MCITEBENCH: A Multimodal Benchmark for Generating Text with Citations

Caiyu Hu[♠] **Yikai Zhang**[♠] **Tinghui Zhu**[♠] **Yiwei Ye**[♢] **Yanghua Xiao**[♠*]
[♠]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[♢]School of Computer Engineering and Science, Shanghai University
{cyhu24,ykzhang22,thzhu22}@m.fudan.edu.cn
yiweiye@shu.edu.cn, shawyh@fudan.edu.cn
https://caiyuhu.github.io/MCiteBench

Abstract

Multimodal Large Language Models (MLLMs) have advanced in integrating diverse modalities but frequently suffer from hallucination. A promising solution to mitigate this issue is to generate text with citations, providing a transparent chain for verification. However, existing work primarily focuses on generating citations for text-only content, leaving the challenges of multimodal scenarios largely unexplored. In this paper, we introduce MCITEBENCH, the first benchmark designed to assess the ability of MLLMs to generate text with citations in multimodal contexts. Our benchmark comprises data derived from academic papers and review-rebuttal interactions, featuring diverse information sources and multimodal content. Experimental results reveal that MLLMs struggle to ground their outputs reliably when handling multimodal input. Further analysis uncovers a systematic modality bias and reveals how models internally rely on different sources when generating citations, offering insights into model behavior and guiding future directions for multimodal citation tasks.

1 Introduction

Multimodal Large Language Models (MLLMs) have shown remarkable progress in integrating external information from diverse modalities, allowing them to generate responses beyond the scope of their internal knowledge (Cho et al., 2024; Li et al., 2024b; Zhang et al., 2024b). Despite the advancements, these models frequently suffer from hallucination (Huang et al., 2023; Bai et al., 2024), undermining the faithfulness of their outputs (Zhu et al., 2024). A natural strategy to alleviate this issue is citation: allowing the model to attribute each generated statement to its source, thereby improving transparency and verifiability.

Existing studies on generating text with citations mainly focus on the textual modality (Gao et al., 2023; Liu et al., 2023a). However, real-world information sources are inherently multimodal, often conveying information that cannot be captured by text alone. Although common in practice, citations from non-textual modalities remain underexplored. Grounding model responses in multimodal sources can improve faithfulness and quality (see Figure 1). At the same time, this task poses several challenges for MLLMs. The model must understand cross-modal content, assess the sufficiency of evidence, and remain robust to irrelevant or distracting input. These challenges are still under exploration. In this paper, we construct a benchmark to systematically evaluate MLLMs in generating text with citation from multimodal input.

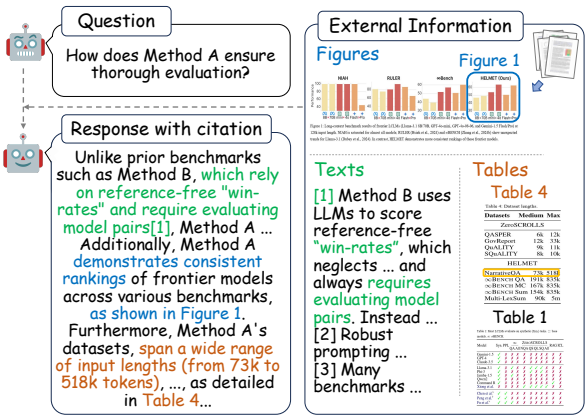

Figure 1: Illustration of the task form in MCITEBENCH. The model takes multimodal corpus and generates responses with explicit citations.

However, building such a benchmark is challenging. First, constructing high-quality question-answer data with multimodal evidence is non-trivial. It requires not only the accurate extraction of heterogeneous evidence (e.g., tables, figures, and passages), but also careful alignment between the evidence and the answer. In cases where multiple pieces of evidence jointly support an answer, it is critical to ensure their mutual consistency and suf-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5949–5966
November 4-9, 2025

ficiency. Second, evaluating MLLMs in this setting introduces additional complexity. A key issue is how to assess cross-modal entailment—whether the cited evidence truly supports the generated answer. Moreover, the citation must correspond closely to the response, ensuring that the retrieved evidence is both necessary and relevant to the output. These challenges highlight the need for a comprehensive evaluation framework that examines multiple dimensions of model performance.

In this paper, we propose MCITEBENCH, the first benchmark for evaluating the ability of MLLMs to generate text with citations in multimodal settings. To address the challenges outlined above, we begin by collecting academic papers and extracting reliable information sources across multiple modalities. These sources are rigorously filtered to form a high-quality attribution corpus. Based on this corpus, we construct question–answer pairs using review–rebuttal interactions, where each answer is supported by evidence. To comprehensively evaluate model performance, we assess models along three axes: citation quality, source reliability, and answer accuracy. Extensive experiments reveal several notable findings: 1) While MLLMs can often answer questions correctly, they struggle to generate accurate citations, particularly when the evidence spans multiple sources. 2) MLLMs are better at attributing citations to textual than to visual evidence, suggesting a potential modality bias.

Our contributions are summarized as follows:

- To the best of our knowledge, MCITEBENCH is the first benchmark that systematically evaluates the ability of MLLMs to generate text with citations from multimodal input.
- MCITEBENCH comprises 3,000 samples of different difficulty levels, including both single- and multi-source evidence, as well as single- and mixed-modality cases. To support comprehensive evaluation, we define multi-dimensional metrics capturing citation quality, source reliability, and answer accuracy.
- We conduct experiments to assess the models' ability to generate text with citations across different modalities. Results reveal that MLLMs exhibit a modality bias, favoring textual over visual sources in citation generation.

2 Related Work

Generating Text with Citations Recent efforts have explored the task of generating text with citations, where models are required to produce responses with explicit references to supporting sources. Gao et al. (2023); Liu et al. (2023a) first introduced this setting to improve the verifiability of model responses. Subsequent works have explored two main paradigms: generating both the response and citations simultaneously (Aly et al., 2024; Huang et al., 2024), and attaching citations in a post-processing step (Slobodkin et al., 2024; Li et al., 2024a). These approaches have also been extended to tasks such as long-context citation (Zhang et al., 2024a) and fine-grained attribution (Xu et al., 2024). Another related line of work is traditional citation text generation, which typically refers to generating citation sentences in academic papers that contain specific scientific claims and cite prior work (Li and Ouyang, 2024; Mandal et al., 2024; Şahinuç et al., 2024). However, existing studies focus almost exclusively on textual evidence, limiting their applicability in real-world multimodal scenarios. In this work, we address this gap by incorporating figure and tabular content as citation sources and evaluating model attribution in multimodal contexts.

Multimodal RAG Multimodal retrieval-augmented generation (mRAG) (Zhao et al., 2023) augments multimodal large language models with retrieved external information, enabling them to answer queries that cannot be resolved using internal knowledge alone. Zhang et al. (2024b) acquire unknown visual knowledge through web search to aid in answering queries, while Li et al. (2024b) builds a self-adaptive retrieval agent to plan the reasoning path. Additionally, Cho et al. (2024) improve multi-page and multi-document understanding through multimodal retrieval. While these approaches integrate retrieval into the generation pipeline, they do not assess whether the generated responses faithfully reflect the retrieved content. In this work, we shift the focus from retrieval itself to attribution: evaluating whether the model can correctly ground its outputs in the provided multimodal sources.

