SIN-Bench: Tracing Native Evidence Chains in Long-Context Multimodal Scientific Interleaved Literature

Anonymous ACL submission

Abstract

Evaluating multimodal large language models (MLLMs) on long-form scientific papers requires measuring not only answer correctness, but also whether answers are supported by traceable cross-modal evidence. Existing answer-only metrics and synthetic "Needle-In-A-Haystack" tests mainly assess answer matching or isolated retrieval, leaving evidence-grounded support under-examined. We introduce "Fish-in-the-Ocean" (FITO), an evidence-chain evaluation lens for native scientific documents, and instantiate it with `SIN-Data`, a scientific interleaved corpus that preserves text–figure coupling. Built on `SIN-Data`, `SIN-Bench` comprises four evidence-auditing tasks: evidence discovery, hypothesis verification, grounded QA, and evidence-anchored synthesis. We further propose "No Evidence, No Score", which credits predictions only when linked to verifiable anchors and diagnoses evidence quality through matching, relevance, and logic. Experiments on nine MLLMs show a clear correctness–grounding gap: Gemini-3-pro obtains the best average overall score (0.566), while GPT-5 reaches the highest SIN-QA answer accuracy (0.767) but lower evidence-aligned scores.

1 Introduction

Long-form multimodal scientific literature provides a demanding testbed for expert-level document reasoning (Xia et al., 2024; Chen et al., 2024; Bai et al., 2025a). Scientific papers contain dense textual arguments, figures, tables, and cross-sectional dependencies, requiring readers to connect dispersed evidence rather than retrieve isolated facts. Recent advances in multimodal large language models (MLLMs) (Bai et al., 2025b; OpenAI, 2025; DeepMind, 2025; xAI, 2025), especially in long-context processing and multimodal alignment, make such document-level analysis increasingly feasible. However, existing evaluations often

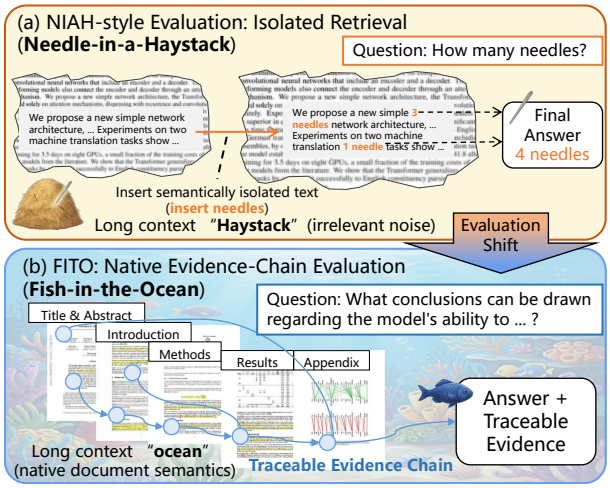

Figure 1: Comparison of long-context multimodal evaluation paradigms. (a) NIAH-style tests insert artificial "needles" into long contexts, primarily probing isolated retrieval and answer matching. (b) The FITO paradigm evaluates native scientific documents, where relevant evidence is distributed across sections and modalities. Rather than assessing answer correctness alone, FITO examines whether the answer is supported by a traceable cross-modal evidence chain.

emphasize answer correctness or isolated retrieval, leaving under-specified whether a prediction is supported by traceable evidence within the document. For scientific document understanding, this distinction is critical: *a plausible answer is valid only when explicitly grounded in text-figure evidence and structured into an auditable reasoning chain.*

Despite rapid progress in long-context evaluation, current benchmarks remain only partially aligned with evidence-intensive scientific document reasoning (Fu et al., 2024). This gap appears in both task construction and evaluation protocols. First, for **task design**, as shown in Fig. 1 (a), NIAH-style evaluations insert artificially constructed and semantically isolated "needles" into long contexts (Wang et al., 2024b). Such synthetic settings are useful for probing retrieval boundaries, but they do not directly test reasoning over native

evidence that is distributed across text, figures, tables, and appendices. For instance, interpreting a result figure often requires linking the reported observations to experimental setups, metric definitions, or ablation conditions described in distant sections. Second, for **evaluation metrics**, answer accuracy alone cannot determine whether a prediction is supported by verifiable document evidence (Mathew et al., 2021; Xia et al., 2024). A model may produce a correct or plausible answer while omitting key evidence, citing irrelevant anchors, or organizing the supporting evidence incorrectly.

To address this limitation, we shift the evaluation focus from answer-only correctness to evidence-supported prediction. As shown in Fig. 1 (b), we introduce "Fish-in-the-Ocean" (FITO), an evidence-chain evaluation lens for native scientific documents. In this lens, the "ocean" denotes the full semantic environment of a scientific paper, and the "fish" denote naturally occurring evidence units rather than artificially inserted facts. The goal is not to prescribe the model's internal reasoning process, but to require an external and auditable provenance interface. That is, a prediction should be accompanied by cross-modal evidence anchors and text spans that can be checked for matching, relevance, and logical consistency. Accordingly, evaluating scientific document reasoning requires assessing whether a model can identify useful evidence distributed across sections and modalities, and organize it into a traceable support chain.

To instantiate the Fish-in-the-Ocean (FITO) paradigm, we build `SIN-Data`, a unified **S**cientific **IN**terleaved data infrastructure for long-form scientific documents. We aggregate papers from arXiv and PMC, and convert each document into a unified interleaved representation with traceable text spans and visual anchors, preserving the semantic coupling between prose and figures. Following the arXiv taxonomy, `SIN-Data` covers more than 12 primary disciplines and over 80 subfields, providing a broad substrate for scientific document evaluation. Building on `SIN-Data`, we develop a scalable pipeline to construct `SIN-Bench`. Each instance consists of a query, answer(s), and an evidence chain grounded to native document anchors. Rather than reproducing the full scientific research workflow, `SIN-Bench` operationalizes document-level evidence auditing through four tasks: (1) SIN-Find for evidence discovery, (2) SIN-Verify for evidence sufficiency verification, (3) SIN-QA for grounded question answering, and (4) SIN-Summary for evidence-anchored synthesis. To move beyond isolated answer-only evaluation, we adopt the **"No Evidence, No Score"** protocol: predictions receive evidence-aligned credit only when they are supported by verifiable anchors. Beyond answer accuracy, we diagnose evidence quality along three dimensions: Matching, Relevance, and Logic.

Empirically, our evaluation of nine widely used MLLMs shows that evidence grounding remains a major bottleneck in long-form scientific document reasoning. Gemini-3-pro achieves the best average overall score (0.566), whereas GPT-5 obtains the highest SIN-QA answer accuracy (0.767) but lower evidence-aligned scores on SIN-Find and SIN-QA. This result highlights a correctness–grounding gap: accurate answers do not necessarily come with matched, relevant, and logically ordered evidence support. Several open-weight models also struggle with structured evidence output under strict schemas.

In summary, the contributions are as follows:

- We introduce FITO, an evidence-chain evaluation lens for traceable support in long-context multimodal scientific documents.
- We build `SIN-Data` and `SIN-Bench`, covering evidence discovery, sufficiency verification, grounded QA, and evidence-anchored synthesis.
- We propose "No Evidence, No Score" with matching, relevance, and logic metrics to evaluate evidence grounding beyond answer accuracy.

2 Related Work

MLLM Long-Context Understanding Tasks. Recent long-context MLLM benchmarks mainly probe retrieval over extended inputs. Early benchmarks, such as MM-NIAH (Wang et al., 2024b) and MMLongCite (Zhou et al., 2025), adopt Needle-In-A-Haystack-style settings that insert synthetic text or image needles into long contexts to test recall and citation behavior. These methods effectively quantify context capacity, but their evidence units are often semantically isolated from the surrounding document. Later benchmarks (Ma et al., 2024; Deng et al., 2025; Wang et al., 2025) improve realism by using web pages, multi-page reports, or real documents. However, they still largely emphasize answer localization, page-level perception, or single-point extraction, leaving limited diagnosis of ordered cross-modal evidence support. Our work utilizes Fish-in-the-Ocean (FITO) as an evidence-chain evaluation lens and instantiates

it with SIN-Bench for scientific interleaved documents.

Evaluating Scientific Reasoning in MLLMs. Scientific literature contains dense multimodal content, explicit figure–text references, and long-range dependencies, making it a useful domain for evidence-grounded document reasoning. Prior efforts convert PDF and LaTeX documents into machine-readable formats, providing important infrastructure for scientific document understanding (Wang et al., 2024a; Zhong et al., 2019; Xia et al., 2024). Recent studies further move toward higher-level tasks: M-DocSum (Yan et al., 2025) studies reference-grounded interleaved summarization, while MMIE (Xia et al., 2025) targets information extraction in knowledge-intensive domains. However, common metrics such as ROUGE or answer accuracy do not directly audit whether predictions are supported by verifiable document evidence. A correct answer may still omit necessary evidence, cite irrelevant anchors, or arrange support in an inconsistent order. Inspired by recent work on deep research (Jiang et al., 2024; Yang et al., 2025b), we introduce "No Evidence, No Score" and diagnose grounding quality through Matching, Relevance, and Logic.

3 The Fish-in-the-Ocean Paradigm

Rather than inserting synthetic "needles" into long contexts (Wang et al., 2024b), Fish-in-the-Ocean (FITO) focuses on native evidence that already exists within long-form multimodal documents. In this view, a scientific paper is treated as an "ocean", and the evidence units that support a conclusion are treated as "fish". FITO is defined by three document properties: (i) **Nativeness**, where evidence is inherently present rather than artificially injected; (ii) **Interconnectivity**, where support spans sections and modalities such as text, figures, and tables; and (iii) **Long-range Dependency**, where conclusions often require linking distant components such as methods, results, and appendices. The goal is therefore not to test whether a model can retrieve an isolated fact, but whether its prediction can be supported by a traceable chain of evidence.

To evaluate this objective, we move beyond simple QA likelihood maximization, $P(A|D,Q)$, where A, D, and Q denote the answer, document, and query, respectively. Instead, FITO evaluates the answer together with an explicit supporting evidence chain E:

$$P(A,E|D,Q) = P(E|D,Q) \cdot P(A|E,D,Q). \quad (1)$$

Here, E is an external provenance interface for evaluation, rather than a claim about the model's internal reasoning trace. The factorization decomposes the evaluation into evidence identification, $P(E|D,Q)$, and evidence-supported answering, $P(A|E,D,Q)$.

4 SIN-Bench: Evidence-based Evaluation

To operationalize the FITO paradigm, we introduce SIN-Bench, an evidence-based benchmark for long-form multimodal scientific documents. Scientific papers provide a natural substrate because they contain dense multimodal evidence, explicit figure–text links, and long-range dependencies. We build a semi-automated framework that converts raw scientific corpora into structured evidence-auditing instances. As shown in Fig. 2, the framework includes a Scientific INterleaved (SIN) representation (Sec. 4.1), evidence-based tasks (Sec. 4.2), evidence-driven metrics (Sec. 4.3), and a scalable construction pipeline (Sec. 4.4).

4.1 SIN-Data Infrastructure

To support evidence-grounded evaluation over long-form multimodal documents, we build SIN-Data, a pipeline that converts scientific papers into a Scientific INterleaved (SIN) format. The goal is to provide a machine-readable representation that preserves traceable text spans, visual anchors, and their semantic links, rather than to replace the original human-readable PDF or source files. As shown in Fig. 2 (a), inspired by recent work (Wang et al., 2024a), our pipeline processes 50k raw source packages from arXiv[1] and PubMed Central (PMC)[2] through three stages.

Stage 1: Element Parsing. We first separate content extraction from document presentation. For arXiv papers, source LaTeX is compiled into responsive HTML with Engrafo, from which we parse text and recover image elements from the DOM tree. For PMC papers, we parse JATS XML files into structured JSON with s2orc-doc2json, preserving tables and citation links while removing decorative artifacts.

