SciVer: Evaluating Foundation Models for Multimodal Scientific Claim Verification

Chengye Wang **Yifei Shen** **Zexi Kuang** **Arman Cohan** **Yilun Zhao**

Yale NLP Lab

Abstract

We introduce SciVer, the first benchmark specifically designed to evaluate the ability of foundation models to verify claims within a multimodal scientific context. SciVer consists of 3,000 expert-annotated examples over 1,113 scientific papers, covering four subsets, each representing a common reasoning type in multimodal scientific claim verification. To enable fine-grained evaluation, each example includes expert-annotated supporting evidence. We assess the performance of 21 state-of-the-art multimodal foundation models, including o4-mini, Gemini-2.5-Flash, Llama-3.2-Vision, and Qwen2.5-VL. Our experiment reveals a substantial performance gap between these models and human experts on SciVer. Through an in-depth analysis of retrieval-augmented generation (RAG), and human-conducted error evaluations, we identify critical limitations in current open-source models, offering key insights to advance models' comprehension and reasoning in multimodal scientific literature tasks.

🤗 **Data** chengyewang/SciVer
Code QDRhhhh/SciVer

1 Introduction

Scientific claim verification has become increasingly vital as the research community grapples with an ever-expanding body of scientific literature across diverse domains (Dasigi et al., 2021; Wadden et al., 2022; Lee et al., 2023; Asai et al., 2024). The accuracy of claim verification in a scientific paper is not merely a matter of cross-checking numerical consistency or validating conclusions—it necessitates a holistic understanding of the paper's context (*e.g.,* textual content, charts, and tables).

Despite the significance of multimodal reasoning, existing benchmarks in scientific claim verification have often treated these components in isolation. Predominantly, prior works have focused either on textual content alone (Wadden et al., 2022) or on verifying claims based on a single table (Lu et al., 2023). While previous multimodal question-answering (QA) benchmarks in scientific literature comprehension incorporate scientific charts, they still remain limited to QA tasks over a single chart (Li et al., 2024d; Wang et al., 2024b; Li et al., 2024c), failing to capture the broader multimodal context of scientific literature. Consequently, the lack of a comprehensive multimodal benchmark restricts the systematic evaluation of foundation models' ability to reason across the diverse and interconnected modalities in scientific literature.

In this work, we introduce **SciVer**, a comprehensive and high-quality benchmark for evaluating multimodal **SCI**entific claim **VER**ification. SciVer consists of 3,000 expert-annotated examples over 1,113 scientific papers spanning diverse domains within computer science. To ensure our benchmark reflects real-world scenarios in scientific literature comprehension, we design four fine-grained tasks (as illustrated in Figure 1): *direct reasoning*, *parallel reasoning*, *sequential reasoning*, and *analytical reasoning*. Each task targets a common reasoning type in multimodal scientific claim verification. Moreover, each example includes expert-annotated supporting evidence, facilitating fine-grained performance evaluation.

We conduct an extensive evaluation on SciVer, covering 21 frontier open-source and proprietary multimodal foundation models. Our experimental results reveal that while state-of-the-art models achieve human-comparable performance on simpler tasks (*e.g.,* direct reasoning), they continue to struggle with more complex challenges. For instance, GPT-4.1, achieves an accuracy of 70.8% on analytical reasoning, falling significantly short of human expert performance (*i.e.,* 90.0%). This demonstrates the challenging nature of SciVer. Furthermore, our analysis of retrieval-augmented generation (RAG) and human-conducted error analyses provide insights for future advancement.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8562–8579
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

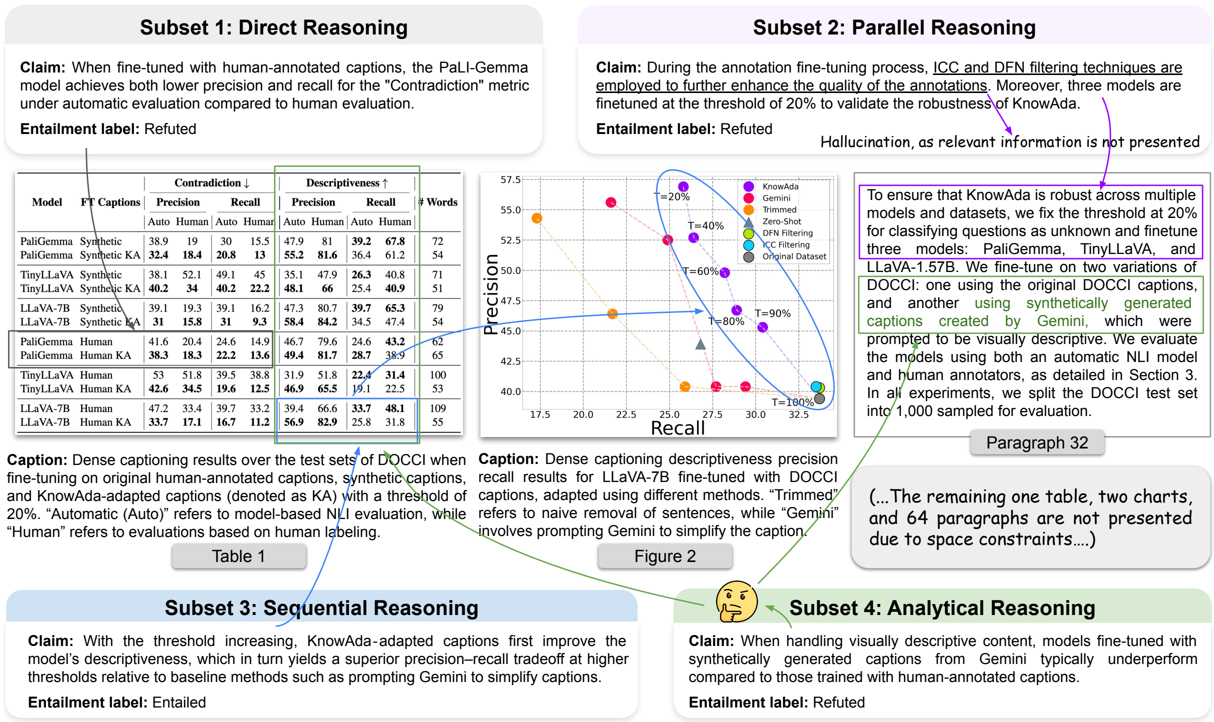


Figure 1: An illustration of the four subsets in the SCIVER benchmark. Our benchmark is designed to evaluate document-grounded scientific claim verification in a multimodal setting. To effectively perform this task, models must go through the full context of a scientific paper—including text, charts, and tables—to locate the appropriate supporting evidence before verifying a claim. The complete data examples are provided in Appendix C.

We summarize our contributions as follows:

- We introduce a new claim verification benchmark to challenge foundation models across diverse reasoning scenarios in multimodal scientific literature comprehension. Each example undergoes expert annotation and strict quality control to ensure benchmark reliability and high standards.
- We conduct an extensive evaluation that encompasses 21 open-source and proprietary foundation models, comprehensively assessing their capabilities and limitations in our task.
- We provide an in-depth analysis of Chain-of-Thought reasoning, RAG settings, and model reasoning errors, offering valuable insights for future advancements and targeted improvements.

2 Related Work

2.1 Claim Verification

Claim verification is a well-established research area that can be categorized into two main settings. The first is the open-domain setting, where an external retriever is used to fetch relevant information from a large corpus to verify claims (Vlachos and Riedel, 2014; Thorne et al., 2018; Aly et al., 2021; Wadden et al., 2022; Rangapur et al., 2024). The second is context-grounded claim verification, where claims are verified based solely on given context, without relying on external retrieval (Chen et al., 2020; Kamoi et al., 2023; Lu et al., 2023; Glockner et al., 2024; Zhao et al., 2024). This work focuses on the latter setting, as it removes variability introduced by retriever performance and enables a more controlled evaluation of foundation models' ability to verify claims within multimodal scientific context. As shown in Table 1, existing multimodal claim verification benchmarks primarily use either a single table (Chen et al., 2020; Gupta et al., 2020; Lu et al., 2023) or single chart (Akhtar et al., 2024) as input context. In real-world scenarios, however, verifying claims in scientific literature requires reasoning across multiple modalities, including textual descriptions, tables, and figures.