3 MCITEBENCH

In this section, we define the task of generating text with citations from multimodal input and describe

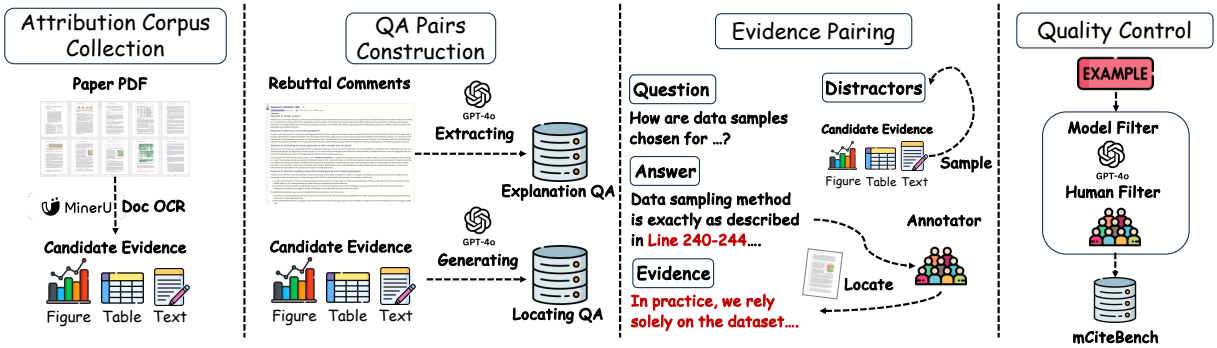

Figure 2: The construction pipeline of MCITEBENCH. Initially, we collect multimodal academic papers along with their corresponding review-rebuttal interactions and then parse the papers to extract candidate evidence. GPT-4o is used to extract explanation QA pairs from the comments and generate locating QA pairs. Next, human annotators match the references in the answers to the relevant content in the original papers. Finally, the data filtered and labeled by the model is manually verified by human annotators to ensure consistency and accuracy.

the construction of our benchmark, MCITEBENCH. As shown in Figure 2, the pipeline consists of four main stages: **Attribution Corpus Collection**, **QA Pairs Construction**, **Evidence Pairing**, and **Quality Control**. We begin by collecting academic papers, which serve as a source of rich multimodal content. Based on these papers, we construct question–answer pairs from review–rebuttal interactions. Human annotators are employed to link answers to their supporting evidence.

3.1 Task Definition

Given a query q and a multimodal evidence set M, where M includes both the ground truth evidence and distractors related to q, the model is required to generate an answer a along with a set of citations C. For each sentence s_i in the answer, the model generates a set of citations $C_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,k_i}\}$, where k_i denotes the number of cited evidence associated with sentence s_i. Each citation $c_{i,j}$ refers to a specific piece of evidence from the multimodal evidence set M.

3.2 Attribution Corpus Collection

To evaluate how well MLLMs generate text with citations, an attribution corpus that includes multimodal information sources and allows for easy verification of cited evidence is needed. In MCITEBENCH, we use academic papers as the attribution corpus because of the following characteristics: 1) Academic papers contain rich content from multiple modalities (*e.g.*, text, figure, and table) that individually or collectively support the arguments. 2) The information sources in academic papers are numbered (*e.g.*, "Figure 1", "Table 2", and text in "Line 10"), making it easy to match them with the cited results. 3) Academic papers cover the latest contents beyond pre-training data, reducing the risk of data leakage.

We collect papers from OpenReview and extract multimodal content using MinerU (Wang et al., 2024a), a state-of-the-art document parsing framework. To avoid contamination from model training data, we focus on ICLR 2025 submissions, which became publicly available in November 2024—after the knowledge cutoff of the evaluated models. ICLR is chosen for its open review process, which includes accessible reviews and author responses, offering reliable structure for citation annotation. From this collection, we obtain a diverse set of multimodal content, including over 400k text paragraphs, 40k images, and 9k tables, which serve as candidate evidence. A subset of this corpus is selected as candidate evidence and distractors for constructing the final 3k evaluation samples.

3.3 QA Pairs Construction

After collecting the attribution corpus, we construct question–answer pairs with explicit references to the supporting evidence. Establishing a reliable correlation between questions and evidence is challenging, as the source of information must be accurately linked to the generated answers.

We divide MCITEBENCH data into two categories: **Explanation** and **Locating**. Explanation questions require in-depth analysis of evidence and often yield long-form responses (*e.g.*, "How is the model's performance evaluated?"). In contrast, Locating questions are straightforward and can be answered by directly identifying the correct evi-

dence (*e.g.*, "Which model performs better on the XYZ benchmark, GPT-4o or GPT-4o-mini?").

For Locating questions, we use GPT-4o to generate structured QA pairs with supporting details. Specifically, we construct QA pairs (Q, A), where each question $q_i \in Q$ is formulated based on specific evidence, and each answer $a_i \in A$ is directly linked to the corresponding source.

However, generating questions that require information from multiple sources remains a challenge for MLLMs. Models often fail to integrate all necessary evidence, resulting in questions that can be answered by a single source rather than all selected evidence. To address this, we leverage review-rebuttal interactions to construct Explanation QA pairs. In this setting, reviewers' questions and authors' responses are used, with responses grounded in multiple evidence segments from the paper (*i.e.*, attribution corpus). From these data, we construct QA pairs (Q, A) by extracting questions q_i and the corresponding answers a_i.[1]

3.4 Evidence Pairing

Review–rebuttal interactions often include rich evidence in the authors' responses to support their claims. For example, when addressing a reviewer's concern about model performance, an author might respond, *"Our approach achieves 85.2% accuracy, as shown in Table 3 and discussed in Section 4.2."* These references provide valuable entry points for identifying the evidence that grounds the answer. Therefore, we extract the supportive evidence $e_i \in E$ from $a_i \in A$ to construct (Q, A, E) triplets. While E provides explicit references (e.g., "Table 3", "Section 4.2"), these references must be resolved to their corresponding content in the source papers before they can be used as input for MLLMs. To achieve this, human annotators manually map each reference to the associated content in the original paper, categorizing the evidence as either text, image, or table.

Distractor Construction. To evaluate whether models can correctly cite relevant sources while ignoring irrelevant ones, we introduce distractor content into the input. These distractors are sampled from the same paper, ensuring a balanced distribution of multimodal content (text, images, tables). Each final sample in MCITEBENCH is formatted as (Q, A, E, D), where Q is the question, A is the correct answer, E is the evidence and D is the distractors.

Statistic	Number
Total questions	3,000
- Explanation	2,000
- Locating	1,000
Evidence sources	
- Single-source	2,538
- Multi-source	462
Evidence modality	
- Text	1,243
- Figure	941
- Table	533
- Mixed	283
Total papers	1,749
Average questions per paper	1.72

Table 1: Statistics of MCITEBENCH.

3.5 Quality Control

After constructing (Q, A, E, D), we apply a quality control pipeline that first uses automated filtering followed by human verification. Initially, GPT-4o assigns quality labels and filters out low-quality samples based on predefined criteria such as relevance, clarity, and evidence alignment. The filtered candidates are then manually verified by annotators to ensure consistency and accuracy, focusing on removing any unclear or incorrect instances.[2]

3.6 Statistics of MCITEBENCH

As shown in Table 1, MCITEBENCH comprises 3,000 data samples for evaluating the ability of MLLMs to generate text with citations, extracted from 1,749 academic papers with an average of 1.72 questions per paper. Among these, 2,000 are Explanation tasks that require detailed evidence analysis and often lead to long-form answers, while 1,000 are Locating tasks that focus on direct evidence identification. The evidence is balanced across modalities, with 1,243 textual, 1,474 visual (including 941 figures and 533 tables), and 283 mixed-modality sources, ensuring diverse multimodal attribution scenarios.

