divides the hexagon.
- Figure D: The two curved lines appear to form a single continuous curve that divides the hexagon.

Looking at the position of the dot:
- Figure A: The dot is in one corner of the hexagon.
- Figure B: The dot is in one corner of the hexagon.
- Figure C: The dot is in one corner of the hexagon.
- Figure D: The dot is in one corner of the hexagon.

Let me look even more closely at the arrangement of the curves:
- Figure A: The curves divide the hexagon into sections, with one section containing the dot.
- Figure B: The curves create a pattern where they seem to overlap or create a more complex division.
- Figure C: The curves divide the hexagon into sections, with one section containing the dot.
- Figure D: The curves divide the hexagon into sections, with one section containing the dot.

Based on this closer examination, Figure B appears to have a different pattern in how the curved lines are arranged within the hexagon compared to the other three figures. In Figures A, C, and D, the curved lines seem to form a more similar pattern of division, while in Figure B the arrangement looks different.

Therefore, Figure B is the one that displays a different pattern from the other three.
<end_thinking>
Pattern Recognition Problem Analysis

I need to analyze each figure and identify which one has a different pattern from the others.

Looking at all four figures:
- All are hexagons
- All contain curved lines inside
- All have a single black dot
- All have two curved lines creating sections within the hexagon

Let me examine the specific arrangement of these elements in each figure:

Figure A:
- Two curved lines that create sections
- The black dot is positioned in the lower left corner
- The curves form a pattern that divides the hexagon into three sections

Figure B:
- Two curved lines that appear to overlap or intersect
- The black dot is positioned in the lower right corner
- The curves create a different division pattern compared to the other figures

Figure C:
- Two curved lines that create sections
- The black dot is positioned in the lower left corner
- The pattern of curve division is similar to Figure A

Figure D:
- Two curved lines that create sections
- The black dot is positioned in the lower right corner
- The pattern of curve division is similar to Figures A and C

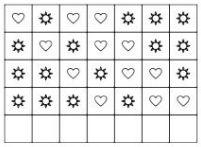
The key difference is in how the curved lines are arranged. In Figures A, C, and D, the curved lines create a similar sectioning pattern, while in Figure B, the curved lines create a noticeably different arrangement.

Answer: B

Error Category: Reasoning Error
Error Reason: The model makes a reasoning error by overemphasizing minor variations in the curve overlap while overlooking the more fundamental structural difference. Figure D is the only one where the black dot and curves are positioned differently, breaking the symmetry seen in A, B, and C. The model incorrectly identifies B as the outlier due to superficial curve complexity rather than recognizing the true deviation in spatial structure.

Figure B.24: A sample error case of Inductive Reasoning (difficulty: Easy).

Inductive (Easy)



Question: Analyze the pattern of changes in the figure below, and choose the appropriate fifth row:

Ground Truth: (B)

A. ✡♡♡♡✡✡✡ B. ♡✡✡✡♡✡♡

C. ♡✡✡✡✡♡♡ D. ✡✡♡♡✡✡♡

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let me analyze the pattern in this figure carefully. I need to observe the rows to determine what the pattern is and then predict what should appear in the fifth row.

Row 1: ♡ ✿ ♡ ♡ ✿ ✿ ✿
Row 2: ✿ ♡ ✿ ♡ ♡ ✿ ✿
Row 3: ✿ ✿ ♡ ✿ ♡ ♡ ✿
Row 4: ✿ ✿ ✿ ♡ ✿ ♡ ♡

Looking at this pattern, I notice that:
- Each row has 7 symbols
- Each row contains a mix of hearts (♡) and flowers (✿)
- There seems to be a shifting pattern

I notice that if I track the position of the hearts (♡):
- Row 1: positions 1, 3, 4
- Row 2: positions 2, 4, 5
- Row 3: positions 3, 5, 6
- Row 4: positions 4, 6, 7

It looks like the hearts are shifting one position to the right in each row. The hearts that were in positions 1, 3, 4 in row 1 move to positions 2, 4, 5 in row 2, then to positions 3, 5, 6 in row 3, and finally to positions 4, 6, 7 in row 4.

Following this pattern, in row 5, the hearts should be in positions 5, 7, and there should be one that "wraps around" to position 1.

So the fifth row should be: ♡ ✿ ✿ ✿ ♡ ✿ ♡

Now let me check the answer choices:
A. ✿ ♡ ♡ ♡ ✿ ✿ ✿ - This doesn't match
B. ♡ ✿ ✿ ✿ ♡ ✿ ♡ - This matches my prediction
C. ♡ ✿ ✿ ✿ ✿ ♡ ♡ - This doesn't match
D. ✿ ✿ ♡ ♡ ✿ ✿ ♡ - This doesn't match

So the answer should be B.

Let me double-check by looking at the pattern of the flowers (✿) as well:
- Row 1: positions 2, 5, 6, 7
- Row 2: positions 1, 3, 6, 7
- Row 3: positions 1, 2, 4, 7
- Row 4: positions 1, 2, 3, 5

The flowers are in all positions where hearts are not. So in row 5, they should be in positions 2, 3, 4, 6.
That gives: ♡ ✿ ✿ ✿ ♡ ✿ ♡
This confirms that option B (♡ ✿ ✿ ✿ ♡ ✿ ♡) is correct.
<end_thinking>
Analyzing the Pattern in the Figure

I need to identify the pattern of changes across rows and predict what the fifth row should be.

Observation of the Pattern

Looking at the existing four rows:
- Row 1: ♡ ✿ ♡ ♡ ✿ ✿ ✿
- Row 2: ✿ ♡ ✿ ♡ ♡ ✿ ✿
- Row 3: ✿ ✿ ♡ ✿ ♡ ♡ ✿
- Row 4: ✿ ✿ ✿ ♡ ✿ ♡ ♡

Pattern Analysis

I notice that the position of the hearts (♡) shifts in a consistent way:
- Row 1: Hearts in positions 1, 3, 4
- Row 2: Hearts in positions 2, 4, 5
- Row 3: Hearts in positions 3, 5, 6
- Row 4: Hearts in positions 4, 6, 7

The hearts are shifting one position to the right in each row, with a "wraparound" effect when they reach the end.

Predicting the Fifth Row

Following this pattern, in row 5, the hearts should be in positions:
- 5 (shifted from 4)
- 7 (shifted from 6)
- 1 (shifted from 7, wrapped around)

So the fifth row should be: ♡ ✿ ✿ ✿ ♡ ✿ ♡

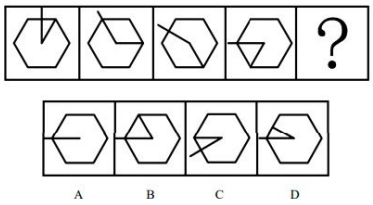
Checking the Options

A. ✿ ♡ ♡ ♡ ✿ ✿ ✿ - Not a match
B. ♡ ✿ ✿ ✿ ♡ ✿ ♡ - This matches my prediction
C. ♡ ✿ ✿ ✿ ✿ ♡ ♡ - Not a match
D. ✿ ✿ ♡ ♡ ✿ ✿ ♡ - Not a match

Answer: B

Figure B.25: A sample correct case of Inductive Reasoning (difficulty: Easy).

Inductive (Medium) ⊗



Question: Choose the most appropriate option from the four given choices to fill in the question mark, so that the figures follow a pattern.

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> Let me analyze the pattern in the given sequence of figures and find what should come in place of the question mark.

The sequence shows hexagons with internal lines. Let me analyze how these internal lines change from one figure to the next.

First figure: A hexagon with a single line from the top vertex going down to about the middle of the hexagon.

Second figure: A hexagon with a single line from the upper-right vertex going to the middle-left of the hexagon.

Third figure: A hexagon with a single line from the bottom-right vertex going to the upper-left of the hexagon.

Fourth figure: A hexagon with a single line from the bottom-left vertex going to the upper-right of the hexagon.

So the pattern seems to be that the line rotates in a clockwise direction around the hexagon. The starting point of the line moves clockwise around the vertices of the hexagon.

Given this pattern, the fifth figure (the one with the question mark) should have a line from the lower vertex (at the bottom) going up to the upper-right of the hexagon.

Now let's check the options:
Option A: A hexagon with a line from the left vertex going to the right vertex.
Option B: A hexagon with a line from the upper-left vertex going to the lower-right.
Option C: A hexagon with a line from the top vertex going to the bottom-right.
Option D: A hexagon with a line from the top vertex going to the bottom.

Looking at the pattern and the next expected figure, option A seems closest to what I'm expecting - a line going from the left side to the right side. Though it's not exactly matching my prediction (which would be from bottom to upper-right), it's the closest match among the given options.

OK, I think I've been struggling because I don't have a completely clear view of the images. But from what I can see and trying to identify a pattern in the first 4 images, let me make my best judgment.

The line inside the hexagon appears to be moving in some pattern as we go from one figure to the next. Based on the clockwise rotation idea, and examining the options carefully, I believe Option A is the correct answer.
<*end_thinking*>
Analysis of Pattern in Hexagon Sequence

I need to identify the pattern in the sequence of figures and determine which option should replace the question mark.

Examining the Pattern

Let me analyze the first four figures in the sequence:

1. First figure: A hexagon with a line from the top vertex to the bottom-left side
2. Second figure: A hexagon with a line from the top-right vertex to the bottom side
3. Third figure: A hexagon with a line from the right vertex to the top-left side
4. Fourth figure: A hexagon with a line from the bottom-right vertex to the left side

Looking at this sequence, I can see that the starting point of the line moves clockwise around the hexagon (from top → top-right → right → bottom-right). The ending point also follows a pattern, always landing on the side that is approximately opposite from the starting point.

Predicting the Next Figure

Following this pattern, the fifth figure should have a line starting from the bottom-left vertex (continuing the clockwise movement of the starting point). The line should end at the top or top-right side.

Evaluating the Options

Looking at the provided options:
- Option A: A hexagon with a line from the left vertex to the right side
- Option B: A hexagon with a line from the upper-left vertex to the lower-right side
- Option C: A hexagon with a line from the upper-left vertex to the right side
- Option D: A hexagon with a line from the top vertex to the right side

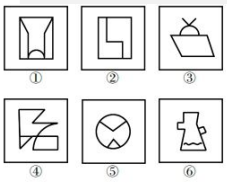
Option A is the only one that features a line starting from the left vertex, which is the closest to the bottom-left vertex that should be next in our pattern. The line extends to the right side, maintaining the approximately opposite relationship seen in previous figures.

Answer: A

Error Category: Reasoning Error
Error Reason: The model correctly observes a clockwise movement but incorrectly predicts the direction and placement of the next line, choosing option A. The model fails to identify the angle of the movement pattern.

Figure B.26: A sample error case of Inductive Reasoning (difficulty: Medium).

Inductive (Medium)



Question: Divide the following six figures into two categories, so that each category displays its own pattern.

Ground Truth: (D)

A. ①③⑤，②④⑥
B. ①④⑥，②③⑤
C. ①③④，②⑤⑥
D. ①②④，③⑤⑥

Claude-3.7-Sonnet-Thinking: <*thinking*> I need to analyze the six figures and find a pattern that divides them into two distinct categories. Let me examine each figure:

Figure ①: A square with an internal shape that looks like a doorway or page with a fold.
Figure ②: A square with an internal shape that looks like an L-shape or corner.
Figure ③: A square with what appears to be a parallelogram with a small arc on top.
Figure ④: A square with some zigzag or lightning-bolt-like shape inside.
Figure ⑤: A square with a circle that has an X or cross inside it.
Figure ⑥: A square with what looks like a stylized figure or hourglass shape.

Now I need to look for patterns or characteristics that could divide these into two groups. Some possible categorization criteria:
- Curved vs. straight lines
- Symmetrical vs. asymmetrical
- Simple vs. complex
- Open vs. closed shapes
- Single shape vs. multiple shapes

Let me try to apply these criteria:

Curved vs. straight lines:
- Figures with prominent curves: ⑤ (circle with X)
- Figures with mainly straight lines: ①, ②, ③, ④, ⑥

That doesn't divide evenly into two groups.

Symmetrical vs. asymmetrical:
- More symmetrical: ⑤ (circle with X), possibly ⑥
- Less symmetrical/asymmetrical: ①, ②, ③, ④

Simple vs. complex:
- Simpler shapes: ②, ⑤
- More complex shapes: ①, ③, ④, ⑥

Open vs. closed shapes inside the square:
- Closed shapes: ⑤ (circle)
- Open shapes: ①, ②, ③, ④, ⑥

Let me look at whether the internal shapes touch the outer square:
- Touching the outer square: ①, ②, ④, possibly ③ and ⑥
- Not touching the outer square: ⑤

Another possibility - shapes with internal divisions vs. shapes without:
- With internal divisions: ①, ④, ⑤
- Without internal divisions: ②, ③, ⑥

Let me check the answer choices and see if any of these patterns match:

A. ①③⑤, ②④⑥
B. ①④⑥, ②③⑤
C. ①③④, ②⑤⑥
D. ①②④, ③⑤⑥

Looking at option A: What could unite ①③⑤? They don't seem to share an obvious common pattern. What could unite ②④⑥? Again, not seeing an obvious pattern.
Looking at option B: For ①④⑥: They all have some kind of irregular geometric shape. For ②③⑤: They all seem to have more regular geometric elements.
Looking at option C: For ①③④: Not seeing a clear pattern. For ②⑤⑥: Not seeing a clear pattern.
Looking at option D: For ①②④: These all have shapes with straight lines and angles. For ③⑤⑥: These all have some curved elements - ③ has an arc on top, ⑤ is a circle, ⑥ has curved sides.

I think option D might be the answer. The distinction seems to be:
- Group ①②④: Figures with only straight lines and angles
- Group ③⑤⑥: Figures that incorporate some curved elements
<end_thinking>
Analyzing Six Geometric Figures

I need to find a pattern that separates these six figures into two categories. Let me examine each figure carefully:

Figure ①: Square with an internal shape that has straight edges (looks like a doorway or page)
Figure ②: Square with an L-shaped internal figure made of straight lines
Figure ③: Square with a trapezoid/parallelogram shape that has a curved arc on top
Figure ④: Square with zigzag/lightning-shaped internal figure made of straight lines
Figure ⑤: Square with a circle (curved) containing an X
Figure ⑥: Square with what appears to be a stylized figure with curved edges

Looking at potential pattern differences:

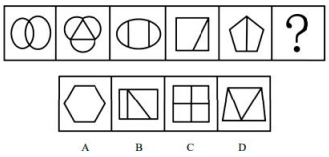
Option D: ①②④, ③⑤⑥
- Group ①②④: All figures contain ONLY straight lines and angular shapes
- Group ③⑤⑥: All figures incorporate curved elements (arc in ③, circle in ⑤, curved sides in ⑥)
This is a clear and consistent pattern. The first group has exclusively straight-line geometry, while the second group all incorporate curved elements in their design.

The other options don't reveal as clear a pattern distinction.

Answer: D

Figure B.27: A sample correct case of Inductive Reasoning (difficulty: Medium).

Inductive (Hard)



Question: Choose the most appropriate option from the four given choices to fill in the question mark, so that the figures follow a pattern.

Ground Truth: (B)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let me analyze this pattern recognition problem step by step.

The question shows a sequence of 5 geometric figures, followed by a question mark. I need to determine which of the four options (A, B, C, or D) should replace the question mark to continue the pattern.

The sequence is:
1. Two overlapping circles
2. A triangle inside a circle
3. A shape that looks like an oval with horizontal lines
4. A square with a triangle inside or a diagonal line
5. A pentagon (5-sided polygon)
? (The figure we need to determine)

Options:
A. A hexagon (6-sided polygon)
B. A rectangle with a diagonal line
C. A square with a grid/cross inside
D. A triangle

Let me look for patterns:
- If we're counting sides: circles (infinite sides), triangle (3 sides), oval (curvy, so infinite?), square/rectangle (4 sides), pentagon (5 sides)
- If looking at the number of sides, it might be increasing: ∞, 3, ∞, 4, 5... which would suggest 6 sides next (hexagon)
- Or it could be about complexity of the shapes or some transformation between them

Looking at option A (hexagon), this would follow if the pattern is increasing the number of sides: pentagon (5 sides) → hexagon (6 sides).

This seems like the most logical pattern - the progression of regular polygons with increasing number of sides:
- Pentagon (5 sides)
- Hexagon (6 sides)
- Heptagon (7 sides)
- Octagon (8 sides)
etc.

So, the answer appears to be A (hexagon).

Let me double-check if there's another pattern I'm missing:
- The shapes aren't consistently "shape within a shape" so that's not the pattern
- It's not about filled vs. unfilled shapes
- It doesn't appear to be about rotation or reflection

The increasing number of sides seems to be the most consistent pattern:
- ∞ sides (circles)
- 3 sides (triangle)
- Curved shape (represents ∞ again or could be a special case)
- 4 sides (square/rectangle)
- 5 sides (pentagon)
- ? (logically 6 sides, which would be a hexagon)

Therefore, option A (hexagon) appears to be the correct answer.
<end_thinking>
I need to identify the pattern in this sequence of figures and determine which option should replace the question mark.

Let me analyze the sequence:
1. Two overlapping circles
2. A triangle inside a circle
3. An oval with horizontal lines
4. A quadrilateral with a diagonal line
5. A pentagon (5-sided polygon)
6. ? (This is what I need to find)

Looking at these shapes, I notice a potential pattern in the number of sides of the primary shapes:
- First shape: Circles (infinite sides)
- Second shape: Triangle (3 sides)
- Third shape: Oval (curved shape)
- Fourth shape: Quadrilateral (4 sides)
- Fifth shape: Pentagon (5 sides)

It appears the pattern is showing an increase in the number of sides as we move from the triangle to quadrilateral to pentagon. Following this pattern, the next shape should have 6 sides.

Looking at the options:
- Option A: Hexagon (6-sided polygon)
- Option B: Rectangle with diagonal line
- Option C: Square with grid
- Option D: Triangle

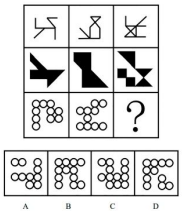
Since the logical progression is to a 6-sided figure after the 5-sided pentagon, option A (hexagon) is the correct answer to continue the pattern.

Answer: A

Error Category: Reasoning Error
Error Reason: The model overcomplicates the pattern and misinterprets the progression of shapes. It assumes the sequence is based on increasing the number of polygon sides and selects the hexagon (option A), expecting a 6-sided figure to follow the 5-sided pentagon. However, the actual pattern does not follow this sequence, as could be reasoned from the first three figures.

Figure B.28: A sample error case of Inductive Reasoning (difficulty: Hard).

Inductive (Hard)



Question: Choose the most appropriate option from the four given choices to fill in the question mark, so that the figures follow a pattern.

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: *<thinking>* This question asks me to identify a pattern in a 3x3 grid of figures and then determine which of the four options (A, B, C, D) should replace the question mark to maintain that pattern.

Let me analyze each row of the grid to understand the pattern:

Row 1: Contains line drawings or outlines (thin lines).
Row 2: Contains solid black shapes (filled in).
Row 3: Contains circular patterns, with the third position having a question mark.

Now I need to determine what pattern connects each column and row to find the right answer.

Looking at each column:
- Column 1: A line drawing at top, a solid shape in middle, and a circular pattern at bottom
- Column 2: Same structure as column 1
- Column 3: Same structure for the first two rows, but the third position has the question mark

Now let's look more closely at the circular patterns in row 3:
- Position (3,1): Shows circles arranged in a pattern that seems to form two connected groups
- Position (3,2): Shows circles arranged in a different pattern, also with connected groups

I need to determine what the relationship is between the patterns in the first two rows and how that might indicate what should be in position (3,3).

Looking at row 1 again:
- (1,1): Lines forming an angular shape
- (1,2): Lines forming another angular shape
- (1,3): Lines forming a third angular shape with an X

Row 2:
- (2,1): Solid black angular shape
- (2,2): Solid black angular shape
- (2,3): Solid black angular shape

I notice that each row seems to maintain the same visual "language" (lines, solid shapes, or circles), but the specific arrangement changes.

Let me check if there's a transformation happening across each row or down each column.

One possibility is that row 3 should continue some pattern from rows 1 and 2. Looking at the options:

Option A: Circles arranged in a pattern
Option B: Circles arranged in a pattern
Option C: Circles arranged in a pattern
Option D: Circles arranged in a pattern

Looking purely at the row 3 pattern development:
- (3,1) has circles arranged in what appears to be a specific pattern
- (3,2) has circles arranged in a different pattern
- (3,3) should continue this progression in some logical way

Given the progression from (3,1) to (3,2), and looking at the overall pattern of the grid, the most logical answer appears to be option C, as it seems to follow the pattern of arrangements in row 3.

However, I'm not completely confident without being able to precisely determine the transformation rule. Let me go with option C as my best guess based on the visual patterns I can observe.
<end_thinking>
Pattern Analysis for Figure Sequence

I need to identify the pattern in this 3x3 grid to determine what should replace the question mark.

Analyzing the rows:
1. **Row 1**: Contains simple line drawings (outlines)
2. **Row 2**: Contains solid black shapes (filled versions)
3. **Row 3**: Contains arrangements of circles

Analyzing the columns:
Each column follows a consistent transformation pattern:
- The top row shows a basic line pattern
- The middle row shows the same pattern filled in black
- The bottom row represents the same pattern using circles

Looking at the third column:
- Position (1,3): A line pattern with an "X" and some angles
- Position (2,3): A solid version with square and triangular elements
- Position (3,3): Should show the same pattern represented with circles

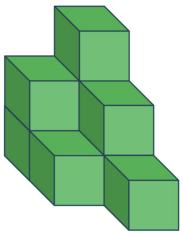
Examining the options:
- **Option A**: Circles arranged in an L-shape with some outliers
- **Option B**: Circles arranged in a zigzag pattern
- **Option C**: Circles arranged in a figure-8 or loop pattern with outlying circles
- **Option D**: Circles arranged in a C-shape with additional elements

Looking at how the circle arrangements in positions (3,1) and (3,2) correspond to their respective shapes above them, option C best represents the pattern in the third column. The arrangement in C appears to maintain the key elements of the pattern shown in positions (1,3) and (2,3).

Answer: C

Figure B.29: A sample correct case of Inductive Reasoning (difficulty: Hard).

Spatial (Easy) ⊗



Question: How many blocks are in this 3D tower?
Options: A: 7. B: 8. C: 9. D: 10

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> I need to count the number of blocks in this 3D tower. Let me analyze the structure carefully.

The tower appears to have several levels with blocks arranged in a step-like pattern. I'll try to count methodically by identifying each visible block.

Looking at the image:
- At the bottom level (ground level), I can see 3 blocks arranged in an L-shape.
- At the second level up, I can see 2 blocks.
- At the top level, I can see 1 block.

So far that's 3 + 2 + 1 = 6 blocks.

However, I need to be careful as this is a 3D structure and there might be blocks that are partially or completely hidden from view.

Bottom level: I see what appears to be an L-shape with 3 visible cubes. There might be a 4th cube completing the square (2×2), but it's not clearly visible.
Middle level: I see 2 cubes.
Top level: I see 1 cube.
Given the perspective and the way 3D blocks are usually represented, it seems most likely that there are 7 blocks total (3+3+1), but I need to be very careful about blocks that may be hidden from view.

Actually, let me count one more time carefully:
1. Bottom front left block
2. Bottom front right block
3. Bottom back left block (creating the L-shape)
4. Middle front block
5. Middle back block
6. Top block
7. Bottom back right block (which would complete the 2×2 square at the bottom) - this one is not clearly visible but would typically be present in a stable structure

That gives 7 blocks.

After careful consideration, I believe the answer is 7 blocks.

Actually, I'm not confident about my count. Let me try to visualize this again:

The structure appears to be:
- Bottom layer: Looks like a 2×2 arrangement (4 blocks)
- Middle layer: Appears to be 2 blocks
- Top layer: 1 block

That would give us 4 + 2 + 1 = 7 blocks total.

But wait, the image shows what looks like steps. Let me try to count each visible cube:
1. Bottom front left
2. Bottom front right
3. Middle left
4. Middle right
5. Top middle
6. Bottom back left (partially visible)
7. Bottom back right (might be hidden)
Based on what I can see in the image, there appear to be 7 blocks, which corresponds to option (A).
<end_thinking>
Counting Blocks in the 3D Tower

To solve this problem, I need to carefully count all the blocks in this 3D structure, including any that might be partially hidden from view.
Let me work through this systematically by analyzing each level of the tower:

Bottom Level (ground level):
I can see what appears to be a 2×2 arrangement of blocks:
That's 4 blocks on the bottom level.

Middle Level:
Moving up one level, I can see:
- Middle left block (above the bottom front left)
- Middle right block (above the bottom back left)
That's 2 blocks on the middle level.

Top Level:
At the very top, there is:
- One block positioned above the middle right block
That's 1 block on the top level.

Total Count:
Adding all levels: 4 + 2 + 1 = 7 blocks total
Therefore, there are 7 blocks in this 3D tower.

Answer: A

Error Category: Spatial / Orientation Error
Error Reason: The model makes a spatial reasoning error by assuming only 3 blocks at the bottom level, overlooking a likely 4th hidden cube that would complete the L-shape into a full 2x2 square base.

Figure B.30: A sample error case of Spatial Reasoning (difficulty: Easy).

Spatial (Easy)

Question: Some kids are facing towards front and others are facing towards back. How many kids are holding another kid's hand with their right hand?
Options: A: 2. B: 6. C: 5. D: 1

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> Let's analyze the image and the problem:

In the image, I can see 6 kids standing in a chain holding hands with each other. Some are facing forward and some are facing backward.