2.2 Scientific Literature Comprehension

With the rapid expansion of research publications, evaluating and applying foundation models for scientific literature comprehension has become increasingly important (Asai et al., 2024; Skarlinski et al., 2024; Li et al., 2024b). Existing benchmarks primarily focus on QA tasks, assessing models on

Dataset	Input Context	Data Construction	# Task / Subsets	Rationale Annotation?
		Scientific Literature Comprehension		
QASPER (Dasigi et al., 2021)	Single NLP paper	Expert annotation	4	Evidence
QASA (Lee et al., 2023)	Single AI/ML paper (text-only)	Expert annotation	3	Evidence
MMSci (Li et al., 2024d)	Multiple figures or charts from STEM papers	GPT-4o generation	2	✗
ArXivQA (Li et al., 2024c)	Single chart from arXiv papers	GPT-4V generation	–	✗
CharXiv (Wang et al., 2024b)	Single Chart from arXiv papers	Expert annotation	2	✗
SCIFACT (Wadden et al., 2020)	Multiple STEM paper abstracts	Expert annotation	–	–
		Claim Verification over Multimodal Context		
INFOTABS (Gupta et al., 2020)	Single wikipedia table	Crowdsourcing	–	✗
TABFACT (Chen et al., 2020)	Single wikipedia table	Crowdsourcing	2	✗
ChartCheck (Akhtar et al., 2024)	Single wikipedia chart	Crowdsourcing	2	Rationale
SCITAB (Lu et al., 2023)	Single scientific table from NLP&ML paper	Expert+InstructGPT	–	✗
SCIVER (ours)	Multiple tables, charts, paragraphs from CS papers	Expert annotation	4	Evidence

Table 1: Comparison of SCIVER with existing claim verification and scientific literature comprehension benchmarks.

their ability to extract or infer information from scientific papers (Dasigi et al., 2021; Lee et al., 2023). While recent efforts have extended QA tasks to incorporate tabular and visual information (Li et al., 2024c; Wang et al., 2024b; Li et al., 2024d), they remain constrained by their single-modality focus, neglecting the rich multimodal context inherent in scientific papers. Claim verification, on the other hand, demands a more comprehensive understanding of scientific literature, as claims are often supported by a combination of textual descriptions, tables, and charts. Additionally, each example in SCIVER includes detailed supporting evidence, facilitating fine-grained evaluation.

3 SCIVER Benchmark

SCIVER is a comprehensive evaluation framework designed to assess the ability of foundation models to verify scientific claims within a multimodal context. Figure 2 provides an overview of the SCIVER construction pipeline. In the following subsections, we detail the benchmark design, data construction process, and quality validation methodology.

3.1 Benchmark Design

We first present the task formulation and the four specialized subsets of our dataset that we designed to evaluate different aspects of model performance.

Task Formulation. We formally define the task of SCIVER within the context of a foundation model FM as follows: Given a scientific claim c and multimodal contexts $\{P, I, T\}$ collected from a scientific paper—where P denotes textual paragraphs, I denotes multiple charts, and T denotes multiple tables—the model is is tasked with determining the entailment label $\ell \in \mathcal{L} = \{\text{“entailed”}, \text{“refuted”}\}$:

$$\ell = \arg\max_{\ell \in \mathcal{L}} P_{\mathbf{FM}}(\ell \mid c, P, I, T) \quad (1)$$

It challenges foundation models to perform complex reasoning by integrating and interpreting textual, tabular, and visual data to verify scientific claims. Since scientific tables often have intricate structures that are difficult to represent in textual format, we follow recent work in multimodal table understanding (Zheng et al., 2024; Deng et al., 2024) by using table screenshots as inputs.

Subset Design. SCIVER includes the following four distinct subsets, each designed to evaluate a specific reasoning type commonly required for scientific claim verification over multimodal context:
(1) *Direct Reasoning*, which evaluates models' ability to extract and interpret a single piece of information to verify a scientific claim.
(2) *Parallel Reasoning*, which evaluates models' ability to simultaneously process and integrate information from multiple distinct sources.
(3) *Sequential Reasoning*, which evaluates models' ability to perform step-by-step inference chains across different modalities. Models are required to establish logical connections between multiple pieces of evidence, where each step's conclusion becomes a premise for subsequent reasoning steps.
(4) *Analytical Reasoning*, which evaluates models' ability to verify claims that require both sophisticated domain knowledge and complex reasoning beyond direct data extraction. Models must not

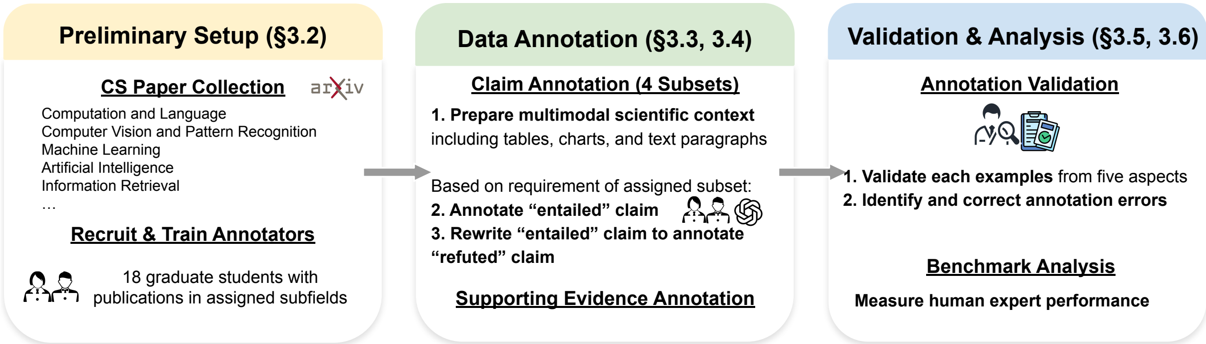


Figure 2: An overview of the SciVer benchmark construction pipeline.

only interpret the provided data but also apply relevant scientific principles and methodological understanding to arrive at valid conclusions.

Appendix C presents detailed examples of each subset. These subsets enable fine-grained evaluation across different reasoning paradigms commonly encountered in scientific literature comprehension.

3.2 Preliminary Setup

We next discuss the preliminary setup for data construction, including the process of scientific paper collection and expert annotator recruitment.

Expert Annotator Recruitment and Training. Existing claim verification datasets primarily rely on crowdsourced data curation (as shown in Table 1). However, our preliminary study suggests that crowd-sourced annotators often lack the necessary domain expertise for our task. To mitigate this, we recruit 18 CS graduate students with relevant subject-specific knowledge, requiring each to have at least two peer-reviewed publications in their assigned subfields. Detailed annotator biographies are provided in Table 5 in Appendix. To further enhance annotation quality and consistency, all selected experts undergo a mandatory two-hour individual training session with one of the authors, ensuring that they are familiar with the annotation guidelines and protocol.

Scientific Paper Collection. SciVer focuses on arXiv papers published between September 1, 2024, and November 15, 2024, covering eight key areas of *computer science*. To ensure high-quality content, we prioritize papers that include comments indicating acceptance by a peer-reviewed venue. For each paper, we extract its multimodal context—including textual content, tables, and charts—from the HTML versions available on the arXiv platform. We filter out papers that contain fewer than two tables or two charts.

3.3 Claim Annotation

Given a paper relevant to their research field, the annotators follow these steps for claim annotation:

Multimodal Scientific Context Preparation. Scientific papers are often lengthy, exceeding the maximum context length of certain foundation models. Including the full text may overwhelm these models and hinder their ability to integrate information effectively across modalities. To address this, annotators refine the paper context by removing textual sections that are not essential to understanding the core research problem, such as related work, acknowledgments, references, and appendix sections.

Entailed Claim Annotation. To reduce bias stemming from the positioning of evidence, the annotation interface randomly selects three charts or tables from the curated context, along with their surrounding textual paragraphs. Annotators are then tasked with writing an entailed claim that aligns with the pre-given reasoning types (*i.e.,* subset). They are required to ensure that verifying the claim requires referencing at least one of the three sampled multimodal elements. Subsequently, annotators identify all relevant supporting evidence, which is later reviewed by a second annotator.

Refuted Claim Annotation. Following established practices in the field (Wadden et al., 2022; Chen et al., 2020; Lu et al., 2023), and given the difficulty of directly obtaining *"refuted"* claims, we instead generate them by perturbing original *"entailed"* claims through a semi-automated annotation process. Specifically, to curate *"refuted"* claims, annotators modify the initially annotated *"entailed"* claim by introducing factual errors that contradict the supporting evidence.

Property (avg.)	**Val**	**Test**
Multimodal Scientific Context		
# Words in text paragraphs	583.6	567.4
# Tables	0.55	0.54
Table caption length	14.2	13.7
# Charts	0.94	0.95
Chart caption length	39.2	40.2
Claim Verification		
Claim length	30.5	33.9
# Entailed	505	995
# Refuted	495	1,005
Supporting Evidence	2.63	2.62
Scientific papers	327	786
Total examples	1,000	2,000

Table 2: Data statistics of SciVer.

Adopted Chain-of-Thought Prompt

{Paper Context (textual paragraphs, tables, charts)}

You are given a multimodal scientific context that includes textual paragraphs, tables, and charts. Your task is to determine whether the given claim is Entailed or Refuted. Be skeptical and cautious: if there is any inconsistency, missing evidence, or ambiguity, consider the claim incorrect.

Claim to verify:
{Claim}

Start by explaining your reasoning process clearly, focusing on identifying potential contradictions, lack of support, or misleading interpretations. Think step by step before answering.

Figure 3: The Chain-of-Thought prompt used.

3.4 Supporting Evidence Annotation

After completing the claim annotation, a second annotator, who is also an expert in the relevant research field, is tasked with annotating the supporting evidence. The annotators are required to carefully review the claim and identifying all relevant paragraphs, tables, and charts that serve as supporting evidence. To ensure consistency and accuracy, we compare the supporting evidence and entailment label annotated in this step with those from the initial claim annotation. If discrepancies arise between the two annotations, a third expert annotator is introduced to adjudicate the differences. Our process achieves an *inter-annotator agreement* of 94.0% for entailment label annotation, demonstrating strong reliability in our annotation.

3.5 Data Validation

Each annotated example undergoes a comprehensive validation process conducted by a different expert annotator within the same research field. The validation focuses on the following five aspects: (1) The claim must be grammatically correct, well-structured, and free of spelling or typographical errors. (2) The claim must align with the annotation requirements of its corresponding subset and should not be verifiable using textual context alone. (3) The claim must be meaningfully situated within the paper context and hold practical significance for scientific literature comprehension. (4) The annotated supporting evidence must be directly relevant to the claim and comprehensive enough to support claim verification without requiring additional, unannotated context.

If an example fails to meet any of these criteria, validators are responsible for making necessary revisions. In practice, 232 initially annotated examples required revisions before being finalized.

3.6 Data Statistics and Analysis

Table 2 presents the data statistics of SciVer. It is randomly divided into the validation and test sets. The validation set contains 1,000 examples and is intended for model development and validation. The test set comprises the remaining 2,000 examples and is designed for standard evaluation.

To approximate **human-expert-level performance** on SciVer, we randomly sampled 10 claims from each subset, totaling 40 claims. Two expert annotators independently evaluated these claims, providing the natural language explanation and final entailment label for each claim. They achieve an average accuracy of 93.8% (Table 3).

4 Experiment

This section first outlines the experiment setup, and then discusses our experiment results and analysis.

4.1 Experiment Setup

We use accuracy as the primary metric to evaluate model performance on SciVer. Following recent benchmark studies (Yue et al., 2024, 2025), we adopt rule-based methods to derive the final entailment label from the model response, which is then compared to the ground-truth label.

We evaluate a broad range of frontier foundation models that support *multiple images* and text as input. Specifically, we evaluate **11 series of open-source models**, including InternVL-2, 2.5,

	Release	Test Set				Avg. Validation	Avg. Test
		Direct	Parallel	Sequential	Analytical		
Baseline Settings							
Human Expert		100.0	95.0	90.0	90.0	–	93.8
Random Guess		50.0	50.0	50.0	50.0	50.0	50.0
Proprietary Models							
o4-mini	2025-04	85.0	80.6	77.6	67.6	79.6	77.7
Gemini-2.5-Flash	2025-05	79.8	76.0	73.2	71.4	76.0	75.1
GPT-4o	2024-11	77.0	71.2	73.6	73.8	72.3	73.9
Gemini-2.0-Flash	2025-02	78.0	72.2	69.4	73.4	73.0	73.3
GPT-4.1	2025-04	77.6	73.2	71.2	70.8	74.3	73.2
GPT-4o-mini	2024-07	71.4	67.6	61.4	62.0	63.8	65.6
Open-source Models							
Mistral-Small-3.1-24B	2025-03	74.8	66.0	68.6	75.6	73.6	71.3
Qwen2.5-VL-72B	2025-01	70.8	69.2	68.2	69.2	70.2	69.4
InternVL3-38B	2025-04	65.8	64.6	65.2	70.4	70.6	66.5
Qwen2-VL-72B	2024-11	70.4	61.0	63.0	67.2	65.9	65.4
InternVL2.5-38B	2024-11	65.0	55.8	62.4	66.8	63.8	62.5
Pixtral-12b	2024-09	60.8	54.6	63.4	65.2	61.1	61.0
InternVL3-8B	2025-04	64.2	54.6	56.0	63.0	58.8	59.5
Qwen2.5-VL-7B	2025-01	55.8	57.4	57.0	60.2	53.5	57.6
InternVL2.5-8B	2024-11	53.8	56.4	53.2	58.2	55.5	55.4
InternVL2-8B	2024-06	54.0	52.6	50.2	54.6	52.9	52.9
Qwen2-VL-7B	2024-11	52.6	54.0	52.0	52.0	52.8	52.7
Llama-3.2-11B-Vision	2024-09	53.6	50.6	51.8	53.2	48.9	52.3
Phi-4-Multimodal	2025-03	50.8	50.8	51.2	51.0	52.1	51.0
LLaVA-OneVision	2024-09	49.8	48.2	49.6	53.6	51.0	50.3
Phi-3.5-Vision	2024-08	46.0	52.0	48.0	49.2	51.5	48.8

Table 3: Model accuracy on SciVer validation and test sets with CoT prompts, ranked by test set performance.

and 3 (Chen et al., 2023, 2024b,a), Qwen2-VL and Qwen2.5-VL (Bai et al., 2023; Wang et al., 2024a), Pixtral (Agrawal et al., 2024), Mistral-Small-3.1 (Mistral AI, 2025), LLaVA-OneVision (Li et al., 2024a), Llama-3.2-Vision (Meta, 2024), Phi-3.5-Vision and Phi-4-Multimodal (Microsoft, 2024; Microsoft et al., 2025). We also evaluate **five series of proprietary models**, including OpenAI o4-mini (OpenAI, 2025a), GPT-4o and GPT-4.1 (OpenAI, 2024, 2025b), Gemini-2.0 and Gemini-2.5 (Google, 2024, 2025). Appendix B details the parameter settings and configurations of the evaluated models. For open-source models, we utilize the vLLM pipeline (Kwon et al., 2023) for model inference; while for proprietary models, we use their official API service.

We evaluate the models with the **Chain-of-Thought** prompt, which is presented in Figure 3.

