M3SciQA: A Multi-Modal Multi-Document Scientific QA Benchmark for Evaluating Foundation Models

Chuhan Li [Y] * **Ziyao Shangguan** [Y] * **Yilun Zhao** [Y] **Deyuan Li** [Y]
Yixin Liu [Y] **Arman Cohan** [Y] [A]
[Y] Yale University [A] Allen Institute for AI
{chuhan.li.cl2575, ziyao.shangguan}@yale.edu

https://github.com/yale-nlp/M3SciQA

Abstract

Existing evaluation benchmarks for foundation models in understanding scientific literature predominantly focus on single-document, text-only tasks. Such benchmarks often do *not* adequately represent the complexity of research workflows, which typically also involve interpreting non-textual data, such as figures and tables, and gathering information across multiple documents and related literature. To address this gap, we introduce M3SciQA, a multi-modal, multi-document scientific question answering benchmark designed for a more comprehensive evaluation of foundation models. M3SciQA consists of 1,452 expert-annotated questions spanning 70 natural language processing paper clusters, where each cluster represents a primary paper along with all its cited documents, mirroring the workflow of comprehending a single paper by requiring *multi-modal* and *multi-document* data. With M3SciQA, we conduct a comprehensive evaluation of 18 frontier foundation models. Our results indicate that current foundation models still significantly underperform compared to human experts in multi-modal information retrieval and in reasoning across multiple scientific documents. Additionally, we explore the implications of these findings for the future advancement of applying foundation models in multi-modal scientific literature analysis.

1 Introduction

In scientific research, the findings presented in a paper often serve as a foundation for further investigation. When studying research papers, researchers typically explore related and cited scholarly works to acquire additional context and insights. Simultaneously, research papers are inherently multi-modal, presenting additional and often important insights in the form of figures and tables. Such properties can pose challenges for AI systems in accurately interpreting and integrating diverse data formats across multiple research papers.

Recent studies have showcased foundation models' remarkable performance across a variety of tasks in scientific literature understanding, including summarization (Goyal et al., 2023; Liu et al., 2023c), document-based question answering (Newman et al., 2023; Zhao et al., 2024; Xu et al., 2024), and scientific figure question answering (Masry et al., 2022; Yue et al., 2023; Lu et al., 2024b). However, current investigations are mostly confined to a *single-document* or *text-only* setting, ignoring the *multi-modal* and *multi-document* nature of scientific research, where insights are often derived from interpreting interconnected texts, figures, and tables across multiple scholarly works.

To address this gap, we introduce M3SciQA, a **M**ulti-**M**odal, **M**ulti-document **Sci**entific **Q**uestion **A**nswering benchmark. This benchmark contains 1,452 expert-annotated questions spanning 70 natural language processing (NLP) paper clusters, encompassing 3,066 papers. Each paper cluster comprises of an anchor paper and all its cited papers. Inspired by the common workflow of comparative analysis in scientific research (as illustrated in Figure 1), our benchmark simulates a process in which a finding, derived from a *scientific image* in the anchor paper, prompts further investigation into a specific referenced paper. This simulation enriches the benchmark by requiring the models to engage in *cross-referencing* among related documents, setting a new testbed for evaluating foundation models in scientific documents understanding and reasoning (Section 2.1).

We evaluate a wide spectrum of *open-source* and *proprietary* large language models (LLMs) and large multi-modal models (LMMs). Our experimental results reveal significant limitations in both open-source and proprietary LMMs, particu-

*Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15419–15446
November 12-16, 2024

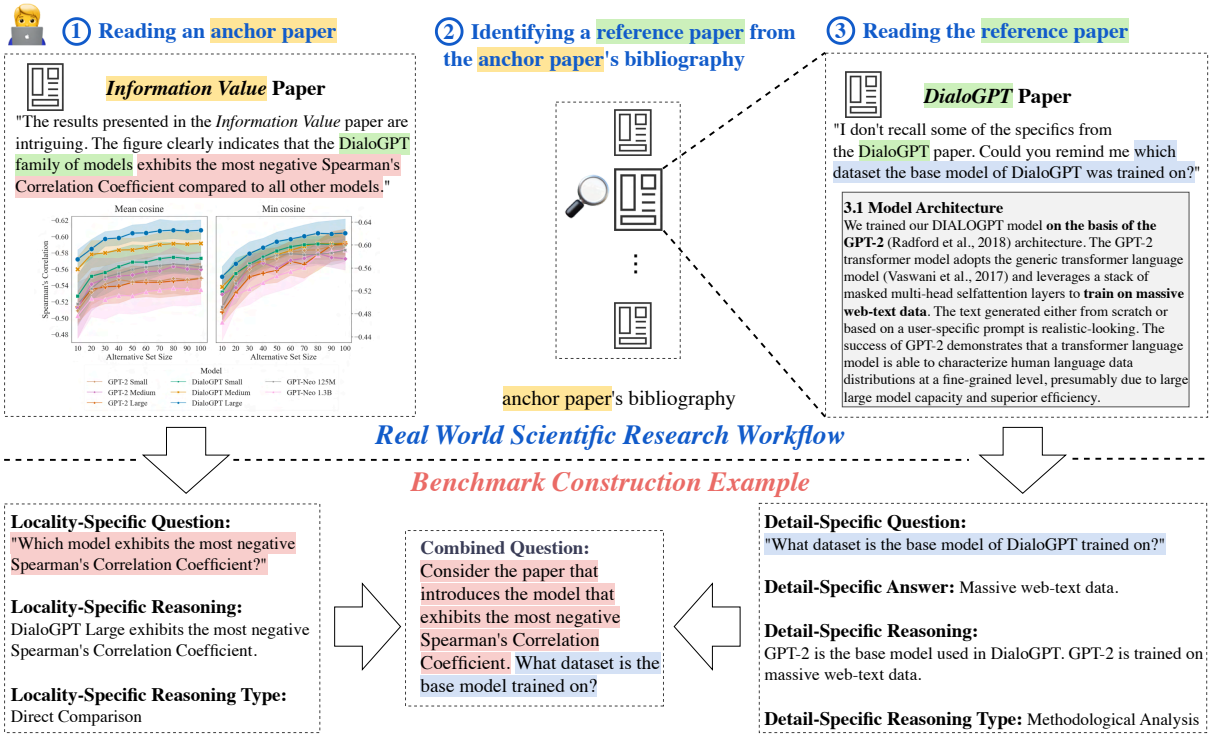


Figure 1: (**Top**) The common workflow of comparative analysis in scientific research, particularly when a result, such as a figure/table in the Information Value paper (anchor paper) (Giulianelli et al., 2023), prompts further examination of related research, such as details from DialoGPT (reference paper) (Zhang et al., 2020b). **(Bottom)** A demonstration of the workflow for constructing a locality-specific question, detail-specific question, and combined question.

larly in their ability to translate and interpret scientific images and perform effective re-ranking based on these images, with the best-performing model, GPT-4o, achieving a Mean Reciprocal Rank (MRR) of 0.488 compared to a human expert score of 0.796, corresponding to a performance gap of 0.308.

Similarly, we observe that both open-source and proprietary LLMs struggle with long-range retrieval tasks, specifically with extracting and analyzing information from one or more academic documents. Here, the best-performing model, Command R+, achieves an accuracy score of 33.25 compared to an human expert accuracy score of 76.56.[1] These findings underscore the challenges that current models face in handling complex, *multi-modal*, *multi-document*, and domain-specific information.

Our main contributions are as follows:

- We introduce M3SCIQA, a comprehensive benchmark designed to evaluate the multi-modal reasoning abilities in interpreting multiple scientific documents.
- We conduct an extensive evaluation covering a wide range of LMMs and LLMs. Our experimental results reveal a noticeable performance gap between foundation models and human experts.
- To better understand the limitations of current foundation models, we conduct a detailed analysis of scientific figure information retrieval, long-context re-ranking, and long-range retrieval, providing valuable insights for future advancements of foundation models.

2 The M3SCIQA Benchmark

2.1 Overview of M3SCIQA

Our objective is to develop a challenging yet realistic QA benchmark that necessitates both *multi-modal* and *multi-document* reasoning over scientific papers. An overview of a benchmark question construction pipeline is shown in Figure 2. From the 70 curated anchor papers, expert annotators are tasked with composing questions from the figures or tables, defined as "locality-specific questions." As detailed in Table 6 in Appendix B.2, these questions are divided into four types of reasoning categories: *comparisons*, *data extraction*, *locations*,

[1]Human expert performance is assessed in the setting where the correct reference paper is known.

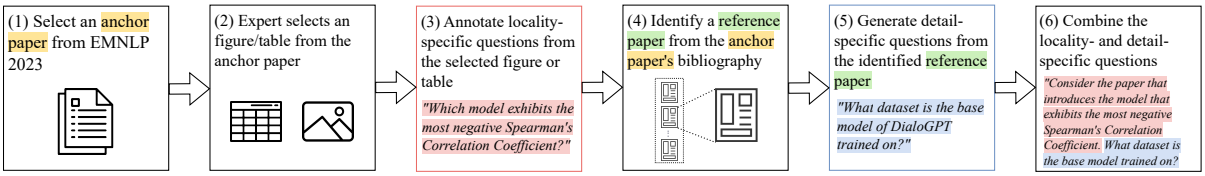


Figure 2: An overview of M3SCIQA question construction pipeline.

and *visual understanding*. By answering a locality-specific question, expert annotators can pinpoint a reference paper that provides further elaboration on the topic from among all the publications cited by the anchor paper. Subsequently, GPT-4[2] is employed to generate questions, defined as "detail-specific questions," from the identified reference paper. GPT-4 is utilized again to rephrase and combine each locality-specific question with each of the related detail-specific questions to form a comprehensive question that embodies both multi-modal and multi-document reasoning. Finally, expert annotators are then tasked with verifying the quality of these GPT-4-assisted questions. Key statistics of the benchmark are listed in Table 1; distributions of reasoning types across locality-specific and detail questions are illustrated in Figure 3.

2.2 Benchmark Construction Principles

To bridge the notable gap in current benchmarks that separately assess either multi-modal or multi-document reasoning, our benchmark, M3SCIQA, aims to encompass both elements in a single QA pair. Therefore, our benchmark construction pipeline adheres to the following guidelines: (1) it includes diverse modalities, such as texts, figures (including line plots, bar plots, scatter plots, etc.), and tables (stored as images to preserve format integrity rather than as plain texts); (2) it necessitates connecting information across multiple documents; (3) it spans a variety of reasoning types, including four types of locality-specific reasoning and five types of detail-specific reasoning; (4) it poses significant challenges in both *multi-modal* comprehension and *multi-document* information retrieval; and (5) it generates realistic QA pairs that reflect the workflows common in scientific literature analysis.