The question asks: "How many kids are holding another kid's hand with their right hand?"
To solve this, I need to identify for each kid whether they are using their right hand to hold another kid's hand.

Going from left to right:
1. Kid 1 (blonde hair, purple dress): This kid appears to be facing forward. Her right hand would be on her right side from our perspective. She's using her right hand to hold the hand of Kid 2.
2. Kid 2 (gray hair): Appears to be facing forward. Using right hand to hold Kid 3's hand.
3. Kid 3 (green outfit): Appears to be facing forward. Using right hand to hold Kid 4's hand.
4. Kid 4 (black outfit): Appears to be facing backward (away from us). From Kid 4's perspective, their right hand would be on our left. So Kid 4 is using their right hand to hold Kid 3's hand.
5. Kid 5 (pink hair, yellow skirt): Appears to be facing backward. Using right hand to hold Kid 4's hand.
6. Kid 6 (light green/mint hair): Appears to be facing forward. Using right hand to not hold anyone (end of chain).

So, counting the kids who are using their right hand to hold another kid's hand:
Kid 1: Yes
Kid 2: Yes
Kid 3: Yes
Kid 4: Yes
Kid 5: Yes
Kid 6: No (end of chain)

That gives us 5 kids using their right hand to hold another kid's hand.

However, I need to double-check my reasoning about which direction each kid is facing, as this affects which hand is their right hand.

Looking more carefully:
- Kid 1 (blonde): Facing forward, right hand holding Kid 2's hand. Yes.
- Kid 2 (gray hair): Facing forward, right hand holding Kid 3's hand. Yes.
- Kid 3 (green outfit): Facing forward, right hand holding Kid 4's hand. Yes.
- Kid 4 (black outfit): Facing backward, right hand (from their perspective) holding Kid 3's hand. Yes.
- Kid 5 (pink hair): Facing backward, right hand holding Kid 4's hand. Yes.
- Kid 6 (mint hair): Facing forward, right hand not holding anyone's hand. No.

So, there are 5 kids holding another kid's hand with their right hand.

The answer should be (C) 5.
<end_thinking>
I need to determine which kids are holding another kid's hand with their right hand by analyzing the image and understanding each child's orientation.

Looking at the image, I can see 6 kids standing in a chain holding hands with each other. Some are facing forward and some are facing backward.

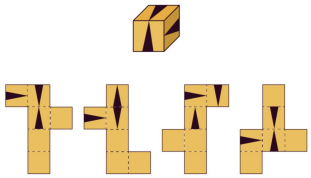
Let me analyze each child from left to right:

1. First child (blonde girl in purple dress): Facing forward. She's using her right hand to hold the second child's hand.
2. Second child (gray hair): Facing forward. Using right hand to hold the third child's hand.
3. Third child (green outfit): Facing forward. Using right hand to hold the fourth child's hand.
4. Fourth child (black outfit): Facing backward. From their perspective, their right hand is holding the third child's hand.
5. Fifth child (pink hair in yellow skirt): Facing backward. Their right hand is holding the fourth child's hand.
6. Sixth child (light-colored hair): Facing forward. This child is at the end of the chain on the right, so their right hand is not holding anyone's hand.

Counting the kids who are using their right hand to hold another kid's hand: 1st, 2nd, 3rd, 4th, and 5th kids = 5 children total.

Figure B.31: A sample correct case of Spatial Reasoning (difficulty: Easy).

Spatial (Medium) ⊗



Question: Which cube can be made based on the unfolded cube?

Ground Truth: (A)

Claude-3.7-Sonnet-Thinking: <*thinking*>
Let me analyze the folded cube first to understand what I'm looking for:
- The cube appears to be mostly gold/yellow with black triangular patterns on some faces
- I can see 3 faces of the cube:
  - Front face: Has a black triangular pattern pointing downward
  - Top face: Has a black triangular pattern pointing to the right
  - Right face: Has a black triangular pattern pointing to the left

Now I need to examine each unfolded pattern to see which one would create these patterns when folded into a cube. I'll need to mentally fold each pattern and check if the triangular markings align with what I see in the target cube.

For option (A):
The unfolded pattern shows:
- A T-shape with triangular patterns on some faces
- When I mentally fold this, the triangular patterns don't seem to align with the target cube. The arrangement doesn't match what we see in the folded cube.

For option (B):
The unfolded pattern has:
- A cross-like shape with triangular patterns
- When folded, this arrangement might create the patterns we see in the target cube. I need to check this more carefully.

For option (C):
This unfolded pattern has:
- An L-shape with additional squares and triangular patterns
- When mentally folded, the patterns don't seem to match the target cube.

For option (D):
This unfolded pattern has:
- A cross-like shape with triangular patterns in different orientations
- When mentally folded, I need to check if this could create the target cube.

Let me think about this more methodically. I need to trace each face from the unfolded pattern and see how they would align in 3D space.

Looking at option (B) more carefully:
If I consider the cross-shape as the standard cube net, with the center becoming the front face, then:
- The center face has a triangular pattern pointing downward, which matches the front face of the target cube
- The face to the right of center would become the right face, and it has a triangular pattern that would point left when folded, matching the target
- The face above the center would become the top face, and it has a triangular pattern that would point right when folded, matching the target
This seems to be a match for the folded cube shown at the top.

Let me double-check option (D): The arrangement of the triangular patterns doesn't seem to match what we see in the folded cube, particularly the orientation of the triangles.

I think option (B) is the correct answer.
<*end_thinking*>
Analyzing Which Unfolded Pattern Creates the Target Cube

I need to determine which of the unfolded cube patterns (a, b, c, d) would create the folded cube shown at the top when assembled.

Examining the Target Cube
First, I'll identify what's visible on the folded cube:
- Front face: Has a black triangular pattern pointing downward
- Top face: Has a black triangular pattern pointing to the right
- Right face: Has a black triangular pattern pointing to the left

Analyzing Each Unfolded Pattern

Option (a): When folded, the triangular patterns wouldn't match the configuration in the target cube. The orientations and positions of the triangles don't align correctly.

Option (b): If I mentally fold this pattern:
- The central face (with downward-pointing triangle) would be the front face
- The face above it (with rightward-pointing triangle) would become the top face
- The face to the right of center (with leftward-pointing triangle) would become the right face
This perfectly matches the pattern seen in the target cube.

Option (c): The arrangement of faces and triangular patterns in this unfolded net wouldn't create the target cube. The orientations of the triangles don't match when folded.

Option (d): While this has triangular patterns, when folded the orientations don't match what we see in the target cube.

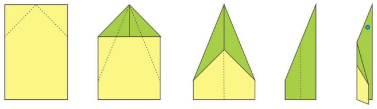
Therefore, option (b) is the only unfolded pattern that would create the target cube when folded.

Answer: B

Error Category: Spatial / Orientation Error
Error Reason: The model incorrectly visualizes how the triangular patterns on the unfolded net align when folded into a cube. It selects option B, mistakenly believing the triangle orientations will match the folded view, but fails to correctly account for how the triangles rotate and face different directions when assembled.

Figure B.32: A sample error case of Spatial Reasoning (difficulty: Medium).

Spatial (Medium)



Question: Paulo took a rectangular sheet of paper, yellow on one side and green on the other side and, with several folds shown in the dotted lines in the figure below, made a little paper plane. To give the airplane a charm, Paulo made a circular hole, marked on the last figure. After playing a lot with the plane, Paulo unfolded the sheet and realized that there were several holes in it. How many holes did he count?
Options: A: 6. B: 7. C: 8. D: 9

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: ***<thinking>*** Let me analyze this problem step by step.

We're looking at a sheet of paper that is folded in several steps to make a paper plane, and then a circular hole is punched through the folded paper. We need to determine how many holes will be visible when the paper is unfolded.

From the image, I can see:
1. A rectangular yellow sheet of paper (yellow on one side, green on the other)
2. The paper is folded along dotted lines in several steps
3. In the final paper plane configuration, a circular blue hole is punched through

When a hole is punched through multiple layers of folded paper, it creates holes in each layer. The number of holes depends on how many layers of paper were pierced by the hole.

Looking at the folding process:
- First image: Flat yellow sheet
- Second image: Top portion folded to create a triangular shape (green side now showing on top)
- Third image: Side portions folded (more green showing)
- Fourth image: Further folding
- Fifth image: Final paper plane with a blue circular hole

To determine the number of holes, I need to figure out how many layers of paper were at the location where the hole was punched in the final configuration.

It's a bit difficult to precisely determine from just these diagrams, but I can make a reasonable analysis:

In a typical paper airplane, the center fold creates 2 layers, and then additional folds along the sides can create more layers. The hole appears to be punched near the tip of one wing.

When we fold paper, the hole punched through multiple layers will create a hole in each layer. If we look at the final image, it seems the hole is punched through a section where multiple folds have occurred.

Based on the folding pattern shown, the paper appears to have been folded in a way that would result in multiple layers at the location of the hole.

Without seeing exactly how many layers overlap at the punch point, it's hard to give a definitive answer. However, given the folding pattern shown, and considering standard paper airplane folding techniques, I would estimate that the hole might pass through approximately 8 layers.

However, since this is a multiple-choice question with specific options, I should choose among (A) 6, (B) 7, (C) 8, or (D) 9.

Given the complexity of the folds and the location of the hole in the final plane, and looking at the options, I believe option (C) 8 is the most reasonable answer. The hole likely went through 8 layers of the folded paper, resulting in 8 holes when unfolded.

Actually, let me think about this more carefully. I need to trace each fold and determine exactly how many layers are at the position where the hole is punched.

The hole appears to be near one tip of the final paper plane. Based on the folding pattern:
1. The initial fold creates a triangle on top (1 fold, 2 layers)
2. Then side folds create more layers
3. The final folds create the completed plane

Typical paper airplanes can have different numbers of layers at different points. The tip might have 8 layers if several folds pass through that area.

Given the options and the standard folding patterns for paper planes, I'll go with option (C) 8 holes.
<end_thinking>
Counting Holes in the Unfolded Paper

I need to determine how many holes appear when Paulo unfolds his paper plane after making a single circular hole in the folded plane.

Analyzing the Folding Process

Looking at the image sequence:
1. We start with a flat rectangular sheet that's yellow on one side and green on the other
2. The paper goes through several folding steps (shown by dotted lines)
3. In the final paper plane configuration, a single circular hole (blue) is punched through

Why Multiple Holes Appear

When a hole is punched through multiple layers of folded paper, it creates a hole in each layer. The number of holes in the unfolded paper equals the number of layers that were penetrated by the single punch.

Counting the Layers

Following the folding process carefully and looking at the final plane:
- The initial rectangular sheet is folded several times
- Each fold creates multiple overlapping layers
- The hole appears to be punched through the right wing of the completed plane

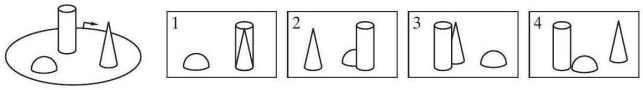
Based on the specific folding pattern shown in the diagrams, when the sheet is completely folded into the paper plane configuration, the location where the hole is punched consists of 8 layers of paper stacked together.

Therefore, when Paulo unfolds the paper, he would find 8 holes in the sheet - one for each layer that was penetrated by the single punch.

Answer: C

Figure B.33: A sample correct case of Spatial Reasoning (difficulty: Medium).

Spatial (Hard) ⊗



Question: Betty walked around the park once, starting from the marked point in the direction of the arrow. She took 4 pictures. In which order did she take the pictures?
Options: A: 1234. B: 1324. C: 2143. D: 2341

Ground Truth: (C)

Claude-3.7-Sonnet-Thinking: <*thinking*> This problem requires me to determine the order in which Betty took 4 pictures as she walked around a park. I need to look at the diagram and figure out the sequence.

First, I'll analyze what I see in the leftmost diagram, which seems to show the park layout:
- There's a cylindrical object (like a column or tree trunk)
- There's a cone-shaped object (like a pine tree or monument)
- There's a dome-shaped object (like a small hill or rock)
- There's an arrow indicating the direction Betty walked

Then I have 4 numbered pictures, each showing different arrangements of these objects:
- Picture 1: Shows the dome shape on the left and the cylinder on the right
- Picture 2: Shows the cone shape on the left and the cylinder on the right
- Picture 3: Shows the cylinder on the left and the dome shape on the right
- Picture 4: Shows the cylinder on the left and both the dome shape and cone shape on the right

I need to determine the order in which Betty would see these views as she walks around the park in the direction of the arrow.

Looking at the main diagram, the arrow indicates she's walking counterclockwise around the park.

If I start at the marked point and move in the direction of the arrow (counterclockwise):
1. Betty would first see the dome on her left and the cylinder on her right - this matches Picture 1
2. As she continues counterclockwise, she would next see the cone on her left and the cylinder on her right - this matches Picture 2
3. Continuing further, she would see the cylinder on her left and the dome on her right - this matches Picture 3
4. Finally, completing the circle, she would see the cylinder on her left with both the dome and cone on her right - this matches Picture 4

So the order of the pictures would be 1, 2, 3, 4, which corresponds to option (A) 1234.

Let me double-check this logic:

From the starting position and moving counterclockwise:
- First view: dome on left, cylinder on right (Picture 1)
- Second view: cone on left, cylinder on right (Picture 2)
- Third view: cylinder on left, dome on right (Picture 3)
- Fourth view: cylinder on left, dome and cone on right (Picture 4)

This confirms that the answer is (A) 1234.
<end_thinking>
Determining Picture Order in Betty's Park Walk

I need to figure out the order in which Betty took 4 pictures as she walked once around the park in the direction of the arrow.

Analyzing the Park Layout
From the first diagram, I can see:
- A cylindrical object (looks like a column/tree trunk)
- A cone-shaped object (like a pointed tree or monument)
- A dome-shaped object (like a small hill or rock)
- An arrow indicating counterclockwise movement

Analyzing Each Picture
1. **Picture 1**: Shows the dome shape on left, cylindrical object on right
2. **Picture 2**: Shows the cone shape on left, cylindrical object on right
3. **Picture 3**: Shows the cylindrical object on left, dome shape on right
4. **Picture 4**: Shows the cylindrical object on left, both dome and cone shapes on right

Determining the Sequence
Starting from the marked point and moving counterclockwise (following the arrow):
1. First, Betty would see the dome on her left and the cylinder on her right → **Picture 1**
2. Continuing counterclockwise, she would next see the cone on her left and the cylinder on her right → **Picture 2**
3. Moving further, she would see the cylinder on her left and the dome on her right → **Picture 3**
4. Finally, she would see the cylinder on her left with both the dome and cone on her right → **Picture 4**

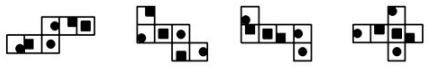
This gives us the sequence 1-2-3-4.

Answer: A

Error Category: Spatial / Orientation Error
Error Reason: The model makes a spatial orientation error by incorrectly interpreting the shifting perspectives as Betty walks around the park, especially at the starting point. It assumes a consistent viewpoint but fails to mentally rotate the scene to match what Betty would actually see from each new position.

Figure B.34: A sample error case of Spatial Reasoning (difficulty: Hard).

Spatial (Hard)



Question: From the flat expansion view of paper boxes, which paper box is different from the other three?

Ground Truth: (A)

Claude-3.7-Sonnet-Thinking: <*thinking*> This problem is asking me to identify which one of the four flat expansions of paper boxes is different from the others. Let me analyze each one carefully.

These patterns represent unfolded 3D boxes (nets), where the squares would fold to create a 3D shape. The black squares and circles represent some kind of marking or pattern on the surface of the box.

Let me analyze the structure of each expansion:
A: This appears to be a linear arrangement of 5 squares with a T-shape extension. The pattern has black regions in various squares.
B: This is a more complex shape with 7 squares arranged in a cross-like pattern. It has black regions in various squares.
C: This also appears to be an arrangement of squares that could fold into a box, with black markings.
D: This is another cross-shaped arrangement of squares with black markings.

To determine which one is different, I need to see if they would all fold into valid 3D boxes and if there's any pattern that sets one apart.

Let me think about how these would fold:
- A cube has 6 faces, so a valid net should have exactly 6 squares
- The squares must be connected in a way that allows folding into a cube

Let me count the squares in each:
A: I count 5 squares
B: I count 7 squares
C: I count 7 squares
D: I count 7 squares

Looking at the number of squares, A has 5 squares while B, C, and D each have 7 squares. A cube needs 6 faces, so none of these would fold into a perfect cube. A is different in that it has fewer squares than the others.

However, let me look more carefully at these patterns to see if there's another difference.

Actually, let me recount:
A: Looking more carefully, I see 6 squares arranged in a T-pattern with an extension
B: I see 7 squares arranged in a cross-pattern with extensions
C: I see 7 squares arranged in a pattern
D: I see 7 squares arranged in a cross-pattern

So option A has 6 squares, while B, C, and D each have 7 squares. Since a cube has 6 faces, A is the only one that could potentially fold into a proper cube. The others have an extra square.

Let me check once more carefully:
A: I count squares from left to right and top to bottom: 1, 2, 3, 4, 5, 6. So there are 6 squares.
B: I count 7 squares.
C: I count 7 squares.
D: I count 7 squares.

So based on the number of squares, A is different from the others because it has 6 squares while the others have 7 squares.

But let me think about whether there could be another difference that's more subtle, like the pattern of black regions or the connectivity of the squares.

Looking at the patterns of black and white regions, they all seem to have black regions and circles in different arrangements.
Looking at the connectivity of the squares, they all have different patterns of how the squares connect.

I think the key difference is the number of squares. A has 6 squares which is the exact number needed for a cube. B, C, and D all have 7 squares, which is one too many for a standard cube - these would need to have one square overlapping or not used when folded.

So my answer is A.
<*end_thinking*>
Analysis of Paper Box Flat Expansions

I need to determine which of these flat expansions (nets) is different from the others by examining their structure and properties.

First, let me count the number of squares in each option:
- Option A: 6 squares arranged in an L-shape with an extension
- Option B: 7 squares arranged in a cross pattern with extensions
- Option C: 7 squares arranged in a T-shape with extensions
- Option D: 7 squares arranged in a cross pattern with extensions

The key difference I notice is that option A has exactly 6 squares, while all the other options (B, C, and D) have 7 squares each.

This is significant because a cube has exactly 6 faces, so a proper net for a cube should have exactly 6 squares. Options B, C, and D all have one extra square, which means they would either:
1. Not fold into a proper cube
2. Have an overlapping square when folded
3. Require leaving one square unused

Therefore, option A is different from the other three because it's the only one with the correct number of squares (6) to form a standard cube.

Answer: A

Figure B.35: A sample correct case of Spatial Reasoning (difficulty: Hard).

Appendix C

Appendix for Chapter 5

C.1 Related Work

Visual Instruction Tuning. Visual instruction tuning is a key technique for enhancing multimodal large language models by aligning visual inputs with textual instructions to improve understanding and generation tasks [79]. Traditionally, these instructions are built using English-language data from visual question answering and other datasets [11, 28, 72, 78, 79, 80, 130, 130, 142, 165]. Researchers often supplement this with synthetic instruction tuning data, generating large volumes of instructional pairs to possibly cover multiple languages too [41, 71, 72]. However, these instruction-tuning datasets have mostly been task-focused and lack conversational capabilities. Further, while translation gives lends to multilingual capabilities, the data remains to be culturally homogeneous. By curating multilingual and multicultural instruction tuning data across various task types, our model is designed to intuitively understand and engage with users from diverse demographics.

Dataset	# Languages	# of Instances	Multicultural	# of Task Types	Open-Sourced
MultiInstruct [142]	1	~235.0K	✗	310	✗
MiniGPT4 [165]	1	5.0K	✗	149	✓
LLAVA [79]	1	1.2M	✗	>100K	✓
InstructBLIP [28]	1	~1.6M	✗	>100K	✗
M^3IT [71]	80	2.4M	✗	400	✓
mBLIP [40]	95	5.1M	✗	68	✓
PALO [110]	10	2.1M	✗	22	✓
Cambrian [130]	1	7.1M	✗	>1M	✓
PANGEAINS (Ours)	39	6.2M	✓	>1M	✓

Table C.1: Comparison of datasets in terms of number of languages, number of instances, whether the dataset is multicultural, number of task types, and open-sourced.

Multilingual Multimodal LLMs. Multilingual MLLMs have evolved from dual-encoder-based models, only capable of understanding and reasoning [56, 94, 156], to encoder-decoder models capable of multilingual text generation as well [23, 41, 117]. Despite their advancements, these models have remained focused on conventional tasks such as VQA and image captioning. Moreover, most efforts have centered around training with multilingual text, while little attention has been given to curating culturally diverse image datasets. Even for text, despite the focus on multilinguality, few attempts have been made to reflect cultural diversity in instructions and captions. As a result, these models tend to reflect a Western-centric bias. By selecting culturally diverse images from LAION and intentionally integrating this diversity into our instructions and captions, our model aims to serve a wide range of users in an inclusive and equitable manner.

C.2 Prompts used in the data construction

In this appendix, we will list the detailed prompts we used when constructing cultural understanding instruction tuning data described in subsection 5.2.2.

Cultural Images LLM Scoring Prompt

You are given an [Alt Text] associated with an image from the web.

[Alt Text]: {Alt Text}

Your goal is to:

1. **Evaluate Text Quality:** Rate the following alt text on a scale from 1 to 5 based on its quality in describing the image, assuming the model does not have access to the image:

- 1 (Very Low Quality): Alt text is vague, irrelevant, misleading, or uses placeholders (e.g., file names).
- 2 (Low Quality): Alt text is overly simplistic, generic, or provides minimal useful information.
- 3 (Moderate Quality): Alt text is somewhat descriptive but lacks detail or relevance, with possible redundancy or ambiguity.
- 4 (High Quality): Alt text is descriptive, clear, concise, and provides sufficient information to understand the image's content.
- 5 (Very High Quality): Alt text is highly specific, detailed, and relevant, with a clear description that conveys all key aspects of the image.

2. **Subject Classification:** Assign a subject/category to the alt text based on its content. Choose from the following categories:

- Vehicles and Transportation
- Cooking and Food
- People and Everyday Life
- Sports and Recreation
- Plants and Animals
- Objects, Materials, and Clothing
- Brands and Products
- Geography, Buildings, and Landmarks
- Tradition, Art, and History
- Public Figure and Pop-Culture
- Others

Cultural Images LLM Scoring Prompt (Continued)

3. **Country/Region Classification:** Decide if the alt text is closely related to a specific country's culture. For example, if the alt text says, "Tokyo Skytree Photo in March with beautiful cherry blossoms", it's strongly related to Japan. If the alt text is not specifically about a certain culture or country, you can say "No specific country". Even if the alt text is written in their official language, it doesn't mean the caption is specifically about the country (e.g., a product page caption is often unlikely to be country-specific).

Output: Provide the final result in the following structured format:

1. **Text Quality Score (1-5):**
2. **Subject Category:**
3. **Country/Region:**

Only generate the final result without any additional descriptions or explanations.

Image Recaption Prompts

We randomly select one recaption prompt from the following:

PROMPT 1:
Please describe the image in detail in {language}. The image might be related to the country: "{country}". The topic might be related to: "{category}". The previous short caption of the image is {text}.

PROMPT 2:
Analyze this image and provide a comprehensive description in "{language}". Consider that it may be associated with "{country}" and the theme could be related to "{category}". If there is cultural significance, please include it. A brief previous description was: {text}.

PROMPT 3:
In "{language}", give a detailed description of what you see in this image. Keep in mind it might be connected to "{country}" and the subject could be about "{category}". If there are culturally relevant details, please include them. An earlier short description stated: {text}.

PROMPT 4:
Examine this image closely and describe its contents in "{language}" in a more structured way. The image might have a connection to "{country}" and could be about "{category}". A previous concise caption mentioned: {text}.

PROMPT 5:
Using "{language}", provide an in-depth and structured description of this image. It may be related to "{country}" and the topic could be associated with "{category}". A prior brief description was given as: {text}.

Instruction Generation Prompt

Task: Generate two **instruction-response pair** based on the visual content of an image. Choose two task types from the list below to guide the rewriting process:

- Coding & Debugging
- Information Seeking
- Creative Writing
- Critical Reasoning
- Planning & Strategy
- Mathematical Thinking
- Text Revision & Editing
- Data Analysis
- Role Playing & Scenarios
- Brainstorming & Ideation
- Advice Seeking & Problem-Solving
- Learning & Understanding
- Cultural Interpretation

Guidelines:

Instruction:

- Select two different task types from the list above.
- Make sure the instruction prompts an interpretation or analysis **directly tied to what can be visually observed in the image**, not just general reasoning.
- The instruction should require a response that **uses details from the image**. Avoid generic instructions that can be answered without visual information.

Response:

- Provide a **very detailed and structured** response that reflects a clear understanding of the implied visual information.
- Offer multiple perspectives, deep analysis, or step-by-step explanations where applicable.
- Avoid general responses that could be inferred without observing the image. Responses must rely on interpreting the visual content.

Instruction Generation Prompt (Continued)

Content:

- Instructions should be varied, challenging, and explore different advanced aspects of the visual scene.
- Responses must showcase a deep understanding of the image's visual context, using thoughtful insights where applicable.

Output:

- Provide the output in JSON format with three keys: "task_type", "instruction" and "response".
- Ensure the instruction and response **do not mention "based on caption"** but instead, refer to the **image** or simply avoid reference to any external description.
- Do not include additional text or explanations beyond what is required.
- Provide both the "instruction" and "response" in {language} but "task_type" in English.

