Who's Adam? Benchmarking Hallucinations in Scientific Dialogue

Zexing Zhang*
zacheryzhang@163.com
National University of Defense Technology
Changsha, China

Tianyang Lei*
leitianyang20@163.com
National University of Defense Technology
Changsha, China

Kewei Yang
kayyang27@nudt.edu.cn
National University of Defense Technology
Changsha, China

Jichao Li†
lijichao09@nudt.edu.cn
National University of Defense Technology
Changsha, China

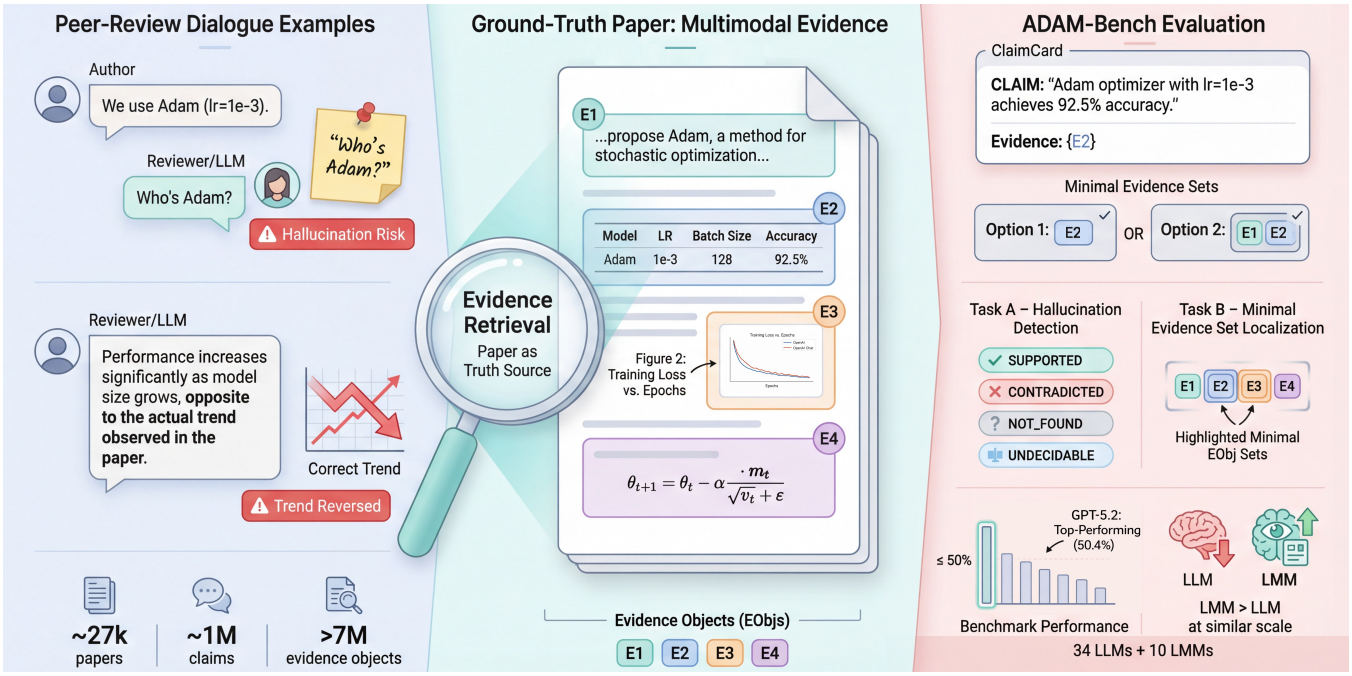

Figure 1: ADAM-Bench aligns atomic claims in scientific dialogue with multimodal evidence objects in the corresponding paper PDF. It enables paper-grounded hallucination detection (4-way grounding) and minimal evidence set localization with set-of-sets gold rationales at scale.

Abstract

LLMs and LMMs are increasingly applied in scientific dialogue, but it remains unclear whether they can reliably ground specific dialogue statements to paper-based evidence. A central challenge is *paper-grounded hallucination* under a paper-as-truth setting: statements that are contradicted by, not found in, or otherwise not decidable from the paper PDF. These hallucinations can be caused by both human misinterpretations and model-generated assertions, ultimately undermining the efficiency, fairness, and credibility of scientific dialogue. Existing benchmarks often overlook this issue, focusing either on subjective macro-level quality assessments or lacking cross-modal evidence localization. We introduce ADAM-Bench (Auditing Dialogue Assertions with Multimodal Evidence), a benchmark for paper-grounded hallucinations in scientific dialogue. Starting from around 27,000 papers, ADAM-Bench is a multi-layer benchmark with three tiers: Scale, Core, and Gold. ADAM-Bench pairs approximately 1 million atomic claims with over 7 million multimodal evidence objects extracted from the corresponding PDFs. We build it through a four-stage pipeline of claim atomization, candidate evidence recall, model-assisted pre-alignment, and human verification. Based on this dataset, we define two tasks: hallucination detection and minimal evidence set localization. Additionally, to avoid the brittleness introduced by single-rationale supervision, we formalize minimal evidence as a set of equivalent evidence sets and evaluate localization by best-matching against multiple gold

*The first two authors contributed equally to this research.
†Corresponding author.

evidence sets. We conduct a comprehensive benchmark of 34 LLMs and 10 LMMs, spanning large proprietary models (Claude-Opus-4-6, GPT-5.2) and open-source models (Qwen3-235B, GLM-4.6V 106B). Results are markedly low (25.2%–51.1%), indicating that grounding conversational hallucinations in real multimodal papers remains far from solved. We hope this benchmark will contribute to building scientific assistants that make calibrated judgments, cite minimal, auditable evidence, and mitigate the impact of hallucinations in scientific discovery evaluation.

Keywords

Scientific Dialogue, Hallucination, Paper-Grounded Verification, Multimodal PDF Evidence, Evidence Localization, Benchmark

1 Introduction

Scientific dialogue shapes high-stakes research decisions, including acceptance, funding, and what directions a field prioritizes. It spans reviews, rebuttals, and editorial discussions, and it increasingly involves LLMs and LMMs as participants or assistants. Recent work has started to evaluate whether LLMs can act as reviewers or provide useful review assistance, signaling rapid adoption but leaving reliability unresolved [21, 37].

A particularly consequential failure mode in scientific dialogue is hallucination. In this paper, a hallucination is a concrete statement about a paper that cannot be grounded to evidence in the paper PDF, and it can be introduced by either humans or models.

For example, a reviewer may ask "Who's Adam?" after a dialogue claims "the paper uses Adam (lr=1e-3)," or a model may confidently invert the direction of an ablation trend; both can derail discussion and trigger unnecessary back-and-forth (Figure 1).

Such hallucinations are rarely blatant. They are typically subtle, arising from small but consequential misreadings of figures, tables, or experimental details, such as confusing a table column or incorrectly claiming that an experiment is missing when it is present. Yet their impact is outsized. Even small hallucinations can waste reviewer and author time, distort judgments of novelty and rigor, and erode efficiency, fairness, and trust in peer review outcomes.

Core motivation and questions. If scientific dialogue is increasingly co-written with models, then fluency is not the bottleneck. **The bottleneck is whether a system can make each dialogue statement *paper-grounded* under the paper-as-truth setting**, i.e., verifiable or falsifiable by concrete evidence inside the paper PDF. This leads to four concrete questions:

- **Hallucination detection (Task A):** Can a system correctly judge whether a claim is supported by the paper, contradicted by the paper, not found in the paper, or not decidable from the PDF, especially distinguishing "not found" from "not decidable" when the PDF is incomplete, ambiguous, or poorly parsed?
- **Evidence (Task B):** Can it localize the *minimal* set of multimodal evidence objects, rather than providing long, non-actionable citations?
- **Layout robustness:** Can it handle layout-dependent semantics, where the meaning of a number depends on a table header, a legend, or an equation reference?
- **Evaluation robustness:** Can we evaluate these abilities objectively and reproducibly while accounting for multiple valid evidence paths within real papers?

Despite its importance, paper-grounded hallucination is rarely treated as an explicit evaluation target in existing datasets and benchmarks. Peer-review datasets and benchmarks largely focus on coarse, outcome-level properties, such as review helpfulness, ratings, acceptance prediction, or overall quality. For example, PeerRead enables large-scale analysis of reviews, but it does not provide claim-by-claim verification grounded in the paper with multimodal evidence localization [9]. Evidence-based claim verification benchmarks such as FEVER and SciFact evaluate support or contradiction with evidence, but they are not grounded in peer-review dialogue and typically operate over non-PDF textual sources or abstracted evidence [22, 23]. Paper-centered QA datasets such as QASPER focus on information-seeking questions, rather than verifying dialogue statements and localizing minimal evidence across PDF modalities [4, 5, 7, 12, 15, 16, 18, 20]. This gap is becoming more urgent as OpenReview-style forums make peer-review dialogue widely accessible and as LLM-assisted reviewing becomes easier to deploy at scale [26]. Moreover, grounding dialogue statements to papers is not merely a textual entailment problem, because decisive evidence often lives in figures, tables, and equations whose semantics depend on layout and cross-modal alignment.

We introduce **ADAM-Bench** (**A**uditing **D**ialogue **A**ssertions with **M**ultimodal evidence), a datasets-and-benchmarks resource for paper-grounded hallucinations in scientific dialogue. ADAM-Bench treats the paper PDF as the sole truth source. It evaluates whether a dialogue statement is SUPPORTED, CONTRADICTED, NOT_FOUND, or UNDECIDABLE with respect to the paper. We construct ADAM-Bench from ICLR OpenReview threads paired with their paper PDFs. We use a multi-stage pipeline that combines claim atomization, candidate evidence recall, LMM pre-alignment, and human arbitration.