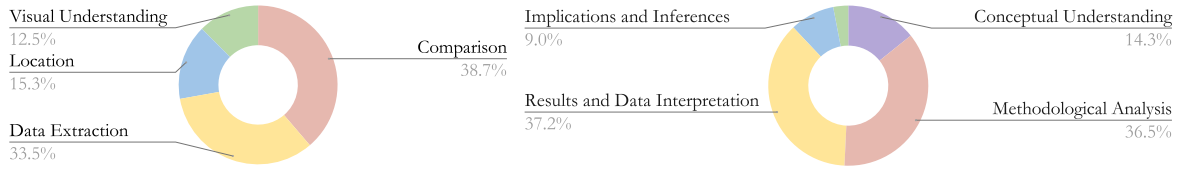
2.3 Benchmark Construction

Expert Annotators. We recruit three computer science graduate students with expertise in the field of NLP, each of whom has authored at least one peer-reviewed publication in top-tier NLP conferences. Their responsibilities include: (1) curating anchor papers from a pool of candidates and composing locality-specific questions; (2) reviewing and verifying the reasoning types of detail-specific questions; (3) resolving discrepancies between answers generated from the two rounds of detail-specific answer generation; and (4) checking consistency, clarity, and redundancy in the combined questions. Further details on annotations are provided in Appendix C.

Anchor Papers. To mitigate the risk of data contamination, where models might rely on pre-trained knowledge to answer the locality questions rather than analyzing the provided scientific images, we curate anchor papers from a recent NLP conference, EMNLP 2023. Among the 1,047 papers accepted by EMNLP 2023, we select 441 papers that were released on ArXiv after October 1st, 2023 as candidate anchor papers.

Locality-Specific QA from Anchor Papers. Two of the expert annotators curate 70 papers by manually examining 441 candidate anchor papers collected. Subsequently, they select 21 figures and 62 tables from the 70 papers to compose 300 locality-specific questions and answers that conform to four visual reasoning types. The ground truth answer to each locality-specific question is the single reference paper to which the locality-specific question directly refers. This facilitates a transition from an anchor paper to a reference paper that elaborates on the subject. The third annotator is responsible for validating the accuracy and relevance of these questions and answers. 371 papers are excluded in this process because they either lack figures or tables that can be analyzed by one of the reasoning types, or transition to a cited paper that is not available on ArXiv. Furthermore, due to the occurrence of identical answers among some locality-specific questions, these 300 questions correspond to only 107 reference papers.

[2] `gpt-4-0125-preview`



(a) Locality-Specific Question

(b) Detail-Specific Question

Figure 3: Distribution of reasoning types of *locality-specific* and *detail-specific* questions in M3SCIQA.

Detail-Specific QA from Reference Papers. By requiring that the 107 reference papers be available on ArXiv, we ensure access to their complete content. This enables us to utilize GPT-4 to generate open-ended, detail-specific questions from the papers. For each reference paper, we create five questions each corresponding to a reasoning type illustrated in Table 7 in B.3. These questions are designed to be answerable in a text-only setting, without the need for visual reasoning or OCR. Considering the possibility that GPT-4 may incorrectly categorize the questions, expert annotators manually examine the reasoning types associated with the questions and reassign when necessary. This process yields a total of 519 detail-specific questions after filtering out duplicates, overly complex questions, questions that do not require specific insights from the paper (*e.g.,* "What is the mathematical expression for calculating the F-1 score?"), and questions that do not belong to any of the five predefined reasoning types. To establish a *gold answer* for each question, we generate answers in two rounds. In the first round, answers are generated concurrently with the questions. In the second round, the model is prompted to answer the questions using the reference paper as context. We employ GPT-4 to determine if the answers from both rounds are consistent. If any discrepancy is identified, expert annotators are enlisted to verify and finalize the answers. Further details can be found in Appendix C.5.

Combined Questions. We utilize GPT-4 to compile the final questions for the benchmark by combining each locality-specific question with its corresponding detail-specific questions. After the combination, expert annotators are tasked with verifying the question validity and rephrasing the GPT-4-assisted combined question when necessary. Overall, we form 1,452 combined questions, each associated with a specific figure or table. Expert annotators then review these combined questions to ensure that each locality-specific question and its corresponding detail-specific questions are logically connected and relevant. They also check for clarity, consistency, and redundancy to maintain the quality and difficulty of the benchmark.

Statistics	***Avg.* Value**
Locality-Specific Question Length *(tokens)*	12.9
Detail-Specific Question Length *(tokens)*	25.95
Combined Question Length *(tokens)*	41.3
Answer Length *(tokens)*	24.9
# Cluster	70
# Anchor Paper per Cluster	1
# Reference Paper per Cluster	42.8
Paper Length *(tokens)*	11.8K
Validation Set Size	452
Test Set Size	1000

Table 1: Key statistics of the M3SCIQA benchmark.

3 Experiments

We evaluate 18 prominent foundation models, including both *open-source* and *proprietary* LMMs and LLMs. For each model, we select the most recent, largest, and best-performing checkpoint as of April 15th, 2024. The evaluation of the M3SCIQA benchmark is structured into two distinct stages: *locality-specific* evaluation and *detail-specific* evaluation.

3.1 Locality-Specific Evaluation

Task Formulation. The locality-specific evaluation with LMMs is defined as follows: Given a locality-specific question Q_{loc}, its correspondent scientific image I, and a list of reference papers $D = \{d_1, d_2, \cdots, d_n\}$, the objective is to determine a ranking of these papers based on their relevance to the question and the image. This ranking is represented by $R = \{r_1, r_2, \cdots, r_n\}$, where r_i denotes the ranking of the paper d_i for each index $i \in \{1, 2, \cdots, n\}$. We input Q_{loc}, I and D into each LMM, denoted by f_{LMM}, and instruct it to generate a ranking R of D based on their relevance

Model	All	Modality		Reasoning Type			
		Table	Figure	COM	DE	LOC	VU
Expert Performance	0.796	0.678	0.765	0.751	0.872	0.711	0.732
Random	0.126	0.134	0.106	0.134	0.130	0.110	0.111
Simple Baselines							
`text-embedding-3-large`	0.297	0.321	0.239	0.267	0.323	0.384	0.218
`text-embedding-3-small`	0.217	0.223	0.205	0.221	0.223	0.267	0.138
`text-embedding-ada-002`	0.180	0.185	0.168	0.200	0.171	0.224	0.096
Contriever	0.184	0.165	0.229	0.196	0.144	0.274	0.142
BM25	0.127	0.138	0.098	0.118	0.128	0.160	0.110
Open-Source Large Multi-modal Models (LMMs)							
InternVL-Chat-V1.1	0.144	0.168	0.084	0.136	0.153	0.170	0.109
Yi-VL-34B	0.091	0.105	0.057	0.101	0.088	0.080	0.086
Qwen-VL-Plus	0.089	0.065	0.131	0.077	0.053	0.148	0.136
LLaVA-1.6	0.056	0.079	0.000	0.088	0.044	0.052	0.000
DeepSeek-VL	0.079	0.075	0.087	0.064	0.081	0.109	0.070
Proprietary Large Multi-modal Models (LMMs)							
GPT-4o	**0.500**	**0.520**	**0.454**	**0.443**	**0.565**	**0.570**	0.418
GPT-4V(ision)	0.400	0.440	0.309	0.383	0.407	0.523	0.288
Claude-3-Sonnet	0.374	0.385	0.369	0.357	0.363	0.395	**0.422**
Claude-3-Opus	0.316	0.256	0.343	0.320	0.362	0.301	0.204
Gemini-Pro-Vision-1.0	0.197	0.217	0.188	0.196	0.160	0.284	0.195
Claude-3-Haiku	0.188	0.189	0.188	0.194	0.201	0.130	0.208

Table 2: Mean reciprocal rank (MRR) on the *test* set of M3SCIQA. The best-performing model in each category is **in-bold**, and the second best is underlined. Reasoning types: **COM**: comparison, **DE**: data extraction, **LOC**: location, **VU**: visual understanding.

Model	Context Window	All	CU	II	RDI	MA	CA
Expert Performance		76.50	72.32	71.11	83.15	76.84	79.17
Open-Source Large Language Models (LLMs)							
†Command R+	128,000	**33.25**	**40.00**	22.73	**33.33**	**37.91**	**39.53**
Llama-3-70B	8192	31.30	31.35	**35.23**	22.84	32.49	35.19
Mistral-7B	32,768	20.45	17.10	24.09	8.89	25.81	26.72
PaLM-2	36,864	23.55	20.73	26.42	16.35	27.65	26.72
DBRX	32,768	19.05	18.13	19.43	13.94	21.30	22.63
Gemma-7B	8,192	12.25	8.89	15.15	1.39	13.95	20.93
Proprietary Large Language Models (LLMs)							
†GPT-3.5	16,385	29.00	22.22	33.33	19.44	32.56	37.21
†GPT-4	128,000	28.50	31.11	21.21	23.61	32.56	31.40
†Claude-3-Haiku	200,000	28.25	28.89	30.88	12.50	29.07	38.10
†Claude-3-Sonnet	200,000	26.50	25.56	21.21	19.44	25.58	38.37
†Claude-3-Opus	200,000	25.00	26.67	18.18	20.83	26.74	30.23
†Gemini-Pro-1.0	30,720	21.75	18.89	19.70	18.06	22.09	29.07

Table 3: LLM-based accuracy score on the *test* set of M3SCIQA in *retrieval* setting from GPT-4o's ranking. The best-performing model in each category is **in-bold**, and the second best is underlined. Human expert performance is assessed in an oracle setting, where the correct reference paper is pre-identified. Reasoning types: **CU**: conceptual understanding, **II**: implications and Inferences, **RDI**: results and data interpretation, **MA**: methodological analysis, **CA**: critical analysis. †: Due to budget constraints, we randomly sampled 200 instances from the *test* set for evaluation.

to Q_{loc} and I:

$$R = f_{LMM}(Q_{loc}, I, d_1, d_2, \cdots, d_n)$$

For comparative analysis, simple baselines presented in Table 2 are also assessed for the ranking task. Other than BM25, these baselines employ cosine similarity between query and document em-

beddings to rank documents. Each query combines the locality-specific question Q_{loc} and its image caption C generated by GPT-4o with one of the documents, represented by its title and abstract. Given a locality-specific question Q_{loc}, its correspondent scientific image I, a list of reference papers $D = \{d_1, d_2, \cdots, d_n\}$, an embedding model *Embed*, and a cosine similarity function *sim*, the ranking process is defined as below:

$$
\begin{aligned}
&C = \textit{GPT-4o}(I) \\
&q = \textit{Embed}(\textit{concat}(Q_{loc}, C)) \\
&\forall d_i \in D, h_i = \textit{Embed}(d_i) \\
&R = \textit{sort}(sim(q, h_1), \cdots, sim(q, h_n))
\end{aligned}
$$

Evaluation Protocol. At the locality-specific evaluation stage, we assess LMMs' ability to accurately retrieve and rank the correct reference paper from a complete list of reference papers. Performance is measured using an established information retrieval metric, Mean Reciprocal Rank (MRR), which effectively gauges a model's ability to identify and prioritize the most relevant reference paper. Additionally, we calculate Recall@k and NDCG@k to further analyze LMMs' retrieval effectiveness, with results detailed in Table 8 and Table 9 in Appendix E.