Caption: {caption}

C.3 Recaptioning Example from LAION-Cultural

Original Alt Text:

Президент и CEO компании The Walt Disney у макета Шанхайского Диснейлэнда

(President and CEO of The Walt Disney Company by a model of Shanghai Disneyland)

Caption with High-quality Alt Text:
На фотографии изображён президент и CEO компании The Walt Disney, стоящий перед макетом Шанхайского Диснейленда. На заднем плане видны красочные цветочные композиции и зелёные кусты, придающие сцене праздничную атмосферу. На переднем плане находится детализированный макет с замком, украшенным множеством башен и шпилей. Президент компании улыбается и разводит руки в стороны, как бы представляя миниатюрный парк перед ним. На нём надет тёмный костюм и полосатая рубашка, что подчёркивает его официальный статус. Фон сцены оформлен синим цветом, что придаёт общей композиции контраст и глубину.

(The photograph features the President and CEO of The Walt Disney Company standing in front of a model of Shanghai Disneyland. In the background, vibrant floral arrangements and green bushes create a festive atmosphere. In the foreground, a detailed model of the park is displayed, with a castle adorned with numerous towers and spires. The President, smiling and extending his arms as if presenting the miniature park, is dressed in a dark suit and a striped shirt, emphasizing his official status. The backdrop is designed in blue, adding contrast and depth to the overall composition.

Caption without Alt Text:

На изображении мужчина в тёмном костюме, светло-голубой рубашке и без галстука, который стоит перед большим макетом замка. Замок имеет ярко выраженные готические элементы с высокими шпилями, что напоминает замки из сказок Диснея. Мужчина улыбается и раскинул руки в приветственном жесте, демонстрируя замок. На заднем плане сцены можно увидеть зелёные растения, цветы в горшках и красный фон с золотыми элементами, что создаёт праздничную атмосферу. Позади замка можно заметить синий задний фон, который придаёт всей сцене яркость.

(The image features a man in a dark suit, light blue shirt, and no tie, standing in front of a large model of a castle. The castle has pronounced Gothic elements with tall spires, reminiscent of Disney fairy tale castles. The man is smiling and has his arms spread in a welcoming gesture, showcasing the castle. In the background, green plants, potted flowers, and a red backdrop with gold elements create a festive atmosphere. Behind the castle, a blue background adds brightness to the entire scene.)

Figure C.1: An example from LAION-Cultural illustrating why the filtered informative alt text helps generate a more informative caption. With the high-quality alt text, the model incorporates important details like *"President and CEO of The Walt Disney Company standing in front of a model of Shanghai Disneyland"* into the generated caption.

C.4 Datasets used in PANGEABENCH

To comprehensively assess the capabilities of PANGEA across diverse languages, cultures, and task types, we developed PANGEABENCH. We list the details of each dataset included in the PANGEABENCH.

C.4.1 Multimodal Datasets

- **xGQA** [100]: A cross-lingual visual question-answering dataset featuring 9,666 questions in eight languages covering five scripts. The dataset includes 300 unique images from Visual Genome [64]. xGQA tests the model's ability to understand and reason about visual content across multiple languages.
- **MaXM** [20]: A VQA dataset in seven languages and five scripts, with questions and answers in the same language. Images are culturally matched to the target language regions. MaXM specifically addresses the challenge of cultural diversity in multimodal understanding.
- **MaRVL** [77]: A Multicultural Reasoning over Vision and Language dataset in five languages and three scripts, featuring 4,914 culturally diverse images matched to respective languages. MaRVL focuses on evaluating models' ability to reason about culturally diverse visual concepts.
- **XM100** [129]: We create a subset of 3600 instances (100 instances per language) from the original XM100 dataset, a large multilingual image captioning dataset comprising 36 languages, with 261,375 captions for 100 unique images per language, culturally matched to each language. XM100 evaluates a model's ability to generate culturally appropriate captions across a wide range of languages. For sampling, we select 100 instances per language, ensuring that all languages share the same set of images for their respective 100 instances. To ensure diversity within our sample, we use Sentence-BERT [112] to cluster the 3600 English instances from the original dataset into 100 groups, and then select one instance from each group. This method ensures that the sampled instances are as diverse as possible. We evaluate models on this new sample of 3600 instances, which allows for a more time-efficient evaluation while still accurately reflecting the multilingual capabilities of models in diverse contexts.
- **M3Exam** [159]: A novel benchmark sourced from real and official human exam questions, featuring 12,317 questions in 9 languages across three educational levels. Approximately 23% of the questions require image processing. M3Exam tests the model's ability to handle complex, multi-step reasoning tasks in an educational context.
- **xMMMU**: MMMU contains multimodal questions from college-level materials across six disciplines and 30 subjects. The dataset features 183 subfields and 30 diverse image types, including charts, diagrams, and chemical structures. We sample 300 questions from the original MMMU validation set and translate them using GPT-4o into xx languages. To ensure the quality, we translated each sampled question multiple times and then back-translated it to English. We select the translation with the highest BLEU score. xMMMU evaluates the model's capacity to understand and reason about specialized academic content across languages and modalities.

C.4.2 Text-Only Multilingual Datasets

- **TyDiQA** [27]: A question answering dataset covering 11 typologically diverse languages with 204K question-answer pairs. Questions are written by native speakers without seeing the answers, ensuring a realistic information-seeking task. TyDiQA is designed to test linguistic diversity and avoid translation artifacts.
- **FLORES** [95]: A machine translation benchmark for 200 languages, including many low-resource languages. It consists of 3,001 sentences from 842 web articles, divided into dev, devtest, and test splits. FLORES-200 includes translations from multiple pivot languages and provides script alternatives for some languages, making it a comprehensive test of translation capabilities.
- **MMMLU** [96]: A human-translated version of MMLU [51], covering 57 subjects across STEM, humanities, social sciences, and more. It ranges in difficulty from elementary to advanced professional levels, testing both world knowledge and problem-solving ability in a zero-shot and few-shot setting across multiple languages.
- **MGSM** [119]: Multilingual Grade School Math Benchmark, featuring 250 grade-school math problems translated into 10 languages. Based on GSM8K, it requires multi-step reasoning and tests the model's ability to solve complex mathematical word problems across languages.

This diverse set of datasets in PANGEABENCH allows for a comprehensive evaluation of PANGEA's capabilities across various languages, cultures, modalities, and task types, providing a holistic assessment of its performance in multilingual and multimodal contexts.

C.5 Explanation of xChatBench

Task Category We first divide into 10 task categories, namely *art_explanation*, *bar_chart_interpretation*, *defeasible_reasoning*, *figurative_speech_explanation*, *iq_test*, *ocr*, *graph_interpretation*, *image_humor_understand* *science_figure_explanation*, *unusual_images*. The task categories are inspired by existing papers that do not use a free-form generation format [48, 52, 58, 88, 153].

Construction Procedure To annotate the instances, we mainly follow the procedure of Kim et al. [60]. Two human annotators first hand-crafted the instances by searching through appropriate images for the task and then hand-crafting each component of the instance. As our motivation for fine-grained evaluation, each instance consists of not only an **instruction**, **reference answer**, but also a unique **evaluation criteria** tailored to each instance (*e.g.*, Does the response effectively explain the humor in the image based on the juxtaposition of a character's portrayal in different scenarios?) and a **description for each score** between 1 and 5 (*e.g.*, score4_description: The response understands the juxtaposition and relates it to the humor involving machine learning models, but may miss some nuances or the related aspect of the humor). During the annotation process, we asked the annotators to not copy-and-paste results from LLM services like ChatGPT or directly from the web. Then, we hire four additional annotators to assess the quality of the instances. Each participant to asked to grade if each instance (1) fits into the devised task category, (2) if the quality of the reference answer is good enough, and (3) if the score rubric is suitable to assess the response. We iteratively ask the annotators who made the instances to revise them if the instance does not satisfy all three criteria. The resulting dataset consists of 50 instructions, reference answers, and evaluation criteria with a corresponding score rubric.

Translation Procedure To assess the multilingual generation capabilities of MLLMs, we translate the hand-crafted 50 instances into 6 different languages, namely Chinese, Hindi, Indonesian, Japanese, Korean, and Spanish. We first use GPT-4o-2024-08-06 to translate the instruction and reference answer of each instance with a naive prompt, "Translate the following sentences into {target_language}. Sentences: {sentences}". Then, the coauthors who are native speakers of each language reviewed the instances and made adjustments if the translated results were unnatural.

Evaluation Pipeline Similar to prior works employing LLM-as-a-Judge, we use GPT-4o-2024-08-06 as the judge model and prompt it in a direct assessment manner. As input, the judge model is given the instruction, the model's response, the reference answer, the evaluation criteria, and the descriptions for each score. As output, the judge generates verbal feedback and an integer score between 1 and 5. For this procedure, we use the `prometheus-eval` library [61] and employ their default hyperparameter setting for evaluation. Lastly, the final score is acquired by averaging the results across the 50 instances for each language. Note that in the main result and breakdown result tables, we normalize the score from 1-5 to 0-100 by $(\text{score} - 1) \times 25$. For the multimodal chat scenarios, we found that many English-centric models tend to respond in English regardless of the query language. This behavior is problematic, as it undermines the fundamental capability of a multilingual model, which should ideally respond in the language

of the query. To address this, we implemented a strict evaluation criterion where such responses were penalized and assigned a score of 0. We believe this is crucial, as users may not understand English, and failing to respond in the appropriate language can hinder effective communication and user experience. Thus, for postprocessing, we use langdetect [1] to identify whether the response is written in the given language and change the score to 1 when it is written in a different language, a phenomenon called *language hallucination* [102, 143].

[1] https://pypi.org/project/langdetect/

C.6 Qualitative Examples from xChatBench

One important application of MLLMs is to answer users' queries in the wild. Here, we show the outputs of PANGEA for the multimodal chat queries from our xChatBench. The examples included the scoring rubric, query, response from our PANGEA, reference answer, and LLM-as-Judge feedback. As shown in Appendix Figure C.2, C.3, C.4, C.5, C.6, C.7, PANGEA successfully interprets the figures in different tasks and generates fluent and readable in certain languages. These qualitative examples further demonstrate the remarkable visual understanding ability of PANGEA in multilingual contexts. On the other hand, we also identified a few bad cases shown in Figure C.8, C.9. Despite generating relevant responses to the queries, the model does not capture the key details of the images due to the lack of knowledge, which points out potential improvement directions in the future.

Score Rubric

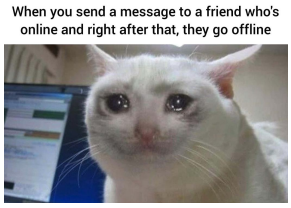
Does the response comprehensively analyze the humorous elements present in the image featuring an animal with human-like emotions?

Score 1: The response does not recognize the anthropomorphism or the situational context, offering no insight into the humor.
Score 2: The response identifies the animal's human-like emotions but fails to connect these to the situational context or the cultural nuances that contribute to the humor.
Score 3: The response notes the anthropomorphism and situational context but provides only a basic explanation of how these elements combine to create humor.
Score 4: The response gives a good explanation of the humor, recognizing both the anthropomorphism and situational context, yet it might miss finer cultural nuances.
Score 5: The response offers a detailed analysis, accurately recognizing the anthropomorphism, situational context, and cultural nuances that all contribute to the humor.

Instruction

動物が人間のような感情を持つ様子をユーモラスな文脈で描かれた画像を見るとき、その画像に存在する文化的なニュアンス、背景情報、または状況的な文脈を考慮しながら、コメディ効果に寄与する根本的な要素を分析してください。

When looking at an image that humorously portrays animals with human-like emotions, analyze the fundamental elements contributing to the comedic effect, while considering the cultural nuances, background information, or situational context present in the image.



Response from Pangea

画像に寄与するコメディ効果にはいくつかの要素があります。第一に、猫は人間のように泣いている表情で描かれており、これはユーモラスな対比です。第二に、猫がスマートフォンを操作している様子は、デジタルコミュニケーションでよく見られる人間の行動を描いています。第三に、猫が「メッセージ」を受け取った後にもうすぐ「オフライン」となりますというこの行動は、猫がメッセージを読んですぐに返事をしようと努力していることを示唆しており、これは人間の行動であり、特に猫はそうしないとされるため、面白い誤解です。これらの要素は、猫がインターネットやメッセージングアプリなどのデジタルコミュニケーションの文脈において、人間と同じように行動し、同じような感情を持っているかのような、ユーモラスで過度化されたシーンを作り出しています。

There are several elements contributing to the comedic effect in the image. First, the cat is depicted with a crying expression, much like a human, which creates a humorous contrast. Second, the cat is operating a smartphone, portraying a behavior often seen in human digital communication. Third, the action of the cat soon going "offline" after receiving a "message" suggests that it is trying to respond immediately, mimicking human behavior, which is amusingly mistaken since cats are not known to do this. These elements together create a humorous, exaggerated scene where the cat acts and feels like a human in the context of internet and messaging apps, which adds to the comedy.

Reference Answer

画像には、涙ぐんだ目によって強調された、悲しみや失望の人間のような表情をした猫が描かれています。ユーモアは、人間ではないものに人間の特性を与えるという一般的なコメディ手法である擬人化から生まれます。「オンラインの友達にメッセージを送った直後に、その友達がオフラインになるとき」というキャプションが付けられています。この状況は、デジタルコミュニケーションにおいてよくあるシナリオであり、相手がすぐにオフラインになることで、交流の試みが無視されたり軽視されたりしたと感じることがあります。これは、即時のコミュニケーションの期待と、しばしば暗黙のデジタルエチケットを利用した状況的アイロニーです。\n\nこの状況の文化的な関連性によって、コメディ効果はさらに高まります。多くの人がデジタル時代における会話の突然の終わりを経験しています。画像は猫の表情を巧みに利用して、その瞬間の感情的なインパクトを誇張し、テクノロジーが私たちの社会的交流に与える影響についての共通理解に訴えかけています。インターネット文化の重要な一部であるミーム形式自体が、しばしばテキストと画像を組み合わせて多面的なユーモアを伝える、共感できるコンテンツを広めるために設計されています。したがって、猫の人間化された表情、一般的なオンライン交流へのキャプションの言及、デジタルコミュニケーションにおけるミームの位置を認識することで、この画像が意図するユーモアの全体像を理解することができます。

The comedic effect is further enhanced by the cultural relevance of this situation. Many people have experienced the abrupt end of a conversation in the digital age. The image cleverly uses the cat's expression to exaggerate the emotional impact of that moment, appealing to a shared understanding of how technology affects our social interactions. The meme format itself, an essential part of internet culture, is often designed to combine text and images to convey multi-layered humor and spread relatable content. Therefore, by recognizing the anthropomorphized expression of the cat, the caption's reference to common online interactions, and the role of memes in digital communication, we can fully grasp the humor intended by this image.

Feedback from GPT-4o

The response identifies various elements that contribute to the comedic effect in the image, such as the cat displaying a human-like crying expression and interacting with a smartphone. It recognizes the anthropomorphism and outlines a scenario where the cat behaves as if it were in a human situation related to digital communication. However, while there is a description of the elements and the anthropomorphic humor, the analysis could be strengthened by delving deeper into cultural nuances and how they enhance humor, as well as the broader social commentary on digital communication. While the response adequately covers the situational context, it doesn't thoroughly explore the cultural aspects to the same extent as the reference answer.

[RESULT] 4

Figure C.2: An xChat example from Japanese subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

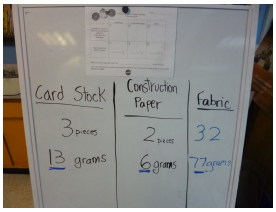
Does the response accurately transcribe all visual information from the image, including specific details such as quantities, weights, underlines, and ink colors?

Score 1: The response recognizes the presence of text but fails to accurately transcribe any of the materials and their corresponding quantities and weights, for example, quantities are present but weights are not present.
Score 2: The response accurately transcribes the information for one of the materials listed, including quantity and weight, but fails to do so for the other two, for example, 'Card Stock' data is correct but 'Construction Paper' and 'Fabric' data are missing or incorrect.
Score 3: The response accurately transcribes the information for two of the materials listed, including quantities and weights, but fails to do so for the third, for example, 'Card Stock' and 'Construction Paper' are correct, but 'Fabric' is not.
Score 4: The response accurately transcribes the quantities and weights for all materials but misses secondary details such as underlines or ink colors, for example, all weights are correct but underlines and ink colors are not mentioned.
Score 5: The response accurately transcribes all primary information (quantities and weights) and secondary details (underlines, ink colors) for all materials, with no omissions.

Instruction

डिज़ाइन चुनौती के दौरान, एक बोर्ड पर विभिन्न सामग्रियों और उनकी संबंधित मात्राएँ और वजन प्रदर्शित किए गए थे रिकॉर्ड रखने के लिए एक तस्वीर ली गई थी हालाँकि, छवि थोड़ी धुंधली है और हस्तलेखन शैली और आकार में भिन्न है छवि की जाँच करें और प्रदर्शित जानकारी का स्पष्ट और सटीक लिप्यंतरण प्रदान करें, जिसमें कार्ड स्टॉक, निर्माण कागज, और कपड़े की मात्राएँ और वजन शामिल हों

During a design challenge, various materials and their respective quantities and weights were displayed on a board. A picture was taken for record-keeping, but the image is slightly blurry, with variations in handwriting style and size. Please examine the image and provide a clear and accurate transcription of the displayed information, including the quantities and weights of card stock, construction paper, and fabric.



Response from Pangea

सामग्री प्रदर्शन स्क्रीन पर निम्नलिखित जानकारी है:\n\n| मात्रा | कार्ड स्टॉक | निर्माण कागज | कपड़े |\n|---|---|---|---|\n| मात्रा | 3 | 2 | 32 |\n| वजन (ग्राम) | 13 | 6 | 77 |\n\nसाइनेज से, हम निष्कर्ष निकाल सकते हैं कि परीक्षण में कार्ड स्टॉक की तीन टुकड़ों का वजन 13 ग्राम, निर्माण कागज की दो टुकड़ों का वजन 6 ग्राम, और कपड़े की अग्रणी छह-अंक संख्या का वजन 77 ग्राम है व्यवस्था बताती है कि प्रत्येक विभाग सामग्री के वजन में योगदान को दर्शाता है, लेकिन वजन आकार चर के रूप में व्यक्त किए गए हैं, जो संभवतः कोई माप या वर्ग का प्रतिनिधित्व करता है सटीक आयाम या सामग्री का उपयोग के बारे में अधिक जानकारी के बिना, यह निर्धारित करना संभव नहीं है कि किस परीक्षण के लिए डेटा प्रस्तुत किया गया है या कौन सी सामग्री को संदर्भित किया गया है

| Quantity | Card Stock | Construction Paper | Fabric |\n|----------|------------|--------------------|--------|\n| Quantity | 3 | 2 | 32 |\n| Weight (grams) | 13 | 6 | 77 |\n\nFrom the signage, we can conclude that three pieces of card stock weigh 13 grams, two pieces of construction paper weigh 6 grams, and the leading six-digit number for the fabric weighs 77 grams. The arrangement suggests that each section represents the material's contribution to the overall weight, but the weight is expressed as a variable, possibly representing some form of measurement or category. Without more precise dimensions or information on the materials used, it is not possible to determine what specific test the data pertains to or which materials are being referenced.

Reference Answer

छवि एक व्हाइटबोर्ड दिखाती है जिस पर मार्कर से जानकारी लिखी हुई है इसमें तीन कॉलम हैं, प्रत्येक एक अलग सामग्री के लिए समर्पित है: कार्ड स्टॉक, निर्माण कागज, और कपड़ा प्रत्येक सामग्री के लिए, मात्रा (टुकड़ों में) और वजन (ग्राम में) दिए गए हैं \n\nकार्ड स्टॉक\n\nमात्रा: 3 टुकड़े\nवजन: 13 ग्राम\nनिर्माण कागज\n\nमात्रा: 2 टुकड़े\nवजन: 6 ग्राम\nकपड़ा\n\nमात्रा: 32\nवजन: 77 ग्राम\nवजनों को जोर देने के लिए रेखांकित किया गया लगता है 'कार्ड स्टॉक' और 'निर्माण कागज' के लिए लिखावट काले स्याही में है, और संख्याएं भी काले रंग में लिखी गई हैं 'कपड़ा' शब्द और इसके संबंधित संख्याएं नीले स्याही में लिखी गई हैं फॉन्ट शैलियाँ आकस्मिक और आकार में विविध हैं, लेकिन जानकारी फिर भी पठनीय है व्हाइटबोर्ड के शीर्ष दाएं कोने में एक मुद्रित फॉर्म भी है, लेकिन इस फॉर्म का पाठ उस कोण और दूरी के कारण पठनीय नहीं है जिस पर फोटो लिया गया था

The image shows a whiteboard with information written in marker. There are three columns, each dedicated to a different material: card stock, construction paper, and fabric. For each material, the quantity (in pieces) and weight (in grams) are given.\nCard Stock:\n * Quantity: 3 pieces\n* Weight: 13 grams\n\nConstruction Paper\n* Quantity: 2 pieces\n* Weight: 6 grams\n\nFabric\n* Quantity: 32\n* Weight: 77 grams\n\nThe weights appear to be underlined for emphasis. The writing for "Card Stock" and "Construction Paper" is in black ink, and the numbers are also written in black. The word "Fabric" and its related numbers are written in blue ink. The font styles are casual and vary in size, but the information is still readable. In the top right corner of the whiteboard, there is a printed form, but the text on the form is not readable due to the angle and distance at which the photo was taken.

Feedback from GPT-4o

The response provides an accurate transcription of the primary information, namely the quantities and weights for all three materials: card stock, construction paper, and fabric. However, it does not mention any secondary details such as whether weights are underlined or the specific ink colors used in the original board, which are valuable details as indicated in the score rubric for a perfect score. This lack of mention of secondary details like underlines and ink colors prevents it from receiving a higher score.

[RESULT] 4

Figure C.3: An xChat example from Hindi subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

Does the response effectively explain the historical context of the phrase 'KEEP CALM AND CARRY ON' and analyze the symbolism of the elements within the image?

Score 1: The response mentions the phrase but fails to provide any historical context or symbolic analysis.
Score 2: The response provides a generic historical context without specific details or a clear explanation of the symbolism.
Score 3: The response gives an accurate historical context or symbolic analysis, but not both. For example, it may mention World War II but not the significance of the crown or color.
Score 4: The response provides a detailed historical context and a basic symbolic analysis. For example, it explains the phrase's origin in World War II and the general significance of the crown, but not the color red or its modern cultural impact.
Score 5: The response offers a comprehensive explanation of both the historical context and the symbolism. For example, it details the phrase's creation during World War II, the authority implied by the crown, the urgency indicated by the color red, and the phrase's modern-day relevance.

Instruction

제시된 시각적 메시지를 분석할 때, 어떤 역사적 맥락이 이러한 문구를 발생시켰을 수 있으며, 이 맥락이 이 이미지 내 요소들의 상징성과 사용과 어떻게 연결되는가?

When analyzing the given visual message, what historical context might have generated this phrase, and how is this context connected to the symbolism and usage of the elements within the image?

Response from Pangea-7B

\"KEEP CALM AND CARRY ON\"이라는 문구는 영국 왕실이 2차 세계대전 중 국가 정신을 높이기 위해 발행한 포스터에서 유래했습니다. 이 문구를 발생시킨 역사적 맥락은 위기 시대에 국민들에게 격려와 안정을 제공하기 위한 것이었습니다. 이 이미지 내 요소들은 전통적인 \"Keep Calm and Carry On\" 메시지를 유지하면서도 현대적인 느낌을 더하여 현대에도 그 목적을 유지하고 있습니다. 굵은, 대문자로 된 텍스트, 굵은 글씨체, 왕관의 상징성은 모두 원래 포스터에 있던 것과 비슷합니다. 이는 문구가 과거의 역사적 맥락에서 온 것으로서 의미가 지속되고 있으며, 현재에도 사용되고 있음을 나타냅니다.

The phrase “KEEP CALM AND CARRY ON” originates from a poster issued by the British royal family during World War II to boost national morale. The historical context that produced this phrase was aimed at providing encouragement and stability to the public during a time of crisis. The elements within this image maintain the traditional “Keep Calm and Carry On” message while adding a modern feel, suggesting that the original purpose remains relevant today. The bold, uppercase text, thick font, and the symbolic crown are all similar to those in the original poster. This indicates that the phrase, rooted in its historical context, continues to hold meaning and is still in use today.

Reference Answer

이미지는 밝은 빨간색 배경에 왕실 문장이 얹혀진 \"KEEP CALM AND CARRY ON\"이라는 대담하고 대문자로 된 문구를 묘사하고 있습니다. 이 상징적인 포스터는 제2차 세계 대전 초기에 영국 역사에 뿌리를 두고 있습니다. 이는 1939년 영국 정부에 의해 잠재적인 전시 재난에 대비하여 대중의 사기를 북돋우기 위한 메시지로 원래 제작되었습니다.\n\n\"KEEP CALM AND CARRY ON\"이라는 은유는 다면적입니다. 이는 대중에게 혼란 속에서도 침착함과 정상성을 유지하라고 촉구하는 회복력과 강인함의 메시지를 담고 있습니다. 왕관의 사용은 단순히 장식적인 것이 아니라 군주제를 상징하며, 영국 국민의 집단적 정체성에 대한 호소와 함께 단결과 국가적 자부심을 불러일으킵니다. 메시지 위의 왕관은 이 지시가 군주제와 관련된 권위와 안정성을 가지고 있다는 것을 암시적으로 나타냅니다.\n\n빨간색은 주목을 끌며 경계와 긴급함과 자주 연관되는데, 이는 전시 중에 적절한 감정이었을 것입니다. 그러나 그 자체로 영령형인 문구는 배경의 경고하는 색조와 대조적으로 침착한 숙고의 감각을 심어주려 합니다. 이 대조는 행동의 긴급함이 정신의 평온함에 의해 완화되는 은유의 깊이를 강조합니다.\n\n현대 문화에서 밈과 장식 모티프로서의 문구의 지속성과 부활은 그 은유적 풍부함을 더욱 강조합니다. 이는 전쟁의 맥락뿐만 아니라 일상적인 도전에 대한 냉철한 대응을 상징하게 되었으며, 그 원래 목적을 초월하는 시대를 초월한 메시지를 담고 있습니다. 여기서 은유는 인생의 어떤 어려움에도 개인의 회복력으로 확장되어 보편적으로 적용 가능한 사고방식을 상징합니다.\n\n본질적으로 이 포스터는 국가의 역사적 회복력을 전달하면서 집단적이고 개인적인 도전에 대한 침착하고 확고한 접근 방식을 제공하는 지속적인 은유를 전달하는 유물입니다.