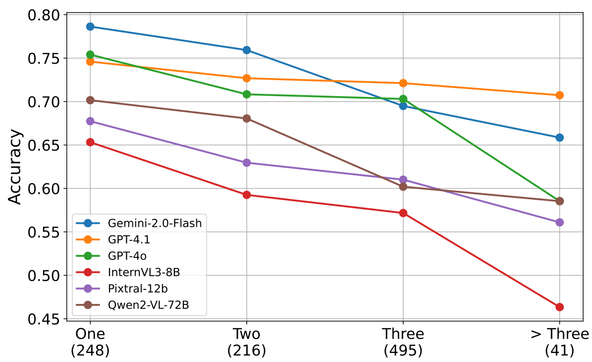


Figure 4: Comparison of model performance on the validation set, with claims requiring varying amounts of annotated supporting evidence. Each piece of evidence is defined as a single table, chart, or paragraph (§3.4).

4.2 Main Findings

Table 3 presents the evaluated models' performance. Our main findings are as follows:

SciVer presents substantial challenges for current models. While the recently released reasoning models, o4-mini and Gemini-2.5-Flash, demonstrate leading performance, other models fall short of human expert capabilities. For instance, GPT-4.1 achieves 73.2% accuracy with CoT prompting, considerably lower than the 93.8% accuracy achieved

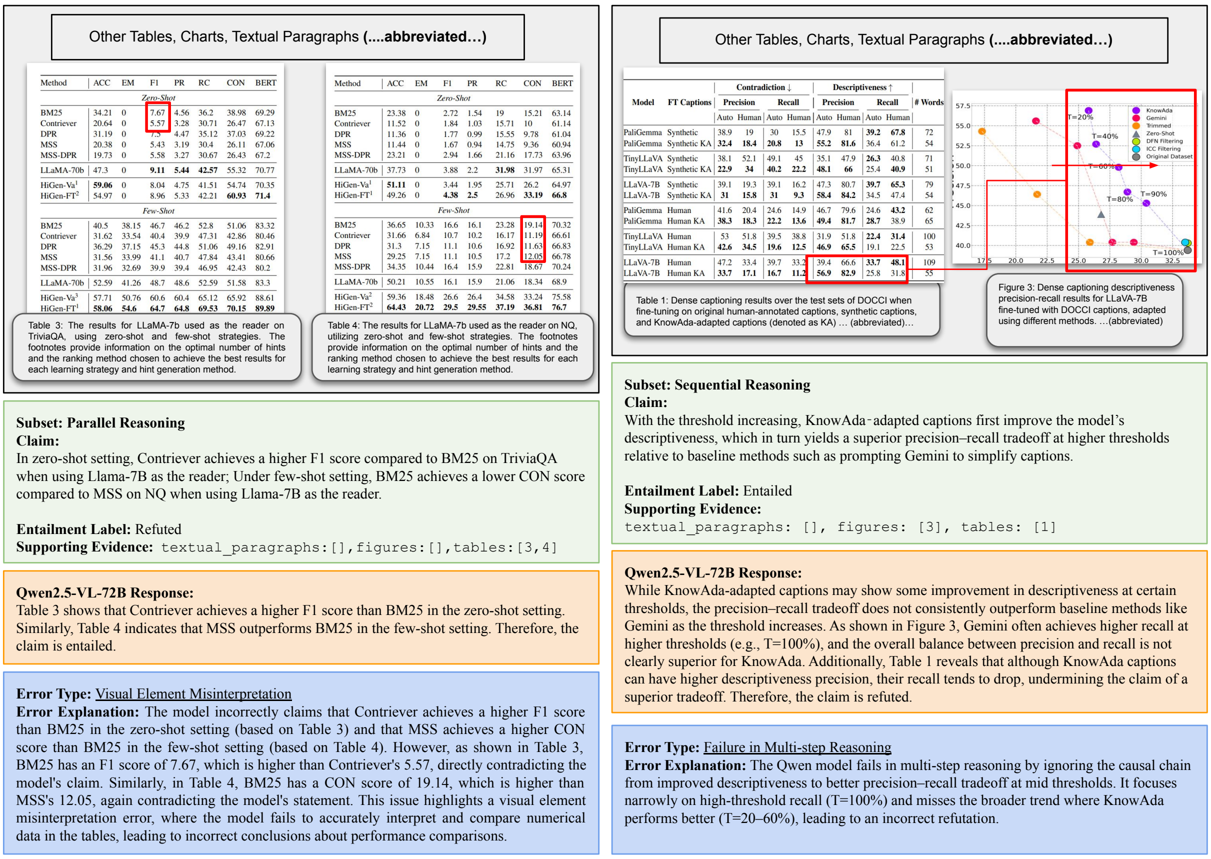


Figure 5: Illustration of two error types: Visual Element Misinterpretation (left) and Failure in Multi-step Reasoning (right). Additional error examples are provided in Appendix C.

by human experts. This performance gap highlights SciVer 's crucial role in advancing and assessing the capabilities of models in multimodal scientific literature comprehension.

Performance of open-sourced models. Open-source models continue to lag behind their proprietary counterparts. However, models such as Mistral-Small-3.1, Qwen2.5-VL, and InternVL3 have achieved competitive performance, narrowing the gap with top proprietary models. These advancements highlight the rapid progress in open-source development. In the following subsections, we provide a detailed analysis of open-source models and offer insights for future improvements.

Model performance declines with increasing evidence requirements. To provide a fine-grained analysis of model performance on multi-hop reasoning in SciVer, we compare frontier models on the validation set across claims that require different numbers of annotated supporting evidence. As shown in Figure 4, model performance consistently declines as the number of ground-truth evidence pieces increases. This trend suggests that current models struggle with multi-hop reasoning and with synthesizing information across multiple multimodal contexts.

4.3 Error Analysis and Case Study

To better understand the limitations of open-source models, we perform a detailed error analysis on Qwen2.5-VL-72B. We randomly select 25 instances from each of the four subsets for evaluation. Through a detailed inspection of model response, we identify five common error types:

- **Failure to Retrieve Relevant Information** (32%), where models fail to retrieve and consider all the key evidence from the provided multimodal context, leading to incomplete reasoning or incorrectly classify verifiable claims as lacking enough information.
- **Visual element misinterpretation** (21%), where models misinterpret charts or tables.
- **Failure in multi-step reasoning** (17%), where models struggle to connect intermediate reasoning steps over extracted information, leading to incorrect entailment predictions.

Prompt for Evidence Filtering

{Single Multimodal Element}

Analyze the given context and determine whether it contains relevant information to verify the following claim: {Claim}
Respond with either "yes" if the context contains the necessary information to verify the claim, or "no" if it does not.

Figure 6: The prompt for evidence filtering in §4.4.

Setting	Recall@5	4o-mini	Qwen2.5-VL
Original	–	63.8	70.2
with RAG			
Contriever	70.7	64.7 ↑0.9	71.8 ↑1.6
BM25	74.3	65.4 ↑1.6	72.2 ↑2.0
OAI Embedding	81.0	67.0 ↑3.2	72.9 ↑2.7
Oracle	–	73.3 ↑9.5	75.3 ↑5.1
LLM Evidence Filter	–	67.5 ↑3.7	74.4 ↑4.2

Table 4: Performance comparison of GPT-4o-mini and Qwen2.5-VL-72B under different RAG settings.

- **Heavy reliance on text modality** (12%), where models focus primarily on textual input, failing to properly integrate crucial information from tables and charts.
- **Domain-specific misconceptions** (10%), where models misapply domain terminology or rely on irrelevant memorized knowledge when verifying the given claims.
- **Other observed errors** include incorrect numerical computations and instances where models refuse to generate a response.

For each error type, we provide illustrative examples and corresponding error analyses in Figure 5 and Appendix C.

4.4 Retrieval-Augmented Generation Analysis

The preceding error analysis highlights that the failure to retrieve relevant information is a primary error type. This finding motivates us to explore how RAG settings can be leveraged to improve model performance on SciVer.