4 Evaluation Metrics

We evaluate the models across three dimensions: **citation quality**, **source reliability**, and **answer accuracy**. Using **Citation F1**, we assess whether the cited evidence accurately and sufficiently supports

[1] Details of prompt design and reference extraction strategies are in Appendix B.1

[2] Details of the human annotation process can be found in Appendix C.2.

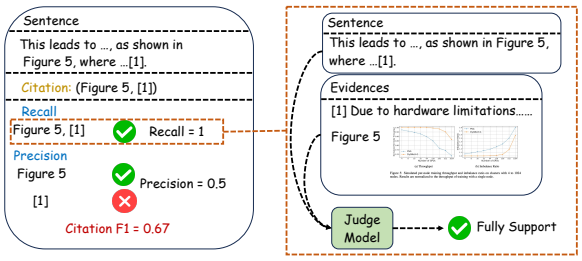

Figure 3: The calculation of Citation F1.

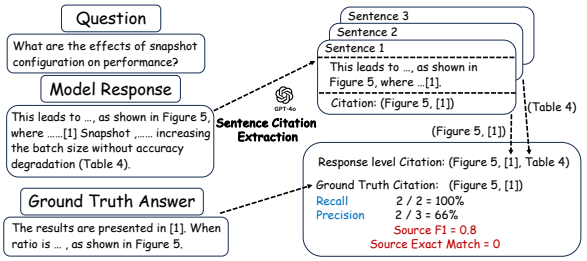

Figure 4: The calculation of Source F1 and Source Exact Match.

the model's response. Source reliability ensures that the model's response cites the ground truth source needed to answer the query. We measure this by comparing the model-generated citation with ground truth citation, using both **Source F1** and **Source Exact Match** scores. Answer accuracy metrics are designed to assess whether the model's response correctly addresses the query.

Citation F1 (C-F1). Citation quality is evaluated using Citation F1, which measures the alignment between cited evidence and the generated response, ensuring that the response is supported by the cited evidence without including irrelevant ones.

As illustrated in Figure 3, a judge model evaluates whether each sentence is supported by its cited evidence. Citation Recall is calculated using a scoring system inspired by LongCite (Zhang et al., 2024a), categorizing citations into three levels: No support, Partially supported, and Fully supported, with corresponding scores of 0, 0.5, and 1. Citation Precision is determined on a binary scale, scored as either relevant (1) or irrelevant (0) to the cited evidence. For sentences citing multiple sources, the final precision score is the average across all cited evidence. Finally, Citation F1 is computed as the harmonic mean of Recall and Precision, providing a balanced measure of the model's citation quality.

Source F1 (S-F1). As shown in Figure 4, Source F1 measures the alignment between citations in the model's response and ground truth citations, evaluating whether the model cites evidence that aids in answering the query.

We first split the model-generated responses into sentence-citation pairs (s_i, c_i) using GPT-4o. These sentence-level citations are then aggregated to form response-level citations, which are compared against the ground truth. The precision, recall, and F1 score are calculated as follows:

$$\text{Source Precision} = \frac{|C_{\text{pred}} \cap C_{\text{gt}}|}{|C_{\text{pred}}|}, \quad (1)$$

$$\text{Source Recall} = \frac{|C_{\text{pred}} \cap C_{\text{gt}}|}{|C_{\text{gt}}|}, \quad (2)$$

We calculate Source F1 by computing the harmonic mean of Recall and Precision. C_{pred} represents the set of citations generated by the model, and C_{gt} denotes the ground truth citations. The intersection $C_{\text{pred}} \cap C_{\text{gt}}$ counts the correctly cited evidence.

Source Exact Match (S-EM). The Source Exact Match metric provides a stricter evaluation, indicating whether the model's response-level citation is the same as the ground truth.

$$\text{Source EM} = \begin{cases} 1, & \text{if } C_{\text{pred}} = C_{\text{gt}} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Accuracy (Acc). We evaluate answer accuracy using the LLM-As-Judge (Zheng et al., 2023; Liu et al., 2023b) framework for both Explanation and Locating questions. The judge model scores each response and reference answer according to criteria specific to each question type, and the scores are then normalized. In Explanation cases, direct comparison with a ground truth answer is not feasible. Instead, we use the authors' responses as the reference and employ a judge model to evaluate the generated answers based on their relevance, logical consistency, and fluency. In Locating scenarios, this evaluation method mitigates issues related to errors caused by minor formatting differences.[3]

5 Experiments

5.1 Evaluation Settings

Implement Details. In this work, we indicate citations from textual content using box brackets (*e.g.*, "[1]"), and refer to figures and tables by the indices in their captions (*e.g.*, "Figure 3", "Table 2"). We conduct an ablation study to assess the

[3]Detailed scoring criteria and judgment prompts are provided in the Appendix B.2.

Models	Explanation								Locating			
	Single-Source				Multi-Source				Single-Source			
	C-F1	S-F1	S-EM	Acc	C-F1	S-F1	S-EM	Acc	C-F1	S-F1	S-EM	Acc
Open-Source Models (7-14B)												
LLaVA-OV-7B	19.93	10.84	5.34	47.79	31.14	22.48	1.26	49.68	26.31	20.93	11.63	60.10
LLaVA-OV-7B-Chat	28.77	13.90	1.43	47.76	35.74	29.82	3.00	49.78	29.58	23.33	4.05	53.85
MiniCPM-V-2.6	49.12	35.23	22.81	51.30	57.90	41.74	5.88	52.60	47.93	52.73	42.94	83.55
Qwen2-VL-7B	58.46	42.98	**35.36**	51.59	58.64	36.62	2.36	53.03	53.99	54.71	46.32	87.45
InternVL2.5-8B	58.47	45.13	33.45	51.53	63.97	45.50	9.86	52.92	55.94	64.17	56.33	83.90
Llama-3.2-Vision-11B	19.65	14.06	9.60	48.63	31.16	25.87	1.22	49.35	26.56	16.56	11.80	61.40
Open-Source Models (>70B)												
Qwen2-VL-72B	53.60	44.81	32.01	52.60	64.66	50.53	8.96	52.38	58.75	68.86	61.48	90.25
InternVL2.5-78B	54.52	42.44	25.40	52.34	71.03	57.65	16.86	54.87	50.57	57.60	52.20	90.10
Llama-3.2-Vision-90B	35.33	28.05	12.30	50.00	46.08	46.73	10.35	51.41	43.69	49.07	32.83	74.75
Proprietary Models												
GPT-4o-mini	43.99	34.42	15.48	52.08	57.81	50.22	8.39	54.22	53.71	58.57	46.56	88.50
GPT-4o	**84.24**	**56.82**	24.50	**54.32**	**89.19**	**67.56**	**21.27**	**56.60**	**91.45**	**85.74**	**69.45**	**90.45**

Table 2: Main results on MCITEBENCH. The highest score is highlighted in **bold**, and the second highest score is underlined. C-F1, S-F1, and S-EM represent Citation F1, Source F1, and Source Exact Match scores, respectively. Acc stands for Accuracy.

impact of including figure captions in the input.[4] For both single-source and multi-source evidence questions, the multimodal corpus M comprises 5 items, including the ground truth evidence and distractors. Distractors are randomly selected from other content within the same paper.