The image depicts the bold, uppercase phrase “KEEP CALM AND CARRY ON” set against a bright red background with a royal crest. This iconic poster is rooted in British history during the early days of World War II. It was originally created in 1939 by the British government to boost public morale in the face of potential wartime disaster.\nThe metaphor of “KEEP CALM AND CARRY ON” is multifaceted. It conveys a message of resilience and strength, urging the public to maintain calm and normalcy amidst chaos. The use of the crown is not merely decorative but symbolizes the monarchy, evoking unity and national pride within the collective British identity. The crown above the message implicitly suggests that the directive carries the authority and stability associated with the monarchy.\nRed is a color that grabs attention and is often linked to caution and urgency, which would have been appropriate during wartime. However, the commanding phrase, in contrast to the warning-like tone of the background, aims to instill a sense of calm reflection. This contrast highlights the depth of the metaphor, where the urgency of action is tempered by a sense of mental composure.\nIn modern culture, the phrase’s persistence and revival as a meme and decorative motif emphasize its metaphorical richness. It has come to symbolize a calm response not only in the context of war but also to everyday challenges, representing a timeless message that transcends its original purpose. Here, the metaphor extends to symbolize personal resilience in the face of life’s difficulties, making it universally applicable.\nEssentially, this poster serves as an artifact that communicates the historical resilience of a nation while offering a calm, steadfast approach to both collective and personal challenges.

Feedback from GPT-4o

The response begins by providing the historical context of the phrase \"KEEP CALM AND CARRY ON,\" correctly stating its origin during World War II, and explaining its purpose to boost national morale in a time of crisis. This establishes a basic historical framework for understanding the significance of the phrase. \n\nIn terms of symbolic analysis, the response does acknowledge elements such as the bold, uppercase text, bold font, and the crown's symbolism. However, the analysis remains somewhat superficial. It does not delve into the specific symbolic significance of the color red or provide a detailed explanation of how these elements function within the historical and modern contexts. Furthermore, it does touch on the modern relevance of the phrase but only in vague terms, without explaining how the symbolism translates into contemporary culture or addressing the broader implications noted in the reference answer.\n\nOverall, while the response demonstrates an understanding of the historical origins and some elements of symbolism, it lacks the depth and comprehensiveness of the reference answer.

[RESULT] 4

Figure C.4: An xChat example from the Korean subset and our model PANGEA’s response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

Does the response insightfully analyze the underlying feelings and interactions at the festive gathering?

Score 1: The response briefly acknowledges the festive setting but fails to delve into the emotions or interactions, for example, noting the Christmas hats but not the expressions or body language of joy and camaraderie.
Score 2: The response recognizes happiness in the gathering but does not connect it to broader social dynamics, for example, mentioning smiling faces but not the shared amusement or sense of belonging.
Score 3: The response describes surface-level interactions and some emotions, such as laughter, but lacks depth in exploring the significance of these in the context of human connection, for example, noting the toasting but not the deeper sense of unity it signifies.
Score 4: The response captures most of the emotional essence and social interactions, like engagement and warmth, but may not fully interpret the broader human experience, for example, recognizing close proximity but not fully discussing the collective mood it creates.
Score 5: The response provides a comprehensive analysis of the feelings and interactions, deeply understanding the joy, belonging, and human connection present at the event, reflecting on how these elements combine to create a rich, shared human experience.

Instruction

Dalam sebuah pertemuan meriah yang tertangkap, apa yang mungkin menjadi perasaan dan interaksi mendasar antara orang-orang yang hadir, dan bagaimana hal ini mencerminkan emosi manusia yang lebih luas dan dinamika sosial pada acara perayaan?

In a festive gathering captured in an image, what might be the underlying feelings and interactions among the people present, and how does this reflect broader human emotions and social dynamics at celebratory events?

Response from Pangea-7B

Berdasarkan gambar, perasaan dan interaksi mendasar di antara orang-orang di sana tampak penuh kebahagiaan dan keterhubungan. Senyum lebar wanita dan gestur mesra pria menunjukkan suasana gembira dan intim yang sering terjadi dalam pertemuan sosial selama perayaan. Interaksi seperti ini mencerminkan emosi manusia universal pada saat mereka merayakan - rasa persahabatan, kesenangan bersama, dan kenikmatan dari pengalaman bersama. Momen-momen semacam itu adalah ciri khas dari acara sosial dan merupakan indikasi hubungan dan empati dalam konteks komunitas atau kelompok. Gambar tersebut menangkap esensi dari dinamika sosial tersebut dengan mengabadikan reaksi spontan manusia dalam lingkungan perayaan yang meriah.

Based on the image, the underlying feelings and interactions among the people seem to be full of happiness and connection. The wide smile of the woman and the affectionate gestures of the man suggest a joyful and intimate atmosphere often found in social gatherings during celebrations. These interactions reflect universal human emotions when celebrating—feelings of friendship, shared enjoyment, and the pleasure of shared experiences. Such moments are typical of social events and are indicative of bonding and empathy within a community or group. The image captures the essence of these social dynamics by immortalizing spontaneous human reactions in a lively festive environment.

Reference Answer

Gambar tersebut memancarkan suasana semarak yang khas dari perayaan liburan, dengan individu-individu yang mengenakan topi Natal, menunjukkan adegan persahabatan yang ceria. Tokoh-tokoh utama, seorang pria dan wanita, terlibat dalam momen hiburan bersama, sebagaimana dibuktikan oleh senyum lebar wanita tersebut dan ekspresi halus dan puas dari pria tersebut. Mereka berdua memegang gelas, tanda bersulang sosial dan kenikmatan pesta, menunjukkan semangat perayaan yang ramah dari acara tersebut.\n\nMata wanita itu bersinar dengan tawa, mencerminkan kegembiraan dan kebebasan yang sering dirasakan selama pertemuan semacam itu. Wajahnya yang terbuka dan ekspresif menandakan rasa memiliki dan kenikmatan, sifat-sifat yang selaras dengan suasana kolektif kehangatan dan keceriaan. Keterlibatan pria dengan wanita tersebut melalui kedekatan dan bahasa tubuh menunjukkan kebutuhan manusia akan hubungan dan perayaan bersama orang lain.\n\nDi latar belakang, peserta pesta lainnya sama-sama terhanyut dalam esensi semangat perayaan. Latar belakang yang kabur dengan lampu berkelap-kelip melambangkan kilauan keajaiban liburan, dan rona emas secara keseluruhan memberikan cahaya yang tampaknya meningkatkan kehangatan emosional dari adegan tersebut. Para peserta tidak hanya hadir; mereka adalah kontributor aktif dalam jalinan pengalaman manusia bersama yang ditenun oleh acara semacam itu.\n\nGambar ini menangkap lebih dari sekadar momen pesta; ia merangkum jalinan kaya dari hubungan emosional yang secara universal beresonansi. Ini adalah pengingat akan kemanusiaan bersama yang dapat disoroti oleh acara perayaan, membawa individu-individu bersama dalam ekspresi kolektif kegembiraan, persahabatan, dan keinginan manusia untuk perayaan komunal.

The image radiates the festive spirit typical of holiday celebrations, with individuals wearing Christmas hats, displaying a scene of cheerful camaraderie. The central figures, a man and a woman, are engaged in a moment of shared amusement, as evidenced by the woman's broad smile and the man's soft, contented expression. Both are holding glasses, a sign of social toasting and party enjoyment, reflecting the friendly celebratory spirit of the event. \n\nThe woman's eyes sparkle with laughter, reflecting the joy and freedom often felt during such gatherings. Her open and expressive face signifies a sense of belonging and enjoyment, emotions that align with the collective warmth and cheerfulness of the occasion. The man's engagement with the woman through proximity and body language highlights the human need for connection and celebration with others.\n\nIn the background, other party attendees are similarly absorbed in the festive spirit. The blurred background with twinkling lights symbolizes the magic of the holiday season, and the overall golden hue adds a glow that seems to enhance the emotional warmth of the scene. The participants are not merely present; they are active contributors to the shared human experience woven together by such events.\n\nThis image captures more than just a party moment; it encapsulates the rich tapestry of emotional connections that resonate universally. It is a reminder of the shared humanity that can be highlighted by celebratory events, bringing individuals together in a collective expression of joy, friendship, and the human desire for communal celebration.

Feedback from GPT-4o

The response insightfully captures the fundamental feelings and interactions of happiness and connection present at the festive gathering. It discusses the significance of smiles and gestures in conveying joy and camaraderie, and it links these interactions to universal human emotions like friendship and shared pleasure. However, while it does touch on dynamics such as empathy and community, it doesn't fully delve into the broader human experience or collective mood in as much depth as the reference answer does. The reference answer provides a more comprehensive interpretation, connecting these elements to a shared human experience with detailed observation of facial expressions and social cues.

[RESULT] 4

Figure C.5: An xChat example from the Indonesian subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

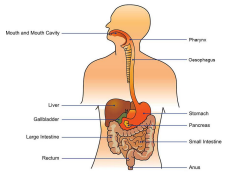
Does this response show a well-rounded preparatory plan for discussing the digestive system in a medical program context?

Score 1: The response does not include any preparatory topics or details relevant to the digestive system.
Score 2: The response mentions one aspect of digestive system study, such as physiology, without additional context or related areas of study.
Score 3: The response includes basic information on several aspects of the digestive system, such as physiology and disease, but lacks depth or the inclusion of diagnostic and treatment approaches.
Score 4: The response provides a good overview, covering multiple aspects like physiology, pathophysiology, diagnostics, and nutrition, but may lack details on pharmacology or interconnected systems.
Score 5: The response thoroughly covers all aspects of the digestive system study, including detailed physiology, pathophysiology, diagnostics, nutrition, pharmacology, and current research, showing an in-depth understanding and preparatory plan.

Instruction

Esta es una figura básica de lo que aprendí en la escuela secundaria. He ingresado a un programa médico como estudiante universitario y debería explicar a un profesor lo que me gustaría aprender más durante la universidad. ¿Podrías ayudarme a añadir más detalles sobre lo que debería buscar y examinar antes de la reunión?

This is a basic figure from what I learned in high school. I have entered a medical program as a college student and should explain to a professor what I would like to learn more about during university. Could you help me add more details on what I should look for and examine before the meeting?



Response from Pangea-7B

Basándonos en los conocimientos básicos que probablemente obtuviste en la escuela secundaria y teniendo en cuenta el contexto médico, es posible que quieras profundizar en los siguientes temas antes de tu reunión:\n\n1. **Función y estructura del sistema digestivo**: Explora cómo funciona cada órgano y cómo contribuye al proceso digestivo. Esto puede incluir la función del esfínter anal, el movimiento gástrico, la absorción de nutrientes en el intestino delgado, la secreción de enzimas pancreáticas y suelto del páncreas, la producción de bicarbonato hepático y la función del cálculo biliar.\n\n2. **Afirmaciones y condiciones del sistema digestivo**: Estudia las enfermedades y afecciones comunes relacionadas con la digestión, como úlceras gastrointestinales, enfermedades inflamatorias del tráqueo (IEP), cáncer gástrico, enfermedad de Crohn y síndrome del intestino irritable. Comprende cómo se diagnóstican y tratan estas afecciones.\n\n3. **Biología molecular y bioquímica**: Aprende más sobre la base química de los procesos digestivos, como la hidrólisis de carbohidratos, proteínas y grasa; el papel de enzimas específicas y las vías metabólicas implicadas en la digestión.\n\n4. **Fisiología médica**: Examina los procesos fisiológicos, como la secreción gastrointestinal, la regulación de los niveles hormonales que rigen los patrones de alimentación y la función de las vías digestivas en respuesta a la nutrición y la digestión.\n\n5. **Evolución y adaptación biológica**: Considera cómo se ha adaptado el sistema digestivo a diferentes formas de alimentación y cómo esta adaptación ha influido en la evolución humana. También puedes investigar sobre otras especies con sistemas digestivos distintos para ver cómo se han adaptado.\n\n6. **Ciencia forense e historia médica**: Examina cómo el estudio del sistema digestivo se utiliza en la autopsia y cómo ha evolucionado a lo largo de la historia. Esto puede ayudarte a entender las diferencias entre la anatomía y la biología modernas y aquellas de períodos históricos.\n\n7. **Práctica clínica y habilidades de atención al paciente**: Aprende sobre los procedimientos médicos relacionados con la digestión, como endoscopias, exploraciones abdominales y terapia con antibióticos para infecciones bacterianas. También debes familiarizarte con las técnicas de comunicación médica y las consideraciones éticas para la atención al paciente.\n\n8. **Cirugía digestiva y endocrinología**: Conoce las cirugías que se realizan para tratar afecciones digestivas y cómo el sistema digestivo interactúa con el sistema endocrino, particularmente con el sistema gastroenterológico.\n\n9. **Nutrición dietética y alimentación saludable**: Aprende sobre la importancia de una dieta equilibrada y cómo las malnutriciones o desequilibrios dietéticos pueden afectar la función digestiva.\n\nAl explorar estos temas, te prepararás no solo para tus clases, sino también para entender mejor el papel complejo del sistema digestivo en el cuerpo humano y su interconexión con otros sistemas.

Based on the basic knowledge you likely obtained in high school and considering the medical context, you might want to dive deeper into the following topics before your meeting:\n\n1. Function and structure of the digestive system: Explore how each organ works and contributes to the digestive process. This could include the role of the anal sphincter, gastric movement, nutrient absorption in the small intestine, pancreatic enzyme secretion, bile production in the liver, and the function of the gallbladder.\n\n2. Digestive system conditions and diseases: Study common digestive diseases and conditions, such as gastrointestinal ulcers, inflammatory bowel diseases (IBD), gastric cancer, Crohn's disease, and irritable bowel syndrome. Understand how these conditions are diagnosed and treated.\n\n3. Molecular biology and biochemistry: Learn more about the chemical basis of digestive processes, such as the breakdown of carbohydrates, proteins, and fats, the role of specific enzymes, and the metabolic pathways involved in digestion.\n\n4. Medical physiology: Examine physiological processes like gastrointestinal secretion, hormonal regulation that controls feeding patterns, and the digestive system's function in response to nutrition and digestion.
5. Evolution and biological adaptation: Consider how the digestive system has adapted to different types of diets and how this has influenced human evolution. You can also research how other species with distinct digestive systems have adapted to their environments.\n\n6. Forensic science and medical history: Examine how the study of the digestive system is used in autopsies and how its understanding has evolved throughout history. This can help you grasp the differences between modern anatomy and biology and those from historical periods.\n\n7. Clinical practice and patient care skills: Learn about medical procedures related to digestion, such as endoscopies, abdominal scans, and antibiotic therapy for bacterial infections. You should also familiarize yourself with medical communication techniques and ethical considerations for patient care.\n\n8. Digestive surgery and endocrinology: Understand the surgeries performed to treat digestive conditions and how the digestive system interacts with the endocrine system, particularly in the context of gastroenterology.
9. Nutritional science and healthy eating: Study the importance of a balanced diet and how malnutrition or dietary imbalances can affect digestive function.\n\nBy exploring these topics, you will be prepared not only for your classes but also to better understand the complex role of the digestive system in the human body and its connections to other systems.

Reference Answer

Absolutamente, prepararse para discutir sus aspiraciones educativas con un profesor en un programa médico requiere un análisis más profundo del tema. Dado que la imagen se refiere al sistema digestivo humano, esto es lo que podría querer investigar y examinar con más detalle:\n\n1. **Fisiología y Función**: Comprenda cómo funciona cada parte del sistema digestivo, no solo anatómicamente sino a nivel celular y molecular. Explore temas como la digestión enzimática en el estómago y los intestinos, la absorción de nutrientes en el intestino delgado y el papel del microbioma en el intestino grueso.\n\n2. **Fisiopatología**: Investigue enfermedades y trastornos comunes que afectan el sistema digestivo, como el reflujo ácido, la enfermedad inflamatoria intestinal, la cirrosis hepática y los trastornos pancreáticos. Comprender cómo estas enfermedades alteran la fisiología normal puede ser crucial.\n\n3. **Correlaciones Clínicas**: Investigue cómo se presentan los síntomas en varias enfermedades digestivas y qué métodos de diagnóstico se utilizan para identificarlas. Esto podría incluir aprender sobre la endoscopia, colonoscopia, técnicas de imagen y pruebas de laboratorio.\n\n4. **Nutrición**: Dado que el sistema digestivo es integral para la nutrición, profundice en cómo se digieren y procesan los diferentes nutrientes. Puede querer entender el impacto de la dieta en la salud digestiva y cómo cambian las necesidades nutricionales en estados de enfermedad.\n\n5. **Farmacología**: Investigue cómo afectan varios medicamentos al sistema digestivo, incluidos aquellos utilizados para tratar trastornos digestivos. Esto incluye comprender los mecanismos de acción, efectos secundarios y el uso terapéutico de estos medicamentos.\n\n6. **Intervenciones Quirúrgicas**: Familiarícese con los procedimientos quirúrgicos relacionados con el sistema digestivo, como apendicectomías, cirugías bariátricas o trasplantes de hígado, y cuándo son necesarias estas intervenciones.\n\n7. **Investigación y Avances**: Manténgase actualizado sobre la última investigación en gastroenterología, incluidos nuevos tratamientos, avances en la investigación del microbioma intestinal y tecnologías emergentes en diagnósticos de salud digestiva.\n\n8. **Sistemas Interconectados**: Explore cómo el sistema digestivo interactúa con otros sistemas del cuerpo, como el sistema nervioso (el eje intestino-cerebro) y el sistema inmunológico.\n\nAl examinar estos temas, podrá tener una discusión más informada y detallada con su profesor sobre sus objetivos de aprendizaje en su programa médico.

Absolutely, preparing to discuss your educational aspirations with a professor in a medical program requires a more in-depth analysis of the subject. Since the image refers to the human digestive system, here is what you might want to research and examine further:\n\n1. Physiology and Function: Understand how each part of the digestive system works, not only anatomically but also at the cellular and molecular levels. Explore topics like enzymatic digestion in the stomach and intestines, nutrient absorption in the small intestine, and the role of the microbiome in the large intestine.\n\n2. Pathophysiology: Investigate common diseases and disorders affecting the digestive system, such as acid reflux, inflammatory bowel disease, liver cirrhosis, and pancreatic disorders. Understanding how these diseases disrupt normal physiology can be crucial.\n\n3. Clinical Correlations: Research how symptoms present in various digestive diseases and what diagnostic methods are used to identify them. This could include learning about endoscopy, colonoscopy, imaging techniques, and laboratory tests.\n\n4. Nutrition: Since the digestive system is integral to nutrition, dive into how different nutrients are digested and processed. You might want to understand the impact of diet on digestive health and how nutritional needs change in states of illness.\n\n5. Pharmacology: Research how various medications affect the digestive system, including those used to treat digestive disorders. This includes understanding the mechanisms of action, side effects, and therapeutic uses of these drugs.\n\n6. Surgical Interventions: Familiarize yourself with surgical procedures related to the digestive system, such as appendectomies, bariatric surgery, or liver transplants, and when these interventions are necessary.\n\n7. Research and Advances: Stay updated on the latest research in gastroenterology, including new treatments, advances in gut microbiome research, and emerging technologies in digestive health diagnostics.\n\n8. Interconnected Systems: Explore how the digestive system interacts with other body systems, such as the nervous system (the gut-brain axis) and the immune system.\n\nBy examining these topics, you will be able to have a more informed and detailed discussion with your professor about your learning objectives in the medical program.

Feedback from GPT-4o

The response attempts to provide a comprehensive preparatory plan by highlighting various topics related to the digestive system. It covers multiple aspects such as the function and structure of the digestive system, common diseases, and certain diagnostic procedures, as well as practical clinical skills and nutrition. However, some key areas necessary for a higher score in the rubric are either vaguely covered or missing. There is no explicit mention of pharmacology, which is an essential part of understanding how medications interact with the digestive system. Additionally, the discussion on interconnected systems, which can include the neurological and immune system interactions with the digestive system, is somewhat lacking in depth. While several areas are addressed, the response does not fully meet the comprehensive criteria for the top score. Overall, the breadth of topics is good, but the depth and inclusion of all necessary topics as per the rubric are limited.

[RESULT] 4

Figure C.6: An xChat example from Spanish subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

Does the response effectively describe the elements that contribute to the mood of the portrait and their interplay to create an emotional experience?

Score 1: The response identifies basic elements of the portrait but does not relate them to the mood, for example, it might mention the subject's gaze but not how it conveys introspection.
Score 2: The response recognizes one or two elements that contribute to the mood but does not integrate them into a cohesive analysis, for example, noting the lighting or color but not how they work together to evoke calmness.
Score 3: The response describes several elements such as gaze, lighting, and color, and begins to relate them to the mood but lacks detail on their interactive effect, for example, it might suggest tranquility but not delve into the emotional depth created by these elements.
Score 4: The response provides a detailed explanation of the elements and their contribution to the mood, but may not fully capture the holistic emotional experience, for example, it describes the serene atmosphere but not the viewer's personal engagement with the portrait.
Score 5: The response offers a comprehensive explanation of the elements and their synergy, thoroughly describing how they interact to create a holistic emotional experience, reflecting a deep understanding of the portrait's figurative language.

Instruction

在观察这幅肖像时，哪些特定元素导致了肖像中氛围，它们如何相互作用以创造整体的情感体验？

When observing this portrait, what specific elements contribute to the atmosphere of the portrait, and how do they interact to create an overall emotional experience?

Response from Pangea-7B

肖像中营造氛围的特定元素包括女人凝视、她的服装以及光影的运用。她直接且略带怀疑的目光营造出一种亲密而引人深思的感觉，邀请观众进行对话。她的服装，以土色为主，传达出低调优雅和谦逊，符合当时的文化规范。光影的微妙运用，通过对比和柔和过渡，增加了画面的深度和立体感，突出了她的面部特征，特别是嘴唇和项链的光线，与背景的阴影形成鲜明对比。这些元素相互作用，创造了一种既沉思又吸引人的整体情感体验。

Specific elements that create the atmosphere in the portrait include the woman's gaze, her attire, and the use of light and shadow. Her direct yet slightly skeptical look fosters a sense of intimacy and contemplation, inviting the viewer to engage in dialogue. Her clothing, primarily in earthy tones, conveys understated elegance and humility, aligning with cultural norms of the time. The subtle manipulation of light and shadow, through contrast and soft transitions, adds depth and dimensionality to the image, highlighting her facial features—especially the light on her lips and necklace, which starkly contrasts with the background's shadows. These elements interact to create an overall emotional experience that is both contemplative and engaging.

Reference Answer

画中人物散发出一种宁静和沉思的氛围，主要通过人物温柔的目光和柔和的光线传达出来。女人的眼睛微微低垂，目光向一侧，引导观众思考她心中可能在想什么，暗示着内省或安静。画面中的光线轻轻照亮她脸的一侧，投下微妙的阴影，营造出宁静的氛围。这种光影的互动为画作整体的平静和深思感受做出了贡献。
此外，色彩的选择增强了这种氛围；蓝色头巾的冷色调和她衣服的温暖土色调实现了平衡和谐的效果，唤起一种平和的感觉。头巾优雅地披在肩上，增加了人物的尊严和反思的气质。孤立的耳环捕捉到一丝光芒，是一个引人注目的细节，将观众的注意力吸引到她沉思的表情上，加深了肖像的情感共鸣
此外，人物与深色背景之间的鲜明对比将观众的注意力集中在她身上，进一步增强了图像的亲密和冥想特质。构图的简洁，没有任何分散注意力的元素，允许更深刻的情感连接，邀请观众在个人层面上与人物互动。这种极简主义的方法鼓励观众考虑人物举止的细微差别，并推测她的内心生活，增强了图像的内省氛围。
总体而言，肖像的情感影响是构图、光线、色彩和细节巧妙结合的结果，共同创造出一种强烈而内敛的内心宁静和反思孤独的描绘。

The figure in the painting exudes an atmosphere of tranquility and introspection, primarily conveyed through her gentle gaze and soft lighting. The woman's eyes are slightly lowered, gazing to the side, guiding the viewer to ponder what might be on her mind, suggesting introspection or quietness. The light in the image softly illuminates one side of her face, casting subtle shadows that contribute to a serene atmosphere. This interplay of light and shadow enhances the overall calm and contemplative feeling of the artwork.\n\nAdditionally, the choice of colors reinforces this ambiance; the cool tones of the blue headscarf and the warm earthy tones of her clothing create a balanced and harmonious effect, evoking a sense of peace. The headscarf drapes elegantly over her shoulders, adding dignity and a reflective quality to the figure. The isolated earring captures a glimmer of light, serving as an eye-catching detail that draws attention to her contemplative expression, deepening the emotional resonance of the portrait.\n\nFurthermore, the stark contrast between the figure and the dark background focuses the viewer's attention on her, further enhancing the image's intimate and meditative qualities. The simplicity of the composition, devoid of distracting elements, allows for a deeper emotional connection, inviting viewers to engage with the figure on a personal level. This minimalist approach encourages the audience to consider the nuances of the figure's demeanor and speculate about her inner life, amplifying the introspective atmosphere of the image.\n\nOverall, the emotional impact of the portrait results from a skillful combination of composition, lighting, color, and detail, collectively creating a strong yet subdued depiction of inner peace and reflective solitude.

Feedback from GPT-4o

The response effectively identifies several key elements of the portrait, such as the woman's gaze, her attire, and the use of light and shadow, contributing to the overall atmosphere. It adequately explains how each element contributes to the mood: the gaze creates intimacy and thoughtfulness, the attire suggests elegance and humility, and the interplay of light and shadow enhances depth and contrast. There is a notable attempt to explain their interactive effect by discussing how these elements work together to create a contemplative and engaging emotional experience. However, while the response delves into specifics such as the gaze inviting dialogue or the warm earth tones conveying norms, it slightly falls short of capturing the holistic emotional experience or the viewer's personal engagement with the portrait, as a Score 5 response would. The level of detail and the relation of elements are thorough but not as comprehensive as required for the highest score. [RESULT] 4

Figure C.7: An xChat example from the Chinese subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

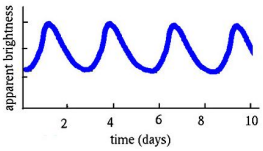
Does this response show an understanding of the Cepheid variable star light curves and their astronomical significance?

Score 1: The response does not recognize the graph as a Cepheid variable light curve nor its relevance in astronomy.
Score 2: The response identifies the graph as related to star brightness but fails to connect it to Cepheid variables or their importance in distance measurement.
Score 3: The response correctly identifies the graph as a Cepheid variable light curve but only partially explains its significance in astronomy.
Score 4: The response identifies the graph as a Cepheid variable light curve and explains its importance in distance measurement but does not mention Henrietta Swan Leavitt's contribution.
Score 5: The response accurately identifies the Cepheid variable light curve, explains the period-luminosity relationship, the concept of \"standard candles,\" and acknowledges Henrietta Swan Leavitt's work.

Instruction

这是我在天文学课上遇到的一张图表，我不知道该如何解释。你能解释一下它与什么相关以及它传达了什么信息吗?

This is a chart I encountered in my astronomy class, and I'm not sure how to interpret it. Can you explain what it relates to and what information it conveys?



Response from Pangea-7B

根据提供的图片，该图表是一个简单的波形图，通常在数学或物理学中用于代表随时间变化的波动现象。在这个情境下，由于图中的横轴是"天"，我们可以推断这个图表可能代表着一个天文学过程。可能的解释是这幅图显示了一个天文事件或天体位置的变化，比如行星或卫星的轨道周期、星系的振荡，或者恒星亮度的变化。这个图表传达了这种现象随时间变化的情况；具体来说，波峰表示该事件处于最大强度，而波谷则表示其最弱的阶段。没有额外的背景信息，我们无法确定具体代表的是什么现象，但它明显是描述了一个随时间变化的天文或数学物理过程。

Based on the provided image, the chart is a simple waveform, commonly used in mathematics or physics to represent fluctuating phenomena over time. In this context, since the horizontal axis is labeled "days," we can infer that this chart likely represents an astronomical process. One possible explanation is that the chart depicts variations in an astronomical event or celestial body's position, such as the orbital period of a planet or satellite, the oscillation of a galaxy, or changes in stellar brightness.

This chart conveys how such phenomena change over time; specifically, the peaks represent the maximum intensity of the event, while the troughs indicate its weakest phase. Without additional background information, we cannot definitively identify what specific phenomenon is being represented, but it clearly describes a time-varying astronomical or mathematical-physical process.

Reference Answer

您提供的图表似乎显示了一个天文物体的表观亮度随时间的周期性振荡。这种模式是造父变星的特征。\n\n造父变星是一类由于外层脉动而表现出亮度规律性变化的恒星。图表显示了这些亮度在十天内的变化。造父变星的关键特征是它们的脉动周期与其本征光度直接相关，这种关系被称为周期-光度关系。这种关系在天体物理学中至关重要，因为它允许天文学家确定造父变星的距离，并进而确定它们所在的遥远星系的距离，从而有助于测量宇宙的尺度。\n\n图表旨在传达：\n- 造父变星亮度变化的周期性特征。\n- 亮度峰值之间的规律时间间隔（周期）。\n- 随时间变化的亮度变化幅度。\n\n这些周期性数据用于根据亨丽爱塔·斯万·勒维特在20世纪初发现的周期-光度关系计算恒星的绝对星等，从而利用这些恒星作为测量天文距离的"标准烛光"来确定其与地球的距离。

The chart you provided appears to show the periodic oscillation of the apparent brightness of an astronomical object over time. This pattern is characteristic of Cepheid variables.\n\nCepheid variables are a class of stars that exhibit regular changes in brightness due to pulsations in their outer layers. The chart displays changes in brightness over a span of ten days. A key feature of Cepheid variables is that their pulsation period is directly related to their intrinsic luminosity, a relationship known as the period-luminosity relation. This relationship is crucial in astrophysics because it allows astronomers to determine the distance to Cepheid variables and, consequently, the distance to the distant galaxies they reside in, helping to measure the scale of the universe.\nThe chart is intended to convey:\n* The periodic nature of the brightness variations in Cepheid variables.\n* The regular time intervals (periods) between the brightness peaks.\n* The amplitude of brightness changes over time.\n\nThese periodic data are used to calculate the absolute magnitude of stars based on the period-luminosity relation discovered by Henrietta Swan Leavitt in the early 20th century, allowing these stars to serve as "standard candles" for measuring astronomical distances to determine their distance from Earth.

Feedback from GPT-4o

The response provided does not identify the graph as a Cepheid variable light curve and instead offers a general interpretation of a waveform graph used in mathematics or physics. It suggests the graph may represent an astronomical process, mentioning possibilities such as orbital periods or brightness variations, but fails to specifically recognize the graph as being related to Cepheid variables. Consequently, it does not address the significance of the period-luminosity relationship in distance measurement nor acknowledge Henrietta Swan Leavitt's contributions. Therefore, the response lacks key elements mentioned in the score rubric.\n\n[RESULT] 1

Figure C.8: An xChat example from the Chinese subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

Score Rubric

Does this response show clear communication of the desired artistic elements for a masterpiece to be displayed in the living room?

Score 1: The response does not specify any of the desired artistic elements for the artwork.
Score 2: The response vaguely mentions abstract elements but lacks detail on style, composition, color, emotional impact, and size.
Score 3: The response includes general descriptions of style and color but does not fully articulate the desired composition, emotional impact, or specific size requirements.
Score 4: The response clearly identifies the desired abstract expressionist style, rich color palette, and emotional impact, with some mention of size and composition preferences.
Score 5: The response provides a detailed and specific description of all desired elements including style, composition, color palette, emotional impact, and size requirements for the artwork.

Instruction

저는 제 거실에 전시된 이 작품과 유사한 미술 작품을 구매하고 싶습니다. 여러 걸작을 판매하는 사람에게 이것을 어떻게 설명해야 할까요? 어떻게 자세히 설명할 수 있을지 잘 모르겠습니다.

I want to buy a piece of art similar to the one displayed in my living room. How should I explain this to someone who sells several masterpieces? I'm not sure how to describe it in detail.

Response from Pangea-7B

이미지를 기반으로 이 작품은 현대 미술에서 종종 발견되는 추상적이고 개념적 특성을 가지고 있습니다. 추상화의 한 형태로 볼 수 있으며, 관찰자의 인식과 예술가의 표현을 교차시키는 방식으로 주제와 의미에 대한 보편적인 접근을 선호하는 탁 트인 해석적 표현과 일치합니다. 색상, 선 및 형태를 사용하여 예술가가 주제에 대한 자신의 감정적 또는 정서적 반응을 나타낼 수 있는 방식입니다. 특정한 장르나 주제가 없으므로 관찰자는 자신의 경험과 감정에 따라 자신의 해석과 연결할 수 있습니다.\n\n구매자에게 이러한 정보를 제공하면 예술가와 작업을 더 잘 이해할 수 있으므로 더욱 풍부하고 즐거운 경험을 할 수 있습니다. 또한 이러한 미술 작품은 종종 개인적인 연관성이나 감정적 호소력을 가질 수 있으며, 이것이 거실이라는 공간에서 예술이 어떻게 작용할 수 있는지에 대해 자세히 설명하는 것이 유용할 수 있습니다.
Based on the image, this piece possesses abstract and conceptual characteristics commonly found in contemporary art. It can be seen as a form of abstraction that aligns with open interpretative expressions, preferring universal approaches to themes and meanings that intersect the observer's perception and the artist's expression. The artist uses color, lines, and shapes to convey their emotional or emotional response to the subject. Since there isn't a specific genre or theme, the observer can connect their interpretation based on personal experiences and feelings.\n\nProviding this information to the seller will help you understand the artist and the work better, leading to a richer and more enjoyable experience. Additionally, such art often has personal connections or emotional appeal, making it useful to explain how the artwork might function within the living space.

Reference Answer

예술품 딜러나 갤러리 소유자와 이미지에 있는 것과 유사한 작품을 구입하려고 할 때, 관심 있는 스타일적 요소와 주제적 요소를 설명해야 합니다. 다음은 선호도를 설명하는 방법입니다:\n\n1. **스타일**: 에너제틱하고 제스처적인 붓질, 생동감 있는 색상, 즉흥적인 느낌이 특징인 추상 표현주의 스타일을 구현한 작품을 찾고 있다고 언급하세요.\n2. **구성**: 복잡하고 층이 있는 구성의 그림을 찾고 있으며, 아마도 비유적인 요소와 추상적인 형태가 얽혀 있는 작품을 원한다고 명시하세요.\n3. **색상 팔레트**: 디자인의 복잡성 내에서 높은 대조와 조화를 이룰 수 있는 풍부하고 다양한 색상 팔레트에 관심이 있다고 표현하세요.\n4. **감정적 영향**: 강한 감정적 또는 극적인 영향을 주는 작품에 끌린다고 전하며, 아마도 움직임이나 혼돈의 느낌이 있는 작품을 원한다고 전하세요.\n5. **크기**: 거실 공간에 맞고 원하는 대담한 표현을 할 수 있는 그림의 크기를 명시하세요.\n\n이러한 세부 정보를 제공함으로써, 예술품 판매자는 귀하의 필요를 더 잘 이해하고 거실과 개인 취향에 맞는 걸작으로 안내할 수 있습니다.

When trying to purchase a piece similar to the one in the image from an art dealer or gallery owner, you should describe both the stylistic and thematic elements that interest you. Here's how you can articulate your preferences:\n\n1. Style: Mention that you are looking for a work that embodies the energetic and gestural brushwork characteristic of abstract expressionism, with vibrant colors and an improvisational feel.\n\n2. Composition: Specify that you are seeking a painting with a complex and layered composition, likely featuring a mix of figurative elements and abstract forms.\n\n3. Color Palette: Express your interest in a rich and varied color palette that can create high contrast and harmony within the complexity of the design.\n\n4. Emotional Impact: Convey that you are drawn to pieces that have a strong emotional or dramatic impact, perhaps conveying a sense of movement or chaos.\n\n5. Size: Indicate the size of the painting that would fit well in your living space and allow for the bold expression you desire.\n\nBy providing these details, the art dealer will better understand your needs and guide you to masterpieces that align with your living room and personal taste.

Feedback from GPT-4o

The response focuses on abstract and conceptual characteristics of the artwork, mentioning its interpretive nature and emotional resonance. However, it does not provide specifics regarding style, composition, or color palette, nor does it address the emotional impact or size. While the response discusses interpretive qualities and the importance of emotional connection, it lacks clear identification or description of the artistic elements such as the abstract expressionist style, preferred composition, or ideal size, which are crucial for a practical purchase discussion with an art dealer. The information is generally abstract, without articulating detailed preferences that would assist a seller in understanding exactly what is needed for the living room setting.\n\n[RESULT] 2

Figure C.9: An xChat example from the Korean subset and our model PANGEA's response. The English part of the instruction, response, and reference answer is additionally added only on the figure to help to understand and was not given nor generated by PANGEA.

C.7 Languages in PANGEAINS

Table C.2 demonstrates the number of instances for each language that we include in PANGEAINS.

Languages	**en**	**multi**	**am**	**ar**	**bg**	**bn**	**cs**	**de**	**el**	**es**
Count	2554.6	4389.5	31.7	162.8	52.7	118.4	4.6	132.2	7.3	126.6
Percentage (%)	36.8	63.2	0.5	2.3	0.8	1.7	0.1	1.9	0.1	1.8
Languages	**fa**	**fr**	**ga**	**hi**	**id**	**ig**	**it**	**iw**	**ja**	**jv**
Count	8.4	193.8	34.7	163.5	178.4	26.1	68.8	168.3	403.3	35.2
Percentage (%)	0.1	2.8	0.5	2.4	2.6	0.4	1.0	2.4	5.8	0.5
Languages	**ko**	**nl**	**mn**	**ms**	**no**	**pl**	**pt**	**ro**	**ru**	**si**
Count	182.5	4.5	37.6	39.2	60.7	8.2	176.5	147.2	211.0	0.6
Percentage (%)	2.6	0.1	0.5	0.6	0.9	0.1	2.5	2.1	3.0	0.1
Languages	**su**	**sw**	**ta**	**te**	**th**	**tr**	**uk**	**ur**	**vi**	**zh**
Count	33.0	132.7	37.1	64.6	160.4	140.7	7.4	156.4	315.5	557.4
Percentage (%)	0.5	1.9	0.5	0.9	2.3	2.0	0.1	2.3	4.5	8.0

Table C.2: Language distribution of PANGEAINS. We demonstrate the number of instances (in thousands) for each language, and the percentage of instances for each language in PANGEAINS

C.8 Training Examples

C.8.1 Machine Translated Instructions

We include a few machine-translated training examples from PANGEAINS in Figure C.10, Figure C.11, Figure C.12, Figure C.13, Figure C.14, Figure C.15.

C.8.2 Multicultural Understanding Instructions

We include a few multicultural understanding instructions from PANGEAINS in Figure C.16, Figure C.17, Figure C.18, Figure C.19.

Translated: Cambrian (Russian)

Human:
Средняя заработная плата в час всех сотрудников компании составляет 20 долларов. Какова почасовая ставка Тома?
(The average hourly wage for all employees in the company is $20. What is Tom's hourly rate?)

GPT:
Невозможно определить почасовую ставку Тома на основе предоставленной информации. Средняя заработная плата в час всех сотрудников компании не обязательно отражает почасовую ставку любого отдельного сотрудника, поскольку индивидуальные оклады могут варьироваться в зависимости от ряда факторов, таких как должность, опыт, образование и производительность. Кроме того, средняя заработная плата в час всех сотрудников компании рассчитывается путем деления общей заработной платы всех сотрудников на количество часов, отработанных всеми сотрудниками, что не обязательно отражает количество часов, отработанных каким-либо отдельным сотрудником.
(It is not possible to determine Tom's hourly wage based on the information provided. The average hourly wage for all employees of a company does not necessarily reflect the hourly wage of any individual employee, as individual salaries can vary based on a number of factors, such as job title, experience, education, and performance. Additionally, the average hourly wage for all employees of a company is calculated by dividing the total wages of all employees by the number of hours worked by all employees, which does not necessarily reflect the number of hours worked by any individual employee.)

Figure C.10: ***Translated Task***: An example from the Cambrian dataset where it discusses the concept of hourly wages based on a given prompt in Russian.

Translated: ALLaVa-LAION (Turkish)

Human:
Dijital ekrandaki zamanlayıcıda görülebilen maksimum süre nedir?
(What is the maximum time that can be seen on the timer on the digital display?)

GPT:
Dijital ekrandaki zamanlayıcıda görülebilen maksimum süre, \"001 00.00.28\" sayılarının gösterdiği gibi 1 dakika, 00 saniye ve 28 milisaniyedir.

(The maximum time that can be seen on the timer on the digital display is 1 minute, 00 seconds and 28 milliseconds, as shown by the numbers \"001 00.00.28\")

Figure C.11: ***Multimodal Translated Task***: An example from the ALLaVa-LAION dataset where the GPT model answers a prompt in Turkish regarding the maximum time displayed on a digital timer.

Translated: GPT-4v (Thai)

Human:
ภาพที่แสดง:
(Image showing:)

GPT:
ชั้นวางไวน์ติดผนังพร้อมขวดไวน์จำนวนมากเรียงรายกันในแนวตั้ง ชั้นวางทำจากแผ่นไม้แนวนอนหลายชั้น เจาะเป็นรูวงกลมสำหรับวางขวดไวน์ในแนวนอน โดยให้คอขวดหันออกด้านนอก มีขวดไวน์หลากหลายรูปแบบ แต่ละขวดมีฉลากและสีต่างกัน บ่งบอกถึงประเภทและยี่ห้อของไวน์ที่หลากหลาย ขวดต่างๆ ถูกจัดวางอย่างเป็นระเบียบและเต็มชั้นวาง พื้นหลังเป็นผนังสีเรียบ ซึ่งตัดกับสีเข้มของชั้นวางไม้และฉลากไวน์หลากสีสัน รูปแบบของภาพดูตรงไปตรงมาและดูเหมือนจะเป็นภาพถ่ายที่เน้นการใช้งานจริง เพื่อแสดงชั้นวางไวน์และขวดไวน์ภายในอย่างชัดเจนและเรียบง่าย
(A wall-mounted wine rack with many wine bottles lined up vertically. The rack is made of several horizontal wooden planks, with circular holes cut out to hold the wine bottles horizontally, with the necks of the bottles facing outward. There are various types of wine bottles, each with different labels and colors, indicating the various types and brands of wine. The bottles are arranged neatly and fill the rack. The background is a plain wall, which contrasts with the dark colors of the wooden racks and the colorful wine labels. The layout of the image is straightforward and seems to be a practical photo, to clearly and simply show the wine rack and the wine bottles inside.)

Figure C.12: ***Multimodal Translated Task***: An example from the ShareGPT-4v dataset where the model describes an image of a wine rack in Thai, detailing its structure and the arrangement of wine bottles.

Translated: OpenHermes (Japanese)

Human:
for ループを関数呼び出しに変換して、コードを最小限に抑えるのを手伝ってもらえますか?
(Can you help me convert the for loop into a function call to minimize the code?)

GPT:
承知しました！ コードを最小限に抑えるには、`map` 関数を使用して同じ結果を得ることができます。次に例を示します。\n\n```javascript\nlet arr = [4, 7, 2, 9];\narr.map((elem) => console.log(elem * 2));\n```\n\nJavaScript の `map` 関数を使用すると、配列を反復処理し、指定された関数に従って各要素を変換できます。この場合、各要素に 2 を掛けてから、コンソールに出力しています。\n\n`map` を使用することで、従来の for ループが不要になり、より簡潔な方法で同じ結果を得ることができます。
*(Sure! To minimize code, you can use the `map` function to achieve the same result. For example:\n\n```javascript\nlet arr = [4, 7, 2, 9];\narr.map((elem) => console.log(elem * 2));\n```\n\nJavaScript's `map` function allows you to iterate through an array and transform each element according to a specified function. In this case, we multiply each element by 2 before printing it to the console.\n\nBy using `map`, we can eliminate the need for a traditional for loop and achieve the same result in a more concise way.)*

Figure C.13: ***Text-only Translated Task***: An example from the OpenHermes2.5 dataset translated into Japanese.

Translated: NuminaMath (Spanish)

Human:
Definir una terna ordenada (A, B, C) de conjuntos como mínimamente intersecantes como se describe: $|A \\cap B| = 1$, $|B \\cap C| = 1$, $|C \\cap A| = 1$, y $A \\cap B \\cap C = \\emptyset$. Determinar el número de tales ternas ordenadas donde cada conjunto es un subconjunto de $\\{1,2,3,4,5,6,7,8\\}$. Cada conjunto debe incluir ahora exactamente un elemento que no se comparte con los otros conjuntos, y ningún conjunto puede estar vacío. Calcular el resto cuando el número de estas ternas se divide entre 1000.

(Define an ordered triple (A, B, C) of sets to be minimally intersecting as follows: $|A \\cap B| = 1$, $|B \\cap C| = 1$, $|C \\cap A| = 1$, and $A \\cap B \\cap C = \\emptyset$. Determine the number of such ordered triples where each set is a subset of $\\{1,2,3,4,5,6,7,8\\}$. Each set must now include exactly one element that is not shared with the other sets, and no set can be empty. Compute the remainder when the number of such triples is divided by 1000.)

GPT:
1. Elegir x, y, z como identificadores únicos para las intersecciones por pares ($A∩B$, $B∩C$, y $C∩A$). Hay $8 \\cdot 7 \\cdot 6 = 336$ maneras de elegir valores distintos para x, y, y z ya que el conjunto $\\{1,2,3,4,5,6,7,8\\}$ contiene 8 números.\n\n2. Cada conjunto (A, B, C) debe incluir exactamente un elemento único aparte de x, y, z. Esto significa seleccionar un elemento único de los 5 números restantes para cada conjunto de manera que ningún elemento se reutilice: \n - Elegir un elemento único para A (no x o y) de los 5 elementos restantes: 5 opciones.\n - Elegir otro elemento único para B (no y o z) de los 4 elementos restantes: 4 opciones.\n - Elegir otro elemento único para C (no x o z) de los 3 elementos restantes: 3 opciones.\n\n3. Los 2 elementos restantes pueden ir a cualquier conjunto o a ninguno, por lo que hay 4 opciones para cada uno de estos 2 elementos (cualquiera de los tres conjuntos o ninguno): $4^2 = 16$.\n\nCombinando estas opciones,\n\\[336 \\times 5 \\times 4 \\times 3 \\times 16 = 107520 \\]\nPor lo tanto, el resto al dividir entre 1000 es $\\boxed{520}$.\n \n".

(1. Choose x, y, z as unique identifiers for the pairwise intersections ($A∩B$, $B∩C$, and $C∩A$). There are $8 \\cdot 7 \\cdot 6 = 336$ ways to choose distinct values for x, y, and z since the set $\\{1,2,3,4,5,6,7,8\\}$ contains 8 numbers.\n\n2. Each set (A, B, C) must contain exactly one unique element other than x, y, z. This means selecting a unique element from the remaining 5 numbers for each set such that no element is reused: \n- Choose a unique element for A (not x or y) from the remaining 5 elements: 5 options.\n- Choose another unique element for B (not y or z) from the remaining 4 elements: 4 options.\n- Choose another unique element for C (not x or z) from the remaining 3 elements: 3 options.\n\n3. The remaining 2 elements can go into any set or none, so there are 4 options for each of these 2 elements (any of the three sets or none): $4^2 = 16$.\n\nCombining these options,\n\\[336 \\times 5 \\times 4 \\times 3 \\times 16 = 107520 \\]\nThus, the remainder when divided by 1000 is $\\boxed{520}$.\n \n")

Figure C.14: ***Text-only Translated Task***: An example from the NumininaMath dataset translated into Spanish.

Translated: Code-Feedback (German)

Human:
Gegeben ein Array von Ganzzahlen, bestimmen Sie den Maximal- und Minimalwert des Arrays. Das Array kann doppelte Werte haben und kann eine beliebige Länge zwischen 1 und 10^6 haben. Außerdem kann das Array unsortiert sein.
(Given an array of integers, determine the maximum and minimum values of the array. The array can have duplicate values and can have any length between 1 and 10^6. Additionally, the array can be unsorted.)

