Can AI Validate Science ? Benchmarking LLMs on Claim → Evidence Reasoning in AI Papers

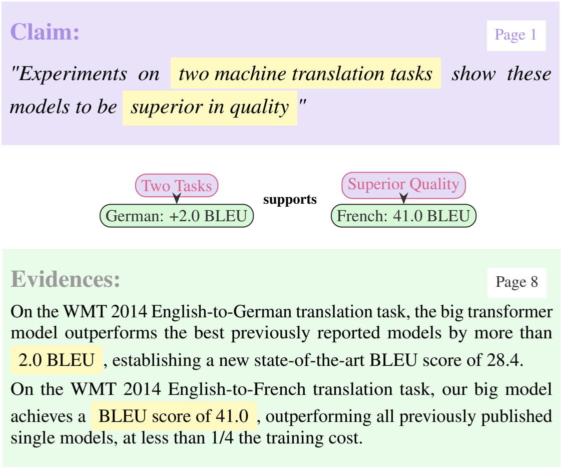
Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly vital tools for academic research. A core ability for these tools is to identify claims and validate them against supporting evidence, and there is yet to be an integrated benchmark to evaluate the claim-evidence reasoning capabilities. To address this gap, we introduce CLAIM-BENCH[1], a new benchmark designed to jointly evaluate three critical skills in claim-evidence reasoning: claim extraction, evidence extraction, and claim-evidence link validation. CLAIM-BENCH contains over 300 manually annotated claim-evidence pairs from AI research papers. We evaluate six LLMs with three prompting strategies using CLAIM-BENCH. We find that closed-source models like GPT-4 and Claude consistently outperform open-source counterparts, though even the best models reach a peak F1-score of only 0.59 on claim identification. This difficulty stems primarily from resolving long-range dependencies, as models struggle to connect claims with evidence dispersed throughout a document. Consequently, we show that iterative prompting strategies, which decompose the task, can boost the number of retrieved claim-evidence pairs by over 4x compared to the baseline single-pass prompt, substantially improving recall but at a significant computational cost. CLAIM-BENCH establishes a much-needed standard for assessing deep scientific comprehension in LLMs, providing both a diagnostic framework to understand current limitations and a path toward building more reliable, deep-reasoning systems.

1 Introduction

Large Language Models (LLMs) have become a pivotal 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. Researchers have increasingly assigned critical tasks to these models—from content summarization (Agarwal et al., 2025) to hypothesis generation (Vladika and Matthes, 2023). Recently, agentic frameworks use LLMs for automated peer review (Checco et al., 2021; Agarwal et al., 2025; Lu et al., 2024; Jin et al., 2024; Sun et al., 2024b). Behind these tasks, a fundamental question emerges: to what extent do these LLMs truly understand scientific papers beyond surface-level pattern recognition? Despite their widespread use and promising outcomes, there remains uncertainty about the depth and accuracy of their reasoning capabilities in the complex context of scientific papers.



"Experiments on two machine translation tasks show these models to be superior in quality"

On the WMT 2014 English-to-German translation task, the big transformer model outperforms the best previously reported models by more than 2.0 BLEU, establishing a new state-of-the-art BLEU score of 28.4.

On the WMT 2014 English-to-French translation task, our big model achieves a BLEU score of 41.0, outperforming all previously published single models, at less than 1/4 the training cost.

Figure 1: Example of a claim and its supporting evidences from Vaswani et al. (2017).

Scientific papers are long documents with intricate relationships. They are structured around claims and are supported by evidence. The ability to accurately identify and reason about these claim-evidence pairs is essential for validating scientific

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

findings and ensuring research integrity, making it a critical test of LLMs' comprehension depth. Unlike surface-level tasks such as summarization, question answering, claim-evidence identification requires global reasoning across paper sections, synthesis of dispersed information, and a nuanced understanding of logical dependencies. The ability to reason about research claims and evidences has been an active research area.

Existing benchmarks evaluate the fact-checking capabilities in various settings. For example, SCIFACT (Wadden et al., 2020) validates expert-written scientific claims using the abstracts of research papers. We defer to the review of Vladika and Matthes (2023). More recent works considered the claim identification and verifications within publications (Lu et al., 2023; Wei et al., 2023), the check-worthiness of claims (Liu et al., 2025), and the retrieval of evidence (Deng et al., 2025). While these benchmarks involve claims and evidences, they do not measure a finer-grained verification task: whether the evidence presented in a full scientific paper supports its claims. This claim-evidence reasoning capability is precisely what we target.

In this paper, we present CLAIM-BENCH. This benchmark consists of a new dataset with over 300 claim-evidence pairs, expert-annotated from full-length AI research papers. It is specifically designed to test the challenging task of long-range scientific argument tracing, where claims must be validated against evidence dispersed throughout a document.

By evaluating six state-of-the-art LLMs on CLAIM-BENCH, we find that larger models (e.g., GPT-4-Turbo, Claude 3.5) maintain high recall on lengthy documents with iterative prompting, while smaller models (e.g., LLaMA, Ministral) see significant performance drops under single-pass strategies. These findings highlight crucial areas for enhancing long-context comprehension and inform the development of reliable AI tools for scientific research. CLAIM-BENCH thus sets a new standard for evaluating deep scientific comprehension in LLMs.

2 Related Work

Claim Extraction and Verification Prior work on scientific claim analysis has largely focused on isolated sub-tasks like citation-reference validation (Zhang and Abernethy, 2024), rather than end-to-end claim-evidence reasoning within a full document. The influential SCIFACT, SciFact-Open benchmarks (Wadden et al., 2020, 2022) test the verification of external claims. Li et al. (2021) focuses on evidence extraction tied to specific discourse elements. Works that engage with full-text articles often stop short of the complete reasoning task. Blake (2010), Achakulvisut et al. (2020), and Wei et al. (2023) developed methods for claim *identification* within publications but didn't operationalize the crucial step of linking claims to dispersed evidence. Similarly, Claimify (Metropolitansky and Larson, 2025) addresses the generation of high-quality claims in isolation, without tracing them back to supporting evidence within a source document . In contrast, CLAIM-BENCH requires this full, integrated reasoning process on complete papers.

AI for Science LLMs have significantly advanced scientific workflows, facilitating tasks such as peer review. Building on early work in AI-assisted peer review (Checco et al., 2021), recent 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.

Benchmarks Long-context benchmarks, such as SCBENCH (Li et al., 2025a), MMLongBench-Doc (Ma et al., 2024), and LongGenBench (Wu et al., 2025), have assessed LLMs' ability to process extended inputs and maintain coherence, focusing on tasks like document summarization and long-form generation. Recent works, including AI Scientist (Lu et al., 2024), LitLLM (Agarwal et al., 2025), and ChatCite (Li et al., 2025b) benchmarked LLMs on tasks such as literature review and hypothesis generation, while ScienceAgentBench (Chen et al., 2025) and SCBENCH (Li et al., 2025a) probe multi-step reasoning and long-context understanding. 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 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. Work in neuroscience, for example, shows LLMs surpassing expert predictions of experimental outcomes (Luo et al., 2025), yet such success doesn't imply reasoning comprehension. Our work focuses on research papers with more complex and dispersed relationships, such as claims supported by evidence across multiple sections. CLAIM-BENCH evaluates how LLMs synthesize these intricate connections, testing their global reasoning and coherence, reflecting the unique demands of scientific texts. This gap is underscored by research from adjacent domains. For instance, works calling for crucial ethical considerations, such as the need for transparency and accountability in AI-driven research (Lissack and Meagher, 2024), or expanding evaluation to include multimodal data (Song et al., 2024), also highlight the absence of a targeted benchmark for claim-evidence validation across long, complex scientific texts—a gap CLAIM-BENCH aims to fill.

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. While these approaches address some limitations of single-pass systems, their primary focus remains on generating content, not validating complex logical relationships. Similarly, tools for hypothesis testing like AIGS (Liu et al., 2024b) and LLM-Assisted Hypothesis Generation (Vladika and Matthes, 2023), and graph-based methods for structured creativity (Leng et al., 2024), fall short of validating inter-linked arguments at scale.

3 Methodology

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. Following specific guidelines (Appendix B.1), annotators selected papers and identified their core scientific claims. The selection criteria for papers were designed to focus the benchmark on text-based reasoning: we chose recent (2024), non-math-intensive articles under 20 pages to ensure a diverse set of current AI/ML topics while avoiding model memorization and bottlenecks from symbolic reasoning.

Statistic	Value
Dataset Overview	
Total Annotations	346
Unique Papers	100
Unique Claims	331
Unique Evidence Passages	335
Duplicate Claims	15
Per-Paper Statistics	
Claims per Paper (Avg/Med/Range)	3.33 / 3 / 1–8
Evidence per Paper (Avg/Med/Range)	3.67 / 3 / 1–9
Content Length (Words)	
Claim Length (Avg/Med/Range)	22 / 20 / 8–43
Evidence Length (Avg/Med/Range)	28 / 25 / 10–40

Table 1: Dataset Summary Statistics

Annotation Tool To facilitate easier annotations, we developed a PDF annotation tool, it lets users load a paper, drag a pointer over any sentence(s) 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 (Appendix B.4).

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 and found moderate to substantial inter-annotator agreement (details in Appendix B.3), confirming that CLAIM-BENCH is a reliable benchmark.

3.2 Evaluation Metrics

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 the prompting techniques.

Precision measures the accuracy of the model's predictions, reflecting its ability to avoid generating spurious claims or evidence from the scientific texts. Recall quantifies the model's ability to identify all relevant spans from the human-annotated ground truth, measuring its comprehensiveness in

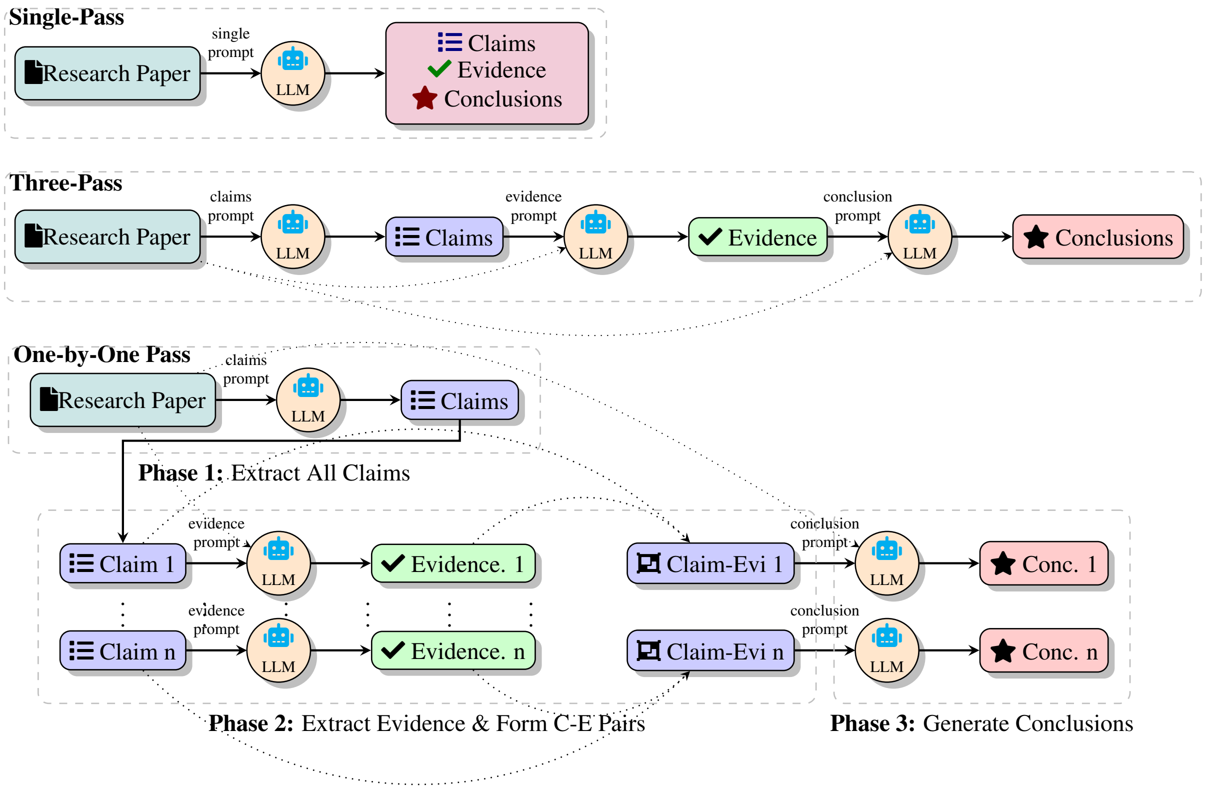


Figure 2: Three methods to prompt LLMs to analyze the papers. **Single-Pass:** Full paper processing with one prompt. **Three-Pass:** Sequential claim → evidence → conclusion extraction. **One-by-One Pass**: Individual evidence retrieval per claim.

response to our prompts. The F1-score, as the harmonic mean of precision and recall, provides a single, balanced metric to compare the overall efficacy of the different LLMs and prompting strategies we test. The sentence_gap metric measures the average sentence-level distance between a retrieved claim and each of its associated retrieved evidence.

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

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 therefore captures how far a model must search across the paper to link a claim with the supporting evidences. It is particularly valuable for quantifying the models' ability to handle textual relationships over extended contexts.

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.

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

Figure 2 shows three distinct prompting methods to assess and enhance model performance on claim-evidence identification tasks.

Single-Pass As a baseline, 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 & conquer" strategy from prior research (Zhang et al., 2024), we then deconstruct the task into sequential stages. In the first stage, the model identifies claims using a dedicated prompt, these claims are supplied to the next stage, where separate prompts elicit corresponding evidences. Finally, we combine the identified claims & 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 3, models exhibit a clear precision-recall trade-off: settings that achieve 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 8), 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. Similar pattern holds in evidence extraction as well.

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 2: The highest performance (across all strategies) for Claim (C) and Evidence (E) extraction. Metrics reported are F1, Precision (P), and Recall (R).

Analyzing claim versus evidence extraction separately reveals distinct performances among LLMs (see Table 2). 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

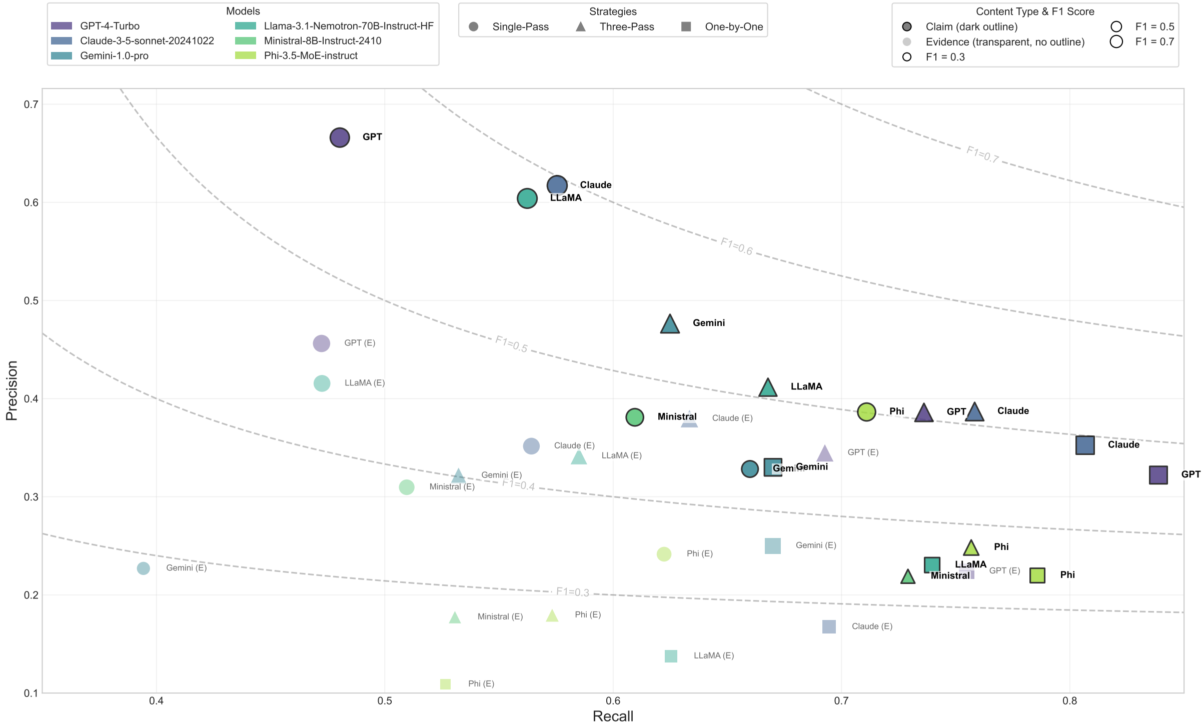


Figure 3: Precision vs. Recall for claim (solid markers) and evidence (transparent markers) identification across models and strategies (shapes: Single-Pass ●, Three-Pass ▲, One-by-One ■). Models show higher precision for claims, higher recall for evidence, with most results below $F_1 = 0.7$.

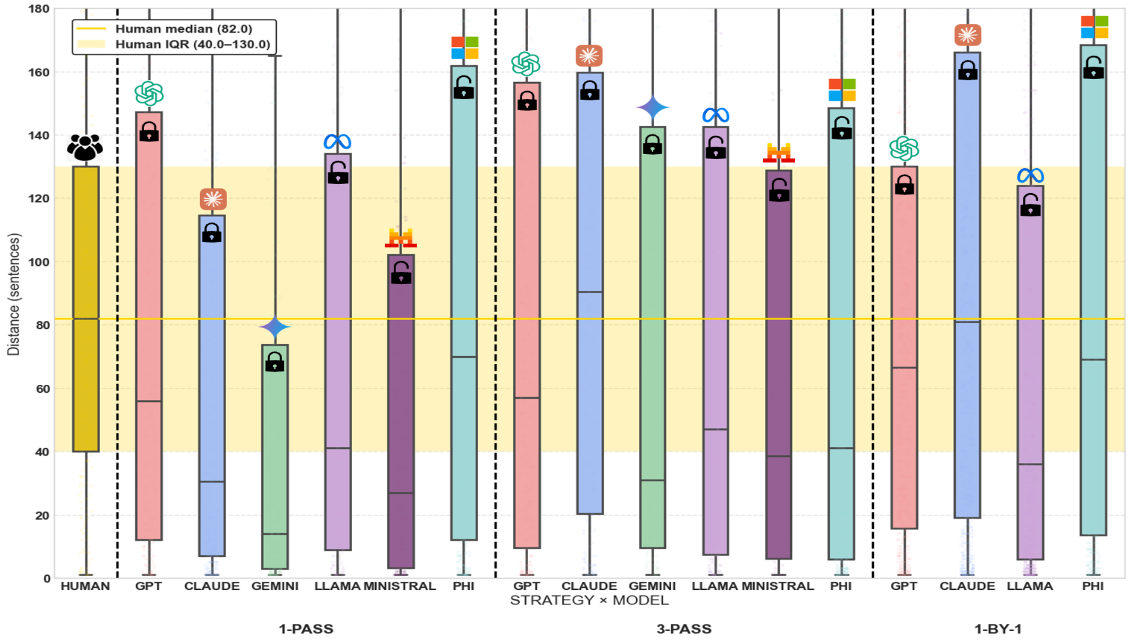


Figure 4: Sentence distance distribution (box plots) between claims and linked evidence vs. Human baseline (*leftmost*). LLMs, especially with iterative strategies, link over longer distances than humans, showing capability but potential noise.

the common trends, the models exhibit distinct patterns. For instance, Claude and LLaMA demonstrate high recall in evidence extraction but with substantial variability in linking distances (Claude: mean 119.4 sentences, SD = 183.5; LLaMA: mean 95.1 sentences, SD = 184.9), suggesting increased noise and inconsistent performance. Conversely, Ministral maintains lower linking distances (mean

75.9 sentences, SD = 89.4), signifying a more cautious and controlled approach.

5.4 Impact of Prompting 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 a significant SD (107.2) 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 SD (Claude: 183.5, LLaMA: 185.0). Phi also achieves substantial coverage (347 pairs) with a notable SD (114.8).

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 (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 8 (in Appendix C).

5.6 Types of Claims and Evidences

To further understand the nature of the claim-evidence reasoning task and the models' behavior, we categorize the claims and the evidences identified by both humans and LLMs. The categorization, developed by synthesizing and extending established types from the scientific validation literature, provides a qualitative lens for our analysis. Full descriptions are in Appendix C.1, and the results are in Table 3 and Table 5.

Many models exhibited a strong bias for "comparative" content over other types. For example, Claude identified 37.3% of claims as comparative, exceeding the human baseline of 23.6%. Rather than being "surface-level", we believe this occurs because comparative claims contain explicit keywords (e.g., "outperforms") that are easy for models to detect. This suggests that iterative prompting strategies, which break the task down, are crucial for calibrating models to look beyond these lexical signals and identify a more balanced set of claims.

Models had different priorities than humans when identifying important claims. GPT and Gemini aligned with humans by prioritizing methodological claims (e.g., GPT: 32.4% vs. human: 42.1%). In contrast, Claude and LLaMA favored claims about empirical results and comparisons.

Models consistently struggled with claims requiring abstract or deep reasoning. They underrepresented theoretical claims (e.g., Claude: 2.5% vs. human: 7.5%) significantly undervalued expert evidence (we define expert evidence as the authors' synthesis or interpretation). Models identified the "expert evidences" less than 3% of the time (vs. 13.9% for humans), suggesting they can extract isolated facts but fail at the higher-order task of connecting data to an author's conclusions, a core component of deep scientific comprehension.

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-

	Meth	Emp	Comp	Theo	Caus		Meth	Emp	Comp	Theo	Caus
Human	**42.1 (1)**	**24.2 (2)**	**23.6 (3)**	**7.5 (4)**	**2.5 (5)**						
Claude						**Llama**					
1	19.7 (3)	34.5 (2)	37.3 (1)	2.5 (5)	6.0 (4)	1	35.5 (1)	26.4 (2)	26.4 (2)	7.8 (4)	3.9 (5)
3	18.6 (3)	37.5 (1)	34.5 (2)	5.0 (4)	4.4 (5)	3	28.9 (2)	30.7 (1)	26.5 (3)	8.0 (4)	5.9 (5)
O	20.9 (3)	33.8 (2)	34.2 (1)	4.8 (5)	6.2 (4)	O	25.2 (3)	33.4 (1)	26.8 (2)	8.6 (4)	6.0 (5)
GPT						**Ministral**					
1	32.4 (1)	24.1 (3)	30.6 (2)	3.7 (5)	9.3 (4)	1	33.2 (1)	24.7 (3)	28.4 (2)	12.0 (4)	1.7 (5)
3	31.4 (1)	30.4 (2)	24.1 (3)	9.8 (4)	4.3 (5)	3	31.0 (1)	26.5 (2)	19.5 (3)	15.6 (4)	7.5 (5)
O	29.5 (1)	28.7 (2)	25.9 (3)	7.5 (5)	8.4 (4)	O	33.0 (1)	26.2 (3)	26.7 (2)	12.0 (4)	2.2 (5)
Gemini						**Phi**					
1	33.5 (1)	29.3 (2)	25.1 (3)	8.5 (4)	3.6 (5)	1	30.6 (2)	31.2 (1)	27.3 (3)	7.3 (4)	3.6 (5)
3	30.6 (1)	29.5 (2)	29.0 (3)	8.2 (4)	2.7 (5)	3	33.3 (2)	34.2 (1)	16.3 (3)	10.8 (4)	5.5 (5)
O	37.2 (1)	31.1 (2)	17.4 (3)	12.3 (4)	1.9 (5)	O	36.7 (1)	17.1 (3)	35.9 (2)	6.5 (4)	3.9 (5)

Table 3: The percentage and rank (in parentheses) of five categories of claims identified by the models employing the strategies, compared to the ground truth identified by humans. **Categories:** Meth=Methodological, Emp=Empirical, Comp=Comparative, Theo=Theoretical, Caus=Causal. **Strategies:** 1=single pass, 3=3-pass, O=one-by-one.

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. A closer look at the models' errors reveals two primary failure modes. The first is over-generation of plausible but incorrect links, prevalent in high-recall models like LLaMA and Claude. These models often identify claim-like and evidence-like sentences in isolation but fail to validate the precise logical connection between them, resulting in low precision. The second failure mode is missed context due to long-range dependencies. This is evident when a claim made in the introduction is supported by a specific result in a table within the results section. Models, especially smaller ones like Ministral or any model using a single-pass prompt, frequently fail to bridge this large sentence_gap, leading to false negatives. These failures underscore that the primary challenge is not just text extraction, but robust, long-distance logical reasoning.

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. Our qualitative analysis further reveals systematic biases and error patterns in current LLM capabilities, underscoring CLAIM-BENCH's critical role in advancing rigorous scientific validation tasks. Moreover, our comparative analysis across three 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.

Titipat Achakulvisut, Tananun Ruangrong, Natasha Bilenko, Chandra Bhagavatula, and Peter Jansen. 2020. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*.

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.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics*, 43(2):173–189.

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.

Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. The next phase of scientific fact-checking: advanced evidence retrieval from complex structured academic papers. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 436–448.

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.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. SciClaimHunt: A large dataset for evidence-based scientific claim verification. *arXiv preprint arXiv:2502.10003*.

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

Xiangci Li, Aixin Sun, and Shafiq Joty. 2021. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562, Online. Association for Computational Linguistics.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. 2025a. 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. 2025b. 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.

Hao Liu et al. 2024a. Retrieval augmented scientific claim verification. *JAMIA Open*, 7(1):ooae021.

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. 2024b. 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.

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. Association for Computational Linguistics.

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*.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. *Preprint*, arXiv:2502.10855.

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

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

Dingjie Song, 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*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

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

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, Lucy Lu Wang, Arman Cohan, 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.

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.

Xin Wei, Yanzhe Zhang, Yifei Liu, and Xuanjing Huang. 2023. ClaimDistiller: Scientific Claim Extraction with Supervised Contrastive Learning. In *CEUR Workshop Proceedings*, volume 3451 of *Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI Informetrics (AII2023)*.

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*.

Boya Zhang et al. 2025. A dataset for evaluating clinical research claims in large language models. *Scientific Data*, 12(1):86.

Tianmai M. Zhang and Neil F. Abernethy. 2024. Detecting reference errors in scientific literature with large language models. *Preprint*, arXiv:2411.06101.

Yizhou Zhang, Lun Du, Defu Cao, Qiang Fu, and Yan Liu. 2024. An examination on the effectiveness of divide-and-conquer prompting in large language models. *Preprint*, arXiv:2402.05359.

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"
        }
    ]
}
```

B Additional Details on Annotation

B.1 Paper Selection

- 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.

B.2 Annotator Guidelines

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.

Annotators followed explicit guidelines for identifying claims and evidence. Claims were annotated based on being novel, specific, and clearly stated scientific assertions, while evidence included supporting sentences explicitly linked to these claims. Annotators were instructed to select the minimal text span that fully conveyed the claim or evidence, avoiding unnecessary contextual sentences.

B.3 Inter-Annotator Agreement Methodology

To evaluate CLAIM-BENCH annotation reliability, we calculated Inter-Annotator Agreement on a subset of 30 papers, each annotated by two different annotators. For claims and evidence, we computed the F1-score treating each annotator alternately as ground truth to ensure symmetry. F1-score was chosen for its relevance to information extraction tasks, balancing precision and recall.

Additionally, we automated Cohen's κ computation using an LLM assistant (Gemini 2.5) on the 30-paper subset. For each paper, the LLM assistant performed four steps clearly defined below: (i) Extracted raw annotation files, (ii) Built binary vectors indicating claim/evidence presence per sentence (1 for presence, 0 otherwise), (iii) Populated the 2×2 contingency table (elements a, b, c, d) where:

- a: sentences marked by both annotators,
- b: sentences marked only by annotator 1,
- c: sentences marked only by annotator 2,
- d: sentences not marked by either annotator,

(iv) Computed Cohen's κ as:

$$P_o = \frac{a+d}{N}, \quad P_e = \frac{(a+b)(a+c)+(c+d)(b+d)}{N^2}, \quad \kappa = \frac{P_o - P_e}{1 - P_e}$$

The automated procedure was validated manually on a sample of 10 papers, confirming arithmetic accuracy. The results yielded $\kappa = 0.66$ (substantial agreement) for claims and $\kappa = 0.30$ (fair agreement) for evidence. The lower agreement for evidence was anticipated, given sparse and dispersed evidence sentences (<0.3% of total text). Minor boundary discrepancies or multiple valid evidence spans legitimately lowered agreement. Nonetheless, these scores affirm CLAIM-BENCH's robustness as a challenging yet reliable benchmark.

Cohen's κ Agreement Prompt

Paper filename: {pdf_name} Total sentences in paper: `{total_sentences}`

You are given two raw annotation lists for *claim identification*—one from Annotator 1 and one from Annotator 2. Follow the steps below **exactly** to compute **Cohen's** κ:

1. **Vector Construction** Build two binary vectors of length $N = \{total\_sentences\}$:
   - `1` if the sentence was marked as a claim by the annotator.
   - `0` if the sentence was *not* marked as a claim.
2. **Contingency Table** Using the two vectors, populate the 2×2 table:

	Ann $2 = 1$	Ann $2 = 0$
Ann $1 = 1$	a	b
Ann $1 = 0$	c	d

3. **Compute** κ

$$P_o = \frac{a+d}{N}$$

$$P_e = \left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right) + \left(\frac{c+d}{N}\right)\left(\frac{b+d}{N}\right)$$

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

4. **Return *only* the JSON below**:

```
{
    "kappa_claims": 0.00
}
```

Raw Annotations – Annotator 1: {`raw_annotations1`}

Raw Annotations – Annotator 2: {`raw_annotations2`}

Example Output: Cohen's κ Calculation

We compute Cohen's κ for **claim identification** on a paper with $N = 667$ sentences.

Annotation statistics

- Annotator 1 marked **5** sentences as claims.
- Annotator 2 marked **6** sentences as claims.
- Overlap (both $claim = 1$): **4** sentences.

Contingency table

	Ann 2 $= 1$	Ann 2 $= 0$	Row Tot.
Ann 1 $= 1$	4	1	5
Ann 1 $= 0$	2	660	662
Col. Tot.	6	661	667

Calculations

$$P_o = \frac{a+d}{N} = \frac{4+660}{667} \approx 0.9955,$$

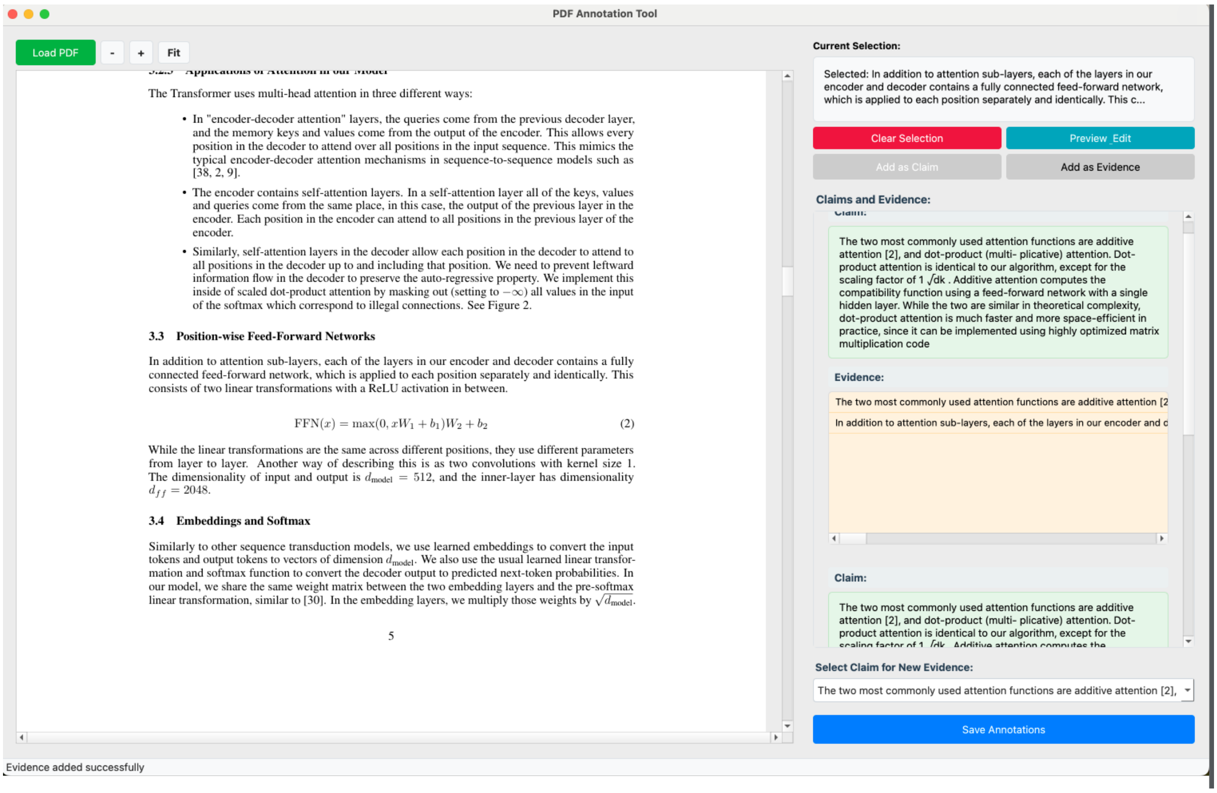
$$\begin{aligned} P_e &= \left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right) + \left(\frac{c+d}{N}\right)\left(\frac{b+d}{N}\right) \\ &= \left(\tfrac{5}{667}\right)\left(\tfrac{6}{667}\right) + \left(\tfrac{662}{667}\right)\left(\tfrac{661}{667}\right) \approx 0.98375, \end{aligned}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0.99550 - 0.98375}{1 - 0.98375} \approx 0.7231.$$

Result JSON

