Open-Sci: Beyond Scaling with High-Density Datasets for Precise Scientific Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have made remarkable progress on reasoning tasks such as coding and mathematics, yet their scientific reasoning ability remains significantly limited, hampered by the scarcity of high-quality scientific datasets. While LLM training has long been guided by scaling laws, attention is shifting toward data quality: high-density supervision complements scale and at the margin can outperform noisier corpora. Constructing such datasets, however, remains challenging. Existing efforts either rely on LLM-generated synthetic data prone to hallucination and logical inconsistency, or on manual curation that struggles with scalability and standardization. To overcome these hurdles, we propose **PrecSci**, a systematic data curation pipeline centred on precision and internal consistency. It extracts knowledge from reliable scientific documents, refines questions for completeness and precision, applies multi-stage filtering to eliminate redundancy and noise, and refines answers through a generator and a critic that scores candidates along multi-dimensional quality criteria, retaining only high-quality samples. Leveraging PrecSci, we build *Open-Sci*, a compact corpus of 186k scientifically precise question–answer pairs. Despite being less than one-sixth the size of the state-of-the-art MegaScience dataset ($\sim 14.88\%$), Open-Sci consistently improves LLM performance, lifting Qwen3-8B by $+7.08$ points on average across diverse scientific benchmarks (e.g., PHYSICS, ChemBench, MMLU-Pro).

1 Introduction

Large Language Models (LLMs) (Xiang et al.; El-Kishky et al., 2025) have made substantial progress on reasoning (Chen et al., 2025a), particularly in coding and mathematics (Shao et al., 2024; Ren et al., 2025), driven by Chain-of-Thought prompting (Wei et al., 2022) and reinforcement learning (Guo et al., 2025; Yu et al., 2025). Their scientific reasoning, however, lags far behind (Wang et al., 2023; Xi et al., 2025) and struggles to apply principles in physics, chemistry or biology, limiting use in scientific research. The primary bottleneck is the scarcity of effective scientific reasoning datasets, which are costly to construct because of verification overhead, complex annotation, and the inherent ambiguity of scientific answers.

To address the scarcity of scientific reasoning data, prior efforts can be categorized into two classes: (1) **Leveraging LLMs to generate synthetic data**, either by directly querying them for question-answer pairs (Wan et al., 2024) or providing reference papers for reasoning data generation (Chen et al., 2025b). While capable of scaling data size rapidly, these approaches often suffer from significant quality issues, including noisy, error-ridden content, susceptibility to model biases, and the risk of hallucinations. Given the stringent demand for precision in the scientific field, these flaws are particularly detrimental in scientific reasoning tasks, where minor inaccuracies can disrupt LLM training. (2) **Extracting human-compiled question-answer pairs** (Saikh et al., 2022; Singh et al., 2024) from scientific documents, *e.g.*, textbooks, examinations, and competitions. While ensuring higher correctness, these methods face data scarcity issues, as existing sources cover limited scenarios and often feature non-standardized questions with unclear requirements and incomplete reasoning in answers, potentially reducing the task to mere factual recall instead of deep scientific reasoning.

We argue that progress in scientific reasoning depends on raising the information density and supervision precision of training data, not on unconstrained scaling. Knowledge drawn from original scientific documents preserves authenticity, while formalized questions and consistent multi-step answers shape the supervision signal toward genuine

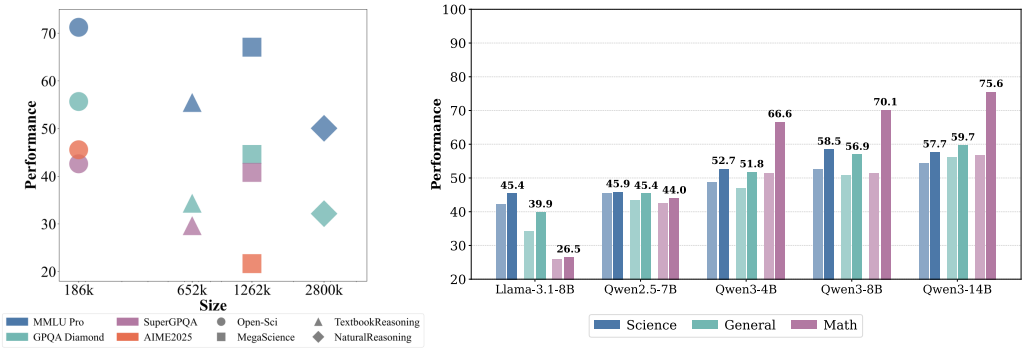


Figure 1: *Left*: Dataset size versus benchmark performance. Each point is a dataset–benchmark pair; colors mark benchmarks and shapes mark datasets. At 185.8k samples, Open-Sci matches or surpasses much larger datasets. *Right*: Per-domain performance: lighter bars are base models, darker bars are the same models fine-tuned on Open-Sci. Open-Sci improves performance across every domain and model.