ADAM-Bench defines two benchmark tasks that jointly test judgment and justification. Task A is hallucination detection as 4-way paper-grounded claim classification. Task B is minimal evidence set localization, where a system outputs one or more minimal sets of evidence objects that justify its label decision. To reflect that real papers often admit multiple valid rationale paths, we annotate gold evidence as a set of alternative minimal evidence sets and evaluate localization by best matching against these alternatives.

We release a large-scale corpus extracted from about 27,000 papers, containing about 1 million dialogue assertions and over 7 million multimodal evidence objects. We also provide a curated gold evaluation subset and score systems using Macro-F1, Evidence-F1, and a FEVER-style metric that rewards both correct labels and sufficient evidence when applicable [22].

Our benchmark results highlight a deeper reliability gap than label accuracy alone can reveal. Across 34 LLMs and 10 LMMs, overall performance remains low, showing that current systems struggle to ground dialogue statements in real multimodal PDFs at benchmark scale. At similar parameter scales, LMMs consistently outperform LLMs, which suggests that access to visual and layout cues is not a cosmetic feature but a functional advantage for evidence-grounded scientific dialogue. At the same time, the remaining gap indicates

that multimodality alone is insufficient, and that scientific assistants must be explicitly optimized for evidence selection, minimal justification, and robustness to ambiguity.

Our contributions and findings are as follows.

- We cast *paper-grounded hallucination* in scientific dialogue as an evidence-centric verification problem under the paper-as-truth setting, and we define two benchmark tasks: hallucination detection and minimal evidence set localization.
- We construct and release ADAM-Bench at scale by aligning atomic claims to a unified multimodal evidence-object representation over real paper PDFs, and we organize the release into three tiers (Scale/Core/Gold) that separate corpus-scale weak alignment from evaluation-grade supervision.
- We design an evaluation protocol for minimal evidence that captures multiple valid rationale paths via set-of-sets gold annotations and best-match scoring, improving robustness and fairness for real-paper evidence localization.
- We benchmark 34 LLMs and 10 LMMs and release a reproducible evaluation toolkit and leaderboard, revealing a persistent gap between hallucination label correctness and minimal evidence localization.

We hope ADAM-Bench will catalyze research on grounding and evidence localization for scientific dialogue and enable tools that reduce hallucinations in scientific discovery evaluation.

2 Related Work

Peer-review corpora and interaction annotations. Public peer-review corpora enable studying review writing, decision making, and reviewer-author interactions at scale [2, 9, 10, 27, 29, 32–34]. PeerRead is an early and widely used resource that pairs papers with accept or reject decisions and review texts [9]. ReviewAdvisor discusses automated scientific reviewing and releases data and code for review-oriented generation and evaluation [33]. DISAPERE annotates discourse relations in review and rebuttal pairs and supports fine-grained interaction analysis [10]. Re^2 provides full-stage peer review threads with multi-turn rebuttal and discussion to model longitudinal dialogue [34]. Analyses of OpenReview further characterize how openness, stages, and participation patterns relate to outcomes [27]. **These corpora primarily target macro properties or discourse phenomena, rather than claim-level verification grounded to the paper as the sole truth source [9, 10, 27].**

LLMs and LMMs for peer review. Recent work evaluates LLMs as reviewers or review assistants, typically measuring helpfulness, completeness, or agreement with human assessments [8, 30, 37]. Zhou et al. provide a comprehensive evaluation of LLMs on automatic paper reviewing tasks [37]. MMReview expands evaluation to multimodal review settings and diverse disciplines with multiple review automation tasks [8]. Another line studies how to evaluate subjective judgments without gold references, which is relevant to peer review as a judgment-generation task [30]. **However, these benchmarks rarely ask whether a specific dialogue assertion is supported, contradicted, missing, or undecidable *with respect to the paper PDF* [8, 37]. They also rarely require minimal and verifiable evidence localization inside the PDF [8, 37].**

Evidence-based verification and rationale selection. Fact verification benchmarks such as FEVER formalize predicting a label together with retrieving supporting evidence and popularize evidence-conditioned scoring [22, 35]. SciFact brings verification to scientific claims, but it typically grounds evidence in abstracts rather than full paper PDFs [24]. FEVEROUS and TabFact extend evidence beyond plain sentences by incorporating structured sources such as tables and cells [1, 3]. SEM-TAB-FACTS further targets scientific tables with both verification and cell-level evidence finding [28]. **These benchmarks motivate our evidence-centric scoring design, but they are not derived from peer-review dialogue and they do not treat the paper PDF as the only truth source [22, 24, 28].**

Paper-grounded QA and multimodal PDF understanding. QASPER anchors questions and answers in research papers and includes evidence spans, figures, or tables for some questions [6]. Document VQA and chart QA benchmarks demonstrate that layout, visual encodings, and graphical conventions are necessary for correct reasoning [13, 14]. Pretrained document models integrate textual content with layout or visual features, which can serve as foundations for evidence access in PDFs [11, 31]. Large-scale layout and table extraction datasets support building unified evidence-object stores over PDFs [17, 19, 36]. **Yet these resources do not connect evidence objects to dialogue assertions from scientific evaluation processes [6, 36].**

Positioning of ADAM-Bench. *Our work bridges peer-review dialogue resources and evidence-based verification by centering on paper-grounded hallucinations in scientific dialogue [6, 9, 22, 24]. Unlike macro-level review evaluation, we require claim-level labels defined only by what is present in the paper PDF [8, 37]. Unlike typical paper QA, we evaluate minimal evidence set localization for verification-oriented dialogue assertions [6]. We also explicitly handle multiple valid rationale paths by annotating alternative minimal evidence sets and scoring by best matching against them.*

3 ADAM-Bench

ADAM-Bench targets paper-grounded hallucinations in scientific dialogue by coupling each claim with multimodal evidence in the corresponding paper PDF under a fixed evidence representation. The construction is organized as four stages that progressively narrow a large candidate space into evaluation-grade supervision, yielding the Scale, Core, and Gold layers. Figure 2 summarizes the four stages and the artifacts exchanged between them. Low-level implementation details, including trigger inventories, prompts, caching, and scripts, are documented on our project page.[1] We provide additional construction details and heuristics in Appendix A.

3.1 Four-stage Construction Pipeline

Step 1: Evidence-first extraction. The pipeline ingests OpenReview forums into a paper manifest and paired discussion threads, retaining essential paper-level metadata such as year, title, PDF link, license, and research area. Each thread is normalized into an ordered utterance sequence with rich attributes (e.g., speaker role,

[1] Project page: https://adam-bench.github.io/Repo/. We will actively maintain this page, which includes the data, code, leaderboard, and evaluation/annotation tools.

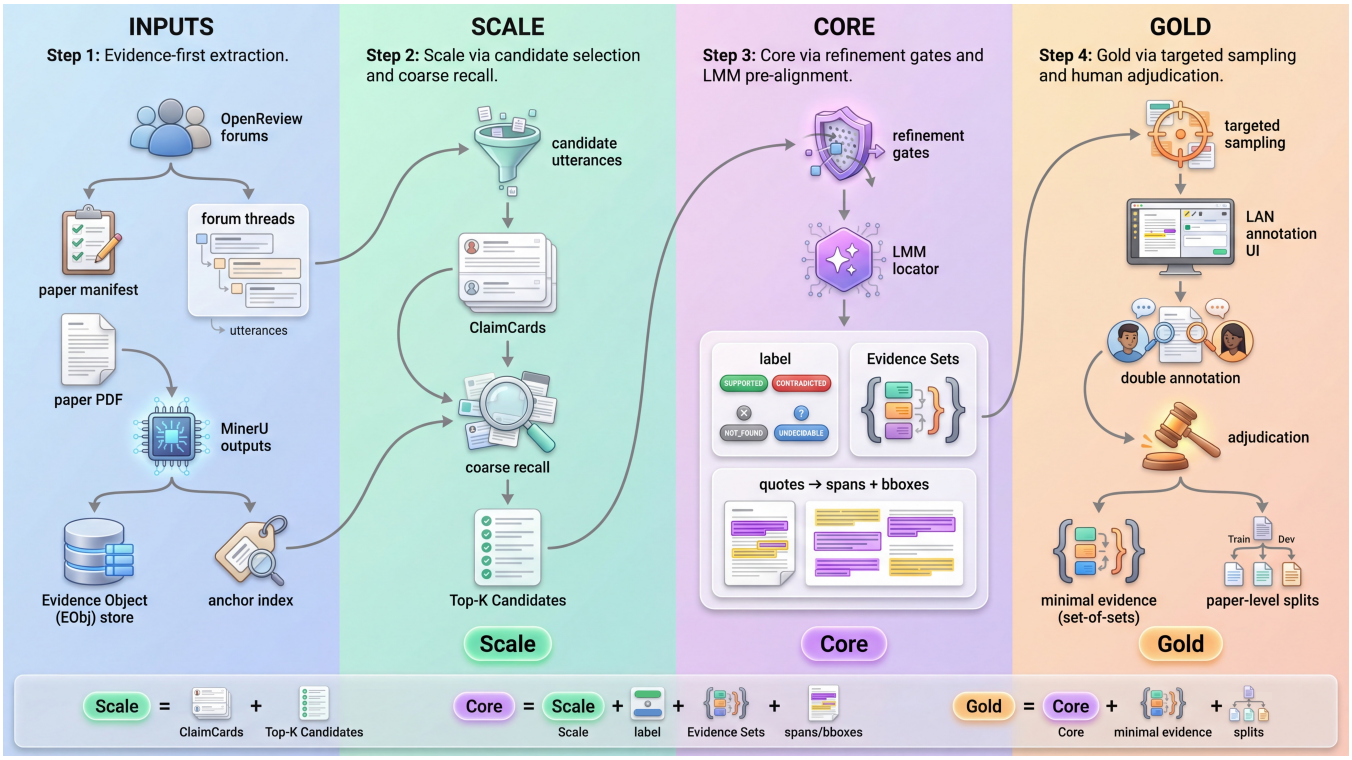

Figure 2: Four-stage construction pipeline and layer refinement: starting from raw PDFs, we extract multimodal evidence objects and align them with dialogue claims, progressively refining supervision from Scale (coarse recall) to Core (model-assisted pre-alignment) and Gold (human-verified minimal evidence).

review stage, timestamps, and text), and we filter utterances by a minimum length to avoid degenerate fragments.