GPT:
Plan\n1. Initialisieren Sie die Variablen min_val und max_val mit dem ersten Element des Arrays.\n2. Iterieren Sie durch das Array und aktualisieren Sie min_val und max_val nach Bedarf.\n3. Geben Sie min_val und max_val zurück
(## Plan\n1. Initialize the variables min_val and max_val with the first element of the array.\n2. Iterate through the array and update min_val and max_val as needed.\n3. Return min_val and max_val)

Human:
Berechnen Sie den Umfang und die Fläche eines Dreiecks mit der gegebenen Seitenlänge. Seitenlängen: 6, 8, 10. Das Programm sollte auch prüfen, ob die angegebenen Seiten ein gültiges Dreieck bilden. Außerdem sollte das Programm Fälle behandeln, in denen die Seiten kein gültiges Dreieck bilden, indem es entsprechende Fehlermeldungen ausgibt

(Calculate the perimeter and area of a triangle with the given side length. Side lengths: 6, 8, 10. The program should also check whether the given sides form a valid triangle. In addition, the program should handle cases where the sides do not form a valid triangle by giving appropriate error messages)

GPT:
Um den Umfang und die Fläche eines Dreiecks mit den gegebenen Seitenlängen (6, 8, 10) zu berechnen, können wir die Formel von Heron verwenden. Bevor wir jedoch mit der Berechnung fortfahren, müssen wir prüfen, ob die gegebenen Seiten ein gültiges Dreieck bilden.\n\nUm festzustellen, ob die gegebenen Seiten ein gültiges Dreieck bilden, können wir den Satz der Dreiecksungleichung anwenden, der besagt, dass die Summe zweier Seiten eines Dreiecks größer sein muss als die dritte Seite.\n\nSchreiben wir ein Programm in Python, das den Umfang und die Fläche eines Dreiecks mit den gegebenen Seitenlängen berechnet und prüft, ob die Seiten ein gültiges Dreieck bilden:\n\n```python\nimport math\n\ndef calculate_perimeter(a, b, c):\n return a + b + c\n\ndef calculate_area(a, b, c):\n s = (a + b + c) / 2 # Berechnen Sie den halben Umfang\n area = math.sqrt(s * (s - a) * (s - b) * (s - c)) # Verwenden Sie die Formel von Heron\n return area\n\ndef is_valid_triangle(a, b, c):\n if a + b > c and b + c > a and a + c > b:\n return True\n return False\n\n# Geben Sie die Seitenlängen des Dreiecks ein\na = 6\nb = 8\nc = 10\n\n# Prüfen Sie, ob die Seiten ein gültiges Dreieck bilden\nif is_valid_triangle(a, b, c):\n perimeter = calculate_perimeter(a, b, c)\n area = calculate_area(a, b, c)\n\n print(\"Umfang:\", perimeter)\n print(\"Fläche:\", area)\nelse:\n print(\"Die angegebenen Seiten bilden kein gültiges Dreieck.\")\n```\n\nWenn Sie dieses Programm ausführen, wird Folgendes ausgegeben:\n\n```\nUmfang: 24\nFläche: 24.0\n```\n\nDa die gegebenen Seitenlängen (6, 8, 10) den Satz der Dreiecksungleichung erfüllen, können wir den Umfang und die Fläche des Dreiecks berechnen. Der Umfang beträgt 24 Einheiten und die Fläche 24 Flächeneinheiten."
(To calculate the perimeter and area of a triangle with given side lengths (6, 8, 10), we can use Heron's formula. However, before we proceed with the calculation, we need to check if the given sides form a valid triangle.\n\nTo determine if the given sides form a valid triangle, we can apply the triangle inequality theorem, which states that the sum of two sides of a triangle must be greater than the third side.\n\nLet's write a program in Python that calculates the perimeter and area of a triangle with the given side lengths and checks if the sides form a valid triangle:\n\n```python\nimport math\n\ndef calculate_perimeter(a, b, c):\n return a + b + c\n\ndef calculate_area(a, b, c):\n s = (a + b + c) / 2 # Calculate half the perimeter\n area = math.sqrt(s * (s - a) * (s - b) * (s - c)) # Use Heron's formula\n return area\n\ndef is_valid_triangle(a, b, c):\n if a + b > c and b + c > a and a + c > b:\n return True\n return False\n\n# Enter the side lengths of the triangle\na = 6\nb = 8\nc = 10\n\n# Check if the sides form a valid triangle\nif is_valid_triangle(a, b, c):\n perimeter = calculate_perimeter(a, b, c)\n area = calculate_area(a, b, c)\n\n print(\"Perimeter:\", perimeter)\n print(\"Area:\", area)\nelse:\n print(\"The given sides do not form a valid triangle.\")\n```\n\nWhen you run this program, the following is output:\n\n```\nPerimeter: 24\nArea: 24.0\n```\n\nSince the given If the side lengths (6, 8, 10) satisfy the triangle inequality, we can calculate the perimeter and area of the triangle. The perimeter is 24 units and the area is 24 units.")

Figure C.15: ***Text-only Translated Task***: An example from the Code-Feedback dataset translated into German.

Multicultural Understanding: LAION-Culture (Portugese)

Human:
Descreva os elementos visuais que compõem o logotipo da Petrobras na imagem.
(Describe the visual elements that make up the Petrobras logo in the image.)

GPT:
O logotipo da Petrobras na imagem é composto por letras metálicas prateadas e brilhantes, fixadas em uma parede de concreto cinza. A palavra 'PETROBRAS' está escrita em letras maiúsculas e maiores, enquanto a sigla 'BR' aparece em letras menores, em um quadrado separado. As letras são tridimensionais, o que lhes confere uma aparência moderna e destacada. O contraste entre o brilho das letras e a superfície opaca da parede cria um efeito visual interessante, destacando a presença da marca.
(The Petrobras logo in the image is made up of shiny silver metallic letters, fixed to a gray concrete wall. The word 'PETROBRAS' is written in larger capital letters, while the acronym 'BR' appears in smaller letters, in a separate square. The letters are three-dimensional, giving them a modern, standout look. The contrast between the brightness of the letters and the opaque surface of the wall creates an interesting visual effect, highlighting the brand's presence.)

Human:
Analise como o design do logotipo da Petrobras reflete a identidade e os valores da empresa.
(Analyze how the Petrobras logo design reflects the company's identity and values.)

GPT:
O design do logotipo da Petrobras reflete a identidade e os valores da empresa de várias maneiras. Primeiramente, as letras metálicas prateadas e brilhantes sugerem inovação e modernidade, valores essenciais para uma empresa de petróleo e gás natural. A tridimensionalidade das letras confere uma sensação de solidez e estabilidade, reforçando a ideia de que a Petrobras é uma empresa forte e confiável. Além disso, o contraste entre o brilho das letras e a parede de concreto cinza pode simbolizar a integração da empresa com o ambiente urbano e industrial. A separação da sigla 'BR' em um quadrado distinto destaca a importância da identidade brasileira para a empresa, reforçando seu papel como uma entidade nacional. Portanto, o design do logotipo não só representa a identidade visual da Petrobras, mas também comunica seus valores e sua conexão com o Brasil.
(The Petrobras logo design reflects the company's identity and values in many ways. Firstly, the shiny silver metallic letters suggest innovation and modernity, essential values for an oil and natural gas company. The three-dimensionality of the letters gives a feeling of solidity and stability, reinforcing the idea that Petrobras is a strong and reliable company. Furthermore, the contrast between the brightness of the letters and the gray concrete wall can symbolize the company's integration with the urban and industrial environment. The separation of the acronym 'BR' into a distinct square highlights the importance of Brazilian identity for the company, reinforcing its role as a national entity. Therefore, the logo design not only represents Petrobras' visual identity, but also communicates its values and its connection with Brazil.)

Figure C.16: ***Multicultural Understanding***: An example from the LAION-Culture dataset where the GPT model describes and analyzes the Petrobras logo in Portuguese.

Multicultural Understanding: LAION-Culture (Bulgarian)

Human:
Какво е значението на японския йероглиф на снимката?
(What is the meaning of the Japanese character in the picture)

GPT:
Йероглифът на снимката представлява символа \"折\", който се чете \"ори\" на японски. Този йероглиф се използва за означаване на думата \"ори\", която може да има различни значения в зависимост от контекста. В случая, йероглифът е показан с черно мастило върху бяла хартия, със сини пунктирани линии, които разделят символа на секции, за да се покаже правилната последователност на ударите при писането. Това е важен елемент от японската калиграфия, която се нарича \"шодо\".
(The hieroglyph in the photo is the character \"折\" which is read \"ori\" in Japanese. This hieroglyph is used to denote the word \"ori\", which can have different meanings depending on the context. In this case, the hieroglyph is shown in black ink on white paper, with blue dotted lines dividing the symbol into sections to show the correct sequence of strokes when writing. This is an important element of Japanese calligraphy, which is called ``shodo''.)

Human:
Какво значение имат сините пунктирани линии в японската калиграфия?.
(What is the meaning of blue dotted lines in Japanese calligraphy?)

GPT:
Сините пунктирани линии в японската калиграфия имат специфично значение и предназначение. Те са използвани за обозначаване на последователността и посоката на движенията на четката при писане на йероглифи. В случая, тези линии разделят йероглифа \"折\" на секции, за да се улесни процесът на писане и да се осигури, че всеки удар ще бъде извършен правилно. Това е част от традицията на японската калиграфия, която се нарича \"шодо\", и е съществена за изучаването и практикуването на този изкуство. Сините линии не само помагат за правилното писане, но и допринасят за естетическото качество на финалния произведение
(The blue dotted lines in Japanese calligraphy have a specific meaning and purpose. They were used to indicate the sequence and direction of brush movements when writing hieroglyphs. In this case, these lines divide the character \"折\" into sections to facilitate the writing process and ensure that each stroke will be executed correctly. It is part of the tradition of Japanese calligraphy, which is called ``shōdo,'' and is essential to the study and practice of this art. Blue lines not only help with correct writing, but also contribute to the aesthetic quality of the final work)

Figure C.17: ***Multicultural Understanding***: Example from the LAION-Culture dataset where the model interprets the meaning of a Japanese character and its representation in calligraphy, described in Bulgarian.

Multicultural Understanding: LAION-Culture (Hebrew)

Human:
Analyze the cultural significance of the prominent landmarks visible in the image of Jerusalem. Answer it in Hebrew.

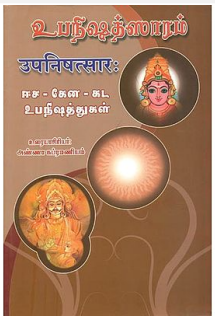
GPT:
התמונה מציגה שלושה מורכבים רב-תווים של ירושלים: חומות העיר העתיקה, כיפת הסלע המוזהבת, ועצמה העיר העתיקה. חומ ות העיר העתיקה, שמקובלות כמורכב בנוי בין השנים 1535 ל-1538 על ידי הסולטן העות'מאני סלימאן המפיק, מסמלות את ההיס טוריה העשירה של העיר ואת המשמעות התרבותית והדתית שלה כמקום של חשיבות רבה ביהדות, נוצרות ואסלאם. כיפת הסלע המוזהבה, בנויה בשנת 691 לספירה, עומדת כסמל נוצץ של האסלאם ומייצגת את המשמעות הרוחנית העמוקה של המקום הזה ב תקופה האיסלאמית. העיר העתיקה בעצמה, עם רחובותיה הצרים ובתיה האבן, מעיינת בין השאר במורשת התרבותית העשירה ש ל ירושלים, בה מתמזגות השפות, הדתות והתרבויות בצורה ייחודית. הנוף הזה, עם העצים והרחובות החיים, מעיין בחיי העיר ובמ שמעות התרבותית וההיסטורית של ירושלים כמרכז רוחני ותיירותי.
(The image shows three multi-character complexes of Jerusalem: the walls of the Old City, the golden Dome of the Rock, and the Old City itself. The Old City Walls, which are accepted as a complex built between 1535 and 1538 by the Ottoman Sultan Suleiman the Magnificent, symbolize the city's rich history and its cultural and religious significance as a place of great importance in Judaism, Christianity and Islam. The golden Dome of the Rock, built in 691 AD, stands as a shining symbol of Islam and represents the deep spiritual significance of this place in the Islamic period. The old city itself, with its narrow streets and stone houses, reflects, among other things, the rich cultural heritage of Jerusalem, where the languages, religions and cultures merge in a unique way. This landscape, with the trees and living streets, is a source of the life of the city and the cultural and historical significance of Jerusalem as a spiritual and touristic center.")

Human:
Imagine you are planning a guided tour of Jerusalem based on the landmarks visible in the image. Outline the key points you would cover in your tour. Answer it in Hebrew.

GPT:
בהסכמה לסיור מודרך בירושלים בהתבסס על המקומות המובילים שמופיעים בתמונה, הייתי מתחיל בחומות העיר העתיקה, שמ סמלות את הגבול ההיסטורי של העיר. הייתי מספר על ההיסטוריה של החומות, על סלימאן המפיק ועל המשמעות התרבותית של הן. משם, הייתי מזיז את הקבוצה לכיפת הסלע המוזהבת, והייתי מספר על הבנייה שלה בשנת 691 ועל המשמעות הרוחנית של ה מקום באסלאם. הייתי מדגיש את הארכיטקטורה הייחודית ואת המשמעות הסמלית של הכיפה. לבסוף, הייתי מוביל את הקבוצה לרחובות הצרים של העיר העתיקה, והייתי מספר על החיי היומיומי, על המזג התרבותי העשיר ועל המשמעות התיירותית של האז ור. הייתי מדגיש את השילוב הייחודי של שפות, דתות ותרבויות בתוך העיר, ואיך זה משפיע על החוויה התיירותית. הייתי מסיים ע ם הצגת הנוף המרהיב והשלווה של העיר, והמשמעות הרוחנית וההיסטורית של ירושלים כמרכז חשוב בעולם."
(Agreeing to a guided tour of Jerusalem based on the leading places that appear in the picture, I would start with the walls of the old city, which symbolize the historical border of the city. I would tell about the history of the walls, about Suleiman the producer and their cultural significance. From there, I would move the group to the golden Dome of the Rock, and I would tell about its construction in 691 and the spiritual significance of the place in Islam. I would emphasize the unique architecture and the symbolic meaning of the dome. Finally, I would lead the group to the narrow streets of the old city, and I would tell about the daily life, the rich cultural temperament and the tourist significance of the area. I would emphasize the unique combination of languages, religions and cultures within the city, and how this affects the tourist experience. I would conclude by presenting the spectacular and peaceful view of the city, and the spiritual and historical significance of Jerusalem as an important center in the world.")

Figure C.18: ***Multicultural Understanding***: LAION-Culture sample in Hebrew.

Multicultural Understanding: LAION-Culture (Tamil)



Human:
வெளிப்பாட்டு படத்தில் உள்ள புத்தகத்தின் கவர்ச்சியான கவர்ச்சிகளை விவரிக்கவும்
(Describe the charms of the book in expressive imagery.)

GPT:
புத்தகத்தின் கவர்ச்சிகள் இந்திய பாரம்பரியத்தை, கலையை, வரலாற்றை ஒருங்கிணைக்கும். புத்தகத்தின் மேற்பரப்பில் 'உபநிஷத் சாரம்: Upanishad Saram' என்ற தலைப்பு தன்னார்வலியாக எழுதப்பட்டுள்ளது. மையப்பகுதியில் ஒரு சூரிய ஒளியை சுற்றியுள்ள வ ட்டங்கள் காட்டும் சிற்பம் உள்ளது, இது உபநிஷத்களில் கொண்டிருக்கும் ஆன்மீக ஒளியுதர்வு அல்லது ஞானத்தின் ஆதாரமாக இரு க்கலாம். வலதுபுறத்தில் ஒரு ஆள் வரைபடம் உள்ளது, இது உபநிஷத்களில் கொண்டிருக்கும் ஞானம் மற்றும் ஆன்மீக அறிவுகளைக் குறிப்பிடுகிறது. இந்த புத்தகம் உபநிஷத்களின் ஆன்மீக அறிவு மற்றும் இறைவன் அல்லது இறைவியற்றின் ஆதாரத்தைப் பிரதிபலி க்கும்.

(The attractions of the book combine Indian heritage, art and history. The title 'Upanishad Saram: Upanishad Saram' is voluntarily written on the surface of the book. In the center there is a sculpture showing circles surrounding a sun, which may be a source of spiritual illumination or wisdom in the Upanishads. On the right is a human diagram, representing the wisdom and spiritual knowledge contained in the Upanishads. This book reflects the spiritual knowledge of the Upanishads and the source of the Lord or God.)

Human:
புத்தகத்தின் கவர்ச்சியான கவர்ச்சிகளைப் பெற்றோர் ஒரு குறுகிய கதை உருவாக்கவும்..
(Create a short story for parents of the book's attractive charms.)

GPT:
ஒரு ஆழ்ந்த காலம் முன்னால், ஒரு சிறு கிராமத்தில், ஒரு பெரிய ஆசாரி ஒரு உபநிஷத் சாரம் என்ற புத்தகத்தை எடுத்துக்கொண்டா ர். அந்த புத்தகத்தில் உபநிஷத்களின் ஆன்மீக அறிவு மற்றும் இறைவன் அல்லது இறைவியற்றின் ஆதாரம் காட்டப்பட்டுள்ளது. ஆசா ரி இந்த புத்தகத்தை எடுத்துக்கொண்டு கிராமத்தின் மக்களுக்கு அறிவு மற்றும் ஆன்மீக அறிவு பகிர்ந்தார். கிராமவாசிகள் இந்த அ றிவுடன் ஆன்மீக ஒளியுதர்வை அனுபவிக்க ஆரம்பித்தனர்.

(A long time ago, in a small village, a great priest picked up a book called an Upanishad Saram. In that book the spiritual knowledge of the Upanishads and the source of the Lord or God is shown. Asari took this book and imparted knowledge and spiritual knowledge to the people of the village. The villagers began to experience spiritual enlightenment with this knowledge.)

Figure C.19: ***Multicultural Understanding***: LAION-Culture sample in Tamil.

C.9 Breakdown Results of Different Languages on PANGEABENCH

C.9.1 xChat

We show the performance of different models on the xChat benchmark in Table C.3.

Models	English	Multi	Spanish	Hindi	Indonesian	Japanese	Korean	Chinese
Gemini-1.5-Pro	71.0	65.6	66.0	62.0	65.5	68.0	66.5	65.5
GPT4o	67.0	65.1	66.0	64.0	65.0	66.5	67.5	61.5
Llava-1.5-7B	22.5	16.7	22.5	3.5	18.0	23.0	12.0	21.0
Llava-Next-7B	40.5	20.4	33.0	1.5	19.0	25.0	15.0	29.0
Phi-3.5-Vision	38.5	21.1	37.0	11.5	10.5	31.0	12.5	24.0
Cambrian-8B	27.5	15.8	22.5	4.0	20.0	20.0	10.5	18.0
Llava-OV-7B	51.0	33.1	45.5	6.5	42.0	36.5	26.0	42.0
Molmo-7B-D	49.5	34.7	45.0	19.5	36.5	36.0	35.0	46.0
Llama3.2-11B	49.0	31.3	42.5	19.5	45.0	26.0	21.0	43.0
PaliGemma-3B	6.0	3.8	4.5	0.5	6.5	6.5	2.0	3.0
PALO-7B	27.0	16.2	23.0	3.0	19.0	20.0	13.5	18.5
mBLIP mT0-XL	2.5	0.5	0.0	0.0	0.5	2.0	0.5	0.0
mBLIP BLOOMZ-7B	4.0	1.7	2.0	2.5	2.5	0.0	0.0	3.0
PANGEA-7B (Ours)	46.0	35.8	43.5	23.5	34.5	39.0	33.5	40.5

Table C.3: Comparison of models on the xChat dataset across different languages.

C.9.2 Multilingual LLaVABench

We show the performance of different models on the Multilingual LLaVABench benchmark in Table C.4.

C.9.3 CVQA

We show the performance of different models on the CVQA benchmark in Table C.5 and Table C.6.

C.9.4 MaRVL

We show the performance of different models on the MaRVL benchmark in Table C.7.

C.9.5 XM100

We show the performance of different models on the XM100 benchmark in Table C.8.

Models	English	Multi	Arabic	Bengali	Chinese	French	Hindi	Japanese	Russian	Spanish	Urdu
Gemini-1.5-Pro	103.4	106.6	112.9	117.1	104.1	115.5	106.2	118.1	95.7	88.2	101.6
GPT4o	104.6	100.4	98.3	111.9	96.5	101.1	99.7	104.0	88.5	100.9	102.5
Llava-1.5-7B	66.1	40.8	26.4	11.9	50.7	63.8	23.2	70.0	46.5	59.2	15.4
Llava-Next-7B	78.9	50.7	24.9	11.2	72.8	91.4	18.0	70.1	71.8	82.9	13.4
Phi-3.5-Vision	70.8	58.0	50.1	35.1	69.2	86.0	35.9	63.0	67.6	75.6	39.3
Cambrian-8B	78.4	61.8	54.1	35.4	80.9	87.3	44.2	64.4	76.4	90.3	23.3
Llava-OV-7B	89.7	55.3	45.5	33.8	90.0	89.4	35.3	70.3	44.7	75.5	13.3
Molmo-7B-D	95.9	13.8	10.1	4.2	0.3	59.6	5.5	6.0	8.7	29.5	0.0
Llama3.2-11B	93.9	58.2	39.4	48.1	47.2	85.6	67.8	53.7	68.5	77.8	35.3
PaliGemma-3B	32.1	31.9	37.3	38.2	29.1	30.0	35.8	33.4	26.1	32.3	25.1
PALO-7B	68.9	71.2	79.1	54.6	71.5	83.9	61.9	66.6	80.9	74.4	68.2
mBLIP mT0-XL	32.7	28.2	33.7	26.2	3.6	39.8	26.9	26.8	34.1	36.9	26.0
mBLIP BLOOMZ-7B	43.5	41.0	48.1	44.1	30.6	53.3	39.1	29.8	38.1	51.5	34.0
PANGEA-7B (Ours)	84.2	89.5	91.0	94.9	94.4	93.8	84.9	92.8	91.2	87.4	75.5

Table C.4: Comparison of models on the Multilingual LLaVABench benchmark across different languages.

C.9.6 xGQA

We show the performance of different models on the xGQA benchmark in Table C.9.

C.9.7 MAXM

We show the performance of different models on the MAXM benchmark in Table C.10.

C.9.8 xMMMU

We show the performance of different models on the xMMMU benchmark in Table C.11.

C.9.9 M3Exam

We show the performance of different models on the M3Exam benchmark in Table C.12.

C.9.10 TyDiQA

We show the performance of different models on the TyDiQA benchmark in Table C.13.

C.9.11 XStoryCloze

We show the performance of different models on the XStoryCloze benchmark in Table C.14.

C.9.12 MGSM

We show the performance of different models on the MGSM benchmark in Table C.15.

Models	ar-es	br-pt	bu-bg	ch-es	ch-zh	co-es	ec-es	eg-ar	et-am	et-or
Llava-1.5-7B	37.8	51.1	35.6	42.4	44.4	50.6	48.6	31.5	27.8	31.8
Llava-Next-7B	52.5	62.3	41.5	59.0	51.1	54.8	50.8	33.5	29.5	36.9
Phi-3.5-Vision	54.0	57.2	36.9	57.7	51.1	52.3	50.1	38.4	27.8	32.2
Cambrian-8B	59.6	60.6	42.0	64.5	59.5	57.7	56.1	40.9	27.8	25.7
Llava-OV-7B	64.5	69.7	49.6	67.1	69.1	66.8	65.5	47.8	32.5	41.1
Molmo-7B-D	61.1	69.0	54.9	60.7	66.2	58.5	54.9	56.7	58.1	60.7
Llama3.2-11B	69.1	74.6	64.2	70.5	73.6	69.3	66.9	68.5	68.4	63.1
PaliGemma-3B	48.7	53.9	39.1	53.4	53.7	50.6	45.3	40.4	24.8	28.0
PALO-7B	50.9	56.7	36.7	55.1	45.3	48.5	46.4	28.6	19.2	32.7
mBLIP BLOOMZ-7B	45.3	51.4	30.5	45.3	51.1	46.9	44.8	35.9	23.9	25.7
mBLIP mT0-XL	40.8	44.4	38.0	44.9	39.9	41.9	42.5	31.0	35.9	26.6
PANGEA-7B (Ours)	68.3	72.9	53.9	70.5	74.0	64.7	63.5	49.3	36.3	35.5

Models	fr-br	in-bn	in-ta	in-te	ind-id	ind-jv	ind-mi	ind-sv	ir-ir	ja-jp
Llava-1.5-7B	29.4	31.1	29.8	28.0	41.7	32.0	32.7	33.5	42.6	37.4
Llava-Next-7B	27.4	31.1	28.8	28.0	42.2	38.7	40.2	35.5	42.6	32.5
Phi-3.5-Vision	29.3	39.0	40.0	36.8	45.0	38.2	38.2	30.8	39.6	39.7
Cambrian-8B	31.6	47.2	38.1	44.0	50.2	43.8	39.4	45.5	47.9	40.9
Llava-OV-7B	34.3	56.3	43.9	46.5	58.0	45.8	45.4	40.5	50.6	49.8
Molmo-7B-D	44.2	61.9	61.2	58.5	52.9	53.9	54.6	55.0	64.4	42.9
Llama3.2-11B	49.4	76.9	80.4	80.5	65.8	60.6	68.9	64.0	76.4	54.2
PaliGemma-3B	29.9	46.2	46.0	43.5	45.4	41.4	39.8	33.0	34.4	43.3
PALO-7B	29.1	37.8	31.2	25.0	41.3	32.3	32.3	32.0	42.9	30.5
mBLIP BLOOMZ-7B	26.7	41.9	40.0	42.0	41.9	35.4	35.1	32.0	29.4	31.0
mBLIP mT0-XL	23.5	36.4	44.2	39.0	37.4	37.4	34.7	31.0	35.3	30.0
PANGEA-7B (Ours)	34.6	59.1	51.9	54.5	62.1	49.5	47.8	53.0	56.4	48.3

Models	ke-sw	ma-my	me-es	mo-mg	ni-ig	no-ng	pk-ur	ph-fi	ro-ro	ru-ru
Llava-1.5-7B	34.4	42.2	42.4	26.9	34.5	47.5	26.4	43.8	47.0	51.0
Llava-Next-7B	46.2	45.7	51.4	33.3	35.0	56.9	36.6	46.8	52.3	53.5
Phi-3.5-Vision	46.0	45.1	46.3	31.9	33.3	50.0	35.2	41.4	47.4	50.5
Cambrian-8B	50.5	52.1	56.7	34.6	36.0	53.5	48.6	47.3	52.0	61.5
Llava-OV-7B	46.5	55.6	59.4	35.9	33.5	62.5	58.3	56.2	60.3	75.5
Molmo-7B-D	73.3	54.6	53.6	51.9	53.0	54.8	67.1	57.6	63.6	61.5
Llama3.2-11B	79.1	72.1	66.6	54.5	61.5	66.9	78.7	70.0	76.8	74.5
PaliGemma-3B	44.0	44.1	47.4	29.2	32.0	52.2	44.9	39.9	50.3	53.5
PALO-7B	35.9	42.5	44.3	28.8	29.5	49.2	44.4	39.4	46.0	47.0
mBLIP BLOOMZ-7B	37.0	42.5	44.8	28.8	33.0	49.2	47.7	31.5	46.0	34.0
mBLIP mT0-XL	45.1	40.6	44.9	29.2	30.5	42.8	40.3	32.0	43.7	42.0
PANGEA-7B (Ours)	64.1	59.7	62.2	42.3	46.0	64.5	66.2	58.6	64.6	74.0

Models	rw-ki	sg-zh	sk-ko	sp-es	sr-si	ur-es	macro
Llava-1.5-7B	31.1	44.3	44.5	56.9	24.9	37.8	38.7
Llava-Next-7B	34.5	44.8	43.4	63.5	29.8	41.0	42.6
Phi-3.5-Vision	31.1	43.9	55.2	62.4	28.0	43.3	42.4
Cambrian-8B	31.9	54.7	54.5	70.4	36.4	45.7	47.5
Llava-OV-7B	35.3	70.3	65.2	79.9	31.6	47.3	53.8
Molmo-7B-D	57.4	69.3	65.2	70.1	68.0	50.8	59.4
Llama3.2-11B	57.9	80.7	73.8	81.4	72.4	52.4	70.1
PaliGemma-3B	27.2	48.6	61.0	60.1	31.6	39.4	43.0
PALO-7B	28.9	45.8	44.5	64.8	28.0	39.4	39.3
mBLIP BLOOMZ-7B	29.4	47.6	33.1	56.6	28.0	39.4	36.9
mBLIP mT0-XL	33.2	36.8	38.3	53.5	31.1	39.1	37.6
PANGEA-7B (Ours)	35.7	65.6	70.7	72.6	39.1	49.8	57.2

Table C.5: Comparison of models on CVQA across different country-language pairs (in local languages). Includes Macro-Acc.