Judge Model. In this study, we use GPT-4o to assess the entailment relationship between model responses and their cited evidence.[5]

Model Choice. For open-source models, we test InternVL-2.5 (8B/78B) (Chen et al., 2024), Qwen2-VL (7B/78B) (Wang et al., 2024b), Llama 3.2-Vision (11B/90B) (Meta, 2024), Llava-OneVision (and its chat version) and MiniCPM-V-2.6 (Yao et al., 2024). For proprietary models, we test GPT-4o (GPT-4o-2024-11-20) and GPT-4o-mini (GPT-4o-mini-2024-07-18) (Hurst et al., 2024).

5.2 Main Results

As shown in Table 2, smaller open-source models achieve lower Citation F1 scores and struggle to select evidence that adequately supports their responses. Furthermore, they also perform poorly in selecting evidence that directly answers the query, as shown by their low Source F1 and Source Exact Match scores. As model size increases, we observe an improvement in citation performance, suggesting that scaling model size enhances attribution capability. In comparison, GPT-4o achieves an 84.24% Citation F1 score on single-source Explanation questions, demonstrating strong citation quality. However, it struggles with source reliability, with Source Exact Match scores remaining low at 24.50% for single-source and 21.27% for multi-source settings. This indicates that even state-of-the-art models struggle to consistently cite evidence that is directly relevant to answering the query, underscoring the difficulty of precise citation in multimodal contexts.

Does Question Difficulty Influence Model Citation Performance? Model performance reflects the difficulty of the questions, with higher accuracy scores observed on locating questions compared to explanation questions, indicating that explanation tasks are more challenging. As shown in Table 2, as question difficulty increases, model citation performance tends to decrease. For instance, GPT-4o achieves 85.74% in Source F1 for single source locating questions but drops to 56.82% for single source explanation questions. Explanation questions place higher demands on citation generation, as they require an in-depth analysis of the inputs.

How Do Multi-Source Scenarios Affect Generating Text with Citations in MLLMs? In multi-source settings, models tend to achieve higher Citation F1 and Source F1 scores, as multiple valid references allow for partial credit. Unlike single-source questions with only one correct citation,

[4] See Table 14 for details in Appendix D.2.

[5] We validate GPT-4o's reliability in Appendix C.3, and further verify in Appendix D.1 that it does not exhibit strong self-preference when evaluating responses in our task.

Model	Overall	By Modality		
		Figure	Table	Text
Open-Source(7-14B)				
Qwen2-VL-7B-Instruct	0.45	0.40	0.38	0.55
InternVL2_5-8B	0.48	0.37	0.42	0.65
Open-Source(>70B)				
Qwen2-VL-72B-Instruct	0.59	0.50	0.57	0.71
InternVL2_5-78B	0.58	0.51	0.50	0.72
Proprietary				
gpt-4o-mini	0.52	0.47	0.48	0.61
gpt-4o-2024-11-20	0.60	0.52	0.55	0.73

Table 3: Model accuracy on identifying the most relevant source for answering a question under the multi-choice setting.

multi-source questions permit credit for correctly identifying any subset of the ground truth, naturally resulting in higher metric values. However, the stricter Source Exact Match metric is lower than in single-source scenarios. This highlights the challenge of citing in multi-source scenarios, where models must correctly include relevant sources while avoiding irrelevant ones.

5.3 Analysis

In this section, we discuss several research questions, revealing the inherent biases in the task.

RQ1: Can MLLMs Accurately Identify the Source Needed to Answer a Question? Generating text with citation can be abstracted into a two-stage process: (1) generating a response, and (2) mapping that response to the appropriate supporting input sources by producing attribution tokens such as "[1]" or "Figure 3".

Instead of requiring the model to generate an answer and then attribute it, we directly evaluate its ability to identify which source would be most helpful in answering a given question. Specifically, we ask: *Can a model identify the correct source needed to answer a given question?*

Settings We construct a probing task based on Single-Source Explanation QA. For each example, we provide the model with a question and 5 candidate sources (1 correct + 4 distractors). The model is tasked with selecting which source would be most helpful in answering the question.[6]

Results Results are presented in Table 3. Importantly, this task directly evaluates the model's ability to identify relevant sources based solely on the question, rather than relying on model-generated answers or intermediate claims. Despite this seemingly simplified setting, no model achieves more than 60% accuracy, highlighting the persistent difficulty in accurately grounding questions in the correct source.

[6]See Table 11 for details in Appendix B.3.

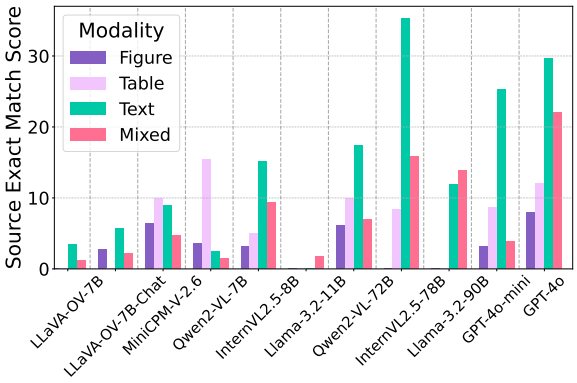

Figure 5: Source Exact Match score of models on the MCITEBENCH benchmark across different modalities, under the multi-source explanation setting with two gold evidence items per question.

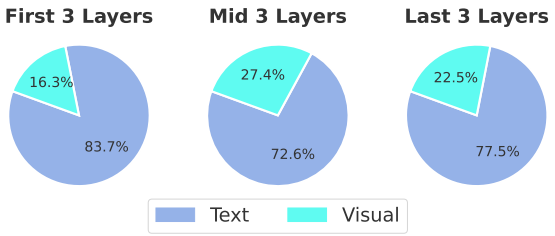

Figure 6: Attention distribution across multimodal sources.

In addition, we observe a consistent performance gap across modalities: models perform better when reasoning over textual sources compared to visual inputs such as figures and tables, which leads to our next research question.

RQ2: Does Modality Influence Citation Performance? We analyze model performance in instances where the evidence modality comes from mixed modalities. The number of evidence is set to 2, and we compare this with data from single modalities with the same number of evidence pieces. As shown in Figure 5, most models achieve high Source EM scores when the ground truth evidence is textual but perform poorly when it is visual. This suggests that although MLLMs can process multimodal inputs, they are better at aligning with textual evidence than accurately citing visual information when generating responses.

To further investigate this, we analyze MLLMs' attention patterns when processing mixed-modality

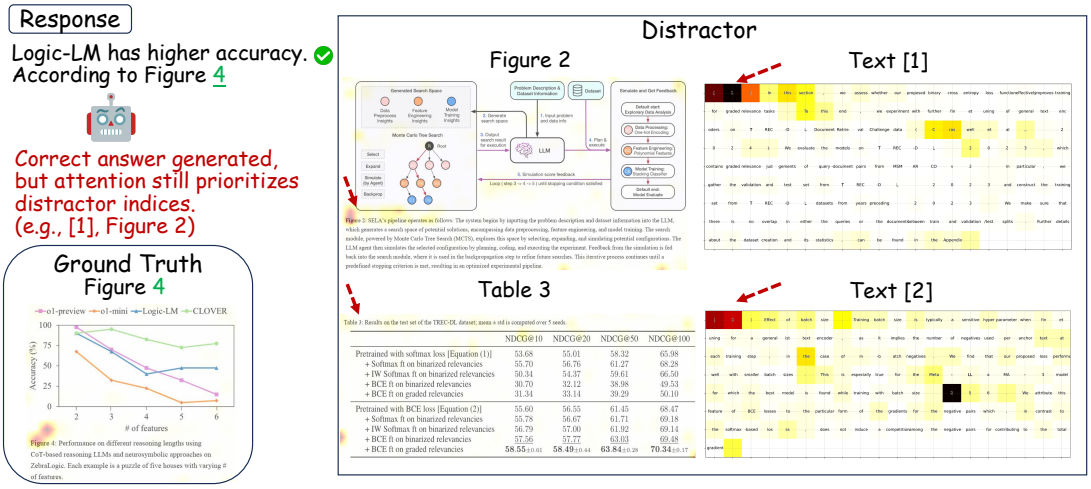

Figure 7: Attention heatmap during source reference generation. The heatmap shows how the model distributes attention when generating the next token in its response, continuing the sentence *"Logic-LM has higher accuracy. According to Figure ◇"*. Although the model answers correctly, its attention in the distractors remains focused on index positions (*e.g.*, "[1]", "Figure 2").

