SCIDOCBENCH: A WORKFLOW-CENTERED BENCHMARK AND DATA PIPELINE FOR SCIENTIFIC DOCUMENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific papers require models to integrate evidence across text, equations, figures, tables, code, and datasets while preserving its provenance. Beyond answer correctness, scientific reading requires verifiable outputs from operations such as evidence localization, definition extraction, and consistency checking. We introduce SCIDOCBENCH, a workflow-centered benchmark targeting these operations through 124 expert-authored and difficulty-screened questions across seven research-assistant capability groups, 19 subtasks, and five scientific domains. Each question is instantiated in four matched settings formed by pairing its bilingual variants with the All Images First and Markdown Interleaved document representations, yielding 496 evaluation instances. The strongest evaluated model, Claude-Opus-5, scores 62.6 out of 100, with remaining gaps in evidence localization, structured information extraction, cross-document synthesis, and robustness to document representation. To convert these diagnostics into scalable training signals, we introduce SCIDOCIR, a structured representation of scientific document objects, layout and cross-reference relations, and provenance. Using SCIDOCIR, we construct SCIDOCDATASET, which contains 4K supervised fine-tuning instances and 10K reinforcement-learning instances, built on 14 verifiable training subtasks. Post-training Qwen3.6-27B on task-aligned data improves its SCIDOCBENCH score from 40.03 to 45.33 with supervised fine-tuning and to 45.74 with subsequent reinforcement learning. Both adapted models preserve DocVQA and InfoVQA performance and improve ChartQA accuracy over the original model by 0.80 and 3.40 points, respectively. Together, SCIDOCBENCH, SCIDOCIR, and SCIDOCDATASET connect capability diagnosis with verifiable training-data construction for scientific-document assistants.[1]

1 INTRODUCTION

Scientific papers communicate claims through text, equations, figures, tables, code, and datasets (Xia et al., 2024b; Li et al., 2024b). A useful scientific assistant must do more than *retrieve passages or summarize prose*: it must locate evidence, recover experimental details, verify numerical relations, compare papers, trace provenance, and produce reusable outputs (Dasigi et al., 2021; Starace et al., 2025). These operations underpin literature review, claim verification, and reproducibility analysis.

Scientific document understanding presents three challenges. **Heterogeneous structure:** scientific objects depend on layout, captions, reading order, and cross-references (Ma et al., 2024). **Compositional reasoning:** verifying one claim can require evidence from several objects, papers, or repositories (Li et al., 2024b). **Verifiable outputs:** answers must preserve numerical fidelity, structure, and provenance, while evaluation must distinguish task competence from sensitivity to language and document representation.

Existing benchmarks cover document QA, layout analysis, scientific QA, and chart reasoning (Mathew et al., 2021; Xu et al., 2020; Dasigi et al., 2021; Masry et al., 2022). Many empha-

[1] Code: `anonymous.4open.science/r/SciDocBench-review-materials-181E/`
Benchmark: `https://anonymous-hf.com/a/19n5senfvsfq/`
Training data: `https://anonymous-hf.com/a/snkurtvsk29n/`.

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.

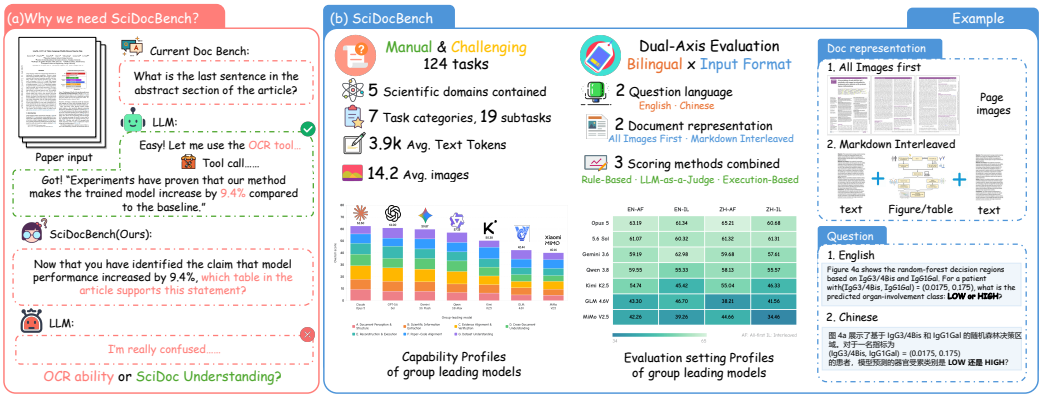


Figure 1: Overview of SciDocBench. (a) From text extraction to evidence grounding. A direct text-extraction prompt can be answered by recovering visible content, whereas the SciDocBench example additionally requires locating the document evidence supporting a scientific claim. **(b) Benchmark and evaluation design.** SciDocBench contains **124 expert-authored and difficulty-screened questions** across five scientific domains, seven capability groups, and 19 subtasks. Its **dual-axis protocol** pairs bilingual question variants with the *All Images First* and *Markdown Interleaved* document representations, producing **four matched settings** per question. Responses are assessed using task-appropriate rule-based, LLM-as-a-judge, or execution-based evaluators. The plots summarize capability-group and evaluation-setting profiles for group-leading models, while the example illustrates the matched question and document-representation variants.

size whether a model returns the correct answer; others also test supporting evidence or individual document operations. A complementary question is whether a model can produce the **verifiable intermediate deliverables** needed in scientific reading: an evidence location, a definition with its context, a consistency check, or an integrated record with provenance (Table 1). These deliverables expose failures that answer correctness alone may leave unresolved.

Figure 1 illustrates the distinction: identifying a claim is only part of the task when answering it also requires locating and interpreting the supporting table. We introduce SciDocBench, a **workflow-centered benchmark** with 124 expert-authored, difficulty-screened questions across five scientific domains, seven capability groups, and 19 subtasks. Its tasks bring evidence-handling operations into the foreground: models must locate support, interpret results under the relevant definitions, and reconcile information across papers. These operations become evaluation targets through question-specific deliverables, such as evidence attribution, structured extraction, or cross-paper synthesis.

Our **matched dual-axis protocol** pairs English and Chinese questions with *All Images First* or *Markdown Interleaved* documents, producing 496 instances with four settings. Task semantics, reference answers, and evaluation criteria remain fixed, isolating language and representation effects. Structured or executable outputs are assessed using rules, semantic judgment, or execution.

To turn diagnosis into training supervision, we introduce SciDocIR, a normalized representation that aligns visual and semantic document objects while preserving layout, cross-references, and provenance. It supports SciDocDataset, which contains approximately 4K SFT and 10K RL instances, built on **14 verifiable training subtasks**. Targets are checked through source records, controlled perturbations, exact operations, or execution. Benchmark questions are not converted into training examples.

Across 13 proprietary and open models, Claude-Opus-5 achieves the highest overall score of **62.6 out of 100**; no model leads more than two capability groups. Scientific extraction and cross-document synthesis remain difficult. For 10 models, the representation gap exceeds the language gap, showing that equivalent evidence can elicit different outcomes depending on its presentation.

We further examine whether task-aligned data improves a scientific-document assistant. Using Qwen3.6-27B, SFT and SFT followed by GRPO raise the SciDocBench score from 40.03 to 45.33 and 45.74, gains of 5.30 and 5.71 points over the original model. Both adapted models preserve general document understanding on DocVQA and InfoVQA and improve ChartQA by 0.80 and 3.40 points, respectively. These experiments demonstrate the training utility of the data pipeline alongside its role in benchmark diagnosis.

Table 1: Focused comparison with representative scientific-document benchmarks. ✓denotes primary coverage, –partial or adjacent coverage, and ✗that the dimension is not an evaluation target. **Cross-doc.** requires integrating multiple documents; **code/data prov.** denotes explicit reasoning over code, datasets, or provenance.

Benchmark	Primary focus	Full paper or PDF	Scientific objects	Cross-doc.	Code/data prov.
DocVQA (2021)	Document-image question answering	✗	✗	✗	✗
QASPER (2021)	Evidence-grounded full-paper QA	✓	–	✗	✗
ChartQA (2022)	Chart question answering	✗	–	✗	✗
SCITAB (2023)	Scientific-table claim verification	✗	✓	✗	✗
MMLongBench-Doc (2024)	Long-context document multimodal QA	✓	–	✗	✗
DocGenome (2024b)	Scientific-document structure and QA	✓	✓	✗	✗
M3SciQA (2024b)	Multimodal multi-document QA	✓	✓	✓	✗
PaperBench (2025)	End-to-end research replication	✓	–	✗	✓
ArXivDoc (2026)	Scientific retrieval across representations	✓	✓	✗	✗
PaperScope (2026)	Agentic multi-document deep research	✓	✓	✓	✗
SciDocBench	**Workflow-centered scientific-document understanding**	✓	✓	✓	✓

Our contributions are as follows: **1) Workflow-centered benchmark.** We introduce SciDocBench, a workflow-centered benchmark that evaluates concrete scientific research tasks across full documents, multiple papers, and associated code or data, using a matched bilingual and dual-representation protocol. **2) Systematic model evaluation.** We evaluate proprietary and open multimodal models, revealing complementary capability profiles, persistent weaknesses in structured extraction and cross-document synthesis, and substantial sensitivity to document representation. **3) Evaluation-to-training resources.** We construct SciDocIR and SciDocDataset, connecting benchmark diagnostics with scalable, verifiable training data. SFT and SFT followed by GRPO improve scientific-document performance while preserving general document question answering.

2 Related Work

Document understanding benchmarks. PubLayNet and LayoutLM support layout annotation and layout-aware representations (Zhong et al., 2019; Xu et al., 2020); DocVQA and DUE evaluate document-image QA (Mathew et al., 2021; Borchmann et al., 2021). MMLongBench-Doc extends evaluation to long multimodal documents (Ma et al., 2024), and ArXivDoc compares image, text, and interleaved scientific-document representations (Khalighinejad et al., 2026). SciDocBench examines how these perceptual and reading capabilities compose within scientific workflows.

Scientific question answering and research assistance. QASPER provides full-paper questions and supporting evidence (Dasigi et al., 2021), while PubMedQA focuses on biomedical abstracts (Jin et al., 2019). M3SciQA and PaperScope address multimodal, multi-document QA and deep research (Li et al., 2024b; Xiong et al., 2026); PaperBench evaluates end-to-end research replication (Starace et al., 2025). SciDocBench complements these resources with explicit tests of claim grounding, notation extraction, consistency checking, and structured result integration.

Chart, table, and scientific-object reasoning. PlotQA and ChartQA evaluate graphical and numerical reasoning (Methani et al., 2020; Masry et al., 2022); SCITAB tests compositional verification over scientific tables (Lu et al., 2023). SciDocBench places such operations within full documents, where evidence may span a figure, caption, equation, appendix, or another paper and must retain its provenance.

Structured scientific-document resources and multimodal training data. S2ORC and DocGenome structure scholarly text, references, and multimodal document objects (Lo et al., 2020; Xia et al., 2024b). Visual instruction tuning and MM-IFEngine use generated multimodal supervision and output constraints (Liu et al., 2023; Ding et al., 2025), while DocSeeker combines SFT with evidence-aware RL (Yan et al., 2026). Our framework emphasizes task-specific verifiability: SciDocIR aligns rendered PDF content with available LaTeX semantics, and SciDocDataset uses source records and controlled operations to construct training targets.

Positioning and scope. Table 1 compares full-document input, scientific-object reasoning, cross-document understanding, and code or data provenance. Partial coverage denotes a related signal rather than a central evaluation target. SciDocBench organizes selected scientific-reading opera-

tions around **evidence requirements and verifiable deliverables**, using matched inputs to diagnose where these operations break down.

3 THE SCIDOCBENCH BENCHMARK

3.1 BENCHMARK OVERVIEW AND DESIGN PRINCIPLES

SCIDOCBENCH contains 124 expert-authored questions in seven capability groups and 19 subtasks. English and Chinese variants are paired with two document representations, yielding 496 matched instances that test evidence localization, scientific-object interpretation, verification, cross-document integration, and connections to code and data.

Three principles guide construction. **Workflow realism:** questions derive from concrete activities such as reviewing claims, integrating results, or reconstructing figures. **Evidence fidelity and controlled variation:** matched settings preserve task-relevant content while varying language and representation. **Verifiable structure:** questions request structured or executable deliverables when the research activity demands them, enabling task-appropriate evaluation.

3.2 TASK TAXONOMY

SCIDOCBENCH organizes its questions into seven capability groups. **Document Perception and Structure**, **Scientific Information Extraction**, and **Evidence Alignment and Verification** evaluate whether models can locate, interpret, extract, and check information within scientific documents. **Cross-Document Understanding** evaluates the integration of notation, datasets, and results across papers. **Reconstruction and Execution**, **Paper–Code Alignment**, and **Dataset Understanding** examine whether models can transform scientific content into reusable artifacts and connect papers with associated computational resources. The complete inventory of 19 subtasks, together with their definitions and evaluation focus, is provided in Appendix B.

3.3 EVALUATION INSTANCES AND SCORING

Model inputs. Each instance supplies all task-relevant source materials followed by an English or Chinese question. All Images First presents every paper page as an image; Markdown Interleaved arranges extracted text and visual elements in document order. Multi-document tasks supply all required papers. Paper–code and dataset tasks may additionally include file trees, code excerpts, or dataset descriptions. Reference answers, task labels, and evaluation metadata are withheld.

Output contract. The question specifies the requested deliverable and any exact format requirements. Outputs include labels, values, spans, ordered lists, JSON, tables, graphs, reconstructed data, LaTeX, or programs. These formats reflect scientific outputs that must be parsed, verified, or reused, rather than arbitrary formatting constraints.

Evaluation and score aggregation. We assign each underlying question one of three evaluators: deterministic rules for 57 questions, a semantic LLM judge for 66, and execution for one image-transformation task. Rules compare specified fields, values, or orderings; the judge applies the question-specific rubric to the reference and final answer; execution compares generated and reference outputs. Appendix B.1 defines partial credit and failure handling and lists the evaluator for every question. Each instance receives a score between zero and one; failed, empty, or unusable responses receive zero. The overall score averages all 496 instances, scaled to 0–100.

Capability-group and setting scores average their respective instances on the same scale. All four variants preserve source content, task and reference-answer semantics, and scoring criteria, enabling paired sensitivity analysis. Scores assess final answers, excluding internal reasoning traces; an explanation explicitly requested in the final answer remains part of its output contract.

3.4 BENCHMARK CONSTRUCTION

Figure 2 summarizes paper selection, expert authoring and review, difficulty screening, and controlled augmentation. Questions must be scientifically meaningful, answerable from the supplied materials, unambiguous, and challenging. Appendix G provides representative examples.

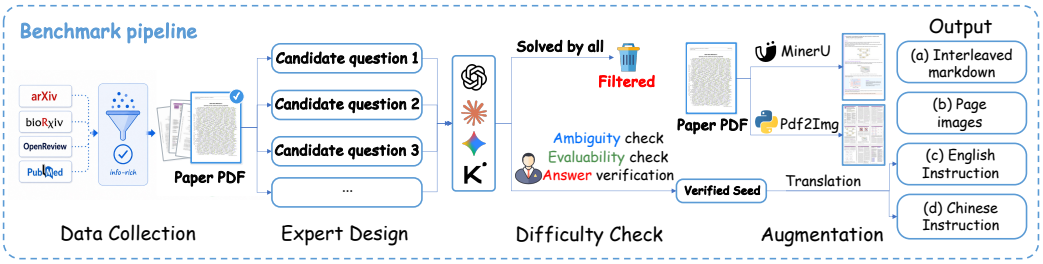


Figure 2: SCIDOCBENCH construction. Information-rich papers support expert-authored research tasks. Difficulty screening and independent verification reject trivial or ambiguous items. Bilingual and dual-representation augmentation produces four matched instances per accepted question.

Paper selection. We collect 116 publicly accessible papers spanning five scientific domains from arXiv, bioRxiv, OpenReview, PubMed, and corresponding journal pages. Selection favors information-rich papers with figures, tables, equations, captions, appendices, and experimental descriptions; we record these object types and counts to characterize document richness.

Papers supporting only generic summarization or unstable answers are excluded. Theoretical papers support notation and equation tasks; empirical papers support experimental-setting extraction, table consistency, and result integration. Appendix B records source papers and construction statistics.

Question authoring and review. Each candidate records its source documents, question, reference answer, task labels, evidence locator, and evaluator. A reviewer other than the author checks answerability, reference correctness, and instruction and output clarity. Structured-output requirements appear in the question itself. Appendix C documents field-level requirements and the independent-review interfaces.

3.5 QUALITY CONTROL AND EVALUATION RELIABILITY

Difficulty screening. To remove questions already solved by strong systems, we evaluate each candidate once using Claude Opus 4.8 (Anthropic, 2026a), Gemini 3.1 Pro (Google DeepMind, 2026a), ChatGPT 5.5 (OpenAI, 2026c), and Kimi K2.5 (Kimi Team, 2026). A candidate is removed as overly easy when all four responses are fully correct under its task-specific scorer.

A retained candidate must have at least one failed screening response. All-model failures still require independent confirmation of sufficient evidence, a stable reference, and an unambiguous output; apparent difficulty alone does not justify acceptance.

Controlled augmentation. English and Chinese questions are paired with both document views from Section 3.3. Each variant is checked for semantic equivalence and answerability while preserving the reference, output contract, and scoring criteria. Translated or reformatted variants are not counted as independently authored questions.

Reliability of semantic evaluation. We audit GPT-5.4-mini (OpenAI, 2026b) judgments on 100 responses spanning the four settings, 11 models, and all seven capability groups. A human expert reviews anonymized responses without access to model identities or GPT judgments. With incorrect, partial, and correct ratings, human–GPT agreement reaches 92.0% and quadratic-weighted Cohen's $\kappa = 0.866$. Including an independent Codex blind review yields ordinal Krippendorff's $\alpha = 0.854$, with a 95% bootstrap confidence interval of $[0.757, 0.924]$. Appendix C.1 reports the sampling protocol, score mapping, binary comparison, and review interface.

4 SCIDOCIR AND SCIDOCDATASET

Scientific objects can appear as LaTeX environments, rasterized PDF regions, or visually separated blocks. This heterogeneity complicates evidence retrieval, controlled editing, and verification. We normalize papers into SCIDOCIR and construct SCIDOCDATASET through verifiable operations over the shared representation (Figure 3).

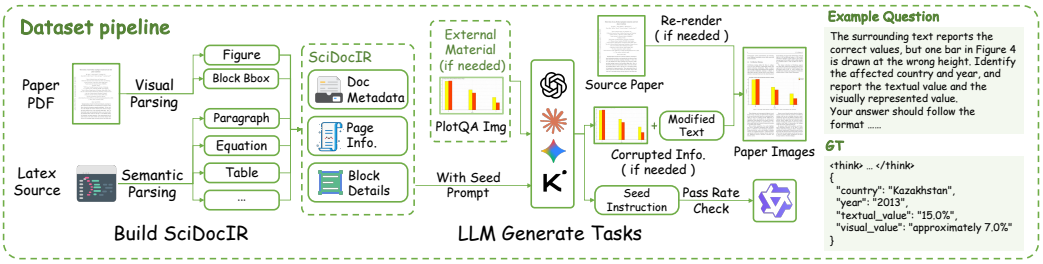


Figure 3: SciDocDataset construction. Aligned PDF and LaTeX records form SciDocIR. Task-specific generation preserves locked source fields, optionally re-renders controlled evidence into page blocks, and validates instructions and targets against source or construction records.

4.1 SciDocIR Construction

PDF pages provide the rendered view: visual parsing identifies text, figures, equations, tables, and captions, recording page indices, bounding boxes, categories, crops, and spatial relations. Available LaTeX sources provide semantic content, section structure, labels, and references. We align these views through text, labels, caption associations, and geometry, preserving both published appearance and source semantics.

Records have three levels: **document metadata** stores identity, domain, and provenance; **page information** stores indices, dimensions, images, and layout; **block details** store content, source code, geometry, captions, reading order, and references. Relations connect continuations, parent–child structures, caption targets, and evidence chains. Appendix D gives the full schema.

4.2 Verifiable Training-Task Design

The 19 benchmark subtasks provide broad diagnostic coverage; the 14 training subtasks are selected for verifiability. They cover layout and reading flow, table logic, chart consistency and readout, cross-document table integration and dataset intersection, notation extraction, symbol disambiguation, and citation roles. Answers are checked through SciDocIR fields, controlled perturbations, exact operations, numerical tolerances, parsing, or execution. This separation supports scalable supervision without converting held-out benchmark questions into training examples. Appendix E lists all construction directions.

4.3 Seeded Generation and Source Authority

Each seed instruction specifies the operation, source records, output schema, and rejection conditions. Retrieval selects relevant SciDocIR pages and blocks; externally sourced or synthetic tasks provide locked values, labels, or formulas with rendered context. GPT naturalizes questions, proposes bounded candidates, or checks semantics and solvability, but cannot alter deterministic fields or construction records. Appendix E.2 specifies model roles and ground-truth authority.

Evidence sources vary by task. Real table integration retrieves compatible tables; simulated integration splits a source table into overlapping, column-permuted views. Chart tasks use PlotQA (Methani et al., 2020) series values and controlled Matplotlib (Hunter, 2007) renderings. Dataset intersection uses a lexicon of approximately 50 scientific resources, while notation tasks use formula and layout templates rendered with XeLaTeX. These sources define exact targets; SciDocIR adds realistic document context where needed.

4.4 Controlled Block-Level Re-rendering

We insert synthetic figures or tables into their recorded SciDocIR bounding boxes, updating the caption or nearby statement when required while preserving unrelated content. This retains the paper's layout and scientific context but gives exact control over task evidence. Unmodified tasks retain original pages. Rendering records store the source, block, geometry, inserted content, textual edits, and expected answer for verification.

Table 2: SCIDOCBENCH leaderboard. Overall is computed over all 496 evaluation instances derived from 124 underlying questions, with failed or unusable responses assigned a score of zero. Evaluation-setting scores pair English or Chinese questions with the All Images First (Images) or Markdown Interleaved (Interleaved) document representation. Capability scores correspond to the seven groups defined in Section 3. Model families are ordered by their best overall score, and models within each family are sorted by overall score. In each score column the best, second, and third values are shaded from dark to light; ties receive the same shading.

Model	Overall	English		Chinese		Capability scores						
		Images	Inter-leaved	Images	Inter-leaved	Document Perception	Information Extraction	Evidence Verification	Cross-Document	Reconstr. & Exec.	Paper–Code	Dataset Underst.
Claude-Opus-5 (Anthropic, 2026b)	**62.60**	**63.19**	61.34	**65.21**	60.68	**61.37**	55.52	74.47	62.81	59.65	51.75	43.75
Claude-Opus-4.8 (Anthropic, 2026a)	53.41	53.63	52.43	54.13	53.46	46.59	**57.12**	58.63	56.65	57.09	44.00	53.12
Claude-Sonnet-4.6 (Anthropic, 2026c)	49.57	52.25	51.18	53.62	41.21	44.80	51.51	54.25	47.43	52.64	52.17	59.38
GPT-5.6-Sol (OpenAI, 2026e)	61.00	61.07	60.32	61.32	**61.31**	54.08	56.78	**78.16**	54.62	**69.70**	46.69	59.38
GPT-5.6-Terra (OpenAI, 2026f)	54.19	51.88	55.69	52.83	56.37	44.01	51.74	66.28	54.01	65.75	61.88	56.25
GPT-5.6-Luna (OpenAI, 2026d)	50.58	47.55	53.99	46.65	54.13	38.68	47.52	65.15	56.07	53.86	54.25	62.50
Gemini-3.6-Flash (Google DeepMind, 2026b)	59.87	59.19	**62.98**	59.68	57.61	54.00	55.19	73.28	62.42	55.42	58.31	56.25
Qwen3.8-Max (Qwen Team, 2026d)	57.15	59.55	55.33	58.13	55.57	48.17	50.79	73.84	**63.02**	53.60	55.75	59.38
Qwen3.7-Plus (Qwen Team, 2026b)	55.88	57.17	53.34	57.57	55.42	49.00	53.84	63.72	59.04	55.31	65.06	**71.88**
Qwen3.8-27B (Qwen Team, 2026c)	52.89	58.65	47.25	54.21	51.46	47.43	50.50	62.87	54.80	54.45	54.06	31.25
Kimi-K2.5 (Kimi Team, 2026)	50.38	54.74	45.42	55.04	46.33	45.69	50.62	53.05	53.61	50.41	**70.25**	40.62
GLM-4.6V (Z.ai, 2025)	42.44	43.30	46.70	38.21	41.56	37.15	46.58	41.63	44.57	43.90	57.27	59.38
MiMo-V2.5 (Xiaomi MiMo Team, 2026)	40.16	42.26	39.26	44.66	34.46	34.11	37.66	49.60	36.40	44.84	55.30	43.75

4.5 VALIDATION AND OUTPUT RECORDS

Task-specific checks reject missing fields, conflicts with locked records, insufficient visual evidence, and inconsistent or non-executable outputs. Accepted samples contain the instruction, document input, and verifiable answer; source identifiers, locked fields, rendering logs, and rejection outcomes remain in separate audit metadata. This preserves a traceable link between model-assisted wording and answer authority.

Task-aligned training release. SCIDOCDATASET contains 3,924 SFT instances from 981 seeds under the four language–representation settings, and 10,143 RL instances combining these questions with 6,219 selected legacy examples. It emphasizes evidence localization, structured extraction, numerical verification, and reusable scientific outputs. Appendix B reports data statistics; Appendix F records training setting used in the Qwen3.6 experiments.

5 EXPERIMENTS

We evaluate 13 proprietary and open multimodal models on all 496 matched SCIDOCBENCH instances, reporting overall, capability-group, and setting-level scores. We then test whether task-aligned post-training improves scientific-document understanding while preserving general document performance. Appendix F provides experimental configurations and evaluation protocols.

5.1 OVERALL LEADERBOARD

Table 2 reports the SCIDOCBENCH leaderboard. Claude-Opus-5 ranks first with an overall score of 62.60, followed by GPT-5.6-Sol with 61.00 and Gemini-3.6-Flash with 59.87. Qwen3.8-Max and Qwen3.7-Plus obtain 57.15 and 55.88, respectively. No evaluated model reaches 63, and overall scores span 22.44 points across the evaluated systems.

5.2 CAPABILITY-LEVEL ANALYSIS

Capability leadership is distributed across model families. Claude models lead Groups A and B, GPT-5.6-Sol leads C and E, Qwen models lead D and G, and Kimi-K2.5 leads F. No model leads more than two groups. Claude-Opus-5's overall lead reflects strength across several groups rather than dominance of the complete taxonomy.

Document perception and scientific extraction form common bottlenecks. Averaged equally across the 13 models, Group A has the lowest score at 46.54, followed by Group B at 51.18. Group C has the highest mean at 62.69. The capability ceilings exhibit a related pattern: the best Group B score is 57.12, while Groups A and D peak at 61.37 and 63.02. These results indicate that locating

document elements, recovering structured scientific information, and integrating evidence across papers remain difficult even when models perform comparatively well on evidence verification.

High overall performance does not imply a balanced capability profile. The gap between each model's strongest and weakest capability groups ranges from 14.58 to 31.62 points. Even Claude-Opus-5 and GPT-5.6-Sol span 30.72 and 31.47 points, with distinct weaknesses in dataset understanding and paper–code alignment, respectively. Overall scores obscure these capability-specific limitations. These contrasting profiles suggest that strong performance on some scientific-document operations does not consistently transfer to others, motivating targeted evaluation and training.

Strong models still lose the scope and completeness of evidence. Three Claude-Opus-5 responses illustrate failures beyond answer formatting. Asked for the ordered metrics in a specified section of Depth Anything 3, it includes all four target metrics but prepends two out-of-scope ones, violating the requested list. On a logarithmic plot, it identifies the correct QAOA range but reads the AQC lower endpoint as 5×10^{-3} instead of 8×10^{-4}. In a table audit, it catches two planted errors but misses another incorrect mean, returning an extra highlighting complaint instead. These examples point to distinct problems in scope control, scale-aware readout, and exhaustive consistency checking, even when much of the response is locally correct. Appendix F.6 gives the references and predicted outputs.

5.3 Language and Representation Robustness

Aggregate language differences conceal model-specific effects. Averaged equally over the 13 models and two document representations, English questions score 53.45 and Chinese questions score 52.72, a difference of 0.72 points before rounding. Individual model effects are considerably larger and differ in direction. Qwen3.7-Plus improves by 1.24 points with Chinese questions, whereas GLM-4.6V and Claude-Sonnet-4.6 decrease by 5.12 and 4.30 points, respectively. Language sensitivity is therefore model dependent rather than a uniform translation penalty.

Document representation changes both scores and rankings. Across all models and languages, All Images First exceeds Markdown Interleaved by 1.96 points on average. However, nine models favor All Images First and four favor Markdown Interleaved. Kimi-K2.5 and Qwen3.8-27B favor All Images First by 9.02 and 7.07 points, whereas GPT-5.6-Luna favors Markdown Interleaved by 6.96 points. Claude-Opus-5 leads both All Images First settings, Gemini-3.6-Flash leads EN-IL, and GPT-5.6-Sol leads ZH-IL. Neither representation is therefore uniformly preferable across systems.

Language and representation effects can interact. For Gemini-3.6-Flash, Markdown Interleaved exceeds All Images First by 3.79 points with English questions, but All Images First exceeds Markdown Interleaved by 2.07 points with Chinese questions. Qwen3.8-27B exhibits a different reversal: Chinese questions reduce its All Images First score by 4.44 points relative to English but improve its Markdown Interleaved score by 4.21 points. These reversals show that the two controlled axes should be analyzed jointly rather than interpreted as independent sources of performance variation.

Cross-setting stability is distinct from peak accuracy. The average range across the four evaluation settings is 6.55 points. GPT-5.6-Sol varies by only 1.00 point and Claude-Opus-4.8 by 1.70 points, whereas Claude-Sonnet-4.6, Qwen3.8-27B, and MiMo-V2.5 exhibit ranges above 10 points. Reporting only the best setting would therefore conflate scientific-document capability with compatibility with a particular language and document representation.

5.4 Post-training with SciDocDataset

We compare the original Qwen3.6-27B dense model (Qwen Team, 2026a) with task-aligned SFT and subsequent group relative policy optimization (GRPO; Shao et al., 2024). These experiments use SCIDOCDATASET. Both stages use lightweight LoRA adaptation (Hu et al., 2022). Data composition, checkpoint selection, optimization, and reward settings are detailed in Appendix F.

Task-aligned data improves scientific-document performance. Table 3 shows gains of 5.30 points for SFT and 5.71 points for SFT followed by GRPO over the original model, reaching 45.33 and 45.74 versus 40.03. SFT improves all seven capability-group aggregates. Large gains occur in reconstruction and execution (42.28 to 52.48) and paper–code alignment (40.31 to 58.94), where the model must convert document evidence into reusable outputs. Relative to the same baseline, the

Table 3: Post-training effectiveness and general-document transfer. Qwen3.6-27B and its adapted variants are evaluated on all 496 SCIDOCBENCH instances and three general document benchmarks. Input settings follow Table 2. DocVQA and InfoVQA report validation ANLS; ChartQA reports test relaxed accuracy. Scores use final answers only; best scores are bold. Improvement rows give absolute score changes from the original model, computed before rounding; green/red indicates gains/drops.

Model	SCIDOCBENCH												General document understanding		
	Overall	Capability scores							English		Chinese				
		Document Perception	Information Extraction	Evidence Verification	Cross-Document	Reconstr. & Exec.	Paper–Code	Dataset Underst.	Images	Interleaved	Images	Interleaved	DocVQA	InfoVQA	ChartQA
Qwen3.6-27B	40.03	38.14	38.55	44.71	38.46	42.28	40.31	34.38	44.61	38.61	35.38	41.53	95.81	90.67	74.60
+ SFT	45.33	**40.11**	**43.55**	**51.40**	43.85	52.48	**58.94**	**40.62**	**46.48**	42.76	45.89	46.20	95.80	**91.02**	75.40
Improvement	+5.30	+1.97	+5.01	+6.69	+5.39	+10.20	+18.62	+6.25	+1.87	+4.15	+10.51	+4.67	-0.01	+0.36	+0.80
+ SFT + GRPO	**45.74**	39.18	42.77	50.43	**46.90**	**61.32**	57.10	**40.62**	44.00	**44.00**	**47.38**	**47.59**	**95.87**	90.89	**78.00**
Improvement	+5.71	+1.04	+4.22	+5.73	+8.45	+19.04	+16.79	+6.25	-0.62	+5.39	+12.00	+6.06	+0.05	+0.23	+3.40

GRPO-adapted model reaches 61.32 in reconstruction and execution and 46.90 in cross-document understanding, gains of 19.04 and 8.45 points. Both adapted models therefore improve scientific-document performance, with distinct capability profiles.

Supervised adaptation transfers across the four input settings. SFT improves both languages and both document representations, rather than only the best-performing interface. The gain ranges from 1.87 points on English All Images First to 10.51 points on Chinese All Images First. Averaged over languages, Markdown Interleaved improves from 40.07 to 44.48 after SFT and to 45.80 after GRPO. The matched variants therefore provide useful supervision for expressing the same scientific operation under different input conditions. Full setting-level results are given in Table 3.

Producing a final deliverable is part of the task. The evaluation permits internal reasoning but scores only the final response against the requested answer contract. Under the same 32,768-token generation budget, the number of instances with a nonempty extracted final answer increases from 323 of 496 for the original model to 461 after SFT and 459 after GRPO. This improvement is relevant to the workflow setting: a model must finish its analysis and return the requested table, structured record, code, or evidence-grounded answer. The training examples pair scientific evidence with explicit output requirements, helping the model translate its analysis into a usable deliverable.

5.5 TRANSFER TO GENERAL DOCUMENT BENCHMARKS

We evaluate all three models on DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), and ChartQA (Masry et al., 2022), without training on their general-document splits. Table 3 shows that DocVQA remains near 95.8 across the three models. Relative to the original model's 90.67 on InfoVQA, SFT and SFT followed by GRPO obtain 91.02 and 90.89. With a shared native VLMEvalKit prompt and the recommended Qwen sampling settings, ChartQA scores reach 75.40 and 78.00 versus 74.60 for the original model, gains of 0.80 and 3.40 points.

The ChartQA gains are consistent with the chart readout, numerical verification, and structured extraction tasks in the training collection. Together, the results show that SCIDOCDATASET supports scientific-document specialization while preserving general document question answering.

6 CONCLUSION

We introduced SCIDOCBENCH, a workflow-centered benchmark for scientific document understanding with 124 questions across seven capability groups, 19 subtasks, and four matched evaluation settings. The strongest model scores 62.60, capability leaders differ across model families, and performance depends on document representation. We further introduced SCIDOCIR and SCIDOCDATASET to connect these diagnostics with verifiable training supervision. Task-aligned SFT and SFT followed by GRPO improve Qwen3.6-27B over the original model by 5.30 and 5.71 points on SCIDOCBENCH, while preserving DocVQA and InfoVQA performance and improving ChartQA by 0.80 and 3.40 points. Scientific document understanding is not OCR, long-document QA, or figure reading in isolation: evidence grounding, structured extraction, verification, reconstruction, and provenance tracking must produce outputs that can be checked and reused in scientific work.

Reproducibility Statement

The benchmark construction protocol, task taxonomy, document representations, evaluation contracts, and training-data generation procedure are specified in Sections 3–5 and the appendices. Appendix B reports dataset statistics, per-question evaluators, and paper-disjoint splits; Appendix F distinguishes the experimental data version and records model adaptations, checkpoint selection, and final-answer evaluation settings. The four benchmark views preserve the identity of each underlying question, allowing results to be compared at both the question and setting levels.

AI Use Statement

AI models were used in the model-assisted construction and rewriting of training examples, bilingual adaptation, reference-conditioned explanation generation, and LLM-based answer evaluation, as described in the paper. The authors are responsible for the scientific claims, reported results, and final submitted content.

References

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

Dong An and Lin Lin. Quantum linear system solver based on time-optimal adiabatic quantum computing and quantum approximate optimization algorithm. *ACM Transactions on Quantum Computing*, 3(2):1–28, 2022.

David Anderson, Greta Panova, and Leonid Petrov. Computation and sampling for schubert specializations. *arXiv preprint arXiv:2603.20104*, 2026.

Anthropic. Introducing Claude Opus 4.8. `https://www.anthropic.com/news/claude-opus-4-8`, May 2026a. Accessed September 4, 2026.

Anthropic. Introducing Claude Opus 5. `https://www.anthropic.com/news/claude-opus-5`, July 2026b. Accessed September 4, 2026.

Anthropic. Introducing Sonnet 4.6. `https://www.anthropic.com/news/claude-sonnet-4-6`, February 2026c. Accessed September 4, 2026.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

D Basilico, G Bellini, J Benziger, R Biondi, B Caccianiga, A Caminata, A Chepurnov, DD Angelo, A Derbin, A Di Giacinto, et al. New limits on the pauli forbidden transitions in 12c nuclei obtained with the complete borexino dataset. *arXiv preprint arXiv:2604.08950*, 2026.

Lukasz Borchmann, Michal Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michal Turski, Karol Szyndler, and Filip Gralinski. DUE: End-to-end document understanding benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Zi Cai and Jinguo Liu. Approximating quantum many-body wave functions using artificial neural networks. *Physical Review B*, 97(3):035116, 2018.

Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. *arXiv preprint arXiv:2512.07733*, 2025.

Yuan Cao, Valla Fatemi, Shiang Fang, Kenji Watanabe, Takashi Taniguchi, Efthimios Kaxiras, and Pablo Jarillo-Herrero. Magic-angle graphene superlattices: a new platform for unconventional superconductivity. *arXiv preprint arXiv:1803.02342*, 2018.

Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025a.

Bo Chen, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Provable failure of language models in learning majority boolean logic via gradient descent. *arXiv preprint arXiv:2504.04702*, 2025b.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025c.

Sully F Chen, Anton Alyakin, Andreas Seas, Eunice Yang, Joanne J Choi, Jin Vivian Lee, Amelia L Chen, Pranav I Warman, Rochelle T Bitolas, Robert J Steele, et al. Llm-assisted systematic review of large language models in clinical medicine. *Nature medicine*, pp. 1–8, 2026.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256.

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*, 2021.

Isaac Davis, Julian Jara-Ettinger, and Yarrow Dunham. Inferring the internal structure of groups through the integration of statistical learning and causal reasoning. *Nature Communications*, 2026.

Selvan Demir, Miguel I Gonzalez, Lucy E Darago, William J Evans, and Jeffrey R Long. Giant coercivity and high magnetic blocking temperatures for n23- radical-bridged dilanthanide complexes upon ligand dissociation. *Nature Communications*, 8(1):2144, 2017.

Dong-Ling Deng, Xiaopeng Li, and S Das Sarma. Quantum entanglement in neural network states. *Physical Review X*, 7(2):021021, 2017.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. MM-IFEngine: Towards multimodal instruction following. *arXiv preprint arXiv:2504.07957*, 2025.

Tuan Dinh, Seon-Kyeong Jang, Noah Zaitlen, and Vasilis Ntranos. Compressing the collective knowledge of esm into a single protein language model. *Nature Methods*, pp. 1–13, 2026.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Amit Agarwal, Zhe Chen, Mo Li, Yubo Ma, Hailong Sun, Xiangyu Zhao, Junbo Cui, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024. URL `https://arxiv.org/abs/2407.11691`.

Jusaina Eyyathiyil, Subhajit Ghosh, Arunima Cheran, Silvano Geremia, Jatish Kumar, Neal Hickey, and Pakkirisamy Thilagar. Axial chirality-induced rigidification in aminoboranes enhances persistent room-temperature phosphorescence and circularly polarized luminescence. *Communications chemistry*, 8(1):126, 2025.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

Sukhpal Singh Gill, Adarsh Kumar, Harvinder Singh, Manmeet Singh, Kamalpreet Kaur, Muhammad Usman, and Rajkumar Buyya. Quantum computing: A taxonomy, systematic review and future directions. *Software: Practice and Experience*, 52(1):66–114, 2022.

Ziyang Gong, Zehang Luo, Anke Tang, Zhe Liu, Shi Fu, Zhi Hou, Ganlin Yang, Weiyun Wang, Xiaofeng Wang, Jianbo Liu, et al. Ace-brain-0: Spatial intelligence as a shared scaffold for universal embodiments. *arXiv preprint arXiv:2603.03198*, 2026.

Google DeepMind. Gemini 3.1 Pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, February 2026a. Accessed September 25, 2026.

Google DeepMind. Gemini 3.6 Flash model card. `https://deepmind.google/models/model-cards/gemini-3-6-flash/`, July 2026b. Accessed September 4, 2026.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13869–13920, 2025.

Jing Han, Zhuochao Zhou, Rongrong Zhang, Yijun You, Zizhen Guo, Jinyan Huang, Fan Wang, Yue Sun, Honglei Liu, Xiaobing Cheng, Yutong Su, Hui Shi, Qiongyi Hu, Jialin Teng, Chengde Yang, Shifang Ren, and Junna Ye. Fucosylation of anti-dsdna igg1 correlates with disease activity of treatment-naïve systemic lupus erythematosus patients. *eBioMedicine*, 77:103883, 2022. doi: 10.1016/j.ebiom.2022.103883.

Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007. doi: 10.1080/19312450709336664.

Serkan Hoşten, Vadym Kurylenko, Elke Neuhaus, and Nikolas Rieke. The euler stratification for $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^n$. *arXiv preprint arXiv:2603.19184*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://arxiv.org/abs/2106.09685`.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Shoham Jacoby, Alex Retzker, and Fernando Pastawski. Stairway codes: Floquetifying bivariate bicycle codes and beyond. *arXiv preprint arXiv:2603.00228*, 2026.

Bin Jiang, Adrien Bouhon, Zhi-Kang Lin, Xiaoxi Zhou, Bo Hou, Feng Li, Robert-Jan Slager, and Jian-Hua Jiang. Experimental observation of non-abelian topological acoustic semimetals and their phase transitions. *Nature Physics*, 17(11):1239–1246, 2021.

Haoyi Jiang, Liu Liu, Xinjie Wang, Yonghao He, Wei Sui, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Spa3r: Predictive spatial field modeling for 3d visual reasoning. *arXiv preprint arXiv:2602.21186*, 2026.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

William R Johnson III, Xiaonan Huang, Shiyang Lu, Kun Wang, Joran W Booth, Kostas Bekris, and Rebecca Kramer-Bottiglio. Impact-resistant, autonomous robots inspired by tensegrity architecture. *arXiv preprint arXiv:2501.15078*, 2025.

Ghazal Khalighinejad, Raghuveer Thirukovalluru, Alexander H Oh, and Bhuwan Dhingra. Document-as-image representations fall short for scientific retrieval. *arXiv preprint arXiv:2604.18508*, 2026.

Hori Kim, Moon-Ki Jeong, Hyuk-Joon Kim, Youngsin Kim, Kisuk Kang, and Joon Hak Oh. Enhanced cycling stability of ncm811 cathodes at high c-rates and voltages via limtfsi-based polymer coating. *Small*, 21(30):2502816, 2025.

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026. URL https://arxiv.org/abs/2602.02276.

Zitai Kong, Yiheng Zhu, Yinlong Xu, Mingze Yin, Tingjun Hou, Jian Wu, Hongxia Xu, and Chang-Yu Hsieh. ProtFlow: Flow matching-based protein sequence design with comprehensive protein semantic distribution learning and high-quality generation. *bioRxiv*, 2026. doi: 10.64898/2026.02.14.705870. URL https://www.biorxiv.org/content/10.64898/2026.02.14.705870v1.

Inhoe Koo, Hyunho Cha, and Jungwoo Lee. Qflownet: Fast, diverse, and efficient unitary synthesis with generative flow networks. *arXiv preprint arXiv:2603.03045*, 2026.

Talya S Kramer, Flossie K Wan, Sarah M Pugliese, Adam A Atanas, Sreeparna Pradhan, Alex W Hiser, Lillie M Godinez, Jinyue Luo, Eric Bueno, Thomas Felt, et al. Neural sequences underlying directed turning in caenorhabditis elegans. *Nature Neuroscience*, pp. 1–17, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Barbara Maria Latacz, Stefan R Erlewein, Markus Fleck, Julia I Jäger, Fatma Abbass, Béla P Arndt, P Geissler, Tomoka Imamura, Marcel Leonhardt, Peter Micke, et al. Coherent spectroscopy with a single antiproton spin. *Nature*, 644(8075):64–68, 2025.

Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed-ensembles for efficient uncertainty estimation. *arXiv preprint arXiv:2210.09184*, 2022.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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, 2024b. doi: 10.18653/v1/2024.findings-emnlp.904. URL https://aclanthology.org/2024.findings-emnlp.904/.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025.

Wangda Li, Xiaoming Liu, Qiang Xie, Ya You, Miaofang Chi, and Arumugam Manthiram. Long-term cyclability of ncm-811 at high voltages in lithium-ion batteries: an in-depth diagnostic study. *Chemistry of Materials*, 32(18):7796–7804, 2020.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, pp. 141261–141285, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e4cd50120b6d7e8daff1749d6bbaa889-Paper-Conference.pdf.

Zhichao Lin, Zhichao Liang, Gaoqiang Liu, Meng Xu, Baoyu Xiang, Jian Xu, and Guanjun Jiang. Quarkmedsearch: A long-horizon deep search agent for exploring medical intelligence. *arXiv preprint arXiv:2604.12867*, 2026b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Junhua Liu, Zhangcheng Wang, Zhike Han, Ningli Wang, Guotao Liang, and Kun Kuang. Twiff (think with future frames): A large-scale dataset for dynamic visual reasoning. *arXiv preprint arXiv:2602.10675*, 2026.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b. URL https://arxiv.org/abs/2503.20783.

Princess Stephanie Llanos, Zahra Ahaliabadeh, Ville Miikkulainen, Jouko Lahtinen, Lide Yao, Hua Jiang, Timo Kankaanpaa, and Tanja M Kallio. High voltage cycling stability of lif-coated nmc811 electrode. *ACS Applied Materials & Interfaces*, 16(2):2216–2230, 2024.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. S2ORC: The Semantic Scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Changze Lv, Yifei Wang, Yanxun Zhang, Yiyang Lu, Jingwen Xu, Xiaohua Wang, Di Yu, Xin Du, Xuanjing Huang, and Xiaoqing Zheng. Biologically plausible learning via bidirectional spike-based distillation. *arXiv preprint arXiv:2509.20284*, 2025.

Shuo Ma, Muhao Chen, Hongying Zhang, and Robert E Skelton. Statics of integrated origami and tensegrity systems. *International Journal of Solids and Structures*, 279:112361, 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. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7775–7803, 2024.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html.

Jonathan Mi, Wenzhe Tong, Yilin Ma, and Xiaonan Huang. Design of a variable stiffness quasi-direct drive cable-actuated tensegrity robot. *IEEE Robotics and Automation Letters*, 2025.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

William A Nyberg, Pierre-Louis Bernard, Wayne Ngo, Charlotte H Wang, Jonathan Ark, Allison Rothrock, Gina M Borgo, Gabriella R Kimmerly, Jae Hyung Jung, Vincent Allain, et al. In vivo site-specific engineering to reprogram t cells. *Nature*, pp. 1–10, 2026.

Tom JN Obey, Mukesh K Singh, Angelos B Canaj, Gary S Nichol, Euan K Brechin, and Jason B Love. A delocalized mixed-valence dinuclear ytterbium complex that displays intervalence charge transfer. *Journal of the American Chemical Society*, 146(42):28658–28662, 2024.

OpenAI. GPT-5.4 model. https://developers.openai.com/api/docs/models/gpt-5.4, 2026a. Accessed September 25, 2026.

OpenAI. GPT-5.4 Mini model. https://developers.openai.com/api/docs/models/gpt-5.4-mini, 2026b. Accessed September 25, 2026.

OpenAI. GPT-5.5 model. https://developers.openai.com/api/docs/models/gpt-5.5, 2026c. Accessed September 25, 2026.

OpenAI. GPT-5.6 Luna model. https://developers.openai.com/api/docs/models/gpt-5.6-luna, 2026d. Accessed September 4, 2026.

OpenAI. GPT-5.6 Sol model. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026e. Accessed September 4, 2026.

OpenAI. GPT-5.6 Terra model. https://developers.openai.com/api/docs/models/gpt-5.6-terra, 2026f. Accessed September 4, 2026.

Mineto Ota, Jeffrey P Spence, Tony Zeng, Emma Dann, Nikhil Milind, Alexander Marson, and Jonathan K Pritchard. Causal modelling of gene effects from regulators to programs to traits. *Nature*, 650(8101):399–408, 2026.

Zhangzhi Peng. Ptm-mamba: a ptm-aware protein language model with bidirectional gated mamba blocks. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 5475–5478, 2024.

Tingyu Qu, Mingxiao Li, Tinne Tuytelaars, and Marie-Francine Moens. Ts-llava: Constructing visual tokens through thumbnail-and-sampling for training-free video large language models. *arXiv preprint arXiv:2411.11066*, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b, April 2026a. Model card: https://huggingface.co/Qwen/Qwen3.6-27B. Accessed September 25, 2026.

Qwen Team. Qwen3.7-Plus. https://qwen.ai/blog?id=qwen3.7-plus, June 2026b. Accessed September 4, 2026.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, August 2026c. Accessed September 4, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, August 2026d. Accessed September 4, 2026.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020. doi: 10.1109/SC41405.2020.00024. URL https://arxiv.org/abs/1910.02054.

Guangbin Ren and Yuchen Zhang. Weighted l^2 theory for the euclidean dirac operator in higher dimensions. *arXiv preprint arXiv:2604.04504*, 2026.

Palas Roy, Naba Kumar Mondal, Shreya Bhattacharya, Biswajit Das, and Kousik Das. Removal of arsenic(III) and arsenic(V) on chemically modified low-cost adsorbent: Batch and column operations. *Applied Water Science*, 3:293–309, 2013. doi: 10.1007/s13201-013-0082-5. URL https://link.springer.com/article/10.1007/s13201-013-0082-5.

Abdulai Salifu, Branislav Petrusevski, Emmanuel S Mwampashi, Iddi A Pazi, Kebreab Ghebremichael, Richard Buamah, Cyril Aubry, Gary L Amy, and Maria D Kenedy. Defluoridation of groundwater using aluminum-coated bauxite: optimization of synthesis process conditions and equilibrium study. *Journal of Environmental Management*, 181:108–117, 2016.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. *arXiv preprint arXiv:2402.09268*, 2024.

Tom Schlegel, Dennis Breu, and Michael Fleischhauer. Imaginary-time evolution of interacting spin systems in the truncated wigner approximation. *arXiv preprint arXiv:2603.03950*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Yanbin Shen, Qinkai Tang, and Yaozhi Luo. A tensegrity-inspired bidirectional quasi-zero stiffness metamaterial for buffering and energy absorption. In *Structures*, volume 79, pp. 109438. Elsevier, 2025.

Sofia Shkunnikova, Anika Mijakovac, Lucija Sironic, Maja Hanic, Gordan Lauc, and Marina Martinic Kavur. Igg glycans in health and disease: Prediction, intervention, prognosis, and therapy. *Biotechnology advances*, 67:108169, 2023.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023. doi: 10.1101/2023.10.01.560349. URL https://www.biorxiv.org/content/10.1101/2023.10.01.560349v1.

Yunfei Su, Guochun Wu, Lei Yao, and Yinghui Zhang. Large time behavior of weak solutions to the inhomogeneous incompressible navier-stokes-vlasov equations in r3. *Journal of Differential Equations*, 402:361–399, 2024.

Zhaochen Su, Jincheng Gao, Hangyu Guo, Zhenhua Liu, Lueyang Zhang, Xinyu Geng, Shijue Huang, Peng Xia, Guanyu Jiang, Cheng Wang, et al. Agentvista: Evaluating multimodal agents in ultra-challenging realistic visual scenarios. *arXiv preprint arXiv:2602.23166*, 2026.

Bambang Suryobroto. Estimation of the biological affinities of seven species of sulawesi macaques based on multivariate analysis of dermatoglyphic pattern types. *Primates*, 33(4):429–449, 1992.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Timothy F. Truong, Jr. and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77379–77415, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4366126eba252699b280e8f93c0ab2f-Paper-Conference.pdf.

Timothy Fei Truong Jr and Tristan Bepler. Understanding protein function with a multimodal retrieval-augmented foundation model. *arXiv preprint arXiv:2508.04724*, 2025.

Antonella Vallenari, Anthony GA Brown, Timo Prusti, Jos HJ De Bruijne, F Arenou, Carine Babusiaux, Michael Biermann, Orlagh L Creevey, Christine Ducourant, Dafydd Wyn Evans, et al. Gaia data release 3-summary of the content and survey properties. *Astronomy & Astrophysics*, 674:A1, 2023.

Hong-Yi Wang, Yu-Qi Li, Qian Wu, and Zhu-Fang Cui. Revisiting p-^{11}b fusion: Updated cross-sections, reactivity, and energy balance. *arXiv preprint arXiv:2601.00241*, 2026a.

Qian Wang, Joo Han Lee, Gregory Nachtrab, Yuan Yuan, Lei Yuan, Wei Qi, Manuel A Mohr, Jing Xiong, Mark A Horowitz, and Xiaoke Chen. Deconstruction of a spino-brain–spinal cord circuit that drives chronic pain. *Nature*, pp. 1–10, 2026b.

Tianyuan Wang and Mark A Post. A symmetric three degree of freedom tensegrity mechanism with dual operation modes for robot actuation. *Biomimetics*, 6(2):30, 2021.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. *arXiv preprint arXiv:2505.23683*, 2025.

S. Wardani, H. Husni, M. I. Sulaiman, H. Desvita, and A. Hadi. Simultaneous removal of heavy metals from aqueous solutions by pineapple crown and avocado peel hydrogel composites. *Global Journal of Environmental Science and Management*, 2026. doi: 10.22034/gjesm.2026.02.01. URL https://doi.org/10.22034/gjesm.2026.02.01. Online first.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Dillon Wong, Kevin P Nuckolls, Myungchul Oh, Biao Lian, Yonglong Xie, Sangjun Jeon, Kenji Watanabe, Takashi Taniguchi, B Andrei Bernevig, and Ali Yazdani. Cascade of electronic transitions in magic-angle twisted bilayer graphene. *Nature*, 582(7811):198–202, 2020.

Shenxi Wu, Haosong Zhang, Xingjian Ma, Shirui Bian, Yichi Zhang, Xi Chen, and Wei Lin. Hyperparameter transfer laws for non-recurrent multi-path neural networks. *arXiv preprint arXiv:2602.07494*, 2026.

Bo Xia, Weimin Zhang, Guisheng Zhao, Xinru Zhang, Jiangshan Bai, Ran Brosh, Aleksandra Wudzinska, Emily Huang, Hannah Ashe, Gwen Ellis, et al. On the genetic basis of tail-loss evolution in humans and apes. *Nature*, 626(8001):1042–1048, 2024a.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. DocGenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024b.

Xiaomi MiMo Team. Xiaomi MiMo-V2.5 series. https://mimo.mi.com/docs/en-US/news/latest/v2.5-open-sourced, 2026. Accessed September 4, 2026.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*, 2025a.

Long Xing, Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jinsong Li, Shuangrui Ding, Weiming Zhang, Nenghai Yu, et al. Scalecap: Inference-time scalable image captioning via dual-modality debiasing. *arXiv preprint arXiv:2506.19848*, 2025b.

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.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

Hao Yan, Yuliang Liu, Xingchen Liu, Yuyi Zhang, Minghui Liao, Jihao Wu, Wei Chen, and Xiang Bai. Docseeker: Structured visual reasoning with evidence grounding for long document understanding. *arXiv preprint arXiv:2604.12812*, 2026.

Z.ai. GLM-4.6V. https://docs.z.ai/guides/vlm/glm-4.6v, December 2025. Accessed September 4, 2026.

Beichen Zhang, Yuhong Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Haodong Duan, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Booststep: Boosting mathematical capability of large language models via improved single-step reasoning. *arXiv preprint arXiv:2501.03226*, 2025.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightWeight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517v4*, 2025. URL https://arxiv.org/abs/2408.05517v4.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, 2024. doi: 10.18653/v1/2024.acl-demos.38. URL https://aclanthology.org/2024.acl-demos.38/.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: Largest dataset ever for document layout analysis. In *International Conference on Document Analysis and Recognition*, 2019.

Zhuochao Zhou, Yuhong Liu, Xiaotong Gu, Haowen Zhang, Panan Zhang, Yue Sun, Honglei Liu, Xiaobing Cheng, Yutong Su, Hui Shi, et al. Glycosylation of anti-dsdna igg correlates with organ involvement in treatment-naïve patients with systemic lupus erythematosus. *Lupus Science & Medicine*, 12(2), 2025.

Table 4: Benchmark and training-data statistics. Training-data totals include validation. Four views of one SFT seed remain in the same split; SFT and RL share the matched-view questions.

Statistic	Value
Data Source	
Source channels	arXiv, bioRxiv, OpenReview, PubMed
Source papers	116
Benchmark Scale	
Underlying questions	124
Evaluation instances	496
Scientific domains	5
Task categories	7
Subtasks	19
Document & Question Characteristics	
Avg. images per paper	14.2
Avg. text tokens per instance	3.9K
Evaluation	
Evaluation methods	3
Training Release	
SFT instances: train / validation / total	3,844 / 80 / **3,924**
SFT seeds: train / validation / total	961 / 20 / 981
Views per SFT seed	2 languages $\times$ 2 document views
RL instances: train / validation / total	10,056 / 87 / **10,143**
RL matched-view / additional QA instances	3,924 / 6,219
Unique image assets (SFT and RL union)	55,603
Train–validation partition	Source-paper-disjoint across SFT and RL

A LIMITATIONS

Several limitations should be noted. First, the current benchmark contains 124 underlying questions. Although they are expert-authored and difficulty-screened, a larger question pool would strengthen statistical reliability and permit finer-grained capability analysis. The 496 evaluation instances are four matched variants of these questions and should not be interpreted as 496 independent scientific problems. Second, domain coverage emphasizes computer science, mathematics, and selected natural and biomedical sciences; social sciences, humanities, and several engineering fields remain underrepresented. Third, cross-document tasks currently involve two to four papers, while real literature reviews may span dozens of documents. Fourth, synthetic evidence enables precise control and verification but cannot reproduce every visual convention or failure mode found in naturally occurring papers. Finally, SCIDOCIR depends on layout parsing, source alignment, and OCR components. Errors in these upstream stages can propagate into both task generation and training-data quality.

B FULL SCIDOCBENCH TASK TAXONOMY AND DATA STATISTICS

Table 4 summarizes SCIDOCBENCH and SCIDOCDATASET . The source documents are drawn from three channels: arXiv (primary), academic journals (including *Nature*, *Journal of the American Chemical Society*, *Chemistry of Materials*, *Organic Letters*, *Biomimetics*, *ACS* family journals, *Global Journal of Environmental Science and Management* and so on), and web-crawled sources (university repositories and Google Scholar).

Release composition and counting. The SFT collection comprises 981 seeds verified against their sources and reference answers. Their four views are also used in RL alongside 6,219 additional verifiable QA instances (6,212 training; seven validation). Connected source papers remain in one split across both stages. Counts precede model-specific length filtering; augmented views and reused questions are not independent tasks.

Figure 4 shows user-prompt characters (including image placeholders, excluding assistant targets), not tokens, and per-instance image references. Appendix F details the training mixtures and experimental settings.

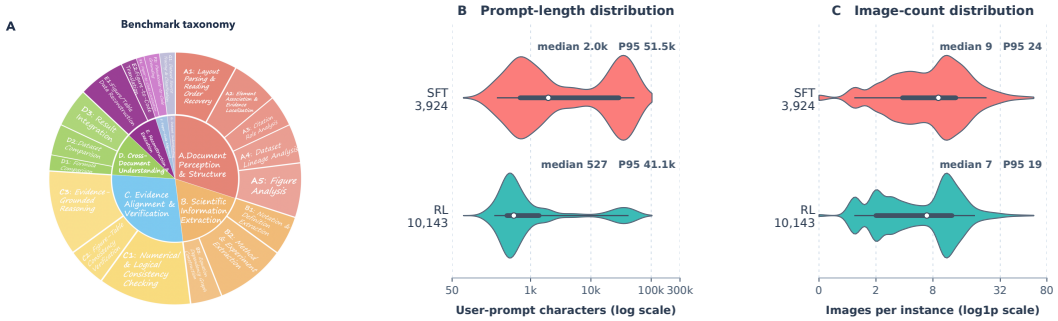


Figure 4: SciDocBench taxonomy and current training-release distributions. A: benchmark taxonomy. B: user-prompt character counts (log scale). C: image references per instance (log-plus-one scale). Panels B–C cover 3,924 SFT and 10,143 RL instances; dots mark medians and thick bars span the interquartile range.

Table 5: Full SciDocBench task taxonomy. The benchmark contains seven capability groups and nineteen subtasks.

ID	Task	Evaluation focus
A. Document Perception & Structure		
A1	Layout Parsing & Reading Order Recovery	Recover local reading flow in two-column PDFs with floating figures, footnotes, and cross-page continuation.
A2	Element Association & Evidence Localization	Given a claim, locate its supporting figure, table, equation, appendix, page, or block.
A3	Citation Role Analysis	Identify whether cited work acts as background, method basis, data resource, baseline, result support, critique, or extension.
A4	Dataset Lineage Analysis	Recover the source, composition, filtering, derivation, and transformation chain of datasets.
A5	Figure Analysis	Interpret scientific figures, subfigures, legends, axes, annotations, and visual evidence supporting figure-level claims.
B. Scientific Information Extraction		
B1	Notation & Definition Extraction	Build notation indices containing symbols, types, definitions, contexts, and first appearances.
B2	Method & Experiment Extraction	Extract datasets, metrics, baselines, instruments, configurations, hyperparameters, and experiment stages.
B3	Equation Dependency Graph Construction	Convert mathematical derivations into equation nodes and dependency edges.
C. Evidence Alignment & Verification		
C1	Numerical & Logical Consistency Checking	Verify whether claims match table values, figure values, rankings, percentages, and derived quantities.
C2	Figure–Table Consistency Verification	Detect whether plotted positions, labels, captions, and tables express the same scientific fact.
C3	Evidence-Grounded Reasoning	Answer questions using document evidence while avoiding unsupported free-form reasoning.
D. Cross-Document Understanding		
D1	Cross-Document Formula Comparison	Compare formula variants, assumptions, and generalization paths across papers.
D2	Cross-Document Dataset Comparison	Identify dataset intersections, source overlap, or near-duplicate data usage across papers.
D3	Cross-Document Result Integration	Merge results from multiple papers into a canonical schema with provenance.
E. Reconstruction & Execution		
E1	Figure/Table Data Reconstruction	Recover data points, structured tables, or plotting code from figures and tables.
E2	Figure-to-Code Translation	Convert scientific figures, diagrams, architectures, or flowcharts into runnable or structured code.
F. Paper–Code Alignment		
F1	Paper-to-Code Localization	Locate modules, scripts, or functions in a repository corresponding to paper descriptions.
F2	Pseudocode-to-Code Grounding	Complete implementation details by aligning pseudocode, derivations, and source code.
G. Dataset Understanding		
G1	Dataset Source Tracing & Classification	Infer dataset sources, families, and provenance classes from fields, structures, samples, and source patterns.

Table 6: Closed-rule scoring beyond recursive JSON comparison. These 19 questions and the 38 recursive-JSON questions comprise the 57 R assignments.

Rule	Questions	Credit and failure conditions
Reading-order ranking	5	Fraction of correctly placed labels; incorrect sequence length scores zero.
Binary decisions	6	Fraction of correct bits in a final bitstring of the required length.
Closed scalar	1	Normalized label match, with the task's answer-field/label extraction.
Grouped choices	1	Four groups weighted 0.30, 0.10, 0.30, 0.30; disallowed choices invalidate the affected group.
Unordered exact records	1	Fraction of reference dictionaries recovered exactly, ignoring key order; one-to-one matching prevents duplicate credit.
Indexed exact records	1	Exact row count and keys are prerequisites; schema credit plus per-row exact-match credit.
Source-tracing records	1	Correct-record credit minus incorrect-record penalties; wrong fields invalidate the response.
Subfigure and choices	1	Wrong subfigure scores zero; otherwise 0.50 plus up to 0.50 for valid choices.
Keyed lists	2	Valid dictionary-of-lists schema plus weighted structured similarity for each required key.

B.1 Evaluation Protocol and Per-Question Assignment

Routing and final-answer scope. The released scorer assigns 57 underlying questions to deterministic rules (R), 66 to a semantic LLM judge (J), and one to execution (X), corresponding to 228, 264, and four evaluation instances. Assignment is fixed by question ID and shared by all four views. The audited ID-to-rule mapping takes precedence over the `eval_method` metadata field: a JSON output requirement alone does not imply exact matching. Unlisted questions use the semantic judge unless explicitly designated for execution. The source-document tables below give R/J/X for every question; the accompanying `question_evaluators.csv` maps each printed ID to its dataset question ID and exact scorer.

Only the final response is evaluated. Delimited internal reasoning is removed; an unfinished reasoning segment, absent final answer, or unusable response receives zero. Explanations requested as part of the final deliverable are still assessed. All evaluator outputs lie in $[0, 1]$; no failed instance is dropped from aggregation.

Deterministic rules (R). Rules are restricted to questions with closed targets and checkable requirements. Thirty-eight questions compare parsed JSON recursively. Scalars are normalized for benign case, spacing, numeric notation, and location/equation aliases. A dictionary is compared by its specified reference fields, assigning zero to missing fields and normalizing by the larger field count. Lists are compared positionally only when the task requires order; otherwise entries are paired one-to-one by their structured similarity. The list score is normalized by the larger list length, so extra entries are not free. Small unordered lists use maximum-weight matching; lists with more than 12 predicted entries use greedy matching. Parsing failures and incompatible types receive zero. These comparisons concern fixed values or labels, not semantic equivalence of arbitrary prose or LaTeX.

The other 19 rule-scored questions use closed rubrics summarized in Table 6. All constants come from the question's scoring contract rather than being inferred from response style. For example, the eight-row indexed-table task awards 0.04 for a valid schema and 0.12 per exact row; the source-tracing task adds 0.33 per correct record and subtracts 0.20 per incorrect record, clipping the result to $[0, 1]$. The two keyed-list tasks use schema credits of 0.10 and 0.09 and per-key similarity weights of 0.10 and 0.07, respectively.

Semantic judgment (J). GPT-5.4-mini receives the question, reference answer, candidate final answer, and task-specific rubric, with temperature zero. It returns a bounded score and a brief assessment. The rubric specifies required content, component weights, prerequisite conditions, and deductions where applicable; the default rubric allocates partial credit by required factual content and penalizes contradictions. Semantically equivalent wording is accepted. JSON key order, surrounding code fences, and harmless whitespace are ignored unless explicitly material to the task.

Table 7: Source documents and evaluators for Categories A1 to A2.

Subtask	ID	Subject	Source Document(s)	Eval.
A1	1	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025)	J
	7	Nuclear Theory	*Revisiting p-^{11}B Fusion: Updated Cross-sections, Reactivity, and Energy Balance* Wang et al. (2026a)	R
	13	Computer Science	*Depth Anything 3: Recovering the Visual Space from Any Views* Lin et al. (2026a)	R
	51	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025)	R
	52	Computer Science	*Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	R
	53	Medicine	*QuarkMedSearch: A Long-Horizon Deep Search Agent for Exploring Medical Intelligence* Lin et al. (2026b)	R
	54	Computer Science	*Hyperparameter Transfer Laws for Non-Recurrent Multi-Path Neural Networks* Wu et al. (2026)	R
	55	Computer Science	*DocSeeker: Structured Visual Reasoning with Evidence Grounding for Long Document Understanding* Yan et al. (2026)	R
	115	Computer Science	*Depth Anything 3: Recovering the Visual Space from Any Views* Lin et al. (2026a)	J
A2	2	Computer Science	*LLaVA-CoT: Let Vision Language Models Reason Step-by-Step* Xu et al. (2025)	J
	8	Environment	*Removal of arsenic (III) and arsenic (V) on chemically modified low-cost adsorbent: batch and column operations* Roy et al. (2013)	R
	14	Computer Science	*Depth Anything 3: Recovering the Visual Space from Any Views* Lin et al. (2026a)	J
	56	Computer Science	*Booststep: Boosting mathematical capability of large language models via improved single-step reasoning* Zhang et al. (2025)	J
	76	Physics	*Cascade of electronic transitions in magic-angle twisted bilayer graphene* Wong et al. (2020)	J
	86	Biology	*Inferring the internal structure of groups through the integration of statistical learning and causal reasoning* Davis et al. (2026)	R
	106	Psychology	*Inferring the internal structure of groups through the integration of statistical learning and causal reasoning* Davis et al. (2026)	R

Required fields, evidence relations, units, ordering, and requested explanations remain assessable. Thus open-ended definitions, explanations, and semantically equivalent code or LaTeX are not automatically assigned a byte-exact rule. The judge compares against the vetted reference rather than independently re-reading the full source documents. Responses are cached by judge configuration and grading inputs to reuse identical judgments.

Execution (X). The sole execution task, `constraintfollow_001`, asks for an image-transformation program. The evaluator extracts Python code and runs it and the reference script on the same input image, with a 30-second timeout per script. Missing output, execution failure, or incorrect dimensions scores zero. Equal output pixels score 1; otherwise mean squared pixel error at most 10 scores 0.5, and larger error scores zero. Pixel comparison uses the reference color mode. Other code-like outputs are not implicitly executed: they retain their listed R or J assignment. Generated programs should be evaluated in an isolated environment.

B.2 Question Sources and Evaluators

In the following tables, R denotes deterministic rules, J semantic judgment, and X execution. IDs identify questions, not papers; several questions can share one source. Their dataset question IDs are supplied in the companion evaluation manifest.

Table 8: Source documents and evaluators for Categories A3 to A4.

Subtask	ID	Subject	Source Document(s)	Eval.
A3	3	Computer Science	*Proximal Policy Optimization Algorithms* Schulman et al. (2017)	R
	9	Biology	*ProtFlow: Flow Matching-based Protein Sequence Design with Comprehensive Protein Semantic Distribution Learning and High-quality Generation* Kong et al. (2026)	R
	15	Computer Science	*Spa3R: Predictive Spatial Field Modeling for 3D Visual Reasoning* Jiang et al. (2026)	R
	77	Physics	*Quantum computing: A taxonomy, systematic review and future directions* Gill et al. (2022)	R
	101	Physics	*Quantum computing: A taxonomy, systematic review and future directions* Gill et al. (2022)	J
	103	Computer Science	*Quantum computing: A taxonomy, systematic review and future directions* Gill et al. (2022)	R
	114	Biology	*ProtFlow: Flow Matching-based Protein Sequence Design with Comprehensive Protein Semantic Distribution Learning and High-quality Generation* Kong et al. (2026)	J
	119	Computer Science	*Spa3R: Predictive Spatial Field Modeling for 3D Visual Reasoning* Jiang et al. (2026)	J
	121	Computer Science	*Proximal Policy Optimization Algorithms* Schulman et al. (2017)	J
A4	10	Biology	*SaProt: Protein Language Modeling with Structure-Aware Vocabulary* Su et al. (2023)	R
	16	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025)	R
	17	Computer Science	*ChartAssistant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning* Meng et al. (2024)	R
	57	Computer Science	*ACE-Brain-0: Spatial Intelligence as a Shared Scaffold for Universal Embodiments* Gong et al. (2026)	R
	87	Biology	*Compressing the collective knowledge of ESM into a single protein language model* Dinh et al. (2026), *ProtFlow: Flow Matching-based Protein Sequence Design with Comprehensive Protein Semantic Distribution Learning and High-quality Generation* Kong et al. (2026), *PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks* Peng (2024)	R
	107	Biology	*Compressing the collective knowledge of ESM into a single protein language model* Dinh et al. (2026), *ProtFlow: Flow Matching-based Protein Sequence Design with Comprehensive Protein Semantic Distribution Learning and High-quality Generation* Kong et al. (2026)	R
	117	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025)	J

Table 9: Source documents and evaluators for Categories A5 to B1.

Subtask	ID	Subject	Source Document(s)	Eval.
A5	58	Biology	*Estimation of the biological affinities of seven species of Sulawesi macaques based on multivariate analysis of dermatoglyphic pattern types* Suryobroto (1992)	J
	78	Chemistry	*A Delocalized Mixed-Valence Dinuclear Ytterbium Complex That Displays Intervalence Charge Transfer* Obey et al. (2024)	J
	79	Computer Science	*Biologically Plausible Learning via Bidirectional Spike-Based Distillation* Lv et al. (2025)	R
	80	Physics	*Design of a variable stiffness quasi-direct drive cable-actuated tensegrity robot* Mi et al. (2025)	R
	88	Biology	*Deconstruction of a spino-brain–spinal cord circuit that drives chronic pain* Wang et al. (2026b)	J
	89	Physics	*New limits on the Pauli forbidden transitions in 12C nuclei obtained with the complete Borexino dataset* Basilico et al. (2026)	R
	90	Physics	*Statics of integrated origami and tensegrity systems* Ma et al. (2023)	R
	104	Computer Science	*Biologically Plausible Learning via Bidirectional Spike-Based Distillation* Lv et al. (2025)	J
	108	Physics	*New limits on the Pauli forbidden transitions in ^{12}C nuclei obtained with the complete Borexino dataset* Basilico et al. (2026)	J
B1	4	Computer Science	*Group Sequence Policy Optimization* Zheng et al. (2025)	R
	11	Physics	*Coherent spectroscopy with a single antiproton spin* Latacz et al. (2025)	J
	12	Maths	*Large time behavior of weak solutions to the inhomogeneous incompressible Navier-Stokes-Vlasov equations in R3* Su et al. (2024)	J
	18	Computer Science	*Hyperparameter Transfer Laws for Non-Recurrent Multi-Path Neural Networks* Wu et al. (2026)	R
	81	Computer Science	*Packed-ensembles for efficient uncertainty estimation* Laurent et al. (2022)	J
	102	Computer Science	*Packed-ensembles for efficient uncertainty estimation* Laurent et al. (2022)	J
	116	Computer Science	*Hyperparameter Transfer Laws for Non-Recurrent Multi-Path Neural Networks* Wu et al. (2026)	J
	118	Computer Science	*ChartAssistant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning* Meng et al. (2024)	J

Table 10: Source documents and evaluators for Category B2.

Subtask	ID	Subject	Source Document(s)	Eval.
B2	5	Computer Science	*Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	R
	19	Medicine	*Glycosylation of anti-dsDNA IgG correlates with organ involvement in treatment-naive patients with systemic lupus erythematosus* Zhou et al. (2025)	R
	20	Biology	*LLM-assisted systematic review of large language models in clinical medicine* Chen et al. (2026)	J
	21	Biology	*In vivo site-specific engineering to reprogram T cells* Nyberg et al. (2026)	J
	59	Maths	*Quantum linear system solver based on time-optimal adiabatic quantum computing and quantum approximate optimization algorithm* An & Lin (2022)	J
	69	Chemistry	*Axial chirality-induced rigidification in aminoboranes enhances persistent room-temperature phosphorescence and circularly polarized luminescence* Eyyathiyil et al. (2025)	J
	82	Chemistry	*A Delocalized Mixed-Valence Dinuclear Ytterbium Complex That Displays Intervalence Charge Transfer* Obey et al. (2024)	J
	91	Physics	*A symmetric three degree of freedom tensegrity mechanism with dual operation modes for robot actuation* Wang & Post (2021)	R
	92	Physics	*Impact-resistant, autonomous robots inspired by tensegrity architecture* Johnson III et al. (2025)	R
	113	Computer Science	*Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	J

Table 11: Source documents and evaluators for Categories B3 to C2.

Subtask	ID	Subject	Source Document(s)	Eval.
B3	6	Computer Science	*Proximal Policy Optimization Algorithms* Schulman et al. (2017)	J
	22	Physics	*Observation of non-Abelian topological acoustic semimetals and their phase transitions* Jiang et al. (2021)	J
	23	Biology	*Poet: A generative model of protein families as sequences-of-sequences* Truong & Bepler (2023)	J
	26	Maths	*Weighted L^2 theory for the Euclidean Dirac operator in higher dimensions* Ren & Zhang (2026)	J
	70	Maths	*Weighted L^2 theory for the Euclidean Dirac operator in higher dimensions* Ren & Zhang (2026)	J
C1	24	Computer Science	*Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	J
	27	Maths	*Learning compositional functions with transformers from easy-to-hard data* Wang et al. (2025)	J
	28	Maths	*Dynamic metastability in the self-attention model* Geshkovski et al. (2024)	J
	60	Environment	*Defluoridation of groundwater using aluminum-coated bauxite: optimization of synthesis process conditions and equilibrium study* Salifu et al. (2016)	J
	61	Computer Science	*Exact recovery in the stochastic block model* Abbe et al. (2015)	R
	62	Maths	*Gaussian differential privacy* Dong et al. (2022)	R
	71	Computer Science	*Spa3R: Predictive Spatial Field Modeling for 3D Visual Reasoning* Jiang et al. (2026)	J
	72	Computer Science	*A unified perspective on the dynamics of deep transformers* Castin et al. (2025)	R
	73	Computer Science	*Transformers, parallel computation, and logarithmic depth* Sanford et al. (2024)	R
	74	Maths	*Benign overfitting in linear regression* Bartlett et al. (2020)	R
	83	Physics	*Magic-angle graphene superlattices: a new platform for unconventional superconductivity* Cao et al. (2018)	J
	84	Computer Science	*Provable failure of language models in learning majority boolean logic via gradient descent* Chen et al. (2025b)	R
C2	25	Computer Science	*Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	J
	29	Physics	*Gaia data release 3-summary of the content and survey properties* Vallenari et al. (2023)	J
	63	Computer Science	*Scalecap: Inference-time scalable image captioning via dual-modality debiasing* Xing et al. (2025b)	J
	64	Biology	*Neural sequences underlying directed turning in Caenorhabditis elegans* Kramer et al. (2026)	J
	93	Computer Science	*Improved baselines with visual instruction tuning* Liu et al. (2024)	R

Table 12: Source documents and evaluators for Categories C3 to D1.

Subtask	ID	Subject	Source Document(s)	Eval.
C3	30	Medicine	*Glycosylation of anti-dsDNA IgG correlates with organ involvement in treatment-naive patients with systemic lupus erythematosus* Zhou et al. (2025)	R
	31	Environment	*Simultaneous removal of heavy metals from aqueous solutions by pineapple crown and avocado peel hydrogel composites* Wardani et al. (2026)	J
	32	Maths	*The Euler Stratification for* $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^n$ Hoşten et al. (2026)	J
	33	Maths	*Computation and sampling for Schubert specializations* Anderson et al. (2026)	R
	34	Maths	*Learning compositional functions with transformers from easy-to-hard data* Wang et al. (2025)	J
	66	Biology	*Neural sequences underlying directed turning in Caenorhabditis elegans* Kramer et al. (2026)	R
	75	Biology	*Deconstruction of a spino-brain–spinal cord circuit that drives chronic pain* Wang et al. (2026b)	J
	85	Chemistry	*Giant coercivity and high magnetic blocking temperatures for N2*$^{3-}$ *radical-bridged dilanthanide complexes upon ligand dissociation* Demir et al. (2017)	J
	94	Physics	*A tensegrity-inspired bidirectional quasi-zero stiffness metamaterial for buffering and energy absorption* Shen et al. (2025)	J
	105	Chemistry	*Giant coercivity and high magnetic blocking temperatures for* N_2^{3-} *radical-bridged dilanthanide complexes upon ligand dissociation* Demir et al. (2017)	J
D1	35	Computer Science	*Proximal Policy Optimization Algorithms* Schulman et al. (2017), *Deepseekmath: Pushing the limits of mathematical reasoning in open language models* Shao et al. (2024), *Group Sequence Policy Optimization* Zheng et al. (2025), *Minimax-m1: Scaling test-time compute efficiently with lightning attention* Chen et al. (2025a)	R
	65	Physics	*Approximating quantum many-body wave functions using artificial neural networks* Cai & Liu (2018), *Quantum entanglement in neural network states* Deng et al. (2017), *Solving the quantum many-body problem with artificial neural networks* Carleo & Troyer (2017)	J
	67	Physics	*Approximating quantum many-body wave functions using artificial neural networks* Cai & Liu (2018), *Quantum entanglement in neural network states* Deng et al. (2017), *Solving the quantum many-body problem with artificial neural networks* Carleo & Troyer (2017)	J

Table 13: Source documents and evaluators for Category D2.

Subtask	ID	Subject	Source Document(s)	Eval.
D2	36	Computer Science	*Eagle 2.5: Boosting long-context post-training for frontier vision-language models* Chen et al. (2025c), *Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale* Guo et al. (2025)	J
	37	Computer Science	*Cambrian-1: A fully open, vision-centric exploration of multimodal llms* Tong et al. (2024), *Llava-onevision: Easy visual task transfer* Li et al. (2024a), *Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale* Guo et al. (2025)	R
	95	Chemistry	*Strategies for pre-training graph neural networks* Hu et al. (2019), *Geometry-enhanced molecular representation learning for property prediction* Fang et al. (2022), *Molecular contrastive learning of representations via graph neural networks* Wang et al. (2022)	R
	96	Biology	*ProtFlow: Flow Matching-based Protein Sequence Design with Comprehensive Protein Semantic Distribution Learning and High-quality Generation* Kong et al. (2026), *PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks* Peng (2024), *Compressing the collective knowledge of ESM into a single protein language model* Dinh et al. (2026)	R
	109	Chemistry	*Molecular Contrastive Learning of Representations via Graph Neural Networks* Wang et al. (2022), *ChemRL-GEM: Geometry Enhanced Molecular Representation Learning for Property Prediction* Fang et al. (2022), *Strategies for Pre-Training Graph Neural Networks* Hu et al. (2019)	J
	120	Computer Science	*Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs* Tong et al. (2024), *LLaVA-OneVision: Easy Visual Task Transfer* Li et al. (2024a), *MAmmoTH-VL: Eliciting Multimodal Reasoning with Instruction Tuning at Scale* Guo et al. (2025)	R

Table 14: Source documents and evaluators for Category D3.

Subtask	ID	Subject	Source Document(s)	Eval.
D3	38	Computer Science	*SpatialDreamer: Incentivizing Spatial Reasoning via Active Mental Imagery* Cao et al. (2025), *Spatialladder: Progressive training for spatial reasoning in vision-language models* Li et al. (2025)	J
	39	Computer Science	*Ts-llava: Constructing visual tokens through thumbnail-and-sampling for training-free video large language models* Qu et al. (2024), *Pllava: Parameter-free llava extension from images to videos for video dense captioning* Xu et al. (2024)	J
	47	Medicine	*Glycosylation of anti-dsDNA IgG correlates with organ involvement in treatment-naive patients with systemic lupus erythematosus* Zhou et al. (2025), *IgG glycans in health and disease: Prediction, intervention, prognosis, and therapy* Shkunnikova et al. (2023)	R
	97	Computer Science	*Caprl: Stimulating dense image caption capabilities via reinforcement learning* Xing et al. (2025a), *Scalecap: Inference-time scalable image captioning via dual-modality debiasing* Xing et al. (2025b)	R
	98	Chemistry	*High voltage cycling stability of LiF-coated NMC811 electrode* Llanos et al. (2024), *Long-term cyclability of NCM-811 at high voltages in lithium-ion batteries: an in-depth diagnostic study* Li et al. (2020), *Enhanced Cycling Stability of NCM811 Cathodes at High C-Rates and Voltages via LiMTFSI-Based Polymer Coating* Kim et al. (2025)	R
	110	Computer Science	*ScaleCap: Inference-Time Scalable Image Captioning via Dual-Modality Debiasing* Xing et al. (2025b), *CapRL: Stimulating Dense Image Caption Capabilities via Reinforcement Learning* Xing et al. (2025a)	R
	111	Chemistry	*Long-Term Cyclability of NCM-811 at High Voltages in Lithium-Ion Batteries: an In-Depth Diagnostic Study* Li et al. (2020), *Enhanced Cycling Stability of NCM811 Cathodes at High C-Rates and Voltages via LiMTFSI-Based Polymer Coating* Kim et al. (2025), *High Voltage Cycling Stability of LiF-Coated NMC811 Electrode* Llanos et al. (2024)	J
	123	Medicine	*IgG glycans in health and disease: Prediction, intervention, prognosis, and therapy* Shkunnikova et al. (2023), *Fucosylation of anti-dsDNA IgG1 correlates with disease activity of treatment-naïve systemic lupus erythematosus patients* Han et al. (2022)	J

Table 15: Source documents and evaluators for Categories E, F, and G.

Subtask	ID	Subject	Source Document(s)	Eval.
E1	40	Computer Science	*Group Sequence Policy Optimization* Zheng et al. (2025)	R
	41	Computer Science	*LLaVA-CoT: Let Vision Language Models Reason Step-by-Step* Xu et al. (2025)	J
	48	Physics	*Stairway Codes: Floquetifying Bivariate Bicycle Codes and Beyond* Jacoby et al. (2026)	J
	49	Physics	*Imaginary-time evolution of interacting spin systems in the truncated Wigner approximation* Schlegel et al. (2026)	R
	68	Biology	*Neural sequences underlying directed turning in Caenorhabditis elegans* Kramer et al. (2026)	J
	99	Biology	*Causal modelling of gene effects from regulators to programs to traits* Ota et al. (2026)	R
	112	Biology	*Causal modelling of gene effects from regulators to programs to traits* Ota et al. (2026)	J
	124	Physics	*Imaginary-time evolution of interacting spin systems in the truncated Wigner approximation* Schlegel et al. (2026)	J
E2	42	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025)	X
	50	Physics	*QFlowNet: Fast, Diverse, and Efficient Unitary Synthesis with Generative Flow Networks* Koo et al. (2026)	J
F1	43	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025)	R
	122	Computer Science	*TwiFF (Think With Future Frames): A Large-Scale Dataset for Dynamic Visual Reasoning* Liu et al. (2026), *AgentVista: Evaluating Multimodal Agents in Ultra-Challenging Realistic Visual Scenarios* Su et al. (2026), *MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025), *Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	R
F2	44	Biology	*On the genetic basis of tail-loss evolution in humans and apes* Xia et al. (2024a)	J
	100	Biology	*On the genetic basis of tail-loss evolution in humans and apes* Xia et al. (2024a)	R
G1	45	Computer Science	*MM-IFEngine: Towards Multimodal Instruction Following* Ding et al. (2025), *Agentvista: Evaluating multimodal agents in ultra-challenging realistic visual scenarios* Su et al. (2026), *TwiFF (Think With Future Frames): A Large-Scale Dataset for Dynamic Visual Reasoning* Liu et al. (2026), *Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning* Liu et al. (2025a)	R
	46	Biology	*Understanding protein function with a multimodal retrieval-augmented foundation model* Truong Jr & Bepler (2025), *Progen2: exploring the boundaries of protein language models* Nijkamp et al. (2023)	J

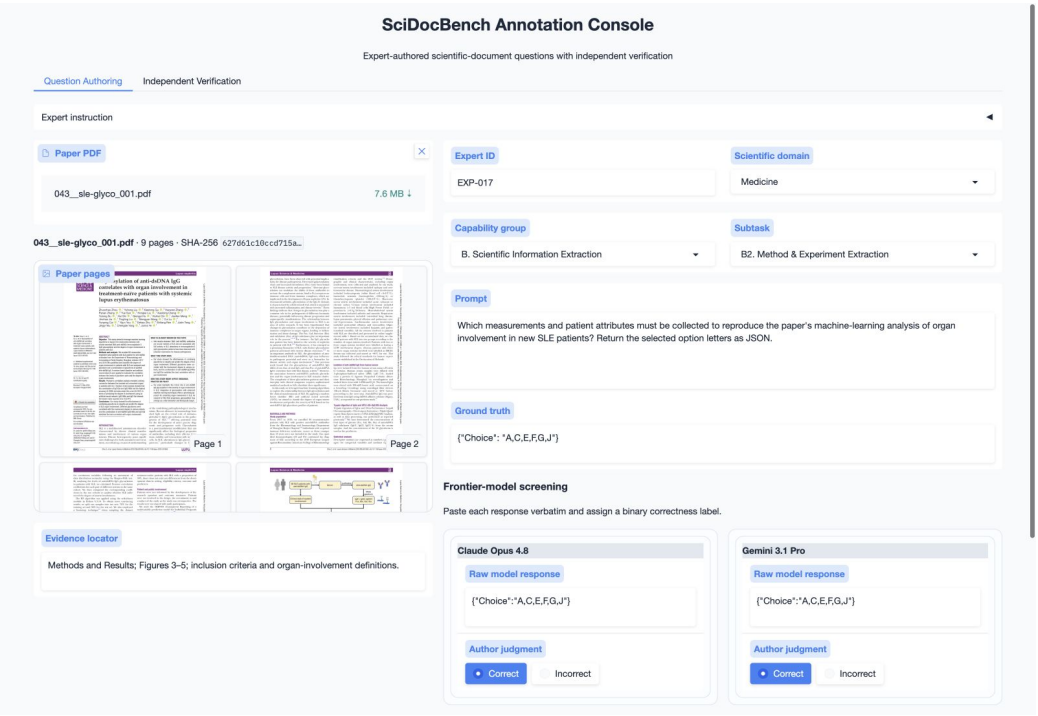


Figure 5: Expert question-authoring interface. The interface presents the source PDF and rendered pages alongside the task metadata, question, ground truth, evidence locator, and verbatim screening-model responses with author-assigned correctness labels. The populated fields constitute a representative interface record used to illustrate the annotation procedure and are not additional leaderboard results.

C Benchmark Annotation and Verification Protocol

Benchmark construction uses two separate interfaces and two annotator roles. The authoring interface records the source PDF, expert identifier, scientific domain, capability group, subtask, finalized question, ground-truth answer, and evidence locator. The evidence locator identifies the page, section, figure, table, equation, appendix, or other document object from which the answer can be verified. Authors then run the finalized item once on each of the four designated screening models, paste the responses without editing, and assign a binary correctness label to every response. An item can be submitted only when all required fields and four screening records are present and at least one screened model is incorrect. If the question, ground truth, or evidence locator changes, the four responses are collected again. Figure 5 shows the authoring interface.

Instructions for Question Authors. Read the supplied paper and create one scientifically grounded question that is answerable from the paper but challenging for current multimodal models.

1. Upload the source PDF; provide the finalized prompt and ground-truth answer; and identify the page, section, figure, table, equation, appendix, or other scientific object that supports the answer.
2. Write a clear and self-contained prompt with an unambiguous output format and a stable answer that can be verified directly from the supplied document. Do not rely on unstated external knowledge.
3. Run the finalized question once on each designated screening model. Paste every response verbatim and label it **Correct** or **Incorrect** according to the ground truth. Do not edit, merge, or selectively rerun individual responses.
4. Submit the item only when every required field and all four screening records are complete and at least one valid screening response is incorrect. If the prompt, ground truth, or evidence locator changes, collect all four responses again.

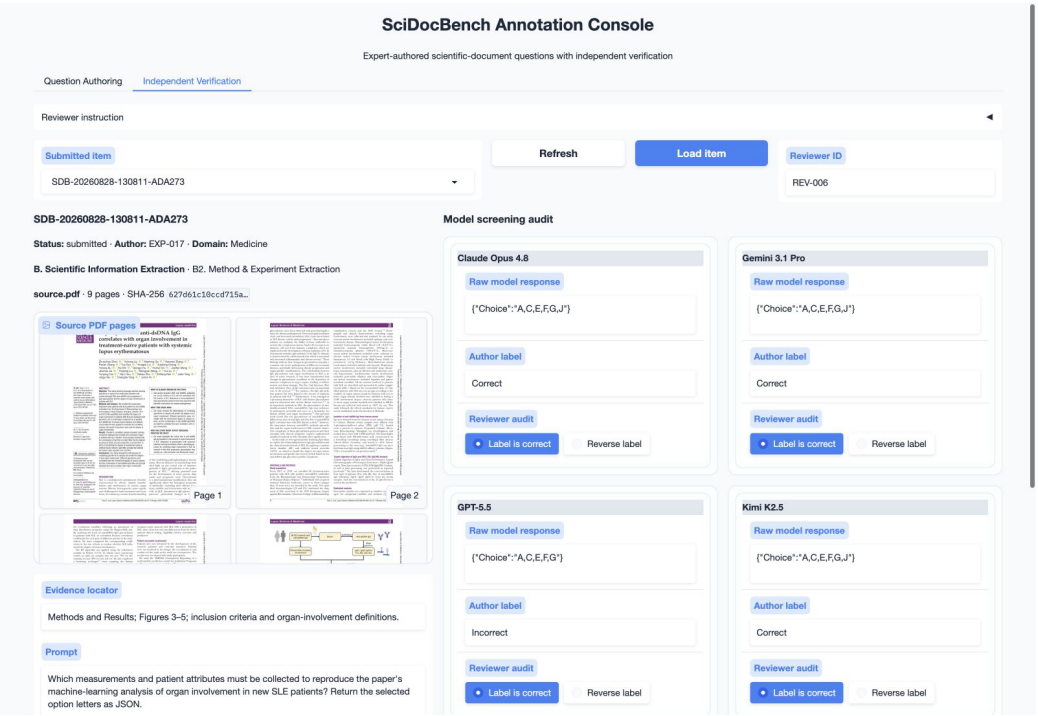


Figure 6: Independent-verification interface. Reviewers inspect the same PDF, task metadata, evidence locator, question, ground truth, and four screening records; audit each correctness label; complete the acceptance checklist; and record an Accept, Revise, or Reject decision.

Every submitted item is subsequently inspected by a reviewer who did not author the question. The reviewer loads the immutable submission and verifies four conditions: the question is answerable from the supplied PDF, the wording and requested output format are unambiguous, the ground truth is correct, and at least one screening model is incorrect. The reviewer also audits every response label and can reverse an author label when it does not agree with the ground truth. An item is accepted only when all four checks pass. Items with correctable defects are returned for revision, while items with unstable evidence or irreparable ambiguity are rejected. Figure 6 shows the independent-verification interface.

The annotation system stores each submission as an atomic record containing a schema version, item identifier, author and reviewer identifiers, timestamps, task metadata, the question and ground truth, the evidence locator, the source-PDF filename and SHA-256 digest, the four verbatim responses, the author labels, reviewer label audits, and the final decision. The PDF digest prevents accidental substitution of the source document after screening. Screening responses are used only to remove items already solved by all designated models; they are not used to construct the ground truth. This separation preserves expert evidence as the basis of correctness while retaining an auditable record of question difficulty and reviewer agreement.

C.1 LLM-AS-A-JUDGE RELIABILITY AUDIT

We assess the reliability of semantic scoring on 100 response-level instances drawn from accepted LLM-as-a-judge evaluation runs. The audit set contains 25 instances from each of the four matched evaluation settings, covers 11 evaluated models, and includes questions from all seven capability groups. Sampling is approximately balanced across the judge's initial incorrect, partial, and correct score bands. Deterministic rule-based and execution-based evaluations are excluded because their correctness is established by programmatic scorers rather than semantic judgment.

The human expert was shown only the question, reference answer, and anonymized model response. Model identity, the GPT-5.4-mini score, and the judge rationale were hidden until all 100 ratings had been submitted. The expert assigned one of three scores: 0 for an incorrect response, 0.5 for a partially correct response, and 1 for a fully correct response. For the ordinal comparison, continuous

Table 16: Pairwise agreement on the 100-instance semantic-judge audit. Binary results threshold the original scores at 0.5. QWK denotes quadratic-weighted Cohen's κ, and MAE denotes mean absolute score difference.

Rater pair	Three-level exact	QWK	Binary exact	Binary κ	MAE
Human expert vs. GPT-5.4-mini	92.0%	0.866	70.0%	0.355	0.169
Codex vs. GPT-5.4-mini	89.0%	0.818	80.0%	0.593	0.128
Human expert vs. Codex	93.0%	0.872	74.0%	0.405	0.145

GPT-5.4-mini and Codex scores are mapped using the same strict endpoint rule: 0 is incorrect, 1 is correct, and every value strictly between 0 and 1 is partially correct. We report quadratic-weighted Cohen's κ (Cohen, 1968) alongside exact agreement. The audit therefore measures consistency in applying the reference-based scoring rubric; it does not constitute an independent re-annotation of each source document.

The average pairwise three-level exact agreement is 91.3%. To summarize all three rating sources jointly, we compute ordinal Krippendorff's α (Hayes & Krippendorff, 2007) over the expert, GPT-5.4-mini, and Codex labels. The resulting α is 0.854, with a 95% confidence interval of $[0.757, 0.924]$ obtained from 10,000 bootstrap resamples of the audit instances. The stronger ordinal agreement than binary agreement reflects the intended use of partial credit: thresholding at 0.5 converts small differences within the intermediate score range into categorical disagreements.

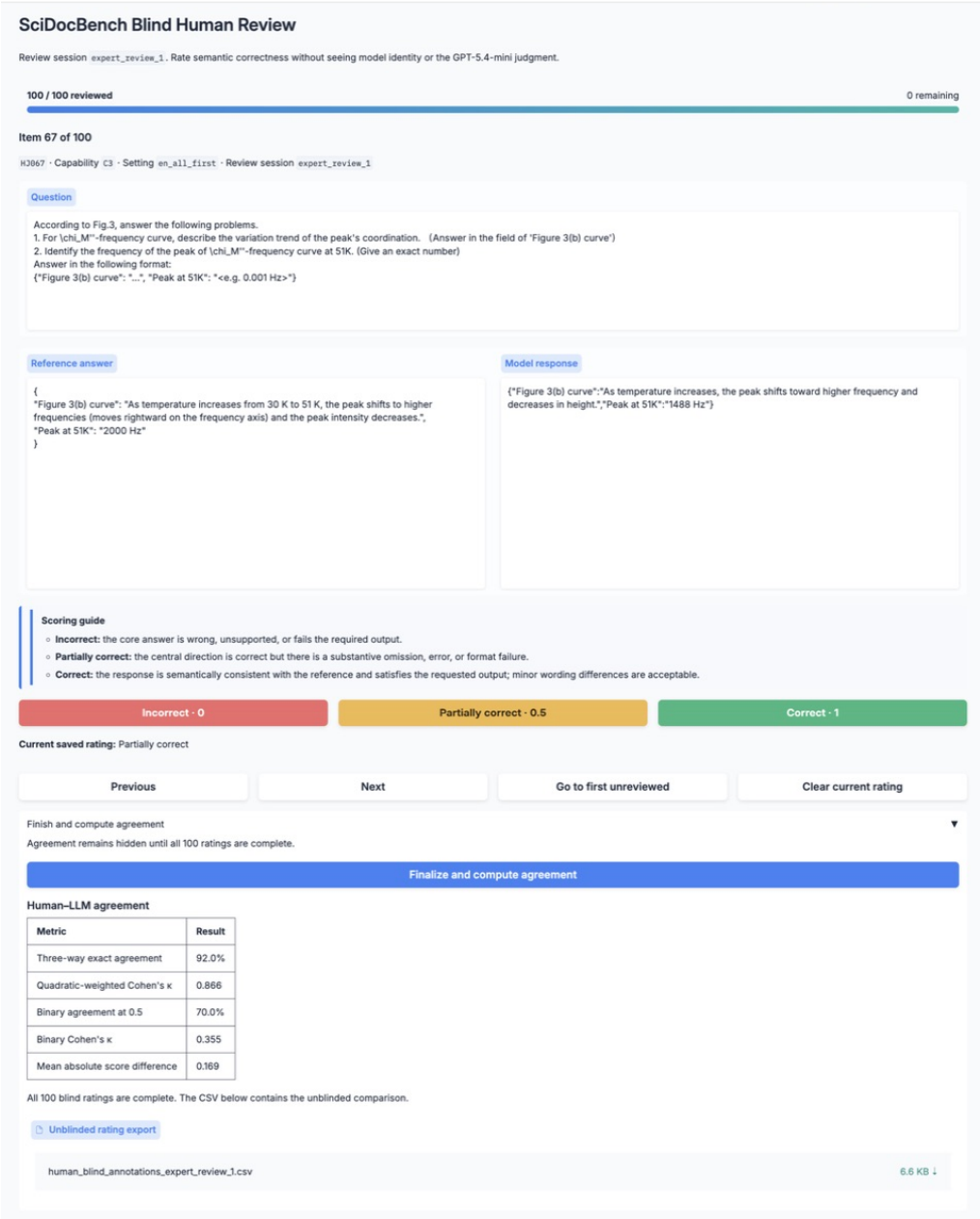


Figure 7: Complete blind expert-review interface for the English audit item HJ067. The interface presents the item metadata, question, reference answer, anonymized model response, three-level rubric, rating controls, navigation controls, and the agreement summary revealed after all 100 ratings were completed. Model identity and the GPT-5.4-mini judgment remain hidden during rating. Browser chrome and unrelated tabs are omitted from the capture.

```
Record SciDocIR
    document  : {doc_id, source_dir, num_pages, provenance}
    categories: List<{id, name, supercategory}>
    pages     : List<Page>
    blocks    : List<Block>
    relations : List<Relation>
    stats     : {num_blocks, num_relations, category_counts}
End Record

Record Page
    page_index : Integer
    width, height : Float
    image_path : String
End Record

Record Block
    block_id, page_index : Integer
    bbox       : {x1, y1, x2, y2, width, height}
    category   : {id, name, supercategory}
    content    : {source_code, format, plain_text, latex}
    structure  : {previous_block, parent_block, next_block}
    references : {labels, outgoing_ref_labels, incoming_ref_from}
    metadata   : Metadata
End Record

Record Metadata
    title_level       : Integer?
    caption_text      : String?
    caption_target    : {type, block_ids}?
    figure_info       : {crop_path, visual_ocr, chart_type, subfigures
    }?
    table_structure   : TableStructure?
    equation_latex    : String?
    nearby_block_ids  : List<BlockId>
    task_fields       : Dict<String, Any>
End Record

Record Relation
    type : "adj" | "sub" | "peer" | "identical" | "continuation"
    from : BlockId
    to   : BlockId
End Record
```

Figure 8: A simplified schema of SCIDOCIR. Optional fields are filled only when useful for generation or verification.

D SCIDOCIR SCHEMA

Figure 8 presents the simplified schema of SCIDOCIR. The representation stores document-level metadata, page-level metadata, block-level records, relations, and task-specific metadata. Optional fields are populated only when they are relevant to task generation or reward verification.

Table 17: Training-data construction directions and their instantiated subtasks.

Direction	Training subtasks
Layout and reading-flow recovery	Block role classification; parent–child linking; reading-order prediction; next-hop reading-target prediction.
Table logic consistency	Table logic consistency check.
Cross-document table integration	Real cross-document table merge; single-document simulated table merge.
Chart readout and recovery	Single-point chart readout; multi-point or series readout.
Chart visual consistency	Chart visual consistency.
Cross-document dataset comparison	Cross-document dataset intersection.
Notation understanding	Notation extraction; symbol ambiguity disambiguation.
Citation-role analysis	A3-lite citation-role classification.

Table 18: Construction sources and GPT-5.4 roles for subtasks 1–7. "None" denotes no external dataset beyond the SCIDOCIR paper collection.

Subtask	SCIDOCIR or document records	External or synthetic source	GPT-5.4 role and GT authority
Block role classification	Page image; target page index, bounding box, type, and layout annotation	None	Light question rewriting. The layout label remains the locked GT.
Parent–child linking	Block text, page and bounding box; heading level; caption pairs; footnote and continuation relations	None	Question rewriting and optional plausibility check. The verified relation remains GT.
Reading-order prediction	Order and reading annotations; block pages and boxes; section and discourse roles	None	Question rewriting only. GT follows the stored order and layout rules.
Next-hop reading target	Current and successor blocks; caption pairs; figure/table references; cross-page and appendix links	None	Question rewriting and optional next-hop plausibility check. The stored successor remains GT.
Table logic consistency	Table LaTeX, normalized JSON, caption, headers, cells, citing context, page, and box	None	Generates a natural corrupted statement, repair, and distractors, then self-checks them. Rules lock the perturbed field and answer.
Real cross-document table merge	Table schemas, rows, columns, values, captions, and nearby text from different papers	None	Polishes context and question. Rules align compatible schemas and compute the merged GT.
Simulated table merge	One real table and its context, split into two partially overlapping tables with permuted columns	Programmatic table splitting and a simulated second paper page	Adjusts table LaTeX, local prose, and question without changing locked cells. Rules compute the complete merged GT.

E TRAINING-DATA DIRECTIONS AND VERIFIABLE SUBTASKS

Table 17 lists the eight construction directions and fourteen verifiable subtasks underlying SCIDOCDATASET. These describe task construction rather than dataset size; instance and split counts are given in Table 4.

E.1 CONSTRUCTION SOURCES AND MODEL ROLES

The seed instruction is not an assertion that every sample is generated end to end by an LLM. It is a task-specific contract governing the part assigned to GPT-5.4 (OpenAI, 2026a). Deterministic labels, values, alignments, and perturbation records are locked before model invocation whenever possible. GPT-5.4 is used for linguistic naturalization, bounded candidate generation, semantic classification, or solvability checks. Tables 18 and 19 identify the records supplied to each template and the authority used for ground truth.

E.2 SEED INSTRUCTION TEMPLATES

The templates below show the shared generation contract and the task-specific seed appended to it. Angle-bracketed expressions are populated by the retrieval, synthesis, or rendering program. A locked field cannot be altered by the model. The model-facing training instance contains the generated question and its associated document input; construction metadata and source identifiers are retained only for generation, verification, and auditing.

```
You construct one scientific-document training instance from supplied records.
```

Table 19: Construction sources and GPT-5.4 roles for subtasks 8–14.

Subtask	SciDocIR or document records	External or synthetic source	GPT-5.4 role and GT authority
Single-point chart readout	Figure and caption blocks; page, box, and nearby text used for page replacement	PlotQA chart, axes, series labels, and source values	Naturalizes the question and performs repeated visual solvability checks. PlotQA values remain GT.
Multi-point or series readout	Same figure placement and context records as the single-point task	Complete PlotQA series data	Rewrites the question and checks the returned list. The source series remains GT.
Chart visual consistency	Figure position, caption, and page context used to reinsert a controlled chart	PlotQA values; Matplotlib radar, line, scatter, and bar renderers	Rewrites the question and checks solvability. Rules move one plotted value and retain the original annotation; the corruption log defines GT.
Dataset intersection	Dataset-like entities mined from text, tables, and captions when available; simulated two-paper text	Internal lexicon of approximately 50 CS, AI, mathematics, and natural-science resources; alias and composition pools	Checks name plausibility and polishes page prose. Exact set intersection defines GT.
Notation extraction	Final batch does not require real SciDocIR records	Mathematics, machine-learning, and physics formula templates; XeLaTeX page rendering	May polish surrounding prose while formulas and definitions stay locked. The saved symbol table defines GT.
Symbol ambiguity disambiguation	Final batch does not require real SciDocIR records	Formula and context templates; one-column, two-column, and spanning layouts	May naturalize local contexts without changing definitions. The rule log for each symbol occurrence defines GT.
A3-lite citation role	Block LaTeX and text; real citation keys; section, page, block identifier, and local context	No external bibliography database	Rules propose a role from lexical cues and section context; GPT-5.4 classifies the observed use. Ambiguous or disagreeing cases are rejected.

```
Inputs may include:
- IR_RECORDS: selected SciDocIR pages, blocks, relations, and metadata.
- DOCUMENT_VIEW: the page image, crop, or rendered document shown to the learner.
- STRUCTURED_SOURCE: source table, chart data, formula template, or entity lists.
- CONSTRUCTION_RECORD: locked labels, values, mappings, perturbations, and tolerances.

Use only the supplied material. Do not invent a scientific fact, citation, identifier,
     value, unit, relation, or source. Preserve every field marked LOCKED. Write one
     self-contained question that is answerable from DOCUMENT_VIEW and that states the
     required answer format. Do not reveal the answer or hidden construction metadata in
     the question.

Return JSON only:
{
  "question": "<user-facing instruction>",
  "answer": <task-specific schema>,
  "source_ids": ["<records used for audit>"],
  "quality_check": {"answerable": true, "single_interpretation": true}
}

If a locked gold answer is supplied, copy it exactly. If the task asks for a bounded
     semantic label, select only from the supplied ontology. Return {"reject": "<reason>"}
     when the evidence is incomplete, ambiguous, illegible, or inconsistent with the
     construction record.
```

E.2.1 Layout and Reading-Flow Recovery

1. Block role classification.

```
Inputs: PAGE_IMAGE, TARGET_BLOCK {page_index, bbox, type}, LAYOUT_ANNOTATION, ROLE_LABELS.

Mark TARGET_BLOCK visibly in the page view. Rewrite the base instruction as a concise
     classification question asking for exactly one label from ROLE_LABELS. Use nearby
     layout only as context and do not mention the stored type. Copy
     LAYOUT_ANNOTATION[TARGET_BLOCK] as the answer. Reject a crop containing several
     inseparable semantic blocks or a label absent from ROLE_LABELS.

Answer schema: "<one label from ROLE_LABELS>"
```

2. Parent–child linking.

```
Inputs: BLOCK_A, BLOCK_B, local page context, heading levels, caption pairs, footnote
     links, and continuation relations; LOCKED_RELATION.

Present A and B with visible labels and enough context to determine their structural
     relation. Ask whether B is the content targeted by A as a caption, a continuation of
     A, or unrelated. Use exactly caption_target, continuation, and none. Naturalize the
     wording without changing block contents. Copy LOCKED_RELATION as the answer and reject
     pairs that require unseen context or admit more than one relation.

Answer schema: "caption_target" | "continuation" | "none"
```

3. Reading-order prediction.

```
Inputs: PAGE_IMAGE, CANDIDATE_BLOCKS with page_index and bbox, ORDER_ANNOTATION,
     READING_ANNOTATION, and discourse roles; LOCKED_ORDER.

Assign visible labels A, B, C, ... in the supplied shuffled order. Ask for the order in
     which the selected blocks should be read. Ensure that the view retains columns,
     headings, captions, and any cross-page cue needed to solve the task. Copy
     LOCKED_ORDER, expressed as a label sequence, as the answer. Reject cyclic,
     disconnected, or visually ambiguous selections.

Answer schema: "<sequence such as A>B>D>C>"
```

4. Next-hop reading-target prediction.

```
Inputs: CURRENT_BLOCK, CANDIDATE_BLOCKS, verified successor, caption pairs, textual
     figure/table references, cross-page relations, and appendix links; LOCKED_NEXT_LABEL.

Label two to six candidates A through F. Ask which block should be read immediately after
     CURRENT_BLOCK in order to continue the document or follow the explicit reference. Keep
     plausible distractors from the same local context. Copy LOCKED_NEXT_LABEL as the
     answer. Reject the sample if two candidates are reasonable next hops or the required
     link is not visible.

Answer schema: "<one candidate letter>"
```

E.2.2 Table Logic and Integration

5. Table logic consistency.

```
Inputs: TABLE_LATEX, TABLE_JSON, caption, headers, cells, nearby citing text, and
     PERTURBATION_RECORD with one locked change.

Using the locked perturbation, write one natural but false statement about a value,
     entity, rank, comparison direction, or metric direction. Also write its minimal
     correction and three plausible statements that are exactly supported by the table.
     Shuffle the four statements, ask which one is inconsistent, and report the
     corresponding letter. Independently check every option against TABLE_JSON. Do not
     create another inconsistency.

Answer schema: {"choice": "A|B|C|D", "repair": "<corrected statement>"}
```

6. Real cross-document table merge.

```
Inputs: TABLE_A and TABLE_B from different papers, captions, nearby text, SCHEMA_MAP, and
     LOCKED_MERGE.

Write a concise cross-paper question asking the learner to merge the compatible rows and
     columns under the canonical schema in SCHEMA_MAP. Preserve method names, dataset
     splits, metric direction, units, and paper provenance exactly. Do not request a
     comparison that the mapping does not support. Copy LOCKED_MERGE as the answer; rules,
     not the language model, determine all aligned keys and values.

Answer schema: {"columns": ["..."], "rows": [{"row_key": "...", "values": {...},
     "sources": {...}}]}
```

7. Single-document simulated table merge.

```
Inputs: one real SOURCE_TABLE, SPLIT_RECORD, two simulated paper contexts, and
     LOCKED_MERGE.

The construction program has split SOURCE_TABLE into TABLE_A and TABLE_B with different
     column orders and partially overlapping rows. Improve the two LaTeX tables and their
     surrounding prose so they read as independent paper excerpts, but preserve every
     locked header and cell. Ask for a complete merge with duplicate rows reconciled by
     SPLIT_RECORD. Copy LOCKED_MERGE as the answer and reject any rewrite that changes a
     value, unit, or row identity.

Answer schema: {"columns": ["..."], "rows": [{"row_key": "...", "values": {...},
     "sources": {...}}]}
```

E.2.3 Chart Readout and Visual Consistency

8. Single-point chart readout.

```
Inputs: a PlotQA chart and source data, TARGET_POINT {series, x, value, unit}, target IR
     figure bbox, caption, and nearby paper text.

After the chart is placed in the paper page, write a natural question identifying one
     unambiguous point by series and x-axis condition. Ask for its value and unit and state
     the allowed precision when needed. Copy TARGET_POINT.value and unit as the answer.
     Perform repeated visual checks that the point, axes, and legend are legible; reject
     overlapping, clipped, or uncertain points.

Answer schema: {"series": "<name>", "x": "<condition>", "value": <number>, "unit":
     "<unit>"}
```

9. Multi-point or series readout.

```
Inputs: a PlotQA chart, FULL_SERIES_DATA, target IR figure bbox, caption, and nearby paper
     text; LOCKED_TARGETS.

Write a question asking for either several specified points or one complete short series
     in displayed x-axis order. State the unit and numeric precision. Preserve the locked
     series and x labels exactly. Copy the requested values from FULL_SERIES_DATA and
     visually verify the complete list after rendering. Reject stacked, occluded,
     discontinuous, or dual-axis cases without an explicit unambiguous mapping. Prefer
     single-point conversion if the full series is not reliably readable.

Answer schema: {"series": "<name>", "x": ["..."], "values": [<numbers>], "unit": "<unit>"}
```

10. Chart visual consistency.

```
Inputs: PlotQA source values, Matplotlib chart rendering, IR figure bbox, caption and page
     context, and CORRUPTION_RECORD for exactly one moved plotted value.

The rendered chart contains one visual error while its label, caption, or surrounding text
     preserves the correct value. Write a question asking for the affected series or
     entity, x-condition, correct value, and visually represented value. Do not reveal
     which mark was moved. Copy all fields from CORRUPTION_RECORD. Check that every other
     mark is consistent and that both values are visually recoverable within TOLERANCE.

Answer schema: {"entity": "<series or entity>", "condition": "<x-label>", "correct_value":
     <number>, "visual_value": <number>, "unit": "<unit>"}
```

E.2.4 Cross-Document Dataset Intersection

11. Cross-document dataset intersection.

```
Inputs: DATASET_LEXICON, ALIAS_MAP, two simulated paper records or IR-mined dataset
     mentions, and LOCKED_DATASET_SETS.

Check that each selected name plausibly denotes a dataset or scientific resource. Write or
     polish two short paper-like passages that mention the locked dataset sets without
     adding another dataset name. Ask for the resources common to both documents and
     require canonical names. Compute the answer by exact set intersection after ALIAS_MAP
     normalization. Reject unresolved family/version distinctions or prose that leaks the
     intersection explicitly.

Answer schema: {"intersection": ["<canonical dataset names>"]}
```

E.2.5 NOTATION UNDERSTANDING

12. Notation extraction.

```
Inputs: GENERATED_FORMULAS, DEFINITION_SENTENCES, SAVED_SYMBOL_TABLE, and a XeLaTeX page
     template. Real IR input is not required.

Polish only the unlocked surrounding prose while preserving every formula, symbol,
     definition, and first-occurrence position. Ask for all symbols explicitly defined in
     the displayed scope, their definitions, and their first visible locations. Do not
     include symbols that are merely used. Copy SAVED_SYMBOL_TABLE as the answer and reject
     a rendering in which a formula or definition is clipped or separated from its scope.

Answer schema: {"notations": [{"symbol": "<LaTeX>", "definition": "<text>",
     "first_location": "<visible location>"}]}
```

13. Symbol ambiguity disambiguation.

```
Inputs: repeated TARGET_SYMBOL, OCCURRENCE_RECORDS with distinct definitions, formula and
     context templates, and a one-column, two-column, or spanning page layout.

Naturalize only unlocked context sentences. Preserve each occurrence of TARGET_SYMBOL and
     its locally defined meaning. Ask the learner to map every labeled occurrence to the
     correct definition using its paragraph, formula, or section context. Do not imply that
     the symbol has one global meaning. Copy the occurrence-to-definition mapping from
     OCCURRENCE_RECORDS and reject contexts that do not uniquely disambiguate every
     occurrence.

Answer schema: {"symbol": "<LaTeX>", "occurrences": [{"label": "A", "definition": "..."}]}
```

E.2.6 CITATION-ROLE ANALYSIS

14. A3-lite citation-role classification.

```
Inputs: real block LaTeX and text, CITATION_KEY, section name, page, block id, local
     citing context, ROLE_ONTOLOGY, and HEURISTIC_ROLE.

Classify how CITATION_KEY is used in the supplied text, not what the cited paper is
     generally about. Select exactly one role from Background, MethodBasis, DataResource,
     Baseline, ResultSupport, Critique, or Extension. Do not consult or invent an external
     bibliography entry. Return a concise question that includes the observed citing
     context and the selected role. Reject the sample if several citations share the same
     grammatical role span, if two roles are equally plausible, or if the classification
     conflicts with the rule-based prior after review.

Answer schema: {"role": "<one role from ROLE_ONTOLOGY>"}
```

The released construction metadata retains source identifiers, locked fields, rendering records, and rejection outcomes. This separates model-assisted linguistic generation from the deterministic or auditable source of each answer.

F Experimental Settings and Additional Results

F.1 Benchmark Evaluation Setup

The leaderboard evaluates 13 proprietary and open multimodal models on the same 124 underlying questions under the four settings in Section 3.3, yielding 496 instances per model. The input contains only the question and its source materials; reference answers, task labels, and evaluator metadata are withheld. Overall scores average all instances, while capability-group and setting scores average the corresponding subsets. Failed or unusable responses receive zero rather than being removed from the denominator.

F.2 Experiment-Specific Training Mixtures

The reported Qwen3.6 post-training experiments use SciDocDataset. Selection emphasizes scientific-document operations and task-appropriate output contracts, including evidence localization, numerical and symbolic extraction, cross-document integration, and structured or executable outputs.

The SFT targets pair a teacher explanation with the reference final answer. The explanation-generation stage is reference-conditioned: the teacher receives the question and its reference answer and produces an explanatory solution. The corresponding teacher endpoint is recorded as `qwen3.6-max-preview`. Explanations are generated at the seed–language level and used with the two document views; they are not four independent solutions sampled from the student model. The construction retains existing explanations when regeneration fails or the replacement does not meet the length checks.

For RL, we use the 4,000 question–answer instances from the SFT mixture, without teacher explanations in the verifier references, together with 6,219 additional verifiable QA instances, yielding 10,219 instances. This mixture covers layout relations, tables, charts, notation, and document evidence. The published collection contains 3,924 SFT and 10,143 RL instances; its composition and splits are reported in Table 4.

F.3 Optimization and Reward

We perform SFT using LlamaFactory (Zheng et al., 2024) and GRPO using ms-swift (Zhao et al., 2025). SFT applies LoRA to the language component while freezing the vision tower and multimodal projector. GRPO starts from the merged SFT checkpoint and learns a new adapter. Both stages use ZeRO (Rajbhandari et al., 2020); the RL loss implementation is Dr. GRPO (Liu et al., 2025b). Table 20 summarizes the training configuration. The selected checkpoints correspond to 50 SFT optimizer steps and 200 additional GRPO optimizer steps; the latter is not a separate RL-only initialization.

Final-answer reward. A separate Qwen3.6-27B verifier compares the policy's final answer with the reference under the question's output requirements. It assigns one of three labels: `CORRECT`, `PARTIAL`, or `INCORRECT`, mapped to rewards 1, 0.2, and -1, respectively. Correctness includes content and material field, format, and ordering requirements; harmless wording differences are permitted. Partial credit requires materially correct progress rather than verbosity alone. The verifier receives the question, reference answer, and candidate final answer, not the policy's internal reasoning. Verifier generation uses temperature zero with thinking disabled. Empty or invalid final answers receive the incorrect reward. The policy itself is allowed to generate a reasoning segment before the final answer.

F.4 Checkpoint Selection and Evaluation Protocol

We compare saved checkpoints on a fixed 48-instance diagnostic subset of SciDocBench and select SFT step 50 and GRPO step 200. The full benchmark report then evaluates each selected checkpoint and the original model on all 496 instances, including those diagnostic instances. All three full-benchmark runs use the same Transformers (Wolf et al., 2020) inference interface, temperature zero, and a maximum of 32,768 newly generated tokens. This budget includes reasoning and the final answer.

Table 20: Post-training configuration. The RL verifier runs on two additional H200 GPUs alongside six training GPUs.

Setting	SFT	GRPO
Model adaptation and optimization		
Initialization	Original Qwen3.6-27B	Merged SFT step 50
Adaptation	LoRA	LoRA
LoRA rank / alpha	32 / 64	16 / 32
Learning rate	1.5×10^{-5}	5×10^{-7}
Learning-rate schedule	Cosine	Cosine
Warmup ratio	0.05	0.03
Compute and memory		
Training GPUs	8 H200	6 H200
Per-device training batch size	1	1
Gradient accumulation steps	2	2
Precision	BF16	BF16
DeepSpeed stage	ZeRO-3	ZeRO-2
Gradient checkpointing	Enabled	Enabled
Input, rollout, and reward optimization		
Configured maximum length	32,768	32,768
Maximum rollout completion	–	4,096 tokens
Maximum image pixels	589,824	262,144
Samples per prompt	–	8
Sampling temperature / top-p	–	0.8 / 0.95
KL coefficient	–	0.01
Loss implementation	Token cross-entropy	`dr_grpo`
Lower / upper clipping	–	0.20 / 0.28
Reward scaling	–	None
Reproducibility and checkpoint selection		
Random seed	42	42
Saved checkpoint interval	25 steps	100 steps
Reported checkpoint	**Step 50**	**Step 200**

The scorer removes the reasoning segment and evaluates only the final response. A truncated reasoning segment without a valid final-answer boundary is not treated as an answer; missing final answers receive zero. Rule-based and execution-based tasks retain their task-specific evaluators. LLM-judged tasks use GPT-5.4-mini with temperature zero and seed 42. All three models use the same final-answer extraction and scoring rules. Raw responses are preserved separately from scored final responses in lossless records.

For the general document benchmarks, we use the standard VLMEvalKit (Duan et al., 2024) implementations on the final answer, with empty answers scored as zero. DocVQA and InfoVQA use their labeled validation splits of 5,349 and 2,801 questions and report average normalized Levenshtein similarity (ANLS). ChartQA uses its 2,500-question test split, including 1,250 human-authored and 1,250 augmented questions. Its relaxed numeric accuracy allows a relative error of 5% for nonzero numeric references; nonnumeric answers use case-insensitive matching. All scores are on a 0–100 scale, and differences are calculated before rounding.

DocVQA and InfoVQA inference uses a 32,768-token generation cap and temperature zero. The original-model runs use Transformers, SFT uses the completed cached runs, and GRPO uses vLLM (Kwon et al., 2023). Missing final answers are retained and scored as zero, including 19 DocVQA and 51 InfoVQA responses for GRPO.

For ChartQA, all three models use the native VLMEvalKit `Qwen3VLChat` Transformers backend and built-in benchmark prompt, without a custom prompt. We use Qwen's recommended sampling parameters: `do_sample=true`, temperature 1.0, top-p 0.95, top-k 20, repetition penalty 1.0, and a maximum of 32,768 new tokens, including reasoning. Final-answer extraction and relaxed-accuracy scoring are identical across models. No additional training on the DocVQA, InfoVQA, or ChartQA training splits is used in these experiments.

F.5 Capability, Matched-Setting, and ChartQA Results

Table 3 jointly reports the seven capability groups, four matched input settings, and general document scores. SFT improves every matched setting over the original model. The GRPO-adapted model improves both Chinese settings and English interleaved performance over the same baseline. Under the common ChartQA configuration, the original, SFT, and SFT+GRPO models score 74.60, 75.40, and 78.00, respectively. Their gains over the original model are 0.80 and 3.40 points.

F.6 Qualitative Failures of the Strongest Model

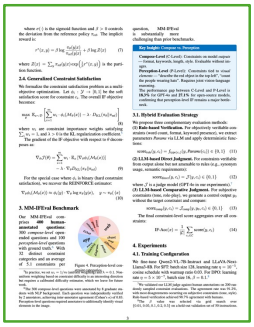
Table 21 examines three completed Claude-Opus-5 responses in the English All Images First setting. These are content errors in nonempty responses, rather than API failures. The examples separate errors in task scope, visual interpretation, and completeness; they are illustrative cases, not estimates of their frequency.

Table 21: Selected Claude-Opus-5 failures. IDs refer to Appendix B; outputs are abbreviated without changing the contrasted values. Scores are on the 0–1 scale.

Task	Reference and predicted output	Failure interpretation
ID 13: ordered metric extraction (A1); score 0	Reference: [Chamfer Distance, Precision, Recall, F1-score]. Prediction: [accuracy, completeness, Chamfer Distance (CD), precision, recall, F1-score].	The model retrieves the target metrics but adds two earlier entries. Since the question restricts the section and order, every aligned position is wrong. This is a scope/ordering failure, not failure to recognize all four metrics.
ID 59: logarithmic chart readout (B2); score 0.30	Reference: AQC(exp) range G (8×10^{-4} to 5×10^{-1}), QAOA range C. Prediction: AQC(exp) range A (5×10^{-3} to 5×10^{-1}), QAOA range C.	The wrong AQC lower endpoint is also asserted in the explanation. Correct QAOA selection earns partial credit; the rubric does not credit an explanation grounded in the incorrect joint range selection.
ID 71: table consistency audit (C1); score 0.67	Reference errors: 48.4, 47.8, 32.4. Prediction: 68.0, 32.4, 48.4. The missing 47.8 should be 47.3, the rounded mean of eight entries.	Two errors are found, including a correctly recomputed mean, but another mean is not checked correctly. The additional highlighting complaint about 68.0 does not recover the missing reference error.

G BENCHMARK EXAMPLES

A. Document Perception & Structure



2504.07957 (Rendered)

Ground Truth

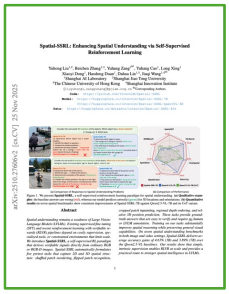
Our MM-IFEval comprises \textbf{400 human-annotated questions}: 300 \textit{compose-level} open-ended questions and 100 \textit{perception-level} questions with ground truth. With 32 distinct constraint categories and an average of 5.1 constraints per question, MM-IFEval is substantially more challenging than prior benchmarks.

Question

You are given a scientific PDF document. Your task is to extract information about the composition and data sources of the MM-IFInstruct-23k dataset described in the paper, and reproduce it as a TikZ radial diagram in LaTeX.
Follow these rules strictly:
Extraction rules:
- Identify all source datasets that contribute to MM-IFInstruct-23k.
- For each source dataset, extract the exact data quantity (e.g., number of samples or instances) if explicitly stated in the paper. If a quantity is not stated, omit the label but still include the node.
- Do not infer or fabricate any information not explicitly present in the document.
Diagram rules:
- Place MM-IFInstruct-23k as a single center node.
- Place each source dataset as a surrounding leaf node, evenly distributed in a radial (spoke) layout around the center.
- Draw arrows pointing from each leaf node toward the center node.
- If a data quantity is available for a source, annotate it on the corresponding arrow as a label.
- Use distinct visual styles for the center node (e.g., filled circle) and the leaf nodes (e.g., rounded rectangle).
Output rules:
- Output a complete, self-contained LaTeX file using the standalone document class.
- Do not include any commentary, explanation, or text outside the LaTeX code.
- Output only the complete LaTeX source, ready to compile.

Figure 9: Representative benchmark example for Category A: Document Perception and Structure.

B. Scientific Information Extraction



2510.27606

Ground Truth


```
{
 "Model": [
  {
   "name": "Spatial-SSRL-3B",
   "base_model": "Qwen2.5-VL-3B"
  },
  {
   "name": "Spatial-SSRL-7B",
   "base_model": "Qwen2.5-VL-7B"
  },
  {
   "name": "Spatial-SSRL-4B",
   "base_model": "Qwen3-VL-4B"
  }
 ],
 ……
```


Question

Please read the attached paper thoroughly and extract the following information in JSON format:
1. Model: List all models proposed in this paper. For each model, provide its name and the base model it is built upon.
2. Data: Provide the name of the training dataset constructed in this paper, and list all source datasets used to build it.
3. Training: List all training stages/methods used in the paper's training pipeline, in order. For each stage, provide its name and all hyperparameters explicitly mentioned in the paper.
Output the result as a single JSON object with exactly three keys: "Model", "Data", and "Training". Follow the schema below strictly:

```
{
 "Model": [
   {
    "name": "<model name>",
    "base_model": "<base model name>"
   }
 ],
 "Data": {
   "name": "<dataset name>",
   "source": ["<source dataset 1>", "<source dataset 2>"]
 },
 "Training": [
   {
    "name": "<training stage name>",
    "parameters": {
      "<param_name>": "<value>"
    }
   }
 ]
}
```

Notes:
- For numerical hyperparameters, use string representation (e.g., "1e-5", "128").
- Only include parameters that are explicitly stated in the paper. Do not infer or compute values.
- List training stages in the order they are applied in the pipeline.
- Output only the JSON, nothing else.

Figure 10: Representative benchmark example for Category B: Scientific Information Extraction.

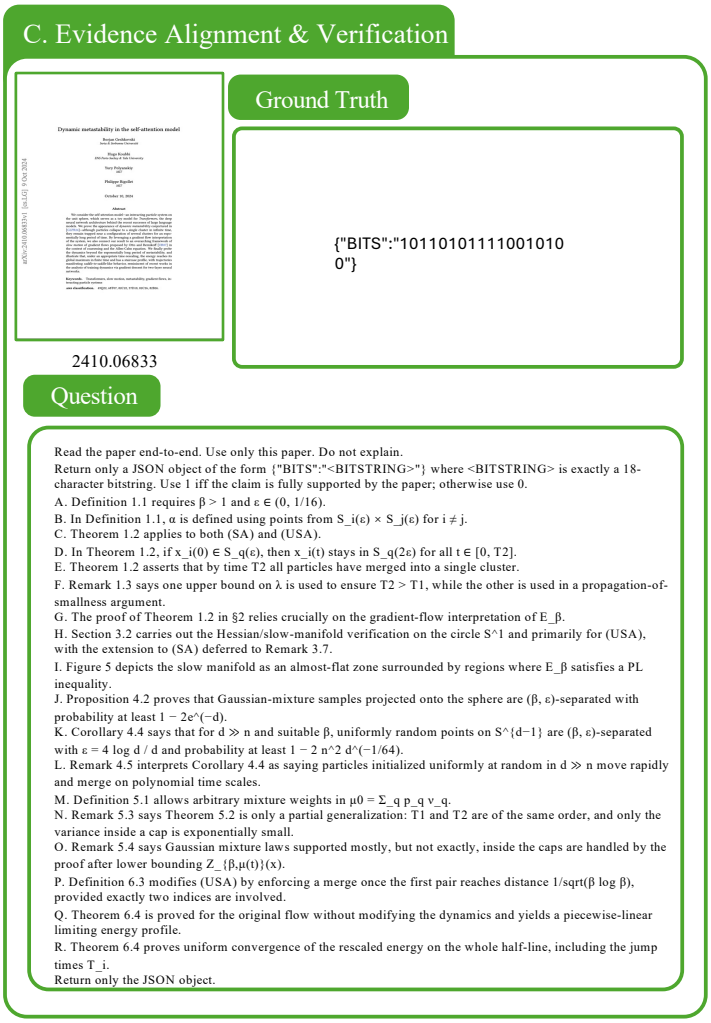


Figure 11: Representative benchmark example for Category C: Evidence Alignment and Verification.

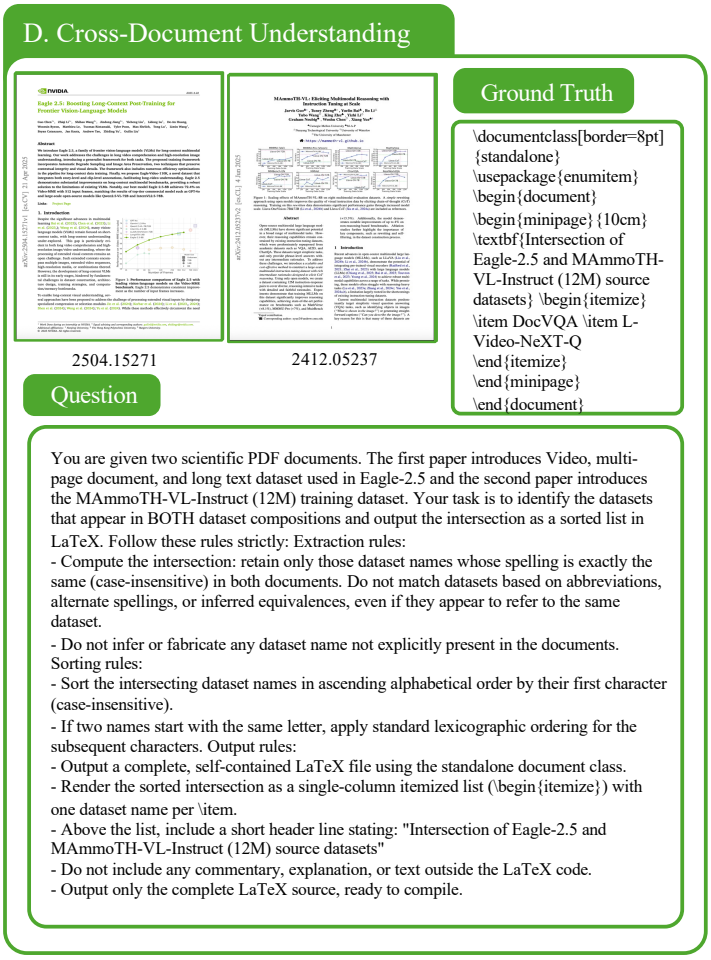


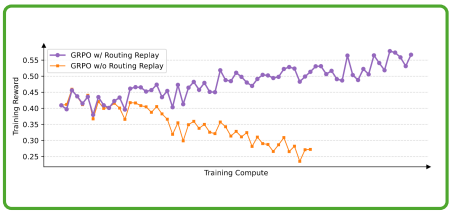
Figure 12: Representative benchmark example for Category D: Cross-Document Understanding.

E. Reconstruction & Execution



2507.18071

Ground Truth



Question

You are an expert Data Visualization Engineer and Python Developer (Matplotlib/Seaborn Specialist).
Objective: Your goal is to Reverse Engineer the provided scientific chart image into an executable Python script.
The Gold Standard:
1.Data Fidelity: The data points in your code must be as close as possible to the pixels in the image.
2. Visual Fidelity: The generated plot must look identical to the original image in terms of: Axis labels (including LaTeX math equations). Legend position and text. Line styles (dashed, solid, dotted), markers (circles, triangles), and colors. Axis limits and scales (log vs. linear). Grid lines and background style. Output: A single, self-contained Python code block.

Figure 13: Representative benchmark example for Category E: Reconstruction and Execution.

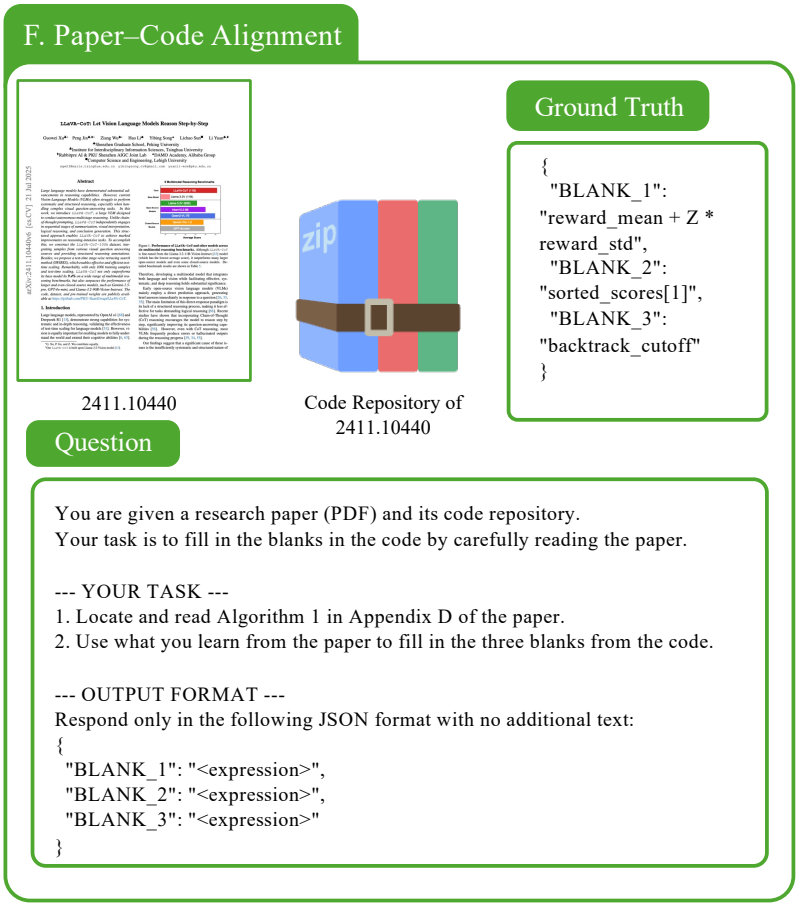


Figure 14: Representative benchmark example for Category F: Paper-Code Alignment.

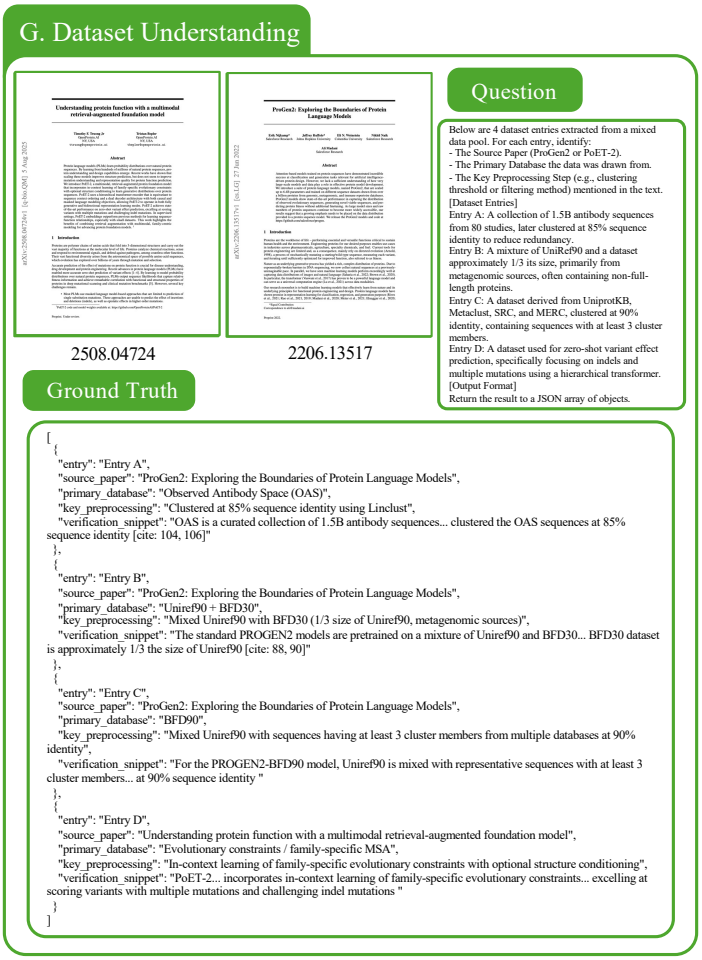


Figure 15: Representative benchmark example for Category G: Dataset Understanding.