Peerify: Benchmarking Peer-Review Claim Verification

Anonymous ACL submission

Abstract

Peer review plays a central role in scholarly publishing, yet verifying whether reviewer claims are supported by manuscript evidence remains a largely manual and time-consuming process. We present PEERIFY, a pipeline for manuscript-grounded verification of peer-review claims. Given a manuscript and a review comment, PEERIFY pipeline decomposes reviews into atomic claims, retrieves relevant manuscript evidence, and determines whether each claim is supported by the paper. To support the development and evaluation of the pipeline, we construct a benchmark of 800 claims derived from authentic peer-review interactions collected from NeurIPS 2024 and ICLR 2024. We evaluate state-of-the-art language models and retrieval strategies within PEERIFY pipeline. Our results demonstrate the importance of retrieval-centered verification and claim decomposition, while highlighting the challenges posed by ambiguous and interpretive reviewer claims.

1 Introduction

Peer review plays a central role in how scientific knowledge is evaluated and incorporated into the scholarly record and help establish confidence in scientific findings (Lee et al., 2013; Bornmann, 2011; Smith, 2006). As submission volumes grow, editors, area chairs, and program committees must assess more reviews under tighter timelines. Maintaining review quality and consistency at scale has therefore become an important operational challenge for scholarly publishing (Mulligan et al., 2013).

A substantial portion of the review process relies on claims made by reviewers which often influence publication outcomes. However, verifying whether such statements are supported by the contents of the manuscript remains largely a manual process. In practice, editorial stakeholders rarely have time to systematically examine every review statement against the corresponding paper. Consequently, unsupported claims and misunderstandings of a manuscript may persist throughout the review process without systematic verification.

The growing adoption of Large Language Models (LLMs) within scholarly workflows further amplifies the importance of this problem (Bender et al., 2021). LLMs are increasingly used to assist with manuscript preparation, review drafting, and editorial support tasks (Liang et al., 2024; Yuan et al., 2022). While these technologies offer substantial opportunities for improving efficiency, they also increase the need for mechanisms that can assess whether statements contained in reviews are faithfully grounded in the manuscript (Maynez et al., 2020; Huang et al., 2025). Reliable verification of review claims is therefore becoming an increasingly important component of research integrity within modern publishing ecosystems.

Although claim verification has been extensively studied in NLP (Dagan et al., 2006; Thorne et al., 2018; Wadden et al., 2020), peer-review claim verification differs in both scope and motivation. Existing work on LLM-assisted peer review focuses on review generation or quality assessment (Yuan et al., 2022; Liang et al., 2024; Hua et al., 2019; Rogers and Augenstein, 2020; Ebrahimi et al., 2025). However, to the best of our knowledge, verifying whether individual reviewer claims are grounded in the manuscript remains underexplored. Scientific claims frequently depend on methodological assumptions, dataset characteristics, and experimental conditions that must be interpreted together rather than in isolation. Furthermore, review state-

ments often combine objectively verifiable assertions with subjective judgments, requiring systems to distinguish between components that can be grounded in evidence and those that reflect reviewer opinion (Metropolitansky and Larson, 2025; Pavlick and Kwiatkowski, 2019; Plank, 2022). Operationalizing review groundedness thus requires robust document understanding, evidence retrieval, multi-hop reasoning, and careful treatment of ambiguity (Pavlick and Kwiatkowski, 2019; Plank, 2022). We provide a more comprehensive discussion of related work in Appendix D.

We identify manuscript-grounded peer-review claim verification as an important but underexplored challenge in scholarly publishing. To address this challenge, we propose PEERIFY, an end-to-end pipeline for peer-review claim verification. PEERIFY pipeline decomposes reviews into atomic, self-contained claims, retrieves relevant manuscript evidence for each claim, and performs groundedness assessment. PEERIFY pipeline was designed with practical deployment requirements in mind, including scalability to large review volumes, transparent evidence attribution, robustness across diverse claim types, and compatibility with existing editorial processes.

To systematically evaluate this task and pipeline, we construct a benchmark of 800 atomic review claims derived from publicly available NeurIPS 2024 and ICLR 2024 submissions, together with their reviews and author–reviewer discussion threads. The benchmark captures diverse verification scenarios, including explicitly supported claims, claims requiring evidence aggregation, ambiguous claims, and claims that cannot be resolved from the manuscript alone.

Using this benchmark, we evaluate PEERIFY pipeline end to end and analyze each of its core components, including claim extraction, evidence retrieval, retrieval configurations, and state-of-the-art language-model verifiers. Our results show that retrieval-centered verification and claim decomposition are important for manuscript-grounded review verification, but also reveal that ambiguous, interpretive, and partially supported reviewer claims remain challenging for current methods. We release PEERIFY pipeline, the benchmark, labels, prompts, and code to support reproducible research at `https://anonymous.4open.science/r/Peerify-CD20/`.

The contributions of this work are summarized as follows:

1. **Task.** We formalize manuscript-grounded peer-review claim verification as an intra-document verification task, where claims from peer reviews are assessed strictly against the reviewed manuscript.
2. **System.** We propose PEERIFY, an end-to-end verification pipeline that decomposes reviews into atomic claims, retrieves relevant manuscript evidence, and predicts explicit evidence attribution.
3. **Benchmark and evaluation.** We construct an 800-claim benchmark from NeurIPS 2024 and ICLR 2024 peer-review interactions, with six evaluation variants and quality-controlled slices to evaluate PEERIFY pipeline.

2 Peerify Pipeline

PEERIFY pipeline is a production-ready pipeline for review-groundedness verification, providing claim-level verification signals with explicit manuscript evidence for editorial quality assurance workflows.

2.1 Problem Definition

Let P denote a scientific manuscript and R denote an associated peer review. The objective of review-groundedness verification is to determine whether verifiable claims contained within R are supported by evidence present in P. Formally, a review R can be represented as a collection of atomic claims $C = \{c_1, \ldots, c_m\}$. For each claim c_i, the system identifies a set of supporting evidence candidates $E_i = \{e_1, \ldots, e_k\}$ from the manuscript and predicts a groundedness label $y_i = f(c_i, E_i)$.

This formulation differs from traditional fact verification and textual entailment settings (Dagan et al., 2006; Thorne et al., 2018; Wadden et al., 2020) in several important respects. First, the evidence space is restricted to a single scientific manuscript rather than an external corpus. Second, evidence supporting a review claim may be distributed across multiple sections, tables, appendices, and more. Third, review claims frequently in-

volve methodological choices, experimental design decisions, and interpretations of empirical findings whose validity depends on contextual qualifiers and evidence aggregation.

