SCIENCEDOC: MULTIDISCIPLINARY MULTIMODAL EVIDENCE-GROUNDED QUESTION ANSWERING FOR SCIENTIFIC DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are increasingly applied to document understanding and scientific question answering (QA). However, current evaluation remains limited by narrow disciplinary coverage, insufficient assessment of multimodal evidence grounding, and limited evaluation of cross-page and multi-document reasoning. We introduce ScienceDoc, an evidence-grounded multimodal question answering benchmark built from 693 scientific documents, containing 2,200 questions across 8 major disciplines and 61 scientific fields. ScienceDoc includes 4 task types, namely General, Unanswerable, Reasoning, and Multi-Document, together with 21 discipline-specific question types, covering multimodal content from text, equations, tables, and figures. All initial questions are designed by human experts, with expert-annotated answers and manually labeled evidence pages, supporting high-quality hierarchical evaluation of answer correctness and evidence grounding. Experiments on 11 representative MLLMs reveal challenges in multi-document QA and large performance variation on unanswerable questions. The study findings also indicate significant differences across scientific disciplines and gaps between answer accuracy and evidence localization, suggesting that conventional answer accuracy alone is insufficient for evaluating trustworthy scientific document question answering.

1 INTRODUCTION

Multimodal large language models (MLLMs) have shown strong performance in document understanding, question answering, and visual reasoning (Li et al., 2023; Fan et al., 2024; Chen et al., 2024b; Abdin et al., 2024; Yao, 2025). Their application to scientific research is also expanding, particularly for understanding scientific papers and reports (Li et al., 2024b). Scientific documents combine long-form text, equations, tables, figures, and experimental results across multiple pages and modalities (Li et al., 2024a). Therefore, scientific document question answering (QA) provides an important setting for evaluating whether models can retrieve, integrate, and reason over complex scientific information (Dasigi et al., 2021; Luo et al., 2023; Zheng et al., 2024; Hannan et al., 2026). However, current MLLMs still face difficulties in locating relevant evidence, integrating information across pages and documents, and generating answers that are fully supported by the source content (Dasigi et al., 2021; Li et al., 2024a; Deng et al., 2025).

Existing studies have explored both scientific question answering and document understanding (Luo et al., 2024). Early document QA datasets such as DocVQA (Mathew et al., 2021) focus on visual question answering over document images, while QASPER (Dasigi et al., 2021) extends question answering to full scientific papers. ScienceQA further explores multimodal reasoning across science topics, while its questions are mainly drawn from educational settings (Lu et al., 2022). With the development of MLLMs, recent works have moved toward more complex document understanding. SPIQA (Pramanick et al., 2024) incorporates figures, tables, and text from scientific papers, while LongDocURL (Deng et al., 2025) evaluates long multimodal document understanding and reasoning. More recent efforts, such as PaperScope (Xiong et al., 2026), explore reasoning across larger collections of scientific papers. These studies show a clear shift from basic document QA toward long-context, multimodal, and trustworthy scientific document understanding.

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.

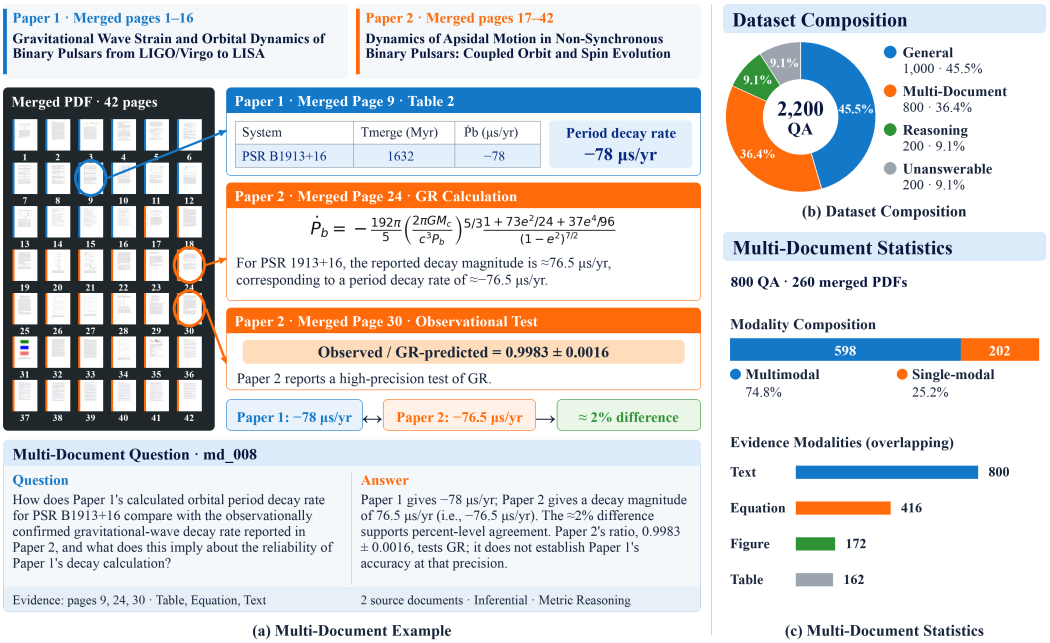


Figure 1: The multi-document question-answer example and dataset statistics.

Despite this progress, three challenges remain. First, scientific fields differ in their knowledge forms and reasoning patterns, making broad and discipline-aware evaluation necessary. Second, a correct answer does not always indicate correct document understanding, since models may rely on prior knowledge or incomplete evidence. Therefore, scientific QA requires explicit evaluation of evidence grounding. Third, scientific questions often require information from multiple pages, modalities, or documents, which makes cross-page and multi-document reasoning essential.

Accordingly, we propose ScienceDoc, an evidence-grounded multimodal question answering benchmark for scientific documents. ScienceDoc comprises 2,200 QA-pairs derived from 693 scientific documents, spanning 8 disciplines and 61 scientific fields. The questions are manually designed according to discipline-specific requirements and verified to assess scientific understanding across text, equations, tables, and figures. ScienceDoc includes 1,000 General, 200 Unanswerable, 200 Reasoning, and 800 Multi-Document questions, covering 21 question types. Reasoning questions require integrating multiple pieces of evidence through comparison, causal reasoning, or multi-step calculation, while multi-document questions require combining complementary evidence from multiple documents. Each question is manually annotated with supporting evidence pages, enabling joint evaluation of answer correctness and evidence grounding. An example is shown in Figure 1.

Extensive evaluation of 11 MLLMs shows that ScienceDoc remains challenging for recent strong models. Multi-document QA is particularly difficult, with 7 models achieving below 50% accuracy, while performance on unanswerable questions varies widely from 2.5% to 72.5%, highlighting the difficulty of recognizing insufficient evidence. Performance also varies across scientific disciplines, and strong answer accuracy does not always correspond to precise evidence localization. These results show that ScienceDoc captures limitations in multi-document reasoning, insufficient-evidence recognition, domain robustness, and evidence grounding beyond conventional answer accuracy.

The main contributions are summarized as follows[1]: 1) We introduce ScienceDoc, a multidisciplinary scientific document QA benchmark covering 8 major disciplines, 61 scientific fields, 4 task types, 21 discipline-specific question types, and multiple modalities. 2) We provide expert-designed questions, expert-annotated answers, and evidence-page annotations, enabling joint evaluation of answer correctness and evidence grounding. 3) We benchmark 11 representative MLLMs under a unified protocol and provide detailed analysis of their performance across task types, scientific disciplines, multi-document reasoning, and evidence localization.

[1]All the code and dataset will be released upon acceptance.

Table 1: Comparison of representative document understanding and scientific QA benchmarks. Sci. Doc: scientific document; Modal.: modality; Multi. Doc.: multi-document; Evid.: evidence annotation; Una.: unanswerable; DS Rea.: domain-specific reasoning; Hie. Eval.: hierarchical evaluation. T/Tb/Fg/Eq: text/table/figure/equation. 3S/26T: 3 subjects/26 topics; 8D/61F: 8 domains/61 scientific fields. ✓/×/△: supported/unsupported/partially supported.

Dataset	Domain & Modality			Task Design				Evaluation
	Sci. Doc	Sci. Domain	Modal.	Multi. Doc.	Evid.	Una.	DS Rea.	Hie. Eval.
DocVQA (Mathew et al., 2021)	×	–	T/Tb/Fg	×	×	×	×	×
QASPER (Dasigi et al., 2021)	✓	NLP	T/Tb/Fg	×	✓	✓	×	×
ScienceQA (Lu et al., 2022)	×	3S/26T	T/Fg	×	×	×	×	×
MP-DocVQA (Tito et al., 2023)	×	–	T/Tb/Fg	×	✓	×	×	×
MMLongBench-Doc (Ma et al., 2024)	△	–	T/Tb/Fg	×	✓	✓	×	×
SPIQA (Pramanick et al., 2024)	✓	CS	T/Tb/Fg	×	✓	×	×	×
LongDocURL (Deng et al., 2025)	△	–	T/Tb/Fg	×	✓	×	×	✓
PaperScope (Xiong et al., 2026)	✓	AI	T/Tb/Fg/Eq	✓	✓	×	×	✓
ScienceDoc (Ours)	✓	8D/61F	T/Tb/Fg/Eq	✓	✓	✓	✓	✓

2 Related Work

Document Understanding Research in document understanding field has gradually evolved from single-page visual question answering to long-document, multi-page, and multimodal reasoning (Han et al., 2025; Hu et al., 2025). Early benchmarks such as DocVQA (Mathew et al., 2021) focus on understanding text, layout, and visual content in document images. MP-DocVQA (Tito et al., 2023) extends this setting to multi-page documents, requiring models to retrieve relevant information from a larger page range before answering questions. More recent benchmarks further increase document length and reasoning complexity. MMLongBench-Doc (Ma et al., 2024) evaluates long multimodal documents with text, figures, and tables, and includes cross-page and unanswerable questions. LongDocURL (Deng et al., 2025) further integrates long-document understanding, numerical reasoning, and content localization into a unified benchmark. These studies have significantly advanced long-context and multimodal document understanding (Chia et al., 2025; Wang et al., 2025a). However, their tasks mainly focus on general or professional documents and usually adopt domain-independent question designs. As a result, they do not capture the distinct reasoning patterns of scientific disciplines, such as equation-based reasoning in mathematics and physics or experiment-oriented reasoning in biology and statistics. In addition, existing benchmarks often evaluate individual abilities, such as answer generation or evidence retrieval, rather than providing a unified evaluation of multimodal understanding, refusal, complex reasoning, multi-document reasoning, and evidence grounding.

Scientific Question Answering Recent works have progressed from general science knowledge to the understanding of real scientific papers and reports (Auer et al., 2023; Wang et al., 2025c; Astaraki et al., 2026). ScienceQA (Lu et al., 2022) introduces a large-scale multimodal benchmark covering natural science, social science, and language science, with questions mainly collected from educational scenarios. QASPER (Dasigi et al., 2021) moves scientific QA to research papers by introducing information-seeking questions over NLP papers together with supporting evidence annotations. Subsequent benchmarks further incorporate richer scientific content and more complex reasoning. SPIQA (Pramanick et al., 2024) further incorporates figures, tables, and text from scientific papers and enables large-scale evaluation of multimodal scientific QA, with its data mainly drawn from computer science. More recently, PaperScope (Xiong et al., 2026) constructs a multimodal document benchmark over AI papers and evaluates retrieval, reasoning, summarization, and problem solving in scientific literature. Although existing benchmarks have extended scientific QA to multimodal and long-document settings (Wang et al., 2024), their disciplinary coverage remains limited, often focusing on NLP, computer science, or AI (Lee et al., 2023). Moreover, their reasoning questions rarely reflect discipline-specific requirements, and existing evaluation frameworks lack a unified assessment of basic QA, unanswerable detection, discipline-specific reasoning, and evidence localization. These limitations hinder comprehensive evaluation of reliable scientific document understanding across diverse disciplines.

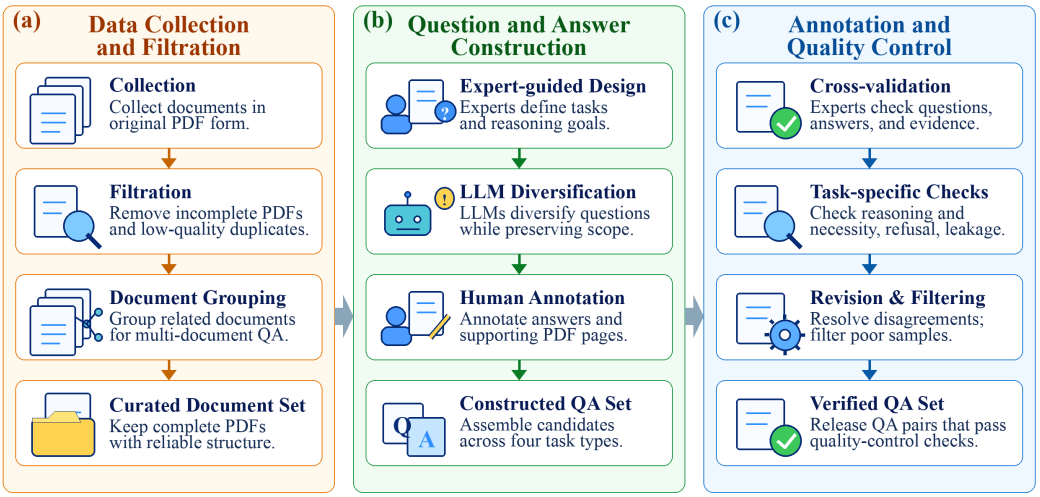


Figure 2: Overview of the ScienceDoc dataset construction pipeline.

3 Dataset

3.1 Data Collection and Filtration

Document collection. We collect scientific documents from multiple open-access sources, including arXiv, conference proceedings, and scientific journals. The collection spans 8 disciplines: *Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, Electrical Engineering and Systems Science, Economics, and Quantitative Finance*. Documents are retained in their original PDF format to preserve complete multimodal content and document structure. To reduce domain imbalance, we monitor the number of documents across both primary disciplines and fine-grained research areas and supplement underrepresented fields during collection. Figure 2 summarizes the overall construction pipeline of ScienceDoc.

Document filtration. We apply a multi-stage filtration process to ensure that the collected documents are suitable for multimodal scientific QA. The quality criteria considers 4 aspects: *file integrity*, *content completeness*, *document quality*, and *duplication*. We first remove files that could not be opened or parsed correctly, as well as documents containing missing, corrupted, or duplicated pages. We then examine document completeness, including the availability of the main body, figures, tables, equations, and experimental content when applicable. Documents with abnormal structures, extremely limited scientific content, or substantial rendering errors are excluded. Finally, we identify and remove duplicate publications, different versions of the same work, and highly overlapping documents based on their metadata and content. There are 693 documents that are retained after filtration. Figure 3 presents their distribution across the 8 disciplines and 61 scientific fields. Details in Appendix A.6.

Multi-Document organization. To construct the multi-document task, we organize two or three source documents into a document group according to their disciplinary relevance and potential information complementarity. We avoid grouping near-duplicate documents or documents whose content is largely redundant. Instead, each group is designed to contain distinct but related evidence that can support comparison, aggregation, or joint reasoning across sources. Candidate groups are further checked during question construction to ensure that each included document contributes necessary information to the corresponding question. This design reduces single-document shortcuts and requires models to identify and integrate evidence across document boundaries. Details in Appendix A.10.

3.2 Question and Answer Construction

Expert-guided question design. ScienceDoc adopts an expert-led and model-assisted construction paradigm. Rather than directly asking large language models to generate QA pairs from scientific

documents, domain experts first define the discipline-specific question types and reasoning requirements for different disciplines. The design considers both the characteristics of scientific documents and the types of information commonly required in scientific reading, such as experimental settings, quantitative results, methodological details, equations, tables, figures, comparisons, and relationships between multiple pieces of evidence. Based on these criteria, experts manually construct the initial questions for each document and identify the information that should be required to answer them. This process ensures that the benchmark reflects meaningful scientific information needs rather than generic questions that can be produced from surface-level document content.

Task types. ScienceDoc contains 4 task types. *General task* assesses information extraction and understanding across text, equations, tables, figures, and multiple pages. *Unanswerable task* is semantically relevant to the document but cannot be answered using the available content, allowing us to evaluate whether models can recognize insufficient evidence rather than generate unsupported answers. Their reference answer is "Unanswerable". *Reasoning task* requires combining multiple facts through comparison, conditional inference, multi-step calculation, constraint satisfaction, or causal reasoning, with questions solvable mainly by common knowledge or question cues excluded. *Multi-Document task* requires complementary evidence from two or three documents, and each source must be necessary for deriving the final answer; otherwise, the question is revised or discarded. Construction details in Appendices A.7–A.10.

Answer construction. ScienceDoc adopts a free-form answer format to reduce the possibility of models exploiting answer choices or other task-specific cues. Reference answers may contain numerical values, units, equations, lists, or concise textual descriptions. All final answers are manually annotated from the source documents rather than generated directly by models. Annotators also identify the minimum evidence set required to derive each answer, enabling the benchmark to evaluate both answer correctness and evidence grounding. More details in Appendices A.3–A.4.

3.3 Annotation and Quality Control

Professional annotation. A total of 10 annotators participate in the professional annotation and cross-validation process. Annotators review the source documents and manually determine the reference answers and their supporting evidence. The annotation protocol emphasizes scientific precision, including numerical values, units, mathematical symbols, experimental conditions, comparison targets, and other context that could affect the interpretation of an answer. For questions involving multiple evidence sources, annotators are required to verify that the complete answer could be derived only when all necessary evidence is considered.

Cross-validation. Each QA instance is independently checked by at least two human experts. The reviewers verify 4 key aspects: question validity, answer correctness, evidence sufficiency, and evidence necessity. Specifically, they examine whether the question is clear and scientifically meaningful, whether the reference answer is fully supported by the document, whether the annotated evidence covers all information required to answer the question, and whether irrelevant evidence has been included. Questions with inconsistent annotations or reviewer disagreement are returned for further inspection and revision before inclusion in the final benchmark.

Task-specific verification. Additional checks are applied according to the task type. For reasoning questions, reviewers verify that the intermediate facts and reasoning dependencies are valid and that removing a required fact would prevent derivation of the complete answer. For multi-document questions, reviewers check the contribution of each source document and reject questions that could be answered using only a subset of the documents. For unanswerable questions, reviewers examine the complete source document to confirm that the required information is genuinely unavailable. We also remove questions containing direct answer leakage, excessive lexical cues, ambiguous conditions, incomplete evidence, or substantial semantic overlap with other questions.

Manual review system. To support consistent annotation and quality control, we develop an internal review interface that jointly presents the question, reference answer, evidence annotations, and source documents. Annotators can directly inspect the supporting content and revise or remove problematic samples. The system records annotation changes and review histories, allowing the complete construction process to remain traceable. Only QA instances that pass the required expert cross-validation are included in the final ScienceDoc benchmark. Review system in Appendices B.1–B.3 and B.5.

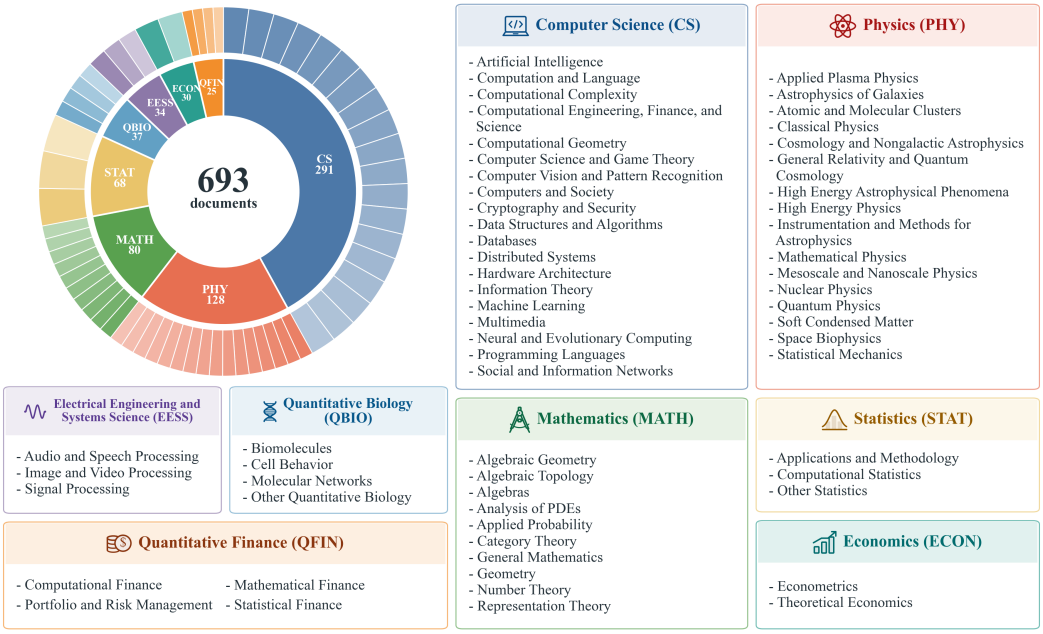


Figure 3: The disciplines and scientific fields of documents in ScienceDoc.

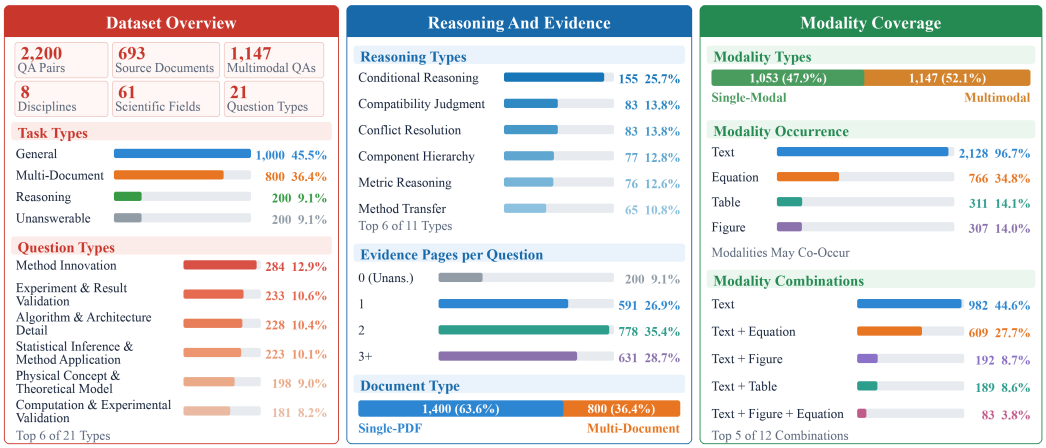


Figure 4: ScienceDoc statistics on question composition, reasoning and evidence, and modality.

3.4 Dataset Statistics

Overall composition. ScienceDoc contains 693 scientific documents and 2,200 questions, including 1,000 General questions (45.5%), 200 Unanswerable questions (9.1%), 200 Reasoning questions (9.1%), and 800 Multi-Document questions (36.4%), as shown in Figure 4. Reasoning and Multi-Document questions together account for 45.5% of the benchmark, providing substantial coverage of complex reasoning and information integration. The dataset spans 8 disciplines and 61 scientific fields (shown in Figure 3), enabling broad evaluation of model performance across diverse scientific domains. Statistics and construction details in Appendices A.1–A.2 and A.5.

Question types and reasoning types. Beyond the 4 task types, ScienceDoc includes 21 discipline-specific question types describing different scientific information needs. The most frequent are Method Innovation (284 questions), Experiment & Result Validation (233), and Algorithm & Architecture Detail (228). The dataset also contains 600 relation-labeled cases, with conditional inference being the most common (155; 25.8%), followed by compatibility judgment and conflict resolution

(83 each; 13.8%). These relations characterize reasoning patterns and are not limited to the 200 reasoning questions, as they also occur in multi-document questions.

Modality and evidence distribution. ScienceDoc contains substantial multimodal and cross-page evidence. Among all questions, 876 (39.8%) involve multiple modalities and 1,324 (60.2%) involve a single modality. Text appears in 2,180 questions, equations in 349, tables in 322, and figures in 246, with overlapping counts for multimodal questions. Of the 2,000 answerable questions, 1,409 (70.5%) require evidence from multiple pages, including all reasoning and multi-document questions. Moreover, 409 General questions also require cross-page evidence. These statistics highlight the need to evaluate evidence localization together with answer correctness.

4 Experimental Setup

Models: We evaluate 11 representative MLLMs to comprehensively assess current capabilities on our benchmark. For analysis and presentation, we divide the evaluated models into two scale groups: *Small-Scale Models* ($<$ 10B) and *Large-Scale Models* ($>$ 10B). The evaluated small-scale models include Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025b), InternVL2.5-8B (Chen et al., 2024a), MiniCPM-V 4.5 (Yu et al., 2026), and MiniCPM-V 2.6 (Yao et al., 2024). We additionally evaluate large-scale models, including Mistral-Small-3.1-24B-Instruct-2503, Gemma 3 27B IT (Team et al., 2025), GLM-4.6V (Hong et al., 2025), Claude Sonnet 5, and DeepSeek-V4-Flash-Vision-Exp(abbreviated as DeepSeek-V4-Flash thereafter) (Xu et al., 2026). These models span multiple model families and parameter scales, providing broad coverage of representative MLLMs. More details of models in Appendix E.

Settings: We evaluate all models under a unified input, prompting, and evaluation protocol. Each PDF is rendered at 144 DPI. All models are evaluated on a single NVIDIA A800 GPU, except for GLM-4.6V, Claude Sonnet 5, and DeepSeek-V4-Flash-Vision-Exp, which are accessed through their official APIs. Prompts and run settings in Appendices F and C.4.

Evaluation: Our benchmark evaluates both answer correctness and evidence localization. Because semantically equivalent answers may differ substantially in wording and formatting, we do not rely on exact string matching. Following existing studies (Zheng et al., 2023; Liu et al., 2023; Zhu et al., 2025), we use the LLM-as-a-Judge method to compare each prediction against the reference answer. For Unanswerable QA, a response is correct only if the model correctly determines that the answer cannot be derived from the provided documents. We further evaluate evidence localization using **E_Precision**, **E_Recall**, and **E_F1**, corresponding to precision, recall, and F1 score computed over PDF page indices. We also report **A_Pages**, the average number of unique pages selected as evidence. Inference and scoring details in Appendices C.1–C.3.

5 Results

5.1 Main Results

Overall accuracy generally improves with model scale, while ScienceDoc remains challenging across different model sizes. As shown in Table 2, larger models generally achieve higher overall accuracy, with GLM-4.6V and DeepSeek-V4-Flash-Vision reaching 80.95% and 79.45%, respectively. Among models below 10B parameters, the Qwen3-VL series performs particularly well: Qwen3-VL-4B and Qwen3-VL-8B achieve 64.91% and 68.32%, respectively, outperforming several larger models. However, none of the evaluated models below 10B exceeds 70% accuracy, indicating that ScienceDoc remains challenging for smaller MLLMs despite recent advances.

Performance varies substantially across task types, revealing distinct challenges in multi-document integration, reasoning, and unanswerable-question recognition. General questions are comparatively easier for most models, with the strongest models exceeding 80% accuracy. multi-document questions remain more challenging: 7 of the 11 models achieve below 50% accuracy, indicating difficulty in integrating complementary evidence across independent documents. Reasoning performance also varies widely, showing large differences in the ability to combine multiple pieces of evidence through multi-step reasoning. Unanswerable questions exhibit a particularly uneven pattern, with accuracy ranging from 2.50% to 72.50%. Several models that perform strongly

Table 2: Overall question answering and evidence localization performance. All denotes accuracy over the complete benchmark; Gen., Unans., Multi-Doc., and Rea. denote general answerable QA, unanswerable QA, Multi-Document QA, and reasoning QA, respectively. Best and second-best results are shown in **bold** and underline, respectively.

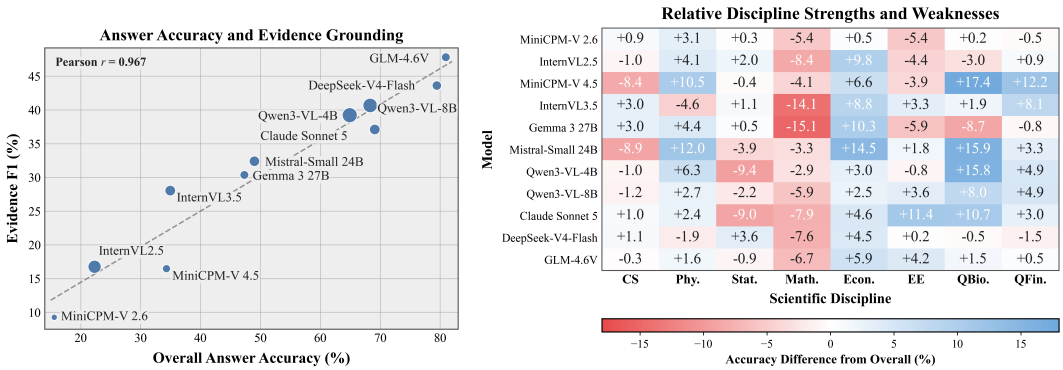
Model	All	Breakdown of Task Types (%)				Evidence Localization			
		Gen.	Unans.	Multi-Doc.	Rea.	E-Prec.	E-Rec.	E-F1	A-Pages
Small-Scale Models (<10 B)									
MiniCPM-V 2.6	15.59	14.70	38.50	12.50	9.50	9.37	9.24	9.26	0.11
InternVL2.5-8B	22.32	24.40	32.00	14.25	34.50	15.39	31.88	16.75	6.69
MiniCPM-V 4.5	34.32	54.60	15.00	10.75	46.50	20.03	15.92	16.47	1.42
InternVL3.5-8B	34.95	36.30	2.50	36.88	53.00	30.86	31.20	28.04	3.72
Qwen3-VL-4B	64.91	80.00	60.00	47.38	64.50	39.35	**53.01**	39.22	9.26
Qwen3-VL-8B	68.32	83.20	42.50	55.38	71.50	43.34	49.63	40.66	8.44
Large-Scale Models (>10 B)									
Gemma 3 27B	47.32	57.00	14.00	43.62	47.00	34.02	30.85	30.36	1.98
Mistral-Small-3.1-24B	49.00	72.40	9.00	24.88	68.50	34.51	36.39	32.39	3.58
Claude Sonnet 5	69.09	70.50	**72.50**	61.00	**91.00**	36.26	44.56	37.10	3.66
DeepSeek-V4-Flash	79.45	88.20	15.50	**83.62**	83.00	44.49	49.22	43.61	2.57
GLM-4.6V	**80.95**	**88.50**	46.00	79.62	83.50	**51.04**	50.47	**47.81**	1.97

Table 3: Accuracy (%) of different MLLMs across 8 disciplines. CS, Phy., Stat., Math., Econ., EE, QBio., and QFin. denote Computer Science, Physics, Statistics, Mathematics, Economics, Electrical Engineering and Systems Science, Quantitative Biology, and Quantitative Finance, respectively. Best and second-best results are shown in **bold** and underline, respectively.

Model	All	Fundamental & Computational Sciences				Applied & Interdisciplinary Sciences			
		CS	Phy.	Stat.	Math.	EE	Econ.	QBio.	QFin.
Small-Scale Models (<10 B)									
MiniCPM-V 2.6	15.59	16.51	18.73	15.85	10.20	10.16	16.06	15.79	15.12
InternVL2.5-8B	22.32	21.34	26.39	24.32	13.88	17.97	32.12	19.30	23.26
MiniCPM-V 4.5	34.32	25.91	44.85	33.88	30.20	30.47	40.88	51.75	46.51
InternVL3.5-8B	34.95	37.99	30.34	36.07	20.82	38.28	43.80	36.84	43.02
Qwen3-VL-4B	64.91	63.89	71.24	55.46	62.04	64.06	67.88	80.70	69.77
Qwen3-VL-8B	68.32	67.11	70.98	66.12	62.45	71.88	70.80	76.32	73.26
Large-Scale Models (>10 B)									
Gemma 3 27B	47.32	50.34	51.72	47.81	32.24	41.41	57.66	38.60	46.51
Mistral-Small-3.1-24B	49.00	40.13	60.95	45.08	45.71	50.78	63.50	64.91	52.33
Claude Sonnet 5	69.09	70.07	71.50	60.11	61.22	80.47	73.72	79.82	72.09
DeepSeek-V4-Flash	79.45	80.54	77.57	**83.06**	71.84	79.69	83.94	78.95	77.91
GLM-4.6V	**80.95**	**80.67**	**82.59**	80.05	**74.29**	**85.16**	**86.86**	**82.46**	**81.40**

on answerable questions achieve much lower accuracy on unanswerable subset, suggesting that recognizing when the available documents do not provide sufficient evidence is a distinct capability. We present case study in Appendix D.1.

Models achieve lower scores on evidence localization than on answer correctness while they differ significantly in the selection of supporting pages. As shown in Table 2, GLM-4.6V achieves the highest Evidence F1 of 47.81%, while most models remain below 40%, despite substantially higher answer accuracy. This gap indicates that generating correct answers does not necessarily imply accurate identification of supporting evidence. Models also exhibit different evidence-selection strategies. Qwen3-VL-4B and Qwen3-VL-8B retrieve 9.26 and 8.44 pages per question on average, achieving higher recall but lower precision. In contrast, GLM-4.6V and DeepSeek-V4-Flash retrieve fewer pages while achieving higher Evidence F1 scores. These results highlight the importance of precise evidence localization over broad retrieval for reliable scientific document QA. Further case studies and error analysis are provided in Appendices D.1 and D.2.



(a) Answer accuracy versus evidence grounding. Marker size indicates the average number of predicted evidence pages.

(b) Model-specific relative strengths and weaknesses across 8 disciplines. Cell values are subject accuracy minus the model's overall accuracy.

Figure 5: Analysis of model performance in evidence grounding and across scientific disciplines.

5.2 Findings and Analysis

Model performance varies substantially across scientific disciplines. Table 3 shows that the overall model ranking remains broadly consistent across disciplines, with GLM-4.6V and DeepSeek-V4-Flash-Vision leading overall and the Qwen3-VL series performing strongest among small-scale (below 10B) models. However, the relative-performance heatmap in Fig. 5b reveals noticeably different domain profiles across models. Mathematics is the most consistent relative weakness, with all evaluated models performing below their own overall accuracy, whereas Physics, Economics, and Quantitative Biology frequently emerge as relative strengths. These variations indicate that similar aggregate accuracy can mask substantially different patterns of discipline-specific robustness.

Answer accuracy is strongly correlated with evidence-localization quality (Pearson $r = 0.967$). As shown in Figure 5a, the correlation between overall accuracy and Evidence F1 suggests that both capabilities depend on effective document understanding and information retrieval. However, clear deviations from this trend indicate that models differ in how precisely they associate answers with supporting evidence. Retrieving more pages also does not necessarily improve performance, as stronger models tend to benefit from more selective and accurate evidence localization rather than broader coverage. These results show that evidence localization provides a complementary measure of model reliability beyond answer accuracy alone. Examples in Appendix D.1, Cases 4 and 7.

Performance on unanswerable questions reveals clear differences in hallucination resistance that are not captured by overall QA accuracy. DeepSeek-V4-Flash-Vision achieves 79.45% overall accuracy but only 15.50% on the unanswerable questions (in Table 2). These questions require models to recognize when the provided documents contain insufficient evidence and avoid generating unsupported answers, making them a useful indicator of hallucination resistance. The results show that stronger QA performance does not necessarily lead to more reliable abstention, providing an important additional view of model reliability under insufficient evidence. The case study in Appendix D.1, Case 10.

6 Conclusion

This paper presents ScienceDoc, a multidisciplinary benchmark for evidence-grounded multimodal question answering over scientific documents. With broad disciplinary coverage, diverse question types, and expert-annotated evidence pages, ScienceDoc enables joint evaluation of answer correctness and evidence grounding. Evaluation of 11 representative MLLMs reveals challenges in multi-document reasoning, unanswerable-question recognition, disciplinary robustness, and evidence localization. The results further demonstrate that high answer accuracy does not always guarantee precise evidence grounding, highlighting the need for hierarchical and comprehensive evaluation of scientific document understanding. We hope ScienceDoc can support future research on trustworthy scientific reasoning and facilitate new scientific discoveries.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with the linguistic diversification of expert-designed questions during dataset construction. Domain experts first defined the discipline-specific question types and reasoning requirements and manually developed the initial questions. Generative AI tools were then used to produce diverse linguistic expressions of these questions. All questions, reference answers, and supporting evidence were manually annotated, reviewed, and validated by human experts. We did not use generative AI tools for research conceptualization, study design, data analysis, or manuscript writing. Other disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ScienceDoc is constructed from publicly accessible scientific documents and will be released under the MIT License for academic research purposes. All human annotators are informed of the dataset's intended use and provided their consent to participate in the annotation process. We respect the intellectual property rights of the original document authors and encourage responsible use of the dataset in accordance with applicable licenses and ethical guidelines.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, Section 3 and Appendix A provide detailed descriptions of the ScienceDoc dataset construction, question design, and annotation procedures. Section 4 and Appendices C, E, F document the experimental setup, evaluation protocol, baseline models, and prompt templates. Appendix B further describes the quality control and dataset release procedures. The dataset and evaluation code will be made publicly available upon acceptance to facilitate independent verification and future research.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Mahdi Astaraki, Mohammad Arshi Saloot, Ali Shiraee Kasmaee, Hamidreza Mahyar, and Soheila Samiee. When iterative rag beats ideal evidence: A diagnostic study in scientific multi-hop question answering. *arXiv preprint arXiv:2601.19827*, 2026.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13 (1):7240, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the association for computational linguistics*, 6:587–604, 2018.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Maojia Song, Chaoqun Liu, Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9244–9261, 2025.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, 2021.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, 2025.

Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106:102306, 2024. ISSN 1566-2535.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025.

Tanveer Hannan, Dimitrios Mallios, Parth Pathak, Faegheh Sardari, Thomas Seidl, Gedas Bertasius, Mohsen Fayyaz, and Sunando Sengupta. Docslm: A small vision-language model for long multimodal document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9337–9346, 2026.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5817–5834, 2025.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124, 2021.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pp. 19036–19052. PMLR, 2023.

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*, pp. 15419–15446, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, 2024b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15630–15640. IEEE, 2024.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*, 2023.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591. IEEE, 2022.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833, 2024.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1776–1826, 2022.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

An-Lan Wang, Jingqun Tang, Lei Liao, Hao Feng, Qi Liu, Xiang Fei, Jinghui Lu, Han Wang, Hao Liu, Yuliang Liu, et al. Wilddoc: How far are we from achieving comprehensive and robust document understanding in the wild? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23002–23012, 2025a.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19162–19170, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2025c.

Lei Xiong, Huaying Yuan, Zheng Liu, Zhao Cao, and Zhicheng Dou. Paperscope: A multi-modal multi-document benchmark for agentic deep research across massive scientific papers. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8015–8040, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Yanyu Yao. A survey on multimodal large language models. In *Computer Science Undergradaute Conference 2025@ XJTU*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Hanwen Zheng, Sijia Wang, Chris Thomas, and Lifu Huang. Advancing chart question answering with robust chart component recognition. *arXiv preprint arXiv:2407.21038*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

APPENDIX CONTENTS

A Dataset Construction and Representation **15**

A.1 Detailed Dataset Statistics 15

A.2 Construction Lineage 16

A.3 PDF and Page Representation 16

A.4 Construction Records and Evaluation Inputs 16

A.5 Construction Stages and Quality Gates 16

A.6 Document Retrieval, PDF Organization, and Basic QA 17

A.7 Construction of Unanswerable Questions 18

A.8 Single-Document Reasoning Candidate Generation 19

A.9 Reasoning Difficulty Filtering and Closed-Book Ablation 19

A.10 Multi-Document Grouping, Generation, and Source Ablation 20

B Quality Control and Dataset Release **21**

B.1 Model-Assisted Dual Review and Evidence Recovery 21

B.2 Human Final Review and Front-End Validation 21

B.3 Freezing the Final 2,200-Item Set 22

B.4 Data Card and Scope 23

B.5 Pre-Release Checklist 24

C Evaluation Protocol and Reproduction **24**

C.1 Formal Model Inference Protocol 24

C.2 Output Validity, Correction, and Raw Records 25

C.3 Semantic Judging and Fixed-Denominator Reporting 26

C.4 Run Configuration and Reproduction Order 29

D Qualitative Cases and Error Analysis **29**

D.1 Qualitative Study Cases 29

D.2 Error Taxonomy and Analysis Method 39

E Baseline Models **40**

F Prompt Templates **41**

A Dataset Construction and Representation

This section supplements the dataset overview in Section 3 with detailed statistics, construction provenance, record representation, and task-specific question construction.

A.1 Detailed Dataset Statistics

Figure 6 provides a task-wise breakdown that complements the composition overview in Figure 4. Panels (a) and (b) count questions rather than source documents: the four task types total 2,200 questions, and 1,455 questions (66.1%) fall outside Computer Science. The discipline distribution therefore differs in unit from the 693-document distribution in Figure 3.

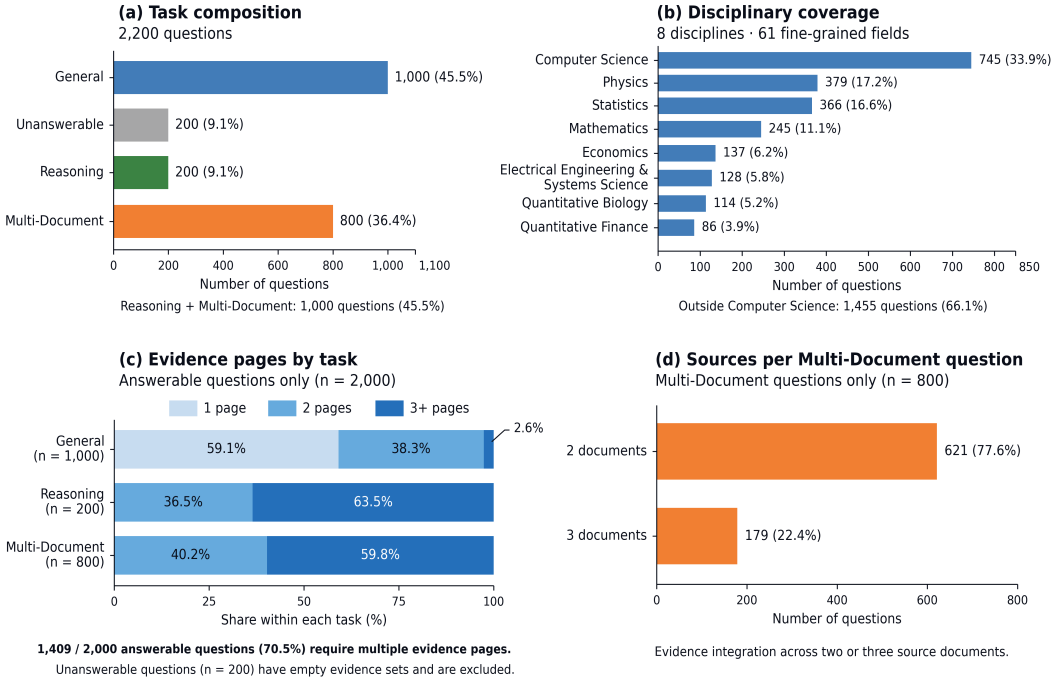


Figure 6: Dataset statistics of ScienceDoc. (a) Question counts by task. (b) Question counts by primary discipline. (c) Annotated evidence-page counts within each answerable task, normalized by its own sample size; the 200 Unanswerable questions have empty evidence sets and are excluded. (d) Number of source documents required per Multi-Document question. Percentages in (a) and (b) use all 2,200 questions, while (d) uses the 800 Multi-Document questions.

Panel (c) normalizes each evidence-page distribution by its task size: 1,000 General, 200 Reasoning, and 800 Multi-Document questions. General questions have one, two, or at least three evidence pages in 591, 383, and 26 cases, respectively. Reasoning questions have two pages in 73 cases and at least three in 127; the corresponding Multi-Document counts are 322 and 478. Thus, 1,409 of the 2,000 answerable questions (70.5%) require multiple evidence pages. The same 1,409 questions represent 64.0% of all 2,200 questions in Figure 4; the difference is solely the denominator. The 200 Unanswerable questions have empty evidence sets and are excluded from panel (c).

Panel (d) uses the 800 Multi-Document questions as its denominator: 621 require two source documents and 179 require three. Source-document counts and evidence-page counts measure different aspects of the task, since each source may contribute multiple evidence pages. This distinction explains why the evidence-page and source-count distributions need not coincide.

A.2 Construction Lineage

The released benchmark is the result of a larger construction pool rather than a random sample of the initial questions. Document acquisition initially produced 6,204 candidate QA pairs from 703 documents. Format normalization, duplicate removal, page-range validation, and evidence review produced a 4,211-item single-document baseline over the retained 693-document corpus. This baseline supplied verified facts for subsequent General, unanswerable, and Reasoning construction, while additional grouping and source-necessity checks produced Multi-Document candidates. These historical counts describe intermediate pools only; they are not evaluation splits and are not used as denominators in the reported experiments.

Source-document counts and evaluation-PDF counts use different units: a merged PDF is an evaluation container, not an additional scientific document. The 693-document total in Section 3 counts source documents, not merged evaluation containers.

A.3 PDF and Page Representation

Source PDFs preserve their original text, equations, tables, figures, and layout. Related document-VQA benchmarks have likewise emphasized visual document understanding across heterogeneous layouts and multiple document pages or images Mathew et al. (2022); Tanaka et al. (2023).For Multi-Document items, two or three source PDFs are concatenated in a fixed order without changing the content of their pages. The construction ledger records the source-document identifier, source-document order, source page, and corresponding merged page. Gold and predicted evidence always use PDF page indices in the evaluation PDF; printed page numbers in headers, footers, or the document body are never used as evidence indices.

Each evaluation record contains a question, concise reference answer, annotated evidence pages, required evidence modalities, task type, question type, and answer format. Reasoning records additionally retain the source facts and their inferential relation. Multi-Document records retain the source-document map and the contribution required from each document. These audit fields support construction and review; the model-facing task remains the question plus PDF pages and the structured answer contract specified in Section C.1.

A.4 Construction Records and Evaluation Inputs

Intermediate generation batches, the single-document baseline, closed-book diagnostics, and candidate-review outputs are construction records. Evaluation uses only the four reviewed components described in the main text. Their release-file identities and freeze checks are specified in Section B.3.

The following construction details expand the expert-led workflow in Section 3. Human–model interaction has also been explored as a mechanism for constructing and refining challenging benchmark instances Kiela et al. (2021). Quality control and release checks are collected in Appendix B, evaluation and reproduction in Appendix C, and qualitative analysis in Appendix D.

A.5 Construction Stages and Quality Gates

Table 4 expands the construction overview in Figure 2 into operational stages. Each stage records input identities, processing dates, retained and rejected counts, model settings where applicable, and reasons for returning an item to revision.

A candidate advances only after passing its stage-specific gate; a returned item may be revised or deleted. Structural checks validate required fields, page bounds, source identifiers, and duplicates keyed by both normalized question text and document identity. The resulting ledger separates candidate generation from human gold annotation: model-proposed answers and pages are review inputs, and evaluation outputs never redefine the reference answer.

Table 4: Construction stages, outputs, and quality gates.

Stage	Input	Executor / model	Key output	Condition for next stage
1. Document retrieval	Open-access platform categories and date ranges	Program + manual review	Document metadata and download list	Domain quotas, duplication checks, and licensing checks pass
2. PDF audit	Downloaded PDFs	PyMuPDF + manual review	Page count, file consistency, anomalous pages	File opens successfully and page indexing is stable
3. Basic QA	Single-document PDFs	Domain experts	Expert-written initial questions, answer requirements, and candidate evidence	Fields complete and evidence exists
4. Short-answer conversion	Mixed-format QA	Rules + manual review	Short-answer QA	No answer-choice leakage
5. Unanswerable	Document topic and missing facts	Domain experts + model-assisted rewriting + full-document review	Exact-refusal items	No sufficient evidence anywhere in the full document
6. Reasoning	Expert-designed question and verified QA from the same document	Domain experts + large language model assistance	Candidates combining necessary facts from 2–4 verified QA items	Every selected fact is indispensable; no conclusion leakage or stable question-only shortcut
7. Multi-Document	Expert-designed question and document groups	Domain experts + model-assisted expansion	Two-/three-document candidates	Every source is indispensable
8. Model-assisted dual review	Candidates + full PDFs	Two independent large language models	Dual-pass result or return reasons (dual KEEP means both independent reviewer models judge the item retainable)	Proceeds to human annotation and cross-validation, not direct release
9. Evidence recovery	Passed candidates + PDFs	Program + model + manual review	page and fact-level ledger	Page indices, sources, and modalities are consistent
10. Human final review	Final candidates	10 annotators; at least two human experts independently verify each item	Keep, modify, delete, and cross-validation records	Every assigned item is closed out
11. Freeze and evaluate	Four official JSON files + PDFs	11 MLLM runs + large language model Judge + fixed-denominator reporting	Per-item outputs for 11 models, answer and evidence-page metrics, strict reports, and exception records	Final 2,200 QA and PDFs pass preflight; missing predictions remain in the denominator; Judge and input identities are traceable

A.6 Document Retrieval, PDF Organization, and Basic QA

The retrieval ledger stores the title, abstract, open-access platform category, assigned domain, filename, download status, and page count. Category quotas guide acquisition; insufficient coverage can trigger a wider date range while retaining deduplication and readability checks. Table 5 specifies the checks applied before semantic review. Field and page representations follow Section A.

Modality labels describe evidence needed for the answer: `text` alone is used when textual evidence is sufficient, while `equation`, `table`, or `figure` is added only when that modality is required.

Table 5: Structural and content checks for documents and basic QA.

Object	Check	Failure handling
File	Openability, page count, duplicate files	Add to exception list; exclude from question construction
Question	Complete, unambiguous, no answer leakage	Rewrite or delete
Answer	Facts, units, precision, direction	Return to the PDF and correct
Evidence pages	Pages, direct support, minimally sufficient	Add missing pages, remove redundant pages, or delete the item
Modality	`text/figure/table/equation`	Relabel according to the actual evidence

Reviewers remove background-only pages and retain a minimally sufficient set of direct supporting pages.

A.7 Construction of Unanswerable Questions

Following prior work on unanswerable QA Rajpurkar et al. (2018), experts anchor each candidate in a method, variable, dataset, or experimental setting explicitly present in the document, then identify a specific missing relation or value. The queried relation must connect to the research objective; absent personal details or unrelated facts are rejected as off-topic. Model assistance is limited to rewriting or proposing variants of this expert-defined information need.

The generation model receives the expert-designed question and missing-information requirement, together with the title, abstract, section headings, entity list, and summaries of verified answerable QA. Its output must include a nonempty question, identifiable `anchored_entities`, a concrete `missing_evidence` description, and an auditable rationale. These fields describe a candidate, not a gold answer. The complete prompt used for expert-guided unanswerable candidate generation is shown in Figure 7.

For retained items, the reference answer is exactly `Unanswerable` and `evidence_pages` is `[]`. Under PDF inference, a refusal with a nonempty page list violates the output contract. The exact label provides a consistent representation of the evidence-insufficiency decision described in the main text.

UNANSWERABLE CANDIDATE GENERATION PROMPT

You assist domain experts in refining an unanswerable question for a scientific-document QA dataset.

Use the expert-designed question and its specified missing-information requirement, together with the supplied document title, abstract, section headings, entity list, and verified QA summaries.

Rewrite the question while preserving the expert's intended scientific meaning, answer scope, and evidence dependency. Do not independently choose a new information need or supply a gold answer.

The question must concern a method, variable, dataset, metric, or experimental condition present in the supplied material.

Do not ask about personal information, publication trivia, or unrelated common knowledge.

Do not invent an entity, premise, or scientific fact.

Flag any possibility that the answer is stated or uniquely derivable by calculation, chart reading, or a standard definition. Summaries alone cannot establish absence from the full document; two human experts must independently inspect the complete PDF before retention.

Return JSON only with question, rationale, anchored_entities, missing_evidence, and rejection_risks. State the precise unavailable relation or parameter in missing_evidence. These are candidate-review fields; humans determine the final reference answer and evidence. Retained Unanswerable items have reference answer "Unanswerable" and evidence_pages [].

Figure 7: Prompt for expert-guided generation of unanswerable-question candidates.

Two human reviewers independently check the entire supplied PDF, including appendices, captions, footnotes, and any included supplementary pages. Each records the topic anchor, searched locations, missing fact, and a nearby statement that does not close the gap. A candidate is rejected as unanswerable if a value can be read reliably from a plot or derived uniquely from the stated parameters, even when the answer is not written verbatim. Disagreement requires renewed inspection of specific pages. Closed-book screening (Section A.9) can additionally reveal question cues or common-knowledge shortcuts; its outputs are retained only as construction diagnostics.

A.8 Single-Document Reasoning Candidate Generation

Experts specify the intended reasoning relation and necessary facts from 2–4 verified QA records belonging to the same document. Related multi-hop QA benchmarks similarly emphasize compositional reasoning over multiple supporting facts and explicit reasoning dependencies Yang et al. (2018); Ho et al. (2020). A large language model receives the expert-designed question, required reasoning relation, verified source questions and answers, and document identity, without PDF pages or evidence indices, and proposes a candidate formulation. Supported relation labels are `causal_chain`, `conditional_inference`, `comparison`, `constraint_intersection`, and `multi_step_calculation`.

Candidate records contain `reasoning_type`, `intermediate_facts`, a proposed `answer`, and `source_qa_ids`. Validation confirms that all source IDs exist, refer to the same document, and contribute necessary facts. The answer must follow through an explicit operation or relation rather than concatenating source answers, and the new question must neither copy a source question nor reveal its conclusion. Human reviewers subsequently determine the reference answer and evidence from the PDF. The prompt used for reasoning-candidate generation is shown in Figure 8.

REASONING GENERATION SYSTEM PROMPT

You assist domain experts in refining a scientific reasoning question within one document.

Inputs are an expert-designed question, its intended reasoning relation, and 2-4 verified source QA records from the same document. You do not receive the PDF. Use only the supplied facts.

Preserve the expert's information requirement, answer scope, and evidence dependency. Propose a formulation that requires the specified comparison, conditional inference, causal chain, constraint intersection, or multi-step calculation.

Every selected source fact must be necessary to derive one coherent conclusion. Reject independent subquestions, concatenated source answers, unsupported assumptions, invented causal links, and ambiguous conclusions. Do not copy a source question or expose source QA IDs in the question or answer.

For numerical reasoning, preserve the stated values, units, conditions, and operation. Do not reinterpret a numerical difference as causal contribution without explicit support. Do not place all answer-bearing premises or the conclusion in the question: the document must remain necessary.

Return JSON only with question, a proposed answer, reasoning_type, intermediate_facts, and source_qa_ids. reasoning_type must be one of causal_chain, conditional_inference, comparison, constraint_intersection, or multi_step_calculation.

The proposed answer is a review aid, not a reference annotation. Human experts must verify the derivation, remove-one-fact necessity, final reference answer, and minimally sufficient evidence pages in the full PDF.

Figure 8: Expert-guided prompt for reasoning-candidate generation.