Each paper PDF is parsed using MinerU [25] outputs to construct a unified Evidence Object (EObj) store that is the only evidence unit used throughout recall, localization, annotation, and evaluation. Each EObj has a stable identifier and stores the minimal information needed for retrieval and inspection, including its modality/type, hierarchical context, page location, region box, canonical text view, and optional visual crops; an anchor index further supports fast linking between references and evidence units. The full EObj schema is deferred to Appendix A.1.

A manifest gate further filters to papers that satisfy licensing constraints and contain both substantive review content and an editorial decision; it also requires PDF reachability and successful MinerU parsing. Normalization, parsing, and manifest gates are detailed in Appendix A.2.

Step 2: Scale via candidate selection and coarse recall. Scale is constructed by selecting candidate utterances, atomizing them into ClaimCards, and then retrieving a per-claim Top-K candidate list from the EObj store within the same paper. Candidate utterances are selected using rule-based triggers that target checkable statements, including anchors, numbers, metric mentions, comparative language, ablation cues, and missing-experiment cues.

Each candidate utterance is atomized into one or more ClaimCards via a heuristic or LLM extractor, then deduplicated.

ClaimCards are assigned to a small set of categories, including method description, experimental result, and missing experiment.

Coarse recall ranks EObjs using lexical retrieval over their canonical text view and returns the Top-K EObjs as candidates, and anchor signals are used to prioritize EObjs that share anchors with the claim when such signals are present. Scale is released as ClaimCards paired with their Top-K Candidates, which reduces the evidence search space per claim while keeping supervision weak. Trigger rules, atomization settings, and retrieval configuration are detailed in Appendix A.3.

Step 3: Core via refinement gates and LMM pre-alignment. Core is obtained by applying stricter inclusion gates to Scale and then running an LMM locator over the Top-K candidate pool for each remaining claim. The Core gates require high-quality papers and stable evidence construction, and they filter to claims that are specific enough for paper-grounded verification under bounded context.

Given a ClaimCard and its Top-K Candidates, the locator predicts a label in {SUPPORTED, CONTRADICTED, NOT_FOUND, UNDECIDABLE} and proposes one or more Evidence Sets, each represented as a set of `eobj_ids`. The locator also emits short quotes, which are deterministically aligned back to EObj character spans and page-localized bounding boxes for auditable inspection. Core is released with ClaimCards, Top-K Candidates, and locator outputs, and span or bbox metadata is auxiliary and does not change the

EObj-keyed evaluation interface. Core gates and locator configuration are detailed in Appendix A.4.

Step 4: Gold via targeted sampling and human verification. Gold is produced by selecting a targeted subset of Core instances and verifying labels and minimal evidence in a lightweight LAN annotation UI. The Gold sampling gates prioritize instances where candidate evidence is non-trivial yet decidable under the paper-as-truth setting, which controls annotation cost while retaining evaluation value. Gold annotations assign the 4-way label and minimal evidence as a set-of-sets of Evidence Sets to allow multiple equivalent minimal rationales. Evidence is evaluated by set-level matching over evidence-object identifiers, while spans and bounding boxes are retained for auditability and UI rendering. Train, development, and test splits are created at the paper level and stratified by year and research area. Sampling gates, annotation guidelines, and split construction are detailed in Appendix A.5.

Table 1: Operational definition of layers by released artifacts.

Artifact	Scale	Core	Gold
Utterances (role, stage, text)	✓	✓	✓
EObj store (`eobj_id` keyed)	✓	✓	✓
ClaimCards (typed, deduplicated)	✓	✓	✓
Top-K Candidates per claim	✓	✓	✓
LLM locator label and evidence sets		✓	✓
Quote-aligned spans and bboxes (aux.)		✓	✓
Human-verified label			✓
Gold minimal evidence (set-of-sets)			✓
Paper-level stratified splits			✓

3.2 Layer Definitions

The three layers share the same fixed evidence representation (EObjs) and differ in refinement strength and supervision, preventing weak alignment from being mistaken as evaluation-grade annotation. Table 1 provides an operational definition of the released layers by artifacts, while Table 2 summarizes their scale.

- Scale provides ClaimCards paired with Top-K Candidates and primarily serves as a reduced search-space corpus for modeling and retrieval research.
- Core further reduces the candidate space by filtering and adds model-produced alignment signals (labels and evidence sets) together with deterministic quote-to-span localization metadata for auditing.
- Gold applies additional filtering and human verification to produce evaluation-grade supervision with minimal evidence set-of-sets, balancing annotation cost with reliability and comparability.

3.3 Tasks and Terminology

- **Task A** (Hallucination Detection) classifies each ClaimCard under the paper-as-truth setting into {SUPPORTED, CONTRADICTED, NOT_FOUND, UNDECIDABLE}.
- **Task B** (Evidence Localization) outputs one or more Evidence Sets as `eobj_id` lists, and may optionally include spans and bounding boxes as auxiliary localization metadata.

Gold evidence is annotated as a set-of-sets of minimal Evidence Sets, and localization is evaluated by best matching against alternative gold sets to avoid brittleness from single-rationale supervision. Official metrics and the evaluation protocol are specified in Section 4.1.

Table 2: Scale statistics of released layers. For Gold, the evidence-object count reports the number of unique cited evidence objects in the gold annotations.

Layer	Papers	Claims	Evidence Objects
Scale	27,532	1,159,103	7,834,351
Core	3,000	114,320	799,289
Gold	563	5,821	69,852

4 Experiments and Discussion

This section evaluates paper-grounded hallucination detection and evidence localization on ADAM-Bench and uses targeted analyses to characterize where current systems fail. The focus is on comparability and diagnostic value rather than incremental modeling tricks.

4.1 Evaluation Protocol

All primary results are reported on the Gold development split, and the Gold test split is reserved for leaderboard evaluation. Each instance consists of a ClaimCard and the corresponding paper PDF represented by Evidence Objects (EObjs), and systems must predict both a label and supporting evidence when applicable. Task A is a 4-way classification over {SUPPORTED, CONTRADICTED, NOT_FOUND, UNDECIDABLE}. Task B outputs one or more Evidence Sets, where each Evidence Set is a set of `eobj_ids`, and Gold evidence is annotated as a set-of-sets of minimal Evidence Sets. Macro-F1 is used for Task A, Evidence-F1 uses set-level best matching against alternative Gold sets, and we additionally report a FEVER-style score [22] that requires both correct labels and sufficient evidence for SUPPORTED/CONTRADICTED.

We use Macro-F1 to treat labels equally under class imbalance, Evidence-F1 to evaluate minimal evidence localization while accounting for multiple valid gold rationales, and FEVER-style scoring to summarize end-to-end correctness by requiring both correct labels and sufficient evidence when applicable. In addition to Macro-F1, we report per-label F1 for SUPPORTED/CONTRADICTED/NOT_FOUND/UNDECIDABLE (F1(S)/F1(C)/F1(NF)/F1(U)) to diagnose asymmetric failure modes, such as confusing NOT_FOUND with UNDECIDABLE. Formal metric definitions are provided in Appendix G. Evidence spans and bounding boxes are retained for auditability and analysis, while the official evidence key remains `eobj_id`. The prediction file schema is shown in Appendix B, and the official evaluator and submission interface are described in Appendix C.1; evaluation commands are provided on the project page.

4.2 Systems and Settings

Experiments consider two evaluation settings that reflect typical deployment constraints. In *provided-evidence* setting, a system receives the released Top-K candidate EObjs per claim and must select

Table 3: Overall performance on ADAM-Bench. Provided-evidence metrics are reported in percent. Avg is the mean of Provided-evidence Macro-F1 and Evidence-F1. Wins counts pairwise wins aggregated over the seven Provided-evidence metrics against all other model baselines (strictly greater; ties not counted), excluding diagnostic rows. Best results within each block are bolded and second best are underlined. Superscript * marks the best open-source model by Avg in each block. Light shading in the Systems column groups models of similar parameter scale across LLMs and LMMs. Closed-book columns report the claim-only setting (no provided candidate evidence).

Systems	Closed-book		Provided-evidence							Summary	
	Macro-F1	FEVER	Macro-F1	Evidence-F1	FEVER	F1(S)	F1(C)	F1(NF)	F1(U)	Avg	Wins
Lower bounds and diagnostics											
R-ONLY (lower bound)	–	–	16.5	25.2	14.9	66.1	0.0	0.0	0.0	–	–
Oracle-in-candidates (upper bound)	–	–	100.0	75.9	81.8	100.0	100.0	100.0	100.0	–	–
Text models (LLMs)											
GPT-5.2	13.4	**36.6**	**45.6**	55.1	**51.0**	67.6	30.3	68.4	16.3	**50.4**	**107**
Claude-Opus-4-6	24.1	32.8	42.8	**57.2**	49.3	69.9	28.6	66.9	5.7	50.0	100
DeepSeek-V3.2	13.7	35.3	41.3	49.5	42.1	70.3	31.6	63.2	0.0	45.4	65
Gemini-3-Flash-Preview	**31.8**	10.7	44.1	53.4	46.6	**73.4**	37.2	65.7	0.0	48.8	93
Qwen3-235B-A22B-Instruct-2507*	16.1	36.4	44.3	54.0	40.5	72.3	**44.0**	60.8	0.0	49.2	85
Qwen2.5-72B-Instruct	14.9	35.0	40.7	49.4	37.2	72.0	36.8	49.5	4.4	45.1	70
Qwen2.5-0.5B-Instruct	25.2	27.5	19.7	30.6	27.8	32.4	0.0	46.3	0.0	25.2	6
Qwen2.5-7B-Instruct	26.0	17.4	36.7	34.5	15.2	68.0	27.7	47.2	3.7	35.6	35
Qwen2.5-14B-Instruct	13.4	**36.6**	36.4	38.0	27.0	72.2	14.8	58.4	0.0	37.2	39
Qwen2.5-32B-Instruct	13.7	**36.6**	39.6	53.0	43.0	71.0	19.4	62.3	5.9	46.3	72
Qwen3-8B	20.1	35.3	33.9	43.2	28.9	71.0	27.6	37.0	0.0	38.6	34
Qwen3-32B	14.3	36.4	35.3	53.2	40.2	70.6	13.8	56.7	0.0	44.2	43
Qwen3-Max	16.0	36.4	43.9	54.0	45.2	73.1	38.3	64.4	0.0	49.0	90
DeepSeek-R1-Distill-Qwen-14B	19.1	30.6	42.6	45.6	36.1	70.0	42.1	52.5	5.7	44.1	69
Grok-4.1	13.4	**36.6**	44.2	54.1	**51.0**	66.0	24.2	**68.8**	**17.9**	49.2	99
Multimodal models (LMMs)											
ChatGPT-4o	21.7	31.1	27.5	38.7	24.8	69.7	0.0	32.0	8.3	33.1	29
Gemini-2.5-Flash	22.7	26.4	29.9	38.1	22.6	68.6	0.0	50.9	0.0	34.0	16
Qwen3-VL-Max	16.4	**36.4**	45.8	54.6	**46.6**	71.4	**45.0**	**66.7**	0.0	50.2	101
Qwen2.5-VL-32B-Instruct*	13.7	**36.4**	40.5	**55.7**	41.6	71.8	24.0	66.2	0.0	48.1	75
GLM-4.6V 106B	**26.0**	30.6	37.3	53.2	38.0	58.1	14.3	66.2	10.5	45.2	62
Qwen3-VL-Plus 235B	13.7	34.4	**47.4**	54.8	43.8	**73.2**	34.3	62.6	**19.5**	**51.1**	**115**

evidence and predict labels. In *closed-book* setting, a system receives the claim (and dialogue context window) only and must predict labels, with empty evidence by definition.

For provided-evidence, both text-only and multimodal variants are evaluated; multimodal variants may additionally consume a capped number of candidate visual crops extracted from the PDF. All reported results use the official coarse-recall candidates produced by the released pipeline, and K is fixed to a single default for the main table. The default K and multimodal caps are specified in Appendix C.2. We report a retrieval-only lower bound (R-ONLY) that ranks candidate evidence by sparse matching (BM25 and TF-IDF with reciprocal-rank fusion), outputs the top evidence object, and assigns SUPPORTED by default to isolate retrieval coverage from reasoning. We additionally report an oracle-in-candidates upper bound that assumes access to Gold evidence when it appears in the released candidate pool, quantifying how much headroom remains beyond coarse recall. We also evaluate a family of prompted LLM or LMM baselines that predict labels and evidence sets from the same candidate interface. System prompts and decoding parameters are fixed across models within each family; we use deterministic decoding (temperature = 0) to emphasize benchmark difficulty rather than prompt engineering. Prompts are included in Appendix C.3 for reproducibility.

4.3 Main Results

Table 3 reports Gold-dev results under both settings and highlights a persistent gap between hallucination labeling and minimal evidence localization. Due to space constraints, Table 3 highlights a representative subset of systems; the full evaluated model suite is listed in Appendix D.

- Closed-book remains low (best 31.8 Macro-F1; 36.6 FEVER), suggesting that paper-grounded verification is not reducible to prior domain knowledge.
- Provided-evidence improves labels (best 47.4 Macro-F1), but evidence localization remains hard (best 57.2 Evidence-F1; 51.0 FEVER), especially for contradicted and undecidable claims.

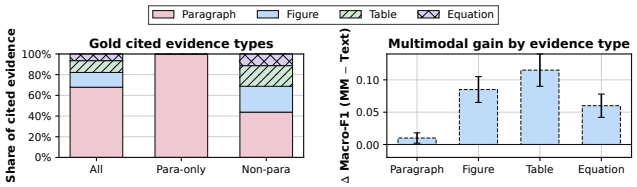

Figure 3: Evidence types and the multimodal advantage.

- Multimodal variants are competitive (best 51.1 Avg) and help on PDF-native hard cases.

Table 3 further provides lower/upper-bound diagnostics. R-ONLY achieves 16.5 Macro-F1 and 14.9 FEVER, whereas oracle-in-candidates reaches 75.9 Evidence-F1 and 81.8 FEVER. The gap to the best baselines (47.4 Macro-F1, 57.2 Evidence-F1, 51.0 FEVER) shows that reasoning and minimal evidence selection, rather than coarse recall, remain the primary bottlenecks.

Finding 1: Retrieval alone is far from sufficient: the retrieval-only lower bound remains weak, showing that ADAM-Bench requires reasoning about claims and selecting minimal evidence beyond coarse candidate recall.

Finding 2: Oracle-in-candidates exposes substantial headroom beyond coarse recall: even when Gold evidence is present among candidates, selecting the right minimal evidence and making correct judgments remains far from solved.

Per-label results concentrate difficulty on contradiction and undecidability. F1(NF) peaks at 68.8, but F1(C) reaches only 45.0 and F1(U) only 19.5, explaining why end-to-end verification remains challenging even under provided-evidence and aligning with Figure 1.

Finding 3: Errors concentrate in nuanced labels: deciding contradiction and undecidability is consistently harder than abstaining with not-found, making these fine-grained distinctions the dominant source of failure.

Finding 4: Multimodal and text models excel on different aspects of verification: cross-modal perception helps on PDF-native evidence, while careful language reasoning remains critical for matching and minimizing evidence.

Scaling helps but saturates. In Qwen2.5, Avg rises from 25.2 (0.5B) to 46.3 (32B) with Evidence-F1 from 30.6 to 53.0, while 72B is comparable (Avg 45.1; Evidence-F1 49.4). In Qwen3, Avg increases from 38.6 (8B) to 49.2 (235B), but Evidence-F1 is non-monotonic (28.9 at 8B, 45.2 at max, 40.5 at 235B), underscoring the minimal-evidence bottleneck.

Finding 5: Scaling improves performance but does not resolve minimal-evidence localization; gains saturate and remain sensitive to model family and tuning.

Open-source models are competitive with closed ones. For text models, the best closed model reaches Avg 50.4 (GPT-5.2) versus 49.2 for the best open model (Qwen3-235B-A22B-Instruct-2507), while FEVER differs more (51.0 vs 40.5). For multimodal models, the best closed model reaches Avg 51.1 (Qwen3-VL-Plus 235B) versus 48.1 for the best open model (Qwen2.5-VL-32B-Instruct*), and the oracle gap still indicates substantial headroom.

Finding 6: Open models are already competitive with closed models on provided-evidence, but end-to-end verification remains the main unresolved bottleneck.

Overall, Table 3 separates retrieval, judgment, and minimal evidence localization; the large oracle gap and weak contradiction/undecidability performance highlight a clear frontier in grounding scientific dialogue to PDF-native evidence with calibrated abstention and minimal rationales.

4.4 Where Multimodality Matters

To understand why multimodal variants help, evidence usage is analyzed by evidence object types and by whether the Gold evidence set contains at least one non-paragraph object. Figure 3 summarizes the distribution of cited evidence types in Gold-dev and the conditional accuracy gap between text-only and multimodal settings. A consistent pattern is that claims grounded in figures, tables, or equations exhibit larger gaps, matching the benchmark's motivating failure cases such as trend reversal and column misreading.

Finding 7: Multimodal inputs yield the largest gains on claims whose minimal evidence includes figures, tables, or equations, suggesting that text-only pipelines systematically miss layout- and visual-dependent cues even when candidate evidence is retrieved.

4.5 Ablations and Diagnostics

This subsection isolates pipeline components that affect benchmark difficulty while staying within the released interfaces. We emphasize two diagnostics: (i) how the candidate pool size K trades off computation with end-to-end verification quality, and (ii) how anchor signals influence retrieval and evidence localization.

Top-K sweep. To make the K effect interpretable, we fix two backbones at a similar scale: Qwen2.5-32B-Instruct and its multimodal counterpart Qwen2.5-VL-32B-Instruct*. We sweep K with a step of 5, i.e., $K \in \{5, 10, 15, 20, 25\}$, where $K = 15$ matches the default used in Table 3. Figure 4 reports performance under the provided-evidence setting: the top row is Provided-evidence Macro-F1 (Task A), and the bottom row is Evidence-F1 (Task B). We additionally stratify claims by the dominant evidence object type in the minimal Gold set (paragraph/text, figure, table, equation) to diagnose where multimodality is most helpful.

Overall, increasing K yields clear gains from $K = 5$ to $K = 15$, but improvements saturate beyond $K = 15$, indicating diminishing returns from enlarging the candidate set alone. The multimodal backbone improves Evidence-F1 more consistently than Macro-F1,

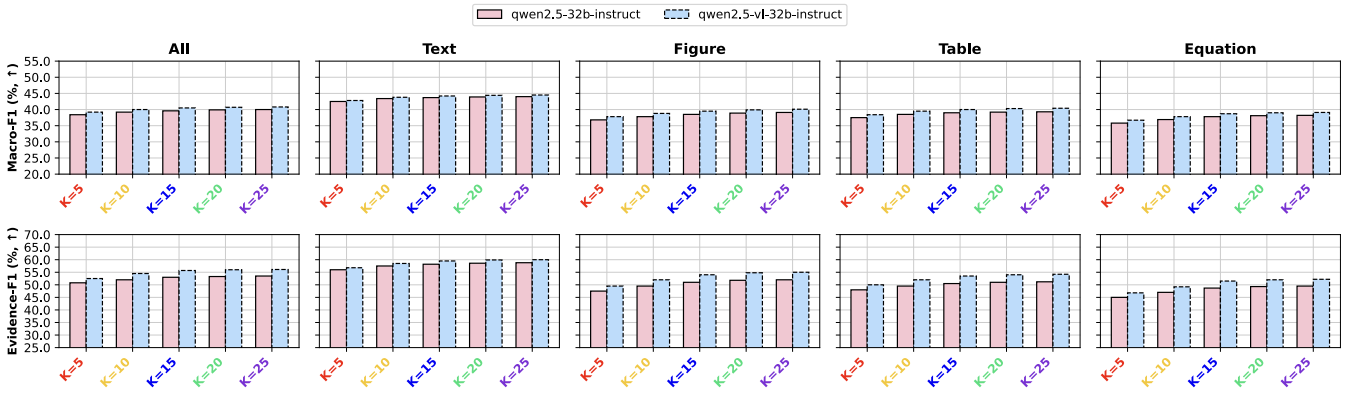

Figure 4: Top-K sweep (step=5) on Gold-dev under the provided-evidence setting. We compare Qwen2.5-32B-Instruct (text-only) and Qwen2.5-VL-32B-Instruct* (multimodal) with $K \in \{5, 10, 15, 20, 25\}$; $K = 15$ corresponds to the main-table default (Table 3). Columns stratify claims by the dominant evidence type in the minimal Gold set (Text/Figure/Table/Equation) plus Overall. Rows report Provided-evidence Macro-F1 (top) and Evidence-F1 (bottom).

and the largest gaps typically appear in Figure/Table/Equation buckets, aligning with the intuition that layout- and vision-dependent cues are hard to recover from text-only representations.

> ***Finding 8:*** Increasing the candidate pool helps only up to a moderate K and then saturates; benchmark difficulty is therefore not relieved by "retrieving more" alone, and multimodal inputs matter most when minimal evidence is non-paragraph (figures/tables/equations).

Anchor ablation. Table 4 reports an anchor ablation that highlights the role of explicit identifiers in scientific dialogue grounding.

Table 4 shows that removing anchor signals degrades both label prediction and evidence localization (e.g., Macro-F1 drops from 0.398 to 0.371; Evidence-F1 drops from 0.407 to 0.381).

> ***Finding 9:*** Anchor signals are a key ingredient for scientific dialogue grounding: explicit identifiers improve both correctness and minimal evidence selection.

Table 4: Anchor ablation on provided-evidence (text-only).

Variant	Macro-F1	Evidence-F1	FEVER
With anchor signals	0.398	0.407	0.329
Without anchor signals	0.371	0.381	0.304

4.6 Error Analysis

Error analysis is performed by inspecting cases where labels are correct but minimal evidence is missing, and cases where evidence is plausible but the label is wrong. Table 5 summarizes frequent failure buckets observed during audit. A recurring pattern is that models often produce "nearby" evidence that is topically related but does not uniquely justify the claim, which is penalized by minimal

> ***Finding 10:*** Minimal evidence evaluation exposes a systematic "near-miss" behavior, where systems retrieve and cite related regions but fail to isolate the decisive evidence needed to justify a label.

Table 5: Common error buckets observed during audit.

Bucket	Description	Share
Near-miss evidence	Evidence is relevant but not sufficient or not minimal	31%
Modal misread	Trend reversed, column misread, or axis misinterpreted	18%
Scope drift	Claim quantifiers or conditions mismatch the paper statement	16%
Missing-exp false alarm	Claim asserts absence while paper contains the experiment	14%
Underspecified	Claim lacks enough detail to be decidable from paper alone	12%
Parser boundary	Evidence crosses object boundaries or is split unexpectedly	9%

5 Conclusion

We present ADAM-Bench, a benchmark for evaluating grounding hallucinations in scientific dialogue under a paper as truth assumption. By aligning atomic dialogue claims to a unified set of multimodal evidence objects extracted from paper PDFs, ADAM-Bench jointly evaluates claim grounding and minimal evidence localization. To reflect the inherent ambiguity of scientific justification, we annotate gold supervision as a set of alternative minimal evidence sets and evaluate localization by best matching, enabling robust comparison across equivalent reasoning paths. The benchmark is released in three layers, Scale, Core, and Gold, which share a fixed evidence representation while decoupling large-scale weak alignment from evaluation-grade supervision. Benchmark results show that current systems still struggle to produce both correct grounding decisions and concise, decisive evidence; multimodal inputs and explicit anchors help on PDF-native hard cases, while increasing candidate evidence provides diminishing returns, highlighting calibration and evidence minimization as key challenges.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC), Grant Nos. 72501301, 72571282, 72231011, and 72421002.

References

[1] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. https://openreview.net/forum?id=h-flVCIlstW

[2] Milena Baumgärtner, Shahbaz Syed, Fei Tony Liu, Kevin Luu, Thomas Markovich, Anurag Mohananey, Lu Chen, Nils Ole Tippenhauer, and Christopher Ré. 2025. PeerQA: A Scientific Question Answering Dataset from Peer Reviews. arXiv:2502.13668 [cs.CL] https://arxiv.org/abs/2502.13668

[3] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. [n. d.]. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations*.

[4] Zhiyu Chen, Yiqiao Liu, Hongliang Liu, Xiaohan Zhang, Xiaoyuan Lu, Songfang Huang, Yu Du, Xiaomeng Huang, and Liang Wang. 2025. VisR-Bench: A Multilingual Benchmark for Retrieval-Augmented Generation with Visual Information. arXiv:2508.07493 [cs.CL] https://arxiv.org/abs/2508.07493

[5] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 4599–4610. doi:10.18653/v1/2021.naacl-main.365

[6] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4599–4610.

[7] Yuxuan Ding, Jiahao Liu, Qi Shen, Hongsheng Li, and Qi Wu. 2024. PDF-MVQA: A Dataset for Multimodal Information Retrieval in PDF-based VQA. arXiv:2404.12720 [cs.CV] https://arxiv.org/abs/2404.12720

[8] Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. MMReview: A Multidisciplinary and Multimodal Benchmark for LLM-Based Peer Review Automation. *arXiv preprint arXiv:2508.14146* (2025).

[9] 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)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 1647–1661. doi:10.18653/v1/N18-1149

[10] Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1234–1249.

[11] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*. Springer, 498–517.

[12] Yaiza López, Juan Pérez, Montse Cuadros, and Nando Garcia. 2025. Enhancing Document VQA Models via Retrieval-Augmented Generation. arXiv:2508.18984 [cs.CV] https://arxiv.org/abs/2508.18984

[13] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*. 2263–2279.

[14] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2200–2209.

[15] Jihun Park, Eunji Kim, Minwoo Kang, Junyoung Kim, Donghyun Lee, and Sangwhan Moon. 2025. DocHop-QA: A Multi-Document and Multi-Hop Benchmark for Document-Level Question Answering. arXiv:2508.15851 [cs.CL] https://arxiv.org/abs/2508.15851

[16] Wenhao Peng, Zechao Li, Shenhao Yao, Xu Han, and Zhiyuan Liu. 2025. UNIDOC-BENCH: A Unified Benchmark for Document-Centric Multimodal RAG. arXiv:2510.03663 [cs.CL] https://arxiv.org/abs/2510.03663

[17] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 3743–3751.

[18] Shraman Pramanick, Jatin Sharma, Sahana Murthy, Balaji V. Srinivasan, and Pranava Madhyastha. 2024. SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. arXiv:2407.09413 [cs.CL] https://arxiv.org/abs/2407.09413

[19] Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4634–4642.

[20] Abhinav Suri, Robert Hronevich, Yizhong Huang, Yuyan Pang, Jungsoo Kim, Douwe Kiela, Mohit Iyyer, and Shiyu Chang. 2025. VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal RAG. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*. https://aclanthology.org/2025.naacl-long.310/

[21] Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737* (2025).

[22] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[23] 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)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7534–7550. doi:10.18653/v1/2020.emnlp-main.609

[24] 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)*. 7534–7550.

[25] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. MinerU: An Open-Source Solution for Precise Document Content Extraction. arXiv:2409.18839 [cs.CV] https://arxiv.org/abs/2409.18839

[26] Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. 2022. What have we learned from OpenReview? *World Wide Web* 26, 2 (Nov. 2022), 683–708. doi:10.1007/s11280-022-01109-z

[27] Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. 2023. What have we learned from OpenReview? *World Wide Web* 26, 2 (2023), 683–708.

[28] Nancy XR Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. 317–326.

[29] Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2189–2198.

[30] Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. 2024. Benchmarking LLMs' Judgments with No Gold Standard. *arXiv preprint arXiv:2411.07127* (2024).

[31] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1192–1200.