Experiment Setup. This stage involves five *open-source* LMMs, including *open-source* models, such as LLaVA 1.6 (Liu et al., 2023a), InternVL-Chat-1.1V (Chen et al., 2024), Yi-VL-34B (AI et al., 2024), DeepSeek-VL (Lu et al., 2024a), and Qwen-VL-Plus (Bai et al., 2023); six *proprietary* LMMs, including GPT-4V(ision) (OpenAI, 2024a), GPT-4o (OpenAI, 2024b), Claude 3 Haiku (Anthropic, 2024), Claude 3 Sonnet (Anthropic, 2024), Claude 3 Opus (Anthropic, 2024), and Gemini Vision Pro 1.0 (Team, 2023); and five *simple baselines*, including BM25, Contriever (Izacard et al., 2021), and OpenAI Embeddings[3] (`Large`, `Small`, and `Ada`).

3.2 Detail-Specific Evaluation

Task Formulation. The detail-specific evaluation is defined as follows: Given a combined question Q_{comb} and a ranking R of the reference papers obtained in the *locality-specific* evaluation stage, the objective is to answer the question based on the top k ranked paper in R, denoted by $\textit{Top}_k(R) = \{R[1], R[2], \cdots, R[k]\}$. Since *combined* questions contain elements from both *locality-specific* and *detail-specific* questions, we instruct LLMs to solely concentrate on the *detail-specific* aspect of Q_{comb}. The prompts used for this instruction are detailed in Table 15 in Appendix F.3. Accordingly, we input Q_{comb} and $\textit{Top}_k(R)$ into LLMs, denoted by f_{LLM}, and instruct LLMs to answer Q_{comb} based on the textual content in top k ranked papers:

$$
\textit{Ans} = f_{LLM}(Q_{comb}, R[1], R[2], \cdots, R[k])
$$

Evaluation Protocol. At the detail-specific evaluation stage, we assesses how LLMs perform on detail-specific questions using the top three ranked papers identified from the locality-specific evaluation stage as context. Specifically, these papers are ranked by GPT-4o, which is highlighted as the most effective retrieval model in Table 2. GPT-4o achieves an MRR of 0.488, suggesting that the correct reference paper typically appears in the 2.1-th position, placing it within the top three ranked papers on average. Given that both detail-specific question and answer generation utilize plain text extracted from TeX files, we employ the same parsed TeX files as input for LLMs to solve the text-only, detail-specific questions.

Generative Response Evaluation. Following effectiveness of LLMs in evaluating the quality of short AI-generated responses (Wang et al., 2023; Lu et al., 2024b; Dubois et al., 2024; Wang et al., 2024), we utilize a strong LLM-evaluator (GPT-4) to evaluate the quality of responses generated in the detail-specific evaluation stage. Specifically, the LLM-evaluator rates answers generated against the *gold answers* using a scoring scale of 0, 0.5, and 1. To more closely align our scoring scale with expert assessments, we compute *Cohen's Kappa* (McHugh, 2012) to assess the agreement between the LLM-evaluator and expert annotators. This comparison is conducted for both the 0-0.5-1 and the 1-2-3-4-5 scales, with prompts utilized for evaluation provided in Table 16 in Appendix F.1. Expert annotators are tasked with rating 200 responses from four different LLMs (Command R+, GPT-4, Mistral, and Gemma) using both scales. Our calculations reveal a Cohen's Kappa value of 0.520 for the 0-0.5-1 scale and 0.444 for the 1-2-3-4-5 scale. These results demonstrate greater consistency with expert evaluations when using the 0-0.5-1 scale. Further details and comparative results

[3] https://platform.openai.com/docs/guides/embeddings

are presented in Appendix F.1. Thus, we adopt the 0-0.5-1 scoring scale for our evaluations. Additionally, we employ established metrics such as ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020a), and AutoACU (Liu et al., 2023b) to further gauge the quality of the generated responses. Detailed results are provided in Table 10, 11, 12 in Appendix E.

Experiment Setup. This stage involves six *open-source* Text-Only LLMs, including Mistral-7B (Jiang et al., 2024), Llama-3-70B (Meta, 2024), DBRX (Databricks, 2024), PaLM-2 (Anil et al., 2023), Gemma (Team et al., 2024), and Command R+ (CohereForAI, 2024); and six *proprietary* LLMs, including GPT-3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2024a), Claude 3 Haiku (Anthropic, 2024), Claude 3 Sonnet (Anthropic, 2024), Claude 3 Opus (Anthropic, 2024), and Gemini-Pro-1.0 (Team, 2023).

3.3 Main Results

Table 2 and Table 3 present our main results for both *open-source* and *proprietary* LMMs and LLMs on the validation and test set of M3SciQA, focusing on *locality-specific* and *detail-specific* questions, respectively. We summarize our key findings as follows:

Finding 1: Challenges in Visual Reasoning and Paper Ranking with M3SciQA. Table 4 provides a breakdown of GPT-4o's performance in answering the locality-specific questions, categorized by both reasoning and ranking outcomes. Despite being the overall best-performing retriever, GPT-4o still struggles with the dual challenges: it fails to correctly interpret 42.4% of the scientific images; even when it does produce correct visual reasoning, it falls short in ranking the associated paper within the top three choices. Notably, one interesting error pattern is the scenario "✗ reasoning ✓ranking@top3," which accounts for 19.7% of the cases for GPT-4o. While this type of error occurs in both open-source and proprietary LMMs, it is more prevalent in the former. Example error analyses are presented in Figure 4, offering a more granular view of these patterns and specific instances where the model underperforms.

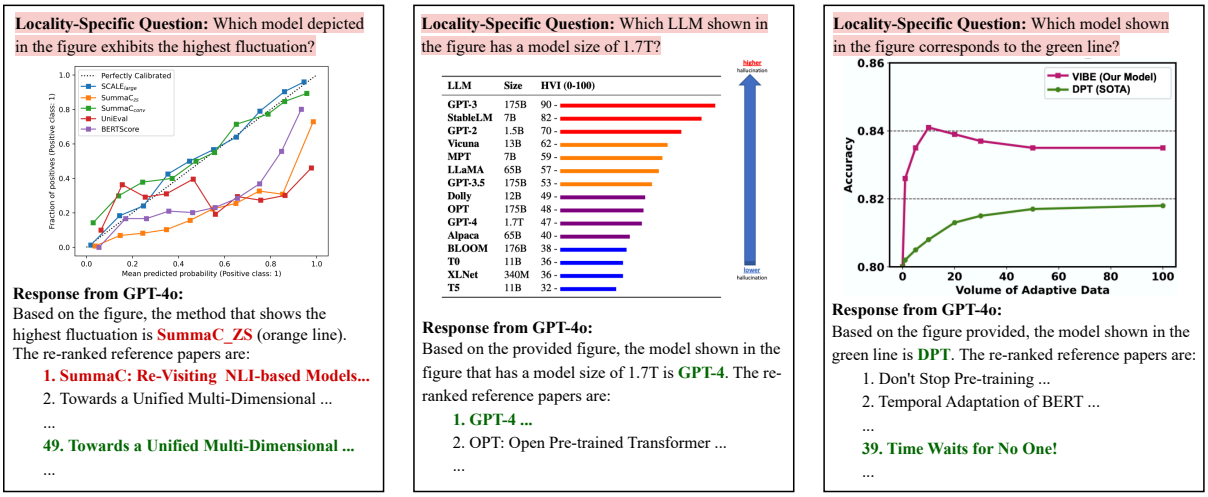
Finding 2: Inherent Limitations of *Open-Source* LMMs in Long-Range Ranking Task. The performance of open-source LMMs in long-range ranking tasks is significantly hindered by their fundamental limitations. We identify three primary challenges: (1) *Limited Context Window*, which necessitates division of large paper clusters into smaller segments, complicating the ranking process and potentially omitting relevant reference papers; (2) *Hallucinations*, characterized by the erroneous generation and prioritization of irrelevant ArXiv webpage URLs, professional NLP terms, repetitive paper IDs, and random numerical values; (3) *Formatting Issues*, where models disregard specified format and list papers in plain text, complicating the integration of results across rankings from segmented paper clusters. These challenges significantly impede the models' ability to provide a comprehensive evaluation of their visual reasoning capabilities, suggesting the need for improvements in their basic functionality to handle more complex reasoning and ranking tasks. A detailed evaluation of open-source LMMs is presented in Appendix G.

Reasoning Correctness	Ranking@Top3	Percentage
✓	✓	33.0%
✓	✗	24.7%
✗	✓	19.7%
✗	✗	22.7%

Table 4: Performance distribution for GPT-4o on locality-specific questions, categorized by **Reasoning Correctness** and **Ranking@Top3**.

Finding 3: Precision-Recall Balance. We evaluate LLMs in retrieval settings using the top k ranked papers from the *locality-specific* evaluation performed by GPT-4o for the values $k \in \{1, 2, 3, 4, 5\}$. As shown in Figure 5, performance generally increases from $k = 1$ to $k = 3$, aligning with an MRR score of approximately 0.488, which places the correct reference paper in the 2.1-th position on average. Beyond this point, as more papers are considered, more noise is introduced. The general decline in performance after $k = 3$ demonstrates models' limitations in retrieval tasks when given more irrelevant information.