Experiment Setup. Implementing RAG for scientific multimodal data presents challenges, as existing open-source retrieval models do not natively support scientific tables and charts. To overcome this limitation, we construct the textual representations for tables and charts as the *concatenation of their original captions and GPT-4o-generated descriptions*. Each representation is indexed as separate evidence alongside the textual paragraphs extracted from the paper. We evaluate three widely used retrieval systems, *i.e.,* BM25, Contriever (Izacard et al., 2021), and OpenAI's text-embedding-3-large, to retrieve the top-5 most relevant evidence for the given claim. The retrieved evidence is then fed into the model in its original form. Additionally, we assess an alternative setting (*i.e.,* Evidence Filtering) where the model first determines, one by one, whether each piece of evidence is relevant to the claim (prompt shown in Figure 6), and then incorporates all confirmed relevant evidence into the final input.

Findings. We evaluate the GPT-4o-mini and Qwen2.5-VL-72B models on the validation set. As shown in Table 4, enhancements in information retrieval quality generally lead to improved entailment classification performance on SciVer. Among the three retrievers tested, the OpenAI embedding model achieves the highest retrieval accuracy, which correlates with the most substantial gains in downstream LLM performance (*i.e.,* 70.2% → 75.3% for Qwen2.5-VL-72B). Additionally, applying an LLM-based evidence filter further boosts overall system performance.

5 Conclusion

This work introduces SciVer, a comprehensive benchmark for evaluating multimodal scientific claim verification. By providing a diverse set of fine-grained, expert-curated examples and a reliable automated evaluation system, SciVer advances the development of foundation models capable of accurately and robustly interpreting real-world scientific texts, tables, and figures. Our experimental results expose significant performance gaps between state-of-the-art foundation models and human experts, revealing key challenges such as reasoning limitations across textual, tabular, and visual data, as well as difficulties in retrieving and integrating relevant multimodal evidence.

Acknowledgement

We are grateful to Google TRC program for providing computing resources and Together AI for granting LLM API credits.

Limitations

While SciVer presents a significant advancement in multimodal scientific claim verification, there are several limitations that we acknowledge, which also point to promising directions for future research. First, SciVer is primarily constructed from computer science papers sourced from arXiv, focusing on verifying claims within this discipline. While this allows us to control for domain expertise in our annotation process and ensures high-quality claim verification, it may limit the generalizability of SciVer to other fields. Second, SciVer primarily focuses on claim verification over textual paragraphs, tables, and charts, as these are the most common multimodal elements in scientific literature. However, some domains rely heavily on other modalities such as equations, figures, or experimental images, which SciVer does not explicitly consider in its current version. Third, SciVer relies on expert annotations with domain expertise, ensuring high-quality labels and reasoning rationales. However, this approach is labor-intensive and may not scale easily to larger datasets.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms.

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. *arXiv preprint arXiv:2308.12966*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

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. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

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.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 407–426, Bangkok, Thailand. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Google. 2025. Gemini 2.5 flash.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

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

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024b. M3SciQA: A multi-modal multi-document scientific QA benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15419–15446, Miami, Florida, USA. Association for Computational Linguistics.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024c. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. 2024d. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Meta. 2024. The llama 3 herd of models.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Mistral AI. 2025. Mistral-small-3.1-24b-instruct-2503. `https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503`. Apache 2.0 License.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2025a. Addendum to openai o3 and o4-mini system card: Openai o3 operator.

OpenAI. 2025b. Introducing gpt-4.1 in the api.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2024. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *CoRR*, abs/2406.18521.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA. Association for Computational Linguistics.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

A SciVer Benchmark Construction

ID	Biography	Assigned Subjects	# Relevant Publications	Author?
1	2nd year PhD	Computer Vision and Pattern Recognition	1-5	✗
2	Final year PhD	Computer Vision and Pattern Recognition	5-10	✗
3	Postdoc	Computer Vision and Pattern Recognition	>10	✗
4	–	Computation and Language	>10	✓
5	–	Computation and Language	1-5	✓
6	–	Computation and Language	1-5	✓
7	3rd year PhD	Robotics	5-10	✗
8	Postdoc	Robotics	>10	✗
9	Final year PhD	Software Engineering	5-10	✗
10	Postdoc	Software Engineering	>10	✗
11	2nd year PhD	Machine Learning	1-5	✗
12	4th year PhD	Machine Learning	5-10	✗
13	3rd year PhD	Artificial Intelligence	5-10	✗
14	Postdoc	Artificial Intelligence	>10	✗
15	Master Student	Information Retrieval	1-5	✗
16	3rd year PhD	Information Retrieval	5-10	✗
17	Final year PhD	Cryptography	5-10	✗
18	Postdoc	Cryptography	>10	✗

Table 5: Biographies of 18 expert annotators involved in SciVer construction (Author biographies are hidden to protect identity confidentiality.