[32] Zecheng Yu, Hongcai Wang, Yufei Liu, Huayi Lin, Haoxiang Wang, Zhenguo Lu, and Huanchen Zhang. 2025. Is Your Paper Being Reviewed by an LLM? A New Benchmark Dataset and Approach for Detecting AI Text in Peer Review. arXiv:2502.19614 [cs.CL] https://arxiv.org/abs/2502.19614

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

[34] Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025. Re^2: A Consistency-ensured Dataset for Full-stage Peer Review and Multi-turn Rebuttal Discussions. *arXiv preprint arXiv:2505.07920* (2025).

[35] Jiong Zhang, Minjoon Seo, Ming Zhou, Zeyi Yao, Shujian Zhang, Min Zhang, and Yue Zhang. 2025. LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-Context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*. https://aclanthology.org/2025.findings-acl.264/

[36] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*. IEEE, 1015–1022.

[37] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*. 9340–9351.

A Construction and Release Details

This appendix provides additional methodological details about construction, evaluation, and annotation for ADAM-Bench. We describe the process in a repository-agnostic way so that the benchmark is understandable to readers without assuming access to, or familiarity with, our codebase.

A.1 Evidence Object (EObj) Schema

Table 6: Evidence Object (EObj) schema used as the fixed evidence representation.

Field	Meaning
eobj_id	Stable identifier within a paper
type	paragraph, heading, equation, figure, table
section_path	Best-effort hierarchical context
page_no	Page index (0-based)
bbox_union_norm	Region box in normalized page coordinates (0–1000): (x_0, y_0, x_1, y_1)
text_concat	Canonical text view for recall and quoting
anchors	Anchor signals and caption references when available
media_path	Optional crop reference for multimodal settings

A.2 Normalization, Parsing, and Manifest Gates

OpenReview ingest and thread normalization. We ingest public OpenReview forums and construct (i) a paper-level manifest and (ii) per-paper discussion threads. The manifest records a stable paper/forum identifier, venue and year, a PDF link, licensing information, and coarse topic metadata. Each forum is normalized into an ordered utterance sequence with speaker roles, review stages, timestamps, and basic length statistics.

Role and stage mapping. Utterances are derived from OpenReview notes and mapped to coarse speaker roles (reviewer, author, area chair, program chair) and coarse stages (e.g., official review, meta-review, decision, comment) using invitation and note metadata. Greeting-only or extremely short fragments are filtered with a minimum-length threshold to avoid degenerate content.

Manifest gates. We apply mandatory manifest gates to ensure the released resource is legally and technically usable. The gates restrict the year scope, enforce licensing constraints for redistribution, and require that each paper contains substantive review content as well as an editorial decision or meta-review. We additionally require PDF reachability and successful parsing so that evidence objects are well-defined for evaluation.

PDF parsing and Evidence Objects (EObjs). Each paper PDF is parsed with a document parsing system (MinerU) to extract multimodal evidence objects. We represent all evidence as a fixed set of Evidence Objects (EObjs). Each EObj has a stable identifier within the paper, a modality/type (paragraph/heading/figure/table/equation), hierarchical context (section path), page index, a region box in normalized coordinates, a canonical text view for retrieval and quoting, optional visual crops when available, and anchor signals (e.g., caption references). We also build an anchor index that links explicit references (e.g., "Figure 3") to the corresponding EObjs within the same paper.

A.3 Candidate Triggers, Claim Atomization, and Coarse Recall

Candidate utterance triggers. We select candidate utterances using conservative rule-based triggers that target checkable content: anchors (Figure/Table/Equation/Section/Appendix), quantitative patterns (numbers, ranges, p-values), metric mentions, comparative language, ablation cues, and missing-experiment cues. A short dialogue context window can optionally be attached to each candidate utterance.

Claim atomization and deduplication. Each candidate utterance is atomized into one or more ClaimCards via a heuristic extractor or an LLM extractor, producing (i) the original claim span in the utterance and (ii) a normalized claim text used for modeling and evaluation. We cap claim length, and we deduplicate near-duplicates with a similarity threshold to avoid inflated counts from repeated phrasing. A lightweight rule augmenter can optionally add high-certainty experiment-result and missing-experiment claims.

Coarse recall and anchor boosting. For each ClaimCard, we retrieve a per-claim Top-K candidate list from the same paper's EObj store. The default retriever uses sparse matching (BM25 and TF-IDF) over the canonical text view and fuses ranks by reciprocal-rank fusion; dense retrieval is optionally supported. When a claim mentions explicit anchors, we boost anchor-resolved EObjs to the top of the fused list. Each candidate entry includes the EObj identifier plus its type, section context, page index, canonical text, and optional visual crop reference (when available).

A.4 Core Refinement and Locator Configuration

Core refinement. Core is obtained by applying stricter inclusion gates to Scale (e.g., higher-quality papers and stable evidence construction) and by running a fixed locator on the released candidate interface. Core therefore separates (i) corpus-scale weak alignment (Scale) from (ii) stable, diagnostic pre-alignment signals used for evaluation and error analysis.

Locator prompt and output constraints. Given a claim and its candidates, the locator outputs exactly one structured prediction containing a 4-way grounding label and one or more evidence sets over EObj identifiers. The interface enforces: (i) evidence must be selected from the provided candidates, (ii) at most three alternative evidence sets, (iii) a preference for minimal evidence, and (iv) optional short quotes that must be exact substrings of candidate text for auditability. For multimodal variants, we optionally attach a capped number of candidate crops with a byte-size limit.

Deterministic quote-to-span alignment. When quotes are present, we deterministically align them back to the canonical text view and propagate the corresponding page/region metadata for auditable inspection. This localization metadata is auxiliary for auditing and UI rendering; the official evidence key remains the EObj identifier.

A.5 Gold Curation, Annotation UI, and Quality Control

Packet export and annotation UI. We export annotation packets by joining ClaimCards, retrieval candidates, and optional locator suggestions. Annotators verify the 4-way grounding label and select one or more minimal evidence sets in a lightweight annotation UI that renders candidate text and optional evidence crops for efficient auditing.

Label guidelines. SUPPORTED/CONTRADICTED require explicit paper evidence that directly supports or refutes the claim under the paper-as-truth setting; NOT_FOUND is used when the needed information is absent from the PDF; UNDECIDABLE is reserved for claims that remain ambiguous or underspecified even after inspecting the PDF (e.g., missing quantifiers/conditions, or subjective judgments).

Quality control and evidence minimization. We apply a QC stage that enforces evaluator-facing invariants: supported/contradicted predictions must include evidence, evidence sets are deduplicated and bounded in size, and an optional evidence-minimization pass removes obviously irrelevant evidence items while preserving at least one cited EObj.

Paper-level splits. Gold splits are constructed at the paper level and stratified by year and research area, preventing leakage between train/dev/test through shared papers and supporting representative evaluation.

A.6 Release Artifacts, Licensing, and Privacy

Released artifacts (what we distribute). We release versioned manifests, threads, evidence objects, ClaimCards, candidates, and the evaluator/leaderboard toolkit. Concretely, the release contains:

- **Manifests:** paper identifiers, metadata, PDF links, and licensing information.
- **Threads:** normalized dialogue utterances with coarse speaker roles and review stages.
- **Evidence:** the fixed EObj representation (with anchor signals) used throughout recall, localization, annotation, and evaluation.
- **ClaimCards and candidates:** atomic claims plus provided-evidence candidate lists for each claim.
- **Gold:** evaluation-grade labels and minimal evidence set-of-sets for Gold splits.

Licensing strategy. Paper PDFs and extracted evidence may be subject to heterogeneous licenses. We therefore (i) preserve licensing information from OpenReview metadata in the manifest, (ii) support filtering to redistribution-safe subsets via an allowlist of licenses, and (iii) provide a manifest-only option where users download PDFs from their original URLs and reconstruct evidence objects locally when redistribution is not permitted.

Privacy and identifiers. Although OpenReview forums are public, we minimize privacy risks by (i) exposing only coarse roles/stages in the released dialogue by default, (ii) omitting personal identifiers where possible (e.g., signatures/emails), and (iii) storing annotator identifiers only as short hashes when human verification is involved. We recommend downstream users avoid drawing conclusions about individuals and treat the resource as a document-grounded benchmark rather than a person-centric dataset.

B Abbreviated Claim Record Schema (ClaimCard)

Each released instance corresponds to an atomic claim (a ClaimCard) paired with evidence candidates from the same paper. Conceptually, a ClaimCard record contains:

- **Identifiers:** a claim identifier and the associated paper identifier.
- **Claim text:** the normalized claim text, optionally with the original span from the source utterance.
- **Dialogue context (optional):** a short window of surrounding utterances, each with a coarse speaker role, stage, and text.
- **Candidate evidence:** a Top-K list of evidence objects from the same paper, each with an EObj identifier, type, section context, page index, canonical text, and optional crop reference.

C Reproducibility and Leaderboard Protocol

This appendix specifies the official evaluator, default settings, and prompt templates used in our released toolkit and leaderboard.

C.1 Official Evaluator and Submission Interface

Prediction file. Submissions are JSON Lines (one JSON object per claim). Each object must contain a claim identifier, a predicted grounding label, and one or more evidence sets. Evidence is keyed by Evidence Object identifiers; optional localization metadata (spans or boxes) is ignored by the official scorer. An abbreviated schema example is included in Appendix B.

Evaluator. The released toolkit includes an official evaluator that reports Macro-F1 for Task A, Evidence-F1 for Task B, and a FEVER-style end-to-end score that requires both correct labels and sufficient evidence when a claim is supported or contradicted. Formal metric definitions are provided in Appendix G.

C.2 Default K and Multimodal Caps

Released candidate pool (provided-evidence). For each claim, the release provides a fixed Top-K candidate list of EObjs from the *same paper*, produced by sparse recall and rank fusion. We fix $K = 15$ for the main benchmark tables. Candidates are ranked by reciprocal-rank fusion over BM25 and TF-IDF ranks, with optional anchor boosting that moves anchor-resolved EObjs to the top of the fused list.