reasoning rather than factual recall. A compact, high-density dataset can therefore reduce the demand for massive scaling while substantially improving scientific reasoning performance.

In this paper, we introduce *PrecSci*, a pipeline that constructs multi-disciplinary scientific reasoning datasets through four stages: (1) Knowledge Extraction from authoritative sources such as textbooks and standardized examinations; (2) Question Formalization, which standardizes raw problems for completeness and clarity; (3) Noise Mitigation through deduplication, decontamination, and cascaded filtering; and (4) Generative-Discriminative Dual Refinement, which couples a strong generator with a critic that scores reasoning chains along multi-dimensional quality criteria. Using this pipeline, we construct *Open-Sci*, a precise and knowledge-dense dataset of 186k scientifically rigorous question–answer pairs. Despite occupying only $\sim 14.88\%$ of the size of the state-of-the-art MegaScience dataset, *Open-Sci* consistently achieves superior performance, yielding an average improvement of $+7.08$ on Qwen3-8B across diverse scientific reasoning benchmarks, including MMLU-Pro, PHYSICS, ChemBench and ProteinLMBench, supporting our hypothesis that precision-first curation can match or surpass substantially larger but noisier corpora.

Our key contributions are summarized as follows:

- We design *PrecSci*, a comprehensive four-stage pipeline that integrates authoritative source collection, question formalization, rigorous noise mitigation, and generative-discriminative dual refinement.
- Using PrecSci, we construct *Open-Sci*, a compact 186k corpus spanning four scientific disciplines. Despite its modest scale, the dataset delivers exceptional per-sample efficiency through systematic precision-oriented curation.
- Despite using only $\sim 14.88\%$ the size of MegaScience, Open-Sci surpasses the MegaScience baseline by an average of $+\mathbf{7.08}$ on scientific benchmarks (e.g., PHYSICS, ChemBench), with additional transfer gains on general and mathematical tasks.

2 Related Work

Scientific Datasets. Scientific datasets have evolved from narrow biomedical resources, such as BioASQ (Nentidis et al., 2022) and PubMedQA (Jin et al., 2019), to cross-domain benchmarks. ScienceQA (Saikh et al., 2022) added chain-of-thought annotations for reasoning, and SciBench (Wang et al., 2023) exposed multi-step reasoning failures on university-level problems in physics, chemistry and mathematics. More recently, TextbookReasoning (Fan et al., 2025) prioritises high-quality sources while NaturalReasoning (Yuan et al., 2025) pursues large-scale expansion via automatic question generation; the former is disciplinarily imbalanced, and the latter suffers from noise and weak difficulty control. Balancing disciplinary coverage with high-quality, reasoning-rich annotations therefore remains an open challenge.

Scientific Models. Early scientific LLMs specialised in single disciplines: BioGPT (Luo et al.,

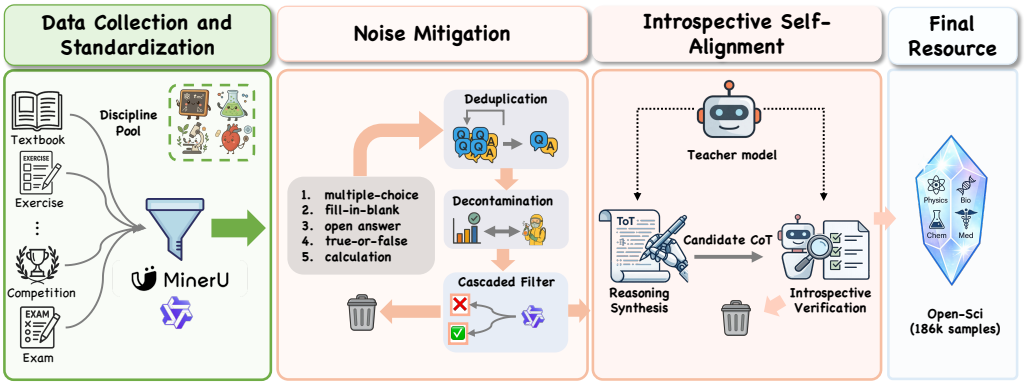


Figure 2: Overview of the PrecSci pipeline for constructing the Open-Sci dataset. The pipeline proceeds through four main stages: (1) collecting data from authoritative sources, (2) extracting and formalizing question–answer pairs, (3) mitigating noise through deduplication, decontamination, and cascaded filtering, and (4) performing generative-discriminative dual refinement, where a generator synthesizes reasoning chains and a critic scores them against multi-dimensional quality criteria. The result is a compact, scientifically rigorous dataset of 185k instances optimized for advanced reasoning.

Table 1: Comparison of representative scientific reasoning datasets. Source: **D**=digital (web / existing datasets), **T**=textbooks, **E**=exams or competitions. **Disc. Ann.**: discipline-level annotations (✓/✗/Partial). **Ans. Len.**: average answer length in tokens. **Open Scope**: ☆=Data, ★=Data+Model, ◆=Data+Model+Pipeline. For Open-Sci, "4(47)" means 4 disciplines and 47 sub-disciplines.

Dataset	Src	Num	Non-math	Ans. Len.	#Disc.	D. Ann.	Open
WebInstruct (Yue et al., 2024)	D	13M	32%	266.4	8	✗	★
NaturalReasoning (Yuan et al., 2025)	D	2.8M	$< 55\%$	766.3	16	✗	☆
Nemotron-Sci. (Bercovich et al., 2025)	D	708.9k	100%	1716.5	1	✗	★
TextbookReason. (Fan et al., 2025)	T	652K	35%	409.7	7	✓	◆
MegaScience (Fan et al., 2025)	D,T	1.25M	-	692.9	7	Partial	◆
Open-Sci (ours)	D,T,E	186k	100%	1122.3	4(47)	✓	◆

2022) and PubMedGPT (Bolton et al., 2024) excelled on biomedical tasks (Jin et al., 2019) but lacked interdisciplinary breadth. Galactica (Taylor et al., 2022) broadened the scope by pre-training on scientific corpora spanning medicine, chemistry and mathematics, yet remained limited in long-chain reasoning. Recent models combine general reasoning with domain expertise: Intern-S1 (Bai et al., 2025) mixes a trillion-token scientific corpus with reinforcement learning to match closed-source models across physics, chemistry, biology and medicine, and LLaMA-Nemotron (Bercovich et al., 2025) attains state-of-the-art results on GPQA (Rein et al., 2024) through reinforcement-tuned reasoning.

3 PrecSci

Scientific reasoning requires precise and high-quality data. Large but noisy corpora often introduce ambiguity and inconsistency, which limit the development of reliable reasoning ability. To overcome this, we design PrecSci, a systematic data curation pipeline built for precision and internal consistency rather than scale. The overview of our pipeline is presented in Figure 2. The process begins with Authoritative Knowledge Extraction to ensure source reliability (Section 3.1) , followed by Question Formalization to improve semantic clarity (Section 3.2). We then perform Multi-stage Noise Mitigation to eliminate redundancy and contamination (Section 3.3) , and finally employ a Generative-Discriminative Dual Refinement mechanism to ensure the logical rigor of reasoning chains (Section 3.4). This closed-loop pipeline results in the Open-Sci dataset, a high-density resource optimized for scientific tasks (Section 3.5). More details are provided in Appendix C

3.1 Authoritative Sources

To construct a dataset emphasizing scientific precision over sheer scale, our pipeline begins with authoritative sources, including textbooks, standardized exams, research exercises, and subject-specific competitions. Textbooks provide systematically organized, expert-authored knowledge; exams and exercises supply well-structured problems that test conceptual accuracy; and competitions introduce advanced challenges requiring creativity and rigor. Drawing on these sources ensures collected data is scientifically reliable, accurate, and inherently high-quality, laying a solid basis for further curation. For PDF-only materials, we employed MinerU (Wang et al., 2024a). For raw text, we directly incorporated it into the pipeline. This dual strategy ensured high recall across heteroge-

neous data formats. To extract question–answer (QA) pairs, we used Qwen2.5-72B (Qwen et al., 2024) with carefully designed extraction prompts to identify QA pairs. This approach differs fundamentally from the reference-driven QA generation strategies, where models are asked to generate new questions given reference documents. Our method extracts existing scientifically valid problems that preserve their original rigor. This ensures that the pipeline benefits from the strengths of large models in pattern recognition while avoiding the risks of hallucination or uncontrolled reformulation (Ji et al., 2023). This process further reinforced the precision of data: the outcome is a curated collection of QA data that faithfully reflects authoritative sources and provides a solid basis for the subsequent stages of formalization and consistency.

3.2 Question Formalization

Raw problems, even from authoritative sources, may still suffer from incomplete context or wrong expression. To ensure precise data representation, we avoid generating new questions and instead formalize the originals to preserve their meaning while improving clarity. Questions are categorized into five types, *i.e.*, multiple-choice, fill-in-the-blank, short-answer, true-or-false, and calculation, using Qwen2.5-72B. We then formalize questions through type-specific prompts with DeepSeek V3-235B (Liu et al., 2024): clarifying options in multiple-choice, ensuring complete context in fill-in-the-blank, and enforcing numerical accuracy in calculation. For short-answer questions, we added minimal clarifications to guarantee that the prompt provides sufficient context for a unique and scientifically valid response. For true-or-false questions, we ensured that each statement is expressed in a precise and unambiguous form, avoiding vague qualifiers and enforcing scientific correctness. This formalization produced uniformly formatted, semantically unambiguous, and scientifically precise questions, strengthening the dataset's reliability for advanced reasoning.

3.3 Noise Mitigation

Although collected from authoritative sources, the raw collection still contained substantial redundancy and problematic items. On the one hand, many questions were near-duplicates that differed only in superficial phrasing, which inflate dataset size without adding new supervision value. Some problems are too simple and offer little challenge for training, while others are too difficult for the model to learn. Additionally, some problems have flaws, such as missing essential conditions. Both types of noise undermine effective learning by promoting shortcut memorization instead of genuine reasoning. To reduce such noise, we first applied locality-sensitive min-hashing (Mou et al., 2023) to remove near-duplicates within each discipline. To safeguard evaluation reliability, we performed strict benchmark decontamination by retrieving nearest neighbors via embedding similarity and confirming semantic overlap with LLM (Fan et al., 2025; Yuan et al., 2025), discarding overlapping items. Finally, we enforced scientific precision through cascaded difficulty filtering. Let $\alpha_M(q,a) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[M(q)_i = a]$ be the empirical accuracy of a probing model M on (q,a) over N samples, and let α_s, α_l denote a small and a strong probing model. We retain

$$\mathcal{D}_{\text{clean}} = \{(q,a) \in \mathcal{D}_{\text{decon}} : \alpha_s < \tau_e, \alpha_l > \tau_h\}, \quad (1)$$

where $\alpha_s < \tau_e$ discards trivial items and $\alpha_l > \tau_h$ discards ill-posed ones ($\tau_e = 7/8$, $\tau_h = 0$ in our setup). As shown in Table 2, these rigorous procedures retained only 32.89% of the original collection, highlighting the selectivity of our pipeline.

Table 2: Dataset retention through the PrecSci pipeline. Starting from $594,930$ raw Q–A pairs, systematic filtering (deduplication, decontamination, cascaded difficulty filtering, and quality-aware selection) yields $185,816$ high-quality instances (31.23% retention), ensuring scientific precision and logical rigor.

Stage	Physics	Chemistry	Biology	Medicine	Total	Retention Rate (%)
Raw Q - A Pairs	144,844	113,018	154,147	182,921	594,930	100.0
+Deduplication	124,566	100,586	110,986	139,020	475,158	79.87
+Decontamination	92,499	85,603	85,793	92,743	356,638	59.95
+Difficulty Filtering	44,601	54,554	45,885	50,641	195,681	32.89
+Dual Refinement	42,808	51,623	43,294	48,091	185,816	31.23

3.4 Generative-Discriminative Dual Refinement

Directly adopting raw reference answers may not be conducive to training: many answers are too concise or lengthy, and some answers may contain errors, which undermines the accuracy of science. Moreover, even after distillation, some generated responses may still contain subtle logical flaws or hallucinations. To address these challenges, we introduce a generative-discriminative dual refinement mechanism that decouples generation from

evaluation: a *generator* produces structured reasoning chains and a *critic* independently scores them along multi-dimensional quality criteria.

We first employ Qwen3-30B-A3B-Instruct (Yang et al., 2025) as a generator G to synthesize high-quality reasoning chains, while concise final answers a^* are extracted from the raw references. For each $(q, a^*) \in \mathcal{D}_{\text{clean}}$, we sample a reasoning chain $\hat{a} \sim G(q)$ and retain it only when its final answer matches a^*:

$$\mathcal{D}_{\text{syn}} = \{(q, \hat{a}) : \hat{a} \sim G(q), \text{ans}(\hat{a}) = a^*\}. \quad (2)$$

This phase transforms verbose or inconsistently formatted raw answers into clear, structured reasoning chains that are compatible with the natural distribution of model outputs, while preserving scientific accuracy. Even if these well-structured responses are obtained, the generative model may still introduce subtle issues. We summarize the potential issues that models may encounter when answering scientific questions, categorizing them into incomplete context, logical ambiguities, excessive repetition, or imprecise formulas. To ensure only the most rigorous samples enter the final dataset, we employ the same model as a critic to perform introspective quality assessment.

Let $\mathcal{F} = \{\text{Context}, \text{Ambig}, \text{Rep}, \text{Form}\}$ denote the four scoring dimensions and $s_d(q, \hat{a}) \in \{0, 1, 2, 3\}$ the score assigned by the critic C along dimension d, where 0 indicates perfect quality and 3 indicates a severe defect. We retain only samples whose composite defect score is zero:

$$\mathcal{D}_{\text{final}} = \{(q, \hat{a}) \in \mathcal{D}_{\text{syn}} : \textstyle\sum_{d \in \mathcal{F}} s_d(q, \hat{a}) = 0\}. \quad (3)$$

This acts as an offline self-rejection sampling mechanism that filters out any instances the critic deems logically insufficient.

Through this generative-discriminative dual refinement, the pipeline reduces the dataset from approximately 196k to 186k instances by removing samples which, although superficially correct, lack the logical rigor required for scientific reasoning. This closed-loop verification ensures that the final dataset achieves both scientific precision and internal coherence, providing reliable supervision signals for model training. We verify in Appendix C.4 that the gains are preserved when the critic or the generator is replaced by a different model family, indicating that the refinement is not biased toward the model used in both roles.

The PrecSci pipeline enforces high selectivity through its sequential filtering stages, as detailed in Table 2. Starting from 595k raw QA pairs, the process retains 186k high-quality instances, ensuring that only the most scientifically accurate and logically rigorous samples are included in the final Open-Sci resource.

3.5 Dataset Description

Using PrecSci, we construct *Open-Sci*, a 186k corpus of scientifically precise question–answer pairs that provides complete non-mathematics coverage with discipline-level annotations (Table 1). By focusing on core natural sciences spanning Physics, Biology, Chemistry, and Medicine across 47 sub-disciplines (visualized in Figure 3 in Appendix A), Open-Sci delivers high-precision, diverse supervision. It delivers answers that are substantially longer and more detailed than most prior datasets, while remaining scientifically accurate and aligned. Compared with previous scientific reasoning datasets such as WebInstruct (Yue et al., 2024) and MegaScience (Fan et al., 2025), which frequently exhibit incomplete context, ambiguous formulations, overly long or repetitive answers, and imprecise formulas, Open-Sci is less prone to these issues through its systematic dual refinement process; concrete side-by-side examples are provided in Appendix A.2. This highlights how careful curation and introspective quality assessment result in higher-quality supervision, reducing noise and enabling more reliable learning. Overall, Open-Sci serves as a compact yet highly efficient dataset that prioritizes precision over scale, offering superior training resources for scientific reasoning. Additional distributions, such as question types, are provided in Appendix A.

4 Experiment

4.1 Experimental Setups

Benchmarks. We systematically assess the effectiveness of our proposed Open-Sci by evaluating trained models on four discipline-specific benchmarks: MedQA (Jin et al., 2020) and PubMedQA (Jin et al., 2019) for medicine, PHYSICS (Feng et al., 2025) for physics, ChemBench (Mirza et al., 2024) for chemistry, and ProