🔥 SPARK: Skeleton-Guided Reasoning Synthesis from Large-Scale Scientific Literature

Yu Li[1,2], Wei Li[1,3], Xin Gao[1], Mengyuan Sun[4], Xiaoyang Wang[1], Qizhi Pei[5], Lijun Wu[1]*
[1]Shanghai AI Laboratory [2]University of Science and Technology of China
[3]East China Normal University [4]Peking University [5]Renmin University of China
liyu1@pjlab.org.cn, lijun_wu@outlook.com
🤗 https://huggingface.co/datasets/OpenDataArena/Spark-234K

Abstract

Scientific reasoning remains challenging for open-source models, largely due to the lack of high-quality scientific reasoning data. Existing datasets are often dominated by factual recall or formulaic problem solving, with limited emphasis on mechanism understanding, evidence-grounded reasoning, and hypothesis evaluation. To address this, we introduce **SPARK** (Scientific Paper Abstracted Reasoning sKeleton), a paper-oriented synthesis framework built on SCI-BASE, a large-scale corpus of research papers spanning 10 scientific disciplines. Instead of directly converting papers into question-answer pairs, SPARK treats the *claim-evidence-derivation* structure of a paper as the fundamental unit of reasoning synthesis. Specifically, SPARK (1) distills each paper into a compact reasoning skeleton capturing its central claims and supporting evidence, enabling self-contained question generation, and (2) synthesizes reasoning tasks from four scientific perspectives: *mechanistic reasoning, hypothesis falsification, quantitative derivation, and boundary calibration*. A final consistency verification stage further removes unsupported or contradictory outputs. Using this framework, we construct Spark-234K, a scientific reasoning dataset with substantially higher difficulty and diversity than existing resources. Experiments show that Spark-234K consistently outperforms existing scientific reasoning datasets while achieving stronger performance with significantly fewer training samples.

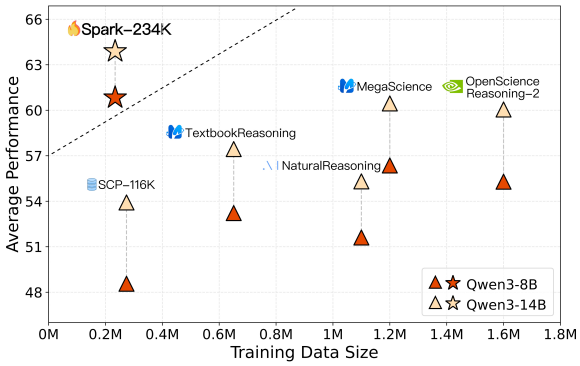

Figure 1: Comparison of the Qwen3-8B base and Qwen3-14B base models trained on baseline datasets versus Spark-234K (benchmark scores in Table 2).

1 Introduction

Recent Large Language Models (LLMs) have achieved strong performance on reasoning tasks such as mathematics and coding (Li et al., 2025; Cai et al., 2025; Pei et al., 2025), but scientific reasoning remains significantly more challenging (Li et al., 2026). A central limitation is the lack of high-quality scientific reasoning data. Existing scientific corpora are often constructed from textbooks, exam problems, or web resources (Lu et al., 2025; Fan et al., 2025), and therefore mainly emphasize factual recall, standard derivations, or routine calculations while underrepresenting the evidence-based and mechanism-driven reasoning processes that characterize real scientific discovery.

Research papers provide a natural source of scientific reasoning supervision because they contain expert-validated claims, supporting evidence, derivations, assumptions, and experimental analysis at scale (Taylor et al., 2022). Recent work has also explored leveraging scientific literature for reasoning-oriented data construction and scientific question-answering (Noorbakhsh et al., 2026; Dong et al., 2024; Liu et al., 2024). However, synthesizing high-quality reasoning data from papers remains difficult. Scientific papers are long, densely structured documents whose reasoning processes are often distributed across multiple sections, figures, tables, and experimental discussions. As a result, naively generating reasoning data from papers frequently produces samples that are not self-contained (Choi et al., 2021; Dasigi et al., 2021a;

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 13858–13881
October 24-29, 2026 ©2026 Association for Computational Linguistics

Jain and Garimella, 2026), omit critical assumptions or evidence, or fail to preserve the underlying claim-evidence reasoning structure (examples are shown in Figure 14).

We introduce **SPARK** (Scientific Paper Abstracted Reasoning sKeleton), a paper-oriented synthesis framework for scientific reasoning data construction. Instead of directly generating reasoning data from sections, chunks, or entire documents, SPARK treats a paper's claim-evidence-derivation chain as the fundamental unit of synthesis.

Given a scientific paper, SPARK first extracts its central claims together with the supporting evidence, assumptions, quantitative relations, and boundary conditions, and organizes them into a compact reasoning skeleton. This reasoning skeleton rewrites the core scientific argument into a self-contained form, reducing long-context interference (Liu et al., 2024) while preserving the underlying reasoning process and mitigating the lack of self-containment commonly observed in document-grounded generation (Choi et al., 2021; Dasigi et al., 2021a; Jain and Garimella, 2026).

Based on the reasoning skeleton, SPARK synthesizes scientific reasoning data from four perspectives: (1) *mechanistic reasoning*, which explains why an observed phenomenon occurs (Machamer et al., 2000); (2) *hypothesis falsification*, which distinguishes competing explanations (Pearl, 2009); (3) *quantitative derivation*, which emphasizes modeling and derivation instead of formula substitution (Sun et al., 2024); and (4) *boundary calibration*, which examines the assumptions and validity limits under which conclusions hold. A final consistency verification stage further removes unsupported or contradictory outputs.

Through the above framework and using 370K frontier scientific papers from the open SCIBASE (OpenDataLab, 2026) corpus, we construct Spark-234K, a high-quality scientific reasoning dataset containing 234K synthesized instances. Despite its compact scale, this dataset exhibits strong data efficiency and overall performance (Figure 1). Notably, it outperforms massive baseline datasets such as MegaScience and OpenScienceReasoning-2 (NVIDIA, 2025), despite these collections being more than five and nearly seven times its size, respectively. Across distinct base model families and parameter scales, our data surpasses the runner-up datasets on multiple benchmarks by substantial average margins of $+3.10$, $+4.47$, and $+3.43$ points.

2 Paper Collection and Preparation

Source corpus. SPARK is built on SCIBASE (OpenDataLab, 2026), an open scientific literature corpus comprising $\sim$3.36M papers across ten disciplines. Each paper is accompanied by metadata and parsed with MinerU2.5 (Niu et al., 2026) into an ordered sequence of paragraphs, tables, captions, and equations, with LaTeX preserved where available. This structured representation facilitates paper-level argument reconstruction while retaining the quantitative information needed for scientific reasoning.

Quality filtering and balanced sampling. To prepare inputs for the pipeline, we apply strict filtering. We restrict the timeframe to January 2024–March 2026 and discard non-English, incompletely parsed, or review articles lacking a clear trajectory of claims and evidence. Crucially, we prioritize papers rich in concrete observables, such as equations, statistical results, tables, and experimental comparisons. Finally, to mitigate the raw corpus's heavy skew toward medicine and life sciences (Table 5), we apply discipline-balanced sampling across the ten major scientific disciplines it spans. The resulting seed pool comprises roughly 370K argument-structured frontier papers.

3 Dataset Construction

Figure 2 illustrates the SPARK pipeline, which converts a screened paper into standalone training instances. The key design choice is to treat a paper as an argument rather than a flat document: SPARK first identifies the core conclusion, reconstructs the evidence trajectory that supports it, and stores this trajectory as a compact *reasoning skeleton*. Questions are then generated from the skeleton under several scientific reasoning perspectives and filtered against the same skeleton for self-containment and source consistency.

3.1 Skeleton Extraction

Skeleton extraction makes a long paper usable for supervision without flattening it into sections or fragmenting it into independent chunks. Directly prompting over the full paper risks *context diffusion*: the central scientific argument is diluted by peripheral details, making generation prone to shallow or unfocused questions. Instead of preserving the paper's original layout, skeleton extraction recovers the claim-centered argument: what conclusion the paper advances, what evidence supports it,

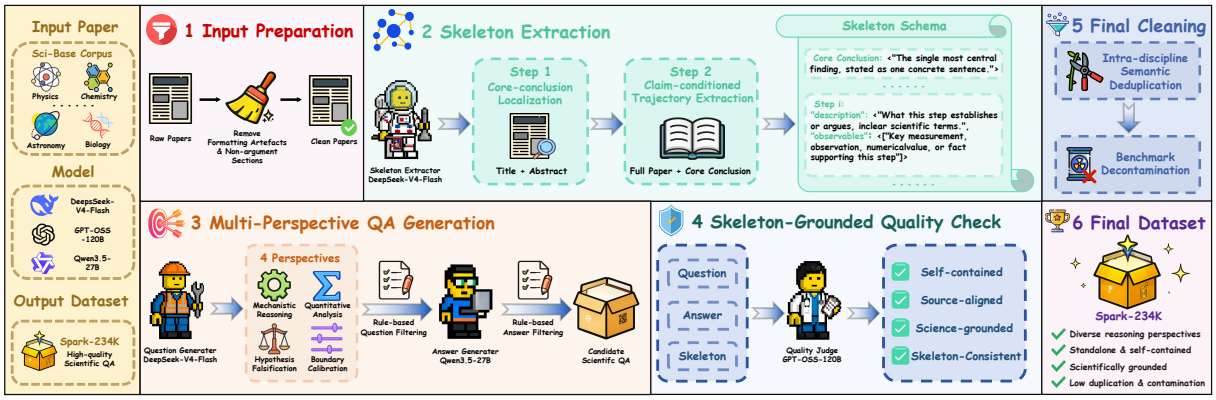

Figure 2: The SPARK data synthesis pipeline. A frontier paper is first condensed into a *reasoning skeleton*. We then generate deep scientific QA pairs via four targeted perspectives, followed by a lightweight consistency check to ensure data quality and self-containment.

and under which assumptions or boundary conditions the conclusion holds.

Input normalization. We first convert the parsed structure into a linearized text representation. This preserves the scientific content vital for reasoning, including equations, table text, captions, and numerical results, while stripping away formatting artifacts (such as headers, footers, and affiliations) and non-argument sections (such as references and acknowledgments).

