Can AI Validate Science ? Benchmarking LLMs for Accurate Scientific Claim → Evidence Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly being used for complex research tasks such as literature review, idea generation, and scientific paper analysis, yet their ability to truly understand and process the intricate relationships within complex research papers, such as the logical links between claims and supporting evidence remains largely unexplored. In this study, we present CLAIM-BENCH, a comprehensive benchmark for evaluating LLMs' capabilities in scientific claim-evidence extraction and validation, a task that reflects deeper comprehension of scientific argumentation. We systematically compare three approaches which are inspired by divide and conquer approaches, across six diverse LLMs, highlighting model-specific strengths and weaknesses in scientific comprehension. Through evaluation involving over 300 claim-evidence pairs across multiple research domains, we reveal significant limitations in LLMs' ability to process complex scientific content. Our results demonstrate that closed-source models like GPT-4 and Claude consistently outperform open-source counterparts in precision and recall across claim-evidence identification tasks. Furthermore, strategically designed three-pass and one-by-one prompting approaches significantly improve LLMs' abilities to accurately link dispersed evidence with claims, although this comes at increased computational cost. CLAIM-BENCH sets a new standard for evaluating scientific comprehension in LLMs, offering both a diagnostic tool and a path forward for building systems capable of deeper, more reliable reasoning across full-length papers. [1]

1 Introduction

Large Language Models (LLMs) have become important tool in academic research, demonstrating impressive capabilities such as automating comprehensive literature reviews, facilitating innovative idea generation, and aiding experimental design. These advancements promise significant improvements in research productivity, creativity, and efficiency, fueling excitement about the transformative potential of AI-driven methodologies in science. However, as researchers increasingly assign critical tasks to these models—from content summarization and hypothesis generation to automated peer review (Checco et al., 2021; Agarwal et al., 2025; Lu et al., 2024)—a fundamental yet overlooked question emerges: how deeply do these models truly understand scientific knowledge beyond surface-level pattern recognition? Despite their widespread use and promising outcomes, there remains uncertainty about the depth and accuracy of their reasoning capabilities, particularly in complex scientific contexts.

Scientific papers are characterized by intricate relationships, primarily structured around claims supported by corresponding evidence. The ability to accurately identify and reason about these claim-evidence pairs is essential for validating scientific findings and ensuring research integrity, making it a critical test of LLMs' comprehension depth. Unlike surface-level tasks such as summarization or question answering, claim-evidence identification requires global reasoning across paper sections, synthesis of dispersed information, and a nuanced understanding of logical dependencies. While existing works have assessed LLMs' capabilities in related research tasks such as summarization (Agarwal et al., 2025), literature synthesis (Lu et al., 2024), and hypothesis generation (Vladika and Matthes, 2023), none have explicitly benchmarked LLM performance on systematically extracting and validating claims with supporting evidence, leaving this area of scientific comprehension underexplored.

Despite the importance of accurately reasoning

[1] To facilitate future research and standardize evaluation in this area, we release CLAIM-BENCH at [repository]

about claims and supporting evidence, no existing benchmarks explicitly assess LLM capabilities for this specific type of high-level scientific reasoning. Benchmarks such as LongGenBench (Wu et al., 2025) and XL2Bench (Ni et al., 2024) have highlighted persistent limitations in LLMs' abilities to process long-context inputs and maintain logical coherence. Similarly, peer review frameworks like MetaWriter (Sun et al., 2024b) and AGENTREVIEW (Jin et al., 2024) evaluate LLMs in automated review contexts but do not specifically test their capability to validate logical relationships such as claims and evidence, a task crucial for rigorous scientific evaluation. Findings from Chain of Evidence (CoE) frameworks (Chang et al., 2024) underscore the complexity of structured, multi-hop reasoning required to integrate and validate information dispersed across documents. All these works evaluate reasoning in the general domains, but the scientific reasoning capability, which imposes unique challenges, is not benchmarked.

Within scientific reasoning, The AI Scientist (Lu et al., 2024), LitLLM (Agarwal et al., 2025), and ChatCite (Li et al., 2025) benchmark LLMs on tasks such as literature review and hypothesis generation, while ScienceAgentBench (Chen et al., 2025) and SCBENCH (LI et al., 2025) probe multi-step reasoning and long-context understanding. However, none of these frameworks explicitly measure the finer-grained ability to verify whether the evidence presented in a full scientific paper truly supports its claims—precisely the claim-and-evidence (C-E) reasoning capability our benchmark targets.

To address these gaps, we present CLAIM-BENCH, a novel benchmark designed to systematically evaluate LLMs' abilities to identify and validate claim-evidence relationships in scientific papers. CLAIM-BENCH challenges LLMs to process entire scientific papers, connect ideas across sections, and reason about them on a high level. In this work, we evaluate six state-of-the-art LLMs across diverse research domains. Our experiments indicate that larger models (e.g., GPT-4-Turbo, Claude 3.5) maintain high recall even with lengthy documents, especially when using iterative prompting, whereas smaller models (e.g., LLaMA, Ministral) experience significant performance drops with increasing document length specially under Single-Pass prompting. These findings highlight crucial areas for enhancing long-context comprehension and inform the development of reliable AI-driven tools for scientific research and peer review.

2 Related Work

AI for Science Large Language Models (LLMs) have significantly advanced scientific workflows, facilitating tasks such as peer review and hypothesis generation. Tools like ReviewerGPT (Liu and Shah, 2023) and ReviewFlow (Sun et al., 2024a) have streamlined peer review processes, while AGENTREVIEW (Jin et al., 2024) simulates collaborative review systems to improve research evaluation workflows. In parallel, fact-checking frameworks, such as Scientific Fact-Checking (Vladika and Matthes, 2023) and Exploring Multidimensional Checkworthiness (Liu et al., 2025), emphasize validating claims in scientific literature. However, these systems primarily focus on localized tasks or prioritization mechanisms, leaving the broader challenge of understanding the connections across entire documents by LLMs unaddressed. Additional work such as AI-assisted peer review (Checco et al., 2021) explores the feasibility of algorithmically approximating peer-review judgments, raising key ethical and practical concerns.

Benchmarks Long-context benchmarks, such as SCBENCH (LI et al., 2025), MMLongBench-Doc (Ma et al., 2024), and LongGenBench (Wu et al., 2025), have assessed LLMs' ability to process extended inputs and maintain coherence, focusing primarily on tasks like document summarization and long-form generation. Specialized benchmarks like U-MATH (Chernyshev et al., 2025) and Leave No Document Behind (Godbole et al., 2024) examine domain-specific reasoning and multi-document synthesis but address relatively structured and localized relationships. The LCFO benchmark (Costa-jussà et al., 2024a) targets summary expansion with varying granularities of content compression, revealing limits in semantic retention. The Y-NQ dataset (Costa-jussà et al., 2024b) exposes disparities in open-book comprehension across low- & high-resource languages, hinting at deeper weaknesses in cross-lingual and low-resource long-context understanding. Data Interpreter (Hong et al., 2024) showcases long-term data analysis workflows with LLM agents, but primarily focuses on task planning and execution rather than deep textual reasoning. In neuroscience, (Luo et al., 2025) show LLMs surpassing expert predictions in future experimental outcomes, yet such success doesn't imply comprehension of reasoning chains. In contrast, our work focuses specifically on research papers, which are characterized by more complex and dispersed relationships, such as claims supported by evidence across multiple sections. CLAIM-BENCH evaluates the ability of LLMs to synthesize these intricate connections, testing their capacity for global reasoning and coherence in a way that reflects the unique demands of scientific texts.