2.2 Operationalization in Peerify

PEERIFY operationalizes review-groundedness verification as a three-stage pipeline:

Step 1: Claim extraction. Review comments often contain several assertions in a single sentence, mixing factual observations with subjective judgments or recommendations. Direct verification is difficult because different portions may require distinct evidence. PEERIFY therefore adopts a claim decomposition strategy, breaking each review comment into atomic, self-contained claims that can be verified independently. This step preserves the reviewer's intended meaning while isolating factual assertions to be checked against the manuscript.

Step 2: Evidence retrieval. The manuscript is segmented into retrieval units such as paragraphs, figure captions, tables, and appendix passages. Given a claim, PEERIFY retrieves the top-k relevant passages based on semantic similarity between the claim and indexed manuscript segments. Since supporting evidence may be distributed across multiple locations, PEERIFY retrieves a set of evidence candidates rather than a single passage, allowing downstream verification to reason over complementary evidence throughout the document.

Step 3: Groundedness assessment. Finally, PEERIFY verifies each claim against the retrieved evidence, building on evidence-based verification paradigms from textual entailment, fact verification, and scientific claim verification (Dagan et al., 2006; Thorne et al., 2018; Wadden et al., 2020). The verifier jointly reasons over the claim and manuscript evidence to determine the extent of support, assigning one of four labels: *Supported*, *Partially Supported*, *Not Supported*, or *Not Determined*. Each prediction is paired with the evidence used for the decision, making the output transparent and traceable for human inspection.

3 Benchmarking Peerify

The PEERIFY pipeline requires an evaluation framework that reflects the challenges of real peer-review workflows. Existing claim-verification resources such as FEVER (Thorne et al., 2018) and SciFact (Wadden et al., 2020) evaluate claims against external corpora, but do not capture manuscript-bounded verification, author–reviewer interactions, or the mix of factual and subjective assertions common in peer review (see Appendix D). We therefore develop the PEERIFY benchmark, which operationalizes review groundedness as an intra-document verification task and supports the development, validation, and evaluation of review-groundedness verification systems. Our goal is to construct a gold-standard collection of review claims paired with groundedness labels under realistic scholarly review conditions. To ensure ecological validity, all claims are derived from authentic peer-review interactions and evaluated against the manuscripts under review. The resulting benchmark contains 800 review-derived claims from publicly available NeurIPS 2024 and ICLR 2024 submissions.

3.1 Source Collection

We collected manuscripts, reviews, author responses, and discussion threads from publicly available submissions in NeurIPS 2024 and ICLR 2024. These venues were selected because they provide complete review artifacts, including reviewer comments and author rebuttals, enabling multiple sources of evidence regarding the validity of review claims. Each instance in the collection consists of a manuscript together with its associated review and discussion history. As reported in Table 1, the collected papers are well balanced across review outcomes, covering both accepted and rejected submissions.

3.2 Claim Construction

Review comments contain summaries, questions, recommendations, methodological critiques, and subjective opinions. Since review-groundedness verification operates at the level of individual factual assertions, raw review text was transformed into atomic claims suitable for independent verification.

To construct these claims, we evaluated three extraction approaches: `Qwen3-4B` (Yang et al., 2025), prompted to produce atomic, decontextualized assertions; Fenice (Scirè et al.,

Table 1: PEERIFY benchmark statistics overview.

Dataset	#Papers	#Claims	Supported	Partially Supported	Not Supported	Not Determined	ICLR	NeurIPS	Reject	Accept
NeurIPS 2024	4236	-	-	-	-	-	-	100%	86.0%	14.0%
ICLR 2024	5778	-	-	-	-	-	100%	-	31.3%	68.7%
PAPERSOURCED (3.4)	100	500	500	0	0	0	52.0%	48.0%	27.0%	73.0%
REBUTTALSOURCED (3.4)	84	800	217	310	189	84	61.0%	39.0%	34.9%	65.1%
LLMJUDGED (3.4)	84	800	207	93	141	359	61.0%	39.0%	34.9%	65.1%
HIGHAGREEMENT (3.4)	71	175	60	39	43	33	63.4%	36.6%	37.1%	62.9%
VERIFIABLE-ONLY (3.4)	67	131	56	29	25	21	64.9%	35.1%	36.6%	63.4%
HUMAN-VERIFIED (3.4)	18	150	44	51	34	21	54.7%	45.3%	21.3%	73.3%

2024), which uses dependency parsing and semantic role labeling to identify proposition structures; and Gemma, a general-purpose instruction-tuned model prompted to list atomic claims from each passage.

We run a controlled comparison of the three extractors and adopt `Qwen3-4B` for all benchmark construction, as it gives the best balance of atomicity, decontextualization, and semantic fidelity. The full evaluation protocol and results are deferred to Appendix A.1. Extracted claims were subsequently post-processed through exact and near-duplicate removal, string normalization, and filtering of rhetorical questions, speculative statements, and non-verifiable meta-commentary.

3.3 Groundedness Annotation

Each extracted claim is assigned one of the four labels of *Supported*, *Not Supported*, *Partially Supported* or *Not Determined*. Each claim is additionally tagged as *factual* or *subjective* (automatically, with `o4-mini`; criteria in Appendix A.3), and the VERIFIABLE-ONLY slice retains only factual claims. This ensures groundedness is evaluated on assertions checkable against the manuscript rather than on inherently subjective reviewer opinions.

3.4 Evaluation Variants

PEERIFY includes four core benchmark variants and two quality-controlled slices, capturing complementary notions of groundedness and label confidence.

PaperSourced. PAPERSOURCED provides a controlled diagnostic setting in which claims are extracted directly from manuscript content. These claims are *verifiable by construction* and are expected to be labeled `Supported`. We use this subset as a sanity check to isolate errors in retrieval and verification from ambiguity in review-derived content.

RebuttalSourced. This variant derives supervision from realistic author–reviewer discussion threads, assigning each claim a label by interpreting the author response signal (agreement, correction, clarification, or explicit dispute). Two language models (`GPT-5-mini` and `Claude-Sonnet-4-6`) label each claim independently from the discussion context, with a human arbitrating disagreements to produce the canonical label (Appendix A.6). This interaction-grounded supervision, captures how authors themselves judge the correctness of reviewer claims.