Stage 2: Semantic-First Formatting. We then convert structured documents into Interleaved Mark-

[1] https://arxiv.org/

[2] https://www.ncbi.nlm.nih.gov/pmc/

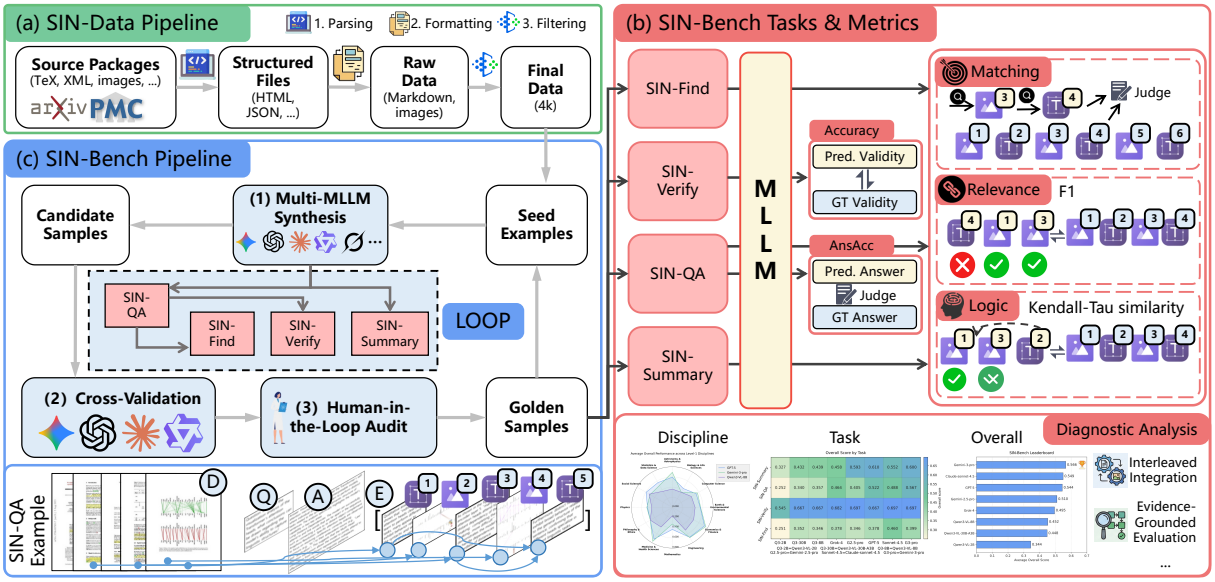

Figure 2: Framework of SIN-Bench. **(a) SIN-Data infrastructure**: source packages are parsed, formatted, and filtered into a Scientific Interleaved representation with text spans and visual anchors. **(b) Tasks and metrics**: four evidence tasks are evaluated under the "No Evidence, No Score" protocol, with evidence chains assessed by Matching, Relevance, and Logic. **(c) Construction pipeline**: candidate samples are generated through multi-MLLM synthesis, filtered by cross-validation, audited by humans, and converted into golden samples.

down. To reduce layout-induced separation between figures and their relevant textual context, we use a citation-driven injection strategy. Each visual component is assigned a unique identifier $\langle x_k \rangle$ and inserted near the paragraph of its first semantic citation, such as before "Figure 3 shows that...". This representation keeps visual evidence close to the prose that invokes it, enabling evidence chains to reference both image anchors and text spans. We also compute quality signals, including the number of text–image interleaved segments and token density, for subsequent filtering.

Stage 3: Quality Filtering. Finally, we filter documents with sparse visual context, broken citation links, or insufficient interleaving quality. We further annotate documents according to the official arXiv taxonomy to maintain broad coverage across scientific domains.

This process yields $4,000$ curated documents. They cover 12 top-level disciplines and over 80 subfields. Additional construction details are provided in App. A.1.

4.2 Evidence-based Hierarchical Tasks

To instantiate FITO at the task level, SIN-Bench defines four evidence-auditing tasks: evidence discovery, evidence sufficiency verification, grounded QA, and evidence-anchored synthesis. Each instance is built on a single document D and uses a unified evidence-chain schema. We represent an evidence chain as $E = [(v_i, t_i)]_{i=1}^{K}$, where v_i is a native visual anchor, such as a figure or table identifier $\langle x_k \rangle$, and t_i is a supporting text span or concise evidence note. This schema provides an external provenance interface shared across tasks, rather than a claim about the model's internal reasoning trace.

SIN-Find (Evidence Discovery). Given a document D and a query Q, the model outputs an ordered evidence chain:

$$E = f_{\text{find}}(D, Q). \tag{2}$$

The query is designed to require semantic evidence search, such as identifying relevant factors or comparing methods, rather than keyword matching. The task therefore evaluates whether the model can locate cross-sectional or cross-modal support and organize it into a traceable chain.

SIN-Verify (Evidence Sufficiency Verification). Given a document D, query Q, answer A, and candidate evidence chain E, the model predicts whether the evidence sufficiently supports the answer:

$$y = f_{\text{verify}}(D, Q, A, E), \quad y \in \{0, 1\}. \tag{3}$$

To test sufficiency rather than superficial overlap, we construct negative samples by perturbing evidence chains, such as omitting necessary premises or mismatching conditions.

SIN-QA (Grounded Question Answering). Given (D, Q), the model outputs both an answer and its supporting evidence chain:

$$(A, E) = f_{\text{qa}}(D, Q). \tag{4}$$

Here, the answer is evaluated together with its auditable support. This setting does not assume that E is the model's internal reasoning process; instead, it makes unsupported or weakly supported answers diagnosable under the evidence metrics.

SIN-Summary (Evidence-Anchored Synthesis). Given a document D, the model generates an evidence-anchored summary consisting of multiple claims, each paired with supporting evidence:

$$\hat{S} = \left\{(a_j, E_j)\right\}_{j=1}^{J} = f_{\text{sum}}(D). \tag{5}$$

Here, each a_j denotes a summary claim, and E_j denotes the evidence chain that supports this specific claim. This claim-wise design avoids treating the summary as a monolithic free-form output, and instead makes each generated statement individually auditable. Equivalently, the summary-level answer and evidence can be written as

$$A = \{a_j\}_{j=1}^{J}, \qquad E = \bigcup_{j=1}^{J} E_j. \tag{6}$$

This task evaluates document-level synthesis with local verifiability: the model must integrate information across the document while grounding each summary claim to explicit evidence anchors.

4.3 Metrics: No Evidence, No Score

SIN-Bench follows a "No Evidence, No Score" protocol to evaluate whether model predictions are supported by verifiable document evidence. As shown in Fig. 2 (b), the evaluation has two layers. For tasks that require evidence-chain generation, i.e., SIN-Find, SIN-QA, and SIN-Summary, we first assess evidence quality with a shared protocol. SIN-QA additionally evaluates answer correctness, while SIN-Verify is evaluated as a binary classification task.

Interleaved Evidence Chain. We represent the ground-truth evidence chain as $E^\star = [(v_i^\star, t_i^\star)]_{i=1}^{K^\star}$ and the predicted chain as $\hat{E} = [(\hat{v}_i, \hat{t}_i)]_{i=1}^{\hat{K}}$. Here, v_i denotes a native visual anchor, such as a figure or table identifier, and t_i denotes the associated supporting text span or evidence note. The pair (v_i, t_i) is treated as the minimal evidence unit for evaluation.

For SIN-Find, SIN-QA, and SIN-Summary, we evaluate evidence-chain quality with three metrics: Matching, Relevance, and Logic. Matching uses an LLM judge to assess semantic consistency between predicted and reference evidence text after anchor matching, while Relevance and Logic are computed deterministically. We use Qwen3-8B (Yang et al., 2025a) as the judge and validate its agreement with human expert ratings in App. C.3.

(M) Matching. Let $V^\star = \{v_i^\star\}_{i=1}^{K^\star}$ denote the set of ground-truth visual anchors. We define the hit set as $H = \{i \in [1, \hat{K}] \mid \hat{v}_i \in V^\star\}$. For each $i \in H$, we find the corresponding ground-truth index $j(i)$ such that $v_{j(i)}^\star = \hat{v}_i$. The LLM judge then scores the semantic consistency between $\hat{t}_i$ and $t_{j(i)}^\star$ as $s_i \in \{0, 1, 2, 3\}$. We normalize it as $\tilde{s}_i = s_i/3$ and define:

$$\text{Match} = \begin{cases} \frac{1}{|H|} \sum_{i \in H} \tilde{s}_i, & |H| > 0, \\ 0, & |H| = 0. \end{cases} \tag{7}$$

(R) Relevance. Relevance measures whether the predicted evidence covers the required evidence units. We define a predicted unit as correct if its visual anchor matches a ground-truth anchor and its text is semantically consistent: $\mathbb{I}_i = \mathbf{1}[\hat{v}_i \in V^\star \wedge \tilde{s}_i \geq \tau]$, where $\tau = 2/3$. Let $TP = \sum_{i=1}^{\hat{K}} \mathbb{I}_i$. We compute precision, recall, and F1 as:

$$\text{Prec} = \frac{TP}{\hat{K}}, \quad \text{Rec} = \frac{TP}{K^\star}, \quad \text{Rel} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}. \tag{8}$$

When $\text{Prec} + \text{Rec} = 0$, we set $\text{Rel} = 0$.

(L) Logic. Logic measures whether the matched evidence units follow the reference order. For the matched anchors, let $\hat{\pi}$ and $\pi^\star$ denote their relative orders in the prediction and ground truth, respectively, and let $K = |H|$. We compute Kendall–Tau similarity (Kendall, 1948) as:

$$\tau = 1 - \frac{2 \cdot \text{inv}(\hat{\pi}, \pi^\star)}{K(K-1)/2}, \quad \text{Logic} = \frac{\tau + 1}{2}. \tag{9}$$

When $K < 2$, we set $\text{Logic} = 1$ if $K = 1$ and 0 if $K = 0$.

SIN-QA: Answer Accuracy. For the predicted answer $\hat{A}$, the LLM judge assigns a semantic correctness score $g \in \{0, 1, 2, 3\}$. We normalize it as $\text{AnsAcc} = g/3$.

SIN-Verify: Accuracy. SIN-Verify is a binary classification task. Given T samples, we compute:

$$\text{Score}_{\text{VERIFY}} = \text{Acc} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}[\hat{y}_i = y_i]. \tag{10}$$

For each task, the overall score is the arithmetic mean of its applicable metrics: Matching, Relevance, and Logic for SIN-Find and SIN-Summary; Answer Accuracy, Matching, Relevance, and Logic for SIN-QA; and Accuracy for SIN-Verify.

4.4 Benchmark Construction Pipeline

Building on the interleaved documents in `SIN-Data`, we construct `SIN-Bench` with an iterative human–model collaborative pipeline. The pipeline aims to scale evidence-chain annotation while filtering out samples with factual errors, weak evidence support, or invalid output structure. As shown in Fig. 2 (c), it consists of four steps.

Seed Examples. We manually create a small set of seed examples for each task. These examples define the task-specific input–output schema, evidence-chain format, and expected granularity of supporting evidence.

(1) Multi-MLLM Synthesis. Given a document D, multiple MLLMs generate candidate instances under the unified evidence-chain schema. We use SIN-QA and SIN-Summary as the main generation pivots. For SIN-QA, the model jointly generates a query Q, an answer A, and a draft evidence chain E, reducing semantic drift between the answer and its support. SIN-Find is derived by asking the model to recover the evidence chain for a given query, while SIN-Verify is derived by pairing answers with either valid evidence or perturbed evidence. Perturbations include omitting key premises, mismatching conditions, shuffling evidence order, or using superficially plausible but insufficient support. For SIN-Summary, we use cite-as-you-write prompting so that each summary claim is paired with evidence anchors. This step produces a pool of raw candidate samples.

(2) Cross-Validation. To reduce synthesis noise, three strong MLLMs independently score each candidate on a 1–5 scale. The validation covers three aspects: question validity, answer correctness, and evidence sufficiency/consistency. We retain a candidate only when it passes majority voting and receives scores of at least 4 on all validation dimensions. This stage removes many candidates with ambiguous questions, unsupported answers, or incomplete evidence chains.