Collaborative Reasoning Collaborative reasoning frameworks offer a complementary perspective, with multi-agent systems like Two Heads Are Better Than One (Su et al., 2025) and iterative feedback mechanisms such as CYCLERESEARCHER (Weng et al., 2025) showing promise in enhancing reasoning capabilities. While these approaches address some limitations of Single-Pass LLM systems, their primary focus remains on generating and refining content rather than validating complex logical relationships. Similarly, tools like AIGS (Liu et al., 2024) and LLM-Assisted Hypothesis Generation (Vladika and Matthes, 2023) explore reasoning and hypothesis testing but do not directly tackle the problem of scientific comprehension. (Leng et al., 2024) introduce a graph-based approach for hypothesis generation and evaluation, demonstrating potential for structured creativity, yet falling short of validating interlinked arguments at scale.

Ethical AI Finally, ethical considerations have been raised in works like Ethical Use of LLMs (Lissack and Meagher, 2024), which stresses the need for transparency and accountability in AI-driven research, and multimodal benchmarks like MileBench (Dingjie et al., 2024), which expand the scope of LLM evaluation to include visual and textual data. These efforts, while addressing important aspects of AI integration in research, highlight the absence of targeted benchmarks that evaluate claim-evidence validation across long, complex scientific texts—a gap CLAIM-BENCH aims to fill.

3 Methodology

In this section, we present the design of CLAIM-BENCH, our benchmark for evaluating how well LLMs identify and analyze claim–evidence relationships in full-length research papers.

3.1 Dataset

Dataset Curation The dataset for this study was curated by 4 PhD students with research experience. Each annotator had at least one first-author conference publication, ensuring familiarity with scientific writing standards. These researchers selected papers according to specific guidelines (Appendix B.1) to ensure relevance and diversity. Selection criteria included: papers from the year 2024, non-math-intensive subjects, length between 0 to 20 pages. The aim was to represent a broad spectrum of current AI/ML research topics within the dataset.

To facilitate easier annotations, we developed a PDF annotation tool, it lets users load a paper, drag a pointer over any sentence or paragraph to mark it as a claim, then click-add evidence additional spans as linked evidence for that claim; each claim–evidence pair is stored in a one-to-many structure and exported as JSON. (see Appendix B.3).

Annotation Quality Check After compiling the initial annotations (100 papers), these were set aside before evaluating the models to ensure an unbiased assessment of their capabilities. To enhance the reliability of our dataset as ground truth, we conducted a validation phase where a different set of annotators re-annotated a subset of 30 papers. We then assessed annotation consistency by calculating Inter-Annotator Agreement (IAA) using the average F1-score across annotator pairs. This analysis yielded substantial agreement for identifying claims (F1 = 0.755) and moderate agreement for identifying evidence (F1 = 0.659) and linked claim-evidence pairs (F1 = 0.617), confirming the dataset's suitability for benchmarking. The detailed methodology used for this IAA is provided in Appendix B.2.

3.2 Evaluation Metrics

In this study, we employ four metrics to evaluate the LLM performance: three established metrics in information retrieval, precision, recall, F1-score, and a novel metric, sentence_gap, to evaluate LLM performance in claim-evidence retrieval tasks and the effectiveness of our various prompting techniques.

Precision (P) Used to measure the proportion of spans the model predicts that are identified by the annotators, reflecting their effectiveness in responding to precise and carefully structured prompts.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{1}$$

where TP (true positive) is the number of correctly retrieved claim/evidence, and FP (false positive) is the number of retrieved "claim"/"evidence" that are not claims/evidences.

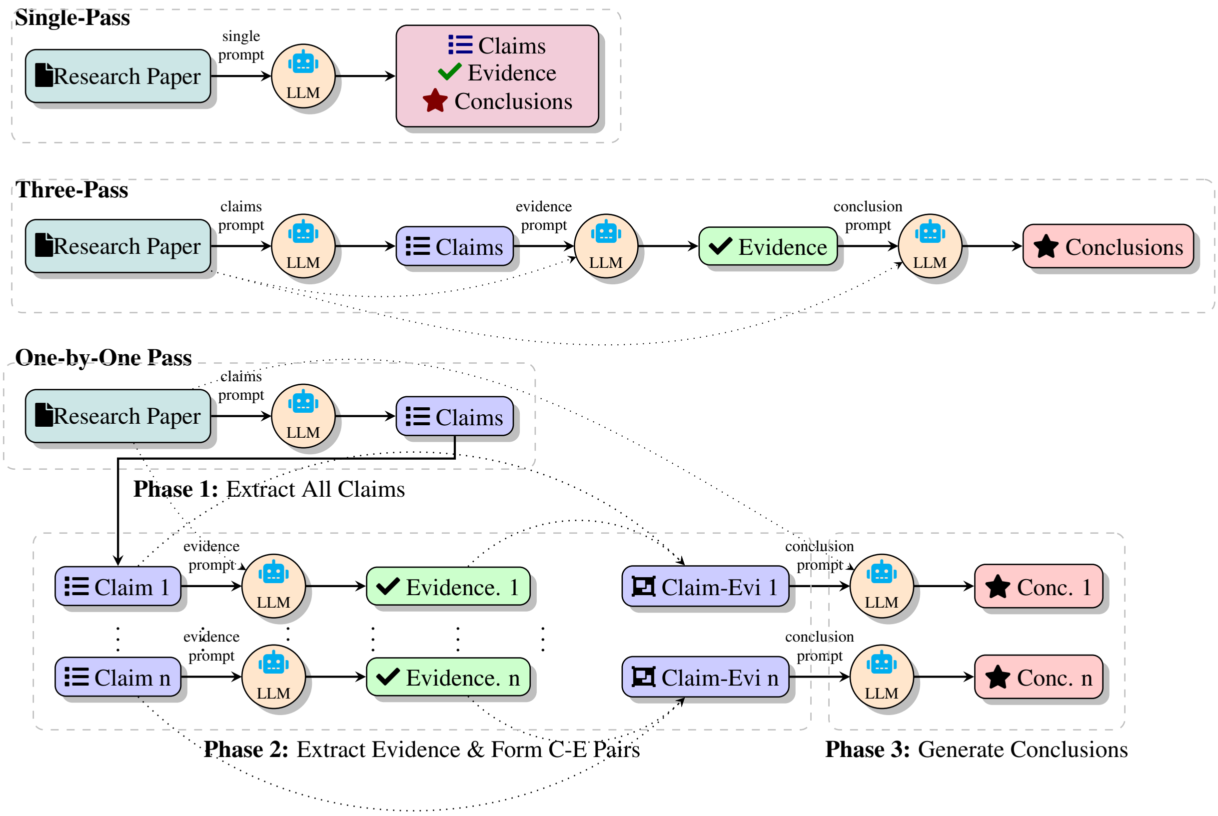


Figure 1: Three methods to prompt LLMs to analyze the papers. **Single-Pass:** Full paper processing with one prompt. **Three-Pass:** Sequential claim → evidence → conclusion extraction. **One-by-One Pass**: Individual evidence retrieval per claim.

Recall (R) Quantifying the portion of claim/evidence that are retrieved. Recall assesses the ability to capture pertinent data, a measure of the model's responsiveness to exhaustive prompt inquiries

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

where FN (false negative) is the number of claims/evidences that are incorrectly missed.

