MMSciFact: A Benchmark for Multimodal Dependency-Aware Scientific Fact-Checking

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) are increasingly used to answer questions about scientific papers with long-form responses grounded in figures, tables, and text. Yet factuality evaluation typically decomposes these answers into atomic claims and verifies them independently. This independence assumption overlooks how observations serve as premises for downstream interpretations, forming a dependency graph within each answer. An interpretation can be supported in isolation yet depend on a premise contradicted by or unverifiable from the source—a failure we call Invalid Support. We introduce **MMSciFact**, a benchmark for dependency-aware multimodal scientific fact-checking comprising 257 question–answer pairs and 2,737 sentences from 45 papers. Model answers are preserved without post-hoc corruption and annotated with expert grounding judgments, sentence roles, and dependencies. Across 12 MLLM fact-checkers and single-page, cross-page, and full-paper evidence, models struggle to identify statements contradicted by or unverifiable from the source, as well as interpretations based on invalid premises. Providing sentence roles and dependency graphs improves the detection of Invalid Support, but scores remain low even with this information, indicating that dependency structure alone is insufficient for reliable verification. These findings suggest that dependency awareness complements claim-level verification in supporting reliable scientific fact-checking.

Code & Data: https://anonymous.4open.science/r/MMSciFact

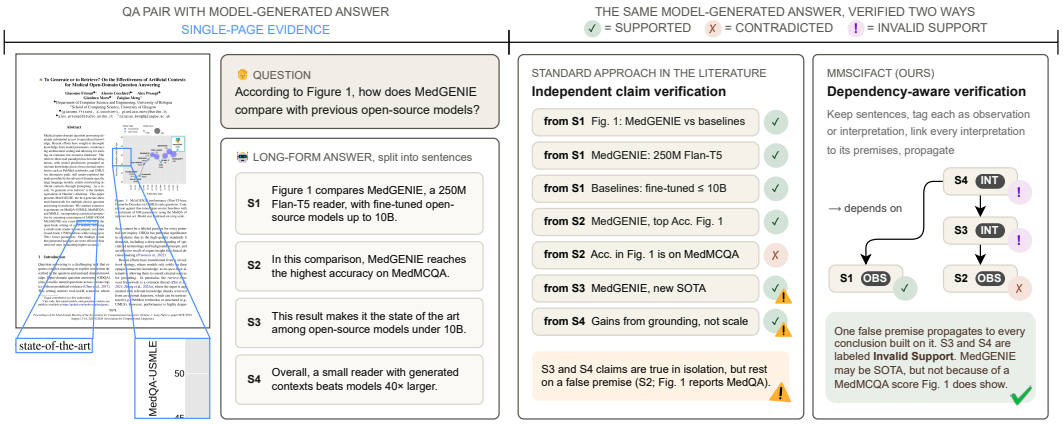


Figure 1: **Independent vs. dependency-aware verification of a long-form scientific answer.** Locally supported conclusions based on a contradicted observation are labeled Supported by independent verification (*left*) but Invalid Support by dependency-aware verification (*right*).

1 Introduction

Multimodal large language models (MLLMs) are increasingly used as an interface for understanding scientific papers, answering questions that require summarizing findings, interpreting figures, and synthesizing evidence across pages (Li et al., 2024; Pramanick et al., 2024; Jung et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This growing use is also reflected in widely available research tools, such as NotebookLM, which let users interact directly with uploaded papers. As generated answers increasingly mediate how readers interpret scientific results, errors can distort not only individual facts but also the conclusions drawn from them. Trusting these answers requires assessing whether they are grounded in the source, not simply factually correct in the open world (Zhao et al., 2024; Jacovi et al., 2025). The dominant paradigm for evaluating the factuality of generated long-form answers, established by FActScore (Min et al., 2023) and adopted by subsequent work (Wei et al., 2024; Marinescu et al., 2025), follows a *break-then-verify* procedure that decomposes an answer into atomic claims and verifies each claim in isolation against the source. Yet, long-form answers are not flat collections of independent claims. They comprise *interpretations* that draw conclusions from *observations* of what the document shows or from other interpretations, forming chains of reasoning within an answer.

Figure 1 illustrates this limitation of verifying each claim in isolation. The long-form answer attributes MedGENIE's highest accuracy to MedMCQA, although the source figure reports MedQA. Independent claim verification marks this observation (S2) as CONTRADICTED but the two downstream interpretations (S3 and S4) as SUPPORTED, because they are supported by the source when evaluated in isolation. It therefore fails to account for the fact that the contradicted observation (S2) is used as a premise for those interpretations. This is critical in science, where readers rely not only on reported findings but also on the evidence and reasoning that support them.

To make this failure measurable, we introduce **MMSCIFACT**, a benchmark for dependency-aware fact-checking of long-form answers grounded in multimodal scientific papers drawn primarily from computer science.[1] It comprises 257 question–answer pairs and 2,737 answer sentences from 45 papers. Each instance pairs a paper and a question with an MLLM-generated answer, preserving naturally occurring grounding failures without post-hoc error injection. A human-validated protocol first assigns sentence roles (observation or interpretation) and identifies their dependencies within each long-form answer, forming a directed acyclic graph linking interpretations to their premises. Expert human annotators then assign grounding labels (SUPPORTED, CONTRADICTED, or NOT ENOUGH INFORMATION (NEI)) to sentences containing a claim. An interpretation is labeled as INVALID SUPPORT when it is supported in isolation but depends, directly or transitively, on a premise sentence labeled as CONTRADICTED or NEI. These annotations reveal INVALID SUPPORT in 148 sentences (5.4% of all sentences), affecting 23% of answers in our benchmark.

MMSCIFACT spans three evidence scopes (*single-page*, *cross-page*, and *full-paper*), reflecting complementary scientific reading settings, from interpreting individual figures and tables to integrating information from multiple pages and synthesizing findings from an entire paper. In each setting, judges must assess sentence-level grounding and localize the supporting PDF regions using bounding boxes. Restricting evidence to the source paper isolates verification failures from external retrieval errors. This remains important in agentic scientific fact-checking, where finding relevant evidence does not ensure that a judge can interpret it or track dependencies among claims.

We evaluate 12 MLLMs under three verification settings that differ in how judges account for dependencies between answer sentences: *independent*, where models ground sentences without modeling dependencies; *holistic*, where models must infer dependencies and account for them during grounding; and *oracle-graph*, where the annotated sentence roles and dependency graph are provided.

Our results first reveal broader weaknesses that arise independently of INVALID SUPPORT. Across three evidence scopes and three verification settings, current MLLMs remain weak scientific fact-checkers: they strongly over-predict SUPPORTED, struggle particularly to recognize NEI, and show limited evidence localization capability. Reasoning on sentence dependencies introduces an additional challenge. In the *holistic* setting, where models must recover the dependency structure themselves, INVALID SUPPORT detection remains poor. In contrast, providing the annotated dependency graph substantially improves its detection while leaving the other grounding labels largely unchanged. However, grounding performance remains limited even when the dependencies are provided. This indicates that recovering and reasoning over sentence dependencies constitutes a distinct bottleneck beyond already weak evidence-grounding capabilities.

Our contributions are threefold: (i) a dependency-aware formulation of scientific claim grounding with a new INVALID SUPPORT label; (ii) the MMSCIFACT benchmark, featuring expert human

[1] Although the phenomenon studied here is domain-general, we focus on scientific question answering to reflect the growing real-world use of MLLMs to understand scientific papers.

Table 1: **Comparison with prior scientific claim-verification benchmarks.** All benchmarks include textual evidence; icons denote additional tables, figures, and charts. *Long context*: complete pages or papers rather than isolated excerpts/figures/tables. *Non-injected errors*: unmodified model outputs. *Cross-sentence dependencies*: annotated links between answer sentences.

	Setting				Annotation	
Benchmark	**Long answer**	**Long context**	**Non-text evidence**	**Non-injected errors**	**Cross-sentence dependencies**	**Grounding labels**
SciFact (Wadden et al., 2020)	✗	✗	✗	✗	✗	S C N
SEM-TAB-FACTS (Wang et al., 2021)	✗	✗	▦	✗	✗	S C N
SciFact-Open (Wadden et al., 2022)	✗	✗	✗	✗	✗	S C N
SCITAB (Lu et al., 2023)	✗	✗	▦	✗	✗	S C N
MuSciClaims (Lal et al., 2025)	✗	✗	🖼	✗	✗	S C N
ClimateViz (Su et al., 2025)	✗	✗	📊	✗	✗	S C N
SciVer (Wang et al., 2025a)	✗	✗	▦ 📊	✗	✗	S C
MMSci-Eval (TFV) (Yang et al., 2025)	✗	✗	▦	✗	✗	S C N
SciAtomicBench (Zhang et al., 2025b)	✗	✗	▦	✗	✗	S C
SciClaimEval (Ho et al., 2026)	✗	✗	▦ 🖼	✗	✗	S C
SciConBench (Jung et al., 2026)	✓	✓	✗	✓	✗	S C N
MMSciFact (ours)	✓	✓	▦ 🖼 📊	✓	✓	S C N !

S UPPORTED C ONTRADICTED N EI ! NVALID SUPPORT

grounding judgments and validated annotations of sentence roles and dependencies; and (iii) a diagnostic evaluation of 12 MLLMs as fact-checkers of long-form scientific answers, separating failures in sentence-level grounding, dependency recovery, and reasoning over known dependencies.

2 RELATED WORK

Scientific claim verification. SciFact (Wadden et al., 2020) and SciFact-Open (Wadden et al., 2022) evaluate the grounding of individual scientific claims against textual evidence (Table 1). Subsequent benchmarks broaden the evidence to scientific tables (Wang et al., 2021; Lu et al., 2023; Yang et al., 2025; Zhang et al., 2025b), figures and charts (Lal et al., 2025; Su et al., 2025), or mixed multimodal contexts (Wang et al., 2025a; Ho et al., 2026). ToolSciVer (Zhou et al., 2026) uses reinforcement learning to train scientific claim verifiers to acquire evidence through table, chart, and region-inspection tools. Despite this broader evidence, verification remains claim-local. Deng et al. (2025) highlight the challenges of evidence retrieval from full-length, multimodal scientific papers. SciConBench (Jung et al., 2026) considers long-form conclusions synthesized from open-web evidence, but verifies their constituent facts independently. MMSCIFACT instead evaluates the grounding of long-form scientific answers while accounting for cross-sentence dependencies.

Factuality beyond atomic facts. Atomic decomposition can introduce noise, remove context, and alter factuality scores (Tang et al., 2024; Hu et al., 2025; Liu et al., 2025; Wanner et al., 2025). Discourse-aware approaches incorporate discourse relations into factual consistency evaluation (Zhang et al., 2025a; Zheng et al., 2025). Graph-based approaches capture structure at other levels. FactReasoner links atomic claims to retrieved contexts (Marinescu et al., 2025), FactCorrector uses this graph as structured feedback for post-hoc correction (Carnerero-Cano et al., 2026), and GraphFC represents entities and relations within individual claims (Huang et al., 2025). MMSCIFACT instead models claim-to-claim dependencies as a directed graph across answer sentences, distinguishing observations from interpretations, and capturing direct and transitive relationships.

Multimodal long-form evaluation. The FACTS Leaderboard (Cheng et al., 2025) includes separate tracks for long-form document grounding (FACTS Grounding (Jacovi et al., 2025)) and image-conditioned factuality (FACTS Multimodal). Related benchmarks evaluate attribution and faithfulness in long-form outputs grounded in documents, images, video, or audio (Jang et al., 2026; Song et al., 2026; Wan et al., 2026). MMCV (Wang et al., 2025b) evaluates individual claims through multi-hop reasoning over textual, visual, and tabular evidence from multiple sources. In science, SciLens (Wang et al., 2026) decomposes claims into atoms grounded in figures and tables, while GAMUT (Chen et al., 2026) represents relations among facts to evaluate answer completeness. Our

focus is whether conclusions in generated answers depend on premises that are grounded in the source, complementing prior work on evidence integration, attribution, and answer completeness.

3 MMSciFact

MMSciFact pairs multimodal scientific documents with model-generated long-form answers, annotated for sentence roles, cross-sentence dependencies, and grounding. We first formalize dependency-aware grounding in Section 3.1, then describe the initial benchmark construction process in Section 3.2, and finally report the annotation and quality-control protocol in Section 3.3.

3.1 Dependency-aware Grounding

Long-form scientific answers contain statements that play different roles in the argument. An *observation* reports information that can be checked directly against the source, such as a value in a table, a trend in a figure, or a methodological detail. An *interpretation* instead draws a conclusion or explanation from such information, using other sentences in the answer as justification. Consider the example in Figure 1. S2—"In this comparison, MedGENIE reaches the highest accuracy on MedMCQA"—is an observation because it reports a claim that can be checked directly against Figure 1, whereas S3—"This result makes it the state of the art among open-source models under 10B"—is an interpretation because it uses the result stated in S2 to support a broader conclusion.

Interpretations thus introduce dependencies between sentences. For each sentence i, let $\mathrm{dep}(i)$ denote the set of other answer sentences used as its immediate premises. In Figure 1, $\mathrm{dep}(\mathrm{S3}) = \{\mathrm{S2}\}$ because the conclusion in S3 is justified by the result reported in S2. Yet dependencies may also pass through other interpretations: S4 builds on S3, so an error in S2 can affect S4 transitively. We thus use $\mathrm{dep}^*(i)$ to denote the full set of direct and transitive premises of i. Premises may occur before or after the sentence they support, and sentences with no cross-sentence premise have $\mathrm{dep}(i) = \varnothing$. Together, these dependencies form a directed acyclic graph over the answer.

For sentence grounding, we first assess each sentence *locally*, evaluating it against the source document independently of the other answer sentences and without propagating errors across sentences. Let r_i denote its role and g_i its grounding label. For compactness, let SUP, CON, NEI, and INV denote Supported, Contradicted, NEI, and Invalid Support, respectively. The grounding label satisfies $g_i \in \{\mathrm{SUP}, \mathrm{CON}, \mathrm{NEI}\}$. Dependency-aware grounding additionally asks whether the premises used to justify an interpretation are themselves grounded. We define:

$$\mathrm{invalid}(i) \iff r_i = \text{Interpretation} \wedge g_i = \mathrm{SUP} \wedge \exists j \in \mathrm{dep}^*(i) : g_j \in \{\mathrm{CON}, \mathrm{NEI}\}.$$

That is, an interpretation receives INV when it is locally supported, but at least one of its direct or transitive premises is contradicted by the source (CON) or unverifiable from it (NEI). The final grounding label y_i is thus set to INV whenever $\mathrm{invalid}(i)$ holds; otherwise, it retains the grounding label g_i. In Figure 1, S3 is locally supported but depends on the contradicted S2 and is therefore labeled as Invalid Support; the same error propagates transitively to S4. Crucially, dependencies encode actual justificatory relations rather than topical similarity: an erroneous sentence affects an interpretation only when it lies on that interpretation's dependency path.

3.2 QA Pair Construction

Figure 2 summarizes the full MMSciFact construction pipeline. We now describe Stage 1, in which we construct question–answer (QA) pairs from scientific PDFs through source-context selection and agentic QA generation. The subsequent stages (structural pre-annotation, human verification, and filtering) are described in Section 3.3. Appendix B indexes all the prompts used.

Source paper selection. Twelve expert human annotators from our team, all senior researchers or Ph.D. students proficient in English and active in Machine Learning or NLP, selected the benchmark source papers. Each annotator nominated up to six papers within their area of expertise, collectively providing coverage across multiple venues and research subfields. This domain familiarity helped annotators interpret technical terminology, experimental conditions, and connections between textual and visual results when verifying claims against explicit source evidence.

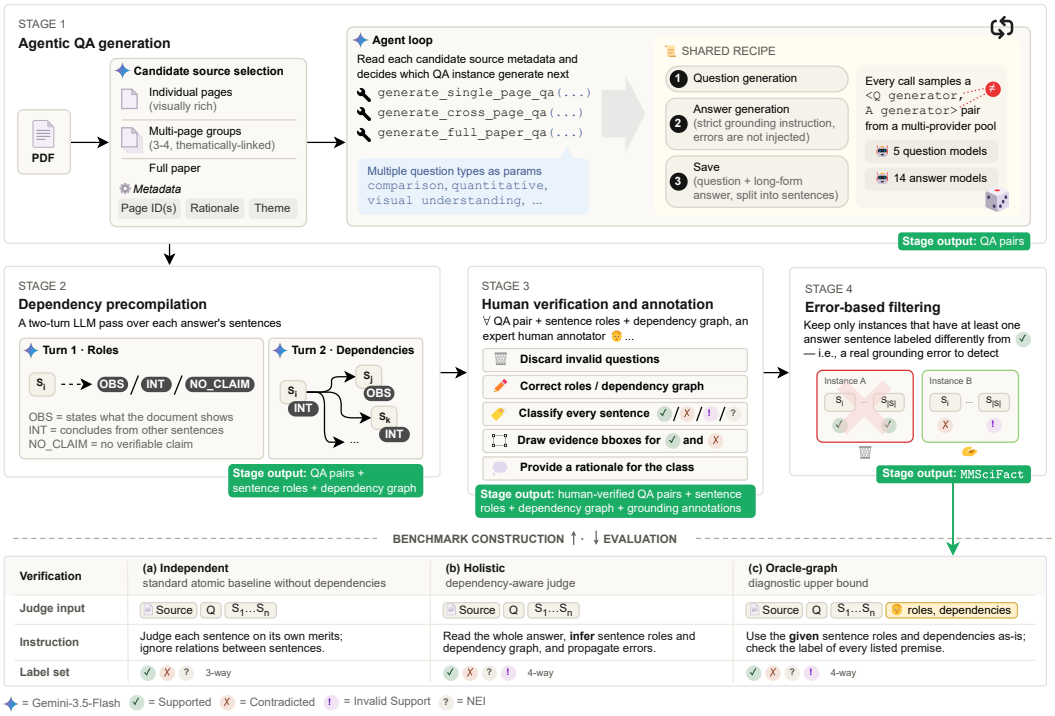


Figure 2: **MMSciFact construction and evaluation pipeline.** *Top:* Agentic QA generation is followed by sentence dependency pre-annotation, human verification and grounding annotation, and error-based filtering. *Bottom:* Evaluation compares independent, holistic, and oracle-graph verification, with sentence dependencies respectively ignored, inferred, or provided.

Evidence scope selection. Once the source papers are established, we use Gemini-3.5-Flash to extract candidate source contexts directly from each PDF, as preliminary experiments demonstrated it offers an optimal balance of efficiency and selection quality. We consider three levels of context granularity to capture different scientific reading settings: *single-page* candidates focus on one visually rich page; *cross-page* candidates comprise 3–4 thematically related pages, with at most two being consecutive; and *full-paper* candidates encompass the entire document. For each candidate, we store its source page ID(s) together with a short rationale and thematic description. These candidate contexts and their metadata are then provided to the agentic QA-generation process.

Agentic QA generation. Because scientific papers distribute background, methods, results, and conclusions across different parts of the document (Teufel & Moens, 2002; Liakata, 2010), we design questions to reflect different scientific reading requirements. *Single-page* questions stress-test native visual reasoning beyond caption reading or single-cell table lookups, while *cross-page* questions follow the synthesis categories of Li et al. (2024) and are checked to ensure they cannot be answered from any single page alone. *Full-paper* questions use the same categories but require aggregating evidence from non-adjacent sections, often spanning the main text and appendices. Both multi-page scopes always integrate at least one visual component with evidence from textual sections. Figure 3 (top) summarizes the five question categories per scope, with examples in Appendix Table 9. Gemini-3.5-Flash orchestrates generation of 15 QA pairs per paper, with five per evidence scope and balanced coverage across question categories. For each pair, we sample a question generator from five MLLMs and a different answer generator from 14 open- and closed-weight MLLMs (Appendix C.2). The selected context is passed to the question generator with a scope- and category-specific prompt (Appendix B), and the resulting question is answered using the same context under a strict instruction to rely only on the source. Unlike approaches that inject or corrupt errors (Table 1), we neither instruct answer generators to make errors nor modify their outputs, preserving naturally occurring grounding failures. Gemini-3.5-Flash then splits each answer into sentences using scientific-text rules (Appendix C.3), yielding units for role, dependency, and grounding annotation.

3.3 ANNOTATION PROTOCOL

After QA-pair construction, each instance undergoes dependency-graph annotation, grounding annotation, and QA-pair filtering, corresponding to Stages 2–4 in Figure 2. Our protocol distinguishes two complementary annotation tasks. *Dependency-graph annotation* assigns each answer sentence a role and, for interpretations, identifies the premises on which it depends, producing the reference dependency graph defined in Section 3.1. *Grounding annotation* instead evaluates each answer sentence against the source document and assigns its grounding label. Both tasks are supported by a dedicated annotation interface described in Appendix G.

Annotator training and effort. Before formal annotation, annotators completed 2–3 one-hour training sessions covering the annotation guidelines and practice examples, with feedback from the guideline authors (Appendix F). Annotating all QA pairs associated with a source paper required approximately three person-hours (PHs) on average, corresponding to about 135 PHs across the 45 papers. Training contributed about 30 additional PHs, for an estimated total of 165 PHs.

Dependency-graph annotation. Motivated by prior work on reducing annotation burden with LLM assistance (Wei et al., 2024), we pre-annotate each long-form answer using two independent LLM runs. Each run first assigns every sentence one of three roles (*observation*, *interpretation*, or *not a claim*), and then identifies the premises of each interpretation. An annotation is accepted automatically only when the two runs agree exactly on both its role and complete dependency set, otherwise, it is resolved by a human annotator (Appendix G, Figures 15–16). We validate three pre-annotation strategies on a private set of 40 QA pairs and 440 sentences (see Section 4): two Gemini-3.5-Flash runs (Gemini×2), two GPT-5.4-mini runs (GPT×2), and one run from each model (Gemini+GPT). LLM–human agreement is measured only on cases where the two LLM runs agree, using Cohen's κ for sentence roles and F1 over directed dependency edges. As shown in Table 2, Gemini×2 achieves the highest agreement on both roles ($\kappa = 0.770$) and dependencies (F1 $= 0.840$), and is therefore used to pre-annotate the full corpus.

Table 2: **Structural annotation agreement.**

Comparison	Role κ	Dep. F1	Auto-res
A vs. B	0.810	0.878	–
C vs. D	0.792	0.818	–
Gemini×2 vs.	0.770	0.840	86.0%
Gemini+GPT vs.	0.710	0.780	82.0%
GPT×2 vs.	0.685	0.760	81.0%

Grounding annotation. For each remaining QA pair, trained human annotators evaluate every answer sentence against the source document and assign its grounding label according to Section 3.1. Grounding labels are assigned directly by humans, without LLM pre-annotation. Annotators additionally provide a grounding rationale and, for SUPPORTED and CONTRADICTED observations, mark the corresponding source evidence with a reference bounding box. For every INVALID SUPPORT case, they also verify the dependency path linking the interpretation to its invalid premise. At least two annotators independently evaluate all QA pairs from each source paper, reaching a mean pairwise Cohen's κ of 0.839. When the two annotators disagree, they discuss and jointly agree the final label and evidence boxes.

QA-pair filtering. We apply two filtering steps, summarized in Figure 2. Before grounding annotation, trained human annotators discard QA pairs whose questions are unrealistic, ambiguous, or not answerable within the assigned evidence scope. After grounding annotation, we retain only QA pairs containing at least one grounding error, removing fully supported answers to obtain an error-enriched diagnostic benchmark.

3.4 CORPUS STATISTICS

The final test benchmark contains 257 QA pairs from 45 papers spanning 18 venues (e.g., ICLR, ACL, NeurIPS) and 12 different topics, as shown in Appendix Figure 4, with approximately six QA pairs per paper on average after filtering. The 257 answers comprise 2,737 sentences, averaging 10.6 sentences per answer, with similar numbers of observations and interpretations. They are divided into 91 single-page, 89 cross-page, and 77 full-paper instances. Figure 3 summarizes sentence roles, grounding labels, input lengths, and question-category distributions across evidence scopes. SUPPORTED is the most frequent grounding label in every scope, while the proportion of NEI sentences increases from 4% in single-page to 10% in full-paper instances. Input length also increases sub-

Single-page		Cross-page / Full-paper	
Comparison	Compare visual values or patterns	*Conceptual*	Connect concepts across the paper
Trend	Track changes across conditions	*Methodological*	Relate methods to performance
Quantitative	Assess values against references or extrema	*Results interpretation*	Synthesize findings across experiments
Visual understanding	Interpret visual structures or processes	*Implication*	Infer consequences or limitations
Adversarial	Check misleading visual interpretations	*Adversarial*	Check misleading cross-page interpretations

Metric	N
Source papers	45
QA pairs	257
single-page	91
cross-page	89
full-paper	77
Sentences	2737
Observation	1358
Interpretation	1297
Not a claim	82
Grounding labels	
SUPPORTED	1445 (52.8%)
CONTRADICTED	866 (31.6%)
NEI	196 (7.2%)
INVALID SUPPORT	148 (5.4%)
NOT A CLAIM	82 (3.0%)
Input tokens (avg/max)	
single-page	5,856 / 6,588
cross-page	9,823 / 12,034
full-paper	49,192 / 97,735

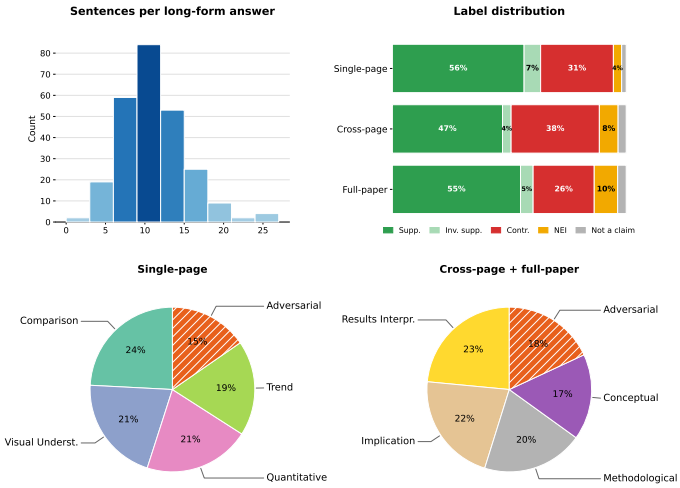


Figure 3: **MMSCIFACT question design, scale, and composition.** *Top*: question categories and reasoning requirements by evidence scope. *Bottom left*: corpus statistics and input lengths. *Bottom right*: distributions of answer length, grounding labels, and question types across evidence scopes.

stantially, averaging about 49K tokens and reaching approximately 98K for the longest full-paper instance. Source-paper coverage across venues and topics is detailed in Appendix C.1.

4 EXPERIMENTAL SETUP

Validation set. In addition to our test benchmark set comprising 257 QA pairs and 2,737 sentences, we construct and annotate a separate private validation set of 40 QA pairs and 440 sentences, with source papers disjoint from the test set. We use this set to (i) validate the role and dependency pre-annotation procedure in Section 3.3, and (ii) compare prompt configurations for each judge and verification setting, selecting the prompt with the highest macro-F1, averaged across the three evidence scopes. Complete validation results are reported in Appendix D.2.

Judge models. We evaluate 12 MLLMs, including proprietary models such as GPT-5.6-Luna with high reasoning effort and Gemini-3.6-Flash with medium thinking effort, together with ten locally hosted open-weight models. These include Gemma 4 (E4B, 26B-A4B, and 31B), Qwen 3.5/3.6/3.8 (9B, 122B-A10B, 397B-A17B, 35B-A3B, and 27B), and GLM-4.6V (Flash-9B and 106B-A12B). Deployment and inference details are provided in Appendix D.1.

Verification settings. All judges receive the source context and the long-form answer and are evaluated under three settings (see Figure 2, bottom). Under *independent* verification, judges assign a grounding label to each sentence independently, without considering dependencies between sentences, following FACTS Grounding (Jacovi et al., 2025). Under *holistic* verification, judges must infer the dependencies between sentences and use them when assigning grounding labels. Under *oracle-graph* verification, judges are provided with the annotated sentence roles and dependencies.

Evidence scopes. We evaluate each verification setting on *single-page*, *cross-page*, and *full-paper* instances. Single- and cross-page inputs are provided as 144 DPI page images following Ma et al. (2024). For full-paper inputs, open-weight models receive all rendered pages, while proprietary models receive the original PDF through their supported document APIs (see Appendix D.1).

Metrics. We report sentence-level per-class F1 and macro-F1. *Holistic* and *oracle-graph* verification distinguish four labels: SUPPORTED, CONTRADICTED, NEI, and INVALID SUPPORT. Because

Table 3: **Grounding performance by judge and evidence scope.** Each model–condition pair uses the prompting strategy selected on held-out validation data. The final column macro-averages per-class F1 across evidence scopes. Within each condition, the best and second-best scores are shown in bold and underlined, respectively. 🔒 denotes proprietary models. Checkpoint names report total parameters and, for mixture-of-experts models, active parameters after A.

	Single-page				Cross-page				Full-paper				
Model	SUP	CON	NEI	INV	SUP	CON	NEI	INV	SUP	CON	NEI	INV	**F1**
Independent													
Qwen3.5-9B	0.773	0.508	0.140	–	0.723	0.452	0.227	–	0.739	0.312	0.064	–	0.444
Qwen3.8-27B	0.838	0.602	0.323	–	**0.778**	0.522	0.269	–	**0.835**	0.484	0.313	–	0.551
Qwen3.6-35B-A3B	0.787	0.554	0.267	–	0.731	0.514	0.168	–	0.766	0.395	0.060	–	0.467
Qwen3.5-122B-A10B	0.796	0.565	0.375	–	0.719	0.485	0.196	–	0.748	0.300	0.200	–	0.485
Qwen3.5-397B-A17B	0.797	0.582	0.311	–	0.737	0.521	0.262	–	0.763	0.408	0.113	–	0.497
Gemma4-E4B	0.711	0.270	0.208	–	0.659	0.181	0.206	–	0.730	0.143	0.241	–	0.375
Gemma4-26B-A4B	0.792	0.571	**0.421**	–	0.732	0.461	**0.297**	–	0.749	0.353	0.131	–	0.493
Gemma4-31B	0.842	0.540	0.327	–	0.768	0.480	0.221	–	0.806	0.353	0.270	–	0.511
GLM-4.6V-Flash (9B)	0.742	0.385	0.175	–	0.704	0.353	0.212	–	0.740	0.179	0.097	–	0.406
GLM-4.6V (106B-A12B)	0.810	0.287	0.250	–	0.711	0.198	0.131	–	0.775	0.088	0.083	–	0.370
🔒 GPT-5.6-Luna (high)	0.846	**0.609**	0.286	–	0.769	**0.559**	0.246	–	0.817	**0.503**	**0.375**	–	**0.560**
🔒 Gemini-3.6-Flash (medium)	**0.849**	0.579	0.400	–	0.757	0.450	0.160	–	0.818	0.436	0.218	–	0.510
Holistic													
Qwen3.5-9B	0.778	0.431	0.111	0.173	0.706	0.324	0.094	0.075	0.739	0.195	0.000	0.054	0.309
Qwen3.8-27B	**0.822**	**0.613**	0.296	**0.299**	**0.769**	**0.569**	**0.279**	0.078	**0.802**	0.480	**0.378**	0.198	**0.471**
Qwen3.6-35B-A3B	0.793	0.543	0.200	0.000	0.716	0.460	0.089	0.038	0.766	0.380	0.062	0.131	0.345
Qwen3.5-122B-A10B	0.791	0.549	0.318	0.212	0.733	0.480	0.162	0.094	0.755	0.310	0.041	0.029	0.374
Qwen3.5-397B-A17B	0.817	0.569	0.244	0.173	0.734	0.497	0.222	0.000	0.761	0.329	0.115	0.159	0.391
Gemma4-E4B	0.740	0.213	0.111	0.051	0.674	0.148	0.183	**0.172**	0.726	0.073	0.092	0.032	0.270
Gemma4-26B-A4B	0.786	0.505	0.333	0.130	0.735	0.409	0.137	0.126	0.761	0.409	0.137	0.167	0.379
Gemma4-31B	0.786	0.484	**0.458**	0.112	0.736	0.412	0.150	0.056	0.769	0.342	0.214	0.154	0.380
GLM-4.6V-Flash (9B)	0.741	0.332	0.151	0.000	0.697	0.320	0.078	0.000	0.738	0.124	0.075	0.000	0.275
GLM-4.6V (106B-A12B)	0.754	0.201	0.182	0.250	0.669	0.186	0.118	0.037	0.715	0.087	0.061	0.000	0.280
🔒 GPT-5.6-Luna (high)	0.805	0.593	0.286	0.205	0.763	0.549	0.210	0.110	0.796	**0.501**	0.317	**0.245**	0.449
🔒 Gemini-3.6-Flash (medium)	0.810	0.511	0.227	0.256	0.761	0.429	0.000	0.085	0.791	0.421	0.128	0.222	0.384
Oracle-graph													
Qwen3.5-9B	0.791	0.429	0.054	0.356	0.775	0.367	0.066	0.097	0.762	0.233	0.061	0.163	0.348
Qwen3.8-27B	0.847	0.574	0.212	**0.456**	**0.816**	0.509	0.274	0.128	**0.827**	0.429	0.186	0.289	0.468
Qwen3.6-35B-A3B	0.831	0.505	0.273	0.329	0.754	0.427	0.124	0.124	0.774	0.308	0.021	0.235	0.386
Qwen3.5-122B-A10B	0.829	0.506	0.162	0.357	0.794	0.423	0.101	**0.181**	0.785	0.343	0.117	0.249	0.403
Qwen3.5-397B-A17B	**0.849**	**0.586**	0.108	0.415	0.805	0.516	0.151	0.169	0.782	0.348	0.081	0.188	0.423
Gemma4-E4B	0.734	0.199	0.237	0.245	0.662	0.161	0.159	0.128	0.727	0.078	0.063	0.193	0.297
Gemma4-26B-A4B	0.808	0.434	0.346	0.292	0.764	0.362	0.158	0.146	0.764	0.314	0.119	0.147	0.382
Gemma4-31B	0.818	0.473	0.311	0.309	0.759	0.415	0.084	0.080	0.777	0.331	0.114	0.177	0.383
GLM-4.6V-Flash (9B)	0.747	0.255	0.204	0.143	0.711	0.204	0.070	0.034	0.736	0.126	0.042	0.145	0.280
GLM-4.6V (106B-A12B)	0.780	0.272	0.182	0.213	0.682	0.105	0.125	0.141	0.734	0.057	0.020	0.000	0.281
🔒 GPT-5.6-Luna (high)	0.823	0.578	0.226	0.405	0.776	**0.534**	**0.300**	0.169	0.798	**0.480**	**0.279**	0.219	**0.471**
🔒 Gemini-3.6-Flash (medium)	0.830	0.467	**0.400**	0.280	0.786	0.443	0.024	0.142	0.806	0.456	0.059	**0.298**	0.405

independent verification does not consider dependencies, the INVALID SUPPORT labels are mapped to their local label (SUPPORTED) and evaluation is performed over the remaining three labels. We thus do not directly compare macro-F1 between independent and dependency-aware verification. INVALID SUPPORT F1 captures dependency-aware grounding performance, while verdict changes between *holistic* and *oracle-graph* quantify the benefit of annotated dependencies. For evidence localization, we compare predicted bounding boxes for correctly classified SUPPORTED and CONTRADICTED observations with the reference boxes using intersection over union (IoU). We also report box precision, the fraction of the predicted box that lies within the reference evidence region.

5 RESULTS

Our results are reported in Table 3. Overall, current MLLMs are weak scientific fact-checkers, with a strong bias toward SUPPORTED, particular difficulty in recognizing NEI, and limited evidence-localization capabilities. Dependency-aware verification adds a further challenge: INVALID SUPPORT is difficult to detect when models must infer dependencies themselves, while providing the annotated dependency graph substantially improves its detection with limited changes to the other labels. Yet performance remains low even with the reference graph, indicating that both source

grounding and dependency reasoning remain important bottlenecks. Appendix E provides further analyses of completion length versus macro-F1, performance by question category and evidence type, and judges' possible advantage on answers generated by models from their own family.

Scientific grounding remains challenging across judges. Table 3 reports per-class F1 for each evidence scope, with the final column giving macro-F1 by averaging the per-class F1 scores across the three evidence scopes. Across the 12 judges, the mean macro-F1 is 0.472 under *independent* verification, with GPT-5.6-Luna achieving the highest score at 0.560. Under *holistic* and *oracle-graph* verification, the corresponding means across judges are 0.359 and 0.377, respectively. Qwen3.8-27B achieves the highest macro-F1 under *holistic* verification, while GPT-5.6-Luna achieves the highest under *oracle-graph*, both at 0.471. The higher scores under *independent* verification are expected because this setting evaluates a simpler task: judges classify each sentence without reasoning over dependencies, and INVALID SUPPORT is mapped to SUPPORTED, reducing the label set from four classes to three. For this reason, macro-F1 under *independent* verification is not directly comparable to the dependency-aware settings, but performance remains modest across all three settings.

Grounding failures are frequently classified as SUPPORTED. Across all 12 judges, errors frequently collapse to SUPPORTED (see Appendix Figure 7): depending on the verification setting, 52–62% of CONTRADICTED and 68–74% of NEI sentences are predicted as SUPPORTED on average. This bias is more pronounced for INVALID SUPPORT: under *holistic* verification, 82% of these sentences are predicted as SUPPORTED on average, compared with 63% under *oracle-graph* verification. This difficulty is not limited to dependency errors. Indeed, among the three labels shared across all settings, NEI has the lowest mean F1 in every setting, including *independent* verification.

Reference dependencies improve INVALID SUPPORT detection. Across evidence scopes, *oracle-graph* verification improves INVALID SUPPORT F1 for all 12 judges on average, with substantial gains for most models (Table 3). A paired bootstrap finds statistically significant improvements for 9 of the 12 judges (Appendix E.8). The gains are concentrated on INVALID SUPPORT, while performance on the other labels changes only modestly and not consistently in the same direction. For some judges, stronger INVALID SUPPORT detection is accompanied by more CONTRADICTED and NEI sentences being misclassified as INVALID SUPPORT (Appendix Figure 7). For example, for Qwen3.8-27B, the substantial gain in INVALID SUPPORT F1 coincides with a slightly lower overall macro-F1 under *oracle-graph* than under *holistic* verification. Overall, providing the dependency graph substantially helps identify conclusions built on invalid premises, but performance remains low, leaving considerable room for improvement in using known sentence dependencies reliably.

Performance varies across evidence scopes. Across judges, CONTRADICTED and NEI F1 generally decrease as the evidence scope expands from single-page to cross-page and full-paper instances (Table 3). INVALID SUPPORT shows a different pattern, with cross-page instances often yielding the lowest scores, particularly under *oracle-graph* verification. These differences should not be attributed to context length alone, since the three evidence scopes also differ in question types, generation criteria, and label distributions.

Correct verdicts often lack precise evidence localization. Even for correctly classified SUPPORTED and CONTRADICTED observations, predicted evidence bounding boxes often align poorly with the reference evidence. Under *oracle-graph* verification, mean IoU ranges from only 0.054 to 0.244 across judges (Appendix Table 7). Because the exact extent of human-drawn reference boxes is subjective (Loison et al., 2026), IoU can penalize a tight prediction that still identifies the correct evidence (Appendix Figure 8). We therefore report box precision, the fraction of the predicted box that lies inside the reference region, which is higher but remains modest at 0.17–0.53 across judges.

6 CONCLUSION

We introduced MMSCIFACT, a benchmark for fact-checking long-form scientific answers. Using sentence roles, dependency graphs, and the INVALID SUPPORT label, we distinguish local support from the validity of supporting premises. Across 12 MLLMs, reference dependencies improve INVALID SUPPORT detection, but performance remains low, while grounding failures are often misclassified as SUPPORTED. Taken together, these results show that reliable scientific fact-checking requires both grounding claims in multimodal evidence and reasoning over the dependencies that connect them. Limitations are discussed in Appendix A.

AI USE STATEMENT

Generative AI tools supported evidence selection, question and answer generation, sentence segmentation, and structural pre-annotation (Sections 3.2 and 3.3). The evaluated MLLM judges produced experimental predictions (Section 4), from which deterministic scoring scripts computed all reported metrics. Benchmark content underwent the human validation and adjudication procedures described in Section 3.3. Generative AI also assisted with scientific writing to improve clarity and fluency. The authors take responsibility for the final text, results, and artifacts.

ETHICS STATEMENT

Human annotation was conducted by the research team, with training and annotation effort documented in Section 3.3. Source documents come from the public scientific literature, thus we discuss possible pretraining contamination in Appendix A. Answers were generated without requesting or injecting errors, and their wording was preserved during annotation. We deliberately retain erroneous answers for diagnostic evaluation. Hence, these outputs should not be treated as reliable scientific guidance. We identify no additional harmful applications specific to our benchmark.

REPRODUCIBILITY STATEMENT

Complete prompts for benchmark construction, structural pre-annotation, and judge evaluation are available in the accompanying repository, indexed in Appendix B. Section 3.3 documents annotator training, annotation procedures, and agreement, while Appendix G illustrates the annotation interface. Section 4 specifies evaluation conditions, evidence scopes, judge models, and validation-based prompt selection. Serving details and prompt-selection results appear in Appendices D.1 and D.2, respectively. Table 3 reports grounding performance by judge, condition, and evidence scope. The benchmark and evaluation code will be publicly released upon publication.

REFERENCES

Javier Carnerero-Cano, Massimiliano Pronesti, Radu Marinescu, Tigran T. Tchrakian, James Barry, Jasmina Gajcin, Yufang Hou, Alessandra Pascale, and Elizabeth M. Daly. FactCorrector: A graph-inspired approach to long-form factuality correction of large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 46278–46307. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.2147. URL https://doi.org/10.18653/v1/2026.acl-long.2147.

Xilun Chen, Zhaleh Feizollahi, Ross Goodwin, Seungwhan Moon, Scott Yih, Pinar Donmez, Babak Damavandi, and Luna Dong. Two-level meta-rubrics for evaluating open-ended generation: GAMUT, a benchmark for factual completeness. *CoRR*, abs/2607.19322, 2026. doi: 10.48550/ARXIV.2607.19322. URL https://doi.org/10.48550/arXiv.2607.19322.

Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, Adam Bloniarz, Aijun Bai, Andrew Wang, Anfal Siddiqui, Arturo Bajuelos Castillo, Aviel Atias, Chang Liu, Corey Fry, Daniel Balle, Deepanway Ghosal, Doron Kukliansky, Dror Marcus, Elena Gribovskaya, Eran Ofek, Honglei Zhuang, Itay Laish, Jan Ackermann, Lily Wang, Meg Risdal, Megan Barnes, Michael Fink, Mohamed Amin, Moran Ambar, Natan Potikha, Nikita Gupta, Nitzan Katz, Noam Velan, Ofir Roval, Ori Ram, Polina Zablotskaia, Prathamesh Bang, Priyanka Agrawal, Rakesh Ghiya, Sanjay Ganapathy, Simon Baumgartner, Sofia Erell, Sushant Prakash, Thibault Sellam, Vikram Rao, Xuanhui Wang, Yaroslav Akulov, Yulong Yang, Zhen Yang, Zhixin Lai, Zhongru Wu, Anca Dragan, Avinatan Hassidim, Fernando Pereira, Slav Petrov, Srinivasan Venkatachary, Tulsee Doshi, Yossi Matias, Sasha Goldshtein, and Dipanjan Das. The FACTS leaderboard: A comprehensive benchmark for large language model factuality. *CoRR*, abs/2512.10791, 2025. doi: 10.48550/ARXIV.2512.10791. URL https://doi.org/10.48550/arXiv.2512.10791.

Xingyu Deng, Xi Wang, and Mark Stevenson. The next phase of scientific fact-checking: Advanced evidence retrieval from complex structured academic papers. In *Proceedings of the 2025 Interna-*

tional ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025, pp. 436–448. ACM, 2025. doi: 10.1145/3731120.3744614. URL https://doi.org/10.1145/3731120.3744614.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, André Greiner-Petter, and Akiko Aizawa. SciClaimEval: Cross-modal claim verification in scientific papers. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference, LREC 2026, Palma de Mallorca, Spain, May 13-15, 2026*, pp. 11060–11071. ELRA Language Resource Association / ACL, 2026. doi: 10.63317/4AP9RG2GNWMF. URL https://doi.org/10.63317/4ap9rg2gnwmf.

Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 6313–6336. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.320. URL https://doi.org/10.18653/v1/2025.naacl-long.320.

Yani Huang, Richong Zhang, Zhijie Nie, Junfan Chen, and Xuefeng Zhang. A graph-based verification framework for fact-checking. *CoRR*, abs/2503.07282, 2025. doi: 10.48550/ARXIV.2503.07282. URL https://doi.org/10.48550/arXiv.2503.07282.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *CoRR*, abs/2501.03200, 2025. doi: 10.48550/ARXIV.2501.03200. URL https://doi.org/10.48550/arXiv.2501.03200.

Youngrok Jang, Hyesoo Kong, Kyunghwan An, Jae Sub Huh, Gyeonghun Kim, and Stanley Jungkyu Choi. VinQA: Visual elements interleaved long-form answer generation for real-world multimodal document QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 41130–41139. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jang_VinQA_Visual_Elements_Interleaved_Long-form_Answer_Generation_for_Real-World_Multimodal_CVPR_2026_paper.html.

Hayoung Jung, Pedro Viana Diniz, José Reinaldo Corrêa Roveda, Abner Fernandes da Silva, Haeun Jung, Enoch Tsai, Aleksandra Korolova, and Manoel Horta Ribeiro. Can AI agents synthesize scientific conclusions? *CoRR*, abs/2606.11337, 2026. doi: 10.48550/ARXIV.2606.11337. URL https://doi.org/10.48550/arXiv.2606.11337.

Yash Kumar Lal, Manikanta Bandham, Mohammad Saqib Hasan, Apoorva Kashi, Mahnaz Koupaee, and Niranjan Balasubramanian. MuSciClaims: Multimodal scientific claim verification. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pp. 3285–3307. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.IJCNLP-LONG.175. URL https://doi.org/10.18653/v1/2025.ijcnlp-long.175.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. M3SciQA: A multi-modal multi-document scientific QA benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, November 2024. URL https://aclanthology.org/2024.findings-emnlp.904/.

Maria Liakata. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP@ACL 2010, Uppsala, Sweden, July 10, 2010*, pp. 1–4. University of Antwerp, 2010. URL https://aclanthology.org/W10-3101/.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. VeriFact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 17908–17925. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.905. URL https://doi.org/10.18653/v1/2025.emnlp-main.905.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel de Souza Pereira Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. Vidore V3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 16570–16600. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.755. URL https://doi.org/10.18653/v1/2026.acl-long.755.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 7787–7813. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.483. URL https://doi.org/10.18653/v1/2023.emnlp-main.483.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ae0e43289bffea0c1fa34633fc608e92-Abstract-Datasets_and_Benchmarks_Track.html.

Radu Marinescu, Debarun Bhattacharjya, Junkyu Lee, Tigran T. Tchrakian, Javier Carnerero-Cano, Yufang Hou, Elizabeth M. Daly, and Alessandra Pascale. FactReasoner: A probabilistic approach to long-form factuality assessment for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 14547–14577. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.785. URL https://doi.org/10.18653/v1/2025.findings-emnlp.785.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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 2023, Singapore, December 6-10, 2023*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d74033a247989e8f6f3bf9e0c9629fb5-Abstract-Datasets_and_Benchmarks_Track.html.

Seokwon Song, Minsu Park, and Gunhee Kim. MAVIS: A benchmark for multimodal source attribution in long-form visual question answering. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 33028–33037. AAAI Press, 2026. doi: 10.1609/AAAI.V40I39.40585. URL https://doi.org/10.1609/aaai.v40i39.40585.

Ruiran Su, Jiasheng Si, Zhijiang Guo, and Janet B. Pierrehumbert. ClimateViz: A benchmark for statistical reasoning and fact verification on scientific charts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 23436–23458. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1196. URL https://doi.org/10.18653/v1/2025.emnlp-main.1196.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 8818–8847. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.499. URL https://doi.org/10.18653/v1/2024.emnlp-main.499.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguistics*, 28(4):409–445, 2002. doi: 10.1162/089120102762671936. URL https://doi.org/10.1162/089120102762671936.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.609. URL https://doi.org/10.18653/v1/2020.emnlp-main.609.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, volume EMNLP 2022 of *Findings of ACL*, pp. 4719–4734. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.347. URL https://doi.org/10.18653/v1/2022.findings-emnlp.347.

David Wan, Han Wang, Ziyang Wang, Elias Stengel-Eskin, Hyunji Lee, and Mohit Bansal. Multimodal fact-level attribution for verifiable reasoning. *CoRR*, abs/2602.11509, 2026. doi: 10.48550/ARXIV.2602.11509. URL https://doi.org/10.48550/arXiv.2602.11509.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. SciVer: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 8562–8579. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.ACL-LONG.420. URL https://doi.org/10.18653/v1/2025.acl-long.420.

Haoran Wang, Aman Rangapur, Xiongxiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. Piecing it all together: Verifying multi-hop multimodal claims. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 7453–7469. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.coling-main.498/.

Nancy Xin Ru Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 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@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pp. 317–326. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.SEMEVAL-1.39. URL https://doi.org/10.18653/v1/2021.semeval-1.39.

Yueming Wang, Tianshi Zheng, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See. SciLens: Multi-modal scientific claim verification with agentic entailment and grounding. *CoRR*, abs/2606.20873, 2026. doi: 10.48550/ARXIV.2606.20873. URL https://doi.org/10.48550/arXiv.2606.20873.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. DnDScore: Decontextualization and decomposition for factuality verification in long-form text generation. In *Proceedings of the*

2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025, pp. 23609–23626. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1205. URL https://doi.org/10.18653/v1/2025.emnlp-main.1205.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Bohao Yang, Yingji Zhang, Dong Liu, André Freitas, and Chenghua Lin. Does table source matter? Benchmarking and improving multimodal scientific table understanding and reasoning. *CoRR*, abs/2501.13042, 2025. doi: 10.48550/ARXIV.2501.13042. URL https://doi.org/10.48550/arXiv.2501.13042.

Kun Zhang, Oana Balalau, and Ioana Manolescu. Structured discourse representation for factual consistency verification. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 820–838. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.FINDINGS-ACL.46. URL https://doi.org/10.18653/v1/2025.findings-acl.46.

Yuji Zhang, Qingyun Wang, Cheng Qian, Jiateng Liu, Chenkai Sun, Denghui Zhang, Tarek F. Abdelzaher, Chengxiang Zhai, Preslav Nakov, and Heng Ji. Atomic reasoning for scientific table claim verification. *CoRR*, abs/2506.06972, 2025b. doi: 10.48550/ARXIV.2506.06972. URL https://doi.org/10.48550/arXiv.2506.06972.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Raghavi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. WildHallucinations: Evaluating long-form factuality in LLMs with real-world entity queries. *CoRR*, abs/2407.17468, 2024. doi: 10.48550/ARXIV.2407.17468. URL https://doi.org/10.48550/arXiv.2407.17468.

Danna Zheng, Mirella Lapata, and Jeff Z. Pan. Long-form information alignment evaluation beyond atomic facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 11007–11027. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.558. URL https://doi.org/10.18653/v1/2025.emnlp-main.558.

Binglin Zhou, Peng Shi, Ryo Kamoi, Nan Zhang, and Rui Zhang. ToolSciVer: Multimodal scientific claim verification with visual tool augmented reinforcement learning. *CoRR*, abs/2607.16131, 2026. doi: 10.48550/ARXIV.2607.16131. URL https://doi.org/10.48550/arXiv.2607.16131.

A LIMITATIONS

MMSCIFACT is an error-enriched diagnostic benchmark. Its label distribution and model performance do not estimate natural error prevalence or deployment accuracy. Its 45 papers were nominated by 12 in-house annotators familiar with the papers, limiting source diversity and potentially introducing selection bias. Evaluation uses supplied pages or documents, covering evidence localization and verification but not external retrieval. Reference dependency graphs are validated silver annotations, and full-paper input interfaces differ across model families. Judge and answer-generator families also overlap. Appendix E.4 reports no consistent accuracy advantage on same-family outputs, but these comparisons do not isolate family effects from differences in instance difficulty. Finally, public source papers may have appeared in model pretraining data (Jacovi et al., 2025), but the questions, answers, dependency structures, and labels are newly created. Although prior source familiarity may affect document-conditioned verification, constructing new evaluation instances reduces the risk of instance-level contamination.

B PROMPT DOCUMENTATION

The complete prompts used for benchmark construction, structural pre-annotation, and judge evaluation are available in the accompanying repository: `https://anonymous.4open.science/r/MMSciFact`. Table 4 lists the repository path for each prompt.

Table 4: **Prompt inventory.** Paths are relative to `artifacts/prompts/`. Combine each group's directory with the listed filename. All files have the extension `.md`.

Construction

Base directory: `construction/`

Prompt	Filename
Evidence selection	`evidence_selection`
Answer generation	`answer_generation`
Sentence segmentation	`sentence_segmentation`
Single-page questions	`question_generation/single_page/`
Comparison	`comparison`
Trend	`trend`
Quantitative	`quantitative`
Visual understanding	`visual_understanding`
Adversarial	`adversarial`
Cross-page questions	`question_generation/cross_page/`
Conceptual	`conceptual`
Methodological	`methodological`
Results interpretation	`results_interpretation`
Implication	`implication`
Adversarial	`adversarial`
Full-paper questions	`question_generation/full_paper/`
Conceptual	`conceptual`
Methodological	`methodological`
Results interpretation	`results_interpretation`
Implication	`implication`
Adversarial	`adversarial`

Annotation and evaluation

Directories relative to the inventory root

Prompt	Filename
Structural annotation	`annotation/`
Role classification	`role_classification`
Dependency assignment	`dependency_assignment`
Independent verification	`evaluation/`
Independent	`independent`
Holistic verification	`evaluation/holistic/`
Baseline	`baseline`
Entailment	`entailment`
Self-graph	`self_graph`
Claim decomposition	`claim_decomposition`
Minimal	`minimal`
Oracle-graph verification	`evaluation/oracle_graph/`
Baseline	`baseline`
Entailment	`entailment`
Minimal	`minimal`
Role only	`role_only`
Terse graph notation	`terse_graph_notation`

C ADDITIONAL BENCHMARK DETAILS

C.1 SOURCE PAPER COVERAGE

Figure 4 shows the venue and topic distribution of the benchmark's 257 QA pairs, drawn from 45 source papers spanning 18 venues and 12 topics.

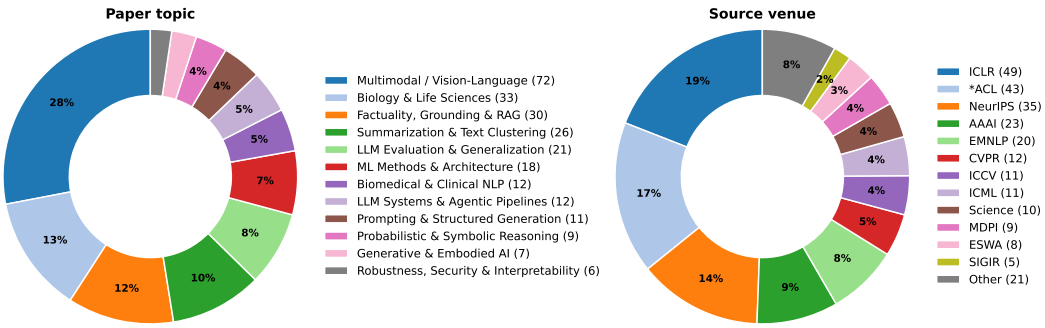


Figure 4: **Source coverage.** Topic (*left*) and venue (*right*) distribution of the benchmark's 257 QA pairs. Percentages are shares of QA pairs and legend values are QA-pair counts. *ACL also includes EACL and NAACL papers, and Other groups IJCAI, Neurocomputing, FAIA, and AACR journal.

C.2 Generator Model Provenance

Figure 5 shows the models used to generate the questions and answers in the annotated corpus. Questions were generated by a curated set of five MLLMs that produced the highest-quality questions in preliminary experiments, thereby reducing the number of ill-posed instances subsequently rejected by annotators. Answers were generated by a broader and more heterogeneous pool of 14 models spanning different families and parameter scales. This diversity allows the benchmark to capture naturally occurring grounding failures across multiple systems rather than single-model failures.

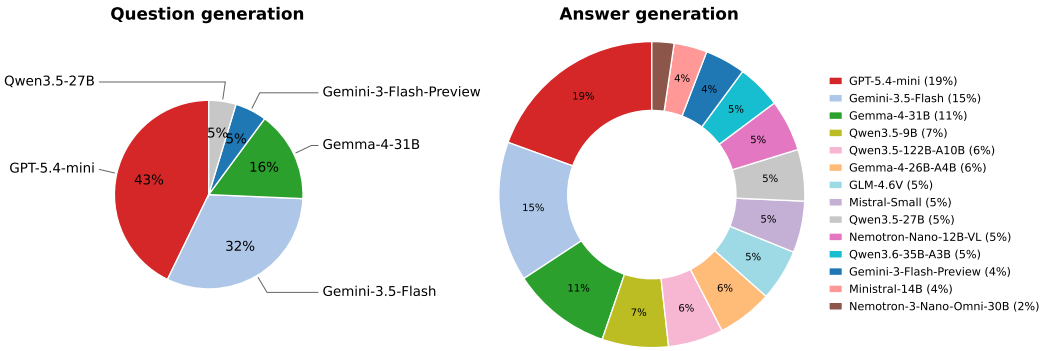


Figure 5: **Generator-model provenance.** Distribution of QA pairs by question-generation model (*left*; 5 models) and answer-generation model (*right*; 14 models).

C.3 Sentence Segmentation

Sentence boundaries define both the units annotated for grounding and the dependency graph nodes. We therefore segment each answer once before annotation and keep the resulting boundaries fixed.

Procedure. Gemini-3.5-Flash splits each answer into a JSON list of verbatim sentences using the prompt indexed in Appendix B. The prompt forbids rephrasing, merging, or dropping text and follows two conventions: semicolons define a boundary only when they separate two statements, while colons never define a boundary. For example, "The method improves accuracy; it also reduces computation." is split into two units, whereas "0.82; 0.79; and 0.71" remains a single sentence.

Verification. After segmentation, we deterministically check that the concatenated sentences, after whitespace normalization, exactly reproduce the original answer. Any altered, added, or missing text causes the split to be regenerated, for up to three attempts. Thus, the model only chooses sentence boundaries and cannot modify the answer content. We also tested standard sentence splitters, but found them less reliable in a small number of cases involving semicolons, abbreviations, and references. We therefore use an LLM, in line with the sentence-level decomposition used in FACTS

Grounding (Jacovi et al., 2025), because it allows us to encode these annotation-specific boundary rules, while the reconstruction check preserves the original answer text exactly.

D Experimental Details

D.1 Implementation Details

Proprietary judges are queried through their respective providers' batch APIs—OpenAI's Batch API[2] and Gemini's Batch API[3]. Open-weight judges are served locally with vLLM on an LSF-managed GPU cluster. Each checkpoint runs in a separate job on a single NVIDIA A100 SXM4 80GB GPU by default. Checkpoints exceeding single-GPU memory use tensor parallelism across up to eight A100 or H100 80GB GPUs, with H100 HBM3 nodes used when available.

D.2 Held-out Validation: Prompt-Variant Selection

As described in Section 4, prompt variants are selected on a held-out validation set before test-set evaluation. All variants receive the same question, answer sentences, and source evidence, and are evaluated against the same target grounding labels. They differ in instruction detail, verification granularity, intermediate representation, and whether answer dependencies are inferred or provided. Prompts are indexed in Appendix B.

Independent prompt. Independent verification follows the general formulation of FACTS Grounding (Jacovi et al., 2025): the judge receives the complete answer and assigns grounding labels to all sentences in a single pass, while treating each sentence independently and ignoring cross-sentence dependencies. The prompt provides detailed label definitions, decision rules, error categories, and grounding examples; because dependencies are not modeled, the judge cannot predict Invalid Support. We also tested a *Minimal* variant with only essential task and output instructions, but it yielded lower validation macro-F1 across all judges and is therefore not reported.

Holistic variants. These variants require the judge to infer and account for answer dependencies without reference structural annotations. *Baseline* performs grounding in a single pass, instructing the judge to read the complete answer and account for relevant relations implicitly. *Entailment* first infers premise–conclusion relations as formal entailment statements, then grounds the sentences using those relations. *Self-graph* instead first predicts sentence roles and explicit dependency edges, then grounds using the predicted graph. *Claim decomposition* splits sentences into atomic claims, grounds them individually, and tracks dependencies at the claim rather than sentence level. Finally, *Minimal* performs dependency-aware grounding in a single pass using concise label definitions and decision rules, without an explicit intermediate structure.

Oracle-graph variants. Four variants receive the complete reference dependency structure, isolating the judge's ability to use dependencies from its ability to recover them. *Baseline* provides each sentence's role and complete dependency ancestry—including direct and indirect premises—as structured JSON. *Entailment* provides the same information as formal premise–conclusion statements. *Minimal* retains the complete reference graph but removes the baseline prompt's extended guidance and examples. *Terse-graph* likewise provides the complete reference structure, but as a compact plain-text block rather than per-sentence JSON. *Role-only* provides reference sentence roles but withholds dependency edges, which the judge must infer before grounding.

Selection. For each judge and dependency-aware condition, we select the variant with the highest validation macro-F1 averaged across the three evidence scopes, excluding variants with more than 5% invalid or missing outputs. Tables 5 and 6 report results for the holistic and oracle-graph conditions. Independent verification uses the detailed prompt above and therefore requires no per-judge variant selection.

[2] https://developers.openai.com/api/docs/guides/batch

[3] https://ai.google.dev/gemini-api/docs/batch-api

Table 5: **Prompt-variant selection for open-weight judges.** Validation macro-F1 by judge, prompt variant, and evidence scope. *Avg* is the mean across scopes and determines selection, while bold marks the selected variant for each judge and condition. Variants producing more than 5% invalid or missing outputs are discarded and struck through.

	Oracle-graph					Holistic				
Model	**Variant**	**Single**	**Cross**	**Full**	**Avg**	**Variant**	**Single**	**Cross**	**Full**	**Avg**
Qwen3.5 9B	baseline	0.406	0.556	0.341	0.434	baseline	**0.499**	**0.317**	**0.274**	**0.364**
	entailment	0.455	0.422	0.392	0.423	claim_decomp	0.336	0.395	0.321	0.351
	minimal	0.439	0.447	0.383	0.423	entailment	0.310	0.319	0.259	0.296
	role_only	0.487	0.382	0.316	0.395	minimal	~~0.386~~	~~0.348~~	~~0.288~~	~~0.341~~
	terse_graph	**0.505**	**0.503**	**0.299**	**0.435**	self_graph	0.393	0.378	0.306	0.359
Qwen3.8 27B	baseline	0.597	0.516	0.570	0.561	baseline	~~0.514~~	~~0.506~~	~~0.401~~	~~0.474~~
	entailment	**0.671**	**0.587**	**0.518**	**0.592**	claim_decomp	0.433	0.430	0.401	0.421
	minimal	0.590	0.539	0.574	0.568	entailment	**0.610**	**0.476**	**0.449**	**0.512**
	role_only	~~0.710~~	~~0.543~~	~~0.811~~	~~0.688~~	minimal	0.407	0.461	0.473	0.447
	terse_graph	0.542	0.526	0.500	0.523	self_graph	~~0.553~~	~~0.489~~	~~0.613~~	~~0.552~~
Qwen3.6 35B-A3B	baseline	**0.520**	**0.513**	**0.502**	**0.511**	baseline	0.386	0.326	0.376	0.363
	entailment	0.404	0.391	0.479	0.425	claim_decomp	0.357	0.435	0.393	0.395
	minimal	0.425	0.564	0.338	0.443	entailment	0.428	0.384	0.409	0.407
	role_only	0.590	0.358	0.440	0.463	minimal	0.348	0.413	0.519	0.427
	terse_graph	0.418	0.533	0.414	0.455	self_graph	**0.593**	**0.364**	**0.380**	**0.446**
Qwen3.5 122B-A10B	baseline	**0.603**	**0.492**	**0.484**	**0.526**	baseline	0.532	0.440	0.340	0.437
	entailment	0.454	0.632	0.454	0.513	claim_decomp	**0.514**	**0.450**	**0.378**	**0.447**
	minimal	0.360	0.489	0.391	0.413	entailment	0.403	0.441	0.472	0.439
	role_only	0.390	0.427	0.417	0.411	minimal	0.324	0.381	0.421	0.375
	terse_graph	0.436	0.610	0.383	0.476	self_graph	0.371	0.502	0.400	0.424
Qwen3.5 397B-A17B	baseline	0.478	0.489	0.384	0.450	baseline	0.375	0.322	0.378	0.358
	entailment	**0.542**	**0.505**	**0.480**	**0.509**	claim_decomp	0.351	0.400	0.356	0.369
	minimal	0.450	0.494	0.391	0.445	entailment	**0.601**	**0.516**	**0.412**	**0.510**
	role_only	0.444	0.428	0.392	0.421	minimal	0.371	0.435	0.364	0.390
	terse_graph	0.473	0.481	0.473	0.476	self_graph	0.391	0.496	0.469	0.452
Gemma4 E4B	baseline	0.290	0.320	0.254	0.288	baseline	0.273	0.372	0.209	0.285
	entailment	**0.216**	**0.408**	**0.328**	**0.318**	claim_decomp	0.315	0.252	0.233	0.266
	minimal	0.253	0.364	0.238	0.285	entailment	**0.241**	**0.360**	**0.305**	**0.302**
	role_only	0.262	0.254	0.288	0.268	minimal	0.243	0.277	0.219	0.247
	terse_graph	0.222	0.364	0.254	0.280	self_graph	0.260	0.345	0.195	0.267
Gemma4 26B-A4B	baseline	0.517	0.545	0.489	0.517	baseline	0.348	0.412	0.388	0.383
	entailment	0.419	0.505	0.482	0.468	claim_decomp	0.329	0.418	0.370	0.372
	minimal	**0.468**	**0.620**	**0.505**	**0.531**	entailment	0.385	0.449	0.497	0.443
	role_only	0.460	0.341	0.483	0.428	minimal	0.380	0.414	0.405	0.400
	terse_graph	0.482	0.426	0.478	0.462	self_graph	**0.436**	**0.549**	**0.416**	**0.467**
Gemma4 31B	baseline	**0.504**	**0.487**	**0.463**	**0.485**	baseline	0.368	0.343	0.407	0.373
	entailment	0.542	0.465	0.405	0.471	claim_decomp	0.327	0.288	0.350	0.322
	minimal	0.329	0.451	0.382	0.387	entailment	0.340	0.347	0.440	0.375
	role_only	0.412	0.456	0.395	0.421	minimal	0.347	0.582	0.425	0.452
	terse_graph	0.479	0.393	0.508	0.460	self_graph	**0.598**	**0.353**	**0.405**	**0.452**
GLM-4.6V Flash	baseline	0.377	0.209	0.235	0.274	baseline	0.323	0.205	0.259	0.263
	entailment	0.351	0.214	0.237	0.268	claim_decomp	0.273	0.258	0.278	0.270
	minimal	**0.286**	**0.336**	**0.259**	**0.294**	entailment	0.254	0.243	0.251	0.249
	role_only	0.249	0.268	0.259	0.259	minimal	**0.323**	**0.238**	**0.259**	**0.273**
	terse_graph	0.335	0.262	0.259	0.285	self_graph	0.243	0.209	0.251	0.234
GLM-4.6V	baseline	**0.287**	**0.344**	**0.314**	**0.315**	baseline	0.262	0.287	0.218	0.256
	entailment	0.308	0.397	0.215	0.307	claim_decomp	0.229	0.210	0.239	0.226
	minimal	0.319	0.326	0.240	0.295	entailment	0.251	0.251	0.217	0.240
	role_only	0.354	0.217	0.255	0.275	minimal	0.240	0.238	0.264	0.247
	terse_graph	0.372	0.269	0.240	0.294	self_graph	**0.343**	**0.242**	**0.214**	**0.267**

Table 6: **Prompt-variant selection for proprietary judges.** Validation macro-F1 by judge, prompt variant, and evidence scope. *Avg* is the mean across scopes and determines selection, while bold marks the selected variant for each judge and condition. Variants producing more than 5% invalid or missing outputs are discarded and struck through.

	Oracle-graph					Holistic				
Model	**Variant**	**Single**	**Cross**	**Full**	**Avg**	**Variant**	**Single**	**Cross**	**Full**	**Avg**
GPT-5.6 Luna	baseline	0.665	0.560	0.553	0.593	baseline	0.624	0.489	0.551	0.555
	entailment	0.514	0.538	0.556	0.536	claim_decomp	0.458	0.344	0.473	0.425
	minimal	0.628	0.586	0.539	0.584	entailment	0.470	0.502	0.554	0.509
	role_only	**0.735**	**0.516**	**0.584**	**0.612**	minimal	0.450	0.514	0.466	0.477
	terse_graph	0.584	0.580	0.577	0.581	self_graph	**0.618**	**0.540**	**0.616**	**0.592**
Gemini-3.6 Flash	baseline	0.393	0.624	0.524	0.513	baseline	0.424	0.460	0.513	0.466
	entailment	**0.499**	**0.570**	**0.556**	**0.541**	claim_decomp	~~0.400~~	~~0.000~~	~~0.000~~	~~0.133~~
	minimal	0.461	0.567	0.499	0.509	entailment	0.417	0.484	0.453	0.452
	role_only	0.383	0.549	0.557	0.496	minimal	0.383	0.488	0.498	0.456
	terse_graph	0.482	0.588	0.406	0.492	self_graph	**0.426**	**0.492**	**0.486**	**0.468**
GPT-5.4 mini	baseline	0.408	0.473	0.524	0.468	baseline	0.307	0.366	0.382	0.352
	entailment	**0.601**	**0.510**	**0.470**	**0.527**	claim_decomp	**0.496**	**0.302**	**0.424**	**0.407**
	minimal	0.378	0.427	0.374	0.393	entailment	0.314	0.350	0.389	0.351
	role_only	0.339	0.357	0.338	0.345	minimal	0.378	0.327	0.427	0.377
	terse_graph	0.396	0.457	0.387	0.413	self_graph	0.336	0.395	0.387	0.372

E Additional Analyses

E.1 Completion Length vs. Accuracy

Figure 6 examines whether judge accuracy is associated with completion length. Across the full model pool, completion tokens correlate positively with macro-F1 (Pearson's $r = 0.60$, 0.62, and 0.39 for oracle-graph, holistic, and independent verification, respectively; the association is weakest under independent verification). However, this association is largely driven by differences between model families: Qwen models generally produce longer completions and outperform GLM models. Within the Qwen family, the relationship is negative under oracle-graph ($r = -0.55$) and independent verification ($r = -0.66$), with Qwen3.5-9B producing the longest completions but achieving the lowest macro-F1. Under holistic verification, the within-family correlation is instead positive ($r = 0.48$). Completion length thus does not explain accuracy among closely related judges.