inputs. Using Qwen2-VL-7B as the test model, we calculate the attention distribution across multimodal inputs by averaging attention head scores and normalizing by input source token length across different layers. As shown in Figure 6, the model allocates fewer attention scores to visual inputs compared to text. In contrast, textual information maintains consistently high attention throughout, with 83.7% in early layers and 77.5% in later layers. This indicates that while the model processes all modalities, it prioritizes textual content and utilizes it more effectively than visual data.

RQ3: What Do Models Look At When Generating Citations? Correctly generating source-identifying tokens (e.g., "[1]", "Figure 2") leads to better performance and higher attribution scores. To better understand how models process ground truth evidence and distractors, we analyze their attention distribution when generating source-identifying tokens.

Settings Specifically, we examine the attention patterns of Qwen2-VL-7B when continuing a partially generated sentence ending in *"According to Figure"*, and tasked with predicting the next token (e.g., "4"). This allows us to assess which input regions the model attends to when making source attribution decisions.

Specifically, we focus on its behavior when predicting the next token after *"According to Figure ◇"* in its response. Notably, the distractors are sampled from unrelated papers, meaning they provide no useful information for answering the question.

Results As shown in Figure 7, the model's attention heatmap reveals an intriguing pattern: even when the response is based entirely on a specific piece of evidence, the model's attention does not solely focus on it. When generating the token after "According to Figure", the model's attention remains high on textual index positions (*e.g.*, "[1]", "[2]"), even though the context suggests the model should focus on figure evidence. This suggests that while the model correctly cites the source, it maintains a broader contextual awareness by attending to multiple potential evidence.

6 Conclusion

In this paper, we introduce MCITEBENCH, a high-quality benchmark built from academic papers and their review–rebuttal interactions, to evaluate the ability of MLLMs to generate text with citations from multimodal input. Leveraging this benchmark, we conduct a detailed evaluation of model performance across multiple dimensions. Through extensive experiments, we find that existing models struggle to accurately attribute their outputs to the correct multimodal sources. Furthermore, we dive deep into the analysis of attention distribution during citation generation and uncover modality bias exhibited by current models. We hope that MCITEBENCH offers valuable insights into generating text with citations and contributes to the development of models capable of producing faithful and verifiable responses.

Limitations

In MCITEBENCH, we construct multi-level questions and build an evaluation pipeline for multimodal inputs. However, the current design has limitations in citation granularity. First, citations are limited to the sentence level, meaning that we do not distinguish between multiple claims within a single sentence. For example, if a sentence contains multiple claims supported by different evidence, we treat it as a full sentence-level citation. Second, MCITEBENCH treats subfigures or subtables (e.g., Figure 1a, 1b) as part of the entire figure or table, without distinguishing between them. These limitations highlight areas for future improvement in handling fine-grained attribution tasks.

Ethical Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

Use of Human Annotations Our institution recruited three annotators to perform the evidence linking, data filtering, and robustness evaluation tasks for MCITEBENCH. We ensure that the privacy rights of the annotators are respected throughout the process. The annotators receive compensation exceeding the local minimum wage and have consented to participate in the tasks for research purposes.

Risks The tasks in MCITEBENCH used in our experiment are created by human annotators, and we conduct additional checks to ensure that the content is free from socially harmful or toxic language. However, the evaluation of data quality relies on common sense, which may differ across individuals with diverse backgrounds.

References

Rami Aly, Zhiqiang Tang, Samson Tan, and George Karypis. 2024. Learning to generate answers with citations via factual consistency models. *arXiv preprint arXiv:2406.13124*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training language models to generate text with citations via fine-grained rewards. *arXiv preprint arXiv:2402.04315*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024a. Citation-enhanced generation for llm-based chatbot. *arXiv preprint arXiv:2402.16063*.

Xiangci Li and Jessica Ouyang. 2024. Related work and citation text generation: A survey. *arXiv preprint arXiv:2404.11588*.

Yangning Li, Yinghui Li, Xingyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S Yu, Fei Huang, et al. 2024b. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023b. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Biswadip Mandal, Xiangci Li, and Jessica Ouyang. 2024. Contextualizing generated citation texts. *arXiv preprint arXiv:2402.18054*.

AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024.

Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with llms: A study in citation text generation. *arXiv preprint arXiv:2407.04046*.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. *arXiv preprint arXiv:2403.17104*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2024. Aliice: Evaluating positional fine-grained citation generation. *arXiv preprint arXiv:2406.13375*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. 2024a. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, and Xiangyu Yue. 2024b. Vision search assistant: Empower vision-language models as multimodal search engines. *arXiv preprint arXiv:2410.21220*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

A Supplementary Results

A.1 Detailed Scores by Metric and Modality

In addition to the overall results reported in Table 2, we provide detailed scores grouped by metric in Table 4.

B Prompt Design

B.1 Data Processing Prompts

We list the prompts used for extracting Explanation QA and generating Locating QA in Table 5, 6.

B.2 Evaluation Metric Prompts

We list the prompts used for evaluating citation recall, citation precision, and the accuracy of explanation and locating questions in Table 7, 8, 9, 10.

B.3 Source Identification Prompt

We list the prompt used to evaluate whether a model can identify the most relevant source for answering a given question in Table 11.

C Human Evaluation

C.1 Evidence Paring

Human annotators map each reference to its corresponding content using the GUI shown in Figure 8.

C.2 Quality Control

Our annotation process involves three students from the artificial intelligence field, with one serving as the annotation lead. The process takes approximately one month to complete, and annotators are compensated at the local minimum hourly wage rate. Regarding inter-annotator agreement, in cases of disagreement about whether to retain specific data points, the annotation lead makes the final decision.

Human annotators verify data quality and filter out bad cases using the GUI shown in Figure 9.

C.3 Agreement Between Human Annotations and GPT-4o

To verify the accuracy of our evaluation pipeline, we conducted a manual annotation study on 75 model-generated responses, comprising 25 objective questions and 50 subjective questions, resulting in over 457 entailment judgments. We then compared these human annotations with the entailment judgments produced by GPT-4o. As shown in Table 12, the results indicate a high degree of agreement between human annotations and GPT-4o's predictions, demonstrating the reliability and correctness of our pipeline. The annotation GUI is shown in Figure 10.

D Ablation Study

D.1 Effect of LLM Judge Choice

To further investigate the robustness of our automatic evaluation setup, we conduct an ablation study using alternative judge models. Specifically, we evaluate model performance on 90 randomly sampled examples (30 *Locating* and 60 *Explanation* questions), comparing scores assigned by GPT-4o and DeepSeek V3-0324.