(3) Human-in-the-Loop Audit. Candidates that pass cross-validation are further reviewed by human auditors. The audit focuses on five criteria: anchored locatability, rigor of support, factual consistency, structural compliance, and discriminability. Auditors verify whether each evidence anchor can be located in the document, whether the associated text span supports the claim, and whether negative examples are genuinely insufficient or inconsistent. Residual factual errors, formatting deviations, and weak support relations are corrected or removed, yielding the final golden samples.

Iterative Loop. Audited golden samples are fed back as updated seeds for later synthesis rounds. This loop gradually stabilizes the output schema and improves the quality of subsequent candidates.

Starting from $4,000$ curated scientific documents, the pipeline produces approximately $3,200$ raw candidate samples. After cross-validation and human auditing, the final `SIN-Bench` contains 490 golden samples, including 159 for SIN-Find, 158 for SIN-QA, 89 for SIN-Summary, and 84 for SIN-Verify. Further details are provided in App. A.2.

5 Experiments

5.1 Experimental Setup

We evaluate 9 representative MLLMs, including 6 proprietary models and 3 open-weight models. The proprietary models include Gemini-3.1-pro (Google DeepMind, 2026), Gemini-3-pro-preview (DeepMind, 2025), Gemini-2.5-pro (Team, 2025), GPT-5 (OpenAI, 2025), Grok 4 (xAI, 2025), and Claude-sonnet-4.5 (Anthropic, 2025). For open-weight models, we select the Qwen3-VL series with Thinking mode, covering both dense and MoE architectures: Qwen3-VL-2B, Qwen3-VL-8B, and Qwen3-VL-30B-A3B (Bai et al., 2025b). All models are evaluated with the same task prompts, output schemas, and metric implementation. Details on model selection and evaluation settings are provided in App. C.1 and C.2.

5.2 Main Results

Overall Performance. As reported in Table 1, Gemini-3-pro achieves the highest Avg. Overall score (0.566), followed by Claude-sonnet-4.5 (0.549) and GPT-5 (0.544). GPT-5 obtains the best SIN-QA answer accuracy (0.767), but its evidence-aligned overall scores on SIN-Find and SIN-QA ($0.378/0.522$) remain lower than those of Gemini-3-pro ($0.399/0.567$). This result highlights a correctness–grounding gap: higher answer accuracy does not necessarily imply stronger matched, relevant, and logically ordered evidence support. Gemini-3.1-pro is competitive but lower in Avg. Overall (0.528), showing no uniform gain under

Model	SIN-Find				SIN-Verify		SIN-QA					SIN-Summary				Avg.
	L	M	R	**O**	Acc	**O**	AnsAcc	L	M	R	**O**	L	M	R	**O**	**Overall**
Proprietary models																
Gemini-3-pro	0.524	0.381	0.294	0.399	0.697	0.697	0.726	0.569	**0.521**	**0.454**	**0.567**	0.784	0.546	0.471	0.600	**0.566**
Claude-sonnet-4.5	0.512	**0.490**	**0.378**	**0.460**	0.697	0.697	0.708	0.550	0.361	0.333	0.488	0.765	0.493	0.399	0.552	0.549
GPT-5	0.534	0.333	0.268	0.378	0.667	0.667	**0.767**	**0.578**	0.406	0.336	0.522	**0.808**	0.538	**0.483**	**0.610**	0.544
Gemini-3.1-pro	0.510	0.367	0.242	0.373	**0.710**	**0.710**	0.669	0.516	0.366	0.276	0.457	0.779	0.505	0.430	0.572	0.528
Gemini-2.5-pro	**0.536**	0.300	0.203	0.346	0.697	0.697	0.684	0.522	0.240	0.197	0.405	0.762	**0.547**	0.472	0.593	0.510
Grok-4	0.529	0.339	0.266	0.378	0.682	0.682	0.688	0.522	0.371	0.289	0.464	0.699	0.399	0.275	0.458	0.495
Open-weight models																
Qwen3-VL-8B	0.500	0.293	**0.245**	0.346	**0.667**	**0.667**	**0.476**	0.514	**0.341**	**0.234**	**0.357**	**0.683**	**0.359**	**0.276**	**0.439**	**0.452**
Qwen3-VL-30B-A3B	**0.522**	**0.307**	0.226	**0.352**	**0.667**	**0.667**	0.441	**0.528**	0.321	0.214	0.340	0.676	0.345	0.275	0.432	0.448
Qwen3-VL-2B	0.485	0.160	0.108	0.251	0.546	0.546	0.337	0.510	0.243	0.164	0.252	0.589	0.228	0.165	0.327	0.344

Table 1: Main results on SIN-Bench across representative proprietary and open-weight MLLMs. We report logic (**L**), matching (**M**), relevance (**R**), answer accuracy (**AnsAcc**), and verification accuracy (**Acc**). **O** denotes the task-specific overall score, and **Avg. Overall** is the arithmetic mean of task-level overall scores. Bold numbers indicate the best score within each model group.

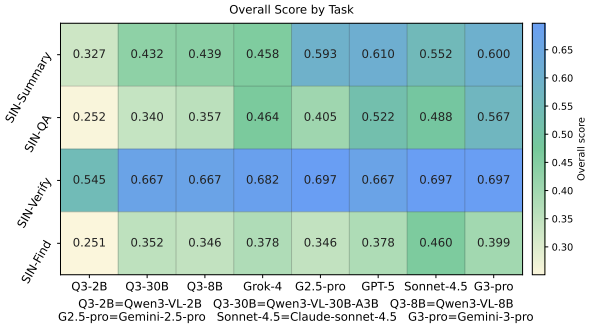

Figure 3: Task-level overall performance heatmap across selected models. Darker cells imply higher scores.

evidence-chain evaluation. Among open-weight models, Qwen3-VL-8B achieves the best Avg. Overall score (0.452), slightly outperforming the larger Qwen3-VL-30B-A3B model (0.448). This suggests that evidence-grounded scientific document evaluation is not determined by parameter count alone. We also observe that structured evidence output remains challenging for open-weight models under strict output schemas, especially in tasks requiring precise anchor–text alignment. More analysis is provided in App. D.

Task-Level Analysis. As shown in Fig. 3, model rankings vary across tasks, indicating that SIN-Bench diagnoses different aspects of evidence-grounded document reasoning.

In **SIN-Find**, Claude-sonnet-4.5 achieves the highest overall score (0.460), mainly due to stronger Matching and Relevance. Gemini-2.5-pro obtains the best Logic score (0.536), suggesting that anchor ordering and anchor selection can favor different models.

In **SIN-Verify**, most strong models obtain similar standard accuracy in Table 1, but Table 2 shows a sharp drop under hard negatives. For example,

Setting	Gemini-3-pro	GPT-5	Qwen3-VL-8B
Easy Negatives	1.000	1.000	1.000
Hard Negatives	**0.250**	0.208	0.044

Table 2: Accuracy on SIN-Verify under different negative-sampling settings. Hard negatives, which use near-miss evidence, lead to near-chance performance.

GPT-5 drops from 1.000 to 0.208, and Qwen3-VL-8B drops from 1.000 to 0.044. This indicates that near-miss evidence remains difficult to reject, even when models handle easy negatives reliably. Detailed experiments appear in App. D.2.

In **SIN-QA**, Gemini-3-pro achieves the best overall score (0.567), whereas GPT-5 achieves the best AnsAcc (0.767). This divergence underscores that answer correctness does not guarantee evidence quality; correct predictions can still stem from incomplete or misaligned evidence.

In **SIN-Summary**, GPT-5 achieves the best overall score (0.610), with the highest Logic and Relevance scores. This suggests that current strong models can often follow the coarse structure of scientific writing, but precise evidence anchoring remains necessary to evaluate whether each summary claim is locally supported.

5.3 Discussion

Importance of Interleaved Input. Fig. 5 shows that preserving the interleaved structure of scientific papers yields clear gains over a separated layout (images→text), improving SIN-QA by +0.102 and SIN-Summary by +0.129 for Gemini-3-pro. Modality ablations further show "Interleaved" > "Text-only (captions)" > "Image-only (rendered pages)". This suggests that captions retain coarse semantic cues but miss fine-grained visual evi-

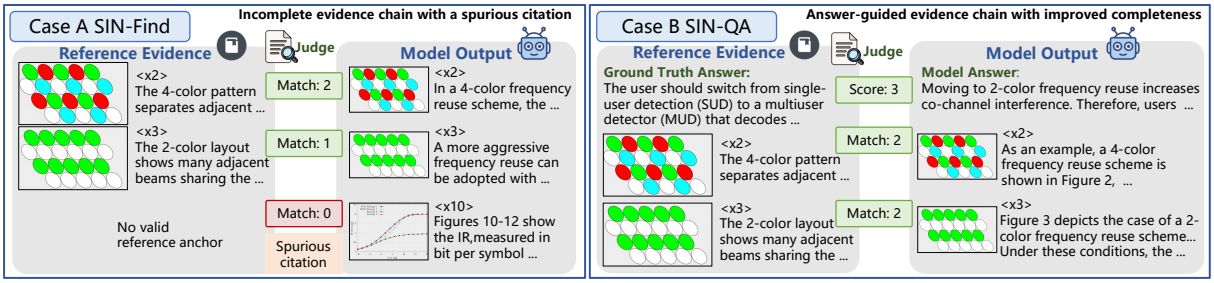

Figure 4: Qualitative example by Gemini-3-pro in SIN-Find and SIN-QA on the same document context.

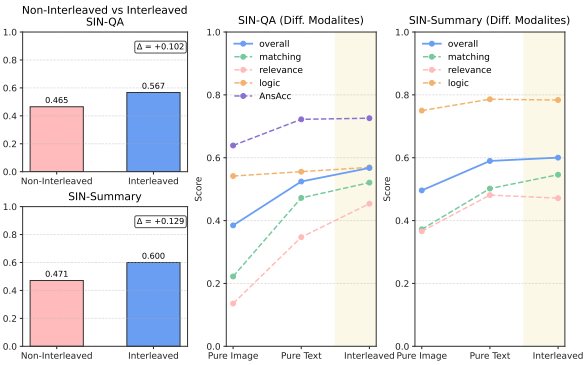

Figure 5: Impact of interleaved input and modality access on SIN-QA and SIN-Summary (Gemini-3-pro).

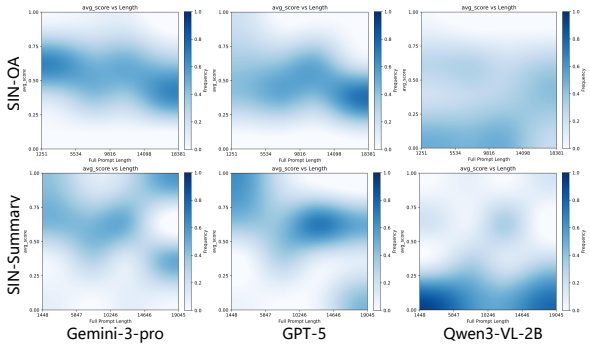

Figure 6: The score density distribution across varying input token lengths for SIN-QA and SIN-Summary.

dence, while raw visuals are most useful when locally grounded by adjacent text.

Effect of Evidence-Chain Output. We examine whether requiring an explicit evidence chain affects answer quality by comparing Gemini-3-pro on SIN-QA with and without evidence-chain output. Answer accuracy increases from 0.694 to 0.726 when evidence output is required. This suggests an external provenance requirement can help align answers with document evidence. However, evidence-chain output does not guarantee faithful support; unsupported or weakly supported claims still need to be diagnosed by Matching, Relevance, and Logic.

Impact of Text Length. Fig. 6 analyzes score distributions for input contexts exceeding 19k text tokens, with further details in App. D.3. Gemini-3-pro and GPT-5 remain relatively stable across long inputs, although GPT-5 shows tail-end degradation on SIN-QA. Qwen3-VL-2B is less stable, with many SIN-Summary samples concentrated near low scores. These results suggest long-context length interacts with both model scale and task format, especially for evidence-anchored synthesis.