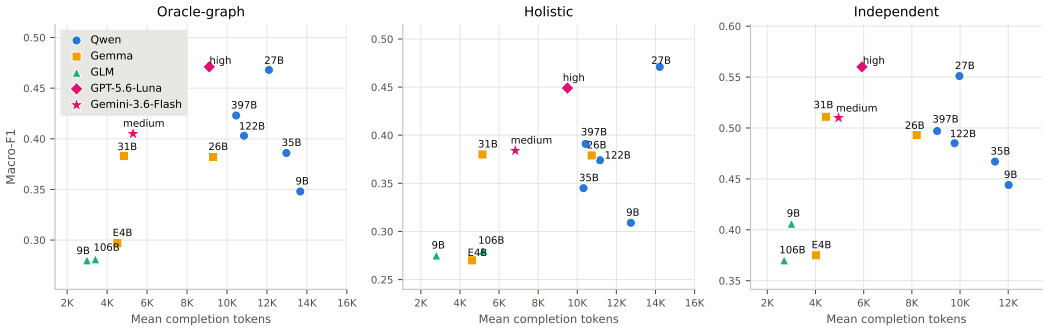


Figure 6: **Completion length vs. grounding performance.** Mean completion tokens per prediction and macro-F1 for each judge under independent, holistic, and oracle-graph verification. Colors denote model families. Labels report total parameters for open-weight models and the configured reasoning or thinking effort for proprietary models, whose parameter counts are undisclosed.

E.2 Bias Toward Supported Across Judges

All 12 judges are biased toward predicting SUPPORTED. Figure 7 shows, for every judge and condition, how sentences of each gold label are predicted, pooled across evidence scopes; the out-

lined column is the prediction of SUPPORTED. Every judge labels at least 88% of truly SUPPORTED sentences correctly, but the errors on the other labels concentrate in the same column: averaged over the 12 judges, 52–62% of CONTRADICTED and 68–74% of NEI sentences are predicted as SUPPORTED, depending on the condition. The bias is strongest for INVALID SUPPORT: the mean fraction predicted as SUPPORTED is 0.82 under Holistic verification and falls to 0.63 with reference dependencies, yet it still ranges from 0.43 to 0.88 across judges under Oracle-graph. Judges tend to accept a sentence even when the source contradicts it, cannot support it, or the premises it relies on are invalid. Independent verification has no INVALID SUPPORT label, so its matrices are 3×3.

E.3 EVIDENCE LOCALIZATION

In addition to a grounding label, each judge returns the document region used as evidence. Under oracle-graph verification, we evaluate localization on observation sentences with a reference bounding box and a gold label of SUPPORTED or CONTRADICTED, restricting evaluation to instances classified correctly by the judge. The resulting scores therefore measure localization conditional on correct grounding classification, and the evaluated subset varies across judges.

Localization is measured using Intersection over Union (IoU), defined as the area shared by the predicted and reference boxes divided by their combined area. Each reference page is scored separately, and a page with no predicted box scores zero. When a page has several reference boxes, the prediction is scored against the best-matching one. We report mean IoU and two threshold-based recall measures: Recall@0.25 and Recall@0.5 are the proportions of scored predictions achieving an IoU of at least 0.25 and 0.5, respectively. Because reference boxes are drawn by hand, their exact extent is subjective: the same evidence can be boxed as a single cell, a row, or the entire table, even though annotators are asked to draw the smallest region that contains it. IoU is symmetric, so it can be low for a prediction that points to the right evidence with a different extent (see Figure 8). We therefore also report two asymmetric measures, computed on the same matched boxes and maximized over the reference boxes on the page as for IoU: *precision*, $|P \cap G|/|P|$, the share of the predicted box P lying inside the reference box G, which does not penalize a box tighter than the reference; and *coverage*, $|P \cap G|/|G|$, the share of the reference box covered by the prediction. The two metrics should be read together, since precision alone rewards arbitrarily small boxes. We thus use IoU as the primary measure and precision and coverage to interpret it.

Table 7 shows substantial variation across judges and generally lower overlap beyond the single-page setting. Although comparisons are conditional on each judge's correctly classified subset, localization rankings do not mirror grounding-classification rankings. Precision is consistently higher than IoU, and judges are ordered similarly under both, with GLM-4.6V the most precise (0.526); part of the low IoU therefore reflects differences in box extent rather than wrong regions. However, 20–44% of scored reference pages contain no predicted box, and these contribute zero to all three measures. Figure 8 illustrates two CONTRADICTED cases.

Table 7: **Evidence localization under oracle-graph verification.** Mean IoU ($[0, 1]$, ↑) by evidence scope and overall (*Avg*), computed on correctly classified SUPPORTED or CONTRADICTED observations with reference boxes. n is the number of scored reference pages. R@t is the proportion with IoU $\geq t$. 🔒 marks proprietary judges.

Model	Single	Cross	Full	Avg	n	R@0.25	R@0.5	Prec.	Cov.
Qwen3.5-9B	0.081	0.055	0.027	0.054	1000	0.082	0.024	0.170	0.102
Qwen3.8-27B	0.199	0.128	0.136	0.156	1081	0.240	0.116	0.376	0.242
Qwen3.6-35B-A3B	0.128	0.072	0.048	0.084	1040	0.129	0.047	0.241	0.135
Qwen3.5-122B-A10B	0.171	0.115	0.057	0.115	1055	0.177	0.084	0.285	0.176
Qwen3.5-397B-A17B	0.176	0.105	0.048	0.111	1086	0.174	0.077	0.308	0.165
Gemma4-E4B	0.087	0.072	0.030	0.062	915	0.087	0.025	0.208	0.113
Gemma4-26B-A4B	0.220	0.236	0.136	0.195	1027	0.302	0.173	0.453	0.278
Gemma4-31B	0.178	0.184	0.137	0.165	1047	0.258	0.138	0.446	0.213
GLM-4.6V-Flash (9B)	0.229	0.204	0.175	0.202	969	0.320	0.175	0.443	0.317
GLM-4.6V (106B-A12B)	0.227	0.275	0.237	0.244	985	0.369	0.232	0.526	0.358
🔒 GPT-5.6-Luna (high)	0.168	0.158	0.161	0.162	1073	0.265	0.130	0.379	0.245
🔒 Gemini-3.6-Flash (medium)	0.302	0.246	0.157	0.234	1070	0.350	0.215	0.473	0.362

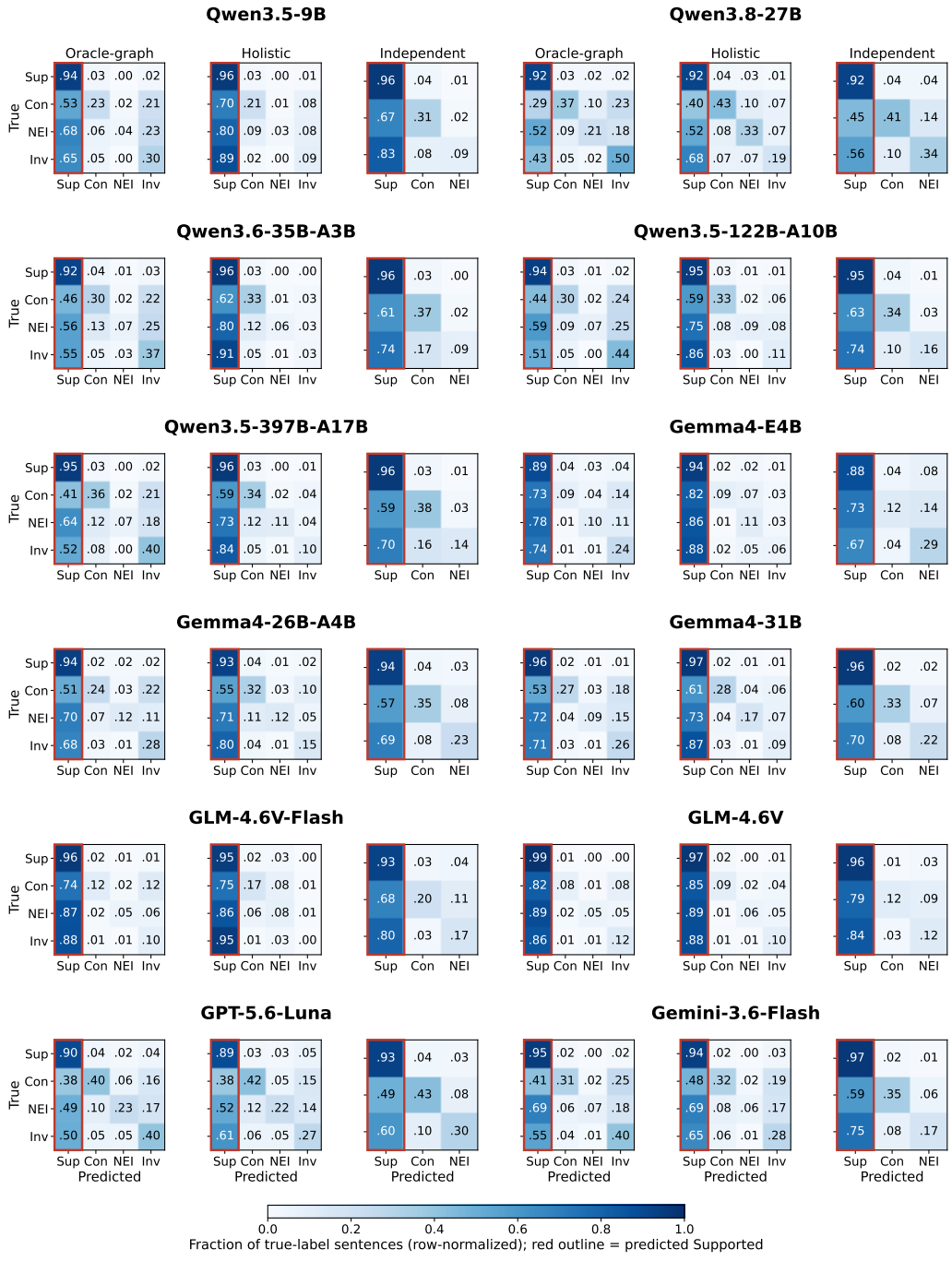
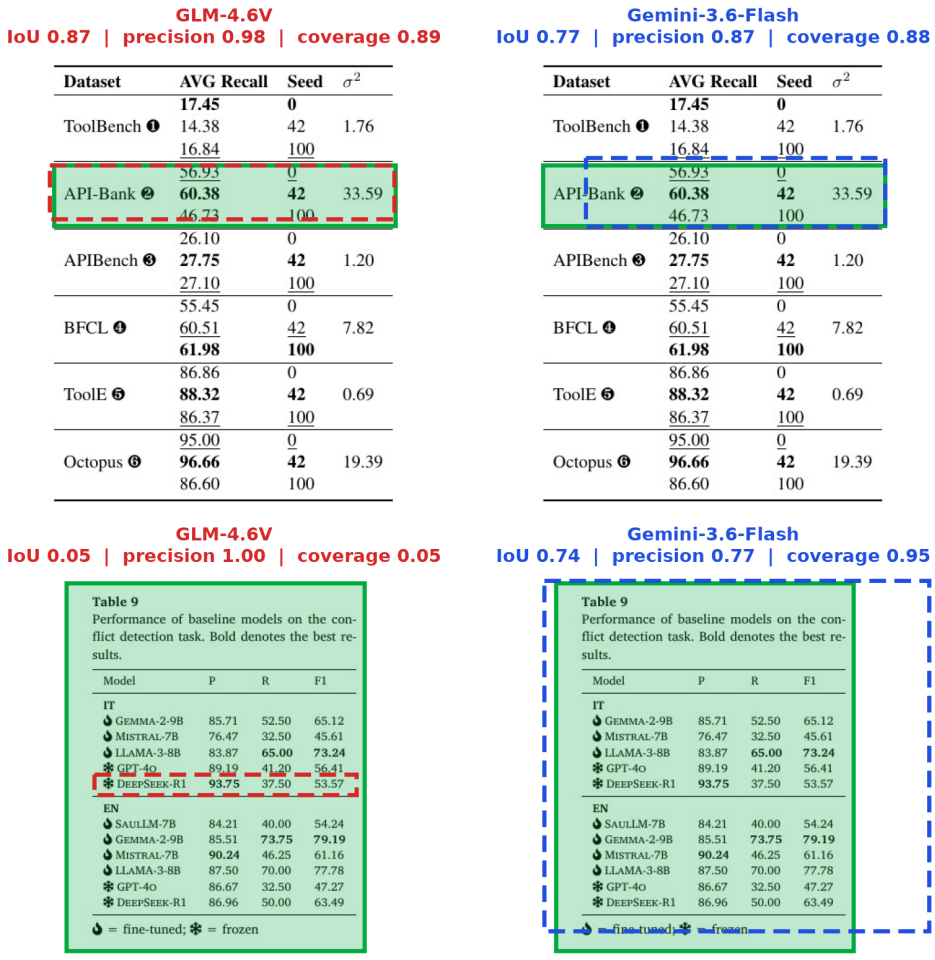


Figure 7: **Bias toward SUPPORTED for all judges.** Row-normalized confusion matrices under *oracle-graph*, *holistic*, and *independent* verification, pooled across evidence scopes. Rows are gold labels and columns are predictions. Each cell is the fraction of the gold label's sentences predicted as that column. The red outline marks predictions of SUPPORTED, which is the most frequent prediction in 91 of the 96 rows with a non-SUPPORTED gold label. Sup, Con, NEI, and Inv denote SUPPORTED, CONTRADICTED, NEI, and INVALID SUPPORT, respectively. Independent verification has no Inv label. Each judge uses its validation-selected prompt.



Dataset	AVG Recall	Seed	σ^2
	17.45	**0**	
ToolBench ❶	14.38	42	1.76
	16.84	100	
	56.93	0	
API-Bank ❷	**60.38**	**42**	33.59
	46.73	100	
	26.10	0	
APIBench ❸	**27.75**	**42**	1.20
	27.10	100	
	55.45	0	
BFCL ❹	60.51	42	7.82
	61.98	**100**	
	86.86	0	
ToolE ❺	**88.32**	**42**	0.69
	86.37	100	
	95.00	0	
Octopus ❻	**96.66**	**42**	19.39
	86.60	100	

Table 9
Performance of baseline models on the conflict detection task. Bold denotes the best results.

Model	P	R	F1
IT			
🔥 GEMMA-2-9B	85.71	52.50	65.12
🔥 MISTRAL-7B	76.47	32.50	45.61
🔥 LLAMA-3-8B	83.87	**65.00**	**73.24**
❄ GPT-4o	89.19	41.20	56.41
❄ DEEPSEEK-R1	**93.75**	37.50	53.57
EN			
🔥 SAULLM-7B	84.21	40.00	54.24
🔥 GEMMA-2-9B	85.51	**73.75**	**79.19**
🔥 MISTRAL-7B	**90.24**	46.25	61.16
🔥 LLAMA-3-8B	87.50	70.00	77.78
❄ GPT-4o	86.67	32.50	47.27
❄ DEEPSEEK-R1	86.96	50.00	63.49

🔥 = fine-tuned; ❄ = frozen

Figure 8: **Localization examples on CONTRADICTED observations in oracle-graph verification.** Green: human reference box; dashed: predicted box. Each title reports IoU, precision (share of the predicted box lying inside the reference box), and coverage (share of the reference box covered by the prediction). *Top* (single-page): both models box the same rows of the results table as the annotators. *Bottom* (full-paper): annotators boxed the whole table; GLM-4.6V boxes only the row containing the contradicting value, which yields a low IoU despite pointing to the right cell, while Gemini-3.6-Flash boxes the whole table. Examples are hand-picked illustrations, not typical cases.

E.4 EVALUATOR-GENERATOR FAMILY OVERLAP

Judge and answer-generator pools overlap in model family (Qwen, Gemma, GLM, GPT, and Gemini). We examine whether grounding accuracy differs between answers generated by the judge's own family and those generated by other families. For each judge, we pool predictions across the three evaluation conditions using their validation-selected prompts and condition-specific reference labels, including the INVALID SUPPORT-to-SUPPORTED projection under Independent verification.

Table 8 shows that the direction of the difference varies across families. Gemma and Gemini judges achieve higher accuracy on same-family answers, whereas Qwen, GLM, and GPT judges achieve lower accuracy. The largest gaps occur for GLM, whose same-family subset contains only 14 QA pairs from one generator. Pooled accuracy is numerically similar for same-family and different-family predictions (0.622 vs. 0.619), although the two groups weight judges differently.

These comparisons are descriptive: generator subsets may differ in difficulty, label distribution, and evidence scope. Moreover, accuracy measures agreement with reference labels, not favorable

treatment of generated answers. The results therefore neither establish self-preference nor rule out effects of family overlap on judge comparisons.

Table 8: **Grounding accuracy by generator-family overlap.** Same-family and different-family accuracy, pooled across evaluation conditions using condition-specific reference labels. n counts sentence–condition predictions, not distinct sentences; #Gen counts same-family generator models. 🔒 marks proprietary judges.