Finding 4: Challenges in Instruction Compliance for LLMs in Retrieval Task. Our evaluation of four models in both a *title-only* setting, where only the title of the reference paper is provided, and a *retrieval* setting, with the top three ranked papers by GPT-4o, highlights variations in instruction compliance. Models are instructed to answer "I don't know" if a definitive answer can-



(a) **Incorrect Reasoning** and **Failed Ranking@Top3**

(b) **Correct Reasoning** and **Successful Ranking@Top3**

(c) **Correct Reasoning** and **Failed Ranking@Top3**

Figure 4: Three examples from GPT-4o in answering locality-specific questions.

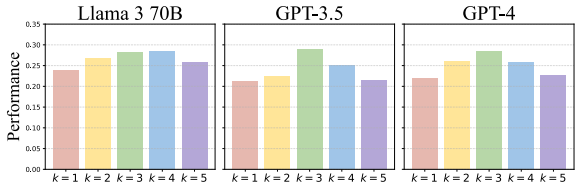


Figure 5: Performance scores of Mistral, Llama 3 70B, GPT-3.5, and GPT-4 in different *retrieval* settings.

not be derived from the given information. This directive tests the models' adherence to instructions, since the task is infeasible with the titles alone and compliant models should exhibit minimal performance. Transition to the *retrieval* setting should reveal a significant increase for the models, as observed with GPT-4 in Table 5. Additionally, employing a LLM-based evaluator to assess generative response overlooks models' confidence levels. Less compliant models, relying on pre-trained knowledge, often produce tangentially relevant responses rather than the instructed "I don't know," leading to disproportionately higher evaluations from the LLM-based evaluator.

Models	GPT-4	GPT-3.5	Llama 3 70B	Mistral
title-only	7.50	13.50	19.75	22.00
retrieval	28.50 (+21.00)	29.00 (+15.50)	28.25 (+8.50)	19.25 (-2.75)

Table 5: Performance of four LLMs in answering detail-specific questions in *title-only* and *retrieval* setting.

4 Related Work

Multi-Modal QA. Multi-modal QA datasets have posed visual reasoning challenges for LMMs. Initially, the focus of benchmarks (Lin et al., 2015; Mobasher et al.; Yagcioglu et al., 2018; Talmor et al., 2021; Lu et al., 2022; Chang et al., 2022; Li et al., 2023; Liu et al., 2023d; Yu et al., 2023) was on conducting QA tasks over simple images, primarily addressing questions such as understanding objects in an image and performing single-hop reasoning. Recently, more complex and nuanced benchmarks (Chen et al., 2022; Lu et al., 2024b) have emerged beyond the scope of understanding simple images to require complex mathematical reasoning over diagrams and plots. Beyond the scope of mathematical reasoning, MMMU (Yue et al., 2023) requires more complex visual reasoning in a diverse range of subjects including science, humanities, and engineering.

Document QA. Document QA is crucial in the field of NLP, focusing on extracting, synthesizing, and analyzing information from structured and unstructured documents. Early document QA benchmarks (Rajpurkar et al., 2016; Bajaj et al., 2018; Yang et al., 2018) involved short document QA, where questions were posed based on content from web pages such as those in Bing's search results or Wikipedia articles. Scientific paper QA benchmarks (Dasigi et al., 2021; Lee et al., 2023) require LLMs to conduct multi-hop reasoning and long-context information processing. However, a notable gap exists in the integration of Multi-modal

QA with Document QA, particularly in the context of scientific research, where it encompasses a blend of textual and visual data alongside complex textual information. M3SCIQA, bridging this gap, is a benchmark for evaluating foundation models' abilities in both multi-modal and multi-document reasoning.

5 Conclusion

In this paper, we introduced M3SCIQA, a novel multi-modal, multi-document scientific QA benchmark for evaluating foundation models. M3SCIQA effectively simulates real-world scientific workflow, thereby providing a more realistic and challenging environment for evaluating the capabilities of both LMMs and LLMs. Our evaluations of various *open-source* and *proprietary* models reveal a significant gap between the performance of current foundation models and human experts. This disparity reveals the current limitations of such models in interpreting *multi-modal*, *multi-document* scientific data and illustrates the need for further improvement in the domain of complex scientific reasoning. In our analysis, we examined numerous areas where LLMs and LMMs fail to correctly solve questions, demonstrating that M3SCIQA is challenging in multiple facets for both *multi-modal* reasoning and *multi-document* processing.

Limitations

The evaluations presented in this study are met with certain limitations due to inherent disparities in the context window of current *open-source* and *proprietary* LLMs and LMMs. There is a significant difference in context window length between models such as GPT-4 Turbo and Claude-3, which can rank all papers in a paper cluster, and models such as InternVL-Chat-V1.1 and QwenVL, which are restricted to handling only two to eight papers in a single prompt. This discrepancy may lead to an "unfair" comparison of their capabilities. Future work could focus on standardizing or extending the context windows in LMMs to mitigate this issue.

Furthermore, as discussed in Section 3.3, prompting an LMM with a set of possible reference papers may be suboptimal due to the challenges models face in ranking a large number of papers. An alternative approach could involve assessing the relevance of each paper individually by encoding the paper into a textual embedding, then comparing it with the textual embedding with of the locality-specific question combined with the image representation of the figure. This method could potentially alleviate the challenges of requiring an LMM to sift through a large set of possible reference papers and would be an interesting area for future research.

Additionally, our approach to ranking papers for certain models, in particular BM25 and Contriever, involves using GPT-4o's textual descriptions of images rather than its direct image embedding, which might not accurately capture the nuances of scientific images. Current image embedding models such as LLaVA (Liu et al., 2023a) and CLIP (Radford et al., 2021), while proficient with natural images, are not trained on scientific images. Developing a specialized LMM trained specifically on scientific images (Li et al., 2024; Wu et al., 2024) could potentially enhance its performance in interpreting scientific plots, figures, and tables, thereby improving its potential usage in scientific applications.

Acknowledgements

This project was supported in part by Tata Sons Private Limited, Tata Consultancy Services Limited, and Titan. We are grateful for the compute support provided by Microsoft Research's AFMR program. We thank Xinyi Han, Zhongjie Wu, Amy Zhao for their help in initial stages of this project.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa

Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. *Preprint*, arXiv:2109.00590.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2022. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *Preprint*, arXiv:2105.14517.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

CohereForAI. 2024. CommandR+.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Databricks. 2024. Dbrx.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. *Preprint*, arXiv:2310.13676.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee, and Moontae Lee. 2023. QASA: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052. PMLR.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *Preprint*, arXiv:2403.00231.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023c. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *Preprint*, arXiv:2311.09184.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023d. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *Preprint*, arXiv:2203.10244.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 06/13/2024.

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. Parsvqa-caps: A benchmark for visual question answering and image captioning in persian.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212, Singapore. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Hello gpt-4o: We're announcing gpt-4o, our new flagship model that can reason across audio, vision, and text in real time. Accessed: 06/13/2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *Preprint*, arXiv:2104.06039.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski,

Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. *Preprint*, arXiv:2305.12421.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Preprint*, arXiv:2406.18521.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval. *Preprint*, arXiv:2401.13478.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. Kiwi: A dataset of knowledge-intensive writing instructions for answering research questions. *Preprint*, arXiv:2403.03866.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

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. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. 2024. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. *Preprint*, arXiv:2407.07053.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. *Preprint*, arXiv:1911.00536.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Contents

A Broader Impacts

The ability to process multi-modal, multi-document content is increasingly critical for LMMs, especially given the growing interest in scientific paper understanding and analysis. LMMs that excel in processing such content can enable both experts and non-experts to quickly grasp complex scientific information. Therefore, evaluating the capabilities of both LLMs and LMMs in this area is crucial to identify existing limitations in handling scientific documents. Although M3SCIQA is primarily focused on Natural Language Processing (NLP), the adaptability of its pipeline suggests potential for extension to other disciplines with appropriate modifications in future research.

Through M3SCIQA, we observe that current LMMs generally struggle with the visual understanding of scientific figures, a limitation also noted by recent studies (Wang et al., 2024; Zhang et al., 2024). Zhang et al. (2024) enhances Llava-1.5-7B's (Liu et al., 2024) performance on abstract image understanding (charts, tables, and graphs) by fine-tuning it with their synthetic scientific figures. This improvement suggests that current LMMs might not be extensively trained with scientific figures, which results in their unsatisfactory performance in this domain. With M3SCIQA, we aim to provide a more nuanced understanding, focusing not only on scientific figure understanding but also on long-range ranking and retrieval tasks, as detailed in Section 3.3.

B Data Collection Guidelines

B.1 Locality-Specific Reasoning Definition

Four locality-specific question reasoning types are defined in Table 6.

B.2 Locality-Specific Reasoning Examples

Four locality reasoning types examples are shown in Figure 6.

B.3 Detail-Specific Reasoning Definition and Examples

Five detail-specific question reasoning types and examples are defined in Table 7.

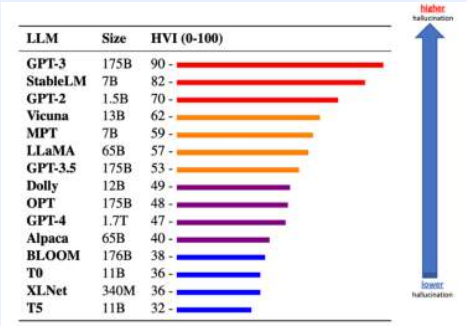
C Expert Annotation Details

C.1 Expert Annotation for Locality-Specific Questions

We employed three computer science graduate students for annotating 300 locality questions. Being provided with the full list of EMNLP 2023 papers, they were required to: (1) check that each anchor paper has ArXiv documentation; (2) find figures or tables that contain comparative information with potential reasoning types described in Table 6; (3) find the potential reference paper in the figure or table and ensure that it has ArXiv documentation; and (4) write the locality-specific question. When they choose a figure or a table, they were required to fill in the corresponding locality-specific reasoning type as well as the "direct answer" to the locality-specific question.

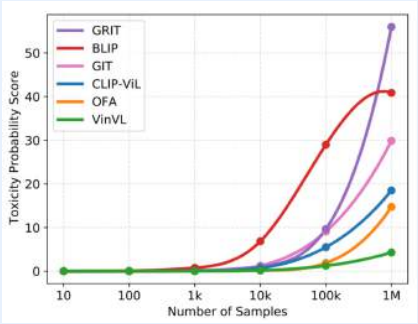
C.2 Bias Mitigation for Locality-Specific Questions Annotation

In preparation for the main annotation process, we conduct a pilot annotation stage where 20 papers where sampled. Annotators are instructed to generate three distinct questions per paper. These questions are subsequently analyzed by the authors and



Locality-Specific Question: Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?