Error Analysis. As shown in Fig. 4, our qualitative review identifies two common failure modes: (1) *Information Deficiency*, where models omit necessary premises or intermediate evidence units; and (2) *Spurious Grounding*, where models cite irrelevant or weakly related anchors, leading to "shotgun citations" and lower precision. We also observe that, under the same document context, SIN-QA responses tend to include more complete evidence chains than standalone SIN-Find outputs. This suggests that the target answer can serve as a semantic anchor for evidence selection, but it does not guarantee sufficient or logically ordered support. Further analysis is provided in App. E.

6 Conclusion

In this work, we introduce the "Fish-in-the-Ocean" paradigm and instantiate it with SIN-Bench, an evidence-based benchmark for long-context, interleaved scientific document reasoning. By combining four evidence-auditing tasks with the "No Evidence, No Score" protocol, SIN-Bench evaluates whether model predictions are supported by matched, relevant, and logically ordered cross-modal evidence. Experiments on representative MLLMs reveal a clear correctness–grounding gap: high answer accuracy does not necessarily imply traceable evidence support. Our analysis further shows that near-miss evidence verification and structured evidence output remain challenging for current models. These findings highlight the need to evaluate scientific document understanding beyond answer correctness with verifiable evidence.

Limitations

SIN-Bench has several limitations. First, model coverage is constrained by current MLLM input interfaces. Although we evaluate strong generalist models such as GPT-5, Gemini-3-pro, and Gemini-3.1-pro, many domain-specialized scientific models do not yet support long-context interleaved text–figure inputs. To facilitate future expansion, we will release evaluation scripts and data-construction code, subject to the licensing constraints of the source documents.

Second, the audited golden split is moderate in scale. This reflects the high cost of evidence-chain supervision: each instance requires cross-model validation and human auditing of evidence anchors, support relations, factual consistency, schema compliance, and discriminability. As summarized in Table 6, real long-document and scientific multimodal benchmarks are often limited to hundreds of documents due to parsing, grounding, and validation costs. SIN-Bench contains 231 documents and 490 golden instances, while further requiring explicit evidence-chain output and evidence-quality scoring. We therefore position it as an audited diagnostic benchmark rather than a large-scale statistical testbed. In addition, although we validate the automatic judge against expert annotations, we do not report an independent human task-performance baseline across all tasks; model scores should thus be interpreted as relative diagnostic results under our evidence-grounded protocol.

Finally, strict filtering improves parsing reliability but reduces data utilization. Documents with minor layout defects, broken citation links, or imperfect interleaving may be excluded even when their scientific content is valuable. Our pipeline exposes configurable filtering thresholds and voting mechanisms, allowing future users to adjust the trade-off between data purity and scale for different evaluation settings.

Ethical considerations

All documents in SIN-Data are collected from open-access repositories, including arXiv and PubMed Central, and are processed under their applicable licenses, such as CC-BY 4.0 when available. We release derived benchmark artifacts with attention to source-document licensing and attribution requirements. A potential misuse risk is that evidence-chain generation could be exploited by paper mills or other actors to produce fabricated scientific content with seemingly plausible provenance (Kendall and da Silva, 2024). Our intention is the opposite: the "No Evidence, No Score" protocol is designed to make scientific claims more auditable by requiring verifiable anchors. We therefore encourage future work to use evidence-chain auditing for detecting unsupported or fabricated claims, and to maintain transparent reporting of data sources, prompts, and evaluation scripts.

References

AI at Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Published April 5, 2025. Accessed 2025-12-30.

Anthropic. 2025. Claude sonnet 4.5 system card. System card, Anthropic.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, Yu Cheng, Pei Chu, Tao Chu, Erfei Cui, Ganqu Cui, Long Cui, Ziyun Cui, Nianchen Deng, and 81 others. 2025a. Intern-s1: A scientific multimodal foundation model. *CoRR*, abs/2508.15763.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025b. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.*, 67(12).

Google DeepMind. 2025. Gemini 3 pro model card. Technical report, Google.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. 2025. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *ACL (1)*, pages 1135–1159. Association for Computational Linguistics.

Chaoyou Fu, Yifan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *CoRR*, abs/2411.15296.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. Technical report, Google DeepMind. Published February 2026.

Goeric Huybrechts, Srikanth Ronanki, Sai Muralidhar Jayanthi, Jack FitzGerald, and Srinivasan Veeravanallur. 2025. Document haystack: A long context multimodal image/document understanding vision LLM benchmark. *CoRR*, abs/2507.15882.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. 2024. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *CoRR*, abs/2409.12959.

Graham Kendall and Jaime A. Teixeira da Silva. 2024. Risks of abuse of large language models, like chatgpt, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learn. Publ.*, 37(1):55–62.

Maurice George Kendall. 1948. Rank correlation methods.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *NeurIPS*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522. Association for Computational Linguistics.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *NeurIPS*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *WACV*, pages 2199–2208. IEEE.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. In *NeurIPS*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Junjie Wang, Yin Zhang, Yatai Ji, Yuxiang Zhang, Chunyang Jiang, Yubo Wang, Kang Zhu, Zekun Wang, Tiezhen Wang, Wenhao Huang, Jie Fu, Bei Chen, Qunshu Lin, Minghao Liu, Ge Zhang, and Wenhu Chen. 2024a. PIN: A knowledge-intensive dataset for paired and interleaved multimodal documents. *CoRR*, abs/2406.13923.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024b. Needle in A multimodal haystack. In *NeurIPS*.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. 2025. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. In *The 39th (2025) Annual Conference on Neural Information Processing Systems*.

xAI. 2025. Grok-4 model card. Model card, xAI.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. 2025. MMIE: massive multimodal interleaved comprehension benchmark for large vision-language models. In *ICLR*. OpenReview.net.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, Shiyang Feng, Bin Wang, Chao Xu, Conghui He, Pinlong Cai, Min Dou, Botian Shi, Sheng Zhou, Yongwei Wang, and 4 others. 2024. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *CoRR*, abs/2406.11633.

Haolong Yan, Kaijun Tan, Yeqing Shen, Xin Huang, Zheng Ge, Xiangyu Zhang, Si Li, and Daxin Jiang. 2025. M-docsum: Do lvlms genuinely comprehend interleaved image-text in document summarization? *CoRR*, abs/2503.21839.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025b. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. *CoRR*, abs/2506.02454.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. In *ICDAR*, pages 1015–1022. IEEE.

Keyan Zhou, Zecheng Tang, Lingfeng Ming, Guanghao Zhou, Qiguang Chen, Dan Qiao, Zheming Yang, Libo Qin, Minghui Qiu, Juntao Li, and Min Zhang. 2025. Mmlongcite: A benchmark for evaluating fidelity of long-context vision-language models. *CoRR*, abs/2510.13276.

Appendix

A Details of Benchmark Creation

A.1 SIN-Data Construction Pipeline

Design Philosophy. SIN-Data targets the Scientific INterleaved (SIN) format for evidence-grounded evaluation over long-form multimodal scientific documents. Large-scale interleaved corpora such as OBELICS (Laurençon et al., 2023) provide broad web-scale sequences, but their image–text relations can be heterogeneous and are not always tied to explicit support relations. Scientific papers provide a more controlled substrate for evidence auditing: figures and tables are usually referenced by surrounding prose, captions define local semantics, and claims often depend on methods, metrics, or conditions described in distant sections. This makes scientific documents suitable for constructing evidence chains that are both cross-modal and verifiable.

The SIN format is designed as a machine-readable evaluation representation, rather than a replacement for the original human-readable PDF or source files. Its goal is to preserve traceable text spans, visual anchors, and the semantic links between them. In scientific documents, the interpretation of a figure may require connecting its caption, the experimental setup, and the corresponding result discussion. We therefore insert visual anchors near their first semantic citation, so that downstream tasks can reference native evidence units while retaining the long-context dependencies of the original paper.

The pipeline also handles structural variation across scientific sources. For arXiv-style documents, source packages are typically LaTeX-based, while biomedical articles in PMC are often distributed as XML-structured archives. SIN-Data unifies these sources into a common interleaved format and records quality signals such as token length, visual density, citation-link validity, and text–image interleaving frequency. These signals support filtering, stratification, and later analysis of model performance across different levels of document and multimodal complexity.

We build on the parsing strategy of PIN (Wang et al., 2024a) and extend it with quality-signal extraction for constructing SIN-Data. As shown in Fig. 2 (a), the workflow consists of three stages.

A.1.1 Stage 1: Element Parsing

The first stage separates document content from presentation layout and converts source files into intermediate structured representations.

- **arXiv stream.** We compile LaTeX source packages into responsive HTML using Engrafo[3], and then extract textual structure and visual elements from the rendered document. Image references are recovered from the DOM tree and re-linked to the corresponding textual contexts through visual matching.
- **PMC stream.** We process JATS XML files with a customized s2orc-doc2json[4] parser. The parser extracts core academic elements while removing stylistic markup, preserving citation links, tables, and other structural information in JSON format.

A.1.2 Stage 2: Semantic-First Formatting and Signal Extraction

We convert the intermediate structured files into a unified Interleaved Markdown format. To reduce layout-induced separation between figures and the text that invokes them, we use a citation-driven injection strategy: each visual component is assigned a unique placeholder $\langle x_k \rangle$ and inserted before the paragraph where it is first cited. This keeps visual anchors close to their semantic context while preserving traceability to the original document. During this conversion, we compute document-level quality signals used for filtering and analysis:

- **Volume and granularity.** We record total_tokens, computed with the Llama-4 tokenizer (AI at Meta, 2025), and avg_segment_length, the average number of tokens per text block.
- **Visual density.** We compute image_count and image_ratio to quantify the amount and proportion of visual content.
- **Interleaving degree.** We track interleave_segments, which measures how frequently text and visual anchors alternate in the document.
- **Layout and visual quality.** We extract format-related signals such as column layout and average image resolution to identify documents with potentially unreliable parsing or low-quality visual evidence.

[3] https://github.com/arxiv-vanity/engrafo
[4] https://github.com/allenai/s2orc-doc2json

A.1.3 Stage 3: Quality Filtering and Taxonomy Alignment

Using the signals from Stage 2, we filter documents with sparse visual context, broken citation links, unreliable image recovery, or extreme lengths. In our implementation, we retain documents within the 32k–1M token range and with sufficient text–image interleaving.

To support domain-balanced evaluation, we annotate documents according to the arXiv taxonomy with Qwen3-VL-2B (Bai et al., 2025b); details are provided in App. A.6. We then apply stratified sampling across taxonomy categories to reduce domain concentration in the final corpus.

Starting from $50{,}000$ source packages, this pipeline yields $4{,}000$ curated scientific documents. The final corpus covers 12 top-level disciplines and over 80 subfields, providing a controlled substrate for constructing evidence-grounded scientific document tasks.

A.2 SIN-Bench Construction Pipeline

Building on the interleaved document representations from `SIN-Data`, we construct `SIN-Bench` through an iterative human–model collaborative pipeline. The pipeline uses MLLMs to scale candidate generation, and then applies cross-model validation and human auditing to filter or correct samples with weak evidence support, factual errors, or invalid output structure. As shown in Fig. 2 (c), the workflow consists of seed example construction, input preparation, multi-MLLM synthesis, cross-validation, human-in-the-loop audit, and golden-to-seed iteration. This design reduces annotation burden while keeping evidence locatability, support validity, and schema compliance explicitly checkable.

Input Preparation: Pruning and Alignment. Scientific papers can exceed the context limits of current MLLMs. When pruning is required, we construct model inputs by preserving core sections such as Abstract, Introduction, Methods, Experiments, and Conclusion, while retaining text spans around the first citation of figures or tables together with their visual anchors $\langle x_k \rangle$. This strategy keeps the main claims and their associated evidence anchors available for candidate generation, while maintaining the text–figure coupling defined in the SIN format.

Seed Examples. We manually curate a small set of seed examples for each task. These examples define the task-specific input–output schema, the evidence-chain format, and the expected granularity of supporting evidence. They are used as demonstrations for subsequent candidate synthesis and are updated in later iterations with golden samples.

(1) Multi-MLLM Collaborative Synthesis. Given a document D, we use multiple MLLMs to generate candidate samples, including Gemini-3-pro, GPT-5, Qwen3-VL, Grok-4, and Claude-sonnet-4.5. We use SIN-QA and SIN-Summary as the main generation pivots because they jointly specify content and evidence. For SIN-QA, the model generates a query Q, an answer A, and a draft evidence chain E within the same context, reducing semantic drift between the answer and its support. For SIN-Summary, we use a cite-as-you-write prompt so that each summary claim is paired with evidence anchors. The remaining tasks are derived from these evidence-grounded outputs within the same generation round:

- **SIN-QA**: The model generates an answer A and an evidence chain E for a reasoning query Q, producing candidate answer–evidence pairs for later validation.
- **SIN-Find**: We convert validated QA or summary instances into evidence discovery tasks by fixing the query or claim and asking the model to recover the supporting evidence chain.
- **SIN-Verify**: We construct discriminative instances by pairing claims with either valid evidence chains or perturbed chains. Negative cases include insufficient evidence, unrelated evidence, mismatched conditions, and disrupted evidence order.
- **SIN-Summary**: The model generates a document-level summary in which each key claim is linked to one or more evidence anchors in D.

This stage yields a pool of raw candidate samples across all four tasks. Because the candidates share the same evidence-chain schema, they can be filtered consistently in the cross-validation and human-audit stages.

(2) Cross-Validation. To filter plausible but weakly supported candidates, we employ a three-model jury system, with validators dynamically sampled from Gemini-3-pro, GPT-5, Claude-sonnet-4.5, and Qwen3-VL-Plus. Each validator independently scores a candidate on a 1–5 scale along three dimensions: *Question Validity*, which assesses clarity, non-triviality, and document support; *Answer Correctness*, which checks factual consistency with

Cross-Validation Auditor Prompt: SIN-Summary

Role:
You are a rigorous academic reviewer auditing the **evidence quality** of a summary generated from a research paper. Your judgment must be **evidence-centric** and **strictly grounded** in the provided excerpt.

Inputs (you will receive):
1) A Markdown excerpt from a research paper (may contain `<figure>` placeholders corresponding to images/figures).
2) A list of summary evidence items. Each item contains (i) an image anchor marker (e.g., `x1`, `x2`, ...) and (ii) a text snippet describing key points.

Task Description:
Evaluate whether the provided evidence items faithfully support the claimed summary points, with explicit verification of **anchor-to-content alignment** (i.e., whether each `xk` correctly maps to relevant visual content and its surrounding text context in the excerpt).

Rubric (Criterionized, 1–5):
Evidence Quality (integer score only)
- **5**: Faithful and complete support; anchors are correctly matched; coverage includes key aspects (method, results, conclusions); organization is coherent.
- **4**: Mostly faithful with minor omissions or slightly weak anchor justification; no critical mismatches.
- **3**: Partial support; noticeable gaps in coverage or unclear anchor relevance; some evidence is loosely connected.
- **2**: Weak support; multiple items are irrelevant, mismatched, or insufficiently grounded in the excerpt.
- **1**: Unsupported or hallucinated; anchors do not match the described content; evidence is largely invalid.

Evaluation Checklist (Evidence-First):
- Do the text snippets accurately reflect key points present in the excerpt?
- Are cited image anchors (`x1`, `x2`, ...) relevant to their corresponding text snippets?
- Do the evidence items cover important aspects (methodology, results, conclusions) rather than superficial details?
- Is the evidence list well-organized and logically ordered (e.g., method → experiments → conclusions)?
- Does the judgment rely **only** on the provided excerpt (no external knowledge)?

Output Format (STRICTLY FOLLOW):
```
Evidence Quality Score: <INTEGER 1-5>
Reason: <ONE PARAGRAPH JUSTIFICATION>
```

Important Constraints:
- The score must be on its own line starting with `Evidence Quality Score:`.
- The score must be an integer from 1 to 5 (no decimals).
- Do not output any additional fields, bullet lists, or extra lines beyond the two required lines.
- Do not use Chinese characters or brackets like `{ }`.
- If an item cites an anchor that is not supported by the excerpt, you must penalize the score accordingly.

Example Output:
Evidence Quality Score: 4
Reason: The evidence items are mostly grounded in the excerpt: item 1 correctly cites x1 to describe the proposed architecture and matches the surrounding text, and item 2 references x2 for the main experimental comparison with accurate reported gains. However, item 3's anchor-to-claim linkage is under-specified because the cited x3 is only loosely related to the stated ablation conclusion, reducing overall faithfulness despite generally coherent organization.

Figure 7: Cross-validation prompt used to audit SIN-Summary evidence chains. It operationalizes an evidence-first rubric with discrete 1–5 tiers and enforces a strict, machine-parsable two-line output schema, enabling scalable filtering of mismatched or weakly grounded image–text evidence.

the provided document; and *Evidence Consistency*, which evaluates relevance, sufficiency, and ordering of the supporting evidence. A candidate is retained only if it receives majority approval and obtains a score of at least 4 on every dimension. This stage serves as an automatic filtering step before human auditing, removing candidates with ambiguous questions, unsupported answers, or inconsistent evidence chains.

To make the validation process parsable and evidence-centered, we use criterionized prompts with fixed output schemas. All validation prompts follow four principles:

- **Evidence-centered review**: validators judge candidates based on evidence anchors and support relations, rather than linguistic fluency alone.
- **Criterionized rubric**: each candidate is scored on fixed dimensions with discrete 1–5 ratings to

reduce calibration drift.

- **Strict I/O schema**: validators output structured fields and integer scores, enabling automatic parsing and threshold-based filtering.
- **Anchored validation**: validators check whether visual identifiers such as $\langle x_k \rangle$ are correctly linked to relevant text spans and claims.

As shown in Fig. 7, the SIN-Summary validation prompt asks the auditor to examine whether summary claims cover key document points, whether evidence anchors are relevant, and whether the summary follows a coherent order.

(3) Human-in-the-Loop Audit. Candidates that pass cross-validation may still contain fine-grained grounding errors, such as locatable but weakly supportive anchors, missing premises, or ambiguous negative evidence. We therefore conduct a human audit with a team of 24 graduate students and doctoral candidates. Each candidate is independently reviewed by 2–3 auditors, and disagreements are adjudicated by senior researchers. The audit follows five criteria:

- **Anchored Locatability**: Each visual anchor $\langle x_k \rangle$ and text span must be identifiable in the SIN document without ambiguity.
- **Rigor of Support**: The evidence chain must directly and sufficiently support the target answer or claim. For SIN-Verify negatives, auditors check that the candidate evidence is genuinely insufficient, mismatched, or inconsistent.
- **Factual Consistency**: Answers and summaries must be supported by the document, including numerical values, comparative statements, experimental conditions, and domain-specific terminology.
- **Structural Compliance**: The sample must follow the task-specific output schema and the interleaved evidence-chain format, with evidence units ordered according to their support relation.
- **Discriminability**: The question or claim should be neither trivial nor undecidable, and the evidence chain should provide diagnostic value for evaluating grounded support.

Samples that fail these criteria are corrected when possible or removed from the golden split. The audited samples are then used both as final benchmark instances and as updated seed examples for subsequent construction rounds.

(4) Bootstrapping Iteration (Golden → Seed). Audited golden samples are fed back as updated seed examples for the next synthesis cycle. This golden-to-seed iteration helps stabilize the task schema, evidence-chain format, and validation criteria across construction rounds.

Scale and Composition. Starting from $4,000$ curated scientific documents, the multi-round synthesis pipeline produces approximately $3,200$ raw candidate samples. After three-model cross-validation and human auditing, the final `SIN-Bench` contains 490 golden samples: 159 for SIN-Find, 158 for SIN-QA, 89 for SIN-Summary, and 84 for SIN-Verify.

The released golden split is moderate in scale because each sample requires evidence-chain validation and human audit. Nevertheless, the semi-automated pipeline is designed to support future expansion: additional samples can be constructed under the same synthesis, validation, and auditing protocol when more computational and human-review resources are available.

A.3 On the Principle of "No Evidence, No Score"

"No Evidence, No Score" is an evidence-centered evaluation principle: a model should not receive high evidence-aligned scores solely from answer correctness when its output cannot be grounded to verifiable document anchors. Accordingly, for evidence-required tasks, evidence quality is treated as a core scoring component rather than an optional explanation.

This principle targets *verifiability*, not rigid citation surface forms. We normalize minor formatting variants such as "Fig. 2", "Figure 2", and "$\langle x_2 \rangle$", as long as the anchor can be recovered and matched to the indexed visual markers in the SIN document. A prediction receives zero evidence credit only when it provides no recoverable anchor, invalid identifiers, or evidence units that cannot be linked back to the document.

The protocol is also consistent with separate answer scoring. For SIN-QA, we report AnsAcc as a standalone diagnostic of answer correctness, but AnsAcc alone is insufficient for a high overall score. When no anchor is matched, evidence metrics are set to zero by definition, including Matching and Relevance; Logic is also set to zero when no matched evidence unit exists. When a single evidence unit is matched, ordering is treated as trivially satisfied, while multi-unit chains are evaluated by Kendall–Tau similarity. Thus, answer-only correctness can be observed, but it is explicitly separated from evidence-grounded performance.

The metric suite further covers common ground-

Task	Avg. Length	Max Length	Min Length
SIN-Find	108.89	186	60
SIN-QA	122.77	203	78
SIN-Verify	243.49	322	160

Table 3: Token-length statistics of input contexts across task types.

Scope	Value
Tokens (Text / Image / Total)	15k / 3k / 18k
Image Ratio	15%
Text Ratio	85%
Avg. #Images per instance	6.6

Table 4: Overall modality composition of benchmark inputs per sample.

Task	Bold Avg.	Italic Avg.	Title Avg.
SIN-Find	70.52	122.46	14.64
SIN-QA	70.83	118.17	15.03
SIN-Summary	37.46	155.74	14.91
SIN-Verify	90.35	155.65	15.94
Overall	**68.01**	**132.81**	**15.03**

Table 5: Markup statistics (bold/italic/title tokens) across tasks, reflecting structural formatting cues in the inputs.

ing failure modes. Over-generating irrelevant anchors lowers precision in Relevance; missing required evidence lowers recall; and cherry-picked or mis-ordered evidence lowers Logic. These cases are handled by the Matching, Relevance, and Logic metrics defined in the main text, without introducing extra task-specific indicators.

A.4 General Statistics

We report descriptive statistics of SIN-Bench from three perspectives: (i) input length, (ii) text–image composition, and (iii) structural markup density.

Table 3 summarizes token-length statistics by task type. Overall, SIN-Verify instances are substantially longer than SIN-Find and SIN-QA, with roughly twice the average token length. This reflects the additional context introduced by the candidate answer and evidence chain, and suggests that evidence sufficiency verification imposes a heavier input-processing burden. In contrast, SIN-Find and SIN-QA show comparable length ranges, since both are built around a document-query interface.

Table 4 reports the modality composition of benchmark instances. On average, image-related tokens account for approximately 15% of total tokens, with 6.6 images per instance. This indicates that the benchmark remains text-dominant while still requiring non-trivial grounding to visual evidence.

Finally, Table 5 reports structural markup statistics, including bold, italic, and title tokens. These features reflect the density of formatting and sectioning cues in scientific documents, which may affect how models identify document structure, locate relevant evidence, and follow long-range dependencies.