LLMJudged. This subset provides scalable supervision by judging each claim directly against the manuscript. The same two labelers (`GPT-5-mini` and `Claude-Sonnet-4-6`) independently assign one of four labels under strict document-bounded evidence constraints, and a human annotator arbitrates the cases where they disagree to produce the canonical label (More details in Appendix A.6). We evaluate label quality against the HUMAN-VERIFIED labels in Appendix A.5.

Human-Verified To provide high-fidelity reference labels and audit automated supervision, we manually annotated a subset of 150 REBUTTALSOURCED instances. Two annotators with formal training in computer science independently labeled each instance following detailed guidelines with concrete examples. Inter-annotator reliability was near-perfect (Cohen's $\kappa \approx 0.96$) for the four-way nominal task, with disagreements resolved through consensus. We further validated automated supervision against these consensus labels: on the instances with the labels from GPT-5-mini, its labels matched the human consensus on over 94% of cases, with divergences confined to ambiguous cases.

Quality-Controlled Slices we additionally release slices that improve label reliability. First, the HIGHAGREEMENT slice retains only instances where independent supervision sources, REBUTTALSOURCED and LLMJUDGED, agree, yielding a high-confidence subset less sensitive to labeling noise. Second, to isolate assertions checkable directly

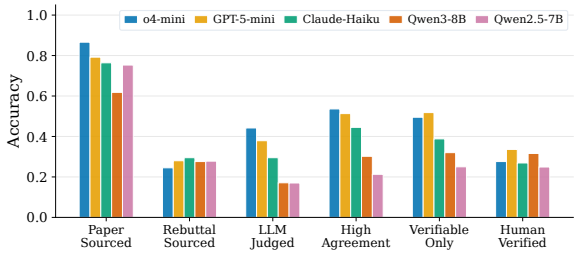

Figure 1: Full-context verification accuracy across all six PEERIFY benchmark variants.

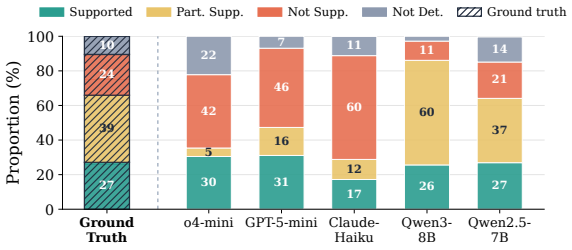

Figure 2: Predicted vs. ground-truth (GT) labels on REBUTTALSOURCED. Models exhibit different calibration biases: some over-predict *Not Supported*, while others over-predict *Partially Supported* or better match the GT distribution.

against the manuscript, the VERIFIABLE-ONLY slice uses `o4-mini` to filter out subjective or normative claims such as novelty, significance, or impact. Specifically, the model predicts whether a claim is (i) grounded in factual content observable in the manuscript, such as missing experiments or absent figures, or (ii) an opinion, recommendation, or forward-looking judgment not directly verifiable from the paper alone.

Table 1 shows the support-level distributions across different variants. PAPERSOURCED is entirely `Supported` (500/500) by construction and serves as a controlled diagnostic. REBUTTALSOURCED is far more heterogeneous (217 Supported, 310 Partially Supported, 189 Not Supported, 84 Not Determined), reflecting corrections, and unresolved disagreement in author–reviewer exchanges. LLMJUDGED shifts heavily toward `Not Determined` (207 Supported, 93 Partially Supported, 141 Not Supported, 359 Not Determined), indicating that many review statements cannot be decisively validated from the manuscript alone.

4 Experimental setup

We evaluate PEERIFY in two deployment configurations. *Full-Context* feeds the whole manuscript and claim to the verifier, serving as an upper bound when the paper fits within the context window; over-length papers are excluded per model. *Retrieval-Augmented* (RAG) retrieves evidence per claim and conditions the verifier only on the top-k passages, with or without a cross-encoder reranker, enabling deployment on long manuscripts. All prompts are listed in Appendix B.

We instantiate PEERIFY pipeline with five state-of-the-art models spanning proprietary and open-weight families: `o4-mini` (200K-context reasoning model), `GPT-5-mini` (compact, long-context), `Claude-Haiku-4.5` (low-latency, 200K context), and the open-weight `Qwen3-8B` and `Qwen2.5-7B`, the last two evaluated mainly under RAG due to context-window limits. All models receive identical label definitions and prompts, and outputs are mapped to the four support level labels.

We evaluate retrieval with recall@k, nDCG@k, and MRR on PAPERSOURCED (BM25 or a dense retriever, optionally reranked by a cross-encoder; retriever and reranker checkpoints, top-k, and chunking are detailed in Appendix C.5), and verification with Accuracy and Macro-F1, taking Macro-F1 as the primary metric since it is robust to label imbalance.

5 Results and Findings

In this section, we evaluate PEERIFY pipeline from several complementary perspectives.

5.1 Full-context Verification

Figure 1 reports verification accuracy across the six evaluation variants, revealing a clear difficulty gradient. Performance is highest on PAPERSOURCED, where claims come directly from manuscript content and evidence is explicit; `o4-mini` reaches 0.866, and four of five verifiers exceed 0.75. Accuracy drops on review-derived settings such as LLMJUDGED and HIGHAGREEMENT, where claims require reasoning over dispersed evidence and reviewer intent. Even the strongest verifier reaches only 0.442 on LLMJUDGED, showing that review-groundedness verification goes beyond simple evidence localization. The hardest setting is REBUTTALSOURCED, where claims come from real author–reviewer interactions and often mix factual observations.

Table 2: RAG accuracy across five models and four retrievers on all six PEERIFY benchmarks.

Model	Retriever	PAPER SOURCED	REBUTTAL SOURCED	LLM JUDGED	HIGH AGREEMENT	VERIFIABLE -ONLY	HUMAN VERIFIED
o4-mini	BM25	0.818	0.234	0.491	**0.514**	0.466	0.260
	BM25+Reranker	0.870	0.234	0.496	**0.514**	**0.519**	0.247
	Dense	0.852	0.244	**0.514**	0.509	0.466	0.220
	Dense+Reranker	**0.904**	0.240	0.499	0.497	0.466	0.247
GPT-5-mini	BM25	0.796	0.289	0.412	0.509	0.496	0.287
	BM25+Reranker	0.856	0.294	0.417	0.491	0.496	0.293
	Dense	0.820	0.294	0.414	0.503	0.511	**0.313**
	Dense+Reranker	0.852	0.286	0.405	0.491	0.466	0.273
Claude-Haiku	BM25	0.786	0.280	0.381	0.440	0.389	0.253
	BM25+Reranker	0.822	0.270	0.372	0.451	0.405	0.213
	Dense	0.808	0.256	0.386	0.429	0.374	0.207
	Dense+Reranker	0.844	0.297	0.374	0.480	0.427	0.267
Qwen3-8B	BM25	0.744	0.274	0.393	0.440	0.382	0.227
	BM25+Reranker	0.816	0.281	0.364	0.440	0.389	0.287
	Dense	0.800	0.287	0.371	0.400	0.321	0.253
	Dense+Reranker	0.814	0.265	0.385	0.423	0.374	0.300
Qwen2.5-7B	BM25	0.844	0.319	0.211	0.263	0.267	0.320
	BM25+Reranker	0.872	0.323	0.211	0.297	0.305	**0.367**
	Dense	0.832	**0.324**	0.209	0.263	0.260	0.320
	Dense+Reranker	0.860	0.295	0.228	0.291	0.305	0.313