Reasoning Type: Comparison



Locality-Specific Question: What method is demonstrated by the solid lavender line?

Reasoning Type: Visual Understanding

Method	Model	10-shot	20-shot	40-shot	80-shot	100-shot
Supervised Method	$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019)	23.101	34.718	43.138	46.182	46.449
	$\text{RoBERTa}_{\text{BASE}}$ (Liu et al., 2019)	21.073	34.157	39.226	45.735	46.942
	$\text{RoBERTa}_{\text{LARGE}}$ (Liu et al., 2019)	29.137	41.425	51.137	55.057	54.172
	PIQN (Shen et al., 2022)	21.750	22.007	28.533	29.339	38.658
	PL-Marker (Ye et al., 2022)	40.902	48.064	52.395	53.249	53.061
Few-shot Method	LSFS (Ma et al., 2022a)	**47.998**	43.269	50.595	51.420	50.366
Proposed Method	CoFiNER	44.951	**51.142**	**56.409**	**56.847**	**57.178**

Locality-Specific Question: Which model achieves a score of 21.073 in 10-shot prompting?

Reasoning Type: Data Extraction

	FinQA		TAT-QA	
Method	EA	PA	EM	F1
PromptPG (Lu et al., 2023)	53.56 ± 3e-3	24.09 ± 1e-3	51.64 ± 0.27	58.86 ± 0.27
Random	55.65 ± 3.35	29.5 ± 17.91	49.7 ± 1.38	57.31 ± 1.24
CSP	60.77 ± 0.14	43.62 ± 0.23	57.47 ± 0.19	65.21 ± 0.13
Fixed set	64.05 ± 0.22	38.15 ± 0.08	66.55 ± 0.18	73.8 ± 0.11
KATE (Liu et al., 2022a)	67.07 ± 0.04	58.65 ± 0.04	68.9 ± 0.07	75.77 ± 0.08
Diverse KATE	67.31 ± 0.24	59.54 ± 0.19	61.53 ± 0.23	68.89 ± 0.21
SEER	68.85 ± 0.04	59.78 ± 0.15	69.68 ± 0.07	76.71 ± 0.07
SEER_{gold}	69.25 ± 0.11	60.16 ± 0.14	71.32 ± 0.07	78.12 ± 0.08
SOTA fine-tuned model	71.07	68.94	73.6	81.3

Locality-Specific Question: Which method is shown in the first row of the table?

Reasoning Type: Location

Figure 6: Examples of four locality-specific reasoning categories in M3SciQA.

Locality-Specific Reasoning	Description
Comparison	It focuses on evaluating and contrasting information presented in tables, figures, or other data formats. To answer questions of this type, one must analyze and compare specific subjects or variables within the given dataset.
Data Extraction	It directly retrieves specific information from a table or figure. This approach focuses on pinpointing exact data points or details.
Location	It is centered on pinpointing spatial or positional information from a table or figure. This involves identifying either relative or absolute locations, such as the placement of items in a figure or row information in a table.
Visual Understanding	It emphasizes understanding visual information from the figure, such as colors, shapes, and marker types. This approach involves analyzing and extracting visual information.

Table 6: Definitions of four locality-specific reasoning categories in M3SCIQA.

Detail-Specific Reasoning	Description & Example
Conceptual Understanding	Evaluate knowledge of essential concepts, basic theories, and critical definitions related to the subject. ***Example: What does the hypernetwork in the proposed Hyperdecoders approach generate?***
Methodological Analysis	Examine and assess the research methodologies and experimental frameworks employed in studies, with an emphasis on their efficacy and constraints. ***Example: What potential application of the Hyperdecoder approach is suggested by its performance on long-context out-of-domain datasets in the MRQA evaluation?***
Results and Data Interpretation	Analyze statistical data, graphs, and tables, focusing on deriving significant insights and conclusions from quantitative and visual information. ***Example: In the experimental results for the GLUE benchmark using $T5_{large}$ v1.1 + LM as the underlying model, which model configuration achieved the highest average score across tasks?***
Implications and Inferences	Infer wider implications and practical uses of study outcomes, concentrating on the extensive impact and prospective significance of the results. ***Example: How does the exponentially weighted pooling method in CET ensure that every embedding receives sufficient training?***
Critical Analysis	Assess the study's reasoning, robustness of evidence, and validity of conclusions critically, with a focus on logical consistency and the support of empirical data. ***Example: How does the unified framework's approach to handling the RefCOCOg task diverge in performance between the VL-T5 and VL-BART models?***

Table 7: Definitions of five reasoning categories in M3SCIQA.

categorized into four distinct reasoning types: *comparison*, *data extraction*, *location*, *and visual understanding*. These categories are comprehensive for scientific image understanding. By following the predefined reasoning type definitions in Table 6, we mitigate the risk of annotator bias driven by their own preferences. Additionally, these reasoning types are not specific to NLP and are carefully chosen such that they are applicable in analyzing scientific images in the broader scientific fields.

C.3 Expert Annotation for Detail-Specific Questions

We require each reference paper to have ArXiv documentation. Then, we use the ArXiv downloader to obtain the full text of the reference paper and generate subsequent detail-specific questions (along with answers, explanations, and evidence) using the prompts described in Section F.5. We test these questions in the oracle setting, use GPT-based evaluators to evaluate if the answer generated in the oracle setting matches the answer generated along with the question. If they do not match, expert annotators proceeded to manually examine these questions and re-write the answers.

C.4 Expert Annotation for Detail-Specific Reasoning

In Section F.5, we automatically assign reasoning types concurrently with the generation of detail-specific questions. To ensure the quality of the generated questions, we prompt GPT-4 with the question and its assigned reasoning type to ask if the question matches the reasoning type. For every question that GPT-4 flags as not matching the assigned reasoning type, expert annotators were instructed to manually examine the reasoning types and correct them when necessary.

C.5 Expert Annotation for Detail-Specific Answers

Following the two-round answer generation process mentioned in Section 2.3, we manually checked 100 questions for which the first and second round answers matched in order to ensure the gold answers were indeed correct. Out of the 100 sampled questions, 96 questions were marked as correct by expert annotators, demonstrating the high-quality of M3SCIQA benchmark.

D More Dataset Analysis

Question Distribution. As illustrated in Table 1, the average question length in M3SCIQA is 41.27 (in tokens), while the maximum number of tokens in a question is 78 (in tokens).

Figure 7 further illustrates the distribution of token counts in all locality-specific, detail-specific, and combined questions, highlighting the diverse distribution of all three types of questions. In these figures, the red solid line represents the median and the blue dashed line represents the mean. From all three distributions, we note that the median and mean are very close in values, implying our dataset is symmetric or only slightly skewed.

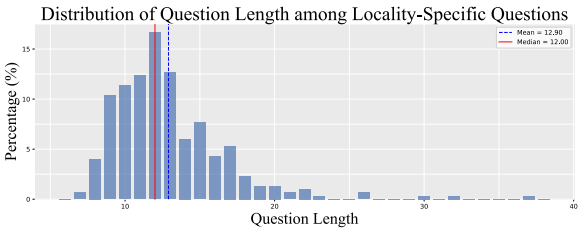


Figure 7: The distribution of the number of tokens per locality-specific question in M3SCIQA- Part 1 of 3.

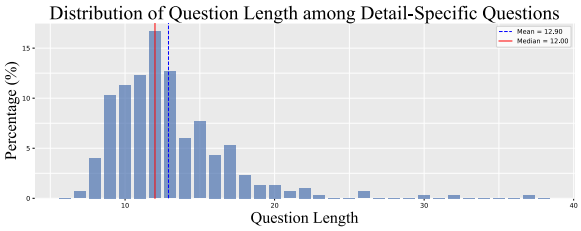


Figure 7: The distribution of the number of tokens per detail-specific question in M3SCIQA- Part 2 of 3.

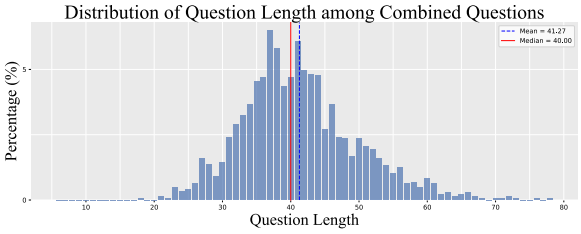


Figure 7: The distribution of the number of tokens per combined question in M3SCIQA- Part 3 of 3.

E More Result Analysis

Recall@k for Locality-Specific Evaluation. In addition to the MRR values shown in Table 2, Recall@k is illustrated in Table 8.

Model	Recall @1	Recall @3	Recall @5
GPT-4o	0.40	0.53	0.57
GPT-4V(ision)	0.30	0.45	0.51
Claude-3-Opus	0.20	0.33	0.44
Claude-3-Sonnet	0.30	0.46	0.57
Claude-3-Haiku	0.09	0.25	0.29
Gemini-Pro-Vision-1.0	0.12	0.21	0.26

Table 8: Recall@k

NDCG@k for Locality-Specific Evaluation. In addition to the MRR values shown in Table 2, NDCG@k is illustrated in Table 9.

Standard Metrics for Detail-Specific Evaluation. In addition to the LLM-based accuracy results shown in Table 3, ROUGE scores are illustrated in Table 10; AutoACU scores (Liu et al., 2023b) are illustrated in Table 12; and each BERTScore (Zhang et al., 2020a) is provided in Table 11.

Model	NDCG @1	NDCG @3	NDCG @5
GPT-4o	0.40	0.51	0.53
GPT-4V(ision)	0.30	0.43	0.45
Claude-3-Opus	0.20	0.31	0.36
Claude-3-Sonnet	0.30	0.44	0.49
Claude-3-Haiku	0.09	0.23	0.25
Gemini-Pro-Vision-1.0	0.12	0.19	0.21

Table 9: NDCG@k

Model	ROUGE-1	ROUGE-2	ROUGE-l
Llama-2-70B	0.125	0.056	0.098
Mistral-7B	0.182	0.086	0.143
PaLM-2	0.197	0.095	0.157
Gemma-7B	0.073	0.032	0.058
DBRX	0.155	0.075	0.122
†Command R+	0.205	0.079	0.176
†GPT-4	**0.237**	**0.127**	**0.202**
†GPT-3.5	0.208	0.101	0.171
†Gemini-Pro-1.0	0.192	0.104	0.162
†Claude-3-Haiku	0.176	0.090	0.143
†Claude-3-Sonnet	0.184	0.086	0.144
†Claude-3-Opus	0.182	0.087	0.140

Table 10: ROUGE score on *test* set of M3SCIQA in *retrieval* setting from GPT-4V(ision)'s retrieval. The best-performing model in each category is **in-bold**, and the second best is underlined.