A.9 Reasoning Difficulty Filtering and Closed-Book Ablation

Filtering checks source identity, answer leakage, and completeness of the derivation. Reviewers decompose a candidate into premises, operations, and conclusion, rejecting unstated assumptions or

multiple possible answers. Per-hop ablation removes one required fact at a time and checks whether the complete conclusion still follows. Evidence recovery checks page support for each remaining step. As reported in Section 3, all retained Reasoning questions require evidence from multiple pages; no additional minimum distance between these pages is imposed.

The `question_only` diagnostic sends only the question to two large language models. Consistently correct closed-book answers flag possible question leakage, common knowledge, or insufficient document dependence and trigger revision or exclusion. Failure in this diagnostic alone does not establish difficulty under PDF input.The closed-book diagnostic prompt is shown in Figure 9.

CLOSED-BOOK DIAGNOSTIC PROMPT

Question: {question}

This is a closed-book diagnostic: no document/PDF is provided. Make a concise best-effort answer from internal knowledge. Do not output Unanswerable merely because the PDF is absent.

If the question itself is intentionally unanswerable, output exactly "Unanswerable"; never use a synonym or case variant. Output only the final answer, with no reasoning or explanation.

Figure 9: Prompt for the closed-book diagnostic.

Closed-book outputs, screening decisions, and revision reasons remain in the construction ledger and are excluded from the main benchmark scores. The reported experiments use the PDF protocol; Full, Oracle, and evidence-page ablations are page-selection policies within that protocol, not additional tasks.

A.10 Multi-Document Grouping, Generation, and Source Ablation

Experts group two or three documents by disciplinary relevance and complementary evidence, then define the information required from each source. A relation ledger records source-document identifiers, document order, title keywords, each source's factual contribution, and the expected relation. Large language models use the expert-designed question, specified source contributions, verified QA text, and document titles to propose candidate formulations; full PDFs are reserved for review and evidence recovery.

Candidate relations include comparisons under a common metric, component compatibility, shared failure conditions, constraint intersections, methodological differences, and cross-study validation. Juxtaposed answers, generic shared background, and links requiring external knowledge are rejected.

For source ablation, reviewers remove each document in turn and record which part of the complete answer becomes unsupported. A three-document candidate must require all three documents; a question answerable from a subset is revised or discarded. The source-to-merged-page map defined in Section A is checked independently so that document order cannot introduce evidence-index errors.The prompt used for multi-document candidate generation is shown in Figure 10.

MULTI-DOCUMENT GENERATION SYSTEM PROMPT

You assist domain experts in refining a Multi-Document scientific question.

Inputs are an expert-designed question, the intended relation and required contribution of each source, verified QA facts, and a map of two or three documents. You do not receive PDFs. Use only the supplied facts.

Preserve the expert's information requirement, answer scope, and evidence dependency. Each candidate must require complementary evidence from every included document to answer the same scientific question.

Valid relations include comparison under a common metric, aggregation, conditional inference, constraint intersection, methodological differences, and cross-study validation. Reject generic shared background, independent document summaries, concatenated source answers, and invented links or assumptions.

For a comparison, each document must supply a necessary comparison target or fact.

For compatibility or method transfer, prerequisites and system properties must be explicitly supported. Do not infer causation from correlation.

Apply a remove-one-document test: if any source can be omitted while the complete answer remains derivable, revise or reject the candidate. A three-document question must require all three sources. Do not expose source QA IDs or leak the conclusion in the question.

Return JSON only with question, a proposed answer, intermediate_facts, and source_qa_ids. Keep the proposed answer concise and sufficient for the question; do not require unrequested reasoning traces. Human experts independently annotate the final answer and minimal evidence pages from the full PDFs and verify source necessity and merged PDF page indices.

Figure 10: Expert-guided prompt for multi-document candidate generation.

B Quality Control and Dataset Release

This section details the review workflow and release checks supporting the professional annotation and independent human cross-validation described in Section 3.

B.1 Model-Assisted Dual Review and Evidence Recovery

Two large language models independently receive the same question, proposed answer, and full PDF pages. They do not see one another's decisions. Each record contains `decision`, `confidence`, `reason`, model identity, and page-level `supported_fact`.

Dual KEEP advances a candidate to evidence recovery. FIX requires revision and resubmission; REJECT routes the item to human inspection or deletion. This is a conjunctive filter, not a majority vote, and never bypasses the human cross-validation required in the main text. Table 6 lists the review dimensions.

Evidence recovery links each intermediate fact to its `PDF_PAGE`; Multi-Document records additionally store `source_doc_number` and `source_page`. A modality audit then applies the evidence-necessity rule in Section A.6.

B.2 Human Final Review and Front-End Validation

The review interface implements the professional annotation and independent cross-validation described in the main text. Reviewers first confirm `dataset_id`, document identity, and `qa_id`, then inspect the question and fields alongside the corresponding PDF (Figure 11).

Question checks cover entities, time ranges, comparison targets, direction, units, and output format. Answer checks cover numerical values, signs, percentages, abbreviations, names, and conditions. Evidence checks require reviewers to state, page by page, which fact in the answer is supported by

Table 6: Independent model-assisted review checks and final gates.

Review item	Large model 1	Large model 2	Final gate
Answer entailment	Independent full-document reading	Independent full-document reading	Both pass
Evidence completeness	Fact-level page indices	Fact-level page indices	No missing hop
Source necessity	Remove-document test	Remove-document test	Every included source is indispensable
Question leakage	Closed-book answerability	Question-cue check	No direct answer leakage
Modality	Visual evidence on pages	Visual evidence on pages	Human confirmation of canonical value

that page; pages that provide only background but do not support the answer are removed. Reasoning items undergo per-hop ablation, whereas Multi-Document items undergo per-source ablation.

Reviewers can keep, modify, delete, or defer an item. Before modification or deletion, the system saves a complete snapshot; after write-back, it records the account, time, previous value, new value, and reason. If the underlying file changes during review, the reviewer must reopen and reassess it rather than overwriting another person's modifications.

Table 7: Manual-review actions and audit records.

Action	When used	System record
Keep	Question, answer, evidence, and dependencies are all correct	Status and reviewer
Modify	The correct value can be determined uniquely from the PDF	Full pre-modification snapshot and updated value
Delete	The question is invalid or cannot be repaired reliably	Deletion tombstone and snapshot
Defer	PDF anomaly, multiple possible repairs, or unclear source	Remains incomplete and is escalated

The interface flags missing fields, out-of-range or duplicate page indices, and noncanonical labels. Semantic acceptance remains a human decision; every modification, deletion, or deferral requires a recorded reason.

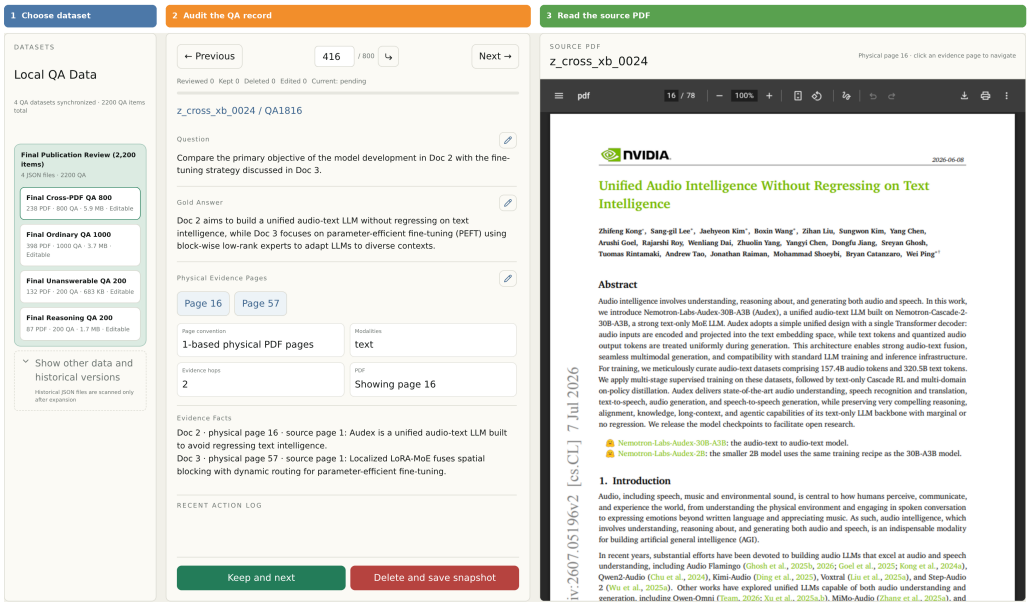
B.3 Freezing the Final 2,200-Item Set

Table 8 maps the components in the main text to canonical release files; the General component retains the filename `ordinary_qa.json`. Each item has a unique, stable identifier within its component, and provenance maps retain the identities used in the reported experiments.

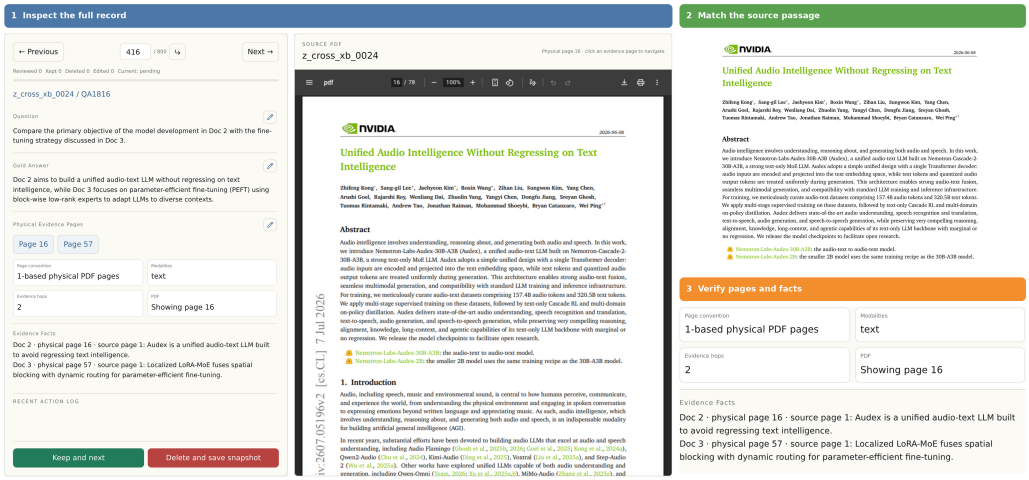
Table 8: Files and roles in the frozen ScienceDoc release.

Component	QA count	Primary capability	Official file
General	1,000	Single-document localization and inference	`ordinary_qa.json`
Unanswerable	200	Evidence absence and refusal	`unanswerable_qa.json`
Reasoning	200	Single-document multi-step reasoning	`reasoning_qa.json`
Multi-Document	800	Two-/three-document relational reasoning	`cross_pdf_qa.json`

Content freezing confirms the reviewed question, answer, minimal evidence, modalities, and source dependencies. Structural freezing verifies shared fields, canonical values, component counts, item identity consistency, PDF readability, and page bounds. The release manifest records file identities and hashes. Subsequent edits require traceable review and refreshed affected statistics, examples, inference outputs, and reports.



(a)



(b)

Figure 11: Manual-review interface. (a) Overview: reviewers select the dataset and item on the left, verify the question and answer in the center, and read the corresponding PDF on the right. (b) Detailed evidence inspection: reviewers check pages, modalities, and source passages that support the answer.

B.4 Data Card and Scope

Consistent with established practices for transparent dataset documentation Bender & Friedman (2018); Gebru et al. (2021); Pushkarna et al. (2022), we document data provenance, release conditions, construction choices, and evaluation considerations to clarify intended use and support reproducibility.

Source PDFs remain subject to their authors' and platforms' permissions. Release materials preserve source metadata and redistribution notes; review accounts, session tokens, and internal logs are excluded.

ScienceDoc combines broad disciplinary coverage with expert-defined task and evidence annotations. Its question distribution reflects source availability and expert design. Alternative valid evidence pages can also receive imperfect set-based scores, so the retained page-level audit supports inspection of such cases. Rendering, font encoding, visual-token limits, and long-document memory requirements can affect model inputs; resource failures are recorded separately from semantic and evidence-selection errors.

B.5 Pre-Release Checklist

Table 9: Pre-release checks for the final ScienceDoc benchmark.

Check	Pass condition
Files	All four official JSON files exist with fixed names
Counts	1000/200/200/800, totaling 2200
Identity	Component and item identifiers are unique, with traceable provenance
Refusal	Only exact `Unanswerable`; page list is empty
PDF	All files exist and file-consistency checks match
Page indices	Positive integers, PDF pages, and within range
Prompt protocol	Only `question_only` and `pdf`
Inference records	Raw outputs and records are complete
Judge binding	Every output is bound to the corresponding inference record
Report	Fixed denominator; illegal outputs are not removed
Human final review	Every assigned item is closed out
Documentation	Statistics and figure captions match the final 2,200-item data

Release checks sample all four components, page-count ranges, and modality combinations, comparing QA content, evidence, captions, and reported statistics. A reviewer independent of the latest edits checks the record before release. Changes to QA, PDFs, prompts, or scoring invalidate the affected inference/Judge bindings and require the corresponding results and reports to be refreshed.

C Evaluation Protocol and Reproduction

This section expands the input protocol, answer and evidence scoring, and reproduction settings used in the main experiments. Construction-time diagnostics remain separate from the reported benchmark results.

C.1 Formal Model Inference Protocol

Long-context benchmarks have highlighted the need to evaluate models over extended single- and multi-document inputs rather than isolated short contexts Bai et al. (2024). The model roster, rendering resolution, hardware, and API deployment follow the main Experimental Settings. Before inference, preflight locks QA and PDF hashes, question identities, model identity, prompt version, and page-selection policy.

PDF pages are provided as images, each preceded by an external `[PDF_PAGE_N]` label. Models may use only these external labels when filling `evidence_pages`; printed page numbers inside the document cannot be copied. The requested output is a single JSON object with no explanation, Markdown, or additional fields. The common response contract below supplements the task-specific templates in Appendix F; source-boundary labels use the index `PDF_PAGE_N`. The common PDF-mode response contract is shown in Figure 12.

PDF-MODE SYSTEM PROMPT

Answer the question using only the supplied scientific-document PDF page images and the task-specific instructions.
External PDF_PAGE_N boundaries identify 1-based PDF pages. Return the integer N, not the label text or page numbers printed inside a document. For concatenated documents, use indices in the merged PDF.
Return exactly one JSON object:

```
{
  "answer_pre": "<complete but concise final answer>",
  "evidence_pages": [<PDF page numbers>]
}
```

Include all and only pages that materially support the requested answer. Use a sorted list of unique positive integers. A consecutive range is justified only if every selected page supplies necessary evidence. Reading scope is not evidence.
Preserve requested entities, units, signs, precision, conditions, and comparison direction. Do not substitute a nearby supported fact for the question as written. Do not include chain-of-thought, derivations, Markdown, additional fields, or text outside the JSON.
For General/Unanswerable QA, determine answerability from the complete supplied document. If the required answer cannot be derived, return exactly:

```
{
  "answer_pre": "Unanswerable",
  "evidence_pages": []
}
```

Reasoning and Multi-Document questions are verified to be answerable from the supplied PDFs. Integrate all required facts before giving the final answer and supporting PDF pages.

Figure 12: Common response contract for PDF-mode inference.

C.2 Output Validity, Correction, and Raw Records

The requested generation contract and the scoring parser have distinct roles. The scoring implementation first reads parsed prediction fields, then attempts recovery from raw generation fields, including supported JSON fragments, code fences, and nested containers. It normalizes whitespace and recovers unique evidence-page indices without using gold answers to repair predictions. Raw generations and parsing provenance remain available for inspection.

A format deviation alone does not force an otherwise recoverable answer to be incorrect. Unrecoverable or empty answers are technical failures and receive zero answer credit. Evidence is evaluated independently before answer judging, so recovered pages can still contribute evidence scores when answer recovery fails. Missing predictions remain in the fixed denominator. Table 10 summarizes this handling.

Table 10: Prediction recovery and independent answer/evidence scoring.

Output issue	**Normalization allowed?**	**Final handling**
Whitespace	Normalize for judging	No automatic answer credit
Repeated evidence indices	Recover a unique page set	Score the recovered set
Large consecutive span	Do not prune using gold pages	Score every recovered page
Extra explanation or Markdown	Recover supported answer/JSON fields	Send recoverable answers to the semantic judge
Refusal with page indices	Recover answer and pages separately	Judge the answer; score the predicted page set
Unrecoverable or empty answer	No answer repair from gold	Zero answer credit; recovered evidence still scored
Missing prediction	No imputation	Zero contribution in the fixed denominator

C.3 SEMANTIC JUDGING AND FIXED-DENOMINATOR REPORTING

Following recent work on LLM-based semantic evaluation Kim et al. (2024), and the semantic evaluation criterion in the main text, a large language model compares each recoverable prediction with its human-annotated reference answer, including Unanswerable items. It returns `CORRECT`, or `INCORRECT`. The question determines which facts, values, units, conditions, and comparisons are required: the reference is a semantic guide, not a checklist of incidental details. Concise sufficient answers are accepted, while missing required components or material contradictions are penalized. No exact-text, alias, numeric-tolerance, or string-similarity shortcut can mark an answer correct before this review. Evidence pages never influence the answer verdict.

For item i, let P_i and G_i be the sets of predicted and gold PDF pages. For nonempty sets,

$$p_i = \frac{|P_i \cap G_i|}{|P_i|}, \qquad r_i = \frac{|P_i \cap G_i|}{|G_i|}, \qquad f_i = \frac{2p_i r_i}{p_i + r_i}.$$

The F1 value is zero when $p_i + r_i = 0$. If both sets are empty, all three values are one; if exactly one is empty, all three are zero. With fixed benchmark size N, E-Precision, E-Recall, and E-F1 are the respective per-item means, reported as percentages, and $\mathrm{A_Pages} = N^{-1} \sum_i |P_i|$. Missing predictions contribute zero.

The complete semantic-judge prompt is shown in Figures 13–15.