A.5 Samples from SIN-Bench

As shown in Fig. 8 and Fig. 9, we provide one audited golden instance for each task in SIN-Bench. Because scientific papers are long, the figures show cropped views for readability rather than the full input context. Each crop highlights the task input, the relevant evidence region, and the corresponding gold output, such as the answer, verification label, or evidence-anchored summary claims. These visualizations are illustrative only: all benchmark instances are evaluated on the full interleaved source files, including Markdown text and associated image assets. Released samples include stable document, image-anchor, and instance identifiers to support exact reconstruction.

A.6 Discipline Taxonomy List

To ensure a comprehensive evaluation of multimodal reasoning across scientific disciplines, we constructed a three-level hierarchical taxonomy referencing the arXiv subject classification.

Our pipeline integrates large model capabilities with expert human verification. We first utilized Qwen3-VL-2B to perform preliminary classification on the collected samples. Subsequently, domain experts reviewed the results to correct misclassifications and refine subfield definitions.

The resulting taxonomy comprises 12 Level-1 major disciplines, branching into 35 Level-2 domains and 84 Level-3 subfields. The hierarchical structure is organized as follows:

- **Astronomy & Astrophysics**:
  - Extragalactic Astronomy & Cosmology: Active Galactic Nuclei & Quasars; Galaxy Formation & Evolution.
  - Galaxy Evolution & Structure: Structure & Evolution of Galaxies.
  - Planetary & Space Science: Planetary Geology & Exploration.
  - Planetary Systems & Formation: Planetary Systems Protoplanetary Disks.
- **Biology & Life Sciences**:

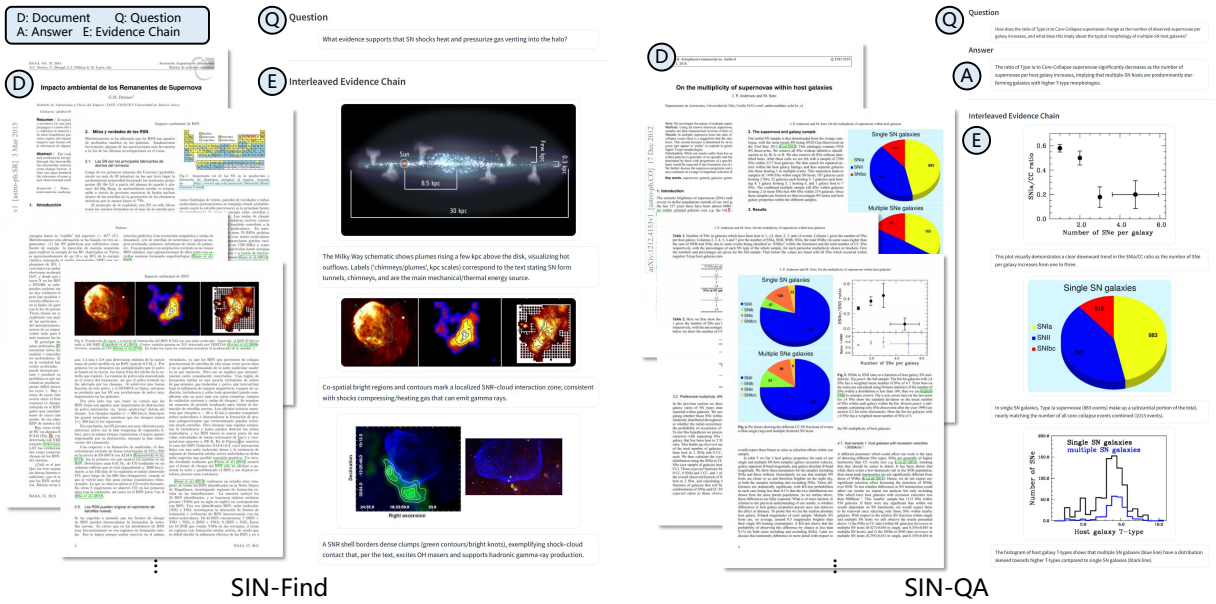

Figure 8: Golden examples from SIN-Bench: SIN-Find and SIN-QA. We visualize one curated instance per task with explicit markers for **D**ocument, **Q**uestion, **A**nswer, and the gold **E**vidence chain.

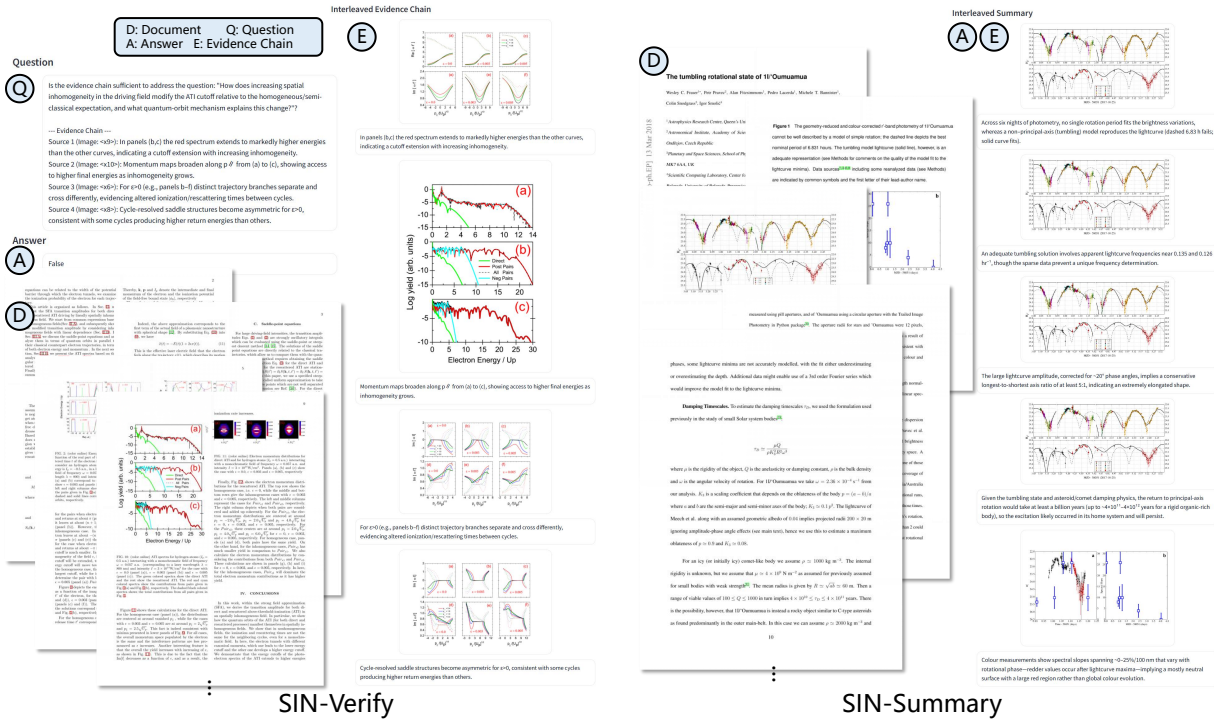

Figure 9: Golden examples from SIN-Bench: SIN-Verify and SIN-Summary. We visualize one curated instance per task with explicit markers for **D**ocument, **Q**uestion, **A**nswer, and the gold **E**vidence chain.

  - Neuroscience: Neuroimaging.
- **Computer Science**:
  - Algorithms & Theory: Combinatorial Algorithms; Graph & Combinatorial Algorithms; Optimization; Quantum Algorithms & Computation Theory.
  - Communication & Networking: Information Theory & Coding; Channel Modeling.
  - Computer Vision: 2D 3D Reconstruction & Modeling; Image Processing & Restoration; Video Understanding & Compression; Vision.
  - Electrical & Computer Engineering: Embedded & Hardware Systems.
  - Human-Computer Interaction: Virtual Augmented Reality & Haptics.
  - Machine Learning & AI: Active Learning;

Clustering & Unsupervised Learning; Deep Learning; Deep Reinforcement Learning; Differential Privacy & Anonymization; Explainable & Fair AI; Federated & Continual Learning; Generative Models; Graph Neural Networks; Multimodal Learning & Fusion; Nonlinear Time Series; Object Detection; Particle Filters & Data Assimilation; Privacy; Quantum Algorithms & Computation Theory; Second Order Methods; Submodular Optimization.
  - Natural Language Processing: Computational Linguistics; Information Retrieval & QA Systems.
  - Robotics: Autonomous & Mobile Robotics; Learning for Robotic Control.
  - Security & Privacy: Adversarial Machine Learning; Blockchain & Distributed Ledger Security; Differential Privacy & Anonymization.
  - Software Engineering & Programming: AI for Software Engineering; Program Synthesis & Verification.
  - Systems & Infrastructure: Edge & Cloud Computing & Serverless.
- **Earth & Environmental Sciences**:
  - Climate & Atmospheric Science: Atmospheric Chemistry & Air Quality; Climate Modeling & Prediction.
  - Geophysics & Remote Sensing: Geospatial Intelligence & GIS; Satellite & Aerial Imaging.
  - Oceanography & Hydrology: Water Resources & Hydrology.
  - Planetary & Space Science: Planetary Geology & Exploration; Space Instrumentation & Missions; Space Weather & Near Earth Environment.
- **Economics & Finance**:
  - Macroeconomics & International Trade: Sovereign Debt & Default Risk.
- **Engineering**:
  - Mechanical & Robotics Engineering: Robotics & Automation.
- **Mathematics**:
  - Pure Mathematics: Analysis & Functional Analysis.
- **Medicine & Health Sciences**:
  - Neurology & Psychiatry: Computational Psychiatry.
- **Philosophy & Ethics**:
  - AI & Technology Ethics: Philosophy of Quantum Physics.
- **Physics**:
  - Astrophysics & Cosmology: Astronomical Instrumentation; Atmospheric Chemistry & Air Quality; Cosmology & Galaxy Formation; Exoplanets & Planetary Science; Gravitational Lensing & Cosmology; Gravitational Lensing & Microlensing; Neutron Stars & Pulsars; Planetary Geology & Exploration; Satellite & Aerial Imaging; Space Instrumentation & Missions; Supernova Remnants & Cosmic Rays; Supernovae & Stellar Evolution.
  - Condensed Matter & Materials: Computational Materials Physics; Condensed Matter Physics; Materials Science & Characterization; Spin Glass & Magnetic Materials; Superconductivity & Magnetism.
  - Optics, Fluids & Dynamics: Laser Physics & Photonics; Photonic Crystals & Band Gap Physics; Plasma & Fusion Physics; Quantum Optics & Atomic Physics .
  - Optics & Photonics: Nonlinear Optics & Optical Processing; Optical Imaging & Sensing.
  - Quantum & Particle Physics: Gravitational Waves & Cosmology; High; Neutrino Astronomy & Astrophysics; Quantum Computing & Information; Quantum Optics & Atomic Physics.
- **Social Sciences**:
  - Communication, Media & Linguistics: Computational Social Media Analysis.
  - Economics & Finance: Econometrics & Applied Microeconomics.
  - Education & Learning Sciences: Educational Technology & AI in Education.
  - Sociology & Social Policy: Social Networks & Computational Social Science.
- **Statistics & Data Science**:
  - Data Science Applications: Health & Social Data Science.
  - Statistical Theory & Methods: Nonparametric & Bayesian Statistics.

This structured distribution ensures that SIN-Bench covers a wide spectrum of knowledge.

A.7 License

Our code and derived annotation files are released under MIT https://opensource.org/licenses/MIT; source documents remain under their original licenses.

Benchmark	Doc type	Scale	Avg. length	Domain coverage	Main tasks	Ev.	Ev.-score
MMLongBench-Doc (Ma et al., 2024)	Long PDFs (multi-domain)	130–135 docs (benchmark)	∼21k tokens	7 domains (e.g., acad./legal/tech)	Long-doc understanding; numerical reasoning; cross-element search	✗	✗
DocGenome (test) (Xia et al., 2024)	Scientific papers (structured docs)	–	–	8 classes +153 subclasses	7 document tasks (e.g., cls./QA/ grounding/layout)	✗	✗
Document Haystack (Huybrechts et al., 2025)	Long PDFs (general)	400 docs	5–200 pp.	General	Needle-in-a-haystack retrieval	✗	✗
SciFIBench (Roberts et al., 2024)	Scientific figures (charts)	–	–	8 figure types (CS-centric)	Figure interpretation	✗	✗
SIN-Bench (ours)	Scientific docs (interleaved)	231 docs / 490 inst.	∼18k tokens	10+ L1 domains; 30+ L2; 80+ L3	4 evidence tasks: Find/Verify/ QA/Summary	✓	✓

Table 6: Comparison with representative long-document and scientific multimodal benchmarks. "Ev." indicates whether the benchmark requires an explicit cross-modal evidence chain in the model output. "Ev.-score" indicates whether evidence quality is explicitly scored beyond answer-only evaluation.