Two-step extraction. Skeleton extraction recovers the core argument rather than the original section order. First, *core conclusion localization* identifies the central claim from the title and abstract, which typically state the research question, main finding, and scope while avoiding distracting body content. Second, *trajectory extraction* reads the normalized full paper conditioned on this conclusion, reconstructing the supporting argument as ordered steps. Each step contains a description of its logical role and a set of observables, such as measurements, comparisons, numerical values, formulas, or stated boundary conditions.

The resulting skeleton comprises the core conclusion and an ordered list of steps. As Figure 3 illustrates, it is not a section-level summary (Edge et al., 2025) but a compact evidence trajectory organized entirely around the paper's central claim.

3.2 Multi-Perspective Question Generation

The next stage generates questions that probe scientific reasoning rather than surface facts. The skeleton exposes the paper's claim, evidence, inferential links, and boundary conditions, which must be connected to understand the argument (Javaji et al., 2025). We therefore generate questions from four complementary perspectives.

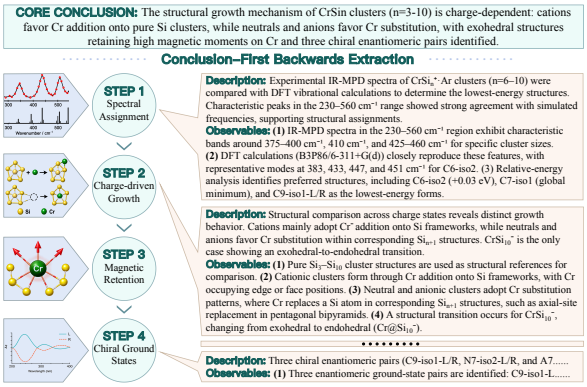

Figure 3: Example skeleton extracted by SPARK from a real paper, condensing the source into a core conclusion and ordered evidence steps with concrete observables.

Mechanistic reasoning turns a reported condition–effect relation into a causal explanation problem, asking why a structure, intervention, or regime produces the phenomenon (Machamer et al., 2000). **Hypothesis falsification** uses contrasts, ablations, or ruled-out alternatives to ask which explanatory model is supported by the evidence (Pearl, 2009). **Quantitative derivation** converts numerical observables, formulas, or scaling relations into modeling-first problems whose difficulty lies in selecting and linking principles, rather than substituting values (Sun et al., 2024). **Boundary calibration** asks where a conclusion ceases to be justified, focusing on the assumptions, regimes, confounders, or approximations that delimit the claim.

Together, these perspectives ask why a result holds, which explanation survives the evidence,

how quantities are related, and when the claim breaks down. This shifts generation away from fact retrieval and plug-and-chug computation toward mechanism-level reasoning. To maintain precision, SPARK follows a conservative yielding strategy: for each paper, it uses only the perspectives recommended during skeleton extraction, generates at most one question per perspective, and skips any perspective unsupported by the evidence.

Filtering. Generated questions undergo a lightweight, rule-based filtering stage to detect malformed formats and source-leakage phrases (detailed in Appendix G). The central constraint is self-containment: a question depending on unstated figures, tables, or source context is discarded. Finally, questions are audited for information sufficiency and intellectual depth, with thresholds adjusted to prevent the over-correction of valid reasoning trajectories.

3.3 Answer Generation

An answer model responds to each accepted question without access to the skeleton or source text. This setup mirrors the downstream fine-tuning environment and empirically verifies that the question is self-contained. We then apply heuristic filters to discard answers that are malformed, excessively short or long, or dominated by repeated n-grams.

3.4 Skeleton-Grounded Quality Check

While rule-based checks catch formatting issues, they cannot assess scientific grounding. We therefore deploy an LLM judge to evaluate each $\langle$question, answer, skeleton$\rangle$ triple against four criteria. For the **question**, it verifies: (i) self-containment (supplying necessary context independent of the paper), and (ii) consistency with the skeleton. For the **answer**, it checks if the reasoning is (iii) coherent and (iv) logically supported by the evidence trajectory. QA pairs failing any criterion are discarded.

The self-containment check acts primarily as a safeguard, as skeleton extraction prevents most dependencies at the source (evidenced by the low discard rate in Table 6). Crucially, by grounding the answer evaluation directly in the extracted skeleton, we bypass the massive majority voting and iterative rewriting typically required in traditional math and reasoning domains (Tong et al., 2024), keeping the pipeline computationally lightweight.

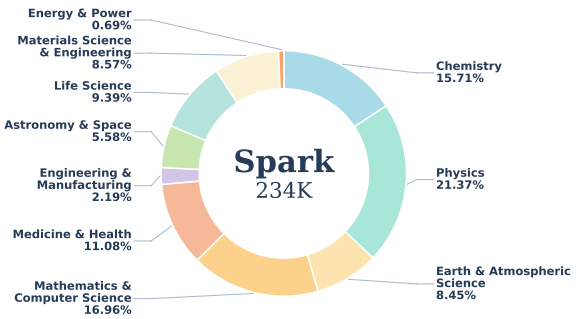

Figure 4: Discipline distribution of Spark-234K.

3.5 Final Data Cleaning

Finally, we apply deduplication and decontamination procedures. We perform semantic deduplication within each discipline using Qwen3-Embedding-8B embeddings (Zhang et al., 2025). To prevent benchmark leakage, we enforce 13-gram exact matching against all downstream evaluation sets, discarding any overlapping records.

4 Dataset Statistics and Analysis

By applying the proposed synthesis pipeline to the 370K filtered seed papers, we construct Spark-234K, a high-quality instruction tuning dataset dedicated to deep scientific reasoning.

4.1 Dataset Statistics

Figure 4 illustrates the disciplinary distribution of Spark-234K. Physics (21.37%), Mathematics & Computer Science (16.96%), and Chemistry (15.71%) constitute the largest segments of the dataset. The final composition reflects both our discipline-balanced sampling of source papers and cross-disciplinary variation in the number of valid reasoning instances retained through the synthesis pipeline. Medicine & Health (11.08%) and Life Science (9.39%) also contribute substantial portions, while the remaining disciplines, ranging from Materials Science & Engineering (8.57%) and Earth & Atmospheric Science (8.45%) to Energy & Power (0.69%), collectively ensure broad thematic coverage across diverse scientific domains.

Beyond disciplinary diversity, all four reasoning perspectives are well represented within the corpus. Mechanistic reasoning (35.6%) and quantitative derivation (30.2%) constitute the majority, aligning with the empirical nature of scientific papers, whereas boundary calibration (20.4%) and hypothesis falsification (13.9%) provide critical diversity in higher-order analytical evaluation.

Table 1: Comparison of dataset length (average token length) and diversity.

Dataset	Size	Length		Diversity	
		Q. Len.	Resp. Len.	Vendi ↑	Cent. Dist. ↑
SCP-116K	274K	122.09	555.68	173.42	0.4923
TextbookReasoning	651K	63.80	409.66	283.75	0.5598
NaturalReasoning	1.1M	76.28	766.33	349.42	0.6163
MegaScience	1.2M	170.77	692.93	373.78	0.6150
OpenScienceReasoning-2	1.6M	217.40	**5691.70**	290.17	0.5610
Spark-234K	234K	**377.61**	3139.45	**384.57**	**0.6183**

The retention statistics in Table 6 indicate that most generated instances pass the subsequent quality-control stages. In particular, the low proportion of rejections attributable to self-containment issues suggests that the reasoning skeleton helps reduce contextual dependencies more consistently during question generation.

4.2 Dataset Analysis

To evaluate the overall effectiveness of Spark-234K, we benchmark it against five established scientific reasoning datasets, SCP-116K, TextbookReasoning, OpenScienceReasoning-2, NaturalReasoning, and MegaScience.

Length. Spark-234K features the longest questions among all datasets, averaging 377.61 tokens (Table 1). This length reflects the contextual information, variables, and boundary conditions often required to formulate well-posed scientific reasoning problems. The responses average 3139.45 tokens, ranking second overall. Together, these statistics indicate that Spark-234K provides detailed reasoning supervision without relying on the longest response traces among the compared datasets.

Diversity. We quantify data diversity using the Vendi Score (Friedman and Dieng, 2022) and Centroid Distance (Suwanda et al., 2020). Despite its compact size (234K) compared with baselines containing up to 1.6M samples (Table 1), Spark-234K achieves the highest scores on both metrics (384.57 and 0.6183, respectively). These results indicate that Spark-234K maintains high semantic diversity relative to the compared datasets and are consistent with the intended effect of our multi-perspective generation design.

Answer correctness. Given the dataset's cross-disciplinary breadth, comprehensive manual verification is prohibitively expensive. We therefore conduct large-scale automated verification on a random sample of 20K questions. For each question, Gemini-3.1 Pro[1] independently generates a solution without access to the reference answer. GPT-OSS-120B (OpenAI et al., 2025) then judges whether the generated solution is semantically equivalent to the original response. This evaluation yields a 93.28% consistency rate, providing additional evidence for answer reliability.

Difficulty. Figure 5a highlights the limitations of existing datasets in eliciting deep reasoning. Using DeepSeek-V4-Flash, we grade 20K sampled questions per dataset on a five-level taxonomy (Appendix H). While baseline datasets contain high proportions of L1 (Fact recall) and L2 (Routine computation) questions, with their combined share exceeding 50% in TextbookReasoning, Spark-234K contains only 0.02% L1 and 1.2% L2 questions. Instead, over 93% of our dataset requires high-order reasoning, dominated by L4 (Multi-step, 58.5%) and L5 (Research-level, 34.6%) problems. Notably, this distribution emerges without explicit difficulty-based filtering, suggesting that skeleton-guided generation naturally tends to produce reasoning-intensive questions.

Self-containment. Naively converting scientific papers into QA pairs often yields unusable questions that rely on unstated, external context. As Figure 5b illustrates, multiple baseline datasets suffer from severe dependency issues, with their proportion of non-self-contained data exceeding 15%. In supervised fine-tuning, introducing such a high ratio of contextually broken data is enough to severely degrade training performance, as it inadvertently teaches models to hallucinate missing variables rather than perform rigorous deduction. To isolate the architectural contribution of our reasoning skeleton in solving this structural flaw, we report Spark-234K both before and after the final quality check. Even prior to the filtering stage, Spark-234K achieves a 98.55% self-containment rate, compared with 92.39% for the strongest baseline. This result suggests that skeleton-guided generation substantially reduces context dependency at the generation stage. The final quality check further increases the self-containment rate to 99.73%.