F1-score This is the harmonic mean of P and R. The F1-score provides a balanced measure of accuracy, crucial for evaluating the efficacy of the prompting techniques in eliciting detailed and relevant responses.

sentence_gap The sentence_gap metric measures the distance between a retrieved claim and each of its associated retrieved evidence. It is particularly valuable for evaluating long-range contextual comprehension by quantitatively assessing models' ability to handle textual relationships over extended contexts. This assessment is crucial for complex prompts designed to challenge such comprehension and is instrumental as we explore how increasing LLM context length capabilities enhance performance in realistic scenarios.

$$\text{sentence\_gap} = \frac{1}{|\mathcal{M}|} \sum_{(p,g)\in\mathcal{M}} \big|s(p) - s(g)\big|, \tag{3}$$

where $\mathcal{M}$ is the set of matched evidence pairs (using Intersection over Union matching rule). $s(\cdot)$ returns the sentence index of a span inside the document. The sentence_gap metric is therefore the average absolute sentence-level distance between each predicted claim span p and its evidence span g, capturing how far a model must reason across the paper to link claims with supporting evidence.

Secondary metrics Additionally, we consider secondary metrics that focus on operational aspects of model performance: the time to generate outputs and how each model's recall changes as input length (token count) increases. These metrics are crucial for understanding efficiency and scalability. They help compare how models manage computational resources and handle large input sizes under varying conditions.

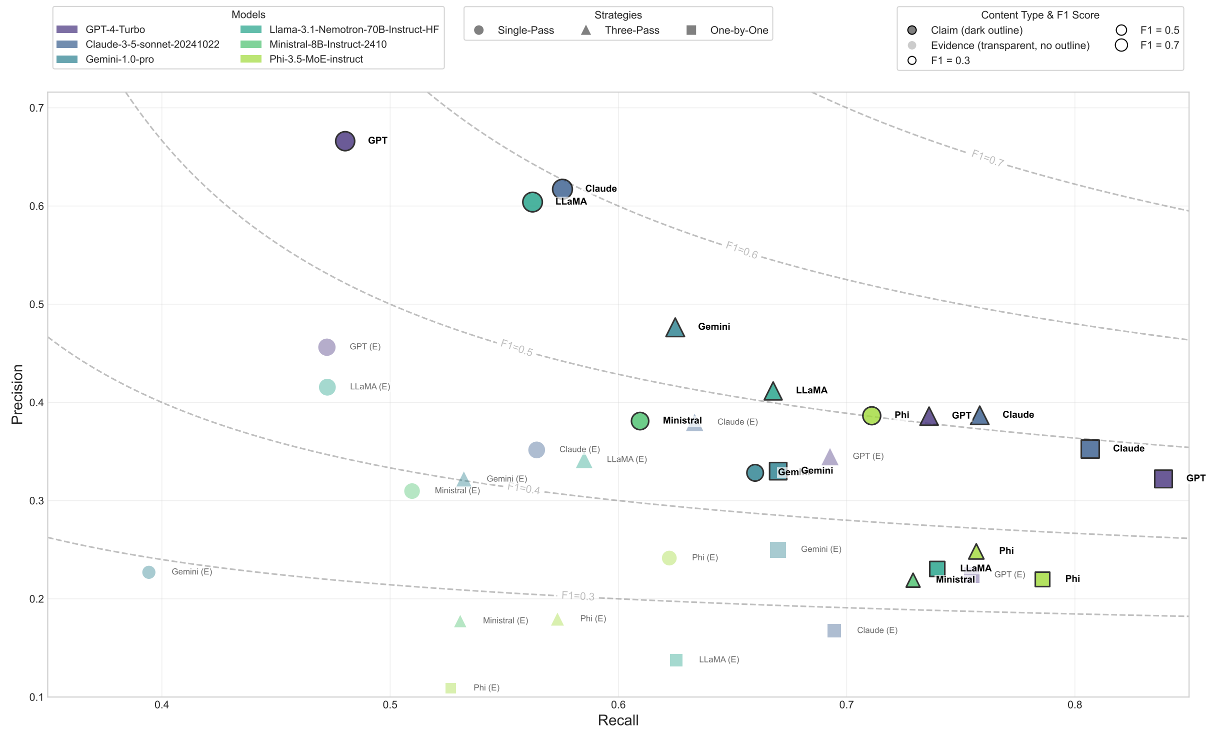


Figure 2: Precision vs. Recall for claim (solid markers) and evidence (transparent markers) identification across models and strategies (shapes: Single-Pass ●, Three-Pass ▲, One-by-One ■). Models show higher precision for claims, higher recall for evidence, with most results below $F_1 = 0.7$.

4 Experimental Setup

We evaluate six state-of-the-art LLMs, chosen to span both licensing regimes and architectural families while sharing a ≥128K-token context window. Open-source include Ministral-8B (Mistral AI, 2024), Phi-3.5-MoE (Abdin et al., 2024), and LLaMA-70B (Wang et al., 2025) and Closed-source includes GPT-4 (OpenAI, 2024), Gemini-Exp_1114 (Gemini Team, 2024), and Claude 3.5 Sonnet (Anthropic, 2025).

4.1 Analysis Methods

As illustrated in Figure 1, we explore three distinct prompting methods to assess and enhance model performance on claim-evidence identification tasks.

Single-Pass Initially, we present the models with a research paper, instructing (Appendix A.1) them to identify claims, evidences, and conclusions in a single comprehensive prompt.

Three-Pass Building on the "divide and conquer" strategy from prior research, we then deconstruct the task into sequential stages. In the first stage, the model identifies claims using a dedicated prompt. Subsequently, these claims are supplied to the next stage, where separate prompts elicit corresponding evidences. Finally, we combine the identified claims and evidences, using another prompt to extract conclusions (Appendix A.2).

One-by-One Pass We adopt a more granular approach where each claim is processed individually to retrieve evidence. This means for n claims, the model runs n times to gather evidence for each, and similarly for conclusions. Although this approach provides detailed analysis, it significantly increases the demand on computational resources and time (Appendix A.3). These methods combine careful prompting with our annotated claim–evidence dataset, allowing us to benchmark each model's extraction accuracy and probe how different prompt strategies improve performance.

5 Results

The following section details the experimental results, highlighting comparative model performance and strategic impacts.

5.1 Precision vs Recall

As shown in Figure 2, models exhibit a clear precision-recall trade-off: settings that achieve

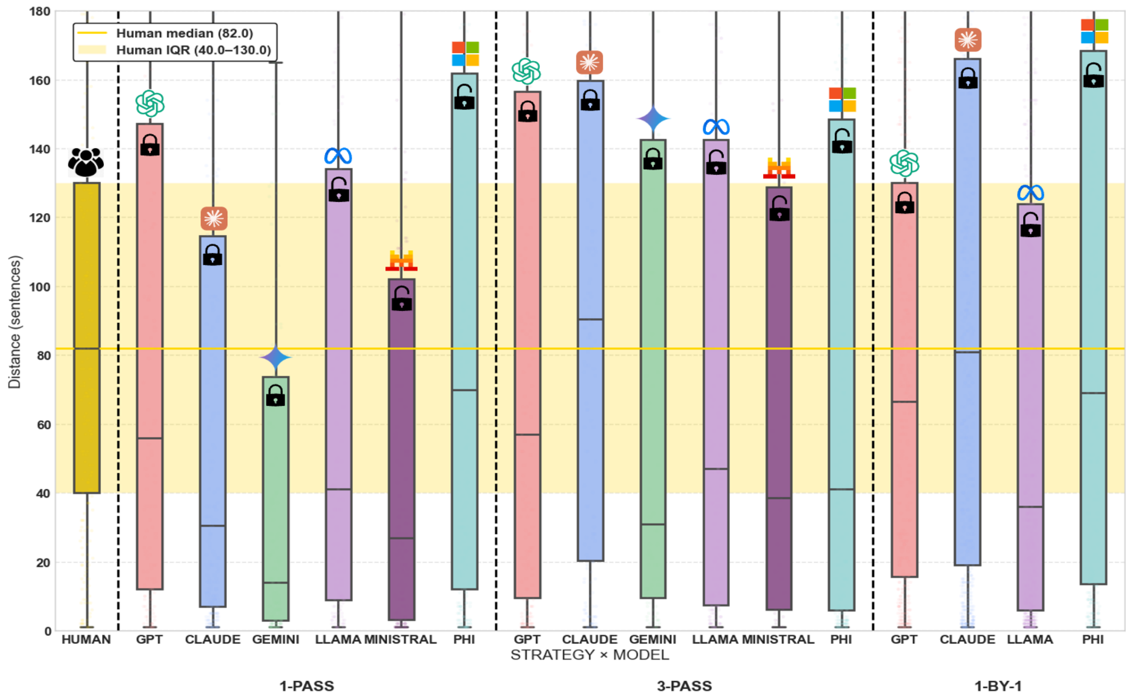


Figure 3: Sentence distance distribution (box plots) between claims and linked evidence vs. Human baseline (*leftmost*). LLMs, especially with iterative strategies, link over longer distances than humans, showing capability but potential noise.

higher recall often incur reduced precision. For instance, Claude and LLaMA achieve high recall but at the cost of extracting numerous false positives, which is evident from their large maximum linking distances (Figure 7), exceeding 2,200 sentences in some cases. Although valuable, such long-range links raise the risk of false claim–evidence pairs. Conversely, models like GPT prioritize precision, maintaining moderate linking distances (around 658–708 sentences) with fewer spurious matches, though this approach slightly limits recall. Ministral offers a balanced precision-recall profile, characterized by consistent, shorter linking distances.

Comparing the precision-recall tradeoff trends between open- and closed-source models, we see that closed-source models balance precision and recall better. Overall, GPT often balances high precision and moderate recall; Claude achieves higher recall rates but exhibits noticeable trade-offs in precision. Gemini remains stable across strategies. Among open-source models, LLaMA came close to matching closed-source recall but with some outliers, also shows variability in precision; Ministral is moderate in both coverage & precision; Phi exhibits the widest swings, at times matching larger models but also dropping in accuracy.

5.2 Smaller vs Larger Models

Larger models, such as GPT-4-Turbo, Claude, Gemini, and LLaMA, generally exhibit strong recall in identifying claims, with GPT-4-Turbo achieving high precision (0.68) and recall (0.81), demonstrating effective balance at different strategies. Claude also shows strong recall (0.83), albeit with a moderate precision drop (0.61). Also, LLaMA achieves similar recall (0.76) but comparative precision (0.60), indicating a tendency to identify extensive and highly precise connections, considering the best cases of each model.

Smaller models, such as Ministral and Phi, typically exhibit lower recall and precision. Ministral shows modest recall (0.60) with precision around 0.38, reflecting a conservative approach to claim-evidence linking. Phi demonstrates similar precision (approximately 0.39) but notably higher recall (around 0.7) in the best cases. These observations highlight a clear trade-off: larger models generally identify broader and more nuanced claim–evidence relationships but often at the cost of precision, whereas smaller models maintain more consistent precision with significantly reduced recall. In both the cases similar pattern holds in evidence extraction as well.

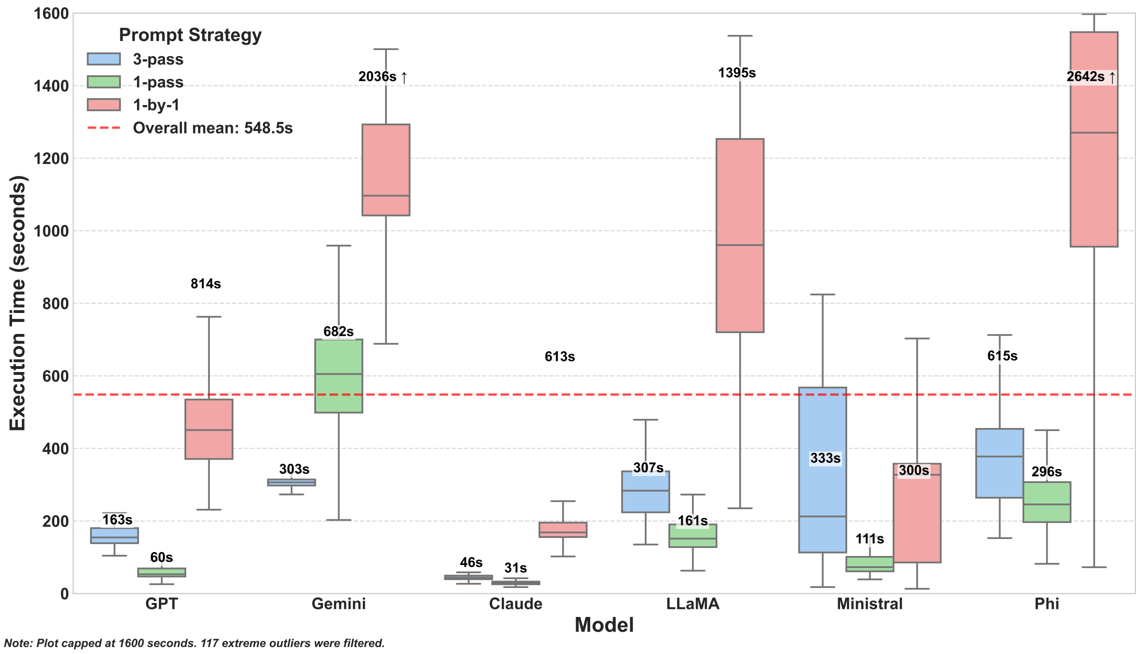


Figure 4: Execution time comparison (box plots): Single-Pass (■) is fastest, One-by-One (■) is slowest. Models vary greatly in speed (e.g., Claude consistently fast; LLaMA/Phi often requiring >1000s).

5.3 Claims vs Evidence Extraction

Model	Best C Performances			Best E Performances		
	F1	P	R	F1	P	R
GPT-4-Turbo	0.56	0.66	0.57	**0.47**	**0.34**	**0.69**
Claude 3.5	**0.59**	**0.62**	**0.60**	0.42	0.33	0.66
Gemini-Exp_1114	0.54	0.48	0.64	0.40	0.30	0.52
LLaMA-70B	0.58	0.60	0.56	0.45	0.42	0.49
Ministral-8B	0.48	0.39	0.61	0.39	0.31	0.52
Phi-3.5-MoE	0.50	0.40	0.72	0.35	0.25	0.63

Table 1: The highest performance (across all strategies) for Claim (C) and Evidence (E) extraction; "P@R" denotes precision at the corresponding recall.

Analyzing claim versus evidence extraction separately reveals distinct performances among LLMs (see Table 1). Across all models, precision is consistently higher for claims than for evidence, indicating the models more readily detect explicit claims compared to the contextually dispersed evidence. Also, the evidence extraction of all models yields higher recall than precision. In addition to the common trends, the models exhibit distinct patterns. For instance, Claude and LLaMA exhibit high recall in evidence extraction but with substantial variability in linking distances (Claude: mean gap of 119.4 sentences, variance of 33,674; LLaMA: mean 95.1 sentences, variance of 34,207), suggesting increased noise and inconsistent performance. Conversely, Ministral maintains lower linking distances (mean 75.9 sentences) with minimal variance, signifying a more cautious and controlled approach.

5.4 Impact of Strategy

The Single-pass strategy is highly efficient but has limited coverage, e.g., GPT-4 produces 152 pairs with a 98.5 average sentence_gap, while Ministral generates 166 pairs (average gap: 64.2). Meanwhile, the Three-pass strategy enhances recall and coverage at moderate computational cost. Claude yields 174 pairs (average gap: 122.2), and Phi captures 279 pairs, albeit with significant variance (11,490.2) in sentence_gap. Finally, the One-by-One strategy maximizes recall but increases computational demand significantly. Claude and LLaMA produce the highest counts (639 and 659 pairs, respectively), with substantial gaps (Claude: 119.4, LLaMA: 95.1) and high variance (Claude: 33,673.9, LLaMA: 34,207.0). Phi also achieves substantial coverage (347 pairs) with notable variance (13,188.2).

5.5 Impact of Token Length on Recall

We observed how the documents' token length affected the models' recall performances. In long documents, we expected performance drops, but these observed drops are tied to the prompting strategy. With the Single-pass strategy, the recall performances dropped as the document length increased. With the iterative prompting strategies (Three-pass or One-by-One), the performance drops are less significant, indicating that the iterative prompting imposes less "processing load" onto the LLMs. Additionally, the recall drops differ by the sizes of the models. Relatively smaller models (LLaMA 70B and Ministral 8B) showed more notable declines, especially with Single-pass, whereas the larger models (Claude and GPT-4) maintained relatively high recalls, underlining the advantage of their long context capabilities. Additional details in Appendix C.

Claude and LLaMA frequently produce the highest pair counts (up to 639 and 659), reflecting broad coverage. This can coincide with their large context window sizes—helpful for capturing distant relationships—yet also introduces potential noise. GPT and Gemini keep moderate distances, suggesting they discovered fewer links. Ministral remains conservative with fewer pairs with shorter distances, while Phi's extreme variance indicates inconsistent linking across long contexts. We include the details in Figure 7 (in Appendix C).

5.6 Execution Time Analysis

As shown in Figure 4, the execution times differ considerably across models and strategies. GPT is highly efficient in the Single-Pass (under 200s) and relatively moderate in one-by-one approaches (∼500s). Gemini exhibits intermediate execution times across all strategies, notably higher for the three-pass (∼600s). Claude consistently achieves the fastest execution across all strategies, maintaining execution times under 200 seconds. LLaMA shows extensive variability, especially with one-by-one strategies frequently exceeding 1,200 seconds, reflecting significant computational demands. Ministral shows relatively balanced execution times, with three-pass and one-by-one strategies averaging around 600–900 seconds. Phi demonstrates the highest computational intensity, especially in one-by-one strategies, often surpassing 1,200 seconds, highlighting the considerable resource investment required for thorough analyses. The execution times recorded for Gemini exhibit some variability, which may partially stem from fluctuations in API response latency during our experiments, combined with the necessary sleep() intervals implemented for rate limiting.

6 Discussion

The insights from CLAIM-BENCH emphasize critical directions for future research and practical applications leveraging the capabilities of LLMs in scientific claim-evidence reasoning. Improving LLMs' ability to accurately validate claim-evidence pairs could enhance their practical use in designing experiments and generating scientifically valid hypotheses. Furthermore, improved claim identification and validation methods provide a foundation for developing sophisticated claim quality scoring tools that can greatly enhance peer-review processes. The capability to systematically link and integrate evidence across multiple scientific papers could lead to powerful retrieval-augmented laboratory assistants and cross-paper evidence graphs, accelerating knowledge discovery. These advancements would not only strengthen the robustness of scientific validations but also facilitate the creation of more sophisticated scientific QA systems, thus laying foundational benchmarks for future scientific text generation and evaluation methods. This research thus serves as a pivotal foundation for transformative applications in scientific inquiry and discourse.

7 Conclusion

Motivated by the limited evaluation in prior literature of LLMs' abilities in scientific reasoning, we introduced CLAIM-BENCH, a novel benchmark specifically designed to evaluate LLMs' capabilities in identifying and validating claim-evidence relationships within scientific texts. We systematically explored diverse LLM architectures and prompting strategies. Our results demonstrate significant limitations in LLMs' comprehension, specifically in their precision and recall balance when processing complex scientific documents. Notably, models showed higher precision in extracting explicit claims, whereas extracting dispersed evidence proved challenging, yielding higher recall but lower precision and increased sentence gaps. Moreover, our comparative analysis across 3 strategies revealed substantial trade-offs between computational efficiency, precision, and coverage. Closed-source models generally displayed more stable performances, while open-source models offered broad yet inconsistent coverage. CLAIM-BENCH provides a framework for the assessment of LLMs in complex scientific contexts, and our study provides useful material and insights for continuing the advancement in LLMs' high-level comprehension and scientific reasoning capabilities.

8 Limitations

While CLAIM-BENCH provides comprehensive insights into the capabilities of LLMs in scientific claim-evidence reasoning. Despite these insights, CLAIM-BENCH has several limitations worth noting. First, the benchmark primarily focuses on recent papers from select domains, which are after the LLMs' knowledge cutoff but might limit the generalizability. Second, the evaluation relies on existing LLM architectures. While we leave the exploration of the impact of model architecture development to future works, CLAIM-BENCH could be a useful material that supports future projects that develop novel LLM architectures that have enhanced long-context language understanding capabilities and scientific reasoning capabilities.

References

Marah Abdin, Jyoti Aneja, and Hany Awadalla et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. LitLLM: A Toolkit for Scientific Literature Review. *arXiv preprint*. ArXiv:2402.01788.

Anthropic. 2025. Claude 3.5 sonnet model card addendum. PDF file. Accessed 12 Apr. 2025.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Qing Wang, Yihao Huang, and Yang Liu. 2024. What External Knowledge is Preferred by LLMs? Characterizing and Exploring Chain of Evidence in Imperfect Context. *arXiv preprint*. ArXiv:2412.12632.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. 2021. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1):1–11.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. In *The Thirteenth International Conference on Learning Representations*.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. 2025. U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. *arXiv preprint*. ArXiv:2412.03205.

Marta R. Costa-jussà, Pierre Andrews, Mariano Coria Meglioli, Joy Chen, Joe Chuang, David Dale, Christophe Ropers, Alexandre Mourachko, Eduardo Sánchez, Holger Schwenk, Tuan Tran, Arina Turkatenko, and Carleigh Wood. 2024a. LCFO: Long Context and Long Form Output Dataset and Benchmarking. *arXiv preprint*. ArXiv:2412.08268.

Marta R. Costa-jussà, Joy Chen, Ifeoluwanimi Adebara, Joe Chuang, Christophe Ropers, and Eduardo Sánchez. 2024b. Y-NQ: English-Yorùbá Evaluation dataset for Open-Book Reading Comprehension and Text Generation. *arXiv preprint*. ArXiv:2412.08279.

Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. MileBench: Benchmarking MLLMs in Long Context. In *First Conference on Language Modeling*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Aditi Godbole, Jabin Geevarghese George, and Smita Shandilya. 2024. Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications. *arXiv preprint*. ArXiv:2409.18454.

Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Chenxing Wei, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Li Zhang, Lingyao Zhang, Min Yang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Xiangru Tang, Xiangtao Lu, Xiawu Zheng, Xinbing Liang, Yaying Fei, Yuheng Cheng, Zhibin Gou, Zongze Xu, and Chenglin Wu. 2024. Data Interpreter: An LLM Agent For Data Science. *arXiv preprint*. ArXiv:2402.18679.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring Peer Review Dynamics with LLM Agents. *arXiv preprint*. ArXiv:2406.12708.

Yan Leng, Hao Wang, and Yuan Yuan. 2024. Llm-Assisted Hypothesis Generation and Graph-Based Evaluation.

YUCHENG LI, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. 2025. SCBench: A KV Cache-Centric Analysis of Long-Context Methods. In *The Thirteenth International Conference on Learning Representations*.

Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. 2025. ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3613–3630, Abu Dhabi, UAE. Association for Computational Linguistics.

Michael Lissack and Brenden Meagher. 2024. Ethical Use of Large Language Models in Academic Research and Writing: A How-To.

Houjiang Liu, Jacek Gwizdka, and Matthew Lease. 2025. Exploring Multidimensional Checkworthiness: Designing AI-assisted Claim Prioritization for Human Fact-checkers. *arXiv preprint*. ArXiv:2412.08185.

Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. *arXiv preprint*. ArXiv:2306.00622.

Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. 2024. AIGS: Generating Science from AI-Powered Automated Falsification. *arXiv preprint*. ArXiv:2411.11910.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint*. ArXiv:2408.06292.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K. Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O. Cohen, Valentina Borghesani, Anton Pashkov, Daniele Marinazzo, Jonathan Nicholas, Alessandro Salatiello, Ilia Sucholutsky, Pasquale Minervini, Sepehr Razavi, Roberta Rocca, Elkhan Yusifov, Tereza Okalova, Nianlong Gu, Martin Ferianc, Mikail Khona, Kaustubh R. Patil, Pui-Shee Lee, Rui Mata, Nicholas E. Myers, Jennifer K. Bizley, Sebastian Musslick, Isil Poyraz Bilgin, Guiomar Niso, Justin M. Ales, Michael Gaebler, N. Apurva Ratan Murty, Leyla Loued-Khenissi, Anna Behler, Chloe M. Hall, Jessica Dafflon, Sherry Dongqi Bao, and Bradley C. Love. 2025. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9(2):305–315.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mistral AI. 2024. Un Ministral, des Ministraux: Introducing the world's best edge models. `https://mistral.ai/news/ministraux`. Accessed 19 May 2025.

Xuanfan Ni, Hengyi Cai, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, and Piji Li. 2024. XL^{2}Bench: A Benchmark for Extremely Long Context Understanding with Long-range Dependencies. *arXiv preprint*. ArXiv:2404.05446.

OpenAI. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. *arXiv preprint*. ArXiv:2410.09403.

Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024a. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 120–137, Greenville SC USA. ACM.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024b. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32.

Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. *arXiv preprint*. ArXiv:2305.16859.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. CycleResearcher: Improving Automated Research via Automated Review. In *The Thirteenth International Conference on Learning Representations*.

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025. LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs. In *The Thirteenth International Conference on Learning Representations*.

A Prompt Templates

A.1 Single-Pass Prompt

Comprehensive Evaluation Prompt

Analyze the research paper and provide a comprehensive evaluation following these guidelines:

1. Identify ALL claims in the paper where each claim:
   - Makes a specific, verifiable assertion
   - Is supported by concrete evidence
   - Represents findings, contributions, or methodological advantages
   - Can be from any section except abstract
2. For each identified claim:
   - Extract ALL supporting or contradicting evidence (experimental results, data, or methodology)
   - Evaluate the evidence strength and limitations
   - Assess how well conclusions align with evidence

Return ONLY the following JSON structure:

```
{
    "analysis": [
        {
            "claim_id": number,
            "claim": {
                "text": "statement of the claim",
                "type": "methodology/result/contribution/performance",
                "location": "section/paragraph",
                "exact_quote": "verbatim text from paper"
            },
            "evidence": [
                {
                    "evidence_text": "specific experimental result/data",
                    "strength": "strong/moderate/weak",
                    "limitations": "specific limitations",
                    "location": "section/paragraph",
                    "exact_quote": "verbatim text from paper"
                }
            ],
            "evaluation": {
                "conclusion_justified": true/false,
                "robustness": "high/medium/low",
                "justification": "explanation of evidence-conclusion alignment",
                "key_limitations": "critical limitations affecting validity",
                "confidence_level": "high/medium/low"
            }
        }
    ]
}
```

Ensure:

- ALL substantive claims are captured
- Evaluations are objective and well-reasoned
- All locations and quotes are precise
- Multiple pieces of evidence per claim are included when present

A.2 Three-Pass Prompt

Claims Extraction Prompt

Paper text: {text}
Task: Identify all statements in the text that meet the following criteria for a claim:

1. Makes a specific, testable assertion about results, methods, or contributions.
2. Represents a novel finding, improvement, or advancement.
3. Presents a clear position or conclusion.

Requirements:

- Include both major and minor claims.
- Don't miss any claims.
- Present each claim as a separate item.

Return ONLY the following JSON structure:

```
{
    "claims": [
        {
            "claim_id": 1,
            "claim_text": "statement of the claim",
            "location": "section/paragraph where this claim appears",
            "claim_type": "Nature of the claim",
            "exact_quote": "complete verbatim text containing the claim"
        }
    ]
}
```

Evidence Identification Prompt

Paper text: {text}
For these claims: `{claims_text}`
Please identify relevant evidence that:

1. Directly supports or contradicts the claim's specific assertion.
2. Is presented with experimental results, data, or concrete examples.
3. Can be traced to specific methods, results, or discussion sections.
4. Is not from the abstract or introduction.

Return ONLY the following JSON:

```
{
    "evidence_sets": [
        {
            "claim_id": number,
            "evidence": [
                {
                    "evidence_id": number,
                    "evidence_text": "specific evidence",
                    "strength": "strong/moderate/weak",
                    "limitations": "key limitations",
                    "location": "section/paragraph",
                    "exact_quote": "verbatim text"
```

```
            }
        ]
    }
  ]
}
```

Conclusion Evaluation Prompt

Analyze these claims and their evidence: `{analysis_text}`
For each claim-evidence pair, evaluate:

1. Whether the evidence justifies the claim.
2. The overall strength of support.
3. Any important limitations.

Return ONLY the following JSON:

```
{
    "conclusions": [
        {
            "claim_id": number,
            "conclusion_justified": true/false,
            "robustness": "high/medium/low",
            "key_limitations": "specific limitations",
            "confidence_level": "high/medium/low"
        }
    ]
}
```

A.3 One-by-One Prompt

Claims Extraction Prompt

Analyze this research paper and extract ALL possible claims made by the authors. *Paper text: {text}*
Your task is to identify all statements in the text that meet the following criteria for a claim:

1. Makes a specific, testable assertion about results, methods, or contributions.
2. Represents a novel finding, improvement, or advancement.
3. Presents a clear position or conclusion.

Make sure to:

- Include both major and minor claims.
- Don't miss any claims.
- Present each claim as a separate item.

Return ONLY the following JSON structure:

```
{
    "claims": [
        {
```

```
            "claim_id": 1,
            "claim_text": "statement of the claim",
            "location": "section/paragraph where this claim appears",
            "claim_type": "Nature of the claim",
            "exact_quote": "complete verbatim text containing the claim"
        }
    ]
}
```

Evidence Analysis Prompt

Paper text: {text}
For the following claim from the paper: "{claim['claim_text']}"
Please identify relevant evidence that:

1. Directly supports or contradicts the claim's specific assertion.
2. Is presented with experimental results, data, or methodology.
3. Can be traced to specific methods, results, or discussion sections.
4. Is not from the abstract or introduction.

If NO evidence is found for the given Claim, return:

```
{
    "claim_id": {claim['claim_id']},
    "evidence": [],
    "no_evidence_reason": "Explain why no evidence was found (e.g., 'Claim is unsupported', '
        ↪ Claim is theoretical without empirical evidence', etc.)"
}
```

ELSE: Return ONLY the following JSON structure:

```
{
    "claim_id": {claim['claim_id']},
    "evidence": [
        {
            "evidence_id": 1,
            "evidence_text": "specific experimental result/data point",
            "evidence_type": "primary/secondary",
            "strength": "strong/moderate/weak",
            "limitations": "stated limitations or assumptions",
            "location": "specific section & paragraph",
            "exact_quote": "verbatim text from paper"
        }
    ]
}
```

Conclusion Analysis Prompt

Paper text: {text}
Analyze the following claim and its supporting evidence: {single_claim_analysis}
Provide a comprehensive conclusion analysis following these guidelines:

1. Evidence Assessment:
    - Evaluate the strength and quality of ALL evidence presented.
    - Consider both supporting and contradicting evidence.
    - Assess the methodology and reliability of evidence.

2. Conclusion Analysis:
   - Determine what the authors concluded about this specific claim.
   - Evaluate if the conclusion is justified by the evidence.
   - Consider the relationship between evidence quality and conclusion strength.
3. Robustness Evaluation:
   - Assess how well the evidence supports the conclusion.
   - Consider methodological strengths and weaknesses.
   - Evaluate the consistency of evidence.
4. Limitations Analysis:
   - Identify specific limitations in both evidence and conclusion.
   - Consider gaps in methodology or data.
   - Note any potential biases or confounding factors.

Return ONLY the following JSON structure:

```
{
    "conclusions": [
        {
            "claim_id": {claim_id},
            "author_conclusion": "detailed description of authors' conclusion based on evidence
                ↪ ",
            "conclusion_justified": true/false,
            "justification_explanation": "detailed explanation of why conclusion is/isn't
                ↪ justified",
            "robustness_analysis": "comprehensive analysis of evidence strength and reliability
                ↪ ",
            "limitations": "specific limitations and caveats",
            "location": "section/paragraph where conclusion appears",
            "evidence_alignment": "analysis of how well evidence aligns with conclusion",
            "confidence_level": "high/medium/low based on evidence quality"
        }
    ]
}
```

780

B Additional Details on Annotation

B.1 Annotator Guidelines

- Select one recent research paper in the field of artificial intelligence or machine learning.
- Prioritize papers published in 2024 to ensure relevance to current developments.
- When possible, select a paper with fewer than 20 pages to facilitate thorough annotation.
- Avoid papers with heavily mathematical content to ensure accessibility.
- Complete all annotation tasks independently, without employing large language models for assistance at any stage of the process.

Task Description
Your task is to identify all statements in the text that qualify as claims under the following criteria:

1. **Specificity**: The statement makes a specific, testable assertion about results, methods, or contributions.
2. **Novelty**: The statement represents a novel finding, improvement, or advancement.
3. **Clarity**: The statement presents a clear position or conclusion.

Requirements

- Include both major and minor claims.
- Ensure no claim is overlooked.
- Present each claim as a separate item.

Evidence Identification
For each identified claim, find and document relevant evidence that:

1. **Relevance**: Directly supports or contradicts the claim's specific assertion.
2. **Concrete Support**: Is presented with experimental results, data, or concrete examples.
3. **Traceability**: Can be traced to specific methods, results, or discussion sections in the text.
4. **Exclusions**: Evidence must not be derived from the abstract or introduction sections of the text.

Conclusion Analysis

- **Justification**: Evaluate whether the conclusions drawn in the text are justified by the evidence provided.

Annotation Format
Each annotation should be formatted as follows:

```
{
    "Claim_id": "<unique_identifier>",
    "Claim_text": "<text_of_the_claim>",
    "Evidence_text": "<text_supporting_or_contradicting_the_claim>",
    "Justification_Conclusion": "<evaluator's_comment_on_evidence_justification>"
}
```

B.2 Inter-Annotator Agreement Methodology

To evaluate the reliability of the CLAIM-BENCH annotations, we calculated Inter-Annotator Agreement on a subset of 30 papers, each annotated by two different annotators on the Claims and the Evidence. For each of the claims and the evidences, we take one set ("set A") as the ground truth and compute the F1-score of the other set ("set B"). Considering the symmetry, we also computed the F1-score swapping sets A and B, and reported the averaged F1-score. We chose F1 because our annotation task (identifying and linking spans) closely parallels standard information extraction tasks, where F1 is a standard evaluation measure balancing precision and recall; this reflects the need for agreement on both the correctness and comprehensiveness of annotations. Furthermore, in our LLM-based assessment, the requested F1-score provides an interpretable measure of semantic concordance derived from the model's understanding of semantic equivalence beyond exact string matches.