Models	ar-es	br-pt	bu-bg	ch-es	ch-zh	co-es	ec-es	eg-ar	et-am	et-or
Llava-1.5-7B	56.2	61.6	52.3	60.2	54.0	55.6	55.5	50.2	51.3	53.3
Llava-Next-7B	53.9	61.3	50.9	59.8	58.8	60.2	52.8	54.7	52.9	58.9
Phi-3.5-Vision	59.2	61.9	54.9	64.1	58.2	59.3	57.5	50.7	54.7	58.4
Cambrian-8B	57.7	66.5	56.1	65.4	64.3	59.3	60.2	56.7	60.3	56.5
Llava-OV-7B	63.0	73.9	59.3	65.8	68.8	65.1	63.3	62.1	59.8	59.3
Molmo-7B-D	57.7	65.8	45.6	63.7	68.5	57.3	55.0	43.8	31.6	38.8
Llama3.2-11B	66.8	72.9	54.4	72.6	72.0	66.4	65.2	56.7	41.9	32.2
PaliGemma-3B	51.7	59.5	49.3	51.7	54.9	54.8	47.2	51.2	52.6	51.4
PALO-7B	50.2	57.0	48.8	53.4	52.1	51.9	53.0	48.3	47.0	52.3
mBLIP mT0-XL	38.1	45.4	39.1	42.7	43.7	41.1	40.9	42.9	34.2	42.1
mBLIP BLOOMZ-7B	46.0	51.4	41.5	44.4	48.9	49.0	45.0	45.3	38.9	46.3
PANGEA-7B (Ours)	67.2	72.9	60.1	68.8	67.2	64.7	61.6	59.1	60.7	56.0
Models	**fr-br**	**in-bn**	**in-ta**	**in-te**	**ind-id**	**ind-jv**	**ind-mi**	**ind-sv**	**ir-ir**	**ja-jp**
Llava-1.5-7B	37.3	52.1	61.4	63.5	47.8	50.8	49.0	44.0	61.3	41.9
Llava-Next-7B	37.5	60.8	61.4	60.5	48.5	48.1	51.4	49.0	66.6	40.9
Phi-3.5-Vision	41.7	58.7	60.5	60.0	51.7	45.5	51.4	47.5	62.6	41.4
Cambrian-8B	40.7	68.5	65.6	63.0	55.1	50.2	58.2	56.0	66.6	42.4
Llava-OV-7B	44.2	69.6	72.0	70.5	59.0	55.9	59.4	58.5	76.4	47.3
Molmo-7B-D	29.6	47.9	36.4	41.5	50.5	45.1	43.4	39.5	43.6	44.8
Llama3.2-11B	36.3	62.9	66.4	66.5	63.6	48.8	58.2	54.0	57.4	58.1
PaliGemma-3B	37.3	59.1	66.0	62.5	49.3	48.1	43.4	46.0	58.3	44.8
PALO-7B	36.8	52.4	53.5	56.5	45.1	45.8	44.2	42.0	55.6	37.4
mBLIP mT0-XL	30.4	43.0	46.0	41.0	38.1	39.1	38.6	32.5	37.4	34.0
mBLIP BLOOMZ-7B	34.6	43.4	52.6	49.5	41.0	44.8	38.2	30.5	42.3	36.5
PANGEA-7B (Ours)	45.2	67.1	71.0	68.0	60.4	57.2	56.9	56.0	72.7	45.8
Models	**ke-sw**	**ma-my**	**me-es**	**mo-mg**	**ni-ig**	**no-ng**	**pk-ur**	**ph-fi**	**ro-ro**	**ru-ru**
Llava-1.5-7B	68.9	52.1	47.9	45.8	51.0	58.5	63.9	52.7	55.6	59.0
Llava-Next-7B	71.1	54.9	51.1	44.2	53.0	57.2	67.1	56.7	62.6	58.5
Phi-3.5-Vision	72.9	57.1	46.3	50.7	53.0	56.2	60.6	57.6	61.9	58.5
Cambrian-8B	74.4	61.9	56.7	48.7	56.5	60.5	73.1	60.1	66.6	61.5
Llava-OV-7B	79.1	65.1	63.2	52.6	57.5	64.2	75.0	64.0	72.5	72.5
Molmo-7B-D	47.6	51.7	55.1	35.9	36.0	49.2	46.8	43.3	52.0	63.5
Llama3.2-11B	61.5	69.2	64.7	41.0	39.5	65.9	65.7	66.0	75.5	74.5
PaliGemma-3B	59.7	54.9	51.7	43.4	46.0	55.2	67.6	48.8	60.9	56.0
PALO-7B	65.9	49.2	53.4	42.9	49.0	54.5	60.6	52.7	55.0	53.5
mBLIP mT0-XL	50.2	41.6	34.7	33.9	39.5	43.1	45.4	36.9	43.7	41.0
mBLIP BLOOMZ-7B	54.6	45.7	39.3	38.1	45.0	47.2	60.6	36.9	50.3	44.0
PANGEA-7B (Ours)	77.2	62.5	61.6	52.9	59.5	64.9	72.2	64.0	71.9	68.5
Models	**rw-ki**	**sg-zh**	**sk-ko**	**sp-es**	**sr-si**	**ur-es**	**macro**			
Llava-1.5-7B	51.1	60.8	56.9	66.0	58.7	42.5	54.2			
Llava-Next-7B	52.8	62.3	60.0	67.6	59.1	38.7	55.7			
Phi-3.5-Vision	52.3	59.4	66.5	66.7	61.3	46.3	56.3			
Cambrian-8B	56.2	66.0	63.1	71.7	63.1	47.0	59.7			
Llava-OV-7B	55.7	73.6	67.9	80.2	72.9	48.9	65.2			
Molmo-7B-D	34.9	66.0	56.9	66.7	31.6	44.8	48.3			
Llama3.2-11B	40.4	73.6	73.1	83.3	51.1	56.2	61.2			
PaliGemma-3B	44.7	59.4	58.3	61.0	62.2	40.6	52.9			
PALO-7B	51.9	56.1	55.9	62.9	54.2	42.2	50.9			
mBLIP mT0-XL	38.3	43.9	41.4	51.9	48.0	34.9	40.5			
mBLIP BLOOMZ-7B	45.1	53.8	46.9	58.5	46.7	34.0	44.9			
PANGEA-7B (Ours)	56.6	71.7	66.6	75.2	70.6	52.7	64.4			

Table C.6: Comparison of models on CVQA across different country-language pairs (in English). Includes Macro-Acc.

Models	English	Multi	Indonesian	Swahili	Tamil	Turkish	Chinese
GPT4o	81.8	82.3	81.9	80.8	80.2	86.4	82.1
Gemini-1.5-Pro	76.4	72.0	71.2	67.8	70.0	75.4	75.8
Llava-1.5-7B	56.2	53.7	56.1	49.8	49.7	55.4	57.5
Llava-Next-7B	62.8	50.9	52.2	50.6	50.5	50.4	50.6
Phi-3.5-Vision	72.1	56.5	58.6	51.4	52.0	58.6	61.7
Cambrian-8B	75.4	61.8	64.7	53.6	56.7	65.2	68.9
Llava-OV-7B	72.7	57.5	60.9	51.2	51.9	63.5	60.0
Molmo-7B-D	65.3	54.9	61.1	49.6	49.6	52.2	62.2
Llama3.2-11B	64.5	58.1	62.7	52.4	54.0	61.6	59.5
PaliGemma-3b	56.5	52.2	53.4	49.6	50.5	56.3	51.3
PALO-7B	63.3	54.2	58.3	50.6	51.9	54.9	55.3
mBLIP mT0-XL	67.3	66.7	64.9	64.8	69.7	68.1	65.9
mBLIP BLOOMZ-7B	62.3	58.6	59.1	56.2	60.3	57.7	59.7
PANGEA-7B	87.0	79.0	81.3	75.1	69.4	84.8	84.3

Table C.7: Comparison of models on the MaRVL dataset across different languages.

C.9.13 MMMLU

We show the performance of different models on the MMMLU benchmark in Table C.16.

Models	English	Multi	Arabic	Bengali	Czech	Danish	German	Greek
Gemini-1.5-Pro	27.6	19.1	1.7	7.5	25.9	32.8	27.6	5.0
GPT4o	27.7	19.1	15.8	13.5	21.1	25.3	19.3	21.1
Llava-1.5-7B	28.6	1.1	0.0	0.0	2.1	1.0	3.1	0.0
Llava-Next-7B	29.3	9.4	5.6	0.1	12.1	15.7	14.4	4.2
Phi-3.5-Vision	30.2	5.2	0.4	2.4	16.6	16.2	0.0	20.7
Cambrian-8B	20.6	9.9	1.4	6.6	7.4	15.1	15.5	4.4
Llava-OV-7B	30.6	7.0	0.2	0.6	5.2	16.8	14.0	0.4
Molmo-7B-D	22.1	9.1	5.4	7.9	5.7	13.8	12.2	4.2
Llama3.2-11B	27.6	4.5	0.0	0.0	1.5	11.8	4.6	1.2
PaliGemma-3B	18.7	0.8	0.0	0.0	1.1	3.1	2.7	0.0
PALO-7B	30.4	0.8	0.0	0.0	2.0	1.0	2.7	0.0
mBLIP mT0-XL	31.9	3.1	3.2	1.6	3.7	2.1	2.9	3.1
mBLIP BLOOMZ	22.5	10.3	9.5	6.4	11.5	15.9	14.5	10.9
PANGEA-7B (Ours)	30.4	14.2	18.1	16.4	16.2	20.7	20.6	11.2

Models	Spanish	Persian	Finnish	Filipino	French	Hebrew	Hindi	Croatian
Gemini-1.5-Pro	39.5	4.2	29.0	28.7	42.4	4.3	2.2	33.8
GPT4o	28.3	26.6	13.1	26.4	23.1	20.4	17.0	19.4
Llava-1.5-7B	3.7	0.0	0.4	1.1	2.0	0.1	0.0	0.3
Llava-Next-7B	23.6	9.4	5.5	9.3	23.0	2.7	10.2	7.5
Phi-3.5-Vision	20.7	0.0	1.0	1.7	21.2	0.3	0.0	0.5
Cambrian-8B	18.6	9.6	5.1	19.6	18.3	5.8	6.8	7.2
Llava-OV-7B	24.9	3.8	1.5	4.2	22.0	0.0	4.4	7.2
Molmo-7B-D	19.8	11.3	3.1	13.0	19.8	8.3	9.4	6.9
Llama3.2-11B	10.2	0.0	2.4	8.4	12.0	0.0	0.2	0.7
PaliGemma-3B	0.7	0.0	0.1	0.1	0.6	0.0	0.0	1.3
PALO-7B	1.5	0.0	0.4	0.9	2.1	0.0	0.0	0.2
mBLIP mT0-XL	8.3	5.5	1.7	2.8	6.4	4.0	1.8	0.9
mBLIP BLOOMZ	18.9	13.8	4.8	7.7	19.1	7.5	10.1	3.2
PANGEA-7B (Ours)	26.2	19.3	3.8	18.9	26.7	18.2	17.4	10.8

Models	Hungarian	Indonesian	Italian	Japanese	Korean	Maori	Dutch	Norwegian
Gemini-1.5-Pro	37.2	55.4	27.6	1.2	8.2	3.8	27.7	36.7
GPT4o	21.8	28.4	21.0	0.0	11.1	26.8	26.4	24.7
Llava-1.5-7B	3.3	0.9	4.3	0.0	0.0	0.2	2.9	3.7
Llava-Next-7B	9.3	14.7	17.6	4.2	5.2	9.2	23.8	16.3
Phi-3.5-Vision	3.4	3.2	17.5	1.6	0.3	0.2	17.2	14.1
Cambrian-8B	6.6	15.7	15.5	7.2	2.0	3.2	20.3	16.0
Llava-OV-7B	3.6	16.4	12.8	0.6	0.0	1.7	24.7	13.9
Molmo-7B-D	3.5	17.2	17.8	5.2	2.4	7.5	15.7	13.8
Llama3.2-11B	12.7	1.2	16.0	0.0	0.0	9.3	22.0	1.1
PaliGemma-3B	2.0	0.2	1.8	0.0	0.0	4.0	2.6	2.3
PALO-7B	3.4	1.1	3.2	0.0	0.0	0.1	3.5	0.7
mBLIP mT0-XL	2.8	6.0	2.8	0.3	2.1	1.5	3.4	3.1
mBLIP BLOOMZ	11.8	16.0	16.5	0.0	4.5	0.1	18.2	14.5
PANGEA-7B (Ours)	7.7	27.9	22.9	2.1	8.1	0.7	26.6	24.9

Models	Polish	Portuguese	Quechua	Romanian	Russian	Swedish	Swahili	Telugu
Gemini-1.5-Pro	35.5	35.7	0.7	31.2	32.4	37.8	10.7	0.0
GPT4o	22.2	28.0	4.4	19.1	20.7	26.0	20.0	12.5
Llava-1.5-7B	0.8	2.5	0.0	1.6	0.5	2.0	0.1	0.0
Llava-Next-7B	13.5	21.3	0.0	11.5	13.5	16.0	3.2	0.0
Phi-3.5-Vision	1.0	21.0	0.4	3.2	0.7	12.5	0.4	0.0
Cambrian-8B	9.3	17.5	0.0	13.4	11.3	17.9	3.7	2.3
Llava-OV-7B	7.4	24.6	0.0	6.8	5.5	15.0	2.0	0.0
Molmo-7B-D	8.2	16.2	0.6	11.6	12.3	14.1	3.8	0.4
Llama3.2-11B	1.0	18.6	0.0	10.1	0.6	7.4	5.8	0.0
PaliGemma-3B	0.9	1.3	0.1	0.8	0.0	2.0	0.0	0.0
PALO-7B	0.8	1.7	0.0	1.1	0.5	0.9	0.2	0.0
mBLIP mT0-XL	3.5	5.8	0.2	2.3	3.1	3.7	3.8	2.7
mBLIP BLOOMZ	11.8	16.5	0.1	13.7	14.5	14.5	8.4	3.0
PANGEA-7B (Ours)	16.2	28.1	0.0	21.4	20.9	19.4	18.7	0.1

Models	Thai	Turkish	Ukrainian	Vietnamese	Chinese
Gemini-1.5-Pro	0.0	0.9	0.0	0.0	0.9
GPT4o	0.0	17.6	16.9	30.9	0.4
Llava-1.5-7B	0.0	0.0	0.0	0.0	0.0
Llava-Next-7B	0.0	0.0	0.3	0.0	6.3
Phi-3.5-Vision	0.5	1.9	0.0	2.2	0.0
Cambrian-8B	0.4	9.3	5.9	17.8	11.3
Llava-OV-7B	0.0	0.0	0.0	0.0	2.9
Molmo-7B-D	0.0	0.0	0.0	0.0	0.0
Llama3.2-11B	0.0	0.0	0.0	0.0	2.9
PaliGemma-3B	0.5	0.0	0.0	0.2	0.0
PALO-7B	0.2	0.0	0.0	0.1	0.0
mBLIP mT0-XL	0.0	3.9	2.0	7.1	0.0
mBLIP BLOOMZ	0.5	1.9	0.0	2.2	0.0
PANGEA-7B (Ours)	0.0	0.0	0.3	0.0	4.9

Table C.8: Comparison of models on the XM100 dataset across different languages.

Models	English	Multi	Bengali	German	Indonesian	Korean	Portuguese	Russian	Chinese
Gemini-1.5-Pro	54.2	48.7	49.4	50.2	48.6	46.4	51.2	44.8	50.2
GPT4o	55.8	51.0	49.4	52.6	50.4	51.0	52.2	50.0	51.4
Llava-1.5-7B	62.0	30.7	15.6	28.4	33.4	38.2	27.5	33.1	38.4
Llava-Next-7B	64.8	37.8	11.5	41.5	37.3	42.5	39.8	43.5	48.2
Phi-3.5-Vision	64.7	38.4	7.7	51.4	36.0	36.3	49.6	46.2	41.4
Cambrian-8B	64.6	39.8	32.3	44.6	36.0	43.6	41.6	44.2	36.2
Llava-OV-7B	64.4	48.2	41.8	49.2	48.8	45.3	52.4	54.0	45.9
Molmo-7B-D	51.5	43.0	25.6	45.9	44.9	44.2	46.5	45.6	48.1
Llama3.2-11B	55.6	45.4	42.9	46.7	46.2	44.5	46.5	44.7	46.1
PaliGemma-3B	59.7	30.5	13.3	44.5	21.3	22.8	34.7	35.8	41.2
PALO-7B	60.5	37.8	42.2	39.1	36.8	41.7	31.7	27.0	46.5
mBLIP mT0-XL	44.2	39.9	39.1	41.1	39.1	39.7	40.7	40.2	39.4
mBLIP BLOOMZ-7B	43.3	36.9	37.7	36.3	39.3	28.5	40.7	36.6	39.1
PANGEA-7B (Ours)	64.7	60.2	58.9	61.6	60.1	58.9	61.8	60.4	59.6

Table C.9: Comparison of models on the xGQA dataset across different languages

Models	English	Multi	French	Hindi	Hebrew	Romanian	Thai	Chinese
Gemini-1.5-Pro	56.4	63.5	60.2	66.5	65.7	57.4	73.9	57.4
GPT4o	60.7	65.4	59.8	68.8	70.0	61.3	76.5	56.3
Llava-1.5-7B	49.8	20.4	32.2	17.3	12.9	15.1	17.2	27.8
Llava-Next-7B	54.9	21.4	33.7	16.2	10.7	15.5	18.3	33.9
Phi-3.5-Vision	55.3	25.0	38.3	31.9	17.5	10.9	24.3	27.4
Cambrian-8B	55.3	28.7	41.7	23.8	17.1	32.0	25.7	31.8
Llava-OV-7B	54.9	34.8	37.9	31.9	17.8	30.2	53.0	37.9
Molmo-7B-D	52.9	37.5	45.5	33.5	30.7	28.9	46.3	40.4
Llama3.2-11B	55.3	43.9	48.1	50.4	41.8	36.6	56.7	30.0
PaliGemma-3B	47.9	19.9	8.0	36.5	19.3	13.4	31.3	10.8
PALO-7B	51.4	16.3	33.7	15.8	12.1	11.3	14.6	10.5
mBLIP mT0-XL	44.7	36.8	36.0	42.7	28.9	30.3	56.3	26.4
mBLIP BLOOMZ-7B	44.7	24.8	33.0	47.3	8.9	16.9	9.7	33.2
PANGEA-7B (Ours)	55.3	53.3	43.6	53.5	59.3	45.8	67.2	50.2

Table C.10: Comparison of models on the MAXM dataset across different languages.

C.10 A Preliminary Exploration of Constructing Multilingual OCR Instructions

Optical Character Recognition (OCR) is a critical capability for multimodal LLMs, enabling them to interpret and process textual information embedded within images. However, most existing OCR training datasets are predominantly English-centric, which limits the models' performance in non-English contexts. To address this gap, we have curated a comprehensive set of 500K multilingual OCR training samples from web user interfaces, spanning 10 languages, with 50K examples per language, sourced from web user interfaces. Webpages naturally serve

Models	English	Multi	Arabic	French	Hindi	Indonesian	Japanese	Portuguese
Gemini-1.5-Pro (0801)	65.8	57.7	57.7	58.1	55.5	60.2	55.0	59.6
GPT4o (0513)	69.1	58.3	56.7	58.1	58.1	59.9	58.0	58.9
Llava-1.5-7B	36.2	31.5	29.5	34.9	27.5	31.6	32.0	33.7
Llava-Next-7B	36.7	34.3	30.5	35.6	30.9	37.0	34.9	37.0
Phi-3.5-Vision	42.6	38.8	35.6	44.0	30.9	36.7	37.9	47.8
Cambrian-8B	41.8	33.2	32.6	34.6	30.9	31.3	33.5	36.0
Llava-OV-7B	46.3	41.0	41.6	43.0	34.7	43.4	40.1	43.4
Molmo-7B-D	42.9	40.4	40.6	42.6	32.6	40.7	43.9	42.1
Llama3.2-11B	39.2	34.0	33.6	39.6	32.3	36.7	29.0	33.0
PaliGemma-3B	26.3	25.2	29.2	23.8	21.6	24.2	24.5	27.6
PALO-7B	33.1	30.5	30.5	33.2	28.9	34.0	27.1	33.3
mBLIP mT0-XL	29.3	30.4	30.2	33.2	28.2	26.9	31.6	32.3
mBLIP BLOOMZ-7B	29.2	30.8	28.5	33.9	27.8	33.3	31.6	29.6
PANGEA-7B (Ours)	45.7	43.7	42.3	45.3	41.6	46.5	40.5	46.1

Table C.11: Comparison of models on the xMMMU dataset across different languages.

as image-rich environments containing abundant text, and by capturing screenshots of websites from various countries in different languages, we were able to gather a substantial number of OCR images.

We utilize URLs from the CC-News-Multilingual[2]dataset [47] to obtain a diverse set of multilingual web pages. Using Playwright[3], we render each website and automatically capture screenshots under various device settings and resolutions to achieve a wide range of image dimensions and aspect ratios. Each screenshot includes a red bounding box that highlights a specific element targeted for OCR extraction. We focus on ten languages for this dataset: English, Chinese, Japanese, Korean, Indonesian, Hindi, Spanish, French, Portuguese, and Arabic. We totally have 1M samples (50K for each language).

We employed the same model architecture as PANGEA but trained it exclusively on these OCR images, reserving a portion of the data as a test set. As shown in Figure C.20 , the results indicate that improving multilingual OCR performance is feasible with an increase in training samples. However, the OCR accuracy for non-Latin scripts (e.g., Chinese, Japanese, and Korean) remains lower than for Latin-based languages. Looking ahead, we aim to further expand the multilingual OCR training dataset to include more languages and integrate this data into PANGEAINS.

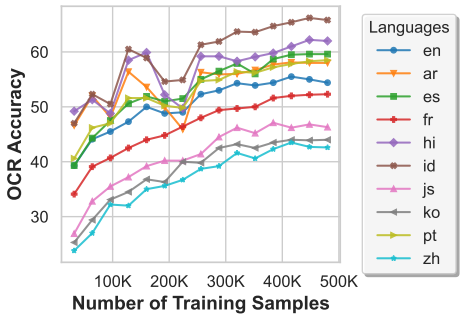


Figure C.20: A preliminary exploration of multilingual OCR.

[2]https://huggingface.co/datasets/intfloat/multilingual_cc_news
[3]https://github.com/microsoft/playwright

Models	English	Multi	Afrikaans	Chinese	Italian	Portuguese	Thai	Vietnamese
Gemini-1.5-Pro	77.4	64.7	80.4	74.1	76.3	61.8	49.9	46.0
GPT4o	68.0	61.0	73.0	68.0	67.0	58.0	52.0	48.3
Llava-1.5-7B	32.3	29.0	28.2	24.3	40.1	28.2	23.7	29.3
Llava-Next-7B	36.5	28.4	28.2	25.4	37.8	27.0	23.7	28.4
Phi-3.5-Vision	55.8	37.2	44.2	40.8	51.4	40.3	25.2	21.6
Cambrian-8B	34.7	33.4	36.8	34.2	45.2	30.3	28.9	25.0
Llava-OV-7B	60.4	45.8	50.3	58.0	57.2	43.8	30.9	34.5
Molmo-7B-D	57.1	39.1	35.6	56.4	49.4	40.2	27.4	25.9
Llama3.2-11B	51.8	36.6	42.3	46.4	45.8	28.4	26.4	30.2
PaliGemma-3B	36.0	25.6	26.4	24.7	32.2	24.3	27.2	19.0
PALO-7B	30.8	27.8	31.9	22.1	36.9	32.3	22.7	20.7
mBLIP mT0-XL	22.8	25.0	16.0	25.6	33.7	21.2	22.4	31.0
mBLIP BLOOMZ-7B	30.3	29.5	28.2	29.8	37.3	28.3	22.9	30.2
PANGEA-7B (Ours)	61.4	42.1	52.1	49.2	54.9	43.3	32.9	19.8

Table C.12: Comparison of models on the M3Exam dataset across different languages.