As shown in Table 13, GPT-4o consistently achieves the highest scores under both judge models. While DeepSeek V3-0324 tends to yield slightly higher absolute scores across all models, the relative ranking remains consistent. This suggests that self-preference bias from GPT-4o does not significantly affect evaluation outcomes, confirming the robustness of our LLM-based evaluation setup.

D.2 Effect of Captions on Citation Performance

To assess the role of visual-textual information in multimodal citation understanding, we conduct a comprehensive ablation study across the full benchmark dataset (3,000 examples), comparing model performance with and without image captions.

As shown in Table 14, we observe that including captions leads to minor changes in performance across most evaluation metrics. Notably, the accuracy improvements are modest for both GPT-4o-mini and GPT-4o. Interestingly, in some cases (e.g., GPT-4o-mini), the inclusion of captions slightly degrades performance in label prediction and citation generation (as measured by F1 and exact match), while GPT-4o exhibits a substantial gain in citation F1.

These results demonstrate that our benchmark does not rely solely on OCR-extracted text, and that the image-caption setting we adopt provides a reasonable and realistic testbed for evaluating MLLMs' citation capabilities. At the same time, the relatively limited gains from caption inclusion highlight that current models still face challenges in grounding their responses effectively, even when textual cues are explicitly embedded in the image.

Models	By Metric															
	C-F1				S-F1				S-EM				Acc			
	Fig.	Tab.	Text	Mix	Fig.	Tab.	Text	Mix	Fig.	Tab.	Text	Mix	Fig.	Tab.	Text	Mix
Open-Source Models (7–14B)																
LLaVA-OV-7B	23.67	26.15	19.52	37.59	15.22	18.09	12.31	29.39	10.20	11.65	3.34	1.02	48.19	47.09	57.72	50.71
LLaVA-OV-7B-Chat	29.61	29.19	28.94	38.69	8.31	11.57	28.60	32.60	1.76	2.35	3.40	1.89	45.64	44.56	55.83	50.18
MiniCPM-V-2.6	59.78	57.27	34.07	66.61	48.03	48.26	33.35	45.54	33.37	31.56	25.29	4.05	57.97	53.38	71.24	53.71
Qwen2-VL-7B	73.11	70.50	31.20	62.16	57.83	61.36	24.07	40.28	50.93	47.97	14.49	1.88	60.20	56.00	71.96	54.24
InternVL2.5-8B	56.61	55.39	58.88	68.26	45.05	47.83	60.36	47.85	33.06	36.30	49.62	8.97	58.55	53.66	71.48	53.18
Llama-3.2-V-11B	23.41	17.10	24.22	34.54	19.56	15.20	12.00	30.46	15.16	12.96	4.36	1.90	51.01	47.65	57.60	49.47
Open-Source Models (>70B)																
Qwen2-VL-72B	40.38	61.12	65.84	69.04	28.62	47.05	78.73	51.06	13.04	38.07	68.26	6.45	61.48	59.57	73.09	52.65
InternVL2.5-78B	34.86	29.94	77.33	72.73	19.01	13.95	85.51	58.64	3.44	2.51	72.38	14.23	61.26	58.35	73.45	47.79
Llama-3.2-V-90B	25.12	23.39	55.52	49.19	12.44	12.24	65.28	50.07	1.76	1.97	40.46	12.19	51.75	49.44	68.95	51.77
Proprietary Models																
GPT-4o-mini	34.11	36.40	63.93	58.71	19.26	22.43	72.70	49.00	5.84	8.78	51.09	3.70	60.63	57.04	72.93	54.95
GPT-4o	81.78	84.14	92.50	90.12	58.25	62.36	78.47	67.12	25.35	34.48	55.63	21.28	62.91	60.69	74.18	57.95

Table 4: Detailed scores grouped by metric.

Task Overview

Your task is to extract valid question-answer-evidence (Q-A-E) triples from rebuttal sections of research papers on OpenReview. The extracted triples must meet the following criteria:

Question: Neutral, logically self-contained, and directly related to the paper's content. The question must not contain explicit citations (e.g., "Section 4.3" or "Figure 2").

Answer: The author's response must include explicit citations to the paper's main body content (e.g., "Section 4.3, Line 39" or "Figure 2, Figure 3").

Evidence: Citations in the answer must be precise and clearly formatted. Multiple references should be separated by commas.

Definitions

Question: A neutral, logically self-contained inquiry related to the paper's content. The question must: Avoid references to specific sections, lines, figures, or tables (e.g., "Can Section 4.3 be clarified?" is invalid). Focus on exploring or clarifying the main body of the paper, excluding appendices.

Answer: The full response provided by the authors, which must: Contain explicit citations to the paper's content (e.g., "Section 4.3, Line 39"). Exclude vague or general references such as "General response" or "Discussion section."

Evidence: Explicit numerical references from the author's response, such as: "Section 4.3, Line 39" "Figure 2, Figure 3" "Table 5"

Evidence must be precise and, if there are multiple references, they should be separated by commas.

JSON Output Format

```
{
    "qas": [
        {
            "question": "Extracted question text.",
            "answer": "Author's response text.",
            "evidence": "Specific reference to the paper"
        }
    ]
}
```

Table 5: Prompt for extracting explanation questions.

Task Overview
You will be provided with a portion of an academic paper, including text, images, tables, etc. Based on this content, generate multiple multiple-choice questions, each with four answer options.
Requirements for Generating Questions:
Grounding Questions from Text:
The question must be directly answerable based on the provided paragraph. Focus on extracting clear, specific, and factual details such as model performance, data, or numerical values mentioned in the text.
Examples:
- "What is the accuracy of Llama3 on the MMLU dataset?"
- "What is the main evaluation metric used for the models?"
- "Which model showed the highest accuracy on the given test?"
- "What value was reported as the accuracy of Llama3 in the study?"
Simple, Fact-based Questions:
Questions should not require external reasoning or inference. They should be straightforward and based solely on the provided content, such as factual details (e.g., accuracy, performance, test results).
Examples:
- "What is the accuracy of the Llama3 model on the MMLU benchmark?"
- "What dataset was used to evaluate the performance of the models?"
- "Which model had the lowest error rate?"
Avoid Reference to External Context: Do not refer to figures, tables, or external sections of the paper. The questions should rely solely on the provided paragraph or text. Ensure that all the information needed to answer the question is contained within the paragraph itself.
Examples:
- "What is the performance of Llama3 on the MMLU dataset?" (without referring to "Table 1" or "Figure 3")
- "What is the reported training time for the model?"
Ensure Directness and Clarity:
The question must be simple and directly related to the paragraph's content, ensuring the answer can be explicitly found in the text.
Examples:
"What performance metric is used to evaluate Model A?"
"What was the result for Model X on the validation set?"
"What is the reported accuracy for Model B?"
Refusal Field Usage:
If the provided content does not contain enough information to generate a valid question, set the **Refusal** field to **True**.
If the question meets the requirements and can be answered directly from the given paragraph, set the **Refusal** field to **False**.
Examples:
Refusal: True (If the paragraph does not contain any measurable data or clear information)
Refusal: False (If the question can be answered based on the paragraph's content)

Table 6: Prompt for generating locating questions.