5.2 Retrieval-Centered Verification.

A key design decision in PEERIFY is separating evidence retrieval from groundedness assessment. Table 2 evaluates four retrieval configurations, including BM25, dense retrieval, and their reranked variants. Retrieval quality substantially affects verification accuracy, but no retriever dominates across all settings. On PAPERSOURCED, where evidence is explicit, reranking consistently improves performance. Dense+Reranker increases o4-mini from 0.852 to 0.904, and BM25+Reranker increases GPT-5-mini from 0.796 to 0.856. On review-derived benchmarks, however, gains are less consistent. For example, Dense retrieval performs best for o4-mini on LLMJUDGED (0.514), while BM25-based retrieval works best for GPT-5-mini (0.417) and Qwen3-8B (0.393). Similar variability is observed on HIGHAGREEMENT and HUMAN-VERIFIED, indicating that retrieval effectiveness depends not only on evidence quality but also on how different verification engines consume retrieved context.

Overall, better retrieval helps most when the task requires explicit evidence localization, but its impact is smaller on REBUTTALSOURCED and HUMAN-VERIFIED, where claims require interpreting reviewer intent and resolving partial support. These results suggest that evidence retrieval is necessary but insufficient. Improving retrieval alone is unlikely to close the gap on the most challenging review-derived settings.

5.3 Error Analysis

The largest error source is REBUTTAL-SOURCED, where reviewer claims mix factual observations with interpretive judgments. Additionally, systematic calibration differences across verification engines (Figure 2) suggest that calibration is as critical as reasoning capability for review-groundedness verification.

A second failure mode emerges on LLM-JUDGED and HIGHAGREEMENT, where verification requires multi-step reasoning over dispersed evidence, limiting even the strongest verifier to 0.442 and 0.536 accuracy, respectively. Together, these findings indicate that the primary bottlenecks for the PEERIFY pipeline are reasoning over distributed evidence and correctly handling interpretive or partially supported claims.

6 Conclusion

We presented PEERIFY pipeline, a production-ready pipeline that decomposes reviews into atomic claims, retrieves manuscript evidence, and assesses whether each claim is grounded, together with a benchmark of 800 claims from authentic NeurIPS and ICLR peer review. Across five LLMs and four retrieval configurations, retrieval-centered verification is effective on grounded claims, but ambiguous, partially supported, and interpretive reviewer claims remain an open challenge. PEERIFY pipeline offers a practical foundation for evidence-based review verification in scholarly publishing.

Limitations

PEERIFY performs **intra-document** verification, assessing review claims only against the corresponding manuscript; claims requiring external literature or background knowledge are out of scope. The benchmark is built from public OpenReview data for NeurIPS 2024 and ICLR 2024, so its distribution may not transfer to other fields or venues with different review norms. Some variants rely on weak supervision (author–reviewer discussions and LLM judgments); we mitigate this with the HIGHAGREEMENT and VERIFIABLE-ONLY slices, but residual ambiguity in review language remains. Finally, models show strong calibration bias (e.g., `Qwen3-8B` predicts `Partially Supported` for 60.5% of REBUTALSOURCED claims vs. a 38.8% ground-truth rate), so accuracy alone can mislead; we recommend reporting macro-F1 alongside accuracy.

Ethics Statement

The benchmark uses only publicly accessible OpenReview submissions, reviews, and discussions from NeurIPS 2024 and ICLR 2024; no private or proprietary data are used, and no reviewer or author is identified beyond what is already public. The 150 HUMAN-VERIFIED instances were labeled by two graduate-level researchers following explicit guidelines after a calibration round; no crowdsourced or low-paid labor was involved. PEERIFY is intended as an *assistive* tool that supplements rather than replaces human judgment, and should not be used to rank or target specific reviewers, authors, or submissions. Because LLM-based judges and extractors can carry systematic biases (Section 5), automated labels should not be treated as ground truth without further validation. We release all data, labels, prompts, and code under open licenses with documentation of provenance and known limitations.

References

Emily Bender and 1 others. 2021. On the dangers of stochastic parrots. *Proceedings of FAccT*.

Lutz Bornmann. 2011. Scientific peer review. *Annual Review of Information Science and Technology*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment*.

Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2025. Rottenreviews: Benchmarking review quality with human and llm-based judgments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 5642–5649.

Mohammad Javad Hosseini, Yang Gao, Tim Baumgärtner, Alex Fabrikant, and Reinald Kim Amplayo. 2024. Scalable and domain-general abstractive proposition segmentation. *arXiv preprint arXiv:2406.19803*.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2131–2137.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *Proceedings of ICLR*.

Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. Resources for brewing beir: Reproducible reference models and statistical analyses. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1431–1440, New York, NY, USA. Association for Computing Machinery.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of EMNLP*.

Carole Lee, Cassidy Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of NeurIPS*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, and 1 others. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*.

Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and Verification (FEVER)*.

Joshua Maynez and 1 others. 2020. On faithfulness and factuality in abstractive summarization. In *ACL*.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims (claimify). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Adrian Mulligan, Louisa Hall, and Ellen S. Raphael. 2013. Peer review in a changing world: An international study measuring the attitudes of researchers. *J. Assoc. Inf. Sci. Technol.*, 64:132–161.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161.

Richard Smith. 2006. Peer review: A flawed process at the heart of science. *Journal of the Royal Society of Medicine*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT*.

Andrew Tomkins, Min Zhang, and William D Heavlin. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2025. Claim extraction for fact-checking: Data, models, and automated metrics.

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 EMNLP*.

David Wadden and Kyle Lo. 2021. Overview and insights from the SciVer shared task on scientific claim verification. In *Proceedings of the Second Workshop on Scholarly Document Processing (SDP)*, pages 124–129.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Hang Xu, Ling Yue, Chaoqian Ouyang, Yuchen Liu, Libin Zheng, Shaowu Pan, Shimin Di, and Min-Ling Zhang. 2026. Factreview: Evidence-grounded reviews with literature positioning and execution-based claim verification. *arXiv preprint arXiv:2604.04074*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Appendix

A Benchmark Construction and Annotation

A.1 Claim Extraction Evaluation

A critical step in the PEERIFY pipeline is accurately decomposing dense, multi-faceted reviewer comments into isolated and verifiable statements. Because reviewer feedback often blends factual assertions with subjective interpretation, the claim extraction module must distill these paragraphs into atomic claims while strictly preserving the original intent of the reviewer. To determine the most robust approach for this task, table 3 compares the performance of three distinct claim extractors (Fenice (Scirè et al., 2024), Gemma (Hosseini et al., 2024), and `Qwen3-4B` (Yang et al., 2025)) across two evaluation axes: semantic alignment with a reference extraction (LLM-as-a-judge) and intrinsic structural quality (reference-free). The two axes are complementary: reference-based metrics capture coverage, while reference-free metrics capture whether individual claims are verification-ready. The evaluation focuses on how effectively each model parses authentic review text from our NeurIPS and ICLR 2024 dataset into self-contained propositions suitable for downstream retrieval and verification.

A.1.1 Reference-Based Evaluation.

To establish a strong comparison signal, we use `o4-mini` as a high-quality reasoning model to generate a reference set of claims that are manually verified to be atomic and decontextualized. We treat this reference extraction as a stronger summarization-style signal and evaluate how well each candidate extraction method aligns with it. Agreement is measured using an LLM-as-a-judge protocol. We employ `GPT-5-mini` as a semantic judge: for each extracted claim, we identify the most similar reference claim and ask the judge whether the two express the same atomic proposition. Based on these semantic equivalence decisions, we compute precision, recall, and F1 scores.

For the reference-based axis, `o4-mini` is used *only* to generate the high-fidelity reference set against which the three candidate extractors are scored; it is not used to produce any of the 800 benchmark claims. The benchmark claims are sourced exclusively from `Qwen3-4B`, which achieves the best overall balance and is therefore used for all benchmark construction, ensuring extraction quality and reproducibility.

Table 3: Claim-extraction quality: LLM-as-a-judge (`GPT-5-mini`, reference-based) and reference-free intrinsic metrics.

		Reference-Based			Reference-Free		
Extractor	**count**	**Prec.**	**Rec.**	**F1**	**Atom.**	**Flue.**	**Deco.**
Fenice	739	0.872	0.538	0.665	0.314	0.788	0.362
Gemma	2464	0.868	**0.762**	**0.811**	0.379	0.781	0.329
Qwen3-4b	923	**0.889**	0.587	0.707	**0.451**	**0.927**	**0.524**

A.1.2 Reference-Free Evaluation.

Regardless of the reference set, a useful claim must meet specific structural and linguistic criteria for downstream verification (Ullrich et al., 2025; Wright et al., 2022). We use `o4-mini` to audit each method's output against three core dimensions. **Atomicity** measures whether a claim expresses exactly one checkable assertion. **Faithfulness** assesses whether the claim remains grounded in the original reviewer's text without introducing hallucinations or scope-creep. **Decontextualization** checks whether the claim is self-contained and does not rely on unresolved references such as "the method" or "Figure 1" that require surrounding context to interpret.

Coverage versus quality tradeoff. The evaluation reveals distinct operational profiles among the extracted models. Fenice exhibits a high-precision, low-recall regime by extracting 739 claims with a precision of 0.872 and a recall of 0.538, effectively minimizing false positives at the expense of comprehensive coverage. Conversely, Gemma prioritizes recall by achieving a rate of 76.2% and an F1 score of 0.811, but it generates a substantially larger volume of 2,464 candidate claims. `Qwen3-4B` optimizes this balance by producing 923 claims with a precision of 0.889, thereby maintaining strict semantic control. In the context of downstream verification, the implications of this tradeoff are asymmetric: while unextracted claims merely reduce total evaluation coverage, poorly formulated claims systematically confound reliable classification.

Intrinsic quality is the discriminating dimension. Beyond semantic alignment, intrinsic structural quality emerges as the critical factor determining extractor viability. `Qwen3-4B` demonstrates significantly superior performance across essential structural metrics. Specifically, it achieves an atomicity score of 0.451, which exceeds the scores of 0.379 for Gemma and 0.314 for Fenice. Furthermore, its fluency score of 0.927 surpasses both Gemma and Fenice as well. Finally, `Qwen3-4B` records a decontextualization score of 0.524, notably outperforming the respective scores of 0.329 and 0.362 achieved by Gemma and Fenice. These structural properties are foundational to downstream verification accuracy. Consequently, the high-fidelity benchmark slices, specifically VERIFIABLE-ONLY and HIGHAGREEMENT, are constructed exclusively using `Qwen3-4B` extractions. Robust claim extraction is a prerequisite for reliable evaluation, meaning these extraction quality differentials directly determine the diagnostic validity of the resulting benchmark partitions.

A.2 Label Definitions and Examples

Each review claim receives one of the four labels defined in Section 2.1, determined by whether the manuscript provides sufficient evidence under a strict document-bounded grounding criterion. Worked examples follow.

- `Supported`: The manuscript provides clear and sufficient evidence that fully substantiates the claim. *Example:* A reviewer states "The authors evaluate on CIFAR-100," and the paper explicitly reports CIFAR-100 experiments.
- `Not Supported`: The manuscript contradicts the claim or the asserted content is demonstrably absent. *Example:* A reviewer states "No ablation study is provided," but the paper contains a dedicated ablation section.
- `Partially Supported`: The manuscript supports only a qualified or incomplete version of the claim. *Example:* A reviewer states "The method outperforms all baselines," but the paper shows it outperforms most but not all.
- `Not Determined`: The manuscript does not contain sufficient information to resolve the claim in either direction. This label applies when (i) the claim references external knowledge or prior work not discussed in the paper, (ii) the claim is too vague or ambiguous to verify, or (iii) the relevant evidence is absent without any contradicting assertion.

A.3 Factual vs. Subjective Claims

Each claim is tagged as *factual* or *subjective* using `o4-mini`: *factual* if its validity is grounded in content observable in the manuscript (e.g., a missing experiment or an absent figure), and *subjective* if it is an opinion, recommendation, or forward-looking judgment about novelty, significance, or impact that cannot be resolved from the paper alone. The VERIFIABLE-ONLY slice retains only factual claims, so verification on that slice targets assertions checkable against the manuscript rather than reviewer opinion.