Models	English	Multi	Arabic	Bengali	Finnish	Indonesian	Korean	Russian	Swahili	Telugu
Vicuna-1.5-7B	59.7	52.7	32.3	68.1	63.0	72.6	58.8	57.6	51.3	18.1
Qwen2-7B-Instruct	72.2	71.2	67.6	75.9	67.1	78.0	64.9	67.2	75.3	73.8
Llava-1.5-7B	66.8	52.8	61.8	33.4	60.2	72.8	63.3	55.0	55.0	20.6
Llava-Next-7B	68.3	52.1	64.5	24.9	63.0	74.3	61.9	58.4	53.1	17.0
Phi-3.5-Vision	75.9	51.3	63.1	24.8	57.3	70.6	60.2	57.5	48.7	28.3
PALO-7B	69.4	50.8	60.9	46.0	61.8	70.6	56.8	56.7	42.5	10.8
PANGEA-7B (Ours)	73.7	66.0	55.5	65.3	66.3	74.5	69.4	60.1	76.6	60.0

Table C.13: Comparison of models on the TyDiQA dataset across different languages.

Models	English	Multi	Arabic	Spanish	Basque	Hindi	Ind.	Burmese	Russian	Swahili	Telugu	Chinese
Vicuna-1.5-7B	78.1	57.4	52.7	69.4	50.8	54.5	61.0	48.4	66.5	52.1	54.5	63.5
Qwen2-7B-Instruct	80.3	61.9	64.0	71.6	51.6	59.6	68.5	50.7	72.7	53.2	55.3	72.1
Llava-1.5-7B	79.1	57.6	52.7	69.2	50.9	54.9	62.6	49.0	65.9	51.7	55.8	63.9
Llava-Next-7B	79.1	57.1	51.7	68.8	50.3	54.5	62.0	46.7	65.5	52.1	55.2	63.8
Phi-3.5-Vision	77.9	54.8	53.7	67.2	50.4	54.9	51.7	47.8	61.3	49.3	52.5	59.5
PALO-7B	77.4	57.2	56.5	68.4	49.8	58.6	58.5	47.4	65.6	51.2	53.1	62.8
PANGEA-7B (Ours)	79.1	61.2	60.5	67.8	50.0	61.8	66.4	48.7	69.4	58.9	60.4	68.2

Table C.14: Comparison of models on the XStoryCloze dataset across different languages.

Models	English	Multi	Bengali	German	Spanish	French	Japanese	Russian	Swahili	Telugu	Thai	Chinese
Vicuna-1.5-7B	17.6	6.4	0.0	14.4	9.6	14.4	2.8	10.8	3.6	0.0	2.0	14.8
Qwen2-7B-Instruct	48.8	40.4	0.0	67.2	67.6	68.8	11.2	71.2	10.8	2.4	45.6	59.2
Llava-1.5-7B	14.8	7.6	0.0	15.2	10.8	18.0	2.8	11.2	0.4	0.0	1.6	15.6
Llava-Next-7B	15.6	7.5	0.0	13.6	13.2	16.0	1.6	12.8	2.0	0.0	1.6	14.0
Phi-3.5-Vision	59.2	33.1	0.0	64.0	59.6	58.0	20.0	54.0	4.0	0.0	18.8	52.4
PALO-7B	13.6	5.8	0.0	11.6	9.6	13.2	1.6	8.8	0.4	0.0	0.0	12.4
PANGEA-7B (Ours)	82.0	47.4	0.0	68.4	74.8	63.2	22.0	68.0	54.0	5.6	49.6	68.0

Table C.15: Comparison of models on the MGSM dataset across different languages.

Models	English	Multi	Arabic	Bengali	Portuguese	Chinese	French	German
Vicuna-1.5-7B	49.5	34.7	30.3	28.5	39.6	36.9	40.4	39.8
Qwen2-7B-Instruct	70.1	53.1	51.0	43.4	60.7	63.8	61.5	57.7
Llava-1.5-7B	50.2	34.9	29.7	28.5	40.3	36.8	40.1	39.8
Llava-Next-7B	52.1	35.6	30.0	28.8	40.7	37.3	41.4	41.4
Phi-3.5-Vision	62.0	39.1	34.9	27.9	47.6	41.5	49.2	45.8
PALO-7B	46.7	32.6	30.3	29.5	36.0	34.2	36.9	35.8
PANGEA-7B (Ours)	68.4	52.2	49.3	44.4	58.9	60.5	58.9	56.7

Models	Hindi	Indonesian	Italian	Japanese	Korean	Spanish	Swahili	Yoruba
Vicuna-1.5-7B	29.8	36.5	39.5	35.9	34.1	40.3	27.9	26.8
Qwen2-7B-Instruct	45.7	57.1	60.8	58.0	54.6	61.9	36.0	31.8
Llava-1.5-7B	29.2	37.1	41.0	35.1	34.1	41.6	28.0	27.3
Llava-Next-7B	29.6	37.5	41.2	36.0	34.2	42.7	28.5	28.7
Phi-3.5-Vision	32.9	38.3	47.0	40.0	36.6	49.6	28.9	27.8
PALO-7B	29.6	33.7	36.4	32.7	30.6	37.0	26.4	27.1
PANGEA-7B (Ours)	45.7	55.4	58.8	55.3	52.7	59.7	42.8	31.3

Table C.16: Comparison of models on the MMMLU dataset across different languages.

Bibliography

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, abs/2404.14219, 2024. URL https://arxiv.org/abs/2404.14219. 5.4.1

[2] Talor Abramovich, Meet Udeshi, Minghao Shao, Kilian Lieret, Haoran Xi, Kimberly Milner, Sofija Jancheska, John Yang, Carlos E. Jimenez, Farshad Khorrami, Prashanth Krishnamurthy, Brendan Dolan-Gavitt, Muhammad Shafique, Karthik Narasimhan, Ramesh Karri, and Ofir Press. Enigma: Enhanced interactive generative model agent for ctf challenges, 2024. URL https://arxiv.org/abs/2409.16165. 2.1.2

[3] Tom Agonnoude and Cyrile Delestre. Table vqa dataset, 2024. URL https://huggingface.co/datasets/cmarkea/table-vqa. 5.2.3

[4] Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proc. of EMNLP*, pages 9904–9923, 2023. URL https://aclanthology.org/2023.emnlp-main.614.pdf. 5.1

[5] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proc. of ACL*, pages 12404–12422, 2024. URL https://aclanthology.org/2024.acl-long.671.pdf. 5.1

[6] https://www.anthropic.com/index/introducing-claude Anthropic. Claude, 2022. URL https://www.anthropic.com/index/introducing-claude. 4.3.1, 4.6

[7] Humanity's Last Exam's Authors. Humanity's last exam. *ArXiv*, abs/2501.14249, 2025. URL https://api.semanticscholar.org/CorpusID:275906652. 4.6

[8] Laura Babcock and Antonino Vallesi. The interaction of process and domain in prefrontal cortex during inductive reasoning. *Neuropsychologia*, 67:91–99, 2015. 4.5.2

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2024. URL https://openreview.net/forum?id=qrGjFJVl3m. 4.3.1

[10] Daniil Belopolskih and Egor Spirin. Gqa-ru, 2024. URL https://huggingface.co/datasets/deepvk/GQA-ru. 5.2.3

[11] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *ArXiv preprint*, abs/2407.07726, 2024. URL https://arxiv.org/abs/2407.07726. 5.4.1, C.1

[12] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *ArXiv preprint*, abs/2405.14782, 2024. URL https://arxiv.org/abs/2405.14782. 5.4.1

[13] Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. Vasr: Visual analogies of situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 241–249, 2023. B.1.2

[14] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In *Proc. of ACL*, pages 5486–5505, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.376. 5.1

[15] S.R.K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. Reinforcement learning for mapping instructions to actions. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 82–90, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1010. 3.1

[16] Aimée K Bright and Aidan Feeney. Causal knowledge and the development of inductive reasoning. *Journal of Experimental Child Psychology*, 122:48–61, 2014. 4.5.2

[17] BUAA. Chinese-llava-med, 2023. URL https://huggingface.co/BUAADreamer/Chinese-LLaVA-Med-7B. Accessed: 2024-10-01. 5.2.3

[18] Andrea Burns, Deniz Arsan, Sanjna Agrawal, Ranjitha Kumar, Kate Saenko, and Bryan A. Plummer. A dataset for interactive vision-language navigation with unknown command feasibility. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, page 312–328, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20073-1. doi: 10.1007/978-3-031-20074-8_18. URL https://doi.org/10.1007/978-3-031-20074-8_18. 2.1.2

[19] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. Visibility into ai agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, 2024. 3.1

[20] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *ArXiv preprint*, abs/2209.05401, 2022. URL https://arxiv.org/abs/2209.05401. 5.3.2, C.4.1

[21] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *ArXiv preprint*, abs/2402.11684, 2024. URL https://arxiv.org/abs/2402.11684. 5.2.3

[22] Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. *arXiv preprint arXiv:2407.07061*, 2024. 3.1

[23] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *ArXiv preprint*, abs/2305.18565, 2023. URL https://arxiv.org/abs/2305.18565. C.1

[24] Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-FLAN: Designing data and methods of effective agent tuning for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 9354–9366, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.557. URL https://aclanthology.org/2024.findings-acl.557. 2.1.2

[25] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin Smith, and Joshua B Tenenbaum. Are deep neural networks smarter than second graders? *arXiv preprint arXiv:2212.09993*, 2022. B.1.2

[26] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A. Smith, and Joshua B. Tenenbaum. Are deep neural networks smarter than second graders? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10834–10844, 2022. URL https://api.semanticscholar.org/CorpusID:254877678. 4.6

[27] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. URL https://aclanthology.org/2020.tacl-1.30. 5.3.3, C.4.2

[28] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA. C.1

[29] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement

learning, 2025. URL https://arxiv.org/abs/2501.12948. 4.1

[30] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *ArXiv preprint*, abs/2409.17146, 2024. URL https://arxiv.org/abs/2409.17146. 5.1, 5.4.1, 5.4.2

[31] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web, 2023. 2.1.1, 2.1.2, 3.7

[32] Khang T Doan, Bao G Huynh, Dung T Hoang, Thuc D Pham, Nhat H Pham, Quan Nguyen, Bang Q Vo, and Suong N Hoang. Vintern-1b: An efficient multimodal large language model for vietnamese. *ArXiv preprint*, abs/2408.12480, 2024. URL https://arxiv.org/abs/2408.12480. 5.2.3

[33] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024. 3.1

[34] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks?, 2024. 3.4.4, A.1

[35] Yu Du, Fangyun Wei, and Hongyang Zhang. Anytool: Self-reflective, hierarchical agents for large-scale api calls. *arXiv preprint arXiv:2402.04253*, 2024. A.1

[36] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783, 2024. URL https://arxiv.org/abs/2407.21783. 4.1, 4.3.1, 4.6, 5.1, 5.2.2, 5.4.1

[37] Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, Demetri Terzopoulos, Ade Famoti, Noboru Kuno, Ashley Llorens, Hoi Vo, Katsu Ikeuchi, Li Fei-Fei, Jianfeng Gao, Naoki Wake, and Qiuyuan Huang. An interactive agent foundation model, 2024. URL https://arxiv.org/abs/2402.05929. 3.1

[38] Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. Autoguide: Automated generation and selection of context-aware guidelines for large language model agents. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=Zu1MihB661. 2.1.2

[39] Jingying Gao, Qi Wu, Alan Blair, and Maurice Pagnucco. Lora: A logical reasoning augmented dataset for visual question answering. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 4.2.2, 4.6

[40] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang, editors, *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.alvr-1.2. URL https://aclanthology.org/2024.alvr-1.2. C.1

[41] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang, editors, *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, 2024. URL https://aclanthology.org/2024.alvr-1.2.pdf. 5.1, 5.4.1, C.1, C.1

[42] Vinod Goel and Raymond J Dolan. Differential involvement of left prefrontal cortexin inductive and deductive reasoning. *Cognition*, 93(3):B109–B121, 2004. 4.5.2

[43] Adam E Green, David JM Kraemer, Jonathan A Fugelsang, Jeremy R Gray, and Kevin N Dunbar. Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral cortex*, 20(1):70–76, 2010. 4.5.2

[44] Yu Gu, Yiheng Shu, Hao Yu, Xiao Liu, Yuxiao Dong, Jie Tang, Jayanth Srinivasa, Hugo Latapie, and Yu Su. Middleware for llms: Tools are instrumental for language agents in complex environments, 2024. URL https://arxiv.org/abs/2402.14672. 2.1.2

[45] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html. 5.1

[46] Izzeddin Gur, Hiroki Furuta, Austin V Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9JQtrumvg8. 2.1.2

[47] Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. newsplease: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223, 2017. URL https://edoc.hu-berlin.de/server/api/core/bitstreams/b83362d0-0fb5-4b65-8370-b31f187223a4/content. C.10

[48] Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. *ArXiv preprint*, abs/2310.10418, 2023. URL https://arxiv.org/abs/2310.10418. C.5

[49] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhen-

zhong Lan, and Dong Yu. WebVoyager: Building an end-to-end web agent with large multimodal models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6864–6890, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.371. URL https://aclanthology.org/2024.acl-long.371. 2.1.2

[50] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024. 3.7

[51] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proc. of ICLR*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ. C.4.2

[52] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proc. of ACL*, pages 688–714, 2023. URL https://aclanthology.org/2023.acl-long.41.pdf. C.5

[53] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents, 2023. URL https://arxiv.org/abs/2312.08914. 3.1

[54] Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, Ashley Llorens, John Langford, Hoi Vo, Li Fei-Fei, Katsu Ikeuchi, and Jianfeng Gao. Position paper: Agent ai towards a holistic intelligence, 2024. URL https://arxiv.org/abs/2403.00833. 3.1

[55] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 4.1, 4.3.1, 4.6

[56] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *ArXiv preprint*, abs/2109.05125, 2021. URL https://arxiv.org/abs/2109.05125. C.1

[57] Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on translating images for cultural relevance. *arXiv preprint arXiv:2404.01247*, 2024. 5.1

[58] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2111.15664. C.5

[59] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sang-

doo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2310.08491. 5.3.2

[60] Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *ArXiv preprint*, abs/2406.05761, 2024. URL https://arxiv.org/abs/2406.05761. 5.3.2, C.5

[61] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *ArXiv preprint*, abs/2405.01535, 2024. URL https://arxiv.org/abs/2405.01535. 5.3.2, C.5

[62] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.50. URL https://aclanthology.org/2024.acl-long.50. 2.1.1, 3.1, 3.7

[63] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. *arXiv preprint arXiv:2407.01476*, 2024. A.1

[64] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. URL https://arxiv.org/abs/1602.07332. C.4.1

[65] Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. Autowebglm: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5295—-5306, 2024. A.1

[66] Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheusvision: Vision-language model as a judge for fine-grained evaluation. *ArXiv preprint*, abs/2401.06591, 2024. URL https://arxiv.org/abs/2401.06591. 5.3.2

[67] Bo Li*, Peiyuan Zhang*, Kaicheng Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL https://github.com/EvolvingLMMs-Lab/lmms-eval. 4.3.1, 5.4.1

[68] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 5.4.1

[69] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4.1, 4.3.1, 4.6

[70] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. https://huggingface.co/AI-MO/NuminaMath-CoT, 2024. https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf. 5.2.1

[71] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^{3}it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. C.1

[72] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. C.1

[73] Yang Li, Jiacong He, Xin Zhou, Yuan Zhang, and Jason Baldridge. Mapping natural language instructions to mobile UI action sequences. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8198–8210, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.729. URL https://aclanthology.org/2020.acl-main.729. 3.1

[74] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models, 2021. URL https://arxiv.org/abs/2112.10668. 5.3.3

[75] LinkSoul-AI. Chinese-llava, 2023. URL https://huggingface.co/spaces/LinkSoul/Chinese-LLaVa. Accessed: 2024-10-01. 5.2.3

[76] Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryTp3f-0-. 3.1

[77] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proc. of EMNLP*, pages 10467–10485, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. URL https://aclanthology.org/

2021.emnlp-main.818. 5.1, 5.3.2, C.4.1

[78] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. URL https://arxiv.org/abs/2310.03744. 4.1, 4.3.1, 5.4.1, C.1

[79] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2304.08485. 5.1, 5.3.2, 5.4.1, C.1

[80] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. URL https://arxiv.org/pdf/2401.13601. 4.3.1, 5.1, 5.2.1, 5.4.1, C.1

[81] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020. 4.2.2

[82] Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding. *ArXiv*, abs/2410.13824, 2024. URL https://api.semanticscholar.org/CorpusID:273403951. 4.6

[83] Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding?, 2024. URL https://arxiv.org/abs/2404.05955. 2.1.2

[84] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv: 2308.03688*, 2023. 2.1.2

[85] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024. 2.1.2

[86] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2023. URL https://api.semanticscholar.org/CorpusID:259837088. 4.6

[87] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 4.1, 4.2.2, 4.6

[88] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math-

ematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02255. C.5

[89] Michael Lutz, Arth Bohra, Manvel Saroyan, Artem Harutyunyan, and Giovanni Campagna. Wilbur: Adaptive in-context learning for robust and accurate web agents, 2024. 2.1.2

[90] Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue, 2024. URL https://arxiv.org/abs/2402.05930. 2.1.2

[91] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, 2019. URL https://api.semanticscholar.org/CorpusID:173991173. 4.6

[92] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.177. 5.2.1

[93] Michael Meng, Stephanie Steinhardt, and Andreas Schubert. Application programming interface documentation: What do software developers want? *Journal of Technical Writing and Communication*, 48(3):295–330, 2018. 3.1

[94] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3P: learning universal representations via multitask multilingual multimodal pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3977–3986. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ni_M3P_Learning_Universal_Representations_via_Multitask_Multilingual_Multimodal_Pre-Training_CVPR_2021_paper.html. C.1

[95] NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841, 2024. URL https://www.nature.com/articles/s41586-024-07335-x. 5.2.1, 5.3.3, C.4.2

[96] OpenAI. Mmmlu dataset, 2024. URL https://huggingface.co/datasets/openai/MMMLU. Accessed: 2024-10-01. 5.3.3, C.4.2

[97] OpenAI. Hello gpt4-o. https://openai.com/index/hello-gpt-4o/, 2024. URL https://openai.com/index/hello-gpt-4o/. 4.1, 4.3.1, 4.6, 5.1

[98] Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents. *Conference On Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=NPAQ6FKSmK.

3.1, 3.4.4, A.1

[99] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023. 3.1, A.1

[100] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.196. C.4.1

[101] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.196. 5.3.2

[102] Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. mmt5: Modular multilingual pre-training solves source language hallucinations. *ArXiv preprint*, abs/2305.14224, 2023. URL https://arxiv.org/abs/2305.14224. C.5

[103] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023. A.1

[104] Qwen Team. Qvq: To see the world with wisdom, December 2024. URL https://qwenlm.github.io/blog/qvq-72b-preview/. 4.1, 4.3.1, 4.6

[105] Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/. 4.1, 4.3.1

[106] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/. 4.1

[107] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5.4.1

[108] Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 66127–66137, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d08b6801f24dda81199079a3371d77f9-Paper-Datasets_and_Benchmarks.pdf. 5.1

[109] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten,

Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d08b6801f24dda81199079a3371d77f9-Paper-Datasets_and_Benchmarks.pdf. 5.1

[110] Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. Palo: A large multilingual multimodal language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*, 2025. URL https://arxiv.org/abs/2402.14818. 5.1, 5.3.2, 5.4.1, C.1

[111] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy P Lillicrap. Androidinthewild: A large-scale dataset for android device control. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=j4b3l5kOil. 3.1

[112] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1410. C.4.1

[113] Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. SciFIBench: Benchmarking large multimodal models for scientific figure interpretation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=HcLFNuQwy5. 4.1

[114] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *ArXiv preprint*, abs/2406.05967, 2024. URL https://arxiv.org/abs/2406.05967. 5.3.2

[115] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms, 2024. URL https://arxiv.org/abs/2407.18416. 2.1.2

[116] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. URL https://arxiv.org/abs/2210.08402. 5.1, 5.2.2

[117] Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-unix2: A unified cross-lingual cross-modal framework for understanding and generation. *ArXiv preprint*, abs/2211.04861, 2022. URL https://arxiv.org/abs/2211.04861. C.1

[118] Haiyang Shen, Yue Li, Desong Meng, Dongqi Cai, Sheng Qi, Li Zhang, Mengwei Xu,

and Yun Ma. Shortcutsbench: A large-scale real-world benchmark for api-based agents, 2024. URL https://arxiv.org/abs/2407.00132. A.1

[119] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2210.03057. 5.3.3, C.4.2

[120] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of bits: An open-domain platform for web-based agents. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3135–3144. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shi17a.html. 2.1.1, 3.7

[121] Paloma Sodhi, SRK Branavan, Yoav Artzi, and Ryan McDonald. Step: Stacked llm policies for web actions. In *First Conference on Language Modeling*, 2024. 3.4.4

[122] Loïc Sokoudjou Sonagu and Yoann Sola. Docvqa dataset, 2024. URL https://huggingface.co/datasets/cmarkea/doc-vqa. 5.2.3

[123] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7584–7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.409. URL https://aclanthology.org/2024.acl-long.409. 2.1.2

[124] Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. GlobalBench: A benchmark for global progress in natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proc. of EMNLP*, pages 14157–14171, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.875. 5.1

[125] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *ArXiv preprint*, abs/2405.11985, 2024. URL https://arxiv.org/abs/2405.11985. 5.2.3

[126] Xunzhu Tang, Kisub Kim, Yewei Song, Cedric Lothritz, Bei Li, Saad Ezzini, Haoye Tian, Jacques Klein, and Tegawende F. Bissyande. Codeagent: Autonomous communicative agents for code review, 2024. URL https://arxiv.org/abs/2402.02172. A.1

[127] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805, 2023. URL https://arxiv.org/abs/2312.11805. 4.3.1, 4.6, 5.1

[128] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. 5.2.1

[129] Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proc. of EMNLP*, pages 715–729, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.45. 5.3.2, C.4.1

[130] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *ArXiv preprint*, abs/2406.16860, 2024. URL https://arxiv.org/abs/2406.16860. 4.3.1, 4.6, 5.2.1, 5.4.1, C.1

[131] Toshi456. Llava-jp-instruct-108k dataset, 2023. Accessed: 2024-10-01. 5.2.3

[132] Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *ACL*, 2024. 3.1

[133] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. B.1.2

[134] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804. 4.6

[135] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*, 2024. 3.4.4, 3.4.4, A.1

[136] Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. Opendevin: An open platform for ai software developers as generalist agents, 2024. URL https://arxiv.org/abs/2407.16741. A.1

[137] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Opendevin: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 2024. 3.4.4

[138] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. What are tools anyway? a survey from the language model perspective. *arXiv preprint arXiv:2403.15452*, 2024. A.1

[139] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024. 3.4.4, 3.7, 3.7

[140] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. 3.1

[141] Nancy Xu, Sam Masling, Michael Du, Giovanni Campagna, Larry Heck, James Landay, and Monica Lam. Grounding open-domain instructions to automate web support tasks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1022–1032, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.80. URL https://aclanthology.org/2021.naacl-main.80. 3.1

[142] Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.641. URL https://aclanthology.org/2023.acl-long.641. C.1

[143] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.41. C.5

[144] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671, 2024. URL https://arxiv.org/abs/2407.10671. 4.3.1, 5.4.1

[145] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022. 3.7, A.1

[146] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents, 2023. URL https://arxiv.org/abs/2207.01206. 2.1.2

[147] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. *ArXiv preprint*, abs/2307.10928, 2023. URL https://arxiv.org/abs/2307.10928. 5.3.2

[148] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proc. of ACL*, pages 417–421,

Vancouver, Canada, 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-2066. 5.2.3

[149] Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui: Grounded mobile ui understanding with multimodal llms, 2024. URL https://arxiv.org/abs/2404.05719. 3.1

[150] Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.273. 5.1, 5.2.2

[151] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Ren Kan, Dongsheng Li, and Deqing Yang. Easytool: Enhancing llm-based agents with concise tool instruction, 2024. URL https://arxiv.org/abs/2401.06201. A.1

[152] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 4.1, 4.2.2, 4.6

[153] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. URL https://arxiv.org/abs/2311.16502. 4.1, 4.6, 5.1, 5.3.2, C.5

[154] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *ArXiv preprint*, abs/2409.02813, 2024. URL https://arxiv.org/abs/2409.02813. 5.1

[155] Xiang Yue, Yueqi Song, Akari Asai, Simran Khanuja, Anjali Kantharuban, Seungone Kim, Jean de Dieu Nyandwi, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=a3g2l4yEys. 4.1, 4.3.1, 4.6

[156] Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *ArXiv preprint*, abs/2206.00621, 2022. URL https://arxiv.org/abs/2206.00621. C.1

[157] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users, 2023. 3.1

[158] Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. Codeagent: Enhancing code genera-

tion with tool-integrated agent systems for real-world repo-level coding challenges, 2024. URL https://arxiv.org/abs/2401.07339. A.1

[159] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505, 2023. URL https://arxiv.org/abs/2306.05179. 5.3.2, C.4.1

[160] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=piecKJ2DlB. 3.1

[161] Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025. 3.7

[162] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023. URL https://arxiv.org/abs/2306.05685. 5.1, 5.3.2, 5.4.1

[163] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. *ArXiv preprint*, abs/2402.14658, 2024. URL https://arxiv.org/abs/2402.14658. 5.2.1

[164] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx. 2.1.1, 2.1.2, 3.1, 3.1, 3.4.1, 3.4.4, A.1

[165] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27. C.1