LLM input caps (baseline + pre-alignment). To control cost and ensure stable comparison, our released locator runner caps the LLM input to the top $M = 12$ candidates with at most 400 characters per candidate text. For multimodal variants, we attach at most 6 candidate images (unique crops), skip files larger than 800,000 bytes, and use low-detail settings by default. All baseline runs use deterministic decoding (temperature = 0) and a strict JSON-only output constraint.

C.3 Prompt Templates

We use prompt templates that (i) extract atomic, paper-checkable claims from a review utterance and (ii) localize minimal evidence for a given claim over a fixed candidate interface. For both stages, we enforce strict structured outputs to make downstream processing deterministic and auditable.

Claim extraction. The extractor is instructed to output a JSON array of verbatim spans that are exact substrings of the utterance. It is explicitly told to skip subjective opinions, speculation, and vague summaries, and to prefer claims that contain concrete anchors, numbers, or metric statements.

Evidence localization. The locator is instructed to output a single JSON object containing a grounding label and one or more minimal evidence sets, where evidence must be selected only from the provided candidates. It is constrained to output a small number of evidence sets/items and may optionally output short quotes that must exactly match candidate text.

D Full Evaluated Model Suite

This appendix lists the full set of evaluated systems referenced throughout the paper (34 text-only LLMs and 10 multimodal LMMs). This enables space-efficient reporting in Table 7 while keeping the experimental scope explicit.

E Annotation Protocol and Quality Assurance

This appendix documents the human verification protocol used to curate Gold and clarifies how *minimal evidence* is operationalized for set-of-sets supervision. The current release focuses on *auditable* annotation with a lightweight UI, and we additionally outline a multi-annotator extension plan (IAA + adjudication) for future versions.

E.1 Annotation Unit and Stored Records

Unit. The annotation unit is a *ClaimCard*: an atomic claim extracted from a single dialogue utterance, optionally accompanied by a short dialogue context window. Annotators inspect the *paper PDF* as the sole truth source and select evidence in the fixed EObj space.

Inputs served by the UI. Each annotation packet joins (i) the claim text, (ii) context utterances, (iii) a per-claim Top-K candidate list, and (iv) optional model suggestions from the locator. Candidates expose an evidence-object identifier, type, section context, page index, canonical text, and optional crops for figure/table/equation objects.

Stored annotation record. The UI stores one JSON object per annotated claim containing at least: a claim identifier, the assigned grounding label, one or more minimal evidence sets, an optional hashed annotator identifier, and free-form adjudication notes. Evidence is stored as a set-of-sets: multiple alternative minimal evidence sets can be provided, and each set is represented as a list of evidence-object identifiers. Optional localization metadata (text spans or boxes) may be included for auditability but is not required for scoring.

E.2 Label Decision Rules and Edge Cases

We annotate grounding labels under a *paper-as-truth* setting:

- **Supported** (*SUPPORTED*): the claim is directly supported by explicit statements, results, or unambiguous evidence in the PDF.
- **Contradicted** (*CONTRADICTED*): the PDF contains explicit evidence that negates the claim (e.g., opposite trend, different number, or incompatible condition).
- **Not found** (*NOT_FOUND*): the required information is absent from the PDF; the claim may be true in general but is not paper-grounded.
- **Undecidable** (*UNDECIDABLE*): the claim is inherently ambiguous or underspecified given the PDF alone (e.g., missing scope/quantifiers, subjective judgments, or multiple plausible interpretations).

Practical disambiguation: NOT_FOUND vs. UNDECIDABLE. If the claim is *well-formed* but the paper does not provide the necessary evidence, use *NOT_FOUND*. If the claim itself lacks crucial conditions or uses subjective language such that no amount of reading resolves it, use *UNDECIDABLE*.

Out-of-paper references. Statements that can only be verified using external sources (e.g., "SOTA on X" without specifying the benchmark split, or claims about other papers) are treated as *NOT_FOUND* unless the paper explicitly provides the needed comparison context.

E.3 Minimal Evidence: Set-of-Sets Annotation Guidance

Sufficiency. An evidence set must be sufficient to justify the assigned label under paper-as-truth. For supported/contradicted cases, the evidence should make the claim verifiable or falsifiable *without* relying on annotator priors.

Minimality. Among sufficient sets, annotators aim to select a *minimal* set: removing any evidence item should break sufficiency (or make the justification materially weaker/ambiguous). To keep evidence comparable and auditable, we bound each evidence set to a small size (default max 4 EObjs in QC).

Alternative rationales (set-of-sets). If multiple *distinct* minimal rationales exist (e.g., a claim is stated both in a paragraph and in a table caption), annotators may provide multiple evidence sets; the evaluator matches predictions against the best gold alternative (Appendix G).

Multi-modal evidence. For figure/table/equation claims, annotators should prefer citing the PDF-native object(s) directly (figure/table/equation EObjs) rather than only nearby narrative text, unless the narrative text alone is sufficient and strictly minimal.

Quality Control and Evidence Minimization

Gold annotations are post-processed by QC to enforce evaluator-facing invariants: (i) supported/contradicted claims must have non-empty evidence, (ii) evidence sets are deduplicated and size-bounded, and (iii) all cited evidence objects must exist in the released EObj store. When enabled, an automatic minimization pass removes obviously irrelevant evidence items using a conservative lexical-overlap filter while preserving at least one cited EObj.

Table 7: Overall performance on ADAM-Bench. Provided-evidence metrics are reported in percent. Avg is the mean of Provided-evidence Macro-F1 and Evidence-F1. Wins counts pairwise wins aggregated over the seven Provided-evidence metrics against all other model baselines (strictly greater; ties not counted), excluding diagnostic rows. Best results within each block are bolded and second best are underlined. Superscript * marks the best open-source model by Avg in each block. Closed-book columns report the claim-only setting (no provided candidate evidence).

Systems	Closed-book		Provided-evidence							Summary	
	Macro-F1	FEVER	Macro-F1	Evidence-F1	FEVER	F1(S)	F1(C)	F1(NF)	F1(U)	Avg	Wins
Lower bounds and diagnostics											
R-ONLY (lower bound)	–	–	16.5	25.2	14.9	66.1	0.0	0.0	0.0	–	–
Oracle-in-candidates (upper bound)	–	–	100.0	75.9	81.8	100.0	100.0	100.0	100.0	–	–
Text models (LLMs)											
GPT-5.2	13.4	**36.6**	**45.6**	55.1	**51.0**	67.6	30.3	68.4	16.3	**50.4**	**198**
Grok-4.1	13.4	**36.6**	44.2	54.1	**51.0**	66.0	24.2	**68.8**	**17.9**	49.2	187
Claude-Opus-4-6	24.1	32.8	42.8	**57.2**	49.3	69.9	28.6	66.9	5.7	50.0	186
Qwen3-Max	16.0	36.4	43.9	54.0	45.2	73.1	38.3	64.4	0.0	49.0	181
Gemini-3-Flash-Preview	**31.8**	10.7	44.1	53.4	46.6	**73.4**	37.2	65.7	0.0	48.8	179
Qwen-Plus	22.5	36.2	42.3	56.5	45.5	73.8	28.6	66.9	0.0	49.4	178
Qwen3-235B-A22B-Instruct-2507*	16.1	36.4	44.3	54.0	40.5	72.3	44.0	60.8	0.0	49.2	174
Qwen-Max	20.0	33.5	39.8	55.9	38.3	72.3	28.6	58.5	0.0	47.9	167
Qwen2.5-32B-Instruct	13.7	**36.6**	39.6	53.0	43.0	71.0	19.4	62.3	5.9	46.3	160
DeepSeek-V3.2	13.7	35.3	41.3	49.5	42.1	70.3	31.6	63.2	0.0	45.4	146
Qwen-Turbo	18.0	35.8	35.4	51.9	44.4	72.1	0.0	59.0	10.5	43.7	143
Qwen3-32B	14.3	36.4	35.3	53.2	40.2	70.6	13.8	56.7	0.0	44.2	132
DeepSeek-R1-Distill-Qwen-14B	19.1	30.6	42.6	45.6	36.1	70.0	**42.1**	52.5	5.7	44.1	130
Qwen3-235B-A22B	20.5	33.0	42.4	45.9	36.6	69.5	38.3	56.5	5.4	44.2	127
Qwen3-30B-A3B-Thinking-2507	18.5	32.0	38.0	46.8	38.0	65.7	25.9	60.3	0.0	42.4	116
Qwen1.5-110B-Chat	21.0	31.5	41.7	46.7	35.0	72.8	32.0	51.1	10.8	44.2	115
Qwen2.5-72B-Instruct	14.9	35.0	40.7	49.4	37.2	72.0	36.8	49.5	4.4	45.1	112
Qwen2-57B-A14B-Instruct	15.0	26.0	32.9	45.7	30.6	70.1	8.3	37.6	15.6	39.3	104
Qwen2-7B-Instruct	16.5	27.0	35.1	40.3	32.0	70.9	20.0	49.5	0.0	37.7	92
Qwen1.5-72B-Chat	18.8	28.0	36.0	43.4	32.5	71.0	12.9	50.4	9.8	39.7	91
Qwen2.5-14B-Instruct	13.4	**36.6**	36.4	38.0	27.0	72.2	14.8	58.4	0.0	37.2	83
Qwen3-30B-A3B-Instruct-2507	16.0	28.5	32.6	46.0	33.6	70.4	14.3	45.7	0.0	39.3	82
Qwen3-8B	20.1	35.3	33.9	43.2	28.9	71.0	27.6	37.0	0.0	38.6	78
Qwen2.5-7B-Instruct	26.0	17.4	36.7	34.5	15.2	68.0	27.7	47.2	3.7	35.6	77
Qwen3-235B-A22B-Instruct-2507*	16.1	36.4	44.3	54.0	40.5	72.3	44.0	60.8	0.0	49.2	174
Qwen1.5-32B-Chat	15.2	19.0	24.1	36.4	20.4	68.4	0.0	27.9	0.0	30.2	63
Qwen2.5-14B-Instruct-1M	17.0	22.5	37.3	33.9	22.0	**73.7**	16.0	59.4	0.0	35.6	60
Qwen-Flash	16.2	24.5	30.5	43.7	31.1	68.1	12.5	41.5	0.0	37.1	58
DeepSeek-R1-Distill-Qwen-1.5B	15.0	19.5	27.2	31.7	22.6	44.8	10.0	43.6	10.3	29.5	55
Qwen2.5-7B-Instruct-1M	15.5	16.0	33.0	33.2	12.9	68.1	27.9	35.8	0.0	33.1	53
Qwen1.5-14B-Chat	14.8	18.0	23.6	28.8	21.2	68.0	0.0	20.1	6.2	26.2	52
Qwen2-1.5B-Instruct	13.0	20.0	22.7	33.3	26.4	30.4	9.6	50.7	0.0	28.0	47
Qwen2.5-3B-Instruct	12.0	24.0	19.7	36.1	37.5	17.2	0.0	55.4	6.2	27.9	44
Qwen1.5-7B-Chat	14.0	14.5	21.9	27.8	15.2	66.7	6.2	8.4	6.2	24.8	36
Qwen2.5-0.5B-Instruct	25.2	27.5	19.7	30.6	27.8	32.4	0.0	46.3	0.0	25.2	22
Multimodal models (LMMs)											
Qwen3-VL-Plus	22.0	34.4	**47.4**	54.8	43.8	**73.2**	34.3	62.6	**19.5**	**51.1**	**48**
Qwen3-VL-Plus 235B	13.7	34.4	**47.4**	54.8	43.8	**73.2**	34.3	62.6	**19.5**	**51.1**	**48**
Qwen3-VL-Max	16.4	**36.4**	45.8	54.6	**46.6**	71.4	**45.0**	**66.7**	0.0	50.2	44
Qwen2.5-VL-32B-Instruct*	13.7	**36.4**	40.5	**55.7**	41.6	71.8	24.0	66.2	0.0	48.1	35
Qwen3-VL-Flash	19.0	30.5	40.5	**55.7**	41.6	71.8	24.0	66.2	0.0	48.1	35
GLM-4.6V 106B	**26.0**	30.6	37.3	53.2	38.0	58.1	14.3	66.2	10.5	45.2	28
Gemini-3-Flash	14.5	18.0	23.5	39.0	40.5	21.4	16.7	55.9	0.0	31.2	14
ChatGPT-4o	21.7	31.1	27.5	38.7	24.8	69.7	0.0	32.0	8.3	33.1	11
Qwen-VL-Plus	15.5	22.0	27.5	38.7	24.8	69.7	0.0	32.0	8.3	33.1	11
Gemini-2.5-Flash	22.7	26.4	29.9	38.1	22.6	68.6	0.0	50.9	0.0	34.0	7