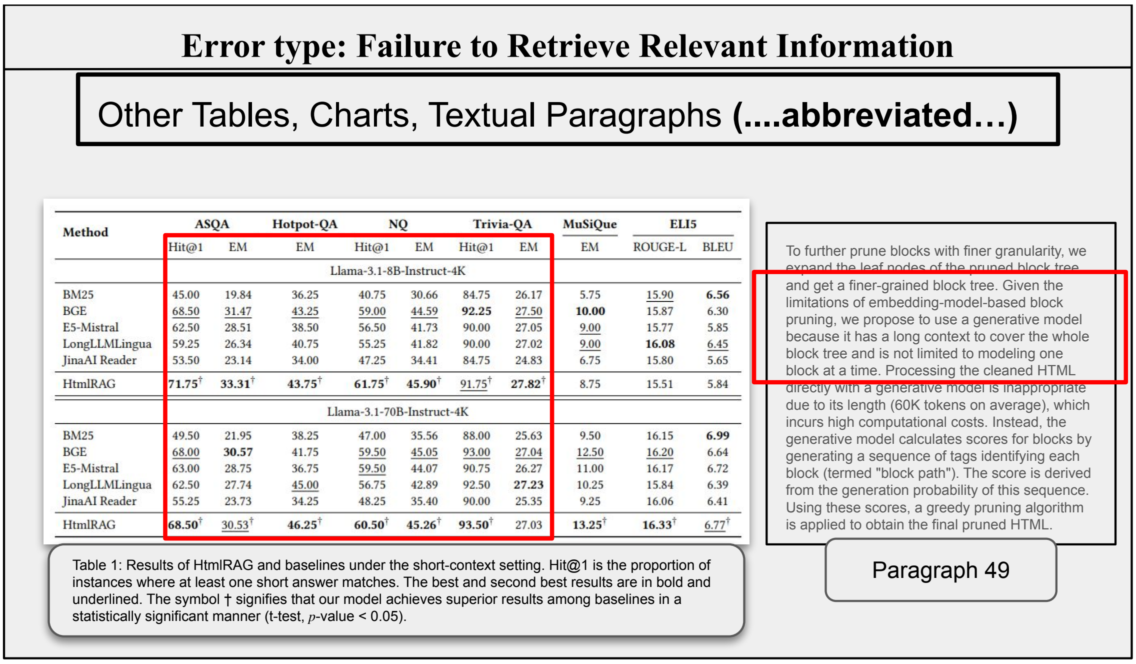
B Configurations of Evaluated Models

Organization	Model	Release	Version	# Inference Pipeline
		Proprietary Models		
OpenAI	o4-mini*	2025-04	`o4-mini-2025-04-16`	API
	GPT-4.1	2025-04	`gpt-4.1-2025-04-14`	
	GPT-4o	2024-08	`gpt-4o-2024-08-06`	
	GPT-4o-mini	2024-07	`gpt-4o-mini-2024-07-18`	
Google	Gemini-2.5-Flash	2025-05	`gemini-2.5-flash-preview-05-20`	API
	Gemini 2.0 Flash	2024-12	`gemini-2.0-flash-exp`	
		Open-source Multimodal Foundation Models		
Alibaba	Qwen2.5-VL-72B	2025-01	`Qwen2.5-VL-72B-Instruct`	vLLM
	Qwen2-VL-72B	2024-09	`Qwen2-VL-72B-Instruct`	
	Qwen2.5-VL-7B	2025-01	`Qwen2.5-VL-7B-Instruct`	
	Qwen2-VL-7B	2024-08	`Qwen2-VL-7B-Instruct`	
Mistral AI	Mistral-Small-3.1	2025-03	`Mistral-Small-3.1-24B`	vLLM
	Pixtral-12B	2024-09	`Pixtral-12B-2409`	
Shanghai AI Lab	InternVL-3-38B	2025-04	`InternVL-3-38B`	vLLM
	InternVL3-8B	2025-04	`InternVL3-8B`	
	InternVL2.5-38B	2024-11	`InternVL2.5-38B`	
	InternVL2.5-8B	2024-11	`InternVL2.5-8B`	
	InternVL2-8B	2024-06	`InternVL2-8B`	
Meta	Llama-3.2-11B-Vision	2024-09	`Llama-3.2-11B-Vision-Instruct`	vLLM
Microsoft	Phi-3.5-Vision	2024-07	`Phi-3.5-Vision-Instruct`	vLLM
	Phi-4-Multimodal	2025-03	`Phi-4-Multimodal`	
Llava Hugging Face	LLaVA-OneVision-7B	2024-09	`llava-onevision-qwen2-7b-ov-chat-hf`	vLLM

Table 6: Details of the multimodal foundation models evaluated in our study. Models are organized by organization and aligned with performance data from the main text.

C Error Analysis

C.1 Failure to Retrieve Relevant Information



Method	ASQA		Hotpot-QA	NQ		Trivia-QA		MuSiQue	ELI5	
	Hit@1	EM	EM	Hit@1	EM	Hit@1	EM	EM	ROUGE-L	BLEU
Llama-3.1-8B-Instruct-4K										
BM25	45.00	19.84	36.25	40.75	30.66	84.75	26.17	5.75	15.90	**6.56**
BGE	68.50	31.47	43.25	59.00	44.59	**92.25**	27.50	**10.00**	15.87	6.30
E5-Mistral	62.50	28.51	38.50	56.50	41.73	90.00	27.05	9.00	15.77	5.85
LongLLMLingua	59.25	26.34	40.75	55.25	41.82	90.00	27.02	9.00	**16.08**	6.45
JinaAI Reader	53.50	23.14	34.00	47.25	34.41	84.75	24.83	6.75	15.80	5.65
HtmlRAG	**71.75**†	**33.31**†	**43.75**†	**61.75**†	**45.90**†	91.75†	**27.82**†	8.75	15.51	5.84
Llama-3.1-70B-Instruct-4K										
BM25	49.50	21.95	38.25	47.00	35.56	88.00	25.63	9.50	16.15	**6.99**
BGE	68.00	**30.57**	41.75	59.50	45.05	93.00	27.04	12.50	16.20	6.64
E5-Mistral	63.00	28.75	36.75	59.50	44.07	90.75	26.27	11.00	16.17	6.72
LongLLMLingua	62.50	27.74	45.00	56.75	42.89	92.50	**27.23**	10.25	15.84	6.39
JinaAI Reader	55.25	23.73	34.25	48.25	35.40	90.00	25.35	9.25	16.06	6.41
HtmlRAG	**68.50**†	30.53†	**46.25**†	**60.50**†	**45.26**†	**93.50**†	27.03	**13.25**†	**16.33**†	6.77†

Table 1: Results of HtmlRAG and baselines under the short-context setting. Hit@1 is the proportion of instances where at least one short answer matches. The best and second best results are in bold and underlined. The symbol † signifies that our model achieves superior results among baselines in a statistically significant manner (t-test, p-value < 0.05).

Subset: Analytical Reasoning
Claim: HtmlRAG excels across multiple datasets compared to baseline methods, with its generative pruning strategy designed to improve the model's capability to handle multimodal contexts.

Entailment Label: Entailed
Supporting Evidence:
`textual_paragraphs: [49], figures: [], tables: [1]`

Explanation:
The claim, "HtmlRAG excels across multiple datasets compared to baseline methods, with its generative pruning strategy designed to improve the model's capability to handle multimodal contexts," is entailed because it is supported by both the table and the textual paragraph. Table 1 provides quantitative evidence showing HtmlRAG consistently outperforms baselines like BM25, BGE, and E5-Mistral on metrics such as Hit@1 and EM across datasets like ASQA, Hotpot-QA, and NQ. The textual paragraph (Paragraph 49) explains the generative fine-grained block pruning strategy, which enhances HtmlRAG's ability to process HTML globally and retain relevant information. Together, these sources confirm that HtmlRAG's superior performance is directly tied to its innovative pruning methodology.

Qwen2.5-VL-72B Response:

While Table 1 provides quantitative performance metrics showing HtmlRAG's superiority in Hit@1 and EM across datasets, it does not directly connect this performance to the generative pruning strategy described in Paragraph 49. The textual paragraph explains how the generative pruning strategy works and its potential benefits for handling multimodal contexts, but it does not explicitly link this methodology to the specific performance improvements shown in the table. Without an explicit connection between the pruning strategy and the quantitative results, the evidence appears insufficient to fully verify the claim. Therefore, this claim is classified as refuted.

Error Explanation:
While both the textual paragraph and Table 1 were retrieved, the model failed to consider all key evidence together. The textual paragraph explains how HtmlRAG achieves better results (methodological improvements), while Table 1 provides quantitative proof of its superior performance. The model's inability to synthesize these two pieces of evidence led to an incomplete reasoning process and an incorrect classification of the claim as refuted.

Figure 7: Illustration of *Failure to Retrieve Relevant Information* with the example from the *Analytical Reasoning* subset.

C.2 Visual element misinterpretation

Error type: Visual Element Misinterpretation

Other Tables, Charts, Textual Paragraphs **(....abbreviated...)**

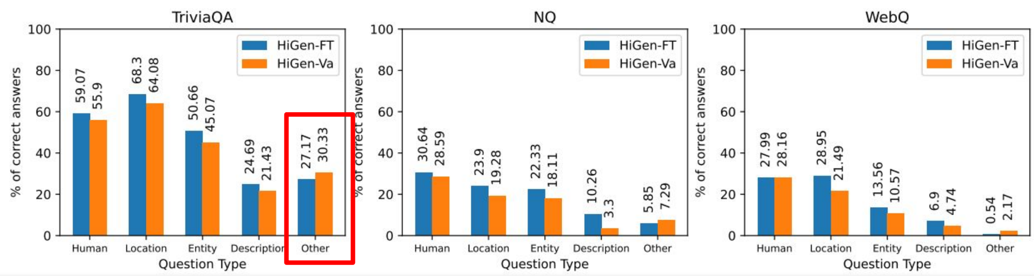


Figure 4: Exact Match values for TriviaQA, NQ, and WebQ datasets categorized by question type, based on the optimal settings for both HiGen-Va and HiGen-FT using few-shot learning on LLaMA-7b.

Subset: Direct Reasoning
Claim:
In TriviaQA dataset, HiGen-FT achieves higher accuracy than HiGen-Va over all types of questions.

Entailment Label: Refuted
Supporting Evidence:
`textual_paragraphs: [], figures: [4], tables: []`

Explanation:
HiGen-FT demonstrates superior accuracy compared to HiGen-Va in the Human, Location, Entity, and Description question types, however, it performs worse in the "Other" category. This contradicts the claim that it achieves higher accuracy across all question types.