```
{
    "kappa_claim": 0.7231
}
```

B.4 Annotation Tool



3.2.3 Applications of Attention in our Model

The Transformer uses multi-head attention in three different ways:

- In "encoder-decoder attention" layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence. This mimics the typical encoder-decoder attention mechanisms in sequence-to-sequence models such as [38, 2, 9].
- The encoder contains self-attention layers. In a self-attention layer all of the keys, values and queries come from the same place, in this case, the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder.
- Similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position. We need to prevent leftward information flow in the decoder to preserve the auto-regressive property. We implement this inside of scaled dot-product attention by masking out (setting to $-\infty$) all values in the input of the softmax which correspond to illegal connections. See Figure 2.

3.3 Position-wise Feed-Forward Networks

In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{\text{model}} = 512$, and the inner-layer has dimensionality $d_{ff} = 2048$.

3.4 Embeddings and Softmax

Similarly to other sequence transduction models, we use learned embeddings to convert the input tokens and output tokens to vectors of dimension d_{model}. We also use the usual learned linear transformation and softmax function to convert the decoder output to predicted next-token probabilities. In our model, we share the same weight matrix between the two embedding layers and the pre-softmax linear transformation, similar to [30]. In the embedding layers, we multiply those weights by $\sqrt{d_{\text{model}}}$.

5

Evidence added successfully

Current Selection:

Selected: In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This c...

Claims and Evidence:

Claim:

The two most commonly used attention functions are additive attention [2], and dot-product (multi- plicative) attention. Dot-product attention is identical to our algorithm, except for the scaling factor of 1 √dk . Additive attention computes the compatibility function using a feed-forward network with a single hidden layer. While the two are similar in theoretical complexity, dot-product attention is much faster and more space-efficient in practice, since it can be implemented using highly optimized matrix multiplication code

Evidence:

The two most commonly used attention functions are additive attention [2

In addition to attention sub-layers, each of the layers in our encoder and d

Claim:

The two most commonly used attention functions are additive attention [2], and dot-product (multi- plicative) attention. Dot-product attention is identical to our algorithm, except for the scaling factor of 1 √dk . Additive attention computes the

Select Claim for New Evidence:

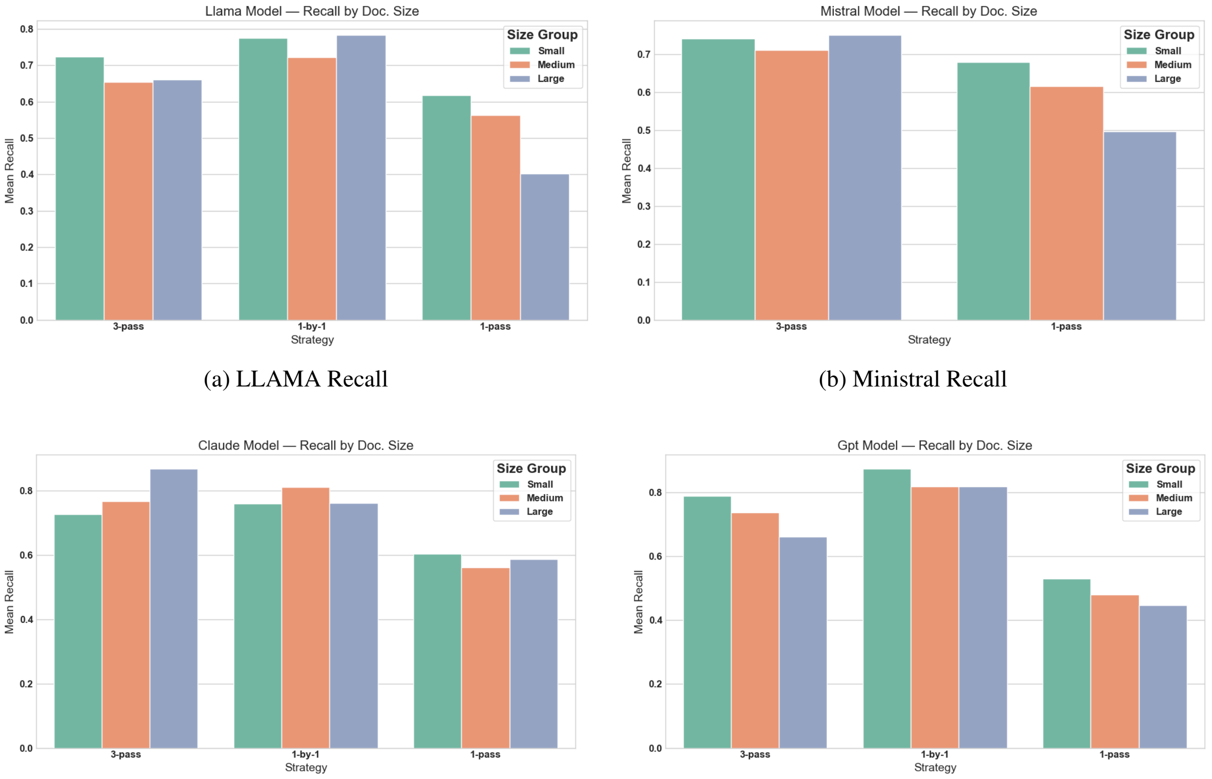
The two most commonly used attention functions are additive attention [2],

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.

B.5 Dataset Statistics

Table 4: Detailed Summary Statistics for the Dataset

Statistic	Value	Statistic	Value
Overall Dataset Statistics			
Total Annotations	346	Avg Claims per Paper	3.33
Unique Papers	100	Median Claims per Paper	3
Unique Claims	331	Min / Max Claims per Paper	1 / 8
Unique Evidence Passages	335	Avg Evidence per Paper	3.67
Duplicate Claims (Total)	15	Median Evidence per Paper	3
		Min / Max Evidence per Paper	1 / 9
Content Characteristics (Length in Words)			
Avg Claim Length	22	Avg Evidence Length	28
Median Claim Length	20	Median Evidence Length	25
Min / Max Claim Length	8 / 43	Min / Max Evidence Length	10 / 40



(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 6 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 Qualitative Analysis Metrics Selection

We selected our claim and evidence categories based on synthesizing and extending established types from prominent scientific validation literature.[2] This categorization draws on prior works, notably Clini-Fact (Zhang et al., 2025), CliVER (Liu et al., 2024a), SCITAB (Lu et al., 2023), and SciClaimHunt (Kumar et al., 2025), ensuring comprehensive coverage and alignment with established standards in scientific claim and evidence categorization. The chosen categories reflect prevalent argumentative structures and evidential forms across multiple domains, enhancing the applicability and robustness of CLAIM-BENCH. Methodological claims highlight innovation and technique advancements, while empirical claims cover observational and experimental findings central to scientific research. Comparative claims are integral to evaluating methodological or result-oriented superiority, whereas theoretical and causal claims capture conceptual advancements and explanatory relationships, respectively.

For evidence, we included experimental and observational evidence to reflect controlled and real-world conditions prevalent in scientific studies. Comparative evidence provides direct performance or outcome comparisons, essential for validation. Statistical evidence captures rigorous quantitative analysis, crucial for establishing scientific credibility, and expert evidence incorporates authoritative insights, emphasizing domain expertise.

Claim Categories:

Methodological claims highlight innovation in techniques or frameworks.
Example: "We propose a novel attention mechanism, `sparse-attention`*, which reduces computational complexity."*

Empirical claims cover observational and experimental findings central to scientific research.
Example: "Our study of 1,000 patients revealed that Drug X lowers blood pressure by an average of 10 mmHg."

Comparative claims are integral to evaluating methodological or result-oriented superiority.
Example: "The BERT-large model achieves a 5% higher accuracy on the SQuAD 2.0 dataset compared to RoBERTa-large."

Theoretical and Causal claims capture conceptual advancements and explanatory relationships, respectively.
Example: "Increased screen time before bed directly causes a measurable delay in sleep onset in adolescents."

Evidence Categories:

Experimental evidence is derived from controlled studies where researchers actively manipulate variables to test a hypothesis.
Example: "The treatment group showed a 95% reduction in infection rates compared to the placebo group under controlled lab conditions."

Observational evidence comes from studies where subjects are observed in their natural setting without researcher intervention.
te *Example: "A cohort study of 5,000 individuals found a positive correlation between high-fiber diets and reduced risk of heart disease."*

[2] The categorization of the outputs themselves was automated using the `claude-3-5-sonnet-20241022` model to ensure consistency.

Comparative evidence provides direct performance or outcome comparisons, essential for validation.

Example: "Table 3 shows our algorithm processed the dataset in 5.2 seconds, while the baseline took 11.8 seconds."

Statistical evidence captures rigorous quantitative analysis crucial for establishing scientific credibility.

Example: "A p-value of < 0.001 indicates that the observed difference in crop yield is statistically significant."

Expert evidence incorporates authoritative insights or the authors' synthesis of findings.

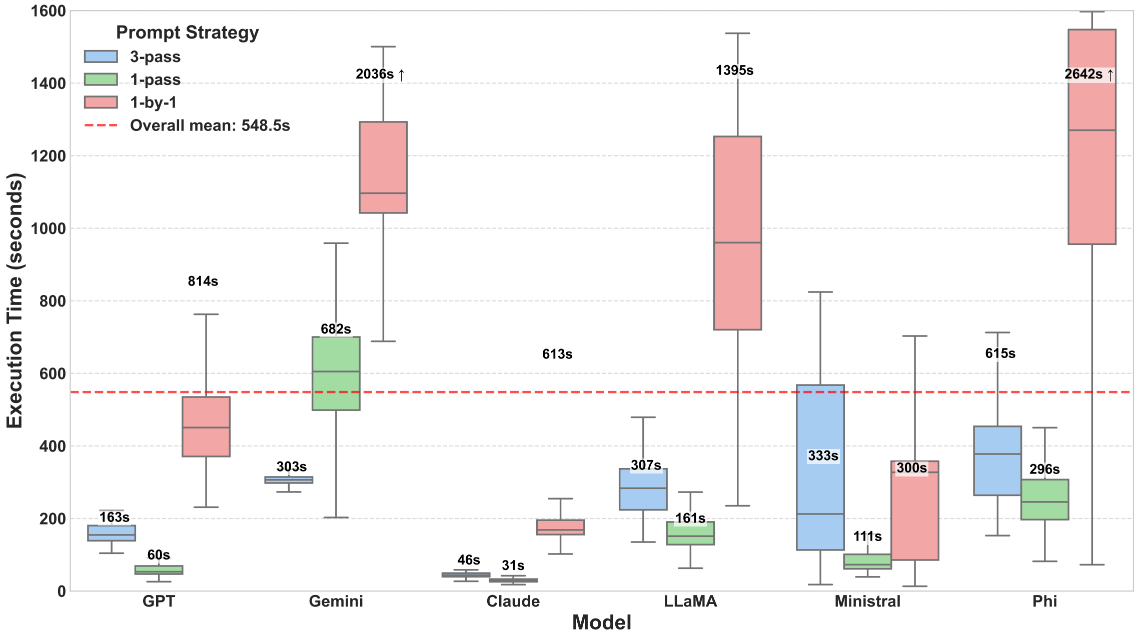
Example: "Based on these findings, we conclude that the geological formations are consistent with those found in other volcanic regions."

C.2 Evidence Qualitative Analysis

Table 5: Evidence Categorization: Percentage (and Rank) across Models and Strategies

Model and Strategy	**Experimental**	**Observational**	**Comparative**	**Statistical**	**Expert**
Human Annotations	23.7 (2)	16.3 (4)	20.8 (3)	25.3 (1)	13.9 (5)
Claude Models					
Claude Single Pass	20.3 (3)	10.9 (4)	34.2 (1)	32.1 (2)	2.5 (5)
Claude 3-Pass	26.8 (2)	13.0 (4)	35.4 (1)	23.6 (3)	1.2 (5)
Claude One-by-One Pass	33.9 (1)	11.4 (4)	29.4 (2)	22.3 (3)	3.1 (5)
GPT Models					
GPT 3-Pass	27.5 (2)	21.5 (3)	28.6 (1)	18.1 (4)	4.4 (5)
GPT All at Once	23.7 (3)	12.2 (4)	33.2 (1)	27.6 (2)	3.3 (5)
GPT One-by-One Pass	39.1 (1)	11.7 (4)	27.7 (2)	17.7 (3)	3.9 (5)
Gemini Models					
Gemini 3-Pass	23.7 (2)	17.3 (4)	31.0 (1)	19.4 (3)	8.6 (5)
Gemini One-by-One Pass	28.3 (1)	27.3 (2)	21.9 (3)	16.0 (4)	6.5 (5)
Gemini Single Pass	27.1 (2)	14.7 (4)	29.1 (1)	22.2 (3)	7.0 (5)
Llama Models					
Llama 3-Pass	26.3 (2)	16.8 (4)	29.9 (1)	22.3 (3)	4.8 (5)
Llama One-by-One Pass	31.3 (1)	16.0 (4)	25.7 (2)	20.7 (3)	6.2 (5)
Llama Single Pass	27.7 (2)	12.1 (4)	28.5 (1)	27.4 (3)	4.2 (5)
Ministral Models					
Ministral 3-Pass	22.9 (2)	31.1 (1)	19.7 (3)	14.2 (4)	12.2 (5)
Ministral One-by-One Pass	13.8 (3)	13.0 (4)	32.5 (2)	34.4 (1)	6.2 (5)
Ministral Single Pass	21.5 (3)	22.9 (2)	21.3 (4)	23.1 (1)	11.3 (5)
Phi Models					
Phi 3-Pass	32.1 (1)	21.7 (3)	23.8 (2)	14.1 (4)	8.2 (5)
Phi One-by-One Pass	27.6 (2)	15.8 (4)	30.0 (1)	20.7 (3)	5.9 (5)
Phi Single Pass	27.2 (2)	17.8 (4)	30.8 (1)	21.2 (3)	3.0 (5)

C.3 Execution Time Analysis

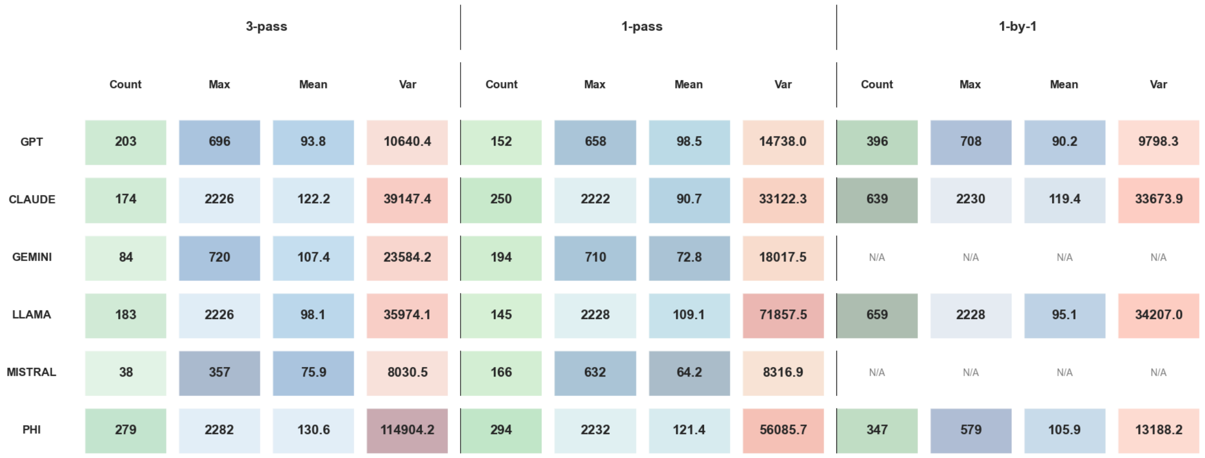


Note: Plot capped at 1600 seconds. 117 extreme outliers were filtered.

Figure 7: 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).

Execution times differ across models and strategies. GPT is highly efficient in the Single-Pass (under 200s) and 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, staying 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.

C.4 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 8: 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 Ministral keep distances short and variance low. "N/A" indicates the model-strategy combination was not executed.