Judge	Family	Same-family acc.	n	Diff-family acc.	n	#Gen
Qwen3.5-9B	Qwen	0.566	1656	0.615	6309	4
Qwen3.8-27B	Qwen	0.633	1656	0.673	6309	4
Qwen3.6-35B-A3B	Qwen	0.619	1656	0.641	6309	4
Qwen3.5-122B-A10B	Qwen	0.599	1656	0.647	6309	4
Qwen3.5-397B-A17B	Qwen	0.625	1656	0.665	6309	4
Gemma4-E4B	Gemma	0.567	1149	0.535	6816	2
Gemma4-26B-A4B	Gemma	0.694	1149	0.608	6816	2
Gemma4-31B	Gemma	0.701	1149	0.621	6816	2
GLM-4.6V-Flash (9B)	GLM	0.399	381	0.583	7584	1
GLM-4.6V (106B-A12B)	GLM	0.352	381	0.572	7584	1
🔒 GPT-5.6-Luna (high)	GPT	0.645	1923	0.663	6042	1
🔒 Gemini-3.6-Flash (medium)	Gemini	0.697	1821	0.628	6144	2
Pooled (all 12 judges)		0.622	16233	0.619	79347	–

E.5 Performance by Question Category

We examine grounding performance by question category (Section 3.2). Single-page instances comprise Comparison, Trend, Quantitative, Visual Understanding, and Adversarial questions. Cross-page and full-paper instances are analyzed jointly using Conceptual, Methodological, Results Interpretation, Implication, and Adversarial categories. For these two scopes, instances marked as adversarial are assigned exclusively to that category, regardless of their original question type. We report each evaluation condition separately because Independent uses three grounding labels, whereas Holistic and Oracle-graph use four.

Across the 12 judges, Visual Understanding achieves the highest mean macro-F1 among single-page categories under every condition, while Trend achieves the lowest (Table 10). Category differences are smaller for cross-page/full-paper instances (Table 11): Implication has the highest mean score under Holistic and Oracle-graph verification and is approximately tied with Adversarial under Independent verification. Adversarial questions do not consistently yield the lowest scores in either group. These comparisons are descriptive: category subsets are small and may differ in grounding-label composition and, for the combined cross-page/full-paper analysis, evidence scope.

E.6 Performance by Evidence Type

We group single-page instances by the source elements recorded in the question metadata: figure-only (49 QA pairs), table-only (30), and mixed evidence (12). Mixed instances combine figures and tables or either with explicitly referenced textual evidence. These groups describe the evidence identified for each question, not the modality required to verify every answer sentence. We restrict this analysis to single-page instances, where the referenced elements provide a clear grouping within a common evidence scope.

Table 12 reports macro-F1 by evidence type and evaluation condition. Across the 12 judges, table-only instances achieve higher mean macro-F1 than figure-only instances under Independent (0.567 vs. 0.530), Holistic (0.435 vs. 0.382), and Oracle-graph verification (0.481 vs. 0.418). Mixed-evidence instances have the lowest mean scores (0.467, 0.376, and 0.352, respectively), although this group contains only 12 QA pairs. These differences may also reflect variation in question categories and grounding-label composition; they do not isolate the contribution of visual information.

Table 9: **Question categories.** Each category is defined by its core reasoning requirement and illustrated with questions drawn verbatim from the benchmark. Cross-page and full-paper questions use the same categories but differ in the scope of evidence required, as shown below.

Category	Core reasoning	Illustrative question (verbatim)
Single-page questions		
Comparison	Compare values, properties, or patterns across visual elements.	"Looking at Figure 2, how does the number of child nodes differ across levels A, B, and C?"
Trend	Identify how a quantity or pattern changes across conditions or over time.	"Looking at Figure 7, how does the learned reward progress over time for the Soccer environment?"
Quantitative	Read exact numeric values from visual evidence and evaluate them against a fixed reference, threshold, or extremum.	"In Figure 2B, look at the hazard ratio point estimates for SDHB, Sex, Nonpseudoglandular variant, and Local recurrence; how do these values compare to the reference line of 1.0, and which one is the lowest?"
Visual understanding	Interpret the structure, components, or process represented by a visual element.	"Can you explain how the process illustrated in Figure 1 works, specifically the flow from the claim and publication content to the final fallacy identification?"
Adversarial	Verify a misleading interpretation against the visual evidence.	"If I'm reading this right, the multitask training approach always outperforms the single-task baseline in terms of R-1f1 across all the different summary length configurations mentioned in the text and Table 5. Is that right?"
Cross-page and full-paper questions		
Conceptual	Connect concepts, phenomena, or design choices described in different parts of the paper.	*Cross-page:* "Looking at Figure 3 and Figure 5, how does the prompt format used in the Prompt Tuning strategy determine which basic objective is being employed?" *Full-paper:* "How does the paper connect the background-first input design for the generator with the way the system selects documents before summarization?"
Methodological	Relate methodological choices or experimental settings to observed performance.	*Cross-page:* "In Figure 4, how does the pruning module's selection flow explain the pattern in Figure 6 where some information-rich parts of speech are kept more often than other token types? Looking at both figures, what structural choice in the architecture seems to drive that selective preservation?" *Full-paper:* "How does the step-wise verification design in the method connect to the performance patterns in the results, and what seems to explain the improvement it produces?"
Results interpretation	Synthesize and relate findings across figures, tables, or experiments.	*Cross-page:* "How does the error magnitude in the top plot of Figure 3 relate to the graph structure shown in Figure 1?" *Full-paper:* "What do Table 3 and Table S2 together reveal about where the different parsing models excel versus where they struggle, looking across both the document types and specific text attributes, and does Table S1 show a similar pattern of model strengths when it comes to table reconstruction?"
Implication	Infer consequences, broader conclusions, or limitations supported by evidence across the paper.	*Cross-page:* "Based on Figure 3 and Table 10, what does the relationship between the approach's resource usage and its performance across different attack types reveal about where its robustness breaks down?" *Full-paper:* "What does Figure 3 show about the generalization failures of the different trained models, and what does the supplementary analysis of the predicted curves reveal about the limitations of learning these binding behaviors?"
Adversarial	Verify a misleading interpretation against evidence from multiple parts of the paper.	*Cross-page:* "If I'm reading Table 5 and Figure 2 correctly, although Bi-GRU outperforms CNN on the MIMIC-III clean dataset overall, Figure 2 (left) shows that at the un-tuned decision boundary of 0.5, CNN actually achieves a higher Micro F1 score than Bi-GRU. Is this interpretation of the threshold curves correct?" *Full-paper:* "I'm trying to digest the dataset composition of ChartMuseum compared to other benchmarks. The abstract mentions that the dataset covers a wide variety of real-world charts from many sources to evaluate complex reasoning. Looking at the chart type distributions in Table 6, this diversity really shows, especially since ChartMuseum has a higher proportion of line and scatter charts compared to CharXiv. Am I reading Table 6 correctly, or did I mix those columns up?"

E.7 Reasoning Structure by Evidence Scope

Table 13 summarizes how the dependency graphs of the annotated answers change with evidence scope. Answers have similar length and a similar number of dependency edges across scopes, but their structure differs. Cross-page and full-paper answers contain more interpretations (51.3% of sentences, versus 39.8% for single-page), and their graphs are deeper: mean depth rises from 2.0 to 2.9 and 3.2, and 58.4% and 55.8% of answers have three or more layers, versus 25.3% for single-page. Single-page interpretations synthesize more premises at once (2.6 direct premises, versus 2.0 and 1.9), whereas in cross-page and full-paper answers more interpretations build directly on

Table 10: **Grounding performance by single-page question category.** Macro-F1 for Comparison ($n = 22$), Trend ($n = 17$), Quantitative ($n = 19$), Visual Understanding ($n = 19$), and Adversarial ($n = 14$) QA pairs. Ind/Hol/Ora denote Independent/Holistic/Oracle-graph verification.

	Comparison			Trend			Quantitative			Visual Underst.			Adversarial		
Model	Ind	Hol	Ora	Ind	Hol	Ora	Ind	Hol	Ora	Ind	Hol	Ora	Ind	Hol	Ora
Qwen3.5-9B	0.43	0.36	0.45	0.43	0.24	0.31	0.45	0.45	0.42	0.60	0.35	0.38	0.43	0.36	0.31
Qwen3.8-27B	0.52	0.53	0.50	0.46	0.36	0.48	0.62	0.46	0.53	0.61	0.55	0.53	0.48	0.47	0.41
Qwen3.6-35B-A3B	0.52	0.33	0.50	0.43	0.29	0.35	0.54	0.43	0.51	0.62	0.40	0.47	0.45	0.34	0.37
Qwen3.5-122B-A10B	0.52	0.50	0.48	0.44	0.30	0.40	0.63	0.45	0.44	0.64	0.48	0.45	0.48	0.37	0.45
Qwen3.5-397B-A17B	0.53	0.46	0.48	0.43	0.27	0.47	0.48	0.43	0.40	0.74	0.53	0.56	0.48	0.45	0.45
Gemma4-E4B	0.32	0.25	0.29	0.34	0.15	0.23	0.43	0.37	0.35	0.54	0.30	0.45	0.30	0.29	0.30
Gemma4-26B-A4B	0.51	0.45	0.50	0.41	0.23	0.22	0.66	0.43	0.54	0.70	0.54	0.50	0.47	0.40	0.31
Gemma4-31B	0.54	0.45	0.48	0.37	0.20	0.21	0.56	0.49	0.55	0.70	0.54	0.60	0.46	0.32	0.34
GLM-4.6V-Flash (9B)	0.36	0.27	0.31	0.42	0.20	0.21	0.44	0.34	0.37	0.62	0.43	0.40	0.38	0.26	0.26
GLM-4.6V (106B-A12B)	0.40	0.33	0.35	0.33	0.26	0.26	0.51	0.33	0.35	0.51	0.40	0.39	0.33	0.27	0.28
🔒 GPT-5.6-Luna (high)	0.55	0.44	0.45	0.50	0.38	0.48	0.55	0.45	0.51	0.64	0.49	0.50	0.47	0.40	0.45
🔒 Gemini-3.6-Flash (medium)	0.54	0.41	0.45	0.42	0.29	0.24	0.67	0.51	0.53	0.66	0.48	0.67	0.47	0.43	0.36

Table 11: **Grounding performance by cross-page/full-paper question category.** Macro-F1 for Conceptual ($n = 28$), Methodological ($n = 33$), Results Interpretation ($n = 39$), Implication ($n = 36$), and Adversarial ($n = 30$) QA pairs. Categories are mutually exclusive; adversarial instances are grouped separately from their original question types. Ind/Hol/Ora denote Independent/Holistic/Oracle-graph verification.

	Conceptual			Methodological			Results Interpr.			Implication			Adversarial		
Model	Ind	Hol	Ora	Ind	Hol	Ora	Ind	Hol	Ora	Ind	Hol	Ora	Ind	Hol	Ora
Qwen3.5-9B	0.45	0.29	0.29	0.37	0.25	0.33	0.44	0.26	0.30	0.43	0.29	0.33	0.48	0.26	0.34
Qwen3.8-27B	0.46	0.44	0.42	0.49	0.40	0.42	0.60	0.44	0.42	0.53	0.48	0.45	0.53	0.40	0.44
Qwen3.6-35B-A3B	0.43	0.35	0.33	0.43	0.28	0.39	0.45	0.30	0.31	0.44	0.35	0.36	0.55	0.35	0.36
Qwen3.5-122B-A10B	0.38	0.28	0.41	0.40	0.29	0.35	0.48	0.37	0.34	0.53	0.33	0.38	0.42	0.33	0.38
Qwen3.5-397B-A17B	0.44	0.30	0.40	0.43	0.39	0.33	0.47	0.34	0.36	0.54	0.39	0.41	0.50	0.33	0.37
Gemma4-E4B	0.37	0.27	0.29	0.33	0.23	0.28	0.35	0.29	0.25	0.39	0.26	0.26	0.39	0.22	0.30
Gemma4-26B-A4B	0.43	0.25	0.36	0.39	0.33	0.30	0.46	0.37	0.31	0.52	0.40	0.36	0.50	0.43	0.37
Gemma4-31B	0.42	0.31	0.34	0.43	0.33	0.27	0.50	0.35	0.34	0.52	0.42	0.38	0.51	0.30	0.34
GLM-4.6V-Flash (9B)	0.40	0.25	0.32	0.40	0.22	0.21	0.41	0.24	0.25	0.40	0.28	0.24	0.40	0.30	0.25
GLM-4.6V (106B-A12B)	0.29	0.24	0.21	0.34	0.19	0.21	0.39	0.24	0.27	0.31	0.21	0.23	0.31	0.30	0.20
🔒 GPT-5.6-Luna (high)	0.50	0.41	0.43	0.54	0.44	0.44	0.57	0.44	0.45	0.54	0.43	0.44	0.53	0.44	0.42
🔒 Gemini-3.6-Flash (medium)	0.42	0.34	0.37	0.46	0.36	0.37	0.44	0.30	0.33	0.50	0.38	0.40	0.53	0.40	0.39

Table 12: **Single-page grounding performance by evidence type.** Macro-F1 for figure-only ($n = 49$), table-only ($n = 30$), and mixed ($n = 12$) QA pairs, grouped using question metadata. Ind/Hol/Ora denote Independent/Holistic/Oracle-graph verification. 🔒 denotes proprietary models.

	Figure			Table			Mixed		
Model	**Ind**	**Hol**	**Ora**	**Ind**	**Hol**	**Ora**	**Ind**	**Hol**	**Ora**
Qwen3.5-9B	0.50	0.32	0.35	0.47	0.41	0.47	0.40	0.34	0.30
Qwen3.8-27B	0.58	0.49	0.51	0.63	0.56	0.59	0.46	0.37	0.41
Qwen3.6-35B-A3B	0.52	0.36	0.42	0.66	0.36	0.63	0.45	0.45	0.32
Qwen3.5-122B-A10B	0.55	0.43	0.47	0.65	0.53	0.51	0.65	0.38	0.36
Qwen3.5-397B-A17B	0.59	0.42	0.50	0.59	0.49	0.49	0.45	0.45	0.40
Gemma4-E4B	0.42	0.23	0.37	0.42	0.29	0.33	0.28	0.36	0.23
Gemma4-26B-A4B	0.60	0.41	0.39	0.57	0.45	0.52	0.64	0.44	0.41
Gemma4-31B	0.54	0.41	0.38	0.66	0.50	0.50	0.44	0.39	0.47
GLM-4.6V-Flash (9B)	0.45	0.28	0.31	0.46	0.33	0.34	0.36	0.34	0.24
GLM-4.6V (106B-A12B)	0.45	0.36	0.34	0.47	0.33	0.37	0.36	0.24	0.27
🔒 GPT-5.6-Luna (high)	0.57	0.48	0.51	0.62	0.51	0.56	0.48	0.36	0.35
🔒 Gemini-3.6-Flash (medium)	0.59	0.42	0.46	0.60	0.46	0.46	0.63	0.40	0.47

other interpretations (54.0% and 57.4%, versus 39.3%). The complexity therefore shifts from wide

synthesis over observations to chains of inference, in which an early error can propagate through several steps.

Table 13: **Reasoning structure of annotated answers by evidence scope.** Depth is the highest layer of an answer's dependency graph, with observations at layer 0 and each interpretation one layer above its deepest premise. Direct premises are averaged over interpretations with at least one premise; "on interpretation" is the share of these that directly build on another interpretation.

Scope	QA pairs	Sentences per answer	Interpretations per answer (share)	Direct premises per interpretation	Depth mean (max)	Depth ≥ 3 (% answers)	On interpretation (% interp.)
Single-page	91	10.2	4.1 (39.8%)	2.6	2.0 (5)	25.3	39.3
Cross-page	89	10.1	5.2 (51.3%)	2.0	2.9 (8)	58.4	54.0
Full-paper	77	11.8	6.0 (51.3%)	1.9	3.2 (9)	55.8	57.4

E.8 Uncertainty in the Invalid Support Improvement

We assess uncertainty in the Oracle-graph–Holistic difference in Invalid Support F1 using a paired bootstrap over the existing test predictions, stratified by evidence scope to match how the point estimate itself is computed (Table 3 reports Invalid Support F1 separately per scope, and the aggregate above is the plain mean of the three). We draw 5,000 samples, resampling each scope's 91/89/77 QA pairs with replacement independently and retaining all sentences within each sampled pair. For each judge, both conditions use the same resampled QA pairs within each scope, and their validation-selected prompts remain fixed. Table 14 reports the original F1 estimates, their difference Δ, its 95% percentile bootstrap confidence interval, and the proportion of resamples with $\Delta > 0$.

Point estimates favor Oracle-graph for all 12 judges, with improvements ranging from 0.022 to 0.173. The individual 95% intervals exclude zero for nine judges, without adjustment for multiple comparisons. For Gemma4-26B-A4B, GLM-4.6V, and Gemini-3.6-Flash, the intervals include zero, indicating greater uncertainty about the direction of the difference. These estimates account for dependence among sentences within a QA pair, but not among different QA pairs from the same paper. They quantify uncertainty in the comparison between validation-selected configurations rather than isolating the effect of reference-graph access.

Table 14: **Uncertainty in the Oracle-graph–Holistic difference in Invalid Support F1.** Intervals are individual 95% percentile confidence intervals from 5,000 paired QA-level bootstrap resamples, without multiplicity adjustment. Δ denotes Oracle-graph minus Holistic F1; Positive (%) is the percentage of resamples with $\Delta > 0$. 🔒 marks proprietary judges.

Judge	Oracle-graph	Holistic	Δ	95% CI	$P(\Delta{>}0)$	Sig.
Qwen3.5-9B	0.205	0.101	+0.105	[+0.024, +0.187]	0.993	✓
Qwen3.8-27B	0.291	0.192	+0.100	[+0.008, +0.188]	0.985	✓
Qwen3.6-35B-A3B	0.230	0.056	+0.173	[+0.069, +0.261]	0.999	✓
Qwen3.5-122B-A10B	0.262	0.112	+0.150	[+0.056, +0.245]	0.999	✓
Qwen3.5-397B-A17B	0.257	0.111	+0.146	[+0.041, +0.245]	0.998	✓
Gemma4-E4B	0.189	0.085	+0.104	[+0.004, +0.203]	0.980	✓
Gemma4-26B-A4B	0.195	0.141	+0.054	[-0.046, +0.145]	0.849	✗
Gemma4-31B	0.186	0.107	+0.079	[+0.007, +0.144]	0.982	✓
GLM-4.6V-Flash (9B)	0.107	0.000	+0.107	[+0.024, +0.196]	0.999	✓
GLM-4.6V (106B-A12B)	0.118	0.096	+0.022	[-0.041, +0.084]	0.771	✗
🔒 GPT-5.6-Luna (high)	0.265	0.187	+0.078	[+0.012, +0.142]	0.989	✓
🔒 Gemini-3.6-Flash (medium)	0.240	0.188	+0.053	[-0.024, +0.132]	0.912	✗

F Annotation Guidelines

Annotators followed written guidelines for each annotation step. The boxes below reproduce the rubric for sentence roles, dependencies, and grounding labels, condensed from the full guidelines that accompany the released annotation tool. Figure 9 shows a fully annotated answer.