You are an expert in evaluating text quality. You will receive a statement from an AI assistant's response based on a paper, along with a part from the document (which could be a text paragraph, image, or table). Your task is to carefully assess whether this statement is supported by the provided part. Please use the following scale to generate your rating:
0: No support — The statement is largely unrelated to the provided part (text, image, or table), or most key points in the statement do not align with the content of the part.
1: Partially supported — More than half of the content in the statement is supported by the part, but a small portion is either not mentioned or contradicts the part.
2: Fully supported — Most information in the statement is supported by or extracted from the part. This applies only to cases where the statement and the part are almost identical.
Ensure that you do not use any information or knowledge outside of the provided part when evaluating. Please return only the rating in JSON format, with 0, 1, or 2.
Statement: **{sentence}**

Table 7: Prompt for evaluating citation recall.

You are an expert in evaluating text quality. You will receive a statement from an AI assistant's response based on a paper, along with a part from the document (which could be a text paragraph, image, or table). Your task is to carefully assess whether the provided part contains some key information of the statement. Please use the following scale to generate your rating:
0: Unrelevant — The statement is almost unrelated to the provided part, or all key points of the statement are inconsistent with the the provided part.
1: Relevant — Some key points of the statement are supported by or extracted from the the provided part.
Ensure that you do not use any information or knowledge outside of the provided part when evaluating. Please return only the rating in JSON format, with 0 or 1.
Statement: **{sentence}**

Table 8: Prompt for evaluating citation precision.

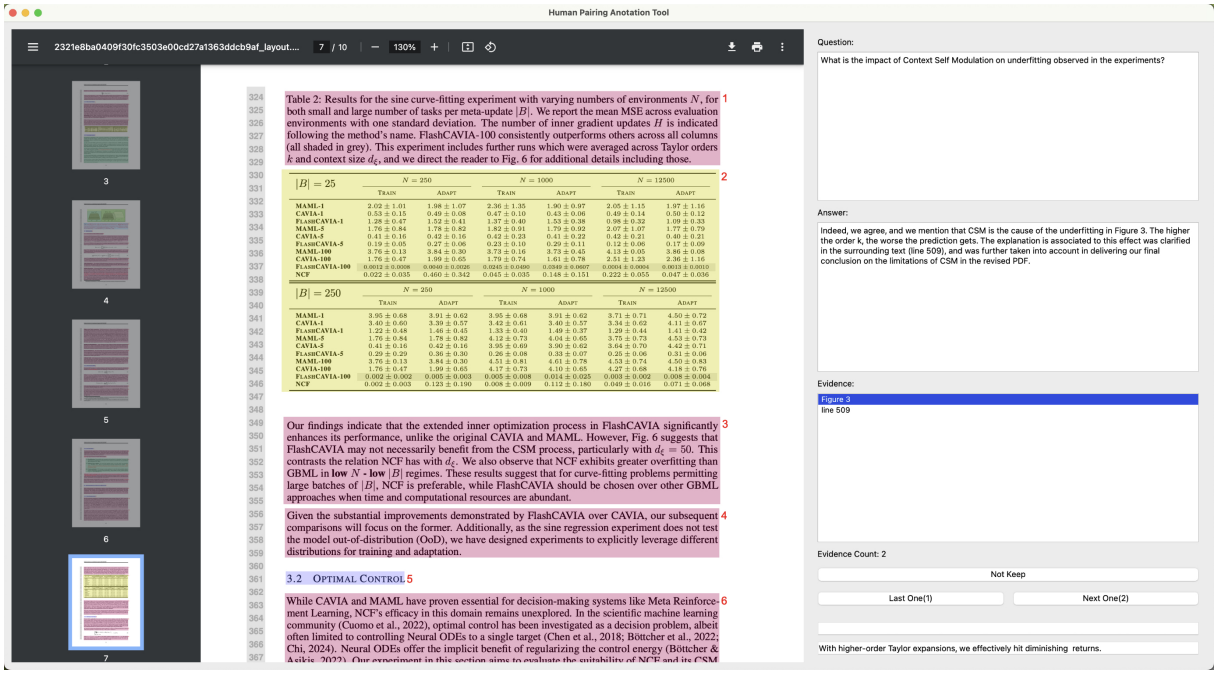

Figure 8: GUI screenshot for human annotators to map each reference to its corresponding content.

You are an assistant skilled in evaluating text quality. Please evaluate the quality of an AI assistant's response to a reviewer's question. Since the response is addressing a reviewer's inquiry regarding a paper, you need to evaluate the answer from the following dimensions:
1. **Similarity with the Author's Response**
- **Definition**: Evaluate how similar the model's response is to the author's response in terms of content, specifically whether the model's answer aligns with the key points and reasoning of the author's reply.
- **Evaluation Criteria**: If the model's response covers the main points of the author's reply and is highly similar in content, score it higher; if the model's response significantly differs from the author's content, score it lower.
2. **Completeness of the Response**
- **Definition**: Evaluate whether the model's response covers all the points raised by the reviewer and fully addresses their question.
- **Evaluation Criteria**: If the model's answer includes all key aspects raised by the reviewer and addresses the question comprehensively, score it higher; if the model misses important points or fails to address key aspects, score it lower.
3. **Logical Coherence**
- **Definition**: Evaluate whether the model's response has a clear logical structure and coherent reasoning.
- **Evaluation Criteria**: If the model's response is logically sound and the reasoning is coherent, score it higher; if there are logical flaws or incoherent reasoning, score it lower.
4. **Clarity and Expression**
- **Definition**: Evaluate whether the model's response is concise, clear, and easy to understand, and if it matches the author's language style.
- **Evaluation Criteria**: If the model's response is straightforward, logically clear, and aligns with the author's style, score it higher; if the response is lengthy, hard to understand, or deviates from the author's language style, score it lower.
Process:
1. Compare the AI assistant's answer with the reference answer, and evaluate the AI's response based on the above dimensions. After evaluating each dimension, provide a score.
2. Your scoring should be strict, and follow these guidelines:
- If the model's response is irrelevant or generates harmful content, the total score must be 0.
- If the model's response shows significant gaps compared to the reference answer or performs poorly in multiple dimensions, the score should be 1.
- If the model's response is similar to the reference answer and performs well in all dimensions, the score should be 2.
- Please return your scores in JSON format.

Table 9: Prompt for evaluating explanation questions

You are asked to evaluate the quality of the AI assistant's answers to user question as an impartial judge, and your evaluation should take into account factors including correctness (high priority), and comprehensiveness (whether the assistant's answer covers all points). Read the AI assistant's answer and compare against the reference answer, and give an overall integer rating in 0, 1, 2 (0 = wrong or irrelevant, 1 = partially correct, 2 = absolutely correct) based on the above principles, strictly in the following format: "answer_rating": 2 (where 2 is just an example). So your JSON output must have the shape "answer_rating": <integer>.

Table 10: Prompt for evaluating locating questions

Please identify the most relevant source of evidence to locate information that could address the following query. Provide your answer by selecting one of the options: A, B, C, D, or E. Begin your response with the selected letter and, if necessary, briefly explain why it is the most relevant source.
Identify where information about
{question}
can be found.
Which source is most relevant?
{options}

Table 11: Prompt used to evaluate source identification ability in the multi-choice setting.