Model	Recall	Precision	F-1
Llama-2-70B	0.852	0.807	0.828
Mistral-7B	0.855	0.832	0.843
PaLM-2	0.855	0.843	0.848
Gemma-7B	0.359	0.355	0.357
DBRX	0.721	0.698	0.709
†Command R+	0.856	**0.862**	**0.859**
†GPT-4	**0.865**	0.851	0.858
†GPT-3.5	0.861	0.842	0.851
†Gemini-Pro-1.0	0.852	0.847	0.849
†Claude-3-Haiku	0.855	0.827	0.840
†Claude-3-Sonnet	0.856	0.829	0.842
†Claude-3-Opus	0.855	0.827	0.840

Table 11: BERTScore on *test* set of M3SCIQA in *retrieval* setting from GPT-4V(ision)'s retrieval. The best-performing model in each category is **in-bold**, and the second best is underlined.

F More Details On the Setup

F.1 LLM-Based Evaluator.

Cohen's Kappa results are detailed in Table 13, illustrating the level of concordance between expert annotators and LLM-evaluators. Our result reveals a Cohen's Kappa value of 0.520 for the 0-0.5-1 scale and 0.444 for the 1-2-3-4-5 scale. While the Cohen's Kappa value of 0.520 only indicates a "weak agreement" with humans (McHugh, 2012), the 0-0.5-1 scale demonstrates stronger agreement compared to the 1-2-3-4-5 scale. Additionally, the evaluation prompts used for both the 0-0.5-1 and 1-2-3-4-5 scales are provided in Table 16.

Model	Recall	Precision	F-1
Llama-2-70B	0.212	0.091	0.111
Mistral-7B	0.176	0.104	0.109
PaLM-2	0.170	0.123	0.113
Gemma-7B	0.097	**0.198**	0.107
DBRX	0.164	0.131	0.111
†Command R+	0.155	0.153	0.115
†GPT-4	**0.226**	0.164	**0.158**
†GPT-3.5	0.195	0.124	0.118
†Gemini-Pro-1.0	0.170	0.134	0.123
†Claude-3-Haiku	0.217	0.113	0.118
†Claude-3-Sonnet	0.215	0.010	0.110
†Claude-3-Opus	0.224	0.108	0.116

Table 12: AutoACU (A3CU) score on *test* set of M3SCIQA in *retrieval* setting from GPT-4V(ision)'s retrieval. The best-performing model in each category is **in-bold**, and the second best is underlined.

	0-0.5-1	1-2-3-4-5
Expert Annotators	0.520	0.444

Table 13: Cohen's Kappa between two LLM-evaluators *w.r.t.* expert annotations.

F.2 Prompt for Evaluating Locality-Specific Question

Prompts used to rank reference papers across multiple LMMs are illustrated in Table 14.

F.3 Prompt for Answering Detail-Specific Question

Prompts used to answer detail-specific questions are illustrated in Table 15.

F.4 Prompt for Answer Evaluation

Prompts used to retrieve answers from each text chunk and aggregate answers are illustrated in Table 16.

F.5 Prompt for Detail-Specific Question Generation

We list our prompt for detail-specific question generation in Table 17.

Model	Prompt
Yi-VL-34B DeepSeek-VL	Answer the question from the figure and the reference papers provided only: {question} Additionally, rerank the following reference papers according to their relevance to this question. Each reference paper consists of an S2_id, a title, and an abstract. {paper_cluster} Format your answer as a python dictionary with keys "question", "answer", and "rank". "rank" should be a list of S2_id. If no relevant reference papers are provided, return an empty list for "rank". Note: The "rank" list should only include **question-relevant** reference papers. Do not include irrelevant ones.
InternVL-Chat-1.1V	You are given a figure, a question, and some paper candidates of titles and abstracts. Your task is to answer the question based on the figure information, then order the paper candidates that I provide to you so that the paper that is more relevant to the question comes first in the list. Provide your answer at the end in a json file of this format using S2_id only:{"ranking": [""] }. Make sure the responded list is in a valid format and that it only contains the S2_id. Do not include the title or abstract in the answer list. <question> {question} </question> <paper candidates> {paper_cluster} </paper candidates>
LLaVa-1.6 Qwen-VL	Answer the question from the figure and the reference papers provided only: {question} Additionally, rerank the following reference papers according to their relevance to this question. Each reference paper consists of an S2_id, a title, and an abstract. {paper_cluster} Format your answer as a python dictionary with keys "question", "answer", and "rank". "rank" should be a list of S2_id. If no relevant reference papers are provided, return an empty list for "rank".
GPT-4o GPT-4V(ision) Gemini-Pro-Vision-1.0 Claude-3-Haiku Claude-3-Sonnet Claude-3-Opus	You are given a figure, a question, and a list of paper candidates of titles and abstracts. Your task is to answer the question based on the figure information and then re-rank the list of paper candidates I provided to you. Provide your answer at the end in a json format using the S2_id only: {"ranking": []}. Only include papers that are relevant. Do not include papers that are irrelevant. Make sure the answer list is properly formatted. <question> {question} </question> <paper candidates> {paper_cluster} </paper candidates>

Table 14: Prompts used to rank reference papers across multiple LMMs.

Stage	Prompt
Answers from text chunk	Answer the below question about a scientific paper. The question is composed of 2 parts, and the second part of the question can be answered from the paper. I will provide you with only a chunk of a paper. Explain your reasoning. Append the answer at the end of the response in a json format {"answer": ""}. You should answer the question in a short-answer form. Do not provide long answers. If you do not know the answer, respond with {"answer": "I don't know"} <QUESTION> {question} </QUESTION> <CHUNK> {chunk} </CHUNK>
Answer aggregation	I will provide you with a set of answer candidates for a question. Aggregate the information from all the candidates and give me one single answer. Note that if one answer candidate is 'I don't know', you can ignore it. Answer the question based on the answer candidates and summarize the final answer into a short answer. <QUESTION> {question} </QUESTION> {answer_candidate_list}

Table 15: Prompts used to generate and aggregate answers from a text chunk.

F.6 Model Parameters for Answering Locality-Specific Question

Model parameters for ranking reference papers from a paper cluster are shown in Table 18.

F.7 Model Parameters for Answering Detail-Specific Question

Model parameters for answering detail-specific questions are exhibited in Table 19.

Evaluator	Prompt
LLM-based Evaluator (0-0.5-1 setting)	I am testing a model's performance on open-ended questions. I want you to help me in checking to see if the candidate answer has the same meaning as the reference answer for a given question. If you think the reference answer and the candidate answer have the same meaning, respond {"selection": "1"}; otherwise, respond by {"selection": "0"}. If you think the candidate is partially correct, respond by {"selection": "0.5"}. If the answer is "I don't know," rate it to 0. <QUESTION> {question} </QUESTION> <REFERENCE> {reference} </REFERENCE> <CANDIDATE> {candidate} </CANDIDATE>
LLM-based Evaluator (1-2-3-4-5 setting)	I am testing a model's performance on open-ended questions. I want you to help me in checking to see if the candidate answer has the same meaning as the reference answer for a given question. Rate the candidate answer from 1, 2, 3, 4, and 5, where 1 means the candidate is the least similar to the reference answer and 5 means the candidate matches to the reference answer perfectly. Respond by {"selection": ""}. If the candidate answer is "I don't know," rate it to 1. Here's some examples you can consider: Question: Why transformer is better than RNN? Reference: Parallel computation Candidate: Computation Rating: 3 Question: What's the major advantage of using ALiBi positional embedding? Reference: Effectively handle sequences of varying lengths, particularly beneficial for very long sequences Candidate: It has more freedom to handle input Rating: 2 Question: What's the model's performance on GSK8K dataset? Refernece: 65.65% Candidate: 44.56% Rating: 1 Question: What specific method does this paper propose to solve LLM searching problem? Reference: MCTS Candiate: Monte Carlo Tree Search is proposed in this paper to solve searching when using decomposed prompting method. Rating: 5 Question: How does the performance change when we switch from CoT to ToT in prompting? Reference: Accuracy from 23.50% to 32.87% Candidate: slightly increase Rating: 4 <QUESTION> {question} </QUESTION> <REFERENCE> {reference} </REFERENCE> <CANDIDATE> {candidate} </CANDIDATE>

Table 16: Prompts used to evaluate answers generated by LLMs.

G A Comparative Study of LMMs in Answering Locality-Specific Questions

In our experiments, we evaluated numerous LMMs in answering locality questions, such as Kosmos2, Fuyu-8B, and Qwen-VL-Chat. Our findings indicate that these models severely suffer from both hallucination and formatting errors when analyzing the scientific figures. Thus, we conclude that they lack the basic capabilities to generate valid rankings, which are crucial for calculating MRR.

G.1 InternVL-Chat-1.1V

InternVL-Chat-1.1V operates with a short context window, a restriction that makes answering locality-specific questions particularly difficult. Although pairwise paper rankings were still possible within the token length restrictions, prompting the model with the entire list of possible reference paper titles and abstracts was not possible. Since the vanilla singular prompting method used to test other models with larger context windows (e.g. GPT-4V) on the locality-specific question dataset could not be applied to InternVL-Chat-1.1V, we used a slightly different prompting scheme.

Three different ranking settings and methodologies were used to determine the rank of the reference paper for each locality-specific question. In the first setting, the model was repeatedly prompted to compare the true reference paper against each of the other papers one at a time in a head-to-head

Model	Prompt
Detail-Specific Question Generation Prompt	Generate 1 short answer question based on the paper's full content below. You should follow the reasoning type of {reasoning_type}, with the definition {reasoning_description}. The short answer question should be as hard as possible, and focus on a single detail from the paper. The target audience of the short answer question is an expert in the field of natural language processing. The question should be hard for GPT-4 to answer. The answer to the question should be short and must be answerable from the content of the paper. Here are some requirements: <REQUIREMENTS> [Question] should make sense and can be answered from the paper's full text. [Answer] should be directly answering the question you generated. [Explanation] should explain why the answer correctly answers the question. [Evidence] should be from the original content from the paper content. This should be an excerpt from the input paper that supports your answer. </REQUIREMENTS> Append the answer at the end of your response in a json-like format: {"question": "", "answer": "", "explanation": "", "evidence":""} <PAPER FULL CONTENT> full_text </PAPER FULL CONTENT>

Table 17: Prompt for detail-specific question generation.