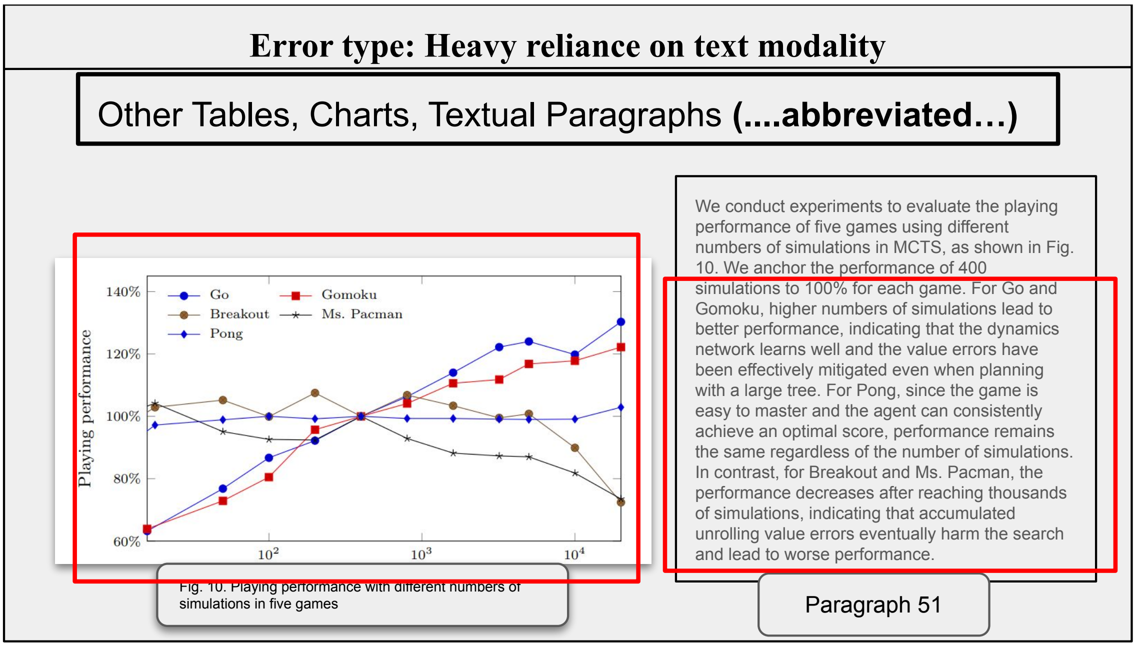
Qwen2.5-VL-72B Response:

To verify the claim, let's analyze the bar chart in the TriviaQA section. The blue bar represents the value of HiGen-FT, while the orange bar corresponds to HiGen-Va. Since all the HiGen-FT bars are consistently higher, the claim is supported.

Error Explanation:
The model incorrectly interprets the bar representing the "Other" category in the bar chart.

Figure 8: Illustration of *Visual element misinterpretation* with the example from the *Direct Reasoning* subset.

C.3 Heavy Reliance on Text Modality



Subset: Analytical Reasoning
Claim: Increasing the number of simulations in Monte Carlo Tree Search (MCTS) consistently improves the performance of all games, regardless of their complexity or specific dynamics, as it universally reduces value errors and enhances decision-making accuracy.

Entailment Label: Refuted
Supporting Evidence:
`textual_paragraphs: [51], figures: [10], tables: []`

Explanation:
The entailment label for the claim is "Refuted" because the claim generalizes the effects of increasing simulations in Monte Carlo Tree Search (MCTS) to **all games**, while the evidence in Section 4.3 and Figure 10 only discusses five specific games. The source shows that performance trends vary: Go and Gomoku improve with more simulations, Pong remains unaffected, and Breakout and Ms. Pacman degrade after a certain point. Since the claim introduces broader implications beyond the scope of the provided evidence, it cannot be verified based solely on the paper's context.

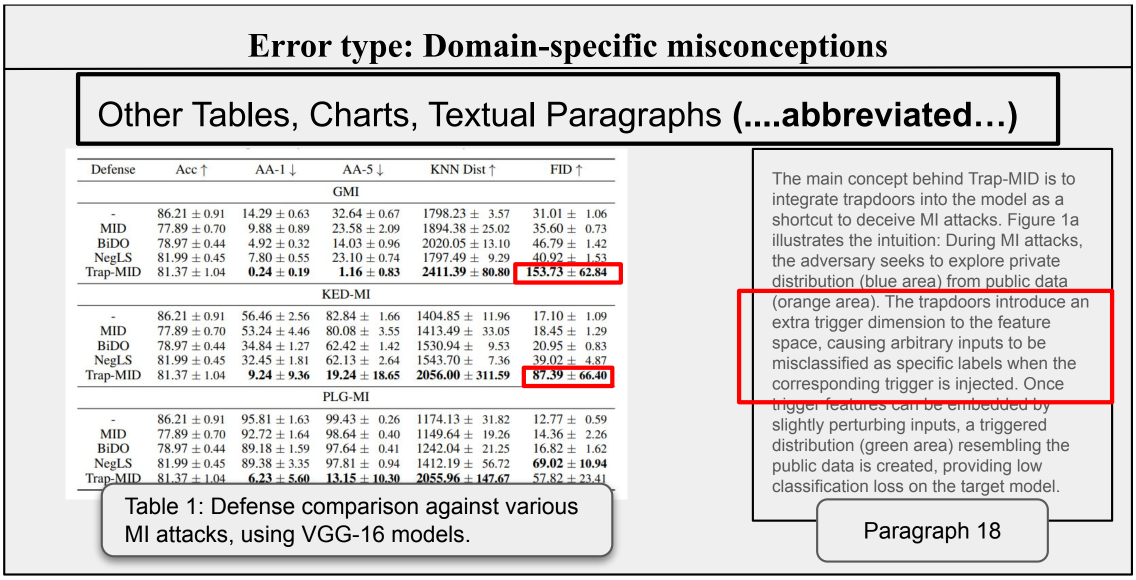
Qwen2.5-VL-72B Response:

The text explains that increasing the number of simulations in Monte Carlo Tree Search (MCTS) improves performance for Go and Gomoku due to reduced value errors and better learning of dynamics. While Pong's performance remains stable, this does not contradict the claim's universality since its simplicity allows for optimal performance regardless of simulations. The upward trend for Go and Gomoku in the figure further supports consistent improvement with increased simulations. Therefore, the claim is entailed.

Error Explanation:
The model's error lies in over-relying on textual descriptions and selectively interpreting visual trends while ignoring critical details from the chart and text. Although Go and Gomoku show improved performance with more simulations, the figure clearly shows that Breakout and Ms. Pacman experience performance degradation after a certain threshold, directly contradicting the claim's universality. Additionally, Pong's performance remains unaffected regardless of simulations, further disproving the claim. The model fails to integrate these crucial insights from the chart and instead focuses narrowly on textual information that partially aligns with the claim, leading to an incorrect entailment label.

Figure 9: Illustration of *Heavy Reliance on Text Modality* with the example from the *Analytical Reasoning* subset.