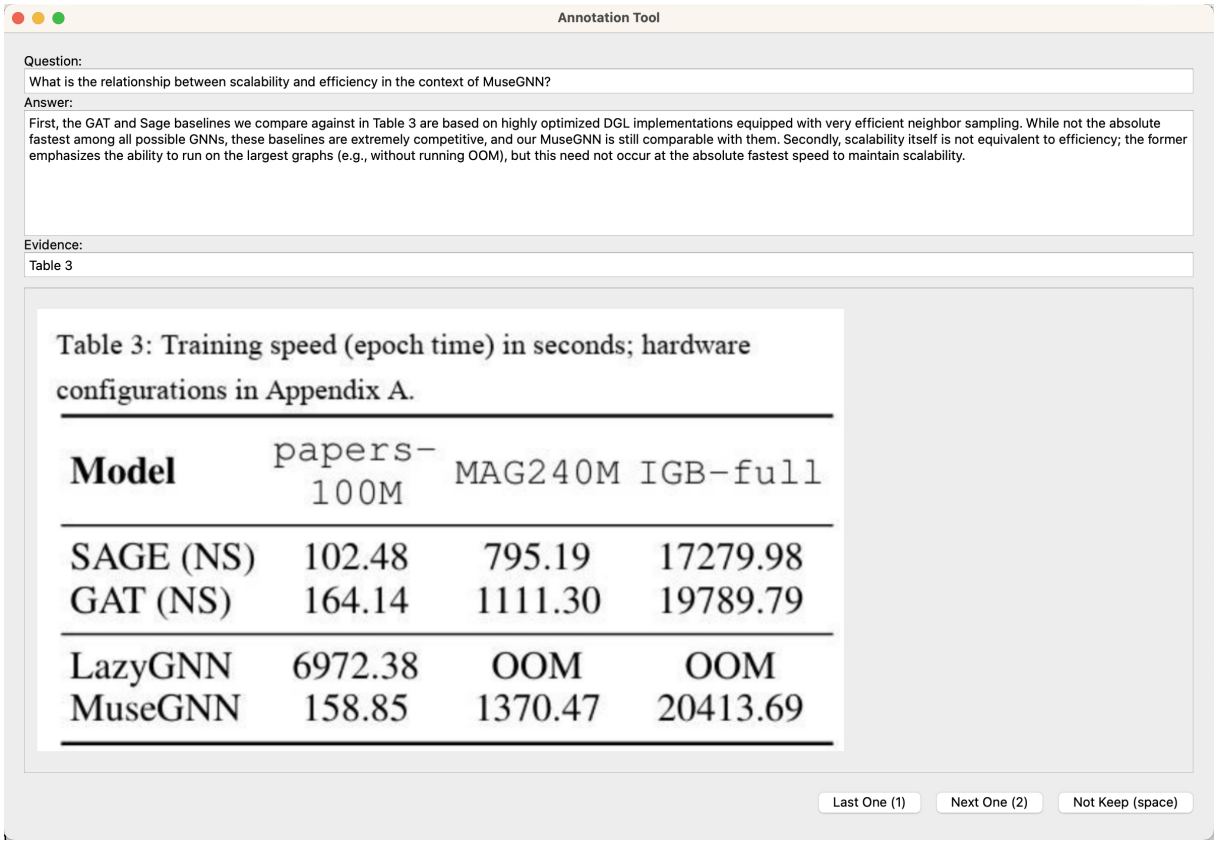

Figure 9: GUI screenshot for verifying filtered QA.

Model	Subjective			Objective		
	F1	Recall	Precision	F1	Recall	Precision
GPT-4o	0.80	0.80	0.79	0.82	0.81	0.83

Table 12: Entailment Judgment Alignment: Model vs. Human Ground Truth

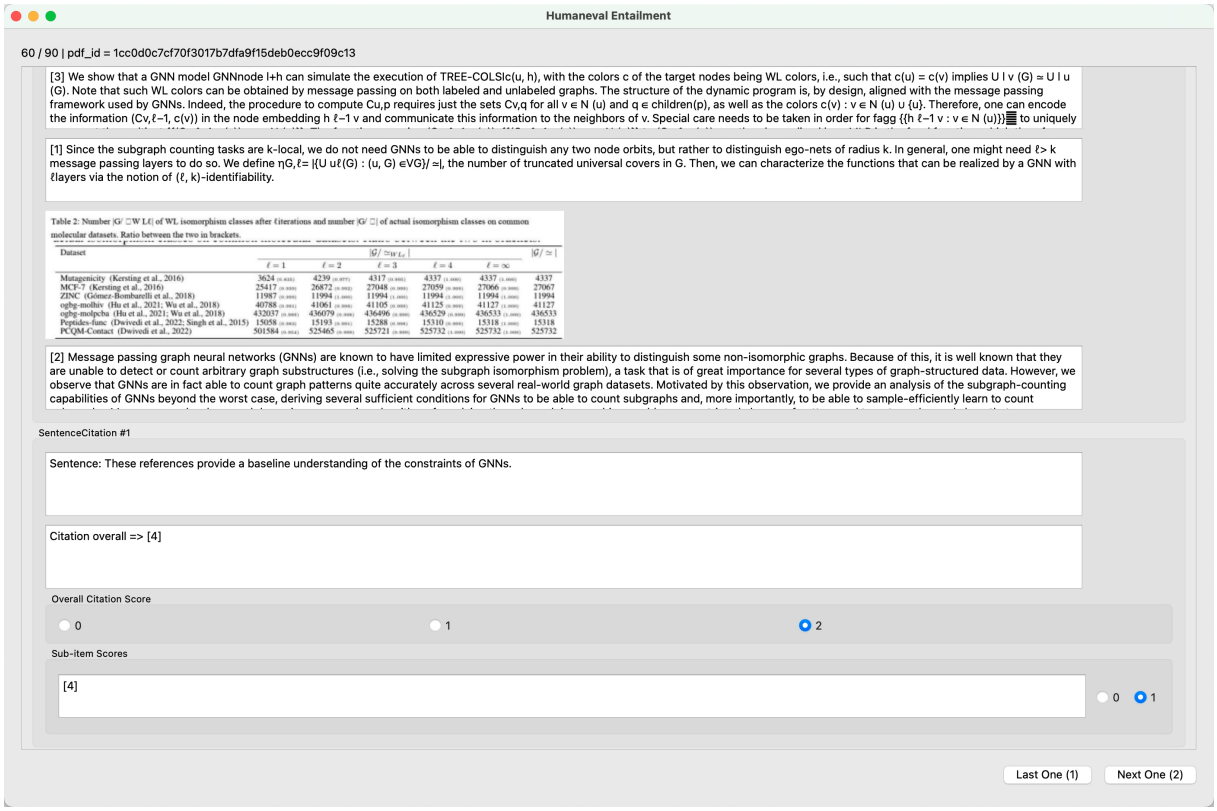

Figure 10: GUI screenshot for human-annotated entailment verification.

Model	GPT-4o Judge	DeepSeek V3 Judge
Proprietary		
GPT-4o-2024-11-20	66.72	76.11
GPT-4o-mini	64.55	72.22
Open-Source(7-14B)		
Qwen2-VL-7B-Instruct	63.77	65.00
InternVL2_5-8B	62.53	67.78

Table 13: Accuracy of citation evaluation across models under different LLM judges on a 90-sample subset.

Model	GPT-4o-mini	GPT-4o
No Cap. Acc	62.25	64.92
With Cap. Acc	64.55	66.72
Acc Impr.	**+2.30**	**+1.80**
No Cap. S-F1	48.63	67.10
With Cap. S-F1	45.34	68.19
S-F1 Impr.	**-3.29**	**+1.09**
No Cap. S-EM	28.98	38.47
With Cap. S-EM	25.34	39.11
S-EM Impr.	**-3.64**	**+0.64**
No Cap. C-F1	55.51	76.54
With Cap. C-F1	49.53	87.43
C-F1 Impr.	**-5.98**	**+10.89**

Table 14: Effect of captions on citation evaluation performance across multiple metrics. The performance with captions is compared to that without captions.