Model	Generation Setup
LLaVa-1.6	model = `llava-v1.6-mistral-7b`, temperature = 0.1, max_tokens = 8192
Yi-VL-6B	model = `Yi-VL-6B`, temperature = 0.1, max_tokens = 8192
DeepSeek-VL	model = `deepseek-vl-7b-chat`, temperature = 0.1, max_tokens = 8192
InternVL-Chat-1.1V	model = `InternVL-Chat-Chinese-V1-1`, temperature = 0.1, max_tokens = 768
Qwen-VL-Plus	model = `qwen-vl-plus`, seed = 1234, max_tokens = 6000
GPT-4o	model = `gpt-4o`, temperature = 0.1, max_tokens = 4096
GPT-4V(ision)	model = `gpt-4-turbo`, temperature = 0.1, max_tokens = 4096
Gemini-Pro-Vision-1.0	model = `gemini-pro-vision`, temperature = 0.1, max_tokens = 4096
Claude-3-Haiku	model = `claude-3-haiku-20240307`, temperature = 0.1, max_tokens = 4096
Claude-3-Sonnet	model = `claude-3-sonnet-20240229`, temperature = 0.1, max_tokens = 4096
Claude-3-Opus	model = `claude-3-opus-20240229`, temperature = 0.1, max_tokens = 4096

Table 18: Parameters of various LMMs in evaluating locality-specific questions.

Model	Generation Setup
Llama-3-70B	temperature = 0.1, max_token = 10,000
Mistral-7B	temperature = 0.1, max_token = 40,000
PaLM-2	temperature = 0.1, max_token = 40,000
Gemma	temperature = 0.1, max_token = 12,000
DBRX	temperature = 0.1, max_token = 40,000
Command R+	temperature = 0.1, max_token = 200,000
GPT-4	model = `gpt-4-0125-preview`, temperature = 0.1, max_tokens = 200,000
Gemini-Pro-1.0	model = `gemini-1.0-pro`, temperature = 0.1, max_tokens = 40,000
Claude-3-Haiku	model = `claude-3-haiku-20240307`, temperature = 0.1, max_tokens = 250,000
Claude-3-Sonnet	model = `claude-3-sonnet-20240229`, temperature = 0.1, max_tokens = 250,000
Claude-3-Opus	model = `claude-3-opus-20240229`, temperature = 0.1, max_tokens = 250,000

Table 19: Parameters of various LLMs in evaluating detail-specific questions.

Models	validation	test
Method 1	0.07	0.07
Method 2	0.218	0.186
Method 3	0.152	0.193

Table 20: MRR for InternVL

model (percentage)	Rank All	Rank Valid (53.1%)	Rank Ground Truth (5.0%)
QwenVL-Plus	0.047	0.089	0.947

Table 21: MRR for QwenVL-Plus on the *test* set across 3 evaluation settings.

ranking. In this setting, we then considered the true reference paper's rank to be one more than the number of papers individually ranked higher than the true reference paper when compared side-by-side. In the second setting, the model was prompted to assign a rating to each of the sampled reference papers; the ratings were then sorted to generate a final ranking among the papers. Finally, in the third setting, the model randomly paired papers together, with each of the higher ranked papers in each pair considered to be ranked higher than every lower ranked papers. By then iteratively pairing papers among the set of higher-ranked papers and also iteratively pairing papers among all the initially lower-ranked ones, a ranking for the true reference paper was generated.

Comparing each pair of sample papers requires a quadratic number of queries to the model, which requires a significant amount of time. Each of the three proposed methods, on the other hand, require a number of model queries that is linear in the total number of sample references.

However, each of the methodologies have their own potential flaws. The first ranking methodology was asymmetric in that the true reference paper was prompted a different number of times; thus, for a method with no reasoning or retrieval capabilities, the true reference paper would have a $1/2^{n-1}$ chance of being ranked first, while it would have a $1/n$ chance of being ranked first in the ranking mechanism used in larger models, if there are n papers to rank. Since MRR heavily favors smaller ranks, the first ranking methodology would bias the observed MRR downward. The second methodology, with zero-shot prompting, was unstable at times; furthermore, the model generally only chose from a set of a few possible ratings (i.e. 0, 80, 90, or 100 out of 100), making it hard to differentiate and rank papers with the same rating. The third method is symmetric in its prompting but yields different results depending on initial pairings; we randomize the papers when pairing, and so this method is unbiased. We report the MRR values from the third method in Table 2. Detailed results are illustrated in Table 20.

G.2 Qwen-VL-Plus

In the locality-specific evaluation stage, only 53.1% of Qwen-VL-Plus's rankings are valid, with a mere 5.0% including the ground truth paper. MRR for QwenVL-Plus is evaluated based on 3 criteria: (1) *Rank All*, assigning a zero value to any invalid rankings; (2) *Rank Valid*, considering only valid rankings for MRR computation; and (3) *Rank Ground Truth*, calculating MRR solely from rankings that include the ground truth. Detailed findings are presented in Table 21, though only *Rank Valid* is reported in Table 3. Additional Error analysis can refer to Figure 8 below.

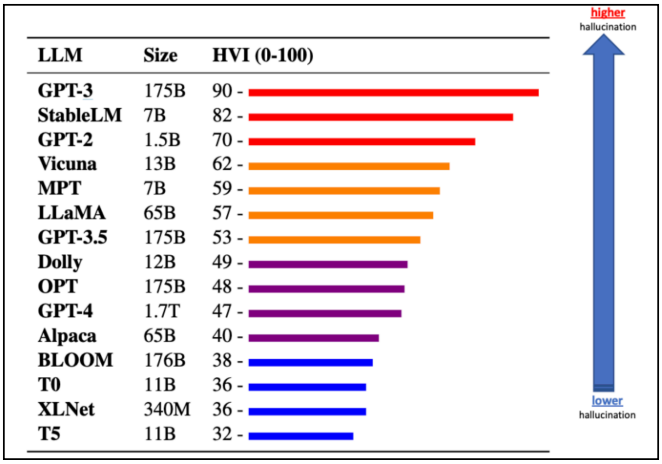
G.3 GPT-4V(ision)

See Figure 9 and Figure 10 below.

G.4 Claude-3-Opus

See Figure 11 and Figure 12 below.

Question: Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?
Image Type: Figure
Reasoning Type: Comparison
Target Paper: GPT-4



LLM	Size	HVI (0-100)
GPT-3	175B	90
StableLM	7B	82
GPT-2	1.5B	70
Vicuna	13B	62
MPT	7B	59
LLaMA	65B	57
GPT-3.5	175B	53
Dolly	12B	49
OPT	175B	48
GPT-4	1.7T	47
Alpaca	65B	40
BLOOM	176B	38
T0	11B	36
XLNet	340M	36
T5	11B	32

Qwen Response (First Round)

Batch 00:
{'question': 'Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?', 'answer': 'Dolly', 'rank': ['c18e13ba65c7247774301314d181c87ee5ebc847']}
Batch 01:
```json
{
  "question": "Which large language model demonstrates lower hallucination than OPT but higher hallucination than Alpa?",
  "answer": "GPT-3",
  "rank": ["964bd39b546f0f6625ff3b9ef1083f797807ef2e"]
}
```
Batch 02:
{
  'question': 'What is the relationship between LLM size and HVI score?',
  'answer': 'A higher LLM size corresponds to a lower HVI score.',
  'rank': ['c10075b3746a9f3dd5811970e93c8ca3ad39b39d',
'b808b6ddf511977e9a33dbe01b412a02b6092ae0']
}
Batch 03:
{
  'question': 'Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpa?',
  'answer': 'Dolly',
  'rank': ['a54b56af24bb4873ed0163b77df63b92bd018ddc',
'7a064df1aeada7e69e5173f7d4c8606f4470365b']
}

Figure 8: Qwen example output on locality-specific question - Part 1 of 3.

Batch 04:
{
'question': 'Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?',
'answer': 'GPT-4',
'rank': ['e952bae3174332b13cbedb35c59c5c2220393f01']
}
Ranking:
['e952bae3174332b13cbedb35c59c5c2220393f01', 'f48287e9ed131ff8ffa79b66717887c5af74f203']

Batch 05:
{
"question": "Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?",
"answer": "T5",
"rank": ["163b4d6a79a5b19af88b8585456363340d9efd04"]
}

Qwen Response (Second Round)

Batch 00:
{'question': 'What is the relationship between hallucinated book and article references?', 'answer': 'The LLM can be said to “know” when it is hallucinating references.', 'rank': ['c18e13ba65c7247774301314d181c87ee5ebc847', '964bd39b546f0f6625ff3b9ef1083f797807ef2e']}

Batch 01:
{
"question": "Which large language model achieves a low HVI score than OPT but high HVI score than Alpaca?",
"answer": "T5",
"rank": ["163b4d6a79a5b19af88b8585456363340d9efd04"]
}

Qwen Response (Third Round)

```python
{
'question': 'Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?',
'answer': 'T5',
'rank': ['c18e13ba65c7247774301314d181c87ee5ebc847',
'964bd39b546f0f6625ff3b9ef1083f797807ef2e']
}
```

The large language model T5 has a lower HVI score than OPT but higher HVI score than Alpaca.

Ranking:
1. c18e13ba65c7247774301314d181c87ee5ebc847 - Do Language Models Know When They're Hallucinating References? [Relevance: High]
2. 964bd39b546f0f6625ff3b9ef1083f797807ef2e - BLOOM: A 176B-Parameter Open-Access Multilingual Language Model [Relevance: Medium-High]
3. 163b4d6a79a5b19af88b8585456363340d9efd04 - GPT-4 Technical Report [Relevance: Low]

Figure 8: Qwen example output on locality-specific question - Part 2 of 3.

Target Paper Ranking: 3

Error Analysis:

(1) **Formatting.** As outlined in Section 3.3, a notable limitation of open-sourced LLMs is their ability to format desired output. During the third round of Qwen response, though the ground truth paper is ranked 3rd in the additional texts, it is not included in the formatted Python dictionary as requested by the prompt. This incident highlights its constrained formatting capabilities. As a corrective measure, each Qwen response is subsequently processed through GPT-3.5/GPT-4 for further formatting before the next round of ranking.