A.4 Benchmark Label Distributions

Table 1 in Section 2 reports the full PEERIFY benchmark statistics overview, including label distributions across all six benchmark variants and dataset provenance (venue and acceptance decision). The PAPERSOURCED variant is all-Supported by construction; the remaining variants show progressively more heterogeneous label distributions as claims become more interaction-grounded and ambiguous. The HUMAN-VERIFIED subset does not include venue or decision metadata because the 150 annotated instances span both ICLR and NeurIPS submissions and the breakdown is not reported separately.

A.5 Inter-Source Agreement

The HIGHAGREEMENT slice is constructed by retaining only instances where REBUTTALSOURCED and LLMJUDGED labels agree. This section describes the agreement analysis underlying this design choice.

Automated supervision sources. REBUTTALSOURCED labels are derived from author–reviewer interaction dynamics: the model infers whether the author's response confirms or refutes the reviewer's claim. LLMJUDGED labels are derived from direct

document-bounded verification: the model checks whether the manuscript supports the claim without consulting the discussion thread. These two sources are therefore *independent* in both mechanism and evidence source, making their agreement a meaningful signal of claim clarity and label reliability.

Across the 800 core review-derived claims, the canonical REBUTTALSOURCED and LLMJUDGED labels agree on 175 instances (21.9%), forming the HIGHAGREEMENT slice. The relatively low raw agreement rate reflects the genuine difficulty of peer-review claims and the different perspectives captured by the two supervision sources: interaction-grounded labels capture how authors interpret their own work, while document-grounded labels capture what the manuscript literally supports.

Automated supervision vs. human labels. To validate automated supervision quality, we compare both REBUTTALSOURCED and LLMJUDGED labels against the 150 HUMAN-VERIFIED consensus labels. Agreement between automated supervision and human annotators exceeds 94% on the HUMAN-VERIFIED subset, with divergences largely confined to borderline and ambiguous cases. Inter-annotator agreement between the two human annotators is near-perfect (Cohen's $\kappa \approx 0.96$), indicating that the high human–automated agreement is not an artifact of low human reliability. These results support the use of REBUTTALSOURCED and LLMJUDGED as reliable large-scale supervision sources, with HUMAN-VERIFIED serving as a gold-standard reference for validation and error analysis.

A.6 Annotator Disagreement and Arbitration

The canonical labels for REBUTTALSOURCED and LLMJUDGED are produced by a two-annotator protocol with human arbitration. Each of the 800 atomic reviewer claims is labeled twice and independently: `GPT-5-mini` (Annotator A) and `Claude-Sonnet-4-6` (Annotator B) receive the same claim and the same source material (the author response for REBUTTALSOURCED and the paper content for LLMJUDGED) under identical prompts. For every record where the two annotators disagree, a human annotator is shown the claim, the relevant source material, and both candidate labels (with model identity anonymized and presentation order randomized per record), and either selects the correct label or assigns a new one when both candidates are wrong. This arbitrated label is taken as canonical.

Table 4: Inter-annotator agreement (`GPT-5-mini` vs. `Claude-Sonnet-4-6`) before arbitration.

Benchmark	n	Annotators agree	Rate
REBUTTALSOURCED	800	480	60.0%
LLMJUDGED	800	459	57.4%

Table 5: Quantitative breakdown of arbitration verdicts on annotator disagreements. Annotator A and B denote the `GPT-5-mini` and `Claude-Sonnet-4-6`, respectively.

Benchmark	#Disagree	A correct	B correct	Neither
REBUTTALSOURCED	320	158 (49.4%)	136 (42.5%)	26 (8.1%)
LLMJUDGED	341	168 (49.3%)	137 (40.2%)	35 (10.3%)

Inter-annotator agreement before arbitration. Table 4 reports raw agreement between the two annotators. `Claude-Sonnet-4-6` is systematically more conservative than `GPT-5-mini`: on REBUTTALSOURCED it reclassifies many `Supported` calls as `Partially Supported` or `Not Supported`, and on LLMJUDGED it shifts confident labels into `Not Determined`. The two models agree almost entirely on `Not Determined` itself (52/67 on REBUTTALSOURCED, 193/251 on LLMJUDGED) but diverge on the confident labels.

Arbitration verdicts. Table 5 shows, among the disagreements, how often each annotator was upheld by the human annotator. `GPT-5-mini` is correct slightly more often (∼49%) than `Claude-Sonnet-4-6` (∼41%), but in roughly one in ten disagreements neither candidate was correct and the annotator supplied a third label.

Canonical distribution and derived slices. Table 6 gives the arbitrated label distribution. The LLMJUDGED distribution is heavily weighted toward `Not Determined`, reflecting that a meaningful share of reviewer claims have no verifiable answer in the paper text alone. After arbitration, the HIGHAGREEMENT slice (canonical REBUTTALSOURCED label equals canonical LLM-

Table 6: Canonical label distribution (post-arbitration).

Label	RebuttalSourced	LLMJudged
Supported	217	207
Partially Supported	310	93
Not Supported	189	141
Not Determined	84	359

JUDGED label) contains 175 instances, and the VERIFIABLE-ONLY slice contains 131.

B Prompts

B.1 Verification Prompt Template

We use the following prompt template for all LLM verifiers in full-context mode. The same label definitions are used across all dataset variants.

Full-Context Verification Prompt

You are a scientific claim verifier. Given a peer review claim and the full text of the reviewed manuscript, determine whether the manuscript supports the claim.

Claim: {claim}
Manuscript: {manuscript}

Classify the claim as exactly one of the following labels based solely on the content of the manuscript:

- **Supported:** The manuscript provides clear and sufficient evidence that fully substantiates the claim.
- **Not Supported:** The manuscript contradicts the claim, or the asserted content is demonstrably absent from the manuscript.
- **Partially Supported:** The manuscript supports only a qualified or incomplete version of the claim (e.g., missing conditions, limited scope, or omitted caveats).
- **Not Determined:** The manuscript does not contain sufficient information to resolve the claim in either direction. Use this label when (1) the claim references external knowledge, (2) the claim is too vague or ambiguous to verify, or (3) the relevant evidence is absent without any contradicting assertion.

For RAG-based verification, the {manuscript} field is replaced with the concatenation of top-k retrieved passages, ranked by the chosen retrieval configuration.