Role definition: Observation vs. Interpretation

- **Task.** Each answer is already split into sentences. For every sentence, annotators assign a *role* and then a grounding label. The role captures the sentence's logical function, not whether it is right or wrong, so any role can receive any label. Reading the whole answer before annotating is recommended.
- **Closed-book judgement.** Roles depend only on the logical relations among the answer's sentences, not on whether the claim can be located in the paper. The test is: *"Does this sentence logically derive from, or depend on, other sentences in this answer?"* If not, it is an Observation; if so, it is an Interpretation and its premises are listed as dependencies.
- **Observation.** A sentence that makes a standalone factual or analytical claim: its meaning and truth can be assessed without using another answer sentence as a premise. This includes direct reads from figures and tables. Typical cues:
  - a named source followed by a direct read ("Table 2 shows…");
  - a specific numeric value from a table cell or figure axis;
  - a concrete example ("For instance, model X achieves Y% on Z");
  - a stated design purpose ("is intended/designed to…");
  - an author-stated finding ("The authors found that…"), only when the content after that frame is itself a standalone, directly verifiable fact;
  - a comparison whose values are both stated in the same sentence ("achieves 67.4, outperforming Y by 4.3 points") is a compound read, hence an Observation.
- **Interpretation.** A sentence that logically derives from, synthesizes, or reasons across other answer sentences: one can draw an arrow from those sentences into it. Premises need not precede it; textual order is irrelevant. Typical cues:
  - deictic openers ("This represents/demonstrates/confirms/indicates…");
  - participial clauses ("…, confirming/suggesting that…");
  - summary openers ("Ultimately/Together/In sum/Overall, …"); closing summaries are always Interpretations;
  - causal framing ("By doing X, the framework allows/enables/proves…");
  - cross-source synthesis ("…when viewed alongside…");
  - a gap, ratio, or ranking computed from values stated in other sentences;
  - "The authors conclude that…" when it synthesizes other answer sentences (rather than reporting a bare standalone statement).
- **Purpose sentences.** A reported reason ("is designed to…") is a standalone claim about design intent, hence an Observation; an inferred purpose ("therefore/thus prevents…") derives from other claims, hence an Interpretation.
- **Mixed sentences.** When a comma, semicolon, or colon joins a factual read to an inference ("Model A achieves 84.2%, confirming that multi-view contexts are beneficial"), the whole sentence is an Interpretation (Interpretation $>$ Observation).
- **NO_CLAIM.** Sentences that make no grounded claim about the document: transitions, meta-commentary, or vague framing. Used sparingly, only for sentences that carry no factual content.
- **Examples.**
  - *"The framework uses a Generator powered by an LLM to produce multi-view artificial contexts."* Observation: a standalone claim about the architecture.
  - *"Model X achieves the best performance on both MedQA and MedMCQA, with 64.7 and 54.1."* Observation: a self-contained performance claim.
  - *"Ultimately, these results highlight the exceptional efficiency and effectiveness of model X across all benchmarks."* Interpretation: a closing synthesis.
  - *"Let me walk you through what the paper shows."* and *"The paper presents interesting findings."* NO_CLAIM: framing without verifiable content.

Dependency graph construction

- **Purpose.** For every Interpretation, annotators list the IDs of the sentences that serve as its premises (`depends_on`, e.g., S2, S5). This makes the reasoning chain explicit, so a flawed interpretation can be traced either to a wrong premise or to an invalid inference. Each answer therefore forms a directed acyclic graph with an edge $i \rightarrow j$ when interpretation i uses sentence j as a premise; these edges define INVALID SUPPORT (Section 3.3).
- **Rules.**
  - Use the sentence IDs shown in the tool (S1, S2, …) and list every sentence the interpretation *directly* depends on.
  - Dependencies are premises actually used in the answer, not merely related content.
  - Premises may come before or after the interpretation: only the logical dependency matters.
  - Observations never have dependencies, since by definition they do not rely on other answer sentences.
  - If an Interpretation depends on another Interpretation, list only that Interpretation and not the Observations behind it: the chain is transitive.
  - If a single sentence fuses a factual read with its own conclusion, its list is empty: the premise lies inside the sentence, not in a separate one.
- **Example 1: a computed gap.** S1 (Observation): *"Model X achieves 74.3 on MedQA, while its closed-book counterpart scores 63.9."* (`depends_on`: none). S2 (Interpretation): *"This represents an accuracy improvement of 10.4 points over the closed-book baseline."* `depends_on`: S1, since the 10.4 figure is computed from the two values in S1.
- **Example 2: the premise comes later.** S1 (Interpretation): *"This indicates that the generate-then-read approach consistently outperforms retrieval-free methods."* `depends_on`: S2. S2 (Observation): *"Model X achieves the best scores on both MedQA (64.7) and MedMCQA (54.1)."* The premise follows the conclusion, which does not change either role.
- **Example 3: a closing summary.** S1 (Observation): *"The model trained with a single context view achieves 45.8% accuracy on MedQA."* S2 (Observation): *"Using five context views raises accuracy to 53.1%."* S3 (Interpretation): *"Together, these results confirm that increasing the number of context views progressively improves performance."* `depends_on`: S1, S2. Closing sentences synthesize all the sentences they summarize.
- **Example 4: a fused sentence.** *"Model A achieves 84.2% on benchmark X, confirming that multi-view contexts are beneficial."* Interpretation, with an empty `depends_on`: the inference relies only on content inside the sentence.
- **Chains.** In the annotated answer of Figure 9, S5 depends on S3 and S4 (a gap computed from two values), S7 on S6, and S9 on S5, S7, and S8, where S9 is a closing summary that relies on two earlier interpretations and one observation. Because dependencies are transitive, S9 also inherits S3, S4, and S6 through S5 and S7; these are not listed (violet in the figure). Figure 10 shows the dependency graphs of a short and a long annotated answer.

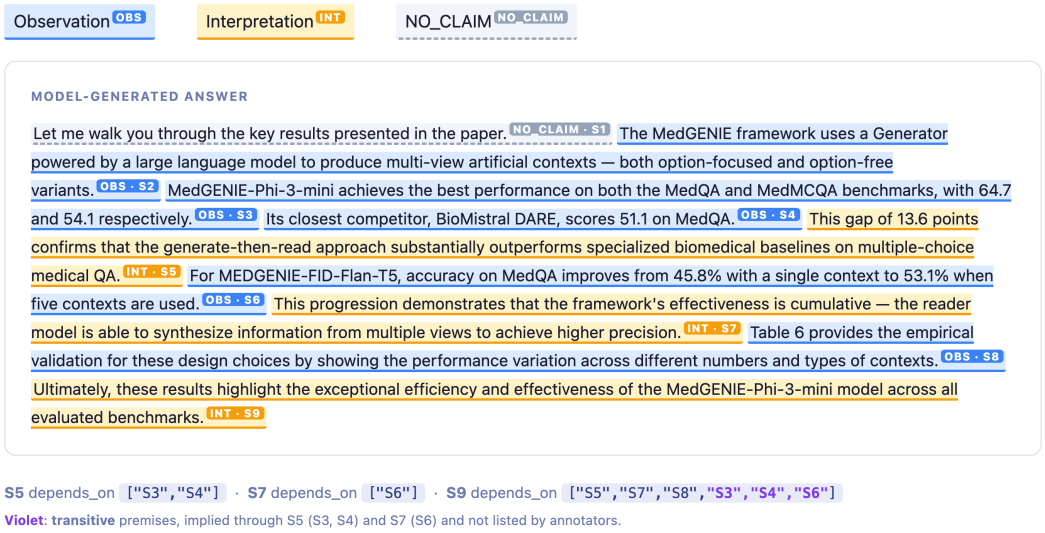


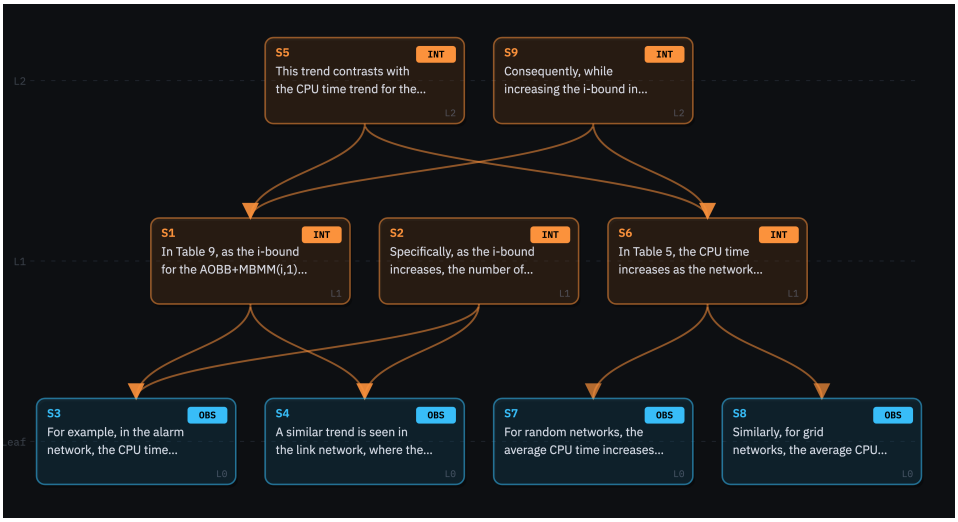
Figure 9: **Role annotation example from the annotator guidelines.** Each sentence of a model-generated answer is colored by role: observations (OBS) report what the source shows, interpretations (INT) draw conclusions from other sentences, and NO_CLAIM sentences contain nothing verifiable. The dependencies below list the direct premises of each interpretation; transitive premises (in violet) are implied and not listed, e.g., S9 also relies on S3, S4, and S6 through S5 and S7.

Grounding labels and error codes

- SUPPORTED. Everything the sentence claims can be verified in the source document, and nothing it claims is contradicted. The sentence need not cite its source explicitly.
- CONTRADICTED. The sentence makes a claim directly falsified by the document: the annotator can point to a specific location that contradicts it.
- NEI (not enough information). The claim cannot be verified from the document: the evidence is absent, the claim requires external knowledge, or the conclusion goes beyond what the document supports. It does not mean "probably wrong": if the document actively disproves the claim, the label is CONTRADICTED. When a sentence mixes both, CONTRADICTED takes precedence over NEI.
- **Evidence boxes.** For each SUPPORTED or CONTRADICTED Observation, annotators draw the smallest region containing the evidence (a table row or cell, a figure panel, a paragraph), never the whole page, with one box per evidence region and at the primary source where the evidence originates. For Interpretations, a box is drawn only when the authors state the same inference; an Interpretation without box or dependencies is valid. No boxes are drawn for NEI.
- **Error codes.** Every CONTRADICTED or NEI sentence must be classified. Span-level codes are attached to the erroneous phrase, one code per span (a sentence may have several spans), optionally with a correction, i.e., the replacement text for that exact span; NEI spans have no correction.
  - **C1, factual misread.** A specific number, name, label, or formula element is stated incorrectly and the correct value is explicit in the document (*"accuracy of 84.2%"* where the table shows 82.4%).
  - **C2, visual misinterpretation.** A trend, ordering, or structure in a figure is described incorrectly; the figure itself is checked, not the caption (*"contexts are generated using a specific option"* where the figure shows all options used simultaneously).
  - **C3, relational error.** The individual observations are read correctly but the stated relation is wrong: direction, degree, invalid comparison, or incorrect arithmetic (*"drops sharply from $k{=}3$ to $k{=}5$"* where the table shows a 0.3-point decrease).
  - **N1, fabricated evidence.** A cited figure, table, result, or statistic does not exist in the document (*"the ablation in Table 4"* when there is no Table 4).
  - **N2, unsupported inference.** The claim relies on external knowledge or on settings the paper does not cover (a mechanism explained through BERT's attention properties, which the document never mentions).
  - **N3, overclaiming conclusion.** The premises are correctly grounded and read, but the conclusion over-extrapolates (*"the definitive state of the art across all clinical settings and task formats"* when only multiple-choice benchmarks are evaluated).
- **Propagated errors.** Two further codes are sentence-level and spanless, for Interpretations only. **C4** (false-premise conclusion) marks a conclusion built on premises that are CONTRADICTED, and **N4** (unverifiable-premise conclusion) one built on NEI premises. The premise sentence carries its own C1–C3 or N1–N2 code; the conclusion carries only C4 or N4, and may do so even if it is correct in substance, since a broken premise breaks the chain. The tool offers them only if at least one listed premise has the matching label. Such sentences are mapped to INVALID SUPPORT, unless they also carry their own C1–C3/N1–N3 error, in which case they keep CONTRADICTED/NEI. Example: an observation misreads a table (C3), and the next sentence concludes *"this confirms that option-free contexts are the dominant factor"* (C4).
- **Disambiguation.** C1 vs. N1: the correct value is in the document (C1) vs. the cited evidence does not exist (N1). C3 vs. N3: the stated relation between correctly read values is wrong (C3) vs. the relation is right but the conclusion over-reaches (N3). N2 vs. N3: even the premises are ungrounded (N2) vs. the premises are grounded but the conclusion oversteps (N3). C4 vs. N4: the upstream premise is contradicted (C4) vs. merely unverifiable (N4). Mixed codes: a span whose problem belongs to the family that lost the CONTRADICTED-over-NEI precedence is still marked, but without a code.

- **Adversarial questions.** Some questions embed a plausible but wrong premise. If the answer accepts and confirms it, the confirming sentences are likely CONTRADICTED; if it identifies and corrects the error, the correcting sentences are annotated on their own grounding.

(a) A short answer: three layers



(b) A long answer: six layers

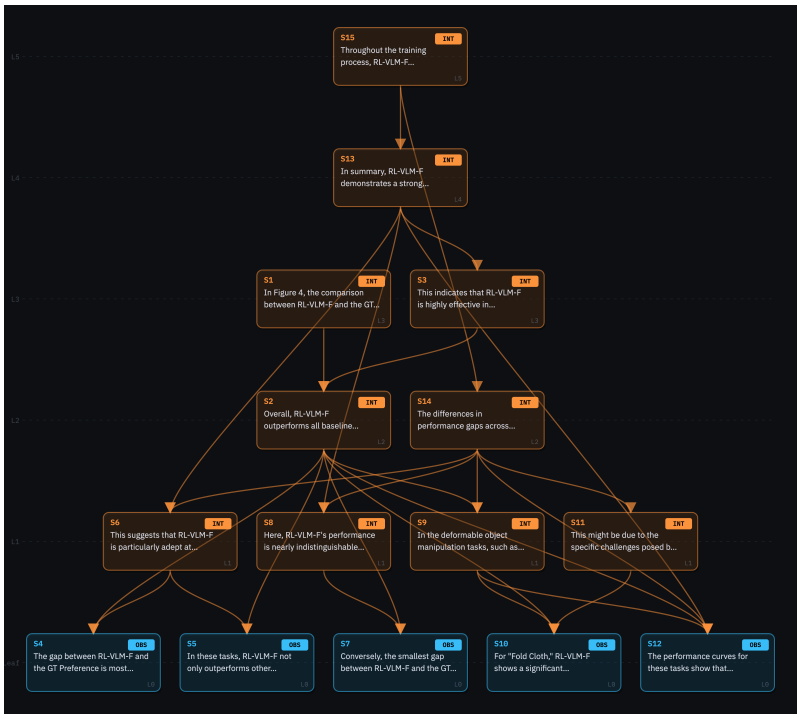


Figure 10: **Dependency graphs of two annotated answers.** Nodes are sentences (orange: interpretation; blue: observation) and each arrow points from an interpretation to a premise it uses. Layers group sentences by depth, from observations (L0) up to interpretations that build on other interpretations. (a) A short answer with three layers. (b) A long answer where chains of interpretations stack up, so upper sentences depend on lower ones only through intermediate steps (transitive dependencies). Both graphs are rendered by the annotation tool from benchmark annotations. The text is truncated by the tool, and the point is the structure.

G HUMAN ANNOTATION INTERFACE

Annotation is performed through a custom browser-based interface requiring no local installation. The interface supports question validation, sentence-role and dependency annotation, sentence-level grounding to the source document, error tagging, and the resolution of disagreements between the two LLM pre-annotation runs and between the two annotators' grounding labels. Figures 11–14 show full-resolution screenshots of the annotation stages, and Figures 15–17 show the resolvers for role, dependency, and grounding disagreements.

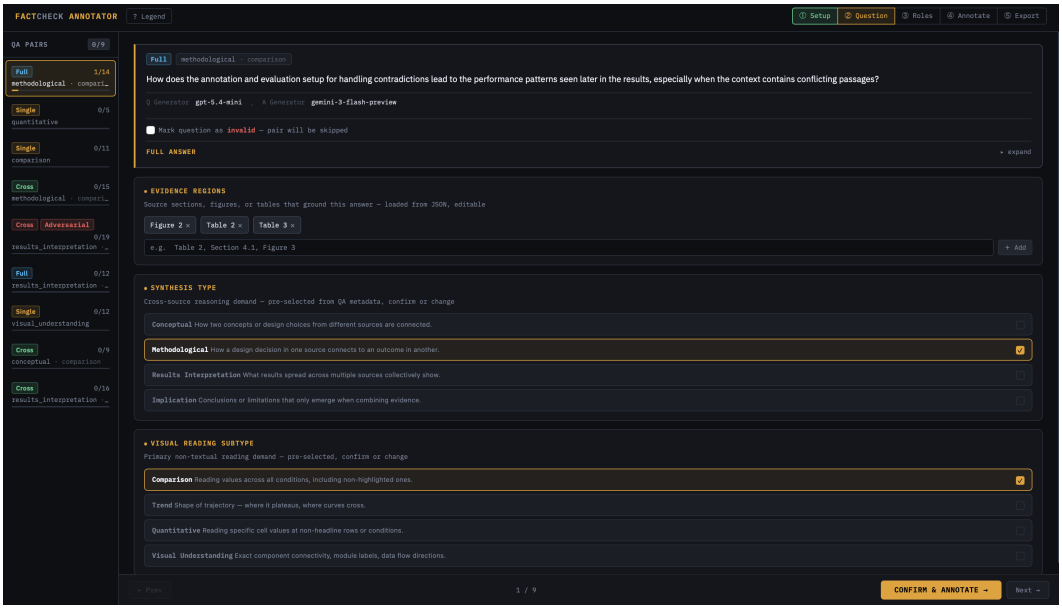


Figure 11: **Question validation.** Annotators verify that the question is answerable from the designated evidence or reject the QA pair. They also confirm or revise metadata recorded during generation, including the relevant evidence regions, synthesis type for cross-page and full-paper questions, and required visual-reading operation. Best viewed if zoomed.

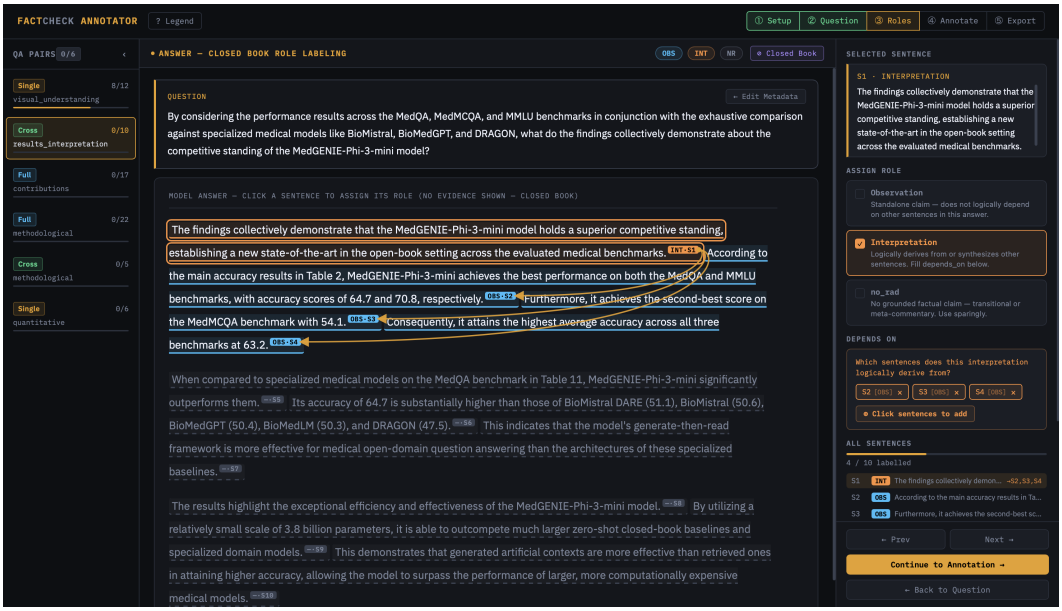


Figure 12: **Structural annotation.** Answer sentences are labeled as either observations, which report what the paper shows, or interpretations, which draw conclusions from other sentences. For each interpretation, annotators record its premise sentences. Here, S1 depends on observations S2–S4, forming part of the dependency graph used to derive INVALID SUPPORT. Best viewed if zoomed.

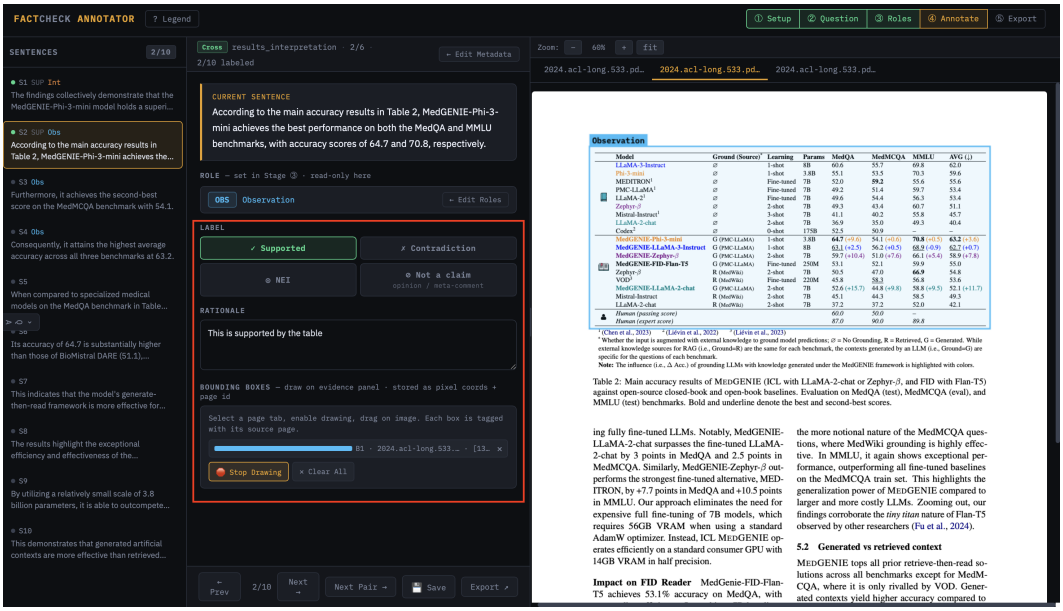


Figure 13: **Sentence grounding.** Annotators assign each sentence a grounding label (SUPPORTED, CONTRADICTED, NEI, or NOT A CLAIM) and provide a rationale. For supported or contradicted observations, they also draw a bounding box locating the relevant source evidence, shown here in blue over Table 2. The red outline highlighting the corresponding interface fields was added to the figure and is not part of the annotation tool. Best viewed if zoomed.

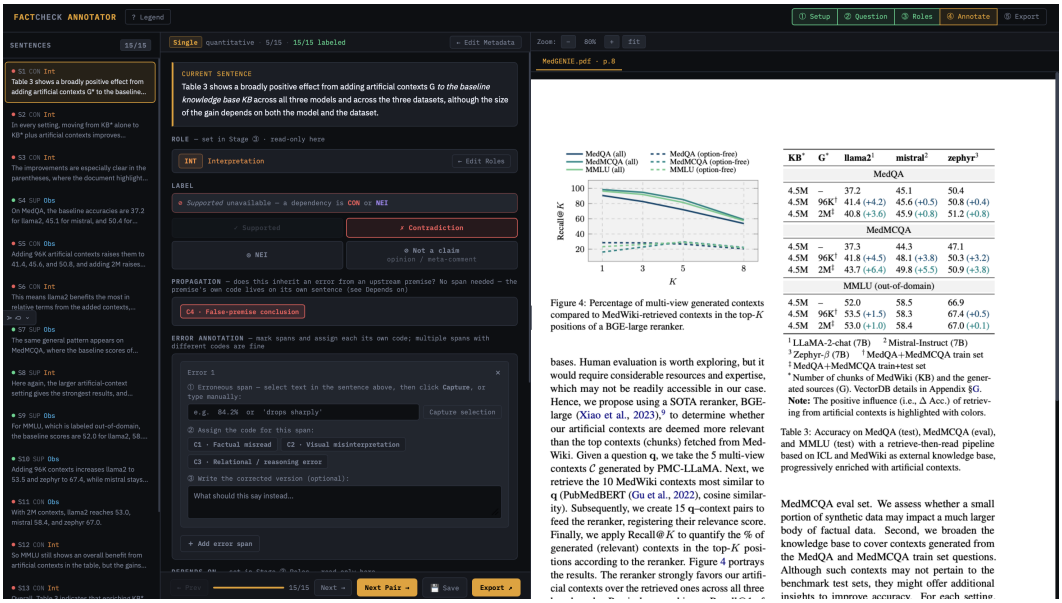


Figure 4: Percentage of multi-view generated contexts compared to MedWiki-retrieved contexts in the top-K positions of a BGE-large reranker.

KB*	G*	llama2[1]	mistral[2]	zephyr[3]
MedQA				
4.5M	–	37.2	45.1	50.4
4.5M	96K[‡]	41.4 (+4.2)	45.6 (+0.5)	50.8 (+0.4)
4.5M	2M[‡]	40.8 (+3.6)	45.9 (+0.8)	51.2 (+0.8)
MedMCQA				
4.5M	–	37.3	44.3	47.1
4.5M	96K[‡]	41.8 (+4.5)	48.1 (+3.8)	50.3 (+3.2)
4.5M	2M[‡]	43.7 (+6.4)	49.8 (+5.5)	50.9 (+3.8)
MMLU (out-of-domain)				
4.5M	–	52.0	58.5	66.9
4.5M	96K[‡]	53.5 (+1.5)	58.3	67.4 (+0.5)
4.5M	2M[‡]	53.0 (+1.0)	58.4	67.0 (+0.1)

[1] LLaMA-2-chat (7B) [2] Mistral-Instruct (7B)
[3] Zephyr-β (7B) [‡] MedQA+MedMCQA train set
[‡] MedQA+MedMCQA train+test set
* Number of chunks of MedWiki (KB) and the generated sources (G). VectorDB details in Appendix §G.
Note: The positive influence (i.e., Δ Acc.) of retrieving from artificial contexts is highlighted with colors.

Table 3: Accuracy on MedQA (test), MedMCQA (eval), and MMLU (test) with a retrieve-then-read pipeline based on ICL and MedWiki as external knowledge base, progressively enriched with artificial contexts.

bases. Human evaluation is worth exploring, but it would require considerable resources and expertise, which may not be readily accessible in our case. Hence, we propose using a SOTA reranker, BGE-large (Xiao et al., 2023),[9] to determine whether our artificial contexts are deemed more relevant than the top contexts (chunks) fetched from MedWiki. Given a question $\mathbf{q}$, we take the 5 multi-view contexts $\mathcal{C}$ generated by PMC-LLaMA. Next, we retrieve the 10 MedWiki contexts most similar to $\mathbf{q}$ (PubMedBERT (Gu et al., 2022), cosine similarity). Subsequently, we create 15 $\mathbf{q}$-context pairs to feed the reranker, registering their relevance score. Finally, we apply Recall@K to quantify the % of generated (relevant) contexts in the top-K positions according to the reranker. Figure 4 portrays the results. The reranker strongly favors our artificial contexts over the retrieved ones across all three

MedQA eval set. We assess whether a small portion of synthetic data may impact a much larger body of factual data. Second, we broaden the knowledge base to cover contexts generated from the MedQA and MedMCQA train set questions. Although such contexts may not pertain to the benchmark test sets, they might offer additional

Figure 14: **Error tagging and INVALID SUPPORT.** Annotators mark erroneous spans and assign each an error code. When an interpretation's own reasoning is sound but it depends on a premise that is itself CONTRADICTED or NEI, annotators tag it **C4** (false-premise conclusion) rather than an independent error, recording that its conclusion would hold if the premise did. This tag is what lets the pipeline map the interpretation from CONTRADICTED to INVALID SUPPORT instead of treating it as an unrelated error. Best viewed if zoomed.

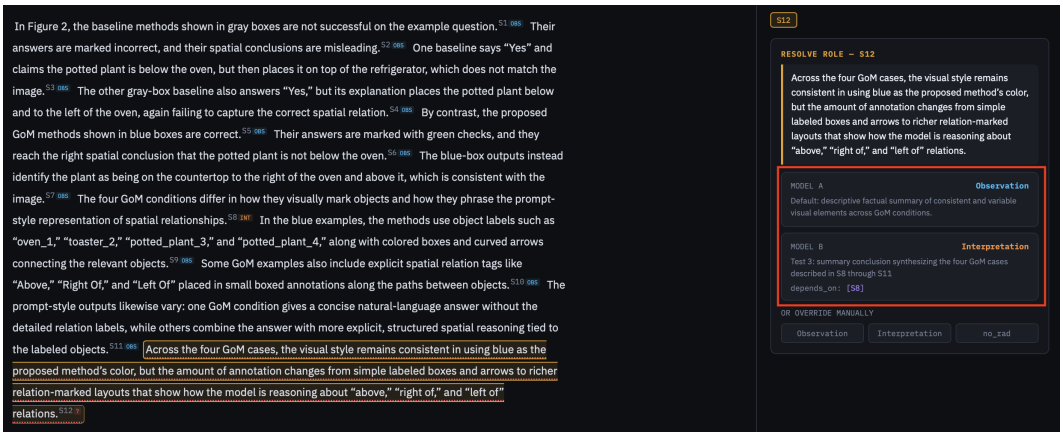


Figure 15: **Resolving a role disagreement.** When the two LLM pre-annotation runs assign different roles to a sentence (here S12, highlighted and marked "?" in the answer), the human annotator sees the sentence in context together with each run's role and rationale: one run treats it as a descriptive Observation, the other as an Interpretation that synthesizes S8–S11 and lists its premise. The annotator accepts one of the two roles or overrides it manually. *Note:* the red outline marks the two runs' proposals; it was added to the figure and is not part of the annotation tool. Best viewed if zoomed.

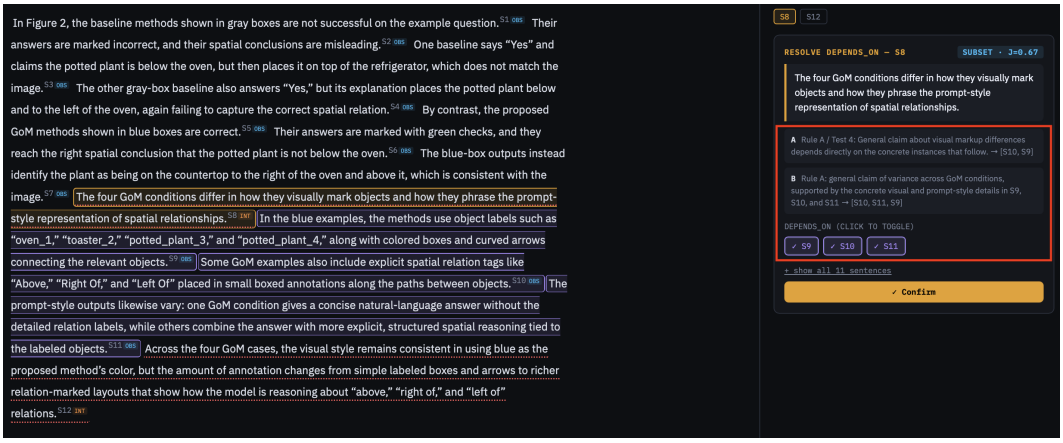


Figure 16: **Resolving a dependency disagreement.** When the two LLM runs agree that a sentence is an Interpretation but list different premises (here S8), the tool shows each run's premise set with its rationale, and the badge reports how the two sets relate: here the first run's premises are a subset of the second's (Jaccard $= 0.67$). Candidate premises, highlighted in the answer, can be toggled to define the final list, which the annotator then confirms. *Note:* the red outline marks the two runs' premise sets and the toggleable candidates; it was added to the figure and is not part of the annotation tool. Best viewed if zoomed.

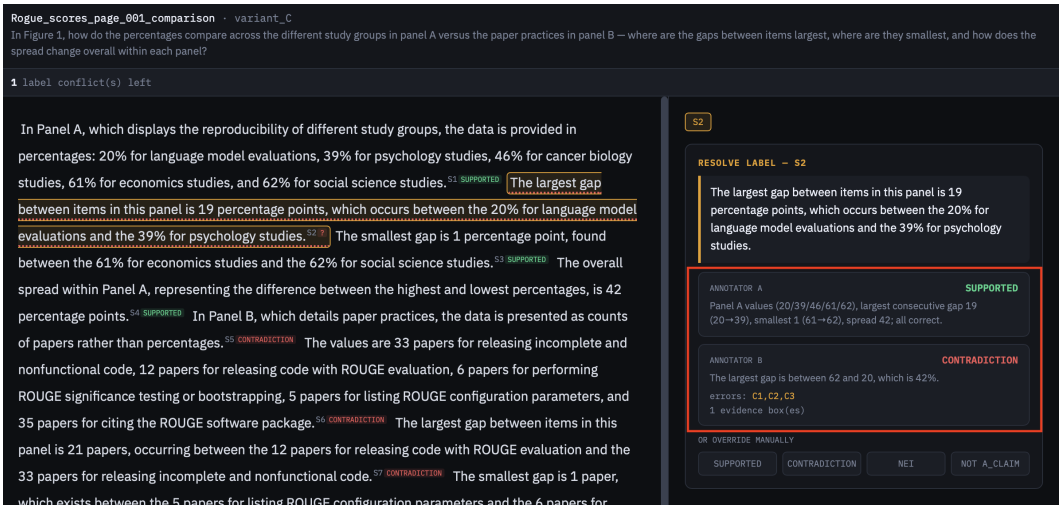


Figure 17: **Resolving a grounding disagreement.** The two human annotators label each sentence independently. When their labels differ (here S2, SUPPORTED vs. CONTRADICTED), the resolver shows the sentence in context with each annotator's label, rationale, error codes, and number of evidence boxes, and the pair settles the final label through discussion, selecting it in the tool or overriding it. *Note:* the red outline marks the two annotators' entries; it was added to the figure and is not part of the annotation tool. Best viewed if zoomed.