(2) **Text Analysis.** Due to Qwen's limited token length, reference papers are divided into batches of 8 for ranking. Each batch requires the model to restate the locality question in its formatted output, which should remain consistent across batches. However, inconsistencies are observed as the question differs in Round 2 Batch 00. Despite Qwen's high performance across several existing benchmarks, it is hypothesized that the scientific figure input has compromised its text analysis capabilities, resulting in hallucinatory results.

(3) **Figure Analysis.** Variations in the responses to the locality question across different rounds and batches suggest that Qwen's scientific figure analysis capabilities are unstable and may be influenced by textual inputs.

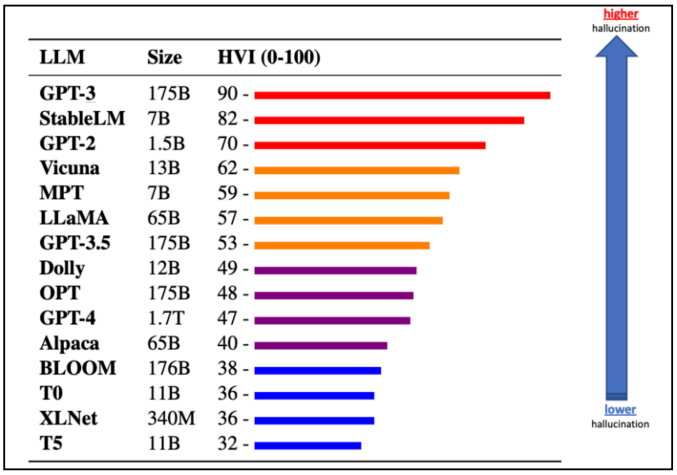
(4) **Prompt Limitation.** The wording used to direct the model to re-rank papers based on their "relevance" to the question may introduce ambiguity in the inference process. This can lead the model to prioritize the topic of the question over the accuracy of the answers. For example, in Round 3, the model potentially emphasized the keyword "hallucination," associated with the Hallucination Vulnerability Index (HVI) score, over "T5," affecting the ranking outcomes.

(5) **Batch Limitation.** The division of reference papers into batches of 8 could result in varying batch size in the last batch, which ranges from 1 to 8 papers. It is conjectured that smaller batches may provide an advantage in ranking accuracy. In this case, with 41 reference papers, the last batch contains only one paper. Although the model incorrectly answered "T5" for the locality-specific question, it still ranked this single paper, leading to a suboptimal outcome where the ground truth paper was ranked third. This factor potentially contributes to the outcome where the ground truth paper was ranked 3rd despite the incorrect answer to the question.

Category: Wrong answer, Suboptimal ranking

Figure 8: Qwen example output on locality-specific question - Part 3 of 3.

Question: Which large language model achieves a lower HVI score than OPT but a higher HVI score than Alpaca?
Image Type: Figure
Reasoning Type: Comparison
Target Paper: GPT-4



LLM	Size	HVI (0-100)
GPT-3	175B	90
StableLM	7B	82
GPT-2	1.5B	70
Vicuna	13B	62
MPT	7B	59
LLaMA	65B	57
GPT-3.5	175B	53
Dolly	12B	49
OPT	175B	48
GPT-4	1.7T	47
Alpaca	65B	40
BLOOM	176B	38
T0	11B	36
XLNet	340M	36
T5	11B	32

GPT-4V(ision) Response: Based on the provided figure, the large language model (LLM) that achieves a lower Hallucination Versatility Index (HVI) score than OPT but a higher HVI score than Alpaca is GPT-4. The HVI score for GPT-4 is 47, which is lower than OPT's score of 48 and higher than Alpaca's score of 40.

Now, let's re-rank the list of paper candidates based on their relevance to the question about the HVI scores of large language models.

Relevant papers will likely discuss the hallucination issue in large language models, methods to mitigate hallucinations, comparisons of different models' propensity to hallucinate, or the development and characteristics of specific models mentioned in the question (GPT-4, OPT, Alpaca).

Target Paper Ranking: 1

Category: Correct reasoning, Optimal ranking

Figure 9: GPT-4V(ision) example output 1 on locality-specific question.

Question: Which Seq2Seq/Tree model has a Test Accuracy of 79.6?
Image Type: Table
Reasoning Type: Data Extraction
Target Paper: Ana-CL

	Model	Test Acc.
Seq2Seq / Tree	GroupAttn(Li et al., 2019)	70.4
	GTS (Xie and Sun, 2019)	71.3
	G2T(Zhang et al., 2020)	72.0
	BERT-T(Liang et al., 2021a)	73.8
	mBERT(Tan et al., 2021)	77.1
	T-Dis†(Liang et al., 2021b)	73.1
	Prototype (Li et al., 2021)	76.3
	Textual-CL†(Shen et al., 2022a)	78
	Ana-CL (Liang et al., 2022)	79.6
Seq2Exp	E-pointer†(Kim et al., 2020)	73.5
	M-Tree†(Wang et al., 2022a)	76.5
	RE-Ext(Jie et al., 2022)	78.6
	M-View♢(Zhang et al., 2022a)	79.5
	Elastic ♣(Zhang and Moshfeghi, 2022)	80.3
	MWP-NAS†(Bin et al., 2023)	79.2
LLM	gpt-3.5-turbo†(OpenAI, 2022)	42.6
	Self-Consistency†(Wang et al., 2022b)	50.7
	Ours	**81.5** ±0.13
	Ours (Layer-Shared)	**81.1** ±0.23

GPT-4V(ision) Response: Based on the figure provided, the Seq2Seq model that has a 79.6 Test Accuracy is "Ana-CL" by Liang et al., 2022.

Target Paper Ranking: 2

Category: Correct reasoning, Suboptimal ranking

Figure 10: GPT-4V(ision) example output 2 on locality-specific question.

Question: What optimization method exhibits an R2 score of 0.191?
Image Type: Table
Reasoning Type: Data Extraction
Target Paper: DPO

	Reward Acc	R1	R2	Rl	Meteor
SALT_{l}	0.368	0.381	0.203	0.371	0.292
SALT_{l+u}	0.591	0.394	0.215	0.383	0.320
$\text{DPO}_{beta=0.1}$	0.484	0.379	0.210	0.369	0.301
$\text{DPO}_{beta=0.5}$	0.532	0.372	0.191	0.361	0.291

Claude-3-Opus Response

Based on the information provided in the image, the optimization method that exhibits an R2 score of 0.191 is DPObeta=0.5.

Here is my ranking of the relevant papers from the provided list, in JSON format:

{"ranking": ["0d1c76d45afa012ded7ab741194baf142117c495",
"58af2d4fcca54c14334d1efd975554b4eb78cd4d", "3ab661db57d924f4ff1706e05ac807873ca00e0a",
"748698bd4387afd08594e0dc8150c2afa210d9ae", "f2b0017ddd77fa38760a18145e63553105a1a236",
"a0a79dad89857a96f8f71b14238e5237cbfc4787", "40c318400809abf5e50aba5a5a80c8012a7715d5",
"8ca62fdf4c276ea3052dc96dcfd8ee96ca425a48", "672491163a327f80e08ce3ef4751e94c78631822",
"a1d3b97f8a6c89bdf4209c714389b9a4dde35353"]}"

Target Paper Ranking: 1

Error Analysis: N/A

Category: Correct reasoning, Correct ranking

Figure 11: Claude-3-Opus example output 1 on locality-specific question.

Question: Which Seq2Exp model is marked with the Club citation symbol?
Image Type: Table
Reasoning Type: Location
Target Paper: Elastic

	Model	Test Acc.
Seq2Seq / Tree	GroupAttn(Li et al., 2019)	70.4
	GTS (Xie and Sun, 2019)	71.3
	G2T(Zhang et al., 2020)	72.0
	BERT-T(Liang et al., 2021a)	73.8
	mBERT(Tan et al., 2021)	77.1
	T-Dis†(Liang et al., 2021b)	73.1
	Prototype (Li et al., 2021)	76.3
	Textual-CL†(Shen et al., 2022a)	78
	Ana-CL (Liang et al., 2022)	79.6
Seq2Exp	E-pointer†(Kim et al., 2020)	73.5
	M-Tree†(Wang et al., 2022a)	76.5
	RE-Ext(Jie et al., 2022)	78.6
	M-View♢(Zhang et al., 2022a)	79.5
	Elastic ♣(Zhang and Moshfeghi, 2022)	80.3
	MWP-NAS†(Bin et al., 2023)	79.2
LLM	gpt-3.5-turbo†(OpenAI, 2022)	42.6
	Self-Consistency†(Wang et al., 2022b)	50.7
	Ours	**81.5** ±0.13
	Ours (Layer-Shared)	**81.1** ±0.23

Claude-3-Opus Response

Based on the information provided in the figure, the Elastic model (Zhang and Moshfeghi, 2022) is marked with the Club citation symbol.

{"ranking": ["e7f63186d9aa212ceb1b65f8d78d646b92960879", "5f19ae1135a9500940978104ec15a5b8751bc7d2", "3f5b31c4f7350dc88002c121aecbdc82f86eb5bb", "473eb062612a17c965eaa62136322f0dec6b1f8e", "fe7ed4acbb7bd02d5b03f9c91f028e8cf56fff77", "e7ad08848d5d7c5c47673ffe0da06af443643bda", "57100e39d0413ee585b381ba9ab366e8a6cf2866", "7b2e65d6bf63217d32c651b514daf02a5a1661b9", "9ccb2beaec722232a84e9a7682c72dcf7de667df", "0ba49945649b40f205503dba3443e2bf550c7115"]}"

Target Paper Ranking: ∞

Error Analysis: The model effectively identifies the model associated with the Club citation marking in a table formatted as a figure, demonstrating its capability in analyzing tables figures. Despite this aspect, the incorrect ranking highlights a deficiency in its reasoning ability. The title of the ground truth paper, "ELASTIC: Numerical Reasoning with Adaptive Symbolic Compiler," clearly indicates its relevance to the Elastic model. However, the top 3 ranked papers lack direct connection to this model, underscoring the model's flawed reasoning in a ranking task.

Category: Correct reasoning, Wrong ranking

Figure 12: Claude-3-Opus example output 2 on locality-specific question.