B.2 Claim Extraction Prompt

The prompted extractor (Qwen3-4B) decomposes each review into atomic claims with the following prompt.

Table 7: PAPERSOURCED Dense+R accuracy vs. the Oracle upper bound (gold paragraph fed directly) for all five models. Full four-retriever results in Table 2.

Model	Dense+Reranker	Oracle
o4-mini	**0.904**	**0.988**
GPT-5-mini	0.852	0.958
Claude-Haiku	0.844	0.860
Qwen3-8B	0.814	0.872
Qwen2.5-7B	0.860	0.816

Claim Extraction Prompt

System: You are an expert at extracting factual claims from academic reviews.

User: Extract factual claims from the review text below. A claim is a specific, verifiable statement about the paper.
Review Text: {review_text}

Each claim should be (1) a single atomic statement, (2) verifiable against the paper or author response, and (3) specific and factual (not an opinion).

B.3 Labeling Prompts

The two automated annotators (GPT-5-mini and Claude-Sonnet-4-6) use the prompts below for the REBUTTALSOURCED and LLMJUDGED variants. The raw Contradicted label corresponds to *Not Supported* in the four-way scheme.

RebuttalSourced Labeling Prompt

System: You are an expert at analyzing academic discourse.

User: Given a reviewer's claim and the author's response, determine how the authors address the claim.
Guidelines: *Supported*: authors clearly agree with or confirm the claim; *Partially Supported*: authors acknowledge some validity but not fully; *Contradicted*: authors explicitly disagree or provide counter-evidence; *Not Determined*: authors do not address the claim, or it is unclear.
Reviewer Claim: {claim} **Author's Response:** {author_response}

LLMJudged Labeling Prompt

System: You are an expert at analyzing academic papers. **User:** Given a reviewer's claim about a paper and the paper content, determine whether the claim is true according to the paper.

Guidelines: *Supported*: the claim is true according to the paper; *Partially Supported*: partially true; *Contradicted*: the paper contradicts the claim; *Not Determined*: the paper does not address the topic, or there is insufficient information. For example, if the claim states "the paper lacks comparison with baseline X" and the paper indeed omits it, the label is *Supported*.
Reviewer Claim: {claim} **Paper Content:** {paper_content}

C Additional Results

C.1 Full-Context Verification Results

Table 8 reports the precise full-context verification accuracy for all five models across the six benchmarks, complementing Figure 1.

C.2 Accuracy vs. Macro-F1

Table 9 reports accuracy and macro-F1 for both full-context and RAG paradigms. In the RAG setting, `Qwen2.5-7B` Dense achieves the highest REBUTTALSOURCED accuracy (0.324, macro-F1 0.277), followed by `GPT-5-mini` Dense (Acc 0.294, F1 0.279) and `o4-mini` Dense (Acc 0.244, F1 0.236). The large accuracy–macro-F1 gaps for some models reflect calibration bias rather than genuine reasoning, motivating macro-F1 as a complementary metric robust to label imbalance.

C.3 PaperSourced RAG Results

Table 7 reports the PAPERSOURCED Oracle upper bound alongside Dense+R for all five models; the full four-retriever PAPERSOURCED results are in the main RAG table (Table 2). Oracle feeds the gold-standard supporting paragraph directly to the verifier, providing an upper bound under perfect retrieval.

The Oracle upper bound reveals that PAPERSOURCED is nearly solvable when retrieval is perfect: `o4-mini` reaches 0.988, `GPT-5-mini` 0.958, `Qwen3-8B` 0.872, and `Claude-Haiku` 0.860. The gap between Oracle and Dense+R directly quantifies accuracy lost to retrieval imperfection: 0.084 for `o4-mini` (0.988→0.904) and 0.106 for `GPT-5-mini`. One anomaly: `Qwen2.5-7B` Dense+R (0.860) exceeds its own Oracle (0.816), because Dense+R supplies ten passages (including adjacent context) while Oracle supplies only the single gold paragraph. For a model with limited long-context capacity, richer surrounding context improves its ability to confirm supported claims even when the exact evidence is present in both conditions.

C.4 Extended RAG Analysis

This section collects secondary RAG findings deferred from Section 5.

Performance gap narrows on interpretive claims. On REBUTTALSOURCED, all retrieval configurations cluster between 0.234 and 0.244 for `o4-mini`, closely tracking the full-context result of 0.245. For LLMJUDGED, the picture inverts: `Qwen2.5-7B` accuracy collapses to 0.209–0.228 with macro-F1 close to accuracy (e.g., 0.200 for Dense), confirming that its REBUTTALSOURCED accuracy advantage does not transfer to semantically harder claims.

Cross-model patterns across benchmarks. On LLMJUDGED, among the models scored by re-scoring REBUTTALSOURCED predictions, `GPT-5-mini` leads (best: BM25-R 0.417), while `Qwen2.5-7B` collapses to 0.209–0.228, a 2× gap confirming a hard capability boundary on semantic entailment. On HUMAN-VERIFIED, `Qwen2.5-7B` recovers sharply: BM25-R reaches 0.367, the highest HUMAN-VERIFIED score across all models and retrievers, consistent with its tendency to over-predict Supported and Partially Supported, the dominant categories in the human-annotated set.

Retriever choice matters most for grounded claims. Across PAPERSOURCED and VERIFIABLE-ONLY, BM25+Reranker is usually the strongest retriever, beating pure BM25 and dense retrieval for every model on PAPERSOURCED and for most on VERIFIABLE-ONLY (the exception is `GPT-5-mini`, where dense retrieval is marginally higher). Reranking improves precision at low ranks, which is critical when the verifier's context is restricted to a small passage set. For LLMJUDGED, dense retrieval surpasses the reranker for `o4-mini` (0.514 vs. 0.499), suggesting that semantic embeddings better capture paraphrased or implicit evidence that lexical matching misses.

Table 8: Full-context accuracy (five models, six benchmarks). `Qwen2.5-7B` excludes 9 over-length papers (Section 4).

Benchmark	o4-mini	GPT-5-mini	Claude-Haiku	Qwen3-8B	Qwen2.5-7B
PaperSourced	**0.866**	0.792	0.764	0.618	0.753
RebuttalSourced	0.245	0.280	**0.295**	0.276	0.278
LLMJudged	**0.442**	0.379	0.295	0.171	0.170
HighAgreement	**0.536**	0.513	0.445	0.302	0.213
verifiable-only	0.495	**0.518**	0.388	0.320	0.250
Human-Verified	0.276	**0.336**	0.269	0.316	0.249

Table 9: Accuracy vs. macro-F1 (full-context FC and best-retriever RAG) on RebuttalSourced and LLMJudged. A large Acc−F1 gap reflects label-distribution bias, not reasoning. `Qwen2.5-7B` FC excludes 9 over-length papers.