Human validation. To further verify the reliability of our automated quality assessments, we conduct a human validation study on 300 stratified samples across ten disciplines. Two paid domain

[1] https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview

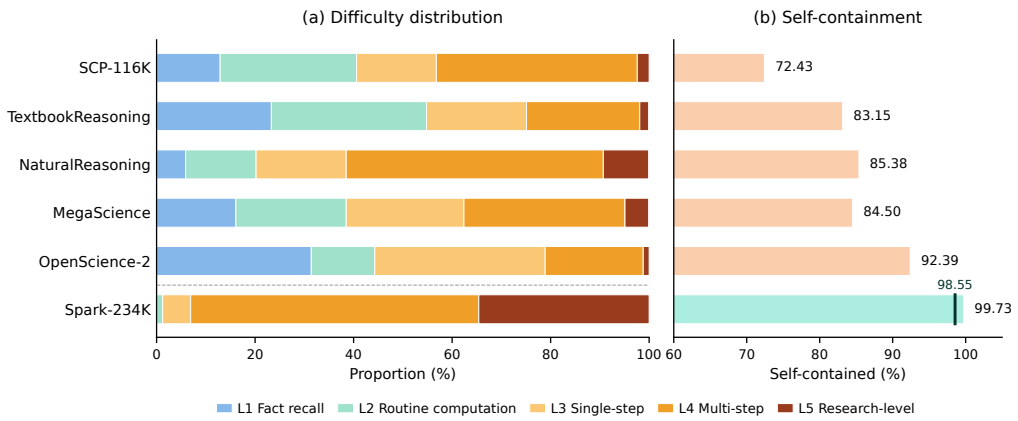

Figure 5: Difficulty distribution over five levels across datasets. Spark-234K shifts mass toward higher difficulty, while baselines concentrate on low-difficulty recall or routine computation.

experts for each discipline independently evaluate the samples across three dimensions: correctness, self-containment, and difficulty (Appendix H). Results show 90.3% agreement with expert-derived answers, 97.7% agreement on self-containment, and 94.3% agreement on difficulty labels, confirming both the reliability of our annotations and the challenging nature of the dataset. Full details are provided in Appendix I.

5 Experiments

5.1 Experimental Setup

Training. Fine-tuning is conducted on Llama3.1-8B-Base (Grattafiori et al., 2024), Qwen3-8B-Base, and Qwen3-14B-Base (Yang et al., 2025). All comparisons share the same training hyperparameters; details are provided in Appendix C.

Evaluation. For fair comparison, all models are evaluated in the zero-shot setting across two groups of benchmarks: **(1) general science reasoning benchmarks** (GPQA-Main [GPQA-M] (Rein et al., 2024), GPQA-Diamond [GPQA-D] (Rein et al., 2024), SuperGPQA [S-GPQA] (Du et al., 2026), SciBench (Wang et al., 2024a), MMLU (Hendrycks et al., 2020), and MMLU-Pro (Wang et al., 2024b)) and **(2) domain-specific benchmarks** (ChemBench (Mirza et al., 2024), CS-Bench (Song et al., 2025), PubMedQA (Jin et al., 2019), MedQA-US (Yao et al., 2026), GSM8K (Cobbe et al., 2021), and MATH-500 (Hendrycks et al., 2021)). We report avg@3 or avg@5 for all benchmarks, with details provided in Appendix D.

5.2 Main Results

Table 2 reports the evaluation on six scientific reasoning benchmarks. The results empirically validate our central claim: preserving a paper's reasoning skeleton provides far denser supervision than the brute-force scaling of synthetic examples.

Data Efficiency. The macro averages highlight the data efficiency of our approach. Across all three backbones, Spark-234K consistently achieves the highest overall performance using only 234K examples—less than one-fifth the volume of million-scale baselines such as MegaScience (1.2M). For instance, fine-tuning Qwen3-8B on Spark-234K yields an average score of 60.83, compared with 56.36 for the strongest baseline. These results show that Spark-234K achieves strong downstream performance with substantially fewer training examples, highlighting the importance of data construction quality in addition to dataset scale.

Performance on Challenging Reasoning Benchmarks. The performance gains are particularly pronounced on benchmarks requiring complex, multi-step reasoning. On the Expert Science MCQ cluster (GPQA series) and the quantitative SciBench, Spark-234K consistently achieves strong performance relative to the compared datasets. For Qwen3-14B, training on Spark-234K yields GPQA-M and GPQA-D scores of 53.71 and 55.05, respectively, outperforming models trained on substantially larger corpora. Improvements on MMLU-Pro show a similar trend. These results are consistent with the high proportion of L4 and L5 questions observed in Spark-234K, suggesting

Table 2: Main results on six scientific reasoning benchmarks. Avg. is the macro average over all six. Within each model group, the best result per column is in **bold** and the second best is underlined.

		Expert Science MCQ			Quant.	Broad Science		
Model / Training Data	**Size**	**GPQA-M**	**GPQA-D**	**S-GPQA**	**SciBench**	**MMLU**	**MMLU-Pro**	**Avg.**
Llama3.1-8B-Base								
Base	—	24.55	24.75	14.41	17.41	54.06	24.39	26.60
+ SCP-116K	274K	26.83	30.71	15.09	20.37	50.71	25.81	28.25
+ TextbookReasoning	651K	30.11	33.95	24.76	31.83	66.25	47.73	39.11
+ NaturalReasoning	1.1M	31.85	32.91	23.80	31.29	60.23	44.31	37.40
+ MegaScience	1.2M	33.71	**38.89**	25.56	36.87	68.79	49.79	42.27
+ OpenScienceReasoning-2	1.6M	29.91	26.77	24.09	27.59	**70.10**	50.61	38.18
+ Spark-234K (ours)	234K	**38.38**	37.37	**28.47**	**43.80**	69.36	**54.83**	**45.37**
Qwen3-8B-Base								
Base	—	32.59	29.29	29.05	30.30	66.22	45.24	38.78
+ SCP-116K	274K	36.93	34.62	33.80	54.13	71.24	60.63	48.56
+ TextbookReasoning	651K	40.75	42.90	37.01	61.31	73.07	64.26	53.22
+ NaturalReasoning	1.1M	40.03	41.29	34.44	59.83	72.15	61.91	51.61
+ MegaScience	1.2M	43.75	44.95	38.66	64.83	79.65	66.31	56.36
+ OpenScienceReasoning-2	1.6M	40.78	39.36	38.55	62.37	81.15	69.55	55.29
+ Spark-234K (ours)	234K	**50.22**	**52.53**	**41.84**	**66.35**	**81.50**	**72.54**	**60.83**
Qwen3-14B-Base								
Base	—	37.95	38.38	33.50	50.16	76.01	59.35	49.23
+ SCP-116K	274K	42.80	43.75	35.29	59.29	77.37	64.99	53.92
+ TextbookReasoning	651K	44.27	46.81	42.15	63.61	79.25	68.53	57.44
+ NaturalReasoning	1.1M	44.09	44.83	34.16	63.27	77.59	67.94	55.31
+ MegaScience	1.2M	47.32	48.99	44.47	67.81	82.44	71.65	60.45
+ OpenScienceReasoning-2	1.6M	46.64	47.83	44.21	65.41	**83.99**	72.20	60.05
+ Spark-234K (ours)	234K	**53.71**	**55.05**	**47.62**	**69.55**	83.65	**73.68**	**63.88**

that skeleton-guided synthesis provides effective supervision for challenging reasoning tasks.

5.3 Domain-Specialized Results

Table 3 evaluates performance across distinct scientific disciplines. Spark-234K consistently achieves the highest macro averages across all three backbones, indicating broad improvements across the evaluated chemistry, computer science, medicine, and mathematics benchmarks.

Notably, Spark-234K maintains a substantial advantage on domain-specific benchmarks such as MedQA-US (e.g., 79.36 vs. 72.58 for MegaScience on Qwen3-14B). This result suggests that the gains from our training data transfer effectively to specialized scientific domains.

5.4 Data Efficiency and Scaling

All experiments in this and the following ablation use Qwen3-8B-Base, evaluated on MMLU, MMLU-Pro, SciBench, and GPQA-Diamond under the same training and evaluation protocol as our main experiments. Here we train on progressively larger random subsets of Spark-234K to measure how performance scales with the amount of synthesized data (Figure 6).

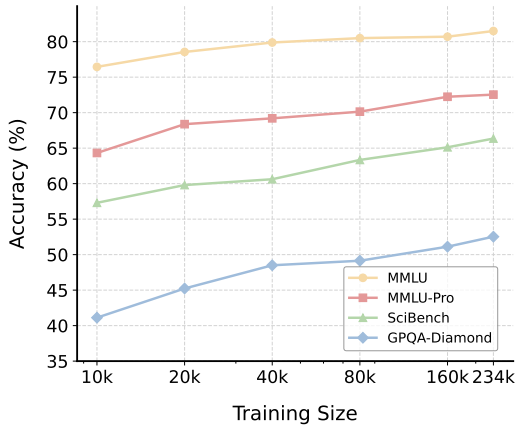

Figure 6: Performance scaling with the number of synthesized training records from Spark-234K.

While the total dataset volume is relatively compact, the scaling curve exhibits a consistent and incremental improvement across benchmarks as the data size expands. This steady upward trajectory indicates that the model has not yet saturated on the provided scientific reasoning patterns. It suggests a highly promising direction for future work, where simply applying our automated extraction pipeline to a broader corpus of academic literature could naturally yield further scaling gains.

Table 3: Domain-specialized transfer results across chemistry, computer science, medicine, and mathematics.

Model / Training Data	Size	ChemBench	CS-Bench	PubMedQA	MedQA-US	GSM8K	MATH-500	Avg.
Llama3.1-8B-Base								
Base	—	23.03	49.90	51.20	48.78	56.40	19.20	41.42
+ MegaScience	1.2M	42.97	59.89	**76.60**	62.03	71.65	45.20	59.72
+ Spark-234K (ours)	234K	**44.51**	**64.29**	74.20	**66.35**	**85.10**	**64.00**	**66.41**
Qwen3-8B-Base								
Base	—	49.60	69.16	61.20	55.54	88.63	70.80	65.82
+ MegaScience	1.2M	53.01	76.07	**77.40**	66.93	93.03	79.60	74.34
+ Spark-234K (ours)	234K	**55.85**	**77.61**	75.60	**75.33**	**94.09**	**81.20**	**76.61**
Qwen3-14B-Base								
Base	—	53.73	75.98	77.00	63.71	93.86	71.60	72.65
+ MegaScience	1.2M	**58.82**	79.92	**78.80**	72.58	94.77	82.00	77.82
+ Spark-234K (ours)	234K	57.63	**81.50**	78.20	**79.36**	**94.82**	**85.37**	**79.48**

5.5 Pipeline Ablation

We ablate the two central designs of SPARK, with all settings trained on a fixed 40K set (Table 4).

Unit of synthesis. We replace skeleton extraction with three alternatives that feed the generator different representations of the same paper: the whole paper, chunked passages, or an LLM-generated summary. The rest of the pipeline remains unchanged. All baseline representations degrade sharply. Whole-paper input performs worst (-9.15), as processing a long document diffuses the central claims and yields shallow questions. Chunking and summarization also lose ground (-5.63, -6.71) by severing cross-section claim–evidence chains or discarding the reasoning skeleton entirely. In contrast, extracting a reasoning skeleton preserves the paper's claim–evidence–derivation trajectory, yielding the highest-quality synthetic data.

Question generation. We fix the skeleton and perform leave-one-out ablations over the four perspectives, resampling from the remaining three to maintain 40K examples so that variants differ only in the included perspectives. With data volume fixed, the performance gaps are smaller, yet removing any perspective still causes a clear drop, indicating that each contributes complementary signal. Quantitative derivation has the largest effect (-4.64), likely because it contributes many computation-intensive questions that strengthen the model's quantitative reasoning ability.

Table 4: Ablation of key SPARK design choices.

Setting	Avg.	Δ
SPARK (full)	63.46	—
Unit of synthesis		
w/o Skeleton (→ whole paper)	54.31	−9.15
w/o Skeleton (→ chunks)	57.83	−5.63
w/o Skeleton (→ summary)	56.75	−6.71
Question generation		
w/o Mechanistic	61.59	−1.87
w/o Falsification	60.30	−3.16
w/o Quantitative	58.82	−4.64
w/o Boundary	61.19	−2.27

6 Related Work

6.1 Scientific Reasoning Data

Scientific post-training data has expanded from curricular sources to large synthetic mixtures. TextbookReasoning (Fan et al., 2025) extracts QA from university textbooks, while domain-specific efforts such as ChemData700K (Zhang et al., 2024) tune models on field-focused instruction data. Broader collections, including SCP-116K (Lu et al., 2025), OpenScienceReasoning-2 (NVIDIA, 2025), NaturalReasoning (Yuan et al., 2026), and MegaScience (Fan et al., 2025), improve scale and coverage. However, scale alone does not guarantee depth. Textbooks favor settled knowledge, while synthetic mixtures often default to factual recall and formulaic problem-solving. This motivates sourcing supervision from texts rich in scientific reasoning, rather than mere facts.

6.2 Data Synthesis from Scientific Literature

Frontier papers naturally fulfill this requirement, yet scalable task conversion remains nontrivial. While human annotations (Dasigi et al., 2021b) resist scaling, LLM pipelines now synthesize tasks from full papers (Wan et al., 2024; Liu et al., 2026), chunked text (Sarthi et al., 2024), or retrieved passages (Wang et al., 2025). However, treating a paper as a monolithic document diffuses its core

claims, while arbitrary chunking severs vital cross-section evidence chains. SPARK instead distills a paper's claims, evidence, and derivations into a compact *reasoning skeleton* as source, yielding self-contained, mechanism-oriented questions.

7 Conclusion

We introduced SPARK to address the limitations of existing scientific reasoning data by treating a paper's *claim–evidence–derivation* structure as the fundamental unit of synthesis. SPARK distills papers into reasoning skeletons, generates questions across four scientific perspectives, and enforces strict consistency. The resulting Spark-234K offers greater difficulty and diversity, and experiments demonstrate it substantially improves the scientific reasoning capabilities of base models.

Limitations

This work has two main limitations. First, constrained by computational resources, our experiments are limited to supervised fine-tuning; we leave the use of Spark-234K for reinforcement learning, including more verifiable reward design, to future work. Second, given the scale and disciplinary breadth of Spark-234K, comprehensive expert verification of the entire dataset is prohibitively expensive. We therefore rely primarily on automated consistency checks for large-scale quality control, complemented by a stratified human validation study on 300 samples across all ten disciplines. Extending expert evaluation to a substantially larger portion of the dataset would provide a more comprehensive assessment of data reliability.

Acknowledgements

This work was supported by Shanghai Artificial Intelligence Laboratory.

References

Mengzhang Cai, Xin Gao, Yu Li, Honglin Lin, Zheng Liu, Zhuoshi Pan, Qizhi Pei, Xiaoran Shang, Mengyuan Sun, Zinan Tang, et al. 2025. OpenDataArena: A fair and open arena for benchmarking post-training dataset value. *arXiv preprint arXiv:2512.14051*.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

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

Kuicai Dong, Derrick Goh Xin Deik, Yi Quan Lee, Hao Zhang, Xiangyang Li, Cong Zhang, and Yong Liu. 2024. MC-indexing: Effective long document retrieval via multi-view content-aware indexing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2673–2691, Miami, Florida, USA. Association for Computational Linguistics.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. 2026. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. 2025. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *arXiv preprint arXiv:2507.16812*.

Dan Friedman and Adji Bousso Dieng. 2022. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Akriti Jain and Aparna Garimella. 2026. Knowing what's missing: Assessing information sufficiency in question answering. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4163–4174, Rabat, Morocco. Association for Computational Linguistics.

Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. 2025. Can AI validate science? benchmarking LLMs on claim →Evidence reasoning in AI papers. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2355–2379, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Yu Li, Zhuoshi Pan, Honglin Lin, Mengyuan Sun, Conghui He, and Lijun Wu. 2025. Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning. *Preprint*, arXiv:2507.17512.

Yu Li, Xiaoran Shang, Qizhi Pei, Yun Zhu, Xin Gao, Honglin Lin, Zhanping Zhong, Zhuoshi Pan, Zheng Liu, Xiaoyang Wang, Conghui He, Dahua Lin, Feng Zhao, and Lijun Wu. 2026. Tracing the roots: A multi-agent framework for uncovering data lineage in post-training LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9606–9625, San Diego, California, United States. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tengxiao Liu, Deepak Nathani, Zekun Li, Kevin Yang, and William Yang Wang. 2026. WildSci: Advancing scientific reasoning from in-the-wild literature. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11677–11695, San Diego, California, United States. Association for Computational Linguistics.

Dakuan Lu, Xiaoyu Tan, Rui Xu, Tianchu Yao, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. 2025. Scp-116k: A high-quality problem-solution dataset and a generalized pipeline for automated extraction in the higher education science domain. *arXiv preprint arXiv:2501.15587*.

Peter Machamer, Lindley Darden, and Carl F Craver. 2000. Thinking about mechanisms. *Philosophy of science*, 67(1):1–25.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Zirui Tang, Boyu Niu, and 40 others. 2026. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 13–42, San Diego, California, USA. Association for Computational Linguistics.

Kimia Noorbakhsh, Joseph Chandler, Pantea Karimi, Mohammad Alizadeh, and Hari Balakrishnan. 2026. Savaal: Scalable concept-driven question generation to enhance human learning.

NVIDIA. 2025. OpenScienceReasoning-2. `https://huggingface.co/datasets/nvidia/OpenScienceReasoning-2`. Hugging Face dataset.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenDataLab. 2026. Sci-base: The largest ai-ready scientific foundation dataset.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, and Rui Yan. 2025. Mathfusion: Enhancing mathematical problem-solving of llm through instruction fusion. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7400–7420.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, et al. 2025. Cs-bench: A comprehensive benchmark for large language models towards computer science mastery. In *International Conference on Learning Representations*, volume 2025, pages 55244–55288.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

Rizki Suwanda, Zulfahmi Syahputra, and Elvi M Zamzami. 2020. Analysis of euclidean distance and manhattan distance in the k-means algorithm for variations number of centroid k. In *Journal of Physics: Conference Series*, volume 1566, page 012058. IOP Publishing.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. DART-math: Difficulty-aware rejection tuning for mathematical problem-solving. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*.

Penghao Wang, Yuhao Zhou, Mengxuan Wu, Ziheng Qin, Bangyuan Zhu, Shengbin Huang, Xuanlei Zhao, Panpan Zhang, Xiaojiang Peng, Yuzhang Shang, et al. 2025. Researchgpt: Benchmarking and training llms for end-to-end computer science research workflows. *arXiv preprint arXiv:2510.20279*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024a. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on Machine Learning*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. 2026. MedQA-CS: Objective structured clinical examination (OSCE)-style benchmark for evaluating LLM clinical skills. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6183–6257, Rabat, Morocco. Association for Computational Linguistics.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason Weston, et al. 2026. Natural-reasoning: Reasoning in the wild with 2.8 m challenging questions. *Advances in Neural Information Processing Systems*, 38.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. 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)*, pages 400–410.

A The Source Corpus: SCI-BASE

SCI-BASE serves as the foundational corpus of open-access scientific literature for our dataset, comprising over 3.36 million parsed and formula-preserving papers with coverage up to March 2026. To ensure strict data quality for downstream synthesis, we applied rigorous filtering criteria to the raw aggregated documents, explicitly removing incomplete papers, such as those missing abstracts or essential metadata.

As detailed in Table 5, the finalized high-quality corpus spans ten major scientific disciplines. While the distribution inherently skews towards Medicine & Health Sciences and Life Sciences (collectively accounting for 61.3%) due to natural open-access publishing volumes, this massive, structured repository provides a highly reliable and diverse pool for our discipline-aware sampling pipeline.

Table 5: Disciplinary distribution of SCI-BASE.

Discipline	# Papers	%
Medicine and Health Sciences	1,303,682	38.8%
Life Sciences	756,011	22.5%
Engineering and Mfg. Science	399,840	11.9%
Earth and Atmospheric Sciences	201,607	6.0%
Physics	181,431	5.4%
Chemistry	181,446	5.4%
Math and Computational Science	134,452	4.0%
Materials Science and Eng.	120,963	3.6%
Astronomy and Space Sciences	63,848	1.9%
Energy and Power Science	16,800	0.5%
Total	3.36M	100%

B Construction Funnel

Table 6 tracks the volume dynamics and rejection rates across the different stages of our data synthesis pipeline.