Table 8: EObj type distribution in the Core snapshot.

Type	paragraph	heading	equation	figure	table
Count	582,485	92,782	72,566	28,246	23,210

Table 9: Claim type distribution in Core (post-augmentation).

Type	method_desc	experiment_result	missing_exp
Count	24,193	84,547	5,580

E.4 Multi-Annotator Extension and IAA (Planned)

The current annotation app provides a single-user workflow for efficient curation. For stronger reliability guarantees, we plan a multi-annotator protocol as follows:

- **Independent double annotation**: sample each claim to two annotators with identical packets (same candidate ordering) and store per-annotator JSONL outputs separately.
- **IAA for labels**: report Cohen's κ (pairwise) and Krippendorff's α (multi-annotator) over 4-way labels, together with confusion patterns (especially Not found vs. Undecidable).
- **IAA for evidence**: treat each annotator's evidence as a set-of-sets and compute best-match Evidence-F1 between annotators; additionally report agreement on evidence *types* (paragraph/table/figure/...) and evidence-set size.
- **Adjudication**: route disagreements to a senior adjudicator who selects the final label and minimal evidence set(s) and records an adjudication note; the released Gold stores only adjudicated outputs.

F Additional Statistics and Distributions

This appendix reports dataset statistics used for sanity checking, ablation design, and reproducibility. Unless stated otherwise, the numbers correspond to the released Core snapshot and the Gold-dev evaluation split, and they can be recomputed from the release artifacts and accompanying summary reports.

F.1 Core Snapshot Composition

Core contains 3,000 papers, 66,317 dialogue utterances (22,167,250 tokens), 114,320 ClaimCards, and 799,289 EObjs (Table 2). EObj type counts in the current snapshot are summarized in Table 8.

F.2 Claim Type Distribution (Core)

Table 9 reports the rule/LLM-derived claim type distribution after augmentation, computed from the released Core snapshot.

In the same Core snapshot, the mean (median) claim length is 24.25 (22.0) tokens and 148.7 (137.0) characters; 53,297 claims mention explicit anchors (Figure/Table/Section/...), 85,859 contain numbers, and 107,212 contain at least one detected entity.

Table 10: Candidate evidence type distribution (Core provided-evidence candidates).

Type	paragraph	figure	heading	table	equation
Count	5,615,447	422,116	331,278	190,097	71,890

F.3 Candidate Pool Statistics (Core)

The released provided-evidence candidates contain 6,630,828 claim–candidate pairs (mean 58 candidates per claim; median 59; $p95 = 60$) with the type mix shown in Table 10.

Candidate texts average 637 characters (median 530), corresponding to about 101 tokens on average (median 83); in this snapshot, 531,652 unique EObj IDs appear in the candidate pool and image-crop availability coverage is 1.0.

F.4 Gold-dev Label and Evidence Statistics

Gold-dev contains 363 claims. Its label distribution is: 179 SUPPORTED, 20 CONTRADICTED, 133 NOT_FOUND, and 31 UNDECIDABLE. Across Gold-dev, the mean number of evidence sets per claim is 0.669 (median 1), and the mean evidence items per evidence set is 1.568 (median 1). Evidence types cited in Gold-dev are: 308 paragraphs, 46 tables, 22 figures, 4 headings, and 1 equation.

G Metric Definitions

This appendix formalizes the official metrics used in Section 4.1.

Macro-F1 and per-label F1 (Task A). Let $\mathcal{L}$ be the Task A label set and N be the number of instances. For each label $\ell \in \mathcal{L}$,

$$\mathrm{P}_\ell = \frac{\mathrm{TP}_\ell}{\mathrm{TP}_\ell + \mathrm{FP}_\ell}, \quad \mathrm{R}_\ell = \frac{\mathrm{TP}_\ell}{\mathrm{TP}_\ell + \mathrm{FN}_\ell}, \quad \mathrm{F1}_\ell = \frac{2\mathrm{P}_\ell \mathrm{R}_\ell}{\mathrm{P}_\ell + \mathrm{R}_\ell}, \tag{1}$$

where ratios with zero denominators are defined as 0. We report both per-label $\mathrm{F1}_\ell$ (F1(S)/F1(C)/F1(NF)/F1(U)) and their macro average:

$$\text{Macro-F1} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathrm{F1}_\ell. \tag{2}$$

Evidence-F1 (Task B). For instance i, let $\mathcal{G}_i = \{G_i^{(1)}, \ldots, G_i^{(m_i)}\}$ be the gold set-of-sets of minimal evidence and $\mathcal{P}_i$ be the predicted set-of-sets, where each G or P is a set of evidence object identifiers. If $\mathcal{G}_i$ is empty, we set

$$\mathrm{EvF1}_i = \mathbf{1}\{\mathcal{P}_i = \varnothing\}. \tag{3}$$

Otherwise, we score by best matching over alternative gold sets and predicted sets (treating an empty prediction as a single empty set):

$$\mathrm{EvF1}_i = \max_{G \in \mathcal{G}_i} \max_{P \in (\mathcal{P}_i \cup \{\varnothing\})} \mathrm{F1}(P, G), \tag{4}$$

where

$$\mathrm{F1}(P, G) = \begin{cases} 1, & P = \varnothing \ \wedge \ G = \varnothing, \\ \frac{2|P \cap G|}{|P| + |G|}, & \text{otherwise.} \end{cases} \tag{5}$$

The reported Evidence-F1 is the mean over instances, $\frac{1}{N} \sum_{i=1}^{N} \mathrm{EvF1}_i$.

FEVER-style score. We adapt the end-to-end FEVER score [22]. For instance i, let y_i and $\hat{y}_i$ denote gold and predicted labels. Define

$$\text{FEVER}_i = \begin{cases} 0, & \hat{y}_i \neq y_i, \\ 1, & y_i \in \mathcal{Y}_{\text{neu}}, \\ \mathbf{1}\Big\{\exists G \in \mathcal{G}_i,\ \exists P \in (\mathcal{P}_i \cup \{\varnothing\}) : G \subseteq P\Big\}, & y_i \in \mathcal{Y}_{\text{ver}}. \end{cases} \tag{6}$$

where $\mathcal{Y}_{\text{neu}}$ denotes neutral labels (NOT_FOUND, UNDECIDABLE), and $\mathcal{Y}_{\text{ver}}$ denotes verifiable labels (SUPPORTED, CONTRADICTED). The reported FEVER-style score is the mean over instances, $\frac{1}{N}\sum_{i=1}^{N} \text{FEVER}_i$.