	Full-Context				RAG (best retriever)			
	RebuttalSourced		LLMJudged		RebuttalSourced		LLMJudged	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1
o4-mini	0.245	0.246	0.442	0.416	0.244_{D}	0.236	0.514_{D}	0.454
GPT-5-mini	0.280	0.275	0.379	0.381	0.294_{D}	0.279	0.414_{D}	0.399
Claude-Haiku	**0.295**	0.294	0.295	0.293	0.297_{DR}	0.275	0.374_{DR}	0.336
Qwen3-8B	0.276	0.268	0.171	0.192	0.287_{D}	0.271	0.371_{D}	0.336
Qwen2.5-7B	0.278	0.269	0.170	0.190	$\mathbf{0.324}_{D}$	0.277	0.209_{D}	0.200

C.5 Retrieval Quality Analysis

Retrieval configuration. The dense retriever uses `all-MiniLM-L6-v2` sentence embeddings, and the cross-encoder reranker uses `cross-encoder/ms-marco-MiniLM-L-6-v2`. Manuscripts are segmented with Docling structure-aware (section/heading-based) chunking, producing variable-length chunks with no fixed token window (median 133 words per chunk, mean 132, p90 188). For each claim we retrieve an initial pool of $k = 20$ candidate passages; when reranking is enabled, these are reranked down to the final top-10 passages fed to the verifier, and non-reranked configurations directly take the top-10.

Table 10 compares four retrieval strategies on PaperSourced, where the ground-truth supporting section is known, enabling intrinsic evaluation. Dense+Reranker achieves the best performance across all metrics (Recall@5: 0.452, Recall@10: 0.530, nDCG@10: 0.684, MRR: 0.346), while BM25 alone is weakest.

Consistent with prior work (Kamalloo et al., 2024), reranking substantially boosts BM25. BM25+Reranker surpasses standalone dense retrieval at Recall@5 (0.434 vs. 0.420) but lags at Recall@10 (0.492 vs. 0.528), suggesting that lexical matching excels at placing relevant chunks at shallow ranks while dense representations improve deeper recall. Despite this, Recall@10 remains below 0.55 across all configurations: the ground-truth section is absent from the top-10 retrieved passages in nearly half of all cases. The Oracle upper bound (Table 7) makes this bottleneck concrete: the 8.4-point gap between Oracle (0.988) and Dense+R (0.904) for `o4-mini` is attributable entirely to retrieval imperfection.

D Related Works

LLMs for Peer Review. A growing body of work explores the use of large language models to support scholarly peer review through tasks such as review generation, review assistance, reviewer recommendation, and manuscript assessment (Yuan et al., 2022; Liang et al., 2024; Tomkins et al., 2017). A related line analyzes peer-review content and dynamics, including review quality, argumentation, sentiment, and reviewer behavior (Hua et al., 2019; Rogers and Augenstein, 2020; Kang et al., 2018; Ebrahimi et al., 2025). These efforts demonstrate that state-of-the-art language models can generate meaningful feedback and shed light on how reviews are

Table 10: Retrieval quality on PaperSourced.

Retriever	Recall@5	Recall@10	nDCG@10	MRR@10
BM25	0.398	0.474	0.570	0.289
BM25 + Reranker	0.434	0.492	0.608	0.331
Dense	0.420	0.528	0.644	0.299
Dense + Reranker	**0.452**	**0.530**	**0.684**	**0.346**

written and how they function socially. However, their primary objective is to produce, evaluate, or characterize reviews rather than to determine whether specific claims made within a review are supported by evidence contained in the reviewed manuscript. As a result, the problem of manuscript-grounded review verification remains largely unexplored.

Scientific Claim Verification. PEERIFY is also related to scientific claim verification and evidence-based fact checking, studied extensively through textual entailment and natural language inference, from early RTE-style formulations (Dagan et al., 2006) to large-scale benchmarks such as FEVER (Thorne et al., 2018), SciFact (Wadden et al., 2020), PubHealth (Kotonya and Toni, 2020), and scientific verification tasks (Malon, 2018; Wadden and Lo, 2021). These datasets have enabled systems that retrieve evidence and determine whether a claim is supported, contradicted, or unsupported, and retrieval-augmented approaches (Lewis et al., 2020; Izacard and Grave, 2021) have further improved evidence-grounded and multi-hop reasoning. While these settings share similarities with review verification, they typically treat claims as standalone statements and assume open-domain retrieval or citation-based reasoning across multiple sources, with evaluation centered on short claims whose supporting evidence is relatively concentrated. In contrast, peer-review claims often contain methodological observations, novelty assessments, and experimental critiques whose supporting evidence may be distributed across multiple sections of a single manuscript. Review verification is therefore a document-bounded grounding problem that requires reasoning over long, highly structured scientific documents while accounting for contextual qualifiers, experimental assumptions, and evidence aggregation.

Claim Decomposition. A further challenge arises from the fact that review comments frequently contain multiple assertions expressed within a single sentence or paragraph. Recent work on claim decomposition has shown that breaking complex statements into atomic claims improves the reliability and interpretability of downstream verification (Metropolitansky and Larson, 2025; Min et al., 2023; Scirè et al., 2024). Peer reviews complicate this step because they often mix subjective assessment with factual assertions and rely on hedging or implicit references. PEERIFY adopts a similar atomic-claim perspective and adapts extraction using multiple supervision signals, enabling more precise evidence attribution and systematic, manuscript-grounded groundedness assessment in an end-to-end evaluation.

Concurrent Work. A closely related concurrent system is FactReview (Xu et al., 2026), which also targets evidence-grounded peer review assessment. FactReview focuses on author-written claims extracted from the submitted manuscript, using literature retrieval and code execution to verify whether reported results are reproducible and positioned correctly relative to prior work. PEERIFY is complementary in scope and emphasis: we focus on *reviewer*-authored claims, statements made in the review about the paper rather than by the paper, and derive supervision from author–reviewer interaction dynamics and multiple independent label sources. This distinction matters because reviewer claims introduce interpretive and normative language absent from author-written text, and their groundedness must be assessed against a single manuscript rather than the broader literature. Our multi-benchmark design, spanning six supervision variants with different reliability and ambiguity profiles, enables systematic evaluation of the task difficulty spectrum.