C.4 Domain-Specific Misconceptions



Defense	Acc ↑	AA-1 ↓	AA-5 ↓	KNN Dist ↑	FID ↑
			GMI		
-	86.21 ± 0.91	14.29 ± 0.63	32.64 ± 0.67	1798.23 ± 3.57	31.01 ± 1.06
MID	77.89 ± 0.70	9.88 ± 0.89	23.58 ± 2.09	1894.38 ± 25.02	35.60 ± 0.73
BiDO	78.97 ± 0.44	4.92 ± 0.32	14.03 ± 0.96	2020.05 ± 13.10	46.79 ± 1.42
NegLS	81.99 ± 0.45	7.80 ± 0.55	23.10 ± 0.74	1797.49 ± 9.29	40.92 ± 1.53
Trap-MID	81.37 ± 1.04	**0.24 ± 0.19**	**1.16 ± 0.83**	**2411.39 ± 80.80**	**153.73 ± 62.84**
			KED-MI		
-	86.21 ± 0.91	56.46 ± 2.56	82.84 ± 1.66	1404.85 ± 11.96	17.10 ± 1.09
MID	77.89 ± 0.70	53.24 ± 4.46	80.08 ± 3.55	1413.49 ± 33.05	18.45 ± 1.29
BiDO	78.97 ± 0.44	34.84 ± 1.27	62.42 ± 1.42	1530.94 ± 9.53	20.95 ± 0.83
NegLS	81.99 ± 0.45	32.45 ± 1.81	62.13 ± 2.64	1543.70 ± 7.36	39.02 ± 4.87
Trap-MID	81.37 ± 1.04	**9.24 ± 9.36**	**19.24 ± 18.65**	**2056.00 ± 311.59**	**87.39 ± 66.40**
			PLG-MI		
-	86.21 ± 0.91	95.81 ± 1.63	99.43 ± 0.26	1174.13 ± 31.82	12.77 ± 0.59
MID	77.89 ± 0.70	92.72 ± 1.64	98.64 ± 0.40	1149.64 ± 19.26	14.36 ± 2.26
BiDO	78.97 ± 0.44	89.18 ± 1.59	97.64 ± 0.41	1242.04 ± 21.25	16.82 ± 1.62
NegLS	81.99 ± 0.45	89.38 ± 3.35	97.81 ± 0.94	1412.19 ± 56.72	**69.02 ± 10.94**
Trap-MID	81.37 ± 1.04	**6.23 ± 5.60**	**13.15 ± 10.30**	**2055.96 ± 147.67**	57.82 ± 23.41

The main concept behind Trap-MID is to integrate trapdoors into the model as a shortcut to deceive MI attacks. Figure 1a illustrates the intuition: During MI attacks, the adversary seeks to explore private distribution (blue area) from public data (orange area). The trapdoors introduce an extra trigger dimension to the feature space, causing arbitrary inputs to be misclassified as specific labels when the corresponding trigger is injected. Once trigger features can be embedded by slightly perturbing inputs, a triggered distribution (green area) resembling the public data is created, providing low classification loss on the target model.

Subset: Analytical Reasoning
Claim: The effectiveness of Trap-MID in misleading Model Inversion attacks depends on achieving a balance between trapdoor naturalness and trapdoor effectiveness.

Entailment Label: Entailed
Supporting Evidence:
`textual_paragraphs: [18], figures: [], tables: [1]`

Explanation:
The entailment label for the claim is **Entailed** because the evidence from Paragraph 18 and Table 1 directly supports the claim. Paragraph 18 explains that the effectiveness of Trap-MID relies on balancing trapdoor naturalness (low visibility) and effectiveness (misleading MI attacks), while Table 1 quantitatively demonstrates this balance through high KNN distances (indicating effective misdirection) and high FID scores (indicating naturalness). These metrics validate the logical connection between the theoretical explanation and empirical results, making the claim a logical conclusion derived from the provided evidence.

Qwen2.5-VL-72B Response:

The table reports high FID and KNN distance values for Trap-MID, suggesting its effectiveness in generating unnatural reconstructions and distancing recovered images from private data. However, the textual paragraph in Paragraph 18 only describes the theoretical importance of trapdoor naturalness and effectiveness without explicitly linking their balance to defense success. The metrics in the table do not quantify how naturalness (low visibility) directly impacts the defense's performance, nor do they establish a causal relationship between these factors. Therefore, this evidence is insufficient to fully support the claim. Entailment Label: refuted.

Error Explanation:
The model incorrectly labels the claim as "refuted" due to a domain-specific misconception. It misunderstands the relationship between the metrics (FID and KNN distance) and the theoretical concept of balancing naturalness and effectiveness. While the textual paragraph in Paragraph 18 highlights the importance of "low visibility" for trapdoor naturalness, and Table 1 provides empirical data on FID and KNN distance, the model fails to integrate these distinct sources of evidence. Instead, it assumes that the lack of explicit causal quantification invalidates the claim, overlooking that the claim logically follows from synthesizing theoretical principles and empirical results.

Figure 10: Illustration of *Domain-Specific Misconceptions* with the example from the *Analytical Reasoning* subset.

C.5 Other Observation Error

Error type: Other observed errors

Other Tables, Charts, Textual Paragraphs **(....abbreviated...)**

Planner	r_{goal}	r_{local}	r_{fail}	# Particles	α_n	γ	κ	Tree Depth	# Sims	Max % of Particles Failed
POMCP	100	-5	-10	1000	N/A	0.999	150	8	2000	N/A
CC-POMCP	"	-3	-100	"	0.001	0.9	200	"	"	10%

TABLE I: Online Planner Parameters. Here, α_n is the step size, γ is the discount factor, and κ is the explore constant.

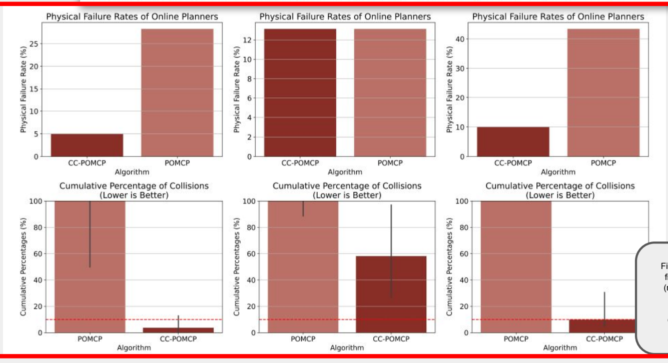


Fig. 5: The Physical Failure Rates (top) and Cumulative Percentage of Collisions (bottom). This figure depicts the results for the ENV-TRAINING (left), ENV-TUNNEL (middle), and ENV-STT (right) environments. Here, the failure rates refer to the number of times the robot left the map or collided with an obstacle while moving or localizing. Each graph shows the mean and standard deviation for each algorithm. Due to POMCP's poor performance, we clipped the cumulative percentages (bottom graphs) to 100%. Finally, the red dashed horizontal lines in the bottom graphs show the collision threshold of 10%.

Subset: Parallel Reasoning
Claim: The CC-POMCP algorithm consistently ensures that the cumulative percentage of collisions remains below the specified collision threshold (10%) in all tested environments. The POMCP algorithm enforces a stricter failure penalty compared to the CC-POMCP algorithm.

Entailment Label: Refuted
Supporting Evidence:
`textual_paragraphs: [], figures: [5], tables: [1]`

Explanation:
This claim is refuted because the numerical results in Figure 5 (Cumulative Percentage of Collisions) clearly show that CC-POMCP exceeded the 10% collision threshold in the ENV-TUNNEL environment. While CC-POMCP performed better than POMCP overall, its average cumulative collision percentage surpassed the threshold in this specific environment. In Table 1, the failure penalty for CC-POMCP is set to -100, significantly stricter than POMCP's penalty of -10.

Qwen2.5-VL-72B Response:

Table 1 shows that POMCP has a stricter failure penalty compared to CC-POMCP. Figure 5 demonstrates that CC-POMCP outperforms POMCP in terms of cumulative collision percentages in most environments, with its average values appearing below the 10% threshold in two out of three environments (ENV-TRAINING and ENV-STT). Therefore, the claim is entailed.

Error Explanation:
The response contains two critical errors: First, the statement "Table 1 shows that POMCP has a stricter failure penalty compared to CC-POMCP" is incorrect because Table I clearly indicates that CC-POMCP has a stricter failure penalty (r_fail = -100) than POMCP (r_fail = -10). Second, the statement "Figure 5 demonstrates that CC-POMCP outperforms POMCP in terms of cumulative collision percentages in most environments, with its average values appearing below the 10% threshold in two out of three environments (ENV-TRAINING and ENV-STT)" is flawed because Figure 5 shows that CC-POMCP exceeds the 10% collision threshold in the ENV-TUNNEL environment, contradicting the claim of consistent performance below the threshold. These errors involve misinterpretation of numerical data and fall under "Other observed errors."

Figure 11: Illustration of *Other Observation Error* with the example from the *Parallel Reasoning* subset.