B Comparisons with Existing Benchmarks

Table 6 compares SIN-Bench with representative long-document and scientific multimodal benchmarks. Prior resources have advanced long-context stress testing and document-centric evaluation, including long-PDF understanding, needle-style retrieval, structured document tasks, and scientific figure interpretation. However, most of them primarily evaluate answer correctness, localization, or task-specific outputs, without requiring models to produce explicit cross-modal evidence chains or scoring the quality of such evidence.

SIN-Bench focuses on this missing evaluation dimension. It uses native interleaved scientific documents and defines four evidence-auditing tasks: evidence discovery, sufficiency verification, grounded QA, and evidence-anchored synthesis. Beyond answer correctness, it evaluates whether model outputs are supported by matched, relevant, and logically ordered evidence anchors. Although the released golden split is moderate in scale, it is produced through multi-stage parsing, cross-model validation, and human auditing. We therefore position SIN-Bench as an audited diagnostic benchmark for evidence-grounded scientific document evaluation, with a construction pipeline that supports future expansion under the same protocol.

C Details of Experimental Setting

C.1 Baselines

We evaluate 9 representative MLLMs, including 6 proprietary models and 3 open-weight models. All evaluations are run once with greedy decoding; for API-based models, we set the temperature to 0 when supported. For reproducibility, we report the proprietary model identifiers used in our experiments:

- Gemini-3-pro: `gemini-3-pro-preview-11-2025`
- Gemini-3.1-pro: `gemini-3.1-pro-preview`
- Gemini-2.5-pro: `gemini-2.5-pro-thinking`
- Claude-sonnet-4.5: `claude-sonnet-4-5-20250929-thinking`
- GPT-5: `gpt-5`
- Grok-4: `grok-4-0709`

The open-weight models are Qwen3-VL-2B, Qwen3-VL-8B, and Qwen3-VL-30B-A3B, all evaluated in Thinking mode. Additional implementation details are released with the benchmark package.

C.2 Models Used in the Benchmark and Experiments

We use all evaluated models for the main benchmark comparison. For controlled analysis experiments, however, compute budget and API cost limit the number of repeated ablations. We therefore use Gemini-3-pro as the primary diagnostic model, since it achieves the best average overall score on SIN-Bench among the evaluated models. This choice provides a strong reference point for analysis, but is not intended to imply that the observed trends are universal across all model families.

In the modality study in Fig. 5 (Sec. 5.3), we isolate the role of direct visual input with a "pure text" setting. Specifically, each image slot in the original interleaved sequence is replaced by a natural-language description of the corresponding figure. The descriptions are generated by Qwen3-VL-8B-Instruct, preserving the document structure while removing direct access to image pixels. This setting is used as an ablation of visual access, and

Task	Metric	Pearson (r)	Spearman (ρ)
SIN-Find	Matching (M)	0.824	0.795
SIN-QA	Matching (M) Answer Acc.	0.812 0.895	0.788 0.862
SIN-Summary	Matching (M)	0.768	0.741
-	**Average**	**0.825**	**0.797**

Table 7: Correlation between expert human ratings and automated scores produced by Qwen3-8B. We report Pearson and Spearman correlations for metrics that require semantic judgment. All correlations are statistically significant with $p < 0.001$.

the generated descriptions may introduce their own approximation errors.

For evaluation components that require semantic judgment, we adopt an LLM-based evaluator rather than n-gram overlap metrics. This is necessary because evidence notes and answers can be semantically correct under diverse surface forms. Concretely, Matching and answer correctness (AnsAcc) are judged by Qwen3-8B with fixed rubrics and structured outputs, while Relevance and Logic are computed deterministically from anchor matching and evidence order. This design follows prior work showing that rubric-guided LLM evaluation can improve agreement with human judgments for open-ended outputs (Liu et al., 2023; Zheng et al., 2023). We further validate the automatic judge against expert annotations in App. C.3, where we report close agreement between the judge scores and human ratings.

C.3 Judges in Metrics

For metrics that require semantic judgment, including Matching and SIN-QA answer accuracy, we use Qwen3-8B (Yang et al., 2025a) as the automatic judge. To assess the reliability of this evaluator, we conduct a judge–human consistency study on the corresponding evaluation items. Each item is scored independently by Qwen3-8B and by 2–3 human annotators with relevant graduate-level research experience. Human scores are aggregated into an expert consensus reference, against which we compute Pearson and Spearman correlations for the automatic judge. As shown in Table 7, Qwen3-8B achieves strong agreement with expert ratings across tasks and metrics, with average Pearson $r = 0.825$ and Spearman $\rho = 0.797$. This result supports the use of Qwen3-8B as a scalable evaluator for semantic evidence matching and answer correctness. We emphasize that this validation assesses the reliability of the automatic scoring procedure; it is not intended as an independent human-performance baseline on SIN-Bench.

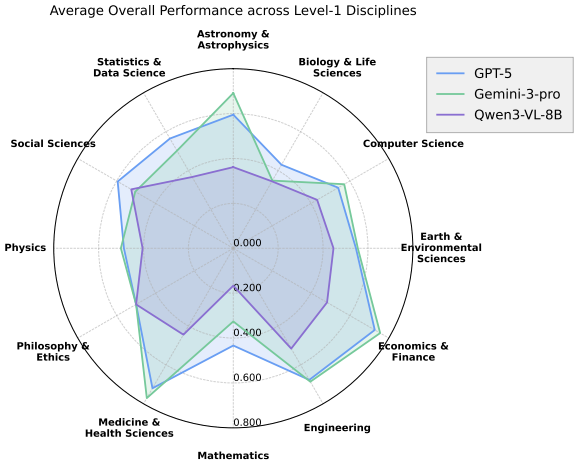

Figure 10: Domain-wise performance on SIN-Bench. The radar plot reports average overall scores across all tasks for GPT-5, Gemini-3-pro, and Qwen3-VL-8B over 12 primary disciplines. Higher values indicate stronger evidence-aligned performance under our metric suite.

D Additional Experiments and Result Analysis

D.1 Discipline-level Performance Analysis

Fig. 10 summarizes discipline-level performance under our Level-1 taxonomy. The results show clear domain-dependent variation under the SIN-Bench evaluation protocol. Aggregated across models, *Economics & Finance* and *Medicine & Health Sciences* obtain relatively high average overall scores (avg. ≈ 0.66 and 0.65), whereas *Mathematics* is the lowest-scoring domain (avg. ≈ 0.31). This suggests that domains requiring dense symbolic manipulation, formal notation, or precise quantitative grounding remain challenging for current MLLMs.

Model rankings also vary across disciplines. Gemini-3-pro performs better in visually intensive or layout-heavy domains such as *Astronomy & Astrophysics* and *Medicine & Health Sciences* (e.g., $+0.01$ and $+0.05$ over GPT-5), where evidence is often conveyed through figures, tables, and chart-like graphics. GPT-5 performs better in more formal or quantitative domains, especially *Mathematics* ($+0.10$ over Gemini-3-pro) and *Statistics & Data Science*. Qwen3-VL-8B generally trails the proprietary models, but remains comparatively competitive in *Social Sciences* and approaches their performance in *Philosophy & Ethics*. Given the moderate number of audited samples per discipline, these results should be interpreted as diagnostic

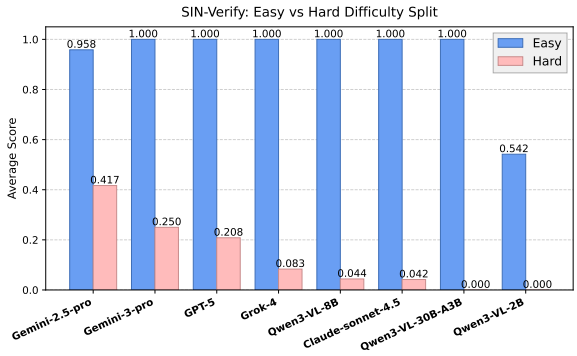

Figure 11: Accuracy on SIN-Verify with two negative sets. Easy negatives are clearly irrelevant evidence; hard negatives are cross-validation near-misses with correct answers but insufficient/ambiguous evidence support.

trends rather than definitive domain rankings.

D.2 An Analysis of Near-Miss Evidence in SIN-Verify

To examine verification difficulty beyond trivial mismatch, we construct two negative sets during cross-validation (App. A.2, step (2)). *Easy negatives* pair a question with clearly irrelevant evidence, so the unsupported relation is easy to identify. *Hard negatives* are near-miss cases: the question and answer are plausible, but the candidate evidence chain is ambiguous, incomplete, or insufficient to support the answer. We sample 24 instances for each setting.

As shown in Fig. 11, MLLMs nearly saturate on easy negatives, indicating that they can detect coarse evidence mismatch. However, accuracy drops sharply on hard negatives, showing that evidence sufficiency verification is substantially more difficult than relevance rejection. These cases require models to check fine-grained support relations, such as numerical values, experimental conditions, and figure–text alignment, rather than relying on topical overlap between the answer and evidence. Gemini-2.5-pro performs best on hard negatives (0.417), while Gemini-3-pro (0.250) and GPT-5 (0.208) remain much lower. This suggests that strong answer-generation ability does not necessarily translate into reliable sufficiency judgment under near-miss evidence. The remaining models approach near-zero accuracy, further highlighting that plausible but insufficient evidence remains a challenging failure mode for current MLLMs.

D.3 Impact of Document Length on Performance

Scientific papers are long-form and multimodal by construction. To analyze how performance varies with input length under interleaved inputs, we examine the joint distribution between prompt length and overall score under two accounting schemes: text-only tokens (Fig. 12) and text+image tokens (Fig. 13). We report results for SIN-Find, SIN-QA, and SIN-Summary. In Sec. 5.3, we focus on SIN-QA and SIN-Summary because they require answer or summary generation together with evidence-chain grounding.

Setup. For each evaluated sample, we record the input length and the corresponding overall score. In Fig. 12, the x-axis counts text tokens only. We use text-only length as the primary comparison axis because image tokenization and visual preprocessing can vary substantially across model APIs and inference pipelines. In Fig. 13, the x-axis counts text+image tokens, where image-token budgets depend on model-specific visual encoding policies, such as native-resolution encoding or fixed-size resizing. For Qwen3-VL-2B, token statistics follow the official configuration. Each panel visualizes the normalized frequency over the joint space of prompt length and overall score. We use coarse uniform binning for visualization and do not apply statistical smoothing to the underlying counts; any smooth appearance comes only from rendering-time interpolation. The inset reports the mean overall score for the corresponding setting.

Text-only length (Fig. 12). Across the three tasks, Gemini-3-pro and GPT-5 maintain substantial probability mass in the mid-to-high score region over a broad text-length range. GPT-5 shows a more visible score decrease at longer prompts than Gemini-3-pro, especially on SIN-QA. However, neither model shows a complete concentration of samples in the low-score region at the longest text lengths. This suggests that, for strong proprietary models, text length alone does not fully explain the remaining errors; evidence selection, anchor matching, and fine-grained support verification remain important factors.

Qwen3-VL-2B shows a different pattern. For SIN-Find and SIN-QA, its scores are concentrated near the lower range across most text lengths, with limited movement toward higher-score regions even for shorter inputs. A similar low-score concentration appears in SIN-Summary. This pattern

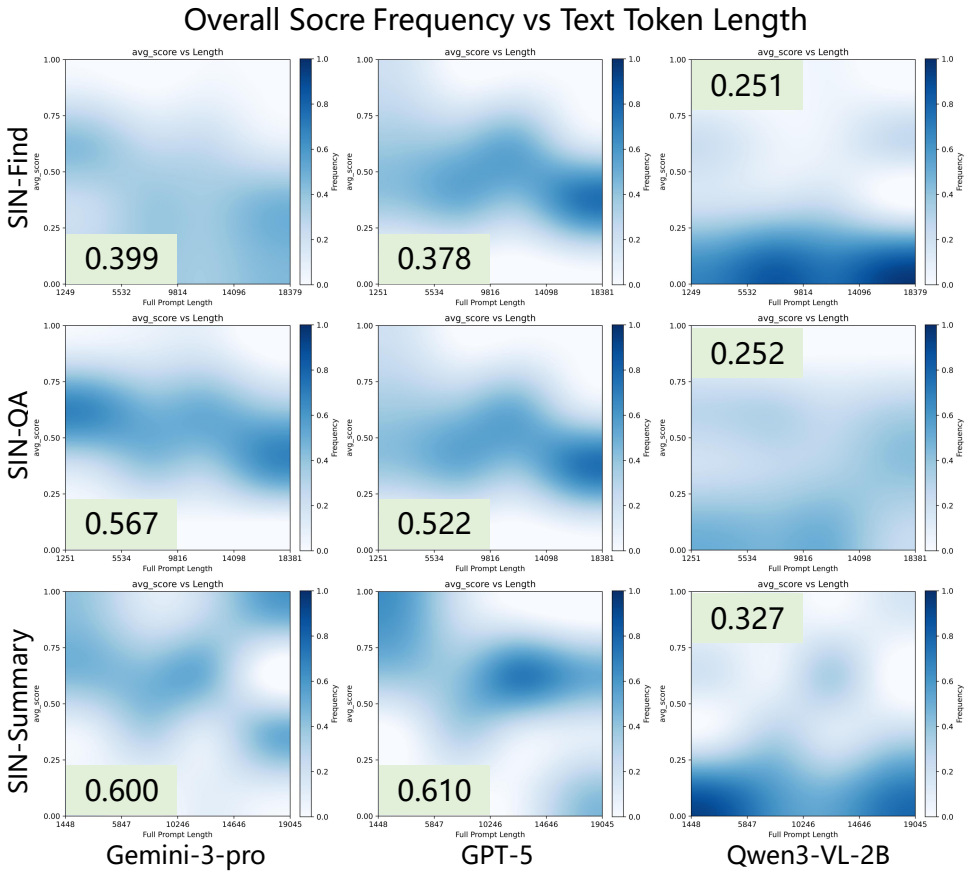

Figure 12: Overall score frequency vs. text-token length for SIN-Find, SIN-QA and SIN-Summary. Normalized 2D frequency of samples binned by text prompt length and overall scores.

suggests that the difficulty for Qwen3-VL-2B is not solely attributable to long-context length, but also reflects challenges in evidence localization, cross-modal grounding, and structured evidence output under interleaved scientific inputs.

Text+image length (Fig. 13). When image tokens are included, the prompt-length range expands substantially, reaching up to ∼87k tokens. This expansion reflects both document-side factors, such as figure count, and model-side factors, such as visual preprocessing and tokenization policy. Compared with Fig. 12, the distributions become more clustered along the length axis, suggesting that total token length is partly discretized by typical image counts, resolutions, and visual encoding choices. Therefore, trends along the text+image length axis should be interpreted cautiously, since this axis conflates document visual content with model-specific encoding behavior.

Gemini-3-pro and GPT-5 achieve similar mean overall scores, but their score distributions differ as total length increases. Gemini-3-pro remains more evenly distributed across mid-to-high score regions, whereas GPT-5 shows a clearer separation between high-score and low-score regions at longer total lengths. This pattern suggests that longer multimodal inputs do not uniformly degrade performance; instead, they may increase instance-level variance. One possible reason is that additional figures can either provide useful supporting evidence or introduce more visually similar anchors, denser evidence chains, and harder figure–text alignment cases. Thus, total multimodal length should be viewed as a mixed indicator of both available evidence and grounding difficulty.

For Qwen3-VL-2B, the distribution remains dominated by low-score regions across tasks. In SIN-Find and SIN-QA, however, some longer text+image inputs still reach moderate scores, suggesting that additional visual anchors can occasionally help when the relevant figure is salient and directly linked to the query. By contrast, SIN-Summary remains concentrated in low-score regions even at large total lengths. This indicates that additional visual tokens alone are insufficient for evidence-anchored synthesis, which requires selecting, organizing, and grounding multiple claims across the document.

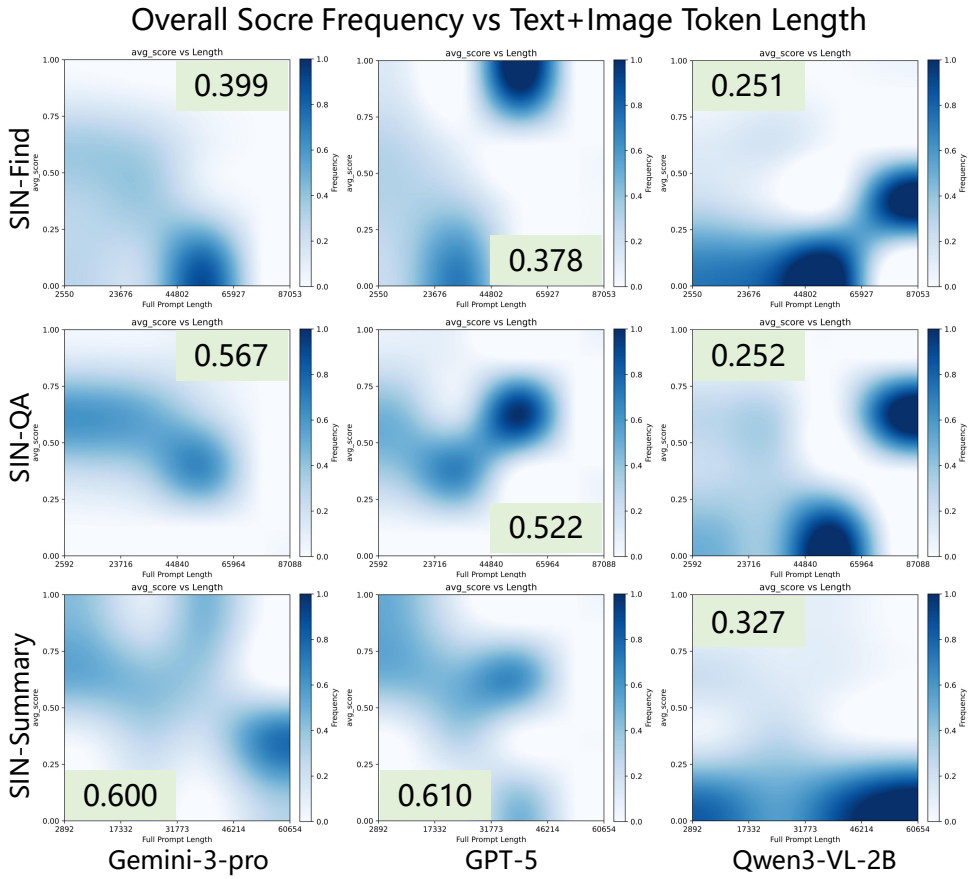

Figure 13: Overall score frequency vs. total (text and image) token length. Same as Fig. 12 but with prompt length computed from text plus image tokens from the API request.

E Case Study

As shown in Fig. 14, we further analyze SIN-QA because it requires both a final answer and an explicit evidence chain, making it useful for inspecting grounding failures beyond aggregate scores. In addition to the two failure modes summarized in Sec. 5.3, namely *Information Deficiency* and *Spurious Grounding*, these cases illustrate two concrete forms of evidence failure.

In Case (a), the model introduces unsupported concepts that are absent from the reference evidence, such as extending a tidal-force explanation into an earthquake-related narrative. Although part of the response remains topically related, the added content shifts the explanatory frame away from the provided evidence. This illustrates *Spurious Grounding*: fluent outputs may cite or mention relevant-looking evidence while still introducing unsupported expansions, thereby reducing precision.

Case (b) reflects a procedural form of *Information Deficiency*. The model fails to preserve the statistical interpretation required by the reference evidence, conflating or oversimplifying the role of standard deviation in variability analysis. As a result, the answer appears quantitative but misses a necessary methodological condition, leading to a conclusion that is not fully supported by the evidence chain.

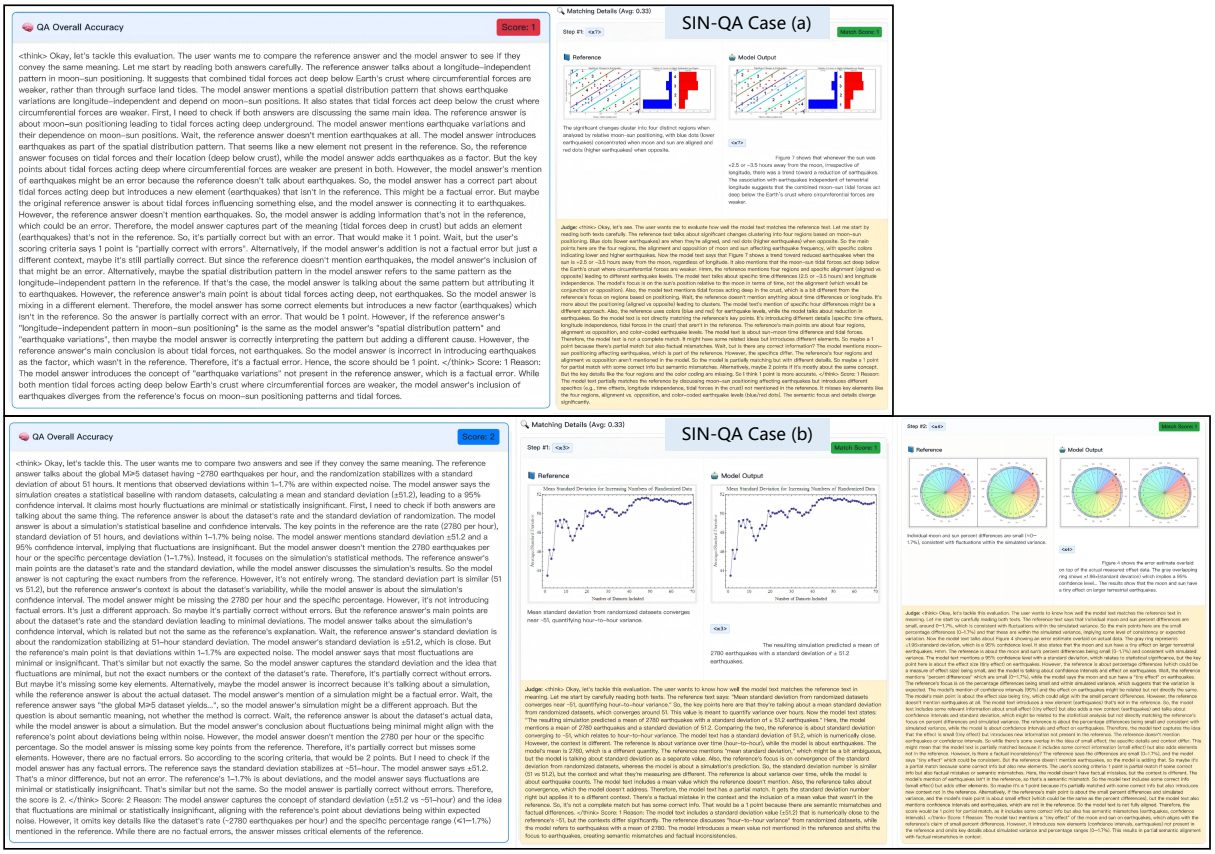

Figure 14: Error cases on SIN-QA.