B.3 Annotation Tool

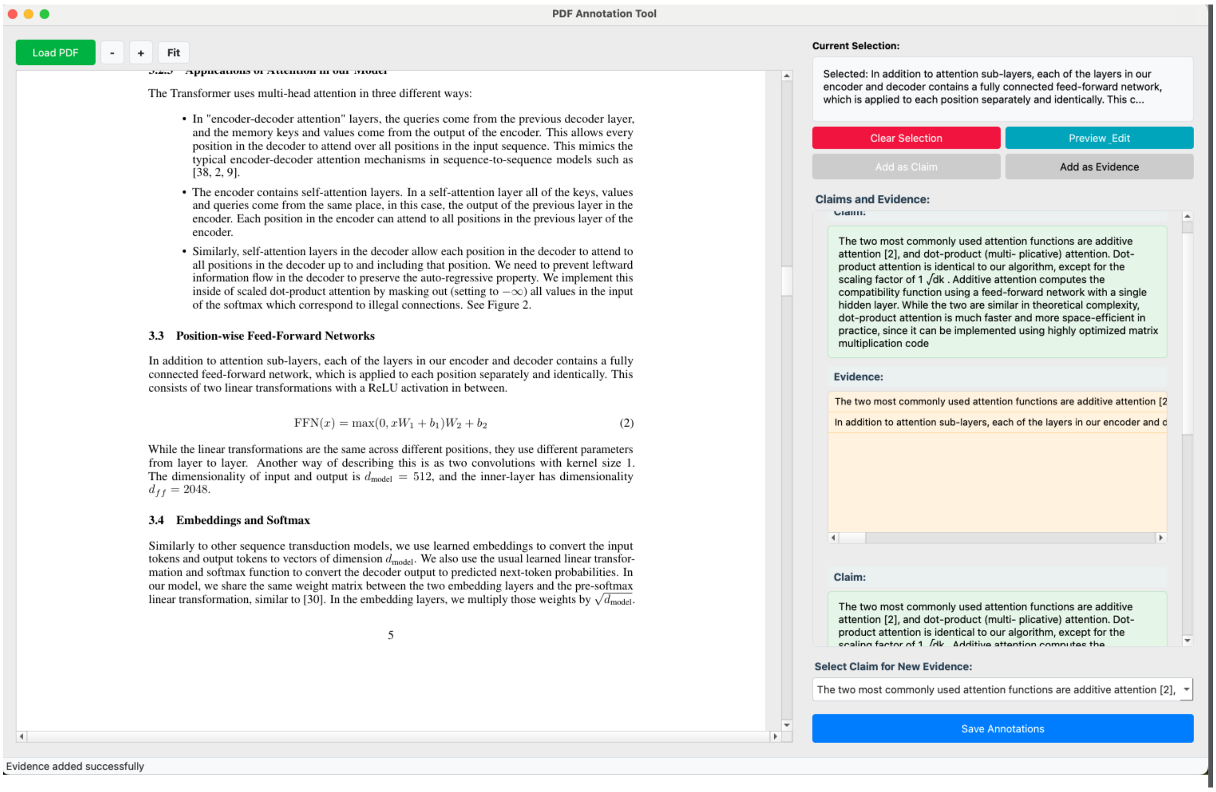
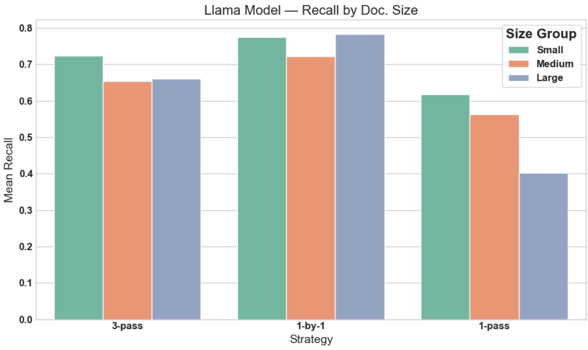
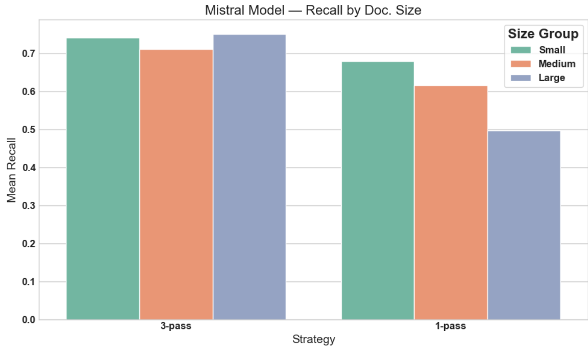


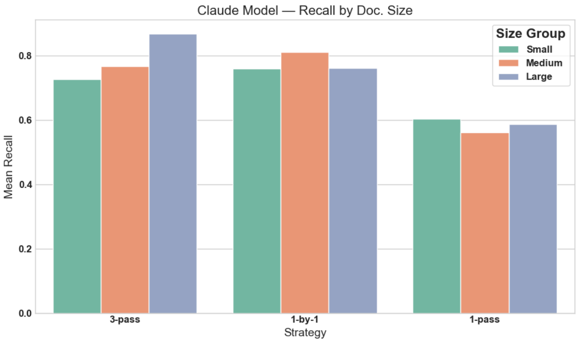
Figure 5: The custom annotation tool interface used for CLAIM-BENCH dataset creation, enabling direct PDF text selection and structured labeling (e.g., 'Add as Claim' button) of claim-evidence pairs.



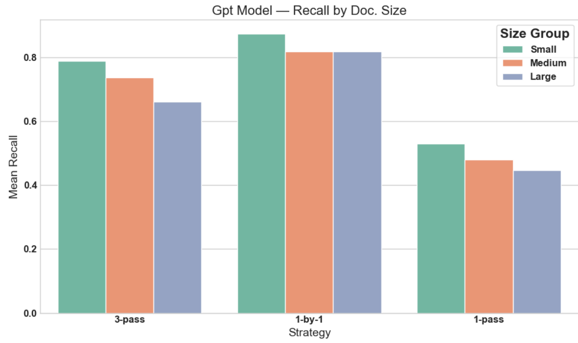
(a) LLAMA Recall



(b) Ministral Recall



(c) Claude Recall



(d) GPT-4 Recall

Figure 6: Mean recall by document size groups (small, medium, large) for different models and prompting strategies, illustrating performance trends across increasing token counts.

C Impact of Documents' Token Length

Figure 6 plots mean recall for three prompting strategies—Three-Pass, One-by-One, and Single-Pass—across three document-length buckets ($<$ 15 k, 15–20 k, $\geq$ 20 k tokens). A closer reading of the bars yields three key observations:

1. **Performance drops are tied to the strategy more than the model size.**
   - For every model, the Single-Pass run shows the steepest decline as documents grow.
   - Example: LLaMA's recall plunges from about 0.60 in small papers to roughly 0.40 in $\geq$20 k-token papers under Single-Pass.
2. **Once an iterative strategy is used, the size-related gap all but disappears.**
   - Iterative prompting (Three-Pass or One-by-One) largely neutralises length effects—even for the smaller models.
   - LLaMA 70B: In One-by-One mode the large-document group matches or exceeds the small-document group ($\approx$ 0.78 vs $\approx$ 0.76).
   - Ministral 8B: Three-Pass recall stays virtually flat ($\sim$ 0.72–0.75) across all three size buckets; the length penalty only appears in Single-Pass.
3. **Larger models still benefit, but their advantage is greatest with fine-grained prompts.**
   - Claude 3.5 Sonnet: Recall rises with document size under Three-Pass ($\approx$ 0.72 $\rightarrow$ 0.85), and remains $\geq$ 0.75 in One-by-One.
   - GPT-4-Turbo: One-by-One keeps recall at or above 0.80 for medium- and large-size papers; the drop to $\sim$ 0.66 for large papers occurs only in Three-Pass, not in Single-Pass.

The figure shows that prompt granularity is the dominant lever for long-context recall. Single-pass prompting amplifies context-window limits—especially in smaller models—but iterative, claim-level prompting (Three-Pass and One-by-One) recovers performance, sometimes even improving it as the text grows. Larger models are naturally more stable, yet they, too, realise their full potential only when given finer-grained, multi-step instructions.

C.1 Sentence Distance Detailed Analysis

	3-pass				1-pass				1-by-1			
	Count	Max	Mean	Var	Count	Max	Mean	Var	Count	Max	Mean	Var
GPT	203	696	93.8	10640.4	152	658	98.5	14738.0	396	708	90.2	9798.3
CLAUDE	174	2226	122.2	39147.4	250	2222	90.7	33122.3	639	2230	119.4	33673.9
GEMINI	84	720	107.4	23584.2	194	710	72.8	18017.5	N/A	N/A	N/A	N/A
LLAMA	183	2226	98.1	35974.1	145	2228	109.1	71857.5	659	2228	95.1	34207.0
MISTRAL	38	357	75.9	8030.5	166	632	64.2	8316.9	N/A	N/A	N/A	N/A
PHI	279	2282	130.6	114904.2	294	2232	121.4	56085.7	347	579	105.9	13188.2

Figure 7: Aggregated statistics of the sentence_gap metric Count, Max, Mean, and Variance (Var)—for each model under the three prompting strategies (Three-Pass, One-pass, and One-by-One). Larger counts and wider gaps (e.g., Claude and LLaMA exceeding 2,200-sentence links in One-by-One) reflect broader retrieval, whereas smaller models such as Mistral keep distances short and variance low. "N/A" indicates the model-strategy combination was not executed.