SEMANTIC-JUDGE PROMPT (1/3)

```
You are a strict semantic reviewer for PDF question answering.
Follow the frozen review protocol below exactly.

[Frozen Review Protocol]
You are performing frozen semantic answer review for a PDF-QA benchmark.

Apply exactly the same standard to every model, dataset, and review session.

You will receive:
- Dataset name
- Question
- Gold answer
- Model prediction answer

Judge ONLY the semantic answer.
Evidence pages are evaluated separately by deterministic code and MUST NOT
influence the answer verdict.

Return exactly ONE label:
CORRECT
INCORRECT

========================================================================
A. CENTRAL CALIBRATION RULE: GOLD IS A REFERENCE, NOT A CHECKLIST
========================================================================

First determine what the QUESTION explicitly requires.
Then use the Gold answer as the semantic reference for those required parts.

The prediction does NOT need to reproduce every detail, example, implementation
detail, explanatory sentence, parenthetical remark, number, method name, or
supporting fact that happens to appear in the Gold answer.

A prediction may be much shorter than the Gold answer and still be CORRECT if
it accurately and sufficiently answers every aspect explicitly required by the question.

Do NOT downgrade an answer merely because the Gold answer is more detailed.

However, do not accept a generic restatement of the question. The prediction
must provide the requested information, distinction, relation, mechanism,
entity, value, direction, condition, or conclusion.
```

Figure 13: Semantic-judge prompt (Part 1 of 3).

SEMANTIC-JUDGE PROMPT (2/3)

```
============================================================
B. CORRECT
============================================================

Output CORRECT when:
- every semantically necessary component explicitly requested by the question
  is answered correctly;
- the answer reaches the same requested conclusion as Gold;
- no material statement in the prediction contradicts the Gold answer.

For comparison questions:
- both requested sides must be addressed;
- the essential requested contrast must be correct;
- supporting Gold details that the question does not require may be omitted.

For "how" / mechanism questions:
- the essential mechanism or process asked for must be stated;
- incidental implementation details present only in Gold are not mandatory.

For yes/no or existence questions:
- the correct yes/no conclusion can be sufficient when that is what the
  question asks, even if Gold additionally explains why.

For numerical/entity questions:
- the requested value/entity must be correct.
- extra Gold context is not mandatory unless the question asks for it.

For multi-hop/reasoning questions:
- a bridge entity, condition, intermediate relation, causal step, comparison
  side, or document-specific relation is mandatory ONLY when the wording of
  the question actually requires that component to answer the target.
- do not require hidden intermediate reasoning that is not requested in the
  answer itself.


============================================================
C. INCORRECT
============================================================

Output INCORRECT when the main answer/direction is substantially correct but at
least one EXPLICITLY REQUIRED component is missing or locally wrong, while the
response still demonstrates meaningful correct understanding.

Typical incomplete-answer cases:
- one required comparison side is incomplete;
- a required condition/qualification is missing;
- the requested mechanism is only partly specified;
- one required sub-answer is omitted;
- a minor numerical/equation/detail error occurs while the main requested
  conclusion remains correct;
- the answer correctly identifies the broad distinction but does not provide
  enough information to satisfy the question as written.

Do NOT use INCORRECT just because the prediction is less detailed than Gold.

Also output INCORRECT when:
- the core requested conclusion is wrong;
- the requested entity/value/sign/direction/ordering is wrong;
- a central requested mechanism or causal/comparative relation is wrong;
- documents, methods, or roles are materially mismatched;
- the answer is unrelated, fabricated, or contradicted by Gold;
- the response is so generic/vacuous that it does not actually answer the
  requested target;
- it merely repeats information already stated in the question without
  supplying the requested answer;
- an answerable question receives no substantive semantic answer.
```

Figure 14: Semantic-judge prompt (Part 2 of 3).

SEMANTIC-JUDGE PROMPT (3/3)

```
================================================================
D. DATASET ANSWERABILITY
================================================================

General / Unanswerable:
- Gold = Unanswerable and prediction = Unanswerable -> CORRECT.
- Gold = Unanswerable but prediction gives a substantive answer -> INCORRECT.
- Gold is answerable but prediction says Unanswerable -> INCORRECT.

Multi-Document / Reasoning:
- every item is answerable;
- any prediction of Unanswerable -> INCORRECT.
================================================================
E. FINAL CONSISTENCY CHECK
================================================================

Before selecting the label, check:

1. What exact information does the QUESTION ask for?
2. Did the prediction answer all of those required parts?
3. Is any required part materially wrong?
4. Am I penalizing the prediction only because Gold contains extra detail?
   If yes, do NOT downgrade for that reason.
5. Am I accepting a vague paraphrase that does not actually answer the target?
   If yes, do NOT mark it CORRECT.

Use CORRECT for a concise but sufficient answer.
Use INCORRECT for a substantively right but explicitly incomplete answer.
Use INCORRECT for a materially wrong, unrelated, or non-answer.

Output exactly one label and nothing else:
CORRECT
INCORRECT

[Dataset]
{dataset_id}

[Question]
{question}

[Gold Answer]
{gold_answer}

[Model Prediction]
{model_prediction}

Return exactly one label and nothing else:
CORRECT
INCORRECT
```

Figure 15: Semantic-judge prompt (Part 3 of 3).

Reporting retains fixed denominators of 1,000 General, 200 Unanswerable, 200 Reasoning, and 800 Multi-Document questions (2,200 overall). The evaluation package reads the frozen experiment result components, keeping the human reference fields separate from prediction fields, and records input, prompt, code, and judge-cache identities. Later release edits must not replace the references used for a reported experiment during reproduction. Missing or unrecoverable answers receive zero answer credit; evidence remains separately scored when recoverable. An unparseable judge response stops scoring rather than being counted as a model error. Answer accuracy and evidence metrics are reported separately as in the main tables; exact page-set match and format diagnostics are not additional headline metrics.

C.4 Run Configuration and Reproduction Order

A new evaluation proceeds from frozen QA/PDF validation to inference, semantic judging, and report validation. Reproduction of the reported tables uses the original experiment inputs and corresponding recorded semantic judgments through the evaluation package; it does not substitute subsequently edited release annotations. A resumed run must match its input and model/prompt identities. Table 11 lists inference and judge settings; deployment follows the main Experimental Settings.

Table 11: Key run parameters and their roles.

Parameter	**Current role**
`input_mode`	`pdf` for reported benchmark results; `question_only` only for construction-time closed-book diagnostics
`page_input_policy`	`full`, `oracle`, or `qa_field`
`dpi`	Fixed 144-DPI rendering
`max_pdf_pages`	`0` means no artificial truncation
`inference_max_new_tokens`	Controls structured-answer length
`judge_max_new_tokens`	32 tokens; one label: `CORRECT` or `INCORRECT`
`expected_qa_sha256`	Locks the QA input
`expected_pdf_corpus_sha256`	Locks the PDF corpus
Judge decoding	Greedy decoding

D Qualitative Cases and Error Analysis

This section complements the main results with qualitative cases and an error taxonomy for interpreting answer correctness and evidence localization.

D.1 Qualitative Study Cases

The main quantitative results in Table 2 use the fixed 2,200-question ScienceDoc benchmark. The ten cases below provide qualitative diagnosis of answer generation and page-level evidence localization; they do not recompute or replace the aggregate results. Each header gives the legacy document and item identity retained in the provenance record. Evidence excerpts come from the cited PDF pages; for Unanswerable QA, a displayed distractor page is not gold evidence. The selected model outputs reproduce recorded page predictions and shortened answers. Correct/incorrect display labels indicate whether the semantic verdict receives answer credit, rather than a separate binary judging protocol.

The cases expose complementary behaviors: whole-document citation, Multi-Document semantic integration, source-confusion in a three-document chain, answer–evidence disagreement in a reasoning question, and over-prediction on a multi-page scientific question. The same model can therefore answer correctly while localizing evidence poorly, which motivates reporting answer correctness and evidence localization separately.

D.1.1 CASE 1: WHOLE-DOCUMENT CITATION IN SINGLE-DOCUMENT QA

Case 1: Document-ID 426 QA-ID gen_685

Document Context

Page.2

where $\varphi > 0$ in the air domain Ω_0 and $\varphi < 0$ in the liquid domain Ω_1. The evolution of the interface under a velocity field $\boldsymbol{u}(\boldsymbol{x}, t)$ is governed by the level set advection equation:

$$\frac{\partial \varphi}{\partial t} + \boldsymbol{u} \cdot \nabla \varphi = 0. \quad (2)$$

This implicit representation naturally accommodates complex topological changes such as merging and splitting without explicit remeshing [1, 47, 49], making level set methods indispensable for a wide range of computational physics applications [44, 51, 17, 18]. Despite their versatility, classical level set methods suffer from two fundamental limitations when solving Eq. (2) numerically. First, numerical diffusion during advection causes substantial volume loss, particularly for thin structures and small-scale features undergoing large deformation [12, 15, 8]. Even with high-order spatial discretizations such as ENO [54] or WENO [36, 25] schemes, the repeated interpolation and reinitialization steps introduce cumulative errors that erode interface fidelity over time. Second, the discrete grid representation lacks the resolution to capture sub-grid geometric details, leading to the gradual erosion of sharp corners and filaments that are crucial for visual and physical fidelity [21, 53].

Particle Level Set. These limitations motivate the development of numerical methods assisted by Lagrangian structures that can better preserve interface geometry over long time horizons. The hybrid particle level set (PLS) method [12] employs marker particles distributed near the interface within a narrow band to correct the signed distance function φ. Each particle p carries its own signed distance value φ_p and is advected along the flow:

$$\frac{d\boldsymbol{x}_p}{dt} = \boldsymbol{u}(\boldsymbol{x}_p, t). \quad (3)$$

When discrepancies arise between the particle values φ_p and the grid values $\varphi(\boldsymbol{x}_p)$, the particles provide correction information to reduce volume loss and maintain sharper interfaces [46, 22, 23, 29, 38, 16]. Subsequent extensions have further improved accuracy and robustness: for instance, oriented and gradient-augmented variants [65, 16] embed additional geometric information such as the interface normal $\mathbf{n} = \nabla\varphi/|\nabla\varphi|$ or gradient $\boldsymbol{g} = \nabla\varphi$ within particles to better preserve curvature and local detail. In addition, the one-layer PLS formulation [23, 65, 73, 74] simplifies particle management by confining particles to a single layer near the interface

2

Page.5

By discretizing the flow map into a set of Lagrangian particles that advect both position and Jacobian information, PFM obtains accurate flow maps implicitly without auxiliary computation or historical velocity buffers. The transfer of information between particles and the Eulerian grid follows the Affine Particle-In-Cell (APIC) style formulation [24], which enables particles to carry higher-order information transferred to grid nodes via weighted Taylor expansion $q_i = \sum_p [q_p + \nabla q_p \cdot (\boldsymbol{x}_i - \boldsymbol{x}_p)]\omega_{ip} / \sum_p \omega_{ip}$, reducing numerical dissipation and achieving high-order accuracy in long-range transport without global map reconstruction.

Recent developments have applied the particle flow map methods, both particle-based and grid-based, to transport impulse and other advected invariants across diverse fluid phenomena, including incompressible flows [75, 40, 11], compressible flows [6], vortical flows [68, 67], and solid-fluid coupling [5, 32]. Methodologically, extensions include Clebsch-gauge formulations [34], gradient-buffer-free transport [35], GPU-accelerated hybrid coupling [66], and differentiable simulation frameworks [33]. However, the application of particle flow maps to interface tracking problems, particularly level set methods with narrow-band confinement and high-order geometric preservation, remains unexplored.

Particle flow map methods achieve long-range transport accuracy through bidirectional flow maps, but have not been applied to level set interface tracking. In contrast to "virtual-particle" approaches [22, 12, 15, 13] that reconstruct flow trajectories through backward tracing incurring significant temporal buffering and substepping overhead PFM establishes a compact, physically consistent bridge between Lagrangian advection and Eulerian projection. This motivates our central question: can we combine the high-order geometric representation of GALS with the long-range flow map accuracy of PFM to achieve high-fidelity level set advection while maintaining the efficiency of narrow-band formulations?

1.2. Particle Flow Map Level Set

Building upon these insights, we present a Particle Flow Map Level Set (PFM-LS) method for high-fidelity level-set interface tracking. By transporting φ, $\nabla\varphi$, and $\nabla^2\varphi$ on particles according to their respective differential forms, we achieve a high-order geometric representation that remains accurate under complex flow deformations. The proposed framework confines particles to a narrow band $\Omega_\epsilon = \{\boldsymbol{x} : |\varphi(\boldsymbol{x})| <$

5

Question

Why does PFM-LS outperform PLS in volume preservation?

Gold Answer

Particles are primary carriers of geometry with flow-map exact transport.

Case Metadata

Dataset:	General QA
Domain:	Mathematics / Applied Probability
Modalities:	Text
Question Type:	Inferential / Proposition Proof & Logical Reasoning
Gold Pages:	2, 5

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Qwen3-VL 4B	**Correct**	[1-51] (all)	PFM-LS transports level-set values, gradients, and Hessians along particle trajectories, preserving geometry ...
Qwen3-VL 8B	**Correct**	[6, 23, 25, ...]	PFM-LS transports values and derivatives along particle trajectories and avoids PLS interpolation errors ...
GLM-4.6V	**Correct**	[1, 2, 6, 44]	The particle flow map tracks high-order information along accurate trajectories and reduces numerical diffusion ...
DeepSeek V4-Flash	**Correct**	[1, 5, 6, 44]	Particles carry higher-order geometry through bidirectional flow maps instead of serving only as local corrections ...

Figure 16: Case 1: whole-document citation in single-document QA.

All four displayed answers capture the central mechanism, but their evidence behavior differs sharply. Qwen3-VL-4B cites all 51 pages, Qwen3-VL-8B cites 13 pages without either gold page, and the two more compact predictions recover only one of the two gold pages. Correct answer semantics therefore do not imply selective or complete evidence localization.

D.1.2 CASE 2: MULTI-DOCUMENT INTEGRATION WITH PAGE-LOCALIZATION ERRORS

Case 2: Document-ID 1675 QA-ID md_430

Document Context

Page.8

allowing all queries to attend to a designated anchor frame or broadcasting the first frame's local attention map to other frames for cross-view interaction.

Query-Probe Top-K Kernel. For vertical-line heads, where attention mass is concentrated on a few critical key tokens (forming several prominent lines in the attention map), we restrict computation to attention scores between all queries and these key tokens, omitting the remaining interactions. As analyzed in Sec. 3, the small set of key tokens is dynamic and input-dependent, which prevents the use of a fixed attention pattern. Therefore, we estimate key positions on the fly using a lightweight pre-computation, as detailed in Sec. 4.3.

DINO Top-K Kernel. Correspondence heads exhibit clear semantic behavior, attending to relevant patches at corresponding spatial locations across frames. We observe that DINO patch features already capture such semantic information and can serve as effective correspondence estimators, as visualized in the appendix. Therefore, we pre-compute the cosine similarity between each patch token and all other patch tokens using the DINO features, and select the Top-K most similar tokens from each frame as the keys and values for attention. Empirically, this strategy provides strong coverage of the full attention distribution with a relatively small K.

Uniform Sampling Kernel. For scanning heads, attention patterns are diverse and broadly distributed across tokens. Unlike other head types that focus on a small subset of keys, scanning heads aggregate information from a wide range of spatial locations without clear structural regularity, making pattern-specific key selection prone to missing important interactions. Therefore, following prior work [23, 26], we adopt a uniform sampling kernel that selects key positions with a fixed stride. This strategy provides an effective approximation by evenly covering the token space. In practice, it performs well at relatively high sampling ratios (*i.e.*, retaining 25%–50% of keys), but its performance degrades rapidly as the ratio decreases.

4.2 Offline Head Profiling

With tailored sparse kernels for each attention type, the remaining challenge is how to assign each attention head to the appropriate category to fully exploit sparsity while maintaining performance. To this end, we propose a simple yet effective sparsity search strategy based on progressive local substitution. Specif-

Page.56

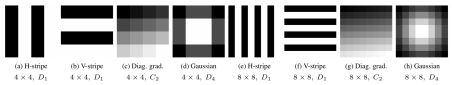


Figure 3: **Synthetic symmetry probes with D_4 subgroup annotations.** Controlled probes for measuring geometric sensitivity and anisotropy in ViT attention under horizontal reflection.

4.5.2 SYNTHETIC AND REAL-IMAGE ATTENTION ANALYSIS

To examine the geometric effect of hidden-state flipping more directly, we consider both controlled synthetic probes and real-image attention visualizations. The synthetic probes isolate geometric structure, while the real-image study tests whether semantic behavior is preserved.

Synthetic probe evaluation. We construct synthetic inputs with explicit geometric structure, including horizontal and vertical stripes, diagonal gradients, and Gaussian-like blobs on 4×4 and 8×8 grids. Using a pretrained ViT-B/16, we compare the attention response of each probe to that of its horizontally flipped counterpart via

$$S^{(\ell)} = \frac{1}{M} \sum_{i=1}^{M} \text{cosine}\Big(A^{(\ell)}(x_i), A^{(\ell)}(T(x_i))\Big), \tag{22}$$

where $A^{(\ell)}$ denotes the Block 5, Head 0 attention field and T is horizontal reflection. As shown in Table 3, stripe and gradient patterns achieve only moderate consistency, whereas Gaussian probes achieve perfect consistency. Together with Fig. 3, these results suggest that the selected ViT attention field responds more consistently to isotropic patterns than to directional ones under horizontal reflection, revealing measurable anisotropy.

Real-image attention perturbation. We next horizontally flip the Block 5 patch-token representation of a pretrained ViT-B/16 and propagate the modified state through the remaining encoder blocks. Figure 4 shows that the resulting attention remains focused on the same semantic object parts, but shifts to the mirrored location. Thus, for this selected attention field, the intervention alters spatial routing without disrupting semantic focus. In all qualitative visualizations, we report Head 0 of the selected transformer block; the rationale is provided in Appendix L.

4.5.3 FLIPPED-HEAD ATTENTION IN DIT: QKV CONSISTENCY VS. MISMATCH

To connect the ViT analysis to Diffusion Transformers (DiTs), we perform a controlled synthetic experiment that isolates the geometric effect of hidden-state flipping inside attention. This is a controlled mechanistic analysis rather than a full DiT benchmark; it illustrates how geometric mismatch can become problematic when attention is applied repeatedly along the denoising trajectory.

We compare the two attention interventions defined earlier: the *attention-inconsistent* output-only flip of Eq. 5, and the *attention-consistent* reference corresponding to Eq. 6, implemented via coherent transformation of the interacting attention pathways. We consider both a symmetric and a non-symmetric synthetic baseline: the former tests geometric stability, while the latter tests geometric correctness under reflection. Figure 5 shows the same qualitative conclusion in both settings. Under the attention-consistent transformation, the attention field either preserves the baseline geometry (symmetric case) or yields the correct mirrored response (non-symmetric case). By contrast, the output-only flip produces shifted and distorted responses, because routing is computed in one coordinate frame while the resulting head output is interpreted in another. The distortion heatmap $|A_{\text{incon}} - A_{\text{con}}|$ makes this deviation explicit. These

Question

How do Doc 1 and Doc 3 characterize the relationship between the semantic features of ViT patch representations and their spatial routing or correspondence behavior?

Gold Answer

Doc 1 shows that semantic ViT patch features support spatial correspondence across views. Doc 3 shows that semantic focus can be preserved while spatial routing follows a geometric transformation such as a horizontal flip.

Case Metadata

Dataset:	Multi-Document QA
Domain:	Computer Science / Computer Vision
Modalities:	Text + Figure
Question Type:	Inferential / Experiment & Result Validation
Gold Pages:	8, 56

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Qwen3-VL 4B	**Incorrect**	[1, 3]	Semantic content and spatial relationships jointly determine heterogeneous attention behavior across heads and layers.
Qwen3-VL 8B	**Incorrect**	[57]	Semantic features are not directly tied to routing; spatial behavior is determined by input geometry ...
GLM-4.6V	**Correct**	[6, 7, 56, 57]	Semantic features guide correspondence, while geometric transformation mirrors routing and preserves semantic focus.
DeepSeek V4-Flash	**Correct**	[7, 56]	Correspondence heads are semantically structured; flipping the hidden state preserves focus but mirrors routing.

Figure 17: Case 2: Multi-Document semantic integration with page-localization errors.

The two Qwen predictions rely on pages 1 and 3 or page 57 and fail to recover the required relation between Doc 1 page 8 and Doc 3 page 56. GLM-4.6V and DeepSeek-V4-Flash-Vision answer correctly and include page 56, but both shift the Doc 1 support to neighboring pages rather than the annotated page 8. Multi-Document evaluation must therefore test both semantic integration and exact source-page attribution.

D.1.3 CASE 3: READING SCOPE DOES NOT ENSURE CORRECT THREE-DOCUMENT SYNTHESIS

Case 3: Document-ID 1529 QA-ID md_198

Document Context

Page.6

is represented by p. Action a_m is the real action applied to regulate the charging and discharging of the ESS.

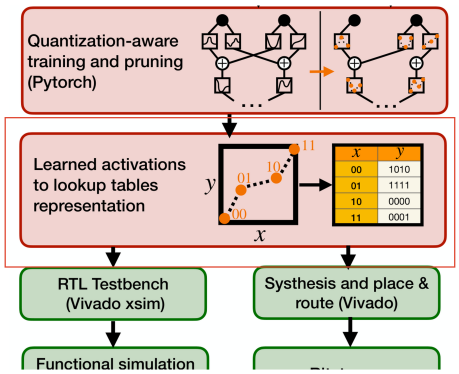
$$\tilde{a}_m = \beta\, a_m \tag{33}$$

$$\tilde{E}_m^{rated} = \beta E_m^{rated} \tag{34}$$

Second, an optimization model that allows independently trained agents to coordinate voltage magnitude regulation by dynamically adjusting actions, mitigating excessive charging or discharging. During online application, we assume that each agent shares information with its neighboring agents Ω, including action a_t and the estimated voltage magnitude $\tilde{V}_t$.

The objective function in (35) is designed to minimise adjustments of actions during online coordination, preserving consistency with the agent's originally planned actions to avoid disrupting the ESS's subsequent charging and discharging schedule, while ensuring global voltage magnitude regulation. The optimization model is subject to constraints, including voltage magnitude limits

Page.14



Page.45

ing, and only transitions that satisfy all constraints under the applied disturbance – thus guaranteeing robustness – are retained. These verified transitions are then used to construct and certify safe control inputs at deployment.

4.4.2. Control Synthesis and Lipschitz Deviation Bounds

Given a query state $\boldsymbol{x}(k)$, the interpolated control input is computed via:

$$\boldsymbol{u}^*(k) = \sum_{(\bar{\boldsymbol{x}},\ \bar{\boldsymbol{u}},\ \bar{\boldsymbol{x}}^+)\in\mathcal{D}(\boldsymbol{x}(k))} w(\bar{\boldsymbol{x}})\cdot\bar{\boldsymbol{u}} \tag{12}$$

where the weights satisfy:

$$w(\bar{\boldsymbol{x}}) \propto \frac{1}{\|\boldsymbol{x}(k)-\bar{\boldsymbol{x}}\|+\epsilon} \tag{13a}$$

$$\sum_{(\bar{\boldsymbol{x}},\ \bar{\boldsymbol{u}},\ \bar{\boldsymbol{x}}^+)\in\mathcal{D}(\boldsymbol{x}(k))} w(\bar{\boldsymbol{x}}) = 1 \tag{13b}$$

with $\epsilon > 0$ a small regularization constant that ensures numerical stability.

The objective is to show that, under mild regularity assumptions on the dynamics, the one-step state resulting from the interpolated input $\boldsymbol{u}^*(k)$ remains close to the verified next-step state vector $\bar{\boldsymbol{x}}^+$. This deviation is bounded using

Question

What online computation or interaction remains in each of the three systems after its efficiency-improving mechanism is applied?

Gold Answer

Distributed RL retains neighbor communication and a local coordination optimization. KANELE retains LUT lookups and pipelined additions. MPC-guided safe RL retains nearest-neighbor lookup, weighted interpolation, a Lipschitz deviation check, and fallback retraining when needed.

Case Metadata

Dataset:	Multi-Document QA
Domain:	Electrical Engineering / Signal Processing
Modalities:	Text + Figure + Table + equation
Question Type:	Inferential / System Optimization & Engineering Implementation
Gold Pages:	6, 7, 14, 20, 38, 45, 46

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Qwen3-VL 4B	**Incorrect**	[1-66] (all)	It conflates the KANELE and safe-RL mechanisms and does not distinguish all residual online operations.
Qwen3-VL 8B	**Incorrect**	[1-66] (all)	It attributes the systems incorrectly and omits KANELE's LUT-only inference distinction.
GLM-4.6V	**Incorrect**	[2, 20, 38]	It finds parts of all three documents but incorrectly describes KANELE as online learning.
DeepSeek V4-Flash	**Correct**	[6, 7, 19, 38, 48]	It distinguishes coordination, LUT evaluation, and interpolation plus safety checking, with fallback retraining.

Figure 18: Case 3: reading-scope coverage versus correct three-document synthesis.

Both Qwen variants cite all 66 pages yet still mischaracterize the residual computation in the three systems, showing that complete reading scope does not guarantee a correct Multi-Document synthesis. GLM-4.6V selects a compact three-page set but incorrectly describes KANELE as retaining online learning. DeepSeek-V4-Flash-Vision gives the correct qualitative distinction, while its five predicted pages still differ from the seven-page gold evidence chain.

D.1.4 CASE 4: ANSWER SEMANTICS AND EVIDENCE COVERAGE CAN DIVERGE IN REASONING QA

Case 4: Document-ID 703 QA-ID rea_199

Document Context

Page.12

3. ONSAGER'S IDEAL TURBULENCE

"My tentative limiting formula for the correlation function in isotropic turbulence is not so obvious that any one student could be expected to find it. However, it seemed very probable to me that somebody would have investigated the line of reasoning, which is not far fetched."
L. Onsager in letter to C.C. Lin [99]

The result of the last section is that, under an observed uniform regularity assumption, the high Reynolds number limit of a Navier-Stokes solution is a locally dissipative weak solution of the Euler equation. Anomalous dissipation is the phenomenon that this local dissipation measure is non-trivial. Moreover, we see that it can be computed in two ways. First, it can be computed as an inviscid limit of the *viscous* local dissipation measure (2.18). But, evidently, it can also be computed on the limiting weak solution itself, see (2.21). As such, we now seek another way to compute the dissipation on the inviscid weak solution u which illuminates its ultimate cause. Recall from our earlier considerations around (2.16) that we expect having exactly $1/3$ of a derivative in $L^3_{t,x}$ to play a special role. The person to envisage irregular weak Euler solutions as describing turbulent dynamics, and to realize this precise regularity threshold was the Nobel Laureate Lars Onsager, see left panel of Figure 5. He wrote a now famous statement to this effect in his 1949 paper on Statistical Hydrodynamics [161]. There he stated that, in order for anomalous dissipation to take place, the fluid cannot satisfy a Hölderian condition of the form

$$|u(x+r)-u(x)| \lesssim |r|^h, \quad \text{for} \quad h > 1/3 \tag{3.1}$$

everywhere in the domain. In fact, Onsager knew much more than what he wrote, having effectively proved a strengthening of this statement in his unpublished notes, see right panel of Figure 5. See [99] for an excellent discussion of this history, before which this fact was apparently unknown. The first rigorous result in the literature towards Onsager's statement was Eyink's work [87], followed by the work of Constantin, E and Titi [42]. Duchon and Robert [63] framed the results in the context of the inviscid limit, as we have done here, and proved a stronger local statement. Their approach is the most similar to Onsager's own unpublished work on the matter.

Following Onsager, we first obtain a point-split energy balance (working on $M = \mathbb{T}^d$):

Proposition 3.1. *Let $u^\nu \in L^3_{t,x}$ be a weak solution to (2.3)–(2.4) for $\nu \geqslant 0$. If $\nu > 0$ we assume the solution is smooth. Denote by $\overline{u}^\nu_\ell = u^\nu * G_\ell$ the space mollification of u^ν*

Page.15

u^ν then $\lim_{\nu\to 0} \varepsilon^\nu[u^\nu] = \varepsilon[u]$ holds in the sense of distributions. As such, Corollary 3.4 identifies the non-linear Euler energy flux with the viscous dissipation:

$$\lim_{\nu\to 0} \varepsilon^\nu[u^\nu] = \lim_{\ell\to 0} \varepsilon_\ell[u]. \tag{3.9}$$

Proof of Corollary 3.4. Since translations are strongly continuous in L^p, it is easy to see

$$(J_\ell[u],\psi)_{L^2_{t,x}} \to (\tfrac{1}{2}|u|^2, \partial_\ell\psi)_{L^2_{t,x}} + (\tfrac{1}{2}|u|^2+p)u;\nabla\psi)_{L^2_{t,x}},$$
$$\big(\tfrac{1}{2}(u\cdot\overline{f}_\ell + \overline{u}_\ell\cdot f),\psi\big)_{L^2_{t,x}} \to (u\cdot f,\psi)_{L^2_{t,x}}.$$

It follows that the distributional limit of $\varepsilon_\ell[u]$ exists and must obey (3.7). □

Thus, we return to Onsager's assertion about energy conservation. The first published results are due to Eyink [87] and Constantin, E and Titi [42]. The final result is

Theorem 3.6 (Energy Conservation). *Fix $p \geqslant 3$ and $\sigma_p > 1/3$. Any weak Euler solution in $u \in L^p(0,T;B^{\sigma_p}_{p,\infty}(\mathbb{T}^d))$ or, equivalently,*

$$S^u_p(\ell) \lesssim \ell^{\zeta_p} \quad \text{with} \quad \zeta_p > \frac{p}{3}, \tag{3.10}$$

has a trivial energy defect $\varepsilon[u] = 0$. If also $u \in C(0,T;L^2(\mathbb{T}^d))$, energy is conserved.

Proof of Theorem 3.6. In view of Corollary 3.4, we bound $\varepsilon_\ell[u]$. Since $p \geqslant 3$ we have

$$\|\varepsilon_\ell[u]\|_{L^1_{t,\ell}} \leqslant \|\varepsilon_\ell[u]\|_{L^{p/3}_{t,\ell}} \leqslant \frac{C}{\ell}\int_{\mathbb{R}^d} |\nabla G(z)| \|\delta_{\ell z}u\|^3_{L^p_{t,x}}\,\mathrm{d}z \lesssim \ell^{3\sigma_p-1}\|u\|^3_{L^p B^{\sigma_p}_{p,\infty}}, \tag{3.11}$$

which vanishes if $\sigma_p > \frac{1}{3}$. Thus $\varepsilon[u] = 0$ is trivial for solutions $u \in L^p(0,T;B^{1/3+}_{p,\infty}(\mathbb{T}^d))$. □

Theorem 3.6 proves a strengthened version of Onsager's original 1949 assertion that the flow cannot possess more that $1/3$ of a derivative and be consistent with anomalous dissipation. The "negative side" of this conjecture was finally proved by Isett [122] where he exhibited a weak solution of Euler with slightly less than $1/3$ of a derivative that did not conserve energy (such solutions can, in fact, be dissipative [27]). This followed a remarkable series of works on the subject of constructing wild weak solutions via a technique called convex integration [171, 175, 51, 52]. See the comprehensive review by Buckmaster and Vicol [26]. The union of these two sides, positive and negative, is termed Onsager's conjecture. However, an important question about the excluded middle remains open (and therefore, so does "Onsager's conjecture"):

Question 3.7. *Suppose $u \in L^\infty(0,T;L^2(\mathbb{T}^d)) \cap L^p(0,T;B^{1/3}_{p,\infty}(\mathbb{T}^d))$ for any $p \geqslant 3$ is a weak solution of Euler. Does it conserve energy?*

Question

What is the specific regularity threshold required by the theoretical framework that posits energy conservation for weak Euler solutions, and what is the numerical value of this exponent?

Gold Answer

Onsager's conjecture identifies a critical regularity threshold of 1/3 (approximately 0.333) for energy conservation in weak Euler solutions.

Case Metadata

Dataset:	Reasoning QA
Domain:	Physics / Classical Physics
Modalities:	Text
Question Type:	Inferential / Physical Concept & Theoretical Model
Gold Pages:	12, 15

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Qwen3-VL 4B	**Incorrect**	[15-60]	The threshold is sigma_p > 1/3, and the numerical value of the exponent is approximately 0.72.
Qwen3-VL 8B	**Correct**	[15, 20]	The required threshold is sigma_p > 1/3, and the numerical value is 1/3.
GLM-4.6V	**Correct**	[12, 15]	The framework requires sigma_p > 1/3, where the critical exponent is 1/3.
DeepSeek V4-Flash	**Correct**	[12]	Onsager's conjecture requires Holder regularity h > 1/3; the exponent is 1/3.

Figure 19: Case 4: divergence between answer semantics and evidence coverage.

This example is especially useful for separating semantic answer judging from evidence localization. The values $1/3$ and 0.333 express the same threshold up to representation, whereas adding an incompatible value such as 0.72 changes the factual content. At the same time, an answer can be semantically correct while only partially covering the required evidence pages. This is precisely the distinction captured by reporting answer accuracy and E-Precision/E-Recall/E-F1 separately.

D.1.5 CASE 5: EVIDENCE OVER-PREDICTION IN A MULTI-PAGE SCIENTIFIC REASONING QUESTION

Case 5: Document-ID 346 QA-ID rea_123

Document Context

Page.2

curvature to thermodynamic quantities such as entropic contributions from the vibrational density of states and to spectroscopic observables, since infrared (IR) and Raman intensities depend on mode-resolved dipole and polarizability derivatives [12]. The validity of the harmonic approximation must be assessed to determine whether macroscopic observables, such as transition rates and temperature-dependent conformational changes, can be described accurately.

One critical limitation of conventional MD approaches based on classical force fields is their neglect of quantum-mechanical effects. Quantum phenomena such as polarization, charge transfer, and anharmonic couplings are not only essential for accurate energy landscapes but also manifest directly in vibrational spectra, where subtle shifts and intensity variations encode information about electronic structure and environmental response. Consequently, molecular mechanics force fields (MMFFs) can only be rigorously compared to experiment in regimes where such quantum contributions are negligible, a condition rarely satisfied in biologically relevant systems.

To understand biomolecular stability and vibrations from first principles, we compare widely used MMFFs with modern machine-learned force fields (MLFFs) trained exclusively on quantum-mechanical forces and energies. While MMFFs typically treat two-body electrostatics at all spatial distances using Ewald-like summations, they usually cover non-Coulomb two-body interactions only up to a distance of 8-10 Å and many-body interactions only for the nearest two or three chemical bonding partners. In contrast, modern MLFFs based on equivariant graph neural networks

Page.11

comparing SO3LR and MMFFs, grouped by bond type. In scatter plots, each point corresponds to the most similar mode matched by the atom index. Source data are provided as a Source Data File.

to reproduce the nuanced packing effects observed in the DFT-level calculations.

In aqueous solution, solvent-dependent shifts arise from a balance of hydrogen-bonding, local electrostatic fields, steric hindrance, and mode coupling. Hydrogen bonding typically elongates and softens O–H and N–H bonds, producing red shifts in their stretching frequencies [69]. Blue shifts can occur in scaffolds exhibiting improper H bonds or when strong local electric fields stiffen the X–H bond—effects showing that electrostatic charge distribution and hydrogen-bond geometry can either red- or blue-shift stretching modes depending on the orientation and nature of the interaction [70, 71].

A comparative analysis of solvation effects in tyrosine, arginine, and leucine, with PBE0+MBD as the reference, shows that SO3LR reliably reproduces solvent-induced vibrational shifts, whereas MMFFs fail to capture these effects (See Supp. Appendix B). For tyrosine, the O–H stretch undergoes a pronounced red shift in both DFT and SO3LR, due to hydrogen bonding of water with the tyrosine hydroxy group, while MMFFs predict minimal change. The N–H groups of arginine exhibit both red and

11

Page.12

ness, while the fixed functional forms of the MMFFs produce nearly identical profiles. Both SO3LR and AMOEBA reproduce an anharmonic bond-stretch profile (Fig. 5d), but the origin of this anharmonicity differs: While SO3LR learns it directly from the QM training data, AMOEBA introduces an analytic anharmonic bond potential [29]. This explains the trend in Fig.5e, where the normalized area between the fully harmonic and computed potentials is compared: the flexible, environment-responsive SO3LR potential captures DFT-like behavior; AMOEBA's built-in anharmonicity improves its shifts relative to non-polarizable MMFFs, though its fixed form still limits adaptability.

MMFFs systematically fail to reproduce these behaviours because they lack explicit polarization, many-body induction, anharmonicity, and environment-sensitive hydrogen-bond descriptions. Their harmonic bond potentials and fixed charges cannot capture the solvent-dependent, cooperative stabilization patterns observed in strongly hydrogen-bonded systems. Since bond-stretch anharmonicity is a primary driver of frequency shifts, harmonic force fields predict little change upon stretching; mode coupling can shift frequencies, but this mechanism alone yields incorrect behavior. Accurately reproducing these shifts requires a quantum-mechanical treatment of the dynamical behavior of electrons and anharmonicity. An important exception among MMFFs is AMOEBA, which incorporates explicit polarization, multipole electrostatics, and anharmonic bond terms. These features allow AMOEBA

12

Question

How does including polarization and many-body effects influence the performance of advanced frameworks relative to classical force fields when describing molecular vibrational environments?

Gold Answer

Advanced frameworks such as SO3LR describe solvent-induced vibrational shifts more accurately because they capture polarization and many-body effects that classical molecular-mechanics force fields omit.

Case Metadata

Dataset:	Reasoning QA
Domain:	Physics / Atomic and Molecular Clusters
Modalities:	Text
Question Type:	Inferential / Physical Concept & Theoretical Model
Gold Pages:	2, 11, 12

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Qwen3-VL 4B	**Correct**	[1-42] (all)	SO3LR captures polarization, anharmonicity, and environment-driven interactions more accurately than classical fields.
Qwen3-VL 8B	**Correct**	[3, 6, 12, 13]	Including polarization and many-body effects yields more accurate and transferable vibrational predictions.
GLM-4.6V	**Correct**	[1, 2]	SO3LR captures quantum effects omitted by classical force fields and reaches DFT-level accuracy.
DeepSeek V4-Flash	**Correct**	[1, 6, 9, 11, 12, 13]	Environment-aware many-body interactions reproduce solvent shifts and protein interactions that MMFFs miss ...

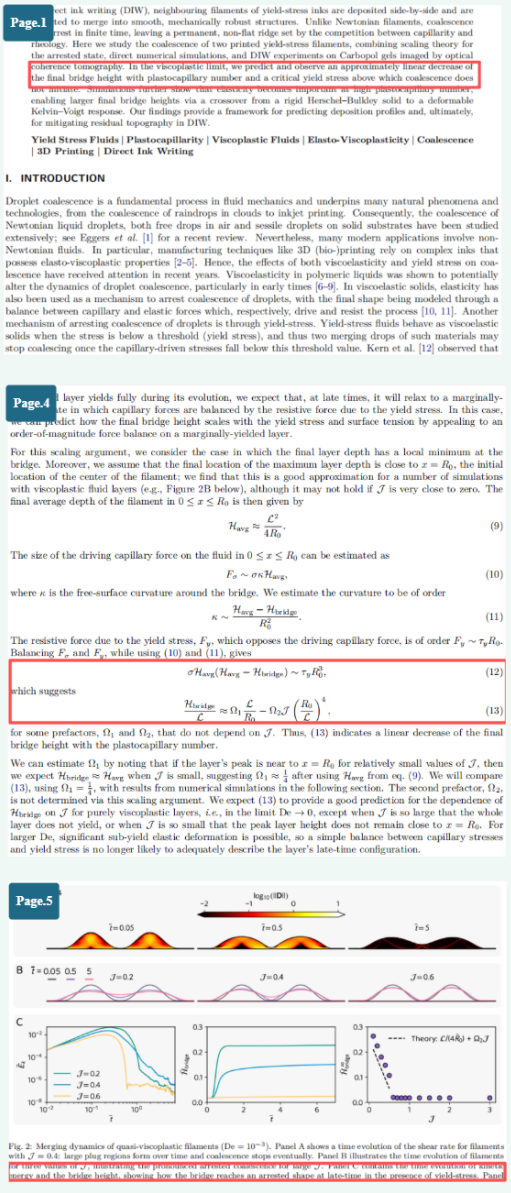
Figure 20: Case 5: evidence over-prediction in multi-page scientific reasoning.

All four answers correctly connect the improved vibrational description to polarization and many-body effects, but none predicts the exact three-page evidence set. Qwen3-VL-4B cites all 42 pages, while the compact GLM-4.6V output includes page 2 but omits the result and causal discussion on pages 11–12. The remaining predictions cover part of that chain with additional neighboring pages, again separating semantic accuracy from grounding quality.

D.1.6 CASE 6: EVIDENCE SELECTIVITY IN SINGLE-DOCUMENT MULTIMODAL QA

Case 6: Document-ID 362 QA-ID gen_6

Document Context



Question

How does the final bridge height change with the plastocapillary number mathcal{J} in the viscoplastic limit?

Gold Answer

Approximately linear decrease

Case Metadata

Dataset:	General QA
Domain:	Physics / Soft Condensed Matter
Modalities:	Text + Image + Equation
Question Type:	Inferential / Computation & Experimental Validation
Gold Pages:	1, 4, 5

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Qwen3-VL-8B	Correct	[4, 5]	The final bridge height decreases linearly with the plastocapillary number \(\mathcal{J}\) in the viscoplastic limit.
InternVL2.5-8B	Correct	[4, 5]	In the viscoplastic limit, the final bridge height decreases linearly with the plastocapillary number mathcal{J}.
GLM-4.6V	Correct	[1, 4, 5, 6, 9, 10]	In the viscoplastic limit, the final bridge height decreases approximately linearly with the plastocapillary number J...
MiniCPM-V-2.6	Incorrect	[]	The final bridge height increases with the increase of the plastocapillary number mathcal{J} in the viscoplastic limit.

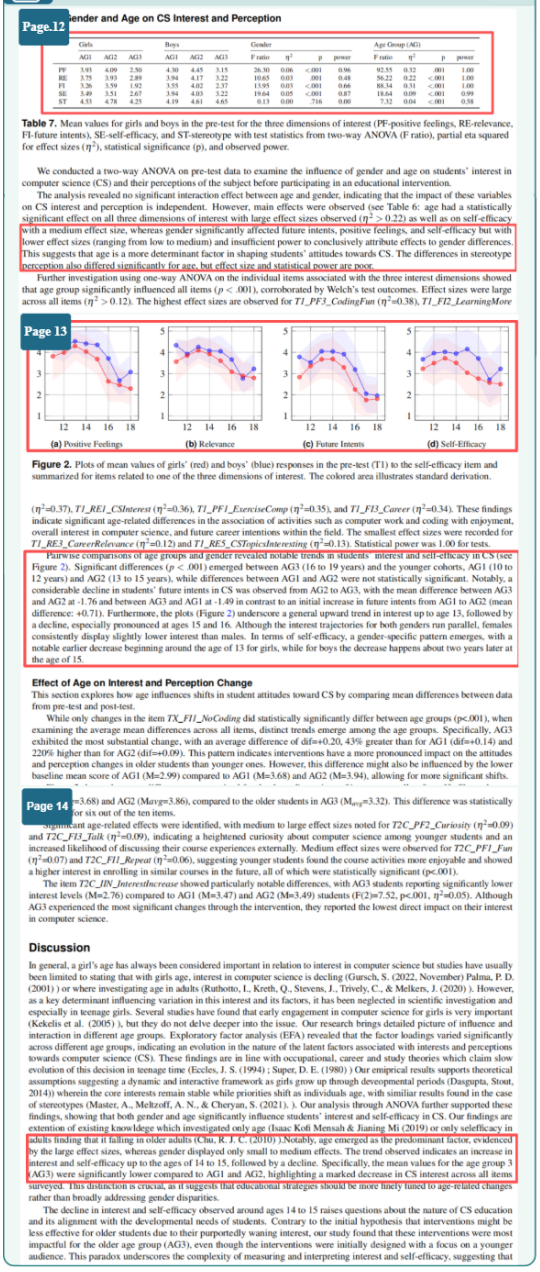
Figure 21: Case 6: evidence selectivity in single-document multimodal QA.

For this single-document General QA example, the models produce broadly similar answers, but none recovers the complete gold evidence set. Qwen3-VL-8B and InternVL2.5-8B both answer correctly and select pages 4 and 5, missing the supporting evidence on page 1. GLM-4.6V also answers correctly, but returns a much broader set of pages (1, 4, 5, 6, 9, and 10), which covers the gold evidence while introducing several redundant pages. MiniCPM-V 2.6, by contrast, provides no evidence pages and gives an incorrect answer. The example therefore separates answer correctness from evidence completeness: correct answers do not necessarily come with complete or selective evidence localization.

D.1.7 CASE 7: CORRECT EVIDENCE DOES NOT GUARANTEE CORRECT INTERPRETATION

Case 7: Document-ID 2 QA-ID gen_7

Document Context Three evidence snippets from the source document



Question

How do the CS interest trajectories of girls and boys change with age?

Gold Answer

rise up to early adolescence then decline

Case Metadata

Dataset:	General QA
Domain:	Economics / Theoretical Economics
Modalities:	Text + Image + Table
Question Type:	Inferential / Causal Inference & Result Interpretation
Gold Pages:	12, 13, 14

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
DeepSeek-V4-Flash	Correct	[12, 13, 14]	CS interest trajectories change with age: girls' positive feelings and relevance decline noticeably after age 13, while boys' decline begins about two years later (around age 15)...
Qwen3-VL-4B	Incorrect	[12, 13, 14]	Girls show a marked decline in enthusiasm during early adolescence, while boys show a more gradual decline, with a notable decrease around age 14 to 15 for both ...
MiniCPM-V-4.5	Correct	[6]	Interest trajectories often rise in late childhood and decline during mid-adolescence for both girls and boys.
Mistral-Small3.1-24B	Correct	[1, 2, 3, 12, 13, 14]	The CS interest trajectories of girls and boys change with age, showing a marked decline in enthusiasm during early adolescence, particularly among girls. Age is a more decisive...

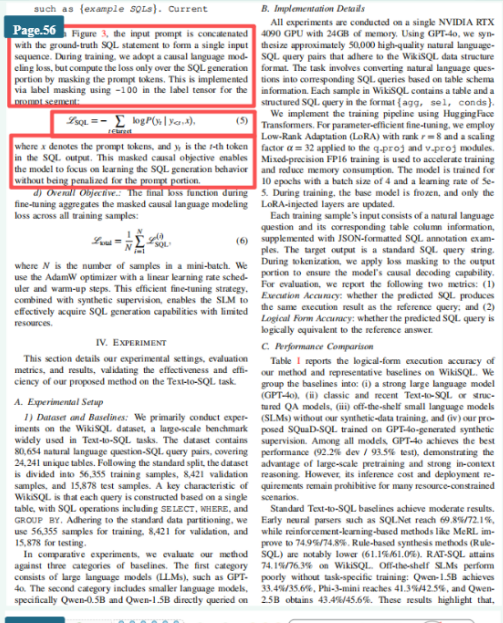
Figure 22: Case 7: correct evidence does not guarantee correct interpretation in single-document QA.

DeepSeek-V4-Flash-Vision, Qwen3-VL-4B, and Mistral-Small-3.1-24B all recover the relevant evidence pages well, with some predictions matching the annotated evidence exactly. However, Qwen3-VL-4B still produces an incorrect answer despite locating the correct evidence. In contrast, MiniCPM-V 4.5 gives the correct answer even though its predicted evidence pages do not match the gold annotations. This case shows that accurate evidence localization does not by itself guarantee correct interpretation, and that answer correctness and evidence grounding should be evaluated separately.

D.1.8 CASE 8: MULTI-DOCUMENT REASONING WITH ANSWER–EVIDENCE INCONSISTENCY

Case 8: Document-ID 0273 QA-ID md_2

Document Context



Question

Compare the training loss formulations of SQuaD-SQL (Doc 2) and COALA (Doc 3) in terms of how they optimize target outputs.

Gold Answer

In SQuaD-SQL (Doc 2), training optimizes SQL generation using a masked causal language modeling loss (L_{SQL}), which computes cross-entropy loss only over the target SQL tokens while ignoring prompt tokens. In COALA (Doc 3), training optimizes the discriminative scoring of biasing entities...

Case Metadata

Dataset:	Multi-Document QA
Domain:	Computer Science / Artificial Intelligence
Modalities:	Equation + Table
Question Type:	Inferential / Experiment & Result Validation
Gold Pages:	56, 60

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
GLM-4.6V	Correct	[56,60]	SQuaD-SQL uses a masked causal language modeling loss (L_SQL) that focuses on SQL generation by masking prompt tokens, computed only over SQL tokens...
Qwen3-VL-8B	Correct	[55,61]	SQuaD-SQL uses a masked causal language modeling loss, where the loss is computed only over the SQL generation...
Claude Sonnet 5	Correct	[54,55,60,61]	SQuaD-SQL uses masked causal language modeling loss for training, where the input prompt is concatenated...
Qwen3-VL-4B	Incorrect	[55,60]	SQuaD-SQL optimizes SQL query generation using a masked causal language modeling loss, while COALA uses a combination of discriminative losses ....

Figure 23: Case 8: Multi-Document reasoning with answer–evidence inconsistency.

For this Multi-Document QA example, the gold evidence is on pages 56 and 60. GLM-4.6V identifies both pages exactly, whereas the other three models select neighboring pages. Among them, only Qwen3-VL-4B gives an incorrect answer. The remaining models therefore reach the correct conclusion despite small page-level localization errors, likely because the adjacent pages contain closely related context. This example also shows that correct answering can tolerate limited evidence-page drift, even when the predicted pages do not exactly match the gold annotations.

D.1.9 CASE 9: MULTI-PAGE REASONING AND RETRIEVAL BREADTH

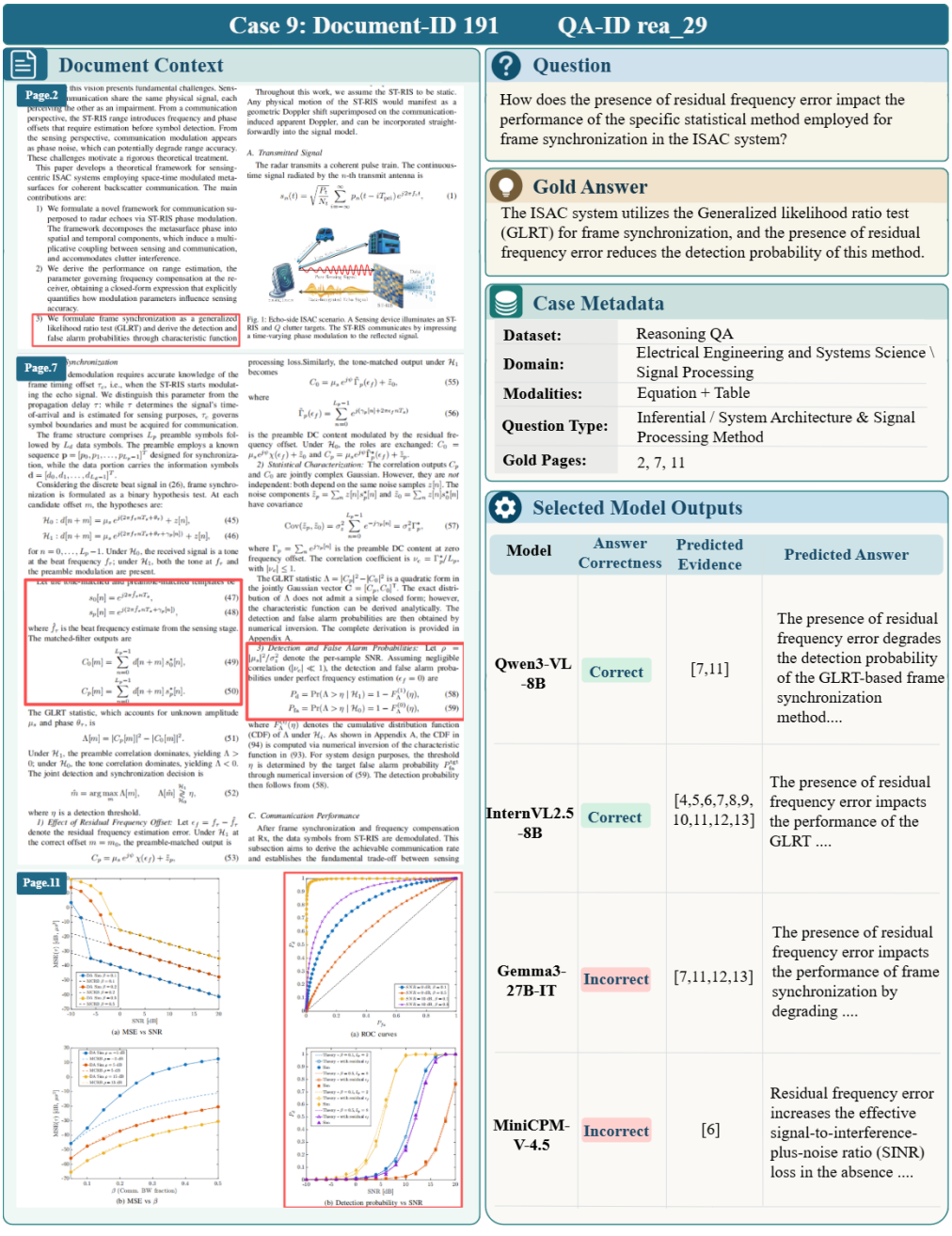


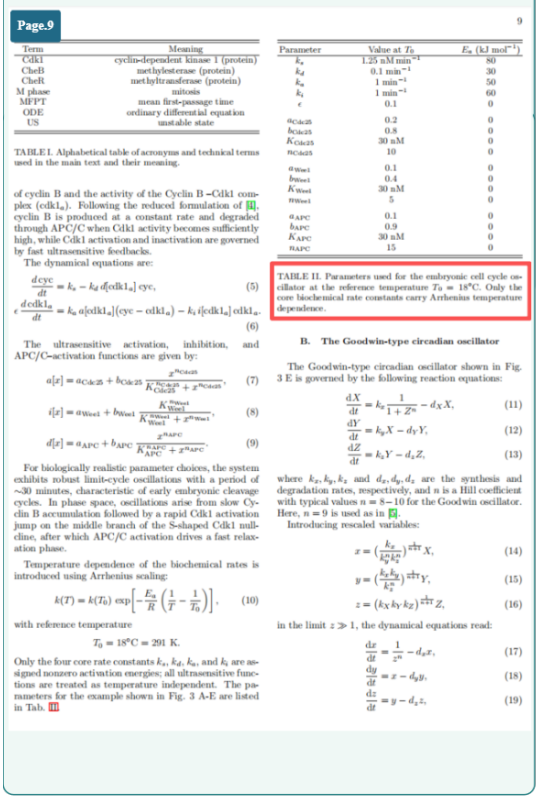
Figure 24: Case 9: multi-page reasoning and retrieval breadth in reasoning QA.

Answering this question requires linking the GLRT-based frame-synchronization formulation with the effect of residual frequency error on detection performance across pages 2, 7, and 11. Qwen3-VL-8B answers correctly from a compact prediction of pages 7 and 11, whereas InternVL2.5-8B also reaches the correct conclusion but retrieves a much broader ten-page range. Gemma 3 27B retrieves several highly relevant pages, including pages 7 and 11, yet fails to derive the required effect on the GLRT-based synchronization method. MiniCPM-V 4.5 retrieves only page 6 and also answers incorrectly. This case highlights that broader retrieval does not necessarily improve reasoning, while access to relevant evidence does not by itself ensure successful multi-page inference.

D.1.10 CASE 10: UNSUPPORTED ANSWERING UNDER A HARD DISTRACTOR

Case 10: Document-ID 90 QA-ID una_90

Document Context


Page.9

9

Term	Meaning
Cdk1	cyclin-dependent kinase 1 (protein)
CheB	methylesterase (protein)
CheR	methyltransferase (protein)
M phase	mitosis
MFPT	mean first-passage time
ODE	ordinary differential equation
US	unstable state

TABLE I. Alphabetical table of acronyms and technical terms used in the main text and their meaning.

of cyclin B and the activity of the Cyclin B–Cdk1 complex ($\mathrm{cdk1_a}$). Following the reduced formulation of [4], cyclin B is produced at a constant rate and degraded through APC/C when Cdk1 activity becomes sufficiently high, while Cdk1 activation and inactivation are governed by fast ultrasensitive feedbacks.

The dynamical equations are:

$$\frac{d\,\mathrm{cyc}}{dt} = k_s - k_d\, d[\mathrm{cdk1_a}]\,\mathrm{cyc}, \tag{5}$$

$$\epsilon\,\frac{d\,\mathrm{cdk1_a}}{dt} = k_a\, a[\mathrm{cdk1_a}](\mathrm{cyc} - \mathrm{cdk1_a}) - k_i\, i[\mathrm{cdk1_a}]\,\mathrm{cdk1_a}. \tag{6}$$

The ultrasensitive activation, inhibition, and APC/C-activation functions are given by:

$$a[x] = a_{\mathrm{Cdc25}} + b_{\mathrm{Cdc25}}\frac{x^{n_{\mathrm{Cdc25}}}}{K_{\mathrm{Cdc25}}^{n_{\mathrm{Cdc25}}} + x^{n_{\mathrm{Cdc25}}}}, \tag{7}$$

$$i[x] = a_{\mathrm{Wee1}} + b_{\mathrm{Wee1}}\frac{K_{\mathrm{Wee1}}^{n_{\mathrm{Wee1}}}}{K_{\mathrm{Wee1}}^{n_{\mathrm{Wee1}}} + x^{n_{\mathrm{Wee1}}}}, \tag{8}$$

$$d[x] = a_{\mathrm{APC}} + b_{\mathrm{APC}}\frac{x^{n_{\mathrm{APC}}}}{K_{\mathrm{APC}}^{n_{\mathrm{APC}}} + x^{n_{\mathrm{APC}}}}. \tag{9}$$

For biologically realistic parameter choices, the system exhibits robust limit-cycle oscillations with a period of ∼30 minutes, characteristic of early embryonic cleavage cycles. In phase space, oscillations arise from slow Cyclin B accumulation followed by a rapid Cdk1 activation jump on the middle branch of the S-shaped Cdk1 nullcline, after which APC/C activation drives a fast relaxation phase.

Temperature dependence of the biochemical rates is introduced using Arrhenius scaling:

$$k(T) = k(T_0)\,\exp\left[-\frac{E_a}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right], \tag{10}$$

with reference temperature

$$T_0 = 18°\mathrm{C} = 291\ \mathrm{K}.$$

Only the four core rate constants k_s, k_d, k_a, and k_i are assigned nonzero activation energies; all ultrasensitive functions are treated as temperature independent. The parameters for the example shown in Fig. 3 A-E are listed in Tab. II.

Parameter	Value at T_0	E_a (kJ mol^{-1})
k_s	1.25 nM min^{-1}	80
k_d	0.1 min^{-1}	30
k_a	1 min^{-1}	50
k_i	1 min^{-1}	60
ϵ	0.1	0
a_{Cdc25}	0.2	0
b_{Cdc25}	0.8	0
K_{Cdc25}	30 nM	0
n_{Cdc25}	10	0
a_{Wee1}	0.1	0
b_{Wee1}	0.4	0
K_{Wee1}	30 nM	0
n_{Wee1}	5	0
a_{APC}	0.1	0
b_{APC}	0.9	0
K_{APC}	30 nM	0
n_{APC}	15	0

TABLE II. Parameters used for the embryonic cell cycle oscillator at the reference temperature $T_0 = 18°\mathrm{C}$. Only the core biochemical rate constants carry Arrhenius temperature dependence.

B. The Goodwin-type circadian oscillator

The Goodwin-type circadian oscillator shown in Fig. 3 E is governed by the following reaction equations:

$$\frac{dX}{dt} = k_x\frac{1}{1+Z^n} - d_X X, \tag{11}$$

$$\frac{dY}{dt} = k_y X - d_Y Y, \tag{12}$$

$$\frac{dZ}{dt} = k_z Y - d_z Z, \tag{13}$$

where k_x, k_y, k_z and d_x, d_y, d_z are the synthesis and degradation rates, respectively, and n is a Hill coefficient with typical values $n \approx 8-10$ for the Goodwin oscillator. Here, $n = 9$ is used as in [5].

Introducing rescaled variables:

$$x = \left(\frac{k_x}{k_y k_z}\right)^{\frac{1}{n+1}} X, \tag{14}$$

$$y = \left(\frac{k_x k_y}{k_z^n}\right)^{\frac{1}{n+1}} Y, \tag{15}$$

$$z = \left(k_X k_Y k_Z\right)^{\frac{1}{n+1}} Z, \tag{16}$$

in the limit $z \gg 1$, the dynamical equations read:

$$\frac{dx}{dt} = \frac{1}{z^n} - d_x x, \tag{17}$$

$$\frac{dy}{dt} = x - d_y y, \tag{18}$$

$$\frac{dz}{dt} = y - d_z z, \tag{19}$$

Question

What is the reference temperature T1 for the embryonic cell cycle oscillator model?

Gold Answer

Unanswerable

Case Metadata

Dataset:	Unanswerable QA
Domain:	Quantitative Biology / Molecular Networks
Modalities:	Text + Equation + Table
Question Type:	Literal / Biological Network & Dynamic Modeling
Gold Pages:	None

Selected Model Outputs

Model	Answer Correctness	Predicted Evidence	Predicted Answer
Claude Sonnet 5	Correct	[]	Unanswerable
Qwen3-VL-4B	Correct	[]	Unanswerable
DeepSeek-V4-Flash	Incorrect	[9]	T0 = 18°C = 291 K
GLM-4.6V	Incorrect	[9]	18°C (291 K)

Figure 25: Case 10: unsupported answering under a hard distractor in unanswerable QA.

For this Unanswerable QA example, the question asks for a reference temperature T_1, but no such parameter is defined anywhere in the document. The only closely related quantity is $T_0 = 18°\mathrm{C} = 291\,\mathrm{K}$. Claude Sonnet 5 and Qwen3-VL-4B correctly recognize that the requested information is unavailable and therefore abstain from answering. DeepSeek-V4-Flash-Vision and GLM-4.6V, however, cite page 9 and return the value of T_0 as if it were the requested T_1. In this case, the models locate a lexically and semantically similar region but fail to verify that the retrieved parameter actually matches the question. This produces a plausible yet unsupported answer, illustrating a concrete form of hallucination under insufficient evidence.

D.2 ERROR TAXONOMY AND ANALYSIS METHOD

Error annotation assigns one dominant mechanism and optional co-occurring labels using Table 12. Reviewers check answer semantics, coverage of necessary facts, excess pages, and formatting or resource status. Consecutive-span expansion uses the longest interval in the predicted page set; page-index drift is verified against the source-to-merged map, and missing chains against uncovered intermediate facts.

Table 12: Error categories used in qualitative and quantitative analysis.

Error type	Decision rule	Typical indicator	Interpretation
Whole-document citation	Predicted pages approach the total PDF page count	High Recall, low Precision	Treats reading scope as evidence
Consecutive-span expansion	Many adjacent pages, most without direct support	Longest consecutive interval	Fails to compress context
Missing key page	At least one necessary hop has no predicted page	Lower Recall	Incomplete reasoning chain
Page-index drift	Cites source-page indices rather than merged pages	Lower evidence precision/recall	Confuses document coordinate systems
Answer–page mismatch	Answer is correct but evidence is wrong	Answer–Evidence gap	Semantic summarization stronger than attribution
Illegal output	JSON/field/refusal contract violation	Illegal-output rate	Instruction-following failure
Resource failure	OOM or PDF rendering failure	`generation_status`	Engineering-resource limitation

E BASELINE MODELS

We evaluate eleven representative MLLMs, including six small-scale models and five large-scale models. Comprehensive multimodal benchmarks such as MME and MMBench have demonstrated the importance of evaluating MLLMs across diverse perception, understanding, and reasoning capabilities Fu et al. (2026); Liu et al. (2024). Brief descriptions of the evaluated baselines are provided below.

Qwen3-VL-4B-Instruct. Qwen3-VL-4B-Instruct is a 4B-parameter vision-language model from the Qwen3-VL family, designed for multimodal instruction following and visual document understanding.

Qwen3-VL-8B-Instruct. Qwen3-VL-8B-Instruct is the larger 8B variant of Qwen3-VL, providing stronger multimodal reasoning and document understanding capabilities.

InternVL3.5-8B. InternVL3.5-8B is an 8B-scale multimodal model from the InternVL family, developed for general-purpose image understanding, visual reasoning, and multi-image tasks.

InternVL2.5-8B. InternVL2.5-8B is an 8B-scale multimodal model from the InternVL2.5 family, developed for high-resolution image understanding, document-oriented visual comprehension, and multi-image reasoning tasks.

MiniCPM-V 4.5. MiniCPM-V 4.5 is a compact multimodal model optimized for efficient visual-language understanding, with particular emphasis on high-resolution images and document-oriented tasks.

MiniCPM-V 2.6. MiniCPM-V 2.6 is an earlier lightweight model in the MiniCPM-V series that supports multi-image visual understanding and serves as a compact baseline for document question answering.

Mistral-Small-3.1-24B-Instruct-2503. Mistral-Small-3.1-24B-Instruct-2503 is a 24B-parameter instruction-tuned multimodal model that combines general language reasoning with visual understanding capabilities.

Gemma 3 27B IT. Gemma 3 27B IT is a 27B-parameter instruction-tuned multimodal model from the Gemma 3 family, supporting image-conditioned language understanding and reasoning.

GLM-4.6V. GLM-4.6V is a general-purpose vision-language model from the GLM family, designed to handle multimodal understanding and reasoning over visual and textual inputs.

Claude Sonnet 5. Claude Sonnet 5 is a proprietary multimodal model from Anthropic that provides strong general-purpose language, visual understanding, and long-context reasoning capabilities.

DeepSeek-V4-Flash-Vision-Exp. DeepSeek-V4-Flash-Vision-Exp is a proprietary multimodal model from DeepSeek, designed for efficient visual-language understanding and reasoning.

F PROMPT TEMPLATES

To ensure a fair comparison across models, we adopt a task-specific prompt template for each evaluation setting. The templates follow a unified overall structure across task settings and differ only in the instructions required by the corresponding task type. Specifically, `answer_pre` stores the final answer, while `evidence_pages` stores a list of PDF page indices. For all tasks, models are required to jointly produce a concise final answer and the PDF pages that materially support it.

The three templates are designed to capture the distinct requirements of the corresponding QA settings. The complete prompt templates are illustrated in Figs. 26–28.

GEN_QA_AND_UNANS_QA_PROMPT

You are answering a scientific PDF question using one supplied complete scientific paper.

This benchmark contains both answerable and deliberately unanswerable questions. Decide answerability only from the supplied PDF pages.

Some unsupported questions may contain a subtle alteration to a year, number, model/version, entity, threshold, experimental condition, or another detail. Do not silently correct the question to a nearby fact in the paper.

Return exactly one valid JSON object:
{"answer_pre":"<complete but concise final answer>","evidence_pages":[<PDF page numbers>]}

IMPORTANT OUTPUT FORMAT:
- evidence_pages must be a JSON array of integer page numbers only.
- Correct example:
  {"answer_pre":"...","evidence_pages":[2,11,18]}
- Never output PDF_PAGE_2, "PDF_PAGE_2", Page 2, "Page 2", 2/20, or any other page-label text inside evidence_pages.
- PDF_PAGE_N_START and PDF_PAGE_N_END are input boundary labels only. In the output JSON, represent that page as the integer N.
- Do not output any text before or after the JSON object.

Rules:
1. Use only information contained in the supplied PDF page images.
2. Inspect the question exactly as written.
3. If the PDF supports the question, answer it concisely and accurately.
4. If the PDF does not support the question as written, return exactly:
   {"answer_pre":"Unanswerable","evidence_pages":[]}
5. Do not guess, repair, reinterpret, or substitute a nearby supported fact for an unsupported question.
6. For answerable questions, evidence_pages must contain the external physical PDF pages that materially support the answer.
7. Use the external PDF_PAGE_N labels to identify physical pages; do not use page numbers printed inside the paper body, header, footer, figure, or table.
8. Evidence page numbers are 1-based physical PDF pages.
9. Include all materially necessary evidence pages, but do not include merely related pages.
10. Preserve requested units, signs, precision, formulas, terminology, conditions, model names, dataset names, and comparison direction.
11. Keep answer_pre complete but concise.
12. answer_pre must contain the final answer, not an outline, plan, or reasoning trace.
13. Do not output chain-of-thought, derivations, Markdown fences, citations, comments, <think> tags, or additional JSON keys.

Figure 26: Prompt template used for General and Unanswerable QA. The model is instructed to answer using the supplied scientific PDF and to return `Unanswerable` when the required evidence is not supported by the document.

MULTI-DOC_QA_PROMPT

You are answering a cross-document scientific reasoning question from one merged PDF.

The merged PDF contains multiple independent scientific papers concatenated in order. References in the question such as "Paper 1"/"Doc 1", "Paper 2"/"Doc 2", and "Paper 3"/"Doc 3" refer to the first, second, and third source papers in that merged PDF.

Identify paper boundaries yourself from the supplied PDF pages using titles, authors, front matter, section restarts, references, and content continuity. No external document-boundary map is provided.

Every question in this Multi-Doc benchmark is answerable from the supplied merged PDF. Do not abstain, refuse, or answer "unanswerable".

A complete answer may require synthesizing definitions, assumptions, methods, mechanisms, formulas, quantitative results, limitations, causal relations, experimental findings, or methodological conditions from two or three source papers.

Inspect every paper requested by the question. Do not stop after locating only one relevant document or one supporting fact.

Return exactly one valid JSON object:
{"answer_pre":"<complete but concise final answer>","evidence_pages":[<merged PDF page numbers>]}

IMPORTANT OUTPUT FORMAT:
- evidence_pages must be a JSON array of integer merged-page numbers only.
- Correct example:
  {"answer_pre":"...","evidence_pages":[5,31,52]}
- Never output PDF_PAGE_5, "PDF_PAGE_5", Page 5, Doc 2 Page 3, 5/80, or any other page-label text inside evidence_pages.
- PDF_PAGE_N_START and PDF_PAGE_N_END are input boundary labels only. In the output JSON, represent that merged physical page as the integer N.
- Do not output source-document-local page numbers unless they are also the corresponding merged page numbers.
- Do not output any text before or after the JSON object.

Rules:
1. Use only information contained in the supplied merged PDF page images.
2. Every question is answerable from the supplied pages. Never output "unanswerable".
3. Resolve Paper/Doc numbering by source-paper order in the merged PDF.
4. Answer every comparison, contrast, mechanism, condition, calculation, transfer, conflict, or conclusion requested.
5. evidence_pages must include the external merged physical pages that materially support all required parts of the answer.
6. Use the external PDF_PAGE_N labels to identify merged pages; do not use page numbers printed inside an individual source paper.
7. Evidence page numbers are pages of the complete merged PDF.
8. Include all materially necessary evidence pages, but do not include pages that are merely related or unnecessary.
9. Preserve technical terms, equation, signs, units, numerical values, theorem names, architecture names, method names, dataset names, and document-specific distinctions.
10. Keep answer_pre complete but concise.
11. answer_pre must contain the final answer, not a plan, hidden reasoning trace, or derivation transcript.
12. Do not output chain-of-thought, derivations, Markdown fences, citations, comments, <think> tags, or additional JSON keys.

Figure 27: Prompt template used for Multi-Document QA. The model is required to identify and integrate supporting evidence across multiple scientific papers concatenated into a single PDF.

REA_QA_PROMPT

You are answering a compositional reasoning question about one supplied scientific PDF.

Every question in this Rea benchmark has a verified answer supported by the supplied PDF. Do not abstain, refuse, or answer "unanswerable".

The answer may require combining multiple facts from non-adjacent PDF pages and then performing one or more reasoning operations, including conditional inference, causal chaining, comparison, multi-step calculation, or constraint intersection.

The benchmark is deliberately constructed so that one isolated premise may be insufficient. Resolve hidden bridge entities, conditions, quantities, mechanisms, or dependencies from the PDF before producing the final answer.

Inspect the complete supplied PDF. Do not stop after finding only one relevant premise. Internally identify every distinct fact needed for the answer and internally perform the required reasoning.

Return exactly one valid JSON object:
{"answer_pre":"<complete but concise final answer>","evidence_pages":[<PDF page numbers>]}

IMPORTANT OUTPUT FORMAT:
- evidence_pages must be a JSON array of integer page numbers only.
- Correct example:
  {"answer_pre":"...","evidence_pages":[4,17,29]}
- Never output PDF_PAGE_4, "PDF_PAGE_4", Page 4, "Page 4", 4/50, or any other page-label text inside evidence_pages.
- PDF_PAGE_N_START and PDF_PAGE_N_END are input boundary labels only. In the output JSON, represent that page as the integer N.
- Do not output any text before or after the JSON object.

Rules:
1. Use only information contained in the supplied PDF page images.
2. Every question is answerable from the supplied PDF. Never output "unanswerable".
3. Do not answer from the question wording alone when the required facts are in the PDF.
4. Combine all required premises before producing the final answer. Do not infer the terminal answer from a single page when the question requires a bridge or dependency.
5. evidence_pages must include the external PDF pages that materially support every required premise.
6. Include all materially necessary evidence pages, but do not include pages that are merely related or were not needed.
7. Use the external PDF_PAGE_N labels to identify pages; do not use page numbers printed inside the PDF page.
8. Evidence page numbers are PDF pages.
9. Preserve requested units, signs, precision, formulas, terminology, conditions, model names, dataset names, and comparison direction.
10. Keep answer_pre complete but concise.
11. answer_pre must contain the final answer, not a plan or reasoning trace.
12. Do not output chain-of-thought, derivations, Markdown fences, citations, comments, <think> tags, or additional JSON keys.

Figure 28: Prompt template used for Reasoning QA. The model is required to perform compositional reasoning over distributed evidence while returning only the final answer and supporting PDF pages.