Skeleton Extraction. The pipeline starts from approximately 370K papers obtained after corpus-level filtering. Of these, roughly 325K are retained after skeleton extraction and lightweight post-processing. The approximately 45K excluded papers fall into three main categories. First, for some papers, no clear target conclusion could be identified, including documents without an explicit conclusion (e.g., certain survey papers) as well as noisy or misclassified entries in the original SCI-BASE corpus, such as speeches or lecture transcripts. Second, some papers failed to yield a valid skeleton after three API retries, typically due to transient network issues or unusually long or short source documents. Third, we apply several lightweight heuristic checks to the extracted skeletons: we discard skeletons containing three or fewer argumentation steps, as well as those that explicitly depend on figures or tables, in order to preserve sufficient reasoning structure and self-containedness.

Table 6: Construction funnel of Spark-234K. The retained size denotes the number of instances remaining after each stage. †Percentages for the four rejection reasons denote their respective shares among the QA pairs rejected during the skeleton-grounded check, rather than removal rates over the full dataset.

Stage	Retained Size	Removal Rate
Initial sampling	370K	–
Skeleton extraction	325K	–
Question generation	293K	–
Answer generation	283K	–
Skeleton-grounded check	241K	14.97%
Question not self-contained	–	7.92%†
Question not paper-grounded	–	13.61%†
Answer inconsistent with skeleton	–	65.87%†
Answer internally incoherent	–	12.60%†
Deduplication & decontamination	234K	3.04%

QA Filtering and Finalization. The skeleton-grounded check removes 14.97% of the generated QA pairs. Among the QA pairs rejected at this stage, 7.92% are rejected because the questions are not self-contained, suggesting that the upstream skeleton extraction and post-processing stages effectively reduce contextual dependency issues. In contrast, 65.87% of the rejected QA pairs are filtered because their answers are inconsistent with the original paper skeleton. This distribution indicates that the quality auditor primarily removes answers that deviate from the source reasoning, rather than questions with insufficient contextual information.

In the subsequent deduplication and decontamination stage, an additional 3.04% of the remaining QA pairs are removed, yielding approximately 234K final instances. These removals arise from semantic duplication, while no instance triggers our 13-gram exact-match decontamination criterion against the downstream evaluation benchmarks considered in this work. This observation indicates low exact lexical overlap with these evaluation sets, reducing the risk of direct benchmark leakage under our decontamination criterion.

Source Provenance and Data Diversity. To characterize the composition of the finalized dataset, Figure 7 illustrates the distribution of source publishing venues and platforms within

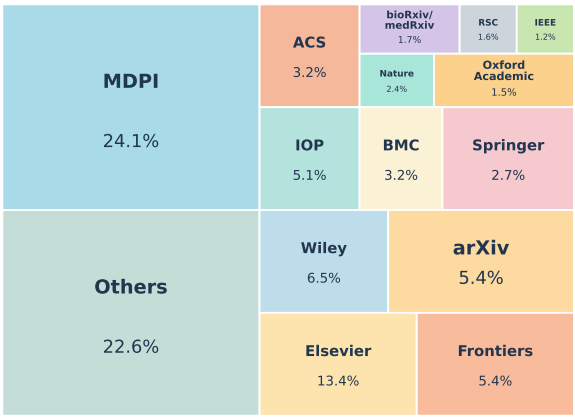

Figure 7: Publisher distribution of the source literature underlying Spark-234K.

Spark-234K. The underlying corpus aggregates literature from a heterogeneous mix of prominent academic publishers and preprint repositories. As shown in the treemap, the dataset draws substantially from major venues including MDPI (24.1%), Elsevier (13.4%), Wiley (6.5%), arXiv (5.4%), and Frontiers (5.4%). This is supplemented by professional societies, publishers, and specialized platforms, including IOP (5.1%), ACS (3.2%), BMC (3.2%), Springer (2.7%), Nature Portfolio (2.4%), bioRxiv/medRxiv (1.7%), RSC (1.6%), Oxford Academic (1.5%), and IEEE (1.2%), alongside a long-tail category of other venues comprising 22.6%.

This broad source provenance provides a diverse foundation for the construction of Spark-234K. The included publishers, academic societies, and preprint repositories span heterogeneous disciplinary communities, vocabulary conventions, and scientific writing practices. Drawing source reasoning structures from this diverse collection reduces dependence on the conventions of any single venue or publisher and broadens the linguistic and disciplinary coverage of the synthesized reasoning instances.

C Training Details

We conduct fine-tuning using LlamaFactory (Zheng et al., 2024). To ensure rigorous evaluation integrity, all baseline datasets undergo a strict 13-gram exact match decontamination process against all downstream test benchmarks prior to training. Unless otherwise specified, experiments utilize 16 GPUs with a per-device batch size of 4 and 2 gradient accumulation steps. For 32-GPU runs, we adjust gradient accumulation to 1 to maintain a consistent global batch size. Table 7 summarizes the remaining shared hyperparameters.

Table 7: Training Configuration.

Configuration	Value
Cutoff length	32,768
Epochs	3.0
Learning rate	5.0e-6
Warmup ratio	0.05
Per-device batch size	4
Gradient accumulation (16 GPUs)	2
Gradient accumulation (32 GPUs)	1
Flash attention	FlashAttention-2
Liger Kernel	Enabled
Thinking mode	Enabled
Average tokens across devices	True
Drop last dataloader batch	False

D Evaluation Details

We organize evaluation benchmarks into two groups. The first group contains general science reasoning benchmarks used in the main comparison, covering expert science QA, quantitative scientific problem solving, and broad multidisciplinary science understanding. The second group contains domain-specific benchmarks for chemistry, computer science, medicine, and mathematics, which evaluate whether scientific fine-tuning transfers to specialized fields. Table 8 summarizes all evaluation sets.

For evaluation implementation, we use the same codebase as MegaScience[2], which standardizes prompting, decoding, and answer extraction across benchmarks. All models are evaluated in the zero-shot setting on 8 GPUs with temperature set to 0, and reported results are avg@3 or avg@5 depending on the benchmark.

E Baseline Data Selection

E.1 Baseline Data Selection Criteria

For baseline comparison, we carefully select the most popular scientific SFT datasets currently available in the open-source community. The selection criteria prioritize datasets that are (1) widely adopted in prior work, (2) publicly accessible, and (3) contain complete CoT reasoning traces suitable for SFT. Based on these criteria, we include SCP-116K (Lu et al., 2025) (we adopt the latest version, which contains 274K

[2] https://github.com/GAIR-NLP/lm-open-science-evaluation

samples; the original authors did not rename the dataset despite the increased size), TextbookReasoning (Fan et al., 2025), NaturalReasoning (Yuan et al., 2026), MegaScience (Fan et al., 2025), and OpenScienceReasoning-2 (NVIDIA, 2025). For all these datasets, we directly use their full original data with the original CoT in our experiments.

E.2 Controlled Comparison with WildSci

WildSci (Liu et al., 2026) is designed for reinforcement learning from verifiable rewards (RLVR). Although it provides verified multiple-choice answers, it lacks the reasoning traces required for SFT, which precludes a direct comparison in Table 2. To enable a more detailed and fair evaluation, we augment WildSci with reasoning responses and present controlled comparisons below.

We first apply n-gram decontamination against all evaluation benchmarks. For each remaining question, Qwen3.5-27B, the answer model used by SPARK, attempts up to four times to generate a valid reasoning response. We retain a response only when its final answer matches WildSci's gold option and discard questions for which no valid response is obtained. We then sample 20K instances from the filtered pool.

We conduct SFT on Qwen3-8B-Base under identical settings, and present the results in Table 9.

Table 9: Spark-234K vs. WildSci SFT comparison.

Training Data	Size	GPQA-M	GPQA-D	S-GPQA	MMLU-Pro	Avg.
WildSci (CoT)	20K	44.66	43.49	37.69	65.05	47.72
Spark-234K (subset)	20K	46.63	45.37	39.75	68.31	50.02

As shown in Table 9, Spark-234K outperforms the SFT-adapted WildSci data on all four overlapping benchmarks, with an average improvement of 2.30 points (50.02 vs. 47.72). Together with the intrinsic comparison presented in the main paper, these results support the benefit of SPARK's skeleton-guided pipeline for question synthesis and answer verification.

Beyond the performance comparison discussed above, our SPARK pipeline also differs from WildSci in several key design aspects:

- **Paper processing.** WildSci processes the full paper directly, whereas SPARK first extracts a paper-level reasoning skeleton and generates questions from the resulting evidence trajectory.
- **Question generation.** WildSci generates three multiple-choice questions per paper. SPARK generates at most one question per perspective and skips perspectives unsupported by the source evidence, adapting its questions to each paper's content.
- **Answer verification.** WildSci uses model voting to identify unanswerable or divergent items, whereas SPARK verifies generated answers against the extracted reasoning skeleton (Section 3.4).

F Pipeline Details

F.1 Model and Decoding Configuration

Table 10 summarizes the API configuration used for skeleton extraction and multi-perspective question generation. Both stages call DeepSeek-V4-Flash with thinking enabled and high reasoning effort.

Table 10: API configuration for skeleton extraction and question generation.

Configuration	Value
Model	DeepSeek-V4-Flash
Thinking	Enabled
Reasoning effort	High

Answer generation and skeleton-grounded quality checking are performed with locally deployed models. Qwen3.5-27B is used to generate answers, while GPT-OSS-120B is used as the quality checker. Table 11 reports the decoding configuration for both stages.

Table 11: Local model configuration for answer generation and skeleton-grounded quality checking.

Configuration	Answer generation	Quality check
Model	Qwen3.5-27B	GPT-OSS-120B
Max new tokens	32,768	16,384
Max input tokens	2048	32,768
Temperature	1.0	0.2
Top-p	0.95	0.95
Top-k	20	20
Min-p	0.0	0.0
Presence penalty	1.5	1.5
Repetition penalty	1.0	1.0

F.2 Prompt Templates

We provide the prompt templates used in the pipeline. Skeleton extraction uses two prompts: Figure 8 localizes the core conclusion from the title and abstract, and Figure 9 reconstructs the claim-conditioned trajectory from the full paper content.

Question generation is implemented as a modular prompt. Figure 11 gives the shared prompt framework, which defines the global objective, self-containment constraints, skip conditions, and output schema. The perspective-specific content is injected through the placeholder `{perspective_guidance}`: at runtime, it is replaced by one of the four definitions in Figure 10. This design keeps the output format and global quality constraints fixed while allowing each generated question to target a different reasoning role, such as mechanism explanation, hypothesis falsification, quantitative derivation, or boundary calibration. Finally, Figure 12 shows the answer-generation prompt, and Figure 13 shows the skeleton-grounded quality-check prompt used to filter QA pairs.

G Source Leakage Patterns

We apply a deterministic source-leakage filter to remove questions that implicitly require access to the original document. The filter consists of 21 case-insensitive regular-expression patterns and is applied to generated questions before they enter the final dataset. Table 12 summarizes the pattern groups. Rather than attempting to judge all forms of self-containment, this filter targets explicit linguistic cues that a question is referring back to the source paper, provided context, or document structure. Any record matching one or more patterns is discarded.

Table 12: Summary of source-leakage pattern groups.

Pattern group	Representative cues
Source-document references	"the paper", "the study", "the authors", "the article", "the passage", "this research"
Provided-context references	"provided text", "provided material", "provided information", "provided context", "according to the provided"
Internal evidence references	"supporting evidence", "supporting facts", "scientific context", "internal context", "internal scientific facts"
Text-reporting constructions	"the text states", "the text shows", "the text reports", "the text demonstrates", and related verb variants
Document-element references	"section N", "figure N", "table N", "reference N", "appendix N"

H Difficulty Taxonomy

We evaluate question difficulty with a five-level reasoning taxonomy. For each dataset, we sample 20K questions and ask DeepSeek-V4-Flash to assign one of the following levels according to the minimum reasoning required to obtain the answer:

- **L1 — Fact Retrieval.** The answer can be directly retrieved as a fact, definition, or given value from knowledge or the source. No reasoning, derivation, or inference is required.
- **L2 — Routine Calculation.** The answer is obtained by applying a known formula or standard procedure and substituting the provided values. No modeling, method selection, or multi-step reasoning is required.
- **L3 — Single-Step Reasoning.** The question requires one non-trivial inference, conceptual link, or causal/mechanistic explanation beyond direct retrieval or mechanical substitution.
- **L4 — Multi-Step Reasoning.** The question requires a chain of reasoning steps, mechanistic analysis, or discrimination among multiple competing explanations or hypotheses.
- **L5 — Research-Level Reasoning.** The question requires constructing a model from principles, synthesizing multiple pieces of evidence, carrying out a derivation, or analyzing boundary conditions, approximations, or counterfactuals.

The resulting labels are used only for aggregate analysis rather than as a training signal. This protocol is intended to compare the reasoning profile of datasets under a consistent rubric, not to provide definitive expert judgments for individual examples.

I Human Validation of Quality Assessments

To validate the reliability of our LLM-based quality assessments, we conduct a stratified human audit of 300 samples, with 30 questions drawn from each of the ten disciplines. Two paid domain experts for each discipline independently evaluate the samples while remaining blind to the original answers and model-assigned labels:

- **Correctness:** Experts independently solve each question, and their adjudicated answer is compared with the original answer in the dataset.

Prompt for Core-Conclusion Localization

```
### Role
You are a senior scientific reviewer identifying the central claim of a research paper.

### Task
Read the title and abstract and articulate the paper's core conclusion. Capture the main finding
     precisely; where a single sentence cannot do it justice, use the supporting fields to record
     the qualifications, scope, or significance that make the claim complete.

### Guidance
- The headline conclusion must be a concrete scientific assertion, not a topic description. Prefer "
     X increases Y under condition Z" over "this work studies the relationship between X and Y".
- Many findings are only meaningful with their conditions, mechanism, or scope attached. Do not
     amputate these to fit one sentence; place them in the supporting fields instead.
- Record only what the abstract actually commits to. Do not invent results, and do not promote
     previewed methods into findings.
- If there is no clear scientific argument (e.g. an editorial, news piece, abstract collection, or
     purely descriptive catalog without a central thesis), output an empty JSON object {{}}.

### Output Schema
Return ONLY a JSON object with this exact structure:
{{
  "core_conclusion": "The single most central finding, stated as one concrete sentence.",
  "key_qualifications": ["A condition, scope limit, or boundary under which the conclusion holds"],
  "significance": "Why this finding matters or what it resolves, in one sentence (empty string if
     the abstract does not state it)."
}}

### Critical Rules
- core_conclusion must be one substantive, self-contained sentence naming the specific subject and
     the asserted effect or mechanism. Keep it focused; offload conditions and scope to
     key_qualifications rather than overloading this sentence.
- key_qualifications captures the decisive conditions, parameters, or limits that the conclusion
     depends on. Use an empty list only when the finding is genuinely unconditional.
- significance is a single sentence and may be an empty string if not stated.
- If the conclusion is vague, hedged beyond recovery, or absent, output {{}}.
- Output ONLY the raw JSON object. Do NOT wrap it in markdown formatting or ```json code fences. No
     explanatory prose.
- Do not refer to "the paper", "the authors", "the study", sections, figures, or tables.

### Title
{paper_title}

### Abstract
{paper_abstract}
```

Figure 8: Prompt for Core-Conclusion Localization.

- **Self-containment:** Experts judge whether the question contains all information needed to be answered independently.

- **Difficulty:** Experts assign one of the five reasoning levels (L1–L5) defined in Appendix H.

Table 13: Human validation results for correctness, self-containment, and difficulty.

Dimension	HA	Agreement	IAA
Correctness	NA	90.3% (271/300)	96.0% (288/300)
Self-containment	98.0% (294/300)	97.7% (293/300)	99.3% (298/300)
Difficulty (L4/L5)	87.0% (261/300)	94.3% (283/300)	91.0% (273/300)

Disagreements are resolved through discussion. Table 13 reports three metrics for each dimension: *Human Assessment* (HA) reports the proportions judged self-contained and assigned L4/L5; it is not applicable to correctness because correctness is evaluated through open-ended problem solving. *Agreement with Original Dataset Result* (Agreement) compares the adjudicated human labels with the original LLM-assigned labels (or the independently derived expert answers with the original answers for correctness). *Inter-Annotator Agreement* (IAA) is computed before adjudication.

The 90.3% agreement with independently derived expert answers supports the reliability of the original answers. Human experts also judge 98.0% of the questions to be self-contained and 87.0% to be L4/L5. Their agreement with the original LLM-assigned labels reaches 97.7% and 94.3%, respectively, supporting the reliability of the automatic assessments. Difficulty is inherently more subjective, as also reflected by its lower inter-annotator agreement; nevertheless, the human results confirm that the dataset predominantly contains challenging questions.

J Diversity Metric Calculations

To rigorously evaluate the semantic span of Spark-234K and baseline data, we use two complementary metrics: Vendi Score (Friedman and Dieng, 2022) and Centroid Distance (Suwanda et al., 2020). Let $X = \{x_1, x_2, \ldots, x_N\}$ be the set of embeddings for the instructions or questions in a dataset, where N is the sample size. All embeddings are computed with `Qwen3-Embedding-8B` (Zhang et al., 2025), and the embedding dimensionality is fixed to 4096 for all samples. Unless otherwise specified, distance is measured by cosine distance, $d(x_i, x_j) = 1 - \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|}$.

Vendi Score. The Vendi Score measures intrinsic diversity by interpreting a dataset's semantic spread as the effective number of independent modes. It is calculated from the eigenvalues of a kernel matrix K, where $K_{ij} = k(x_i, x_j)$ denotes the similarity between samples. We use the cosine-similarity kernel. The Vendi Score is defined as the exponential of the Shannon entropy of the normalized eigenvalues:

$$\text{Vendi}(X) = \exp\left(-\sum_{i=1}^{N} \lambda_i \ln \lambda_i\right), \quad (1)$$

where $\lambda_1, \ldots, \lambda_N$ are the normalized eigenvalues of K/N. A higher Vendi Score indicates that the dataset contains a larger number of effective independent semantic clusters. In practice, we compute Vendi Score using the reference implementation released by the original authors.[3]

Centroid Distance. Centroid Distance measures geometric dispersion in the embedding space. We first compute the global centroid μ of the dataset:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i. \quad (2)$$

Centroid Distance is then defined as the complement of the average cosine similarity between each sample x_i and the centroid μ:

$$\text{Dist}_{\text{cent}}(X) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{x_i \cdot \mu}{\|x_i\|\|\mu\|}. \quad (3)$$

A higher Centroid Distance indicates that samples are more widely dispersed around the dataset center, covering a broader semantic region rather than clustering tightly around a narrow set of topics.

[3] https://github.com/vertaix/Vendi-Score

K Self-Containment Bad Cases

Figure 14 shows representative self-containment failures observed in baseline data. These examples illustrate why a fluent question can still be unusable for standalone reasoning: it may refer to missing definitions, omitted equations, or unstated variables.

L Qualitative Examples

To provide qualitative insight into the output of our automated data construction pipeline, we present examples of a generated paper skeleton and corresponding generated questions and answers in Figure 15. The figure illustrates the clear structure extracted for a sample paper, highlighting key entities, processes, and findings. Linked to specific points in this skeleton are sequential examples of different question types—entity-based, procedural, causal, and more—each accompanied by a grounded answer derived directly from the skeleton content. Subtle visual links demonstrate the grounding and diversity of the generated data, ensuring a robust and factually anchored fine-tuning dataset for scientific reasoning.

M Use of AI Assistants

We used a large language model only for light writing assistance, limited to grammar checking and minor clarity edits on a small number of sentences. All technical content, experiments, and analyses in this paper were written by the authors.

Table 8: Benchmark details for our evaluation.

Benchmark	Size	Report	Category	Description
General science reasoning benchmarks				
GPQA-Main (Rein et al., 2024)	448	avg@5	Expert science QA	Graduate-level, Google-proof multiple-choice questions across biology, chemistry, and physics.
GPQA-Diamond (Rein et al., 2024)	198	avg@5	Expert science QA	The higher-confidence GPQA subset selected for expert agreement and difficulty.
SuperGPQA (Du et al., 2026)	26,529	avg@3	Broad graduate science	Large-scale graduate-level benchmark spanning many disciplines and subfields.
SciBench (Wang et al., 2024a)	640	avg@5	Quantitative science	Aggregated evaluation over the tested chemistry, mathematics, and physics subsets.
MMLU (Hendrycks et al., 2020)	14,042	avg@3	Broad knowledge	Multitask benchmark covering broad academic and professional subjects, including science.
MMLU-Pro (Wang et al., 2024b)	12,032	avg@3	Broad knowledge	More challenging multitask benchmark with harder options and stronger reasoning demands.
Domain-specific benchmarks				
ChemBench (Mirza et al., 2024)	2,788	avg@5	Chemistry	Aggregated evaluation over the tested multiple-choice and string-match subsets.
CS-Bench (Song et al., 2025)	1,778	avg@5	Computer science	Aggregated evaluation over the tested assertion and multiple-choice subsets.
PubMedQA (Jin et al., 2019)	500	avg@5	Medicine	Biomedical question answering over PubMed-derived contexts.
MedQA-US (Yao et al., 2026)	1,273	avg@5	Medicine	Medical licensing-style multiple-choice questions from the US subset.
GSM8K (Cobbe et al., 2021)	1,319	avg@5	Mathematics	Grade-school math word problems requiring multi-step arithmetic reasoning.
MATH-500 (Hendrycks et al., 2021)	500	avg@5	Mathematics	A 500-problem subset of competition-style mathematical reasoning problems.

Prompt for Claim-Conditioned Trajectory Extraction

```
### Role
You are a senior scientific reviewer reconstructing the logical structure that supports a paper's
     stated conclusion.

### Task
The core conclusion has already been established. Treat it as the fixed endpoint of the argument.
     Your work has two parts:
1. Read the full content and reconstruct the step-by-step argumentation chain that leads to this
     conclusion, identifying key logical steps (not sections) and focusing on how the conclusion is
     reached through experiments, observations, derivations, or reasoning.
2. Judge which training perspectives the reconstructed trajectory genuinely supports, using the
     definitions below.

### Core Conclusion (fixed target)
{core_conclusion}

### Training Perspectives
Match perspectives to the nature of the trajectory itself:
- mechanistic_reasoning -> the work centers on a new mechanism or causal account of why/how
     something happens.
- hypothesis_falsification -> the work hinges on testing competing explanations and ruling some out.
- quantitative_derivation -> the work is computation-heavy, rich in values, units, and formulas.
- boundary_calibration -> the work is careful about what its evidence does and does not justify.

### Critical Rules
- steps must capture the logical flow from evidence to the given conclusion. Do NOT produce a
     section-by-section summary.
- Each step must describe one logical unit of the argument (an experiment that tests a hypothesis, a
      comparison that rules out an alternative, a derivation that connects observations to a
     mechanism, etc.).
- The "description" MUST explicitly mention the specific experimental conditions, chemical names,
     mathematical operators, or equipment setups used. Avoid generic phrasing like "an analysis was
     conducted".
- The "observables" MUST contain concrete anchors: exact numerical values with units, specific
     formulas/variables, statistical bounds, key entities, or distinct phenomenological signs. Never
      write placeholder text like "did an experiment and got a result".
- recommended_perspectives: list 1-4 perspectives from the four above (exact strings), ordered best-
     fitting first, picking those that match the character of this trajectory. Do not default to a
     fixed mix.
- If the full content does not actually support a coherent chain toward the given conclusion, output
      an empty JSON object {{}}.
- Output ONLY the raw JSON object. Do NOT wrap it in markdown formatting or ```json code fences. No
     explanatory prose.
- Do not refer to "the paper", "the authors", "the study", sections, figures, or tables.

### Paper Content
{paper_content}

### Output Schema
Return ONLY a JSON object with this exact structure:
{{
  "steps": [
    {{
      "step_number": 1,
      "description": "What this step establishes or argues, in clear scientific terms.",
      "observables": ["Key measurement, observation, numerical value, or fact supporting this step"]
    }}
  ],
  "recommended_perspectives": [
    "mechanistic_reasoning",
    "hypothesis_falsification"
  ]
}}
```

Figure 9: Prompt for Claim-Conditioned Trajectory Extraction.

Perspective Guidance for Question Generation

```
[mechanistic_reasoning]
Target: Probe the reasoning link between a condition, intervention, or structure and an observed
     phenomenon. Frame this link as a universal scientific problem, abstracting away paper-specific
     identifiers so that solving it requires general mechanistic reasoning rather than reading
     comprehension.
Style: Answerable in 2-5 paragraphs of mechanistic reasoning without extensive math. Focus entirely
     on the system/phenomenon itself.
Construction:
- State the setup using generic descriptors (e.g. "a two-component system under periodic forcing").
- The expected answer should be a chain of mechanistic steps, each justified by a general principle,
      not a single recalled fact.
Perspective Guardrails:
- Anti-Stipulation: Do NOT stipulate a counterintuitive finding and ask "why". The solver must
     engage via general principles to deduce the outcome.
- Anti-Comprehension: Must not be a mere reading-comprehension question. If it requires having read
     the specific paper to start reasoning, reject it.
- Single-Mechanism Focus: Center one dominant causal mechanism. Do not bundle several loosely
     related effects into one prompt.

[hypothesis_falsification]
Target: Test whether the evidence rules out plausible but competing explanatory models, forcing the
     solver to identify which single model the evidence is consistent with.
Style: MCQ format. The `question` string MUST end with four options labeled (A), (B), (C), (D)
     separated by newlines. Do NOT indicate the correct answer in the question.
Construction:
- Present a concrete body of evidence in the stem, then offer four candidate explanations.
- Each distractor must be defeated by a DIFFERENT, identifiable piece of the evidence, so
     elimination requires genuine discrimination rather than plausibility ranking.
Perspective Guardrails:
- Structural Rivalry: Options must be structurally different explanatory models or historical
     misconceptions, NOT paraphrases or lexical variants of the same answer.
- Lexical Decoupling: Avoid technical terms in the stem that act as "giveaways" via direct
     vocabulary matching with the correct option. The correct option must not be identifiable by
     surface word overlap alone.
- Balanced Distractors: Keep options comparable in length, specificity, and technical register so
     none stands out by form rather than content.

[quantitative_derivation]
Target: Convert numerical observables into a modeling-first problem. The solver must formulate the
     relevant physical, chemical, or statistical relations and derive an analytical expression,
     bound, scaling relation, or key calculated quantity.
Style: A rigorous graduate-level theoretical problem. State knowns objectively and demand derivation
     .
Construction:
- Provide only the observables genuinely needed; let the solver decide which principles couple them.
- Ask for a symbolic result first (expression, bound, or scaling), with any numerical evaluation as
     a final, minor step.
Perspective Guardrails:
- Modeling-First: The intended difficulty MUST lie in constructing the model and connecting
     principles, NOT in tedious arithmetic or plug-and-chug substitution.
- Anti-Fragmentation: Focus on ONE core modeling challenge. Avoid sub-questions that are purely
     arithmetic continuations.
- No Over-Specification: Do NOT pre-state the governing equation, the regime, or the solution method
     . Handing the solver the premise removes the modeling step that defines this perspective.

[boundary_calibration]
Target: Target the qualifiers, assumptions, parameter regimes, or confounders under which the claim
     would no longer be justified.
Style: Concise critique targeting a specific assumption/approximation breakdown (e.g., "At what
     parameter regime does this approximation break down?").
Construction:
- Anchor on one specific claim, then ask where its supporting evidence stops holding.
- Frame the answer around a concrete switch: a parameter crossing a threshold, an assumption being
     violated, or a confounder becoming non-negligible.
Perspective Guardrails:
- Sharp Critique: Point to a SPECIFIC, scientifically reachable approximation, regime, or confounder
     . Do not ask vague "what could go wrong" questions.
- No Sci-Fi: The boundary must be physically or operationally reachable, not an absurd hypothetical
     (e.g., halving gravity).
- Justified-vs-Unjustified: The item should hinge on distinguishing what the evidence does support
     from what it does not, not on listing every conceivable caveat.
```

Figure 10: Perspective Guidance for Question Generation.

Shared Prompt for Multi-Perspective Question Generation

```
### Role & Objective
You are an expert data architect crafting high-quality Supervised Fine-Tuning data to train next-
     generation reasoning models.
Distill a PhD-level scientific problem from the 'Paper Trajectory' based on the requested
     perspective. The item must function as a rigorous logic stress test requiring complex, multi-
     step deduction.

### Input Context
- Requested Perspective: {requested_perspective}

### Perspective Focus
{perspective_guidance}

### Global Guardrails & Structural Bans (CRITICAL)
To prevent dataset mode collapse, enforce these rules strictly:
1. Universal Abstraction (No Source Referencing): The problem MUST be entirely self-contained. NEVER
      use labels like "In this study," "the authors," "the trajectory," or result-reporting verbs ("
     found," "measured").
2. No Hyperlocal Naming: Replace sample-specific IDs, specific instrument names, or isolated
     regional names with generic scientific descriptors.
3. No Meta/Lazy Evaluations: NEVER end with meta-questions (e.g., "Is this justified?", "What
     remains unknown?").
4. Target Difficulty: Difficulty must stem from intricate mechanistic chains, subtle tradeoffs, or
     advanced modeling rather than niche trivia.

### Global Skip Conditions
Output EXACTLY `[]` (an empty array) with no other text if ANY of the following apply:
- The trajectory lacks the fundamental depth, alternatives, or theoretical bounds required for the
     `{requested_perspective}`.
- The core conclusion is just a descriptive observation/high-level summary with no "black box" left
     to interrogate.
- The question cannot be made self-contained without relying on paper-specific data.
- The question degrades into a trivial textbook variant of a standard field problem.

### Output Schema
[
  {{
    "question": "The fully formulated scientific problem (append (A)-(D) options here if MCQ).",
    "evidence": ["Concise scientific fact 1 from trajectory", "Fact 2"],
    "solvability_justification": "Rigorous proof the question is well-posed. For open math: explain
     theoretical steps. For MCQs: state correct option AND detailed elimination logic for
     distractors."
  }}
]

### Paper Trajectory
{trajectory_text}
```

Figure 11: Shared Prompt for Multi-Perspective Question Generation.

Prompt for Answer Generation

```
{question}
Solve the above problem step by step.
```

Figure 12: Prompt for Answer Generation.

Prompt for Skeleton-Grounded Quality Checking

```
### Role & Objective
You are a scientific QA auditor. Rule-based checks already catch format failures; your job is the
     part rules cannot decide: whether this QA pair is scientifically grounded in the paper
     trajectory. Judge it for reasoning-model SFT training and return a single KEEP/DISCARD decision
     .

### What You Receive
- A Paper Trajectory: the extracted core conclusion (the scientific direction) plus the step-by-step
      evidence chain (the evidence anchors).
- A Generated Question and a Generated Answer to audit against that trajectory.

### Audit Dimensions

#### Dimension 1: Question
1. Self-Containment
The question must provide the conditions, variables, and context needed to be answered WITHOUT the
     original paper. All custom variables, experimental settings, and boundary conditions must be
     stated. Common universal scientific constants need not be restated. Flag any question that
     silently depends on unstated paper-specific content.

2. Source Match
The question must correspond to what the trajectory actually establishes. Flag a source mismatch
     when the question concerns an entity, quantity, or claim the trajectory does not support, or
     fabricates premises absent from it.

#### Dimension 2: Answer
1. Reasoning Coherence
The reasoning chain must be coherent and complete: no logical jumps, no meaningless repetition or
     loops, no interrupted or truncated output. The final answer must be definite and well-organized
     .
*Rigorous self-correction, re-evaluation of earlier steps, and explicitly resolving uncertainty
     within the text are valuable and fully permitted.*

2. Trajectory Grounding
The answer must be consistent with the scientific direction of the trajectory's core conclusion AND
     not be contradicted by its evidence chain (steps and observables). Equivalent academic
     statements, synonymous expressions, and reasonable paraphrased reasoning are acceptable; only
     essential conflicts with the established direction or evidence are disqualifying.

### Hard Elimination Rules
1. Dependency Defect: Missing essential custom variables; unanswerable without extra paper context.
     -> DISCARD
2. Source Mismatch: Question targets claims/entities the trajectory does not support, or invents
     absent premises. -> DISCARD
3. Generation Defect: Incoherent reasoning -> meaningless loops, arbitrary guessing without
     deduction, truncated output, or serious hallucination. -> DISCARD
4. Grounding Conflict: Answer's core direction is inconsistent with the trajectory's conclusion, or
     is contradicted by its evidence chain. -> DISCARD

Supplementary Rule: Standard multi-step inference, causal analysis, and logical self-correction all
     count as qualified responses.

### Output Specification
Output a single, strictly valid JSON object. Do NOT add any extra symbols, markdown formatting
     outside the JSON, or conversational text. Use the exact schema below:

{{
  "reasoning_steps": "First, write your thorough step-by-step reasoning and analysis here.",
  "checks": {{
    "is_question_self_contained": true,
    "is_question_source_matched": true,
    "is_answer_coherent": true,
    "is_answer_trajectory_grounded": true
  }},
  "fatal_flaw_details": "Concrete flaw reason based on the Hard Elimination Rules, or 'none'.",
  "final_decision": "KEEP"
}}
// Note: "final_decision" must be strictly either "KEEP" or "DISCARD".

### Paper Trajectory
{trajectory_text}

### Generated Question
{question}

### Generated Answer
{answer}
```

Figure 13: Prompt for Skeleton-Grounded Quality Checking.

Self-Containment Bad Cases from Baseline Data

MegaScience.
Case 1: Missing definitions.

> Show that if a binary sequence $\langle X_n \rangle$ is R5-random (as per Definition R5), and if $\langle s_n \rangle$ is any computable sequence as in Definition R4, then $\overline{\Pr}(X_{s_n} = 1) \geq \frac{1}{2}$ and $\Pr(X_{s_n} = 1) \leq \frac{1}{2}$.

Reason: The question refers to Definition R5 and Definition R4, but the definitions are not provided. A solver cannot independently interpret the assumptions.

Case 2: Missing referenced equation.

> Prove why the last equality holds true for a function in the Dirichlet space.

Reason: The phrase "the last equality" refers to an omitted equation, so the target statement to be proved is unavailable.

Case 3: Missing required variable.

> A roller coaster is moving at an initial velocity of 10 m/s and comes to a stop after a distance of 50 m. Calculate the average force supplied by the brakes to stop the roller coaster. Assume a constant acceleration and neglect any frictional forces. Use the following equations: $V_1 = 10$ m/s, $\Delta D = 50$ m, $V_2 = 0$ m/s, $F_{\text{net}} = F_{\text{brakes}} = ma$.

Reason: The mass of the roller coaster is not specified, so the average braking force cannot be determined.

SCP-116K.
Case 4: Missing circuit diagram.

> Two AC signals, V_1 and V_2, are to be combined such that $V_{\text{out}} = \frac{3}{2}V_2 - \frac{5}{2}V_1$. The subtracting amplifier circuit shown is used. What must be the values of R_1, R_2, R_3, and R_4? Circuit diagram: V_1 connected via R_1; V_2 connected via R_2; operational amplifier with feedback resistance R_3; ground connection via R_4. (Professional Publications, Inc.)

Reason: The question refers to an external circuit diagram. Without the actual diagram, the resistor topology and value relationships are under-specified.

Case 5: Missing section dependency.

> Using the results of Section 7.5, find the transformation $x = Tw$ that puts the system $\frac{dx}{dt} = \begin{pmatrix} 1 & -1 & 2 \\ 0 & -2 & 1 \\ -1 & 3 & 0 \end{pmatrix} x + \begin{pmatrix} 1 & 3 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} u$ into the form $\frac{dw}{dt} = \begin{pmatrix} 0 & -3 & 6 \\ -1 & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix} w + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} u$.

Reason: The question depends on results from Section 7.5, but those results are not provided. A solver cannot recover the required transformation rule from the question alone.

Case 6: Missing referenced formulas.

> Calculate the Hilbert transforms of $\cos(\omega_0 t)$ and $\sin(\omega_0 t)$. Hint: Use Euler's identity in the defining integral (7.10.12) for the Hilbert transform of a time signal, and you may also find that recalling (7.9.3) is helpful.

Reason: The prompt relies on formula numbers (7.10.12) and (7.9.3), but the formulas themselves are omitted.

NaturalReasoning.
Case 7: Undefined property.

> What should be the cardinality of A_n to ensure that the subset of infinite binary sequences having a certain property is countable/uncountable, considering the set R_n of binary sequences of length n and the subset A_n of sequences having that property?

Reason: The question never defines "a certain property", so the solver cannot determine which sequences belong to A_n or what condition is being analyzed.

Case 8: Missing humidity information.

> Given the temperature of 18 degrees Celsius and the atmospheric pressure of 782 mm Hg, calculate the air density using the ideal gas law equation, considering the effects of humidity. Show your work and explain your reasoning.

Reason: The question asks the solver to account for humidity but provides no relative humidity, water-vapor partial pressure, or equivalent humidity-related quantity.

Figure 14: Representative self-containment failures from baseline data. Each example appears plausible but lacks information required for standalone solution.

Dataset Synthesis: From Structured Extraction to Deep Reasoning QA

Phase 1: Extracted Paper Framework
Core Conclusion: A tail wave formed behind density cusps in nonlinear laser wakefields driven by a Gaussian laser pulse provides simultaneous focusing and acceleration for externally injected positrons, enabling nearly 100% trapping efficiency with terawatt-class laser systems.

- **Step 1: Defocusing mechanism & Tail wave proposal.**
  - *Description:* Establish that nonlinear wakefields naturally defocus positively charged particles. Propose that a tail wave forms behind density cusps when laser intensity marginally exceeds the nonlinear threshold, converging plasma electrons into an on-axis filament.
  - *Observables:* $a_0 = 1.6$, $w_0 = 12\lambda_0$, peak intensity 3.5×10^{18} W/cm^2. Regime thresholds based on $n(0)/n_{e0}$: I (≥ 0.2), II (0.1–0.2), III (≤ 0.1).
- **Step 2: Unloaded dynamics.**
  - *Description:* Using 3D PIC simulations (OSIRIS), demonstrate tail wave dynamics without positron loading. The focusing region for positrons expands from a limited cusp to about half a bubble as laser intensity increases.
  - *Observables:* Grid $200 \times 200 \times 2400$. Transverse field maps show focusing region expanding.
- **Step 3: Beam loading & Field reversal.**
  - *Description:* Load a uniform positron beam into the second bubble. The beam's space-charge attracts tail wave electrons, converting a defocusing transverse field into a focusing field. However, tail wave III causes transverse loss.
  - *Observables:* Positron $n_{+0} = 0.5n_{e0}$, initial energy 20 MeV. Trapping efficiency $\sim 60\%$ due to tail wave III.
- **Step 4: Channel optimization.**
  - *Description:* Increase the plasma channel radius to weaken relativistic self-focusing and suppress tail wave III, maintaining the required on-axis electron density throughout propagation.
  - *Observables:* Channel radius $r_0 = 14\lambda_0 > w_0$. Density $n(0)/n_{e0} \geq 0.1$ (avoiding regime III). Trapping efficiency $> 95\%$ after $1000\lambda_0$.
- **Step 5: Beam evolution.**
  - *Description:* Characterize the accelerated positron beam, showing reduced transverse size, limited emittance growth, and stable longitudinal acceleration.
  - *Observables:* Transverse size reduced to $\sim 1/3$. Acceleration gradient ~ 100 GV/m. Final energy peak at 177 MeV (average gain 110 MeV) with 3% FWHM energy spread.
- **Step 6: Parameter robustness.**
  - *Description:* Vary drive laser parameters and beam offsets to map the viability of the scheme.
  - *Observables:* $a_0 \in [1.2, 1.8]$, $w_0 \in [8\lambda_0, 12\lambda_0]$ for high trapping. Longitudinal offset tolerance within $2\lambda_0$; transverse tolerance $\sim 0.5\lambda_0$.

Phase 2: Synthesized SPARK-QA

In a laser-driven plasma wakefield used for positron acceleration, the transverse field that a witness positron beam will experience is predicted from a simulation of the *bare* wake—one run without the beam present (an "unloaded" field map). A positron beam of density $n_+ = 0.5\, n_{e0}$ (n_{e0} = ambient plasma density) is then injected into an accelerating bucket of this wake, and its focusing or defocusing is read off from the sign of the unloaded transverse field at the beam's location. Which of the following is the most fundamental limitation of using the unloaded field map to predict the focusing experienced by the *loaded* beam?

(A) As the drive laser propagates, relativistic self-focusing reshapes the pulse and therefore the wake's longitudinal profile, so a field map taken at one propagation distance no longer applies further downstream.

(B) The guiding plasma channel imposes a transverse density gradient that the bare-wake simulation does not reproduce, so the off-axis transverse field is mis-estimated across the beam's radial extent.

(C) The positron beam's own space charge draws plasma electrons toward the axis, adding on-axis negative charge that the unloaded map omits; this can convert a region that is defocusing for positrons in the bare wake into a focusing one.

(D) The transverse field varies strongly with longitudinal phase within a single bucket, so a map read at one slice of the bucket mis-predicts the field at the precise phase the beam occupies.

Figure 15: Dataset synthesis pipeline. Phase 1 details the extraction of a logically closed paper framework, structurally decoupling qualitative descriptions from quantitative observables. Phase 2 leverages this structured context to generate a self-contained, expert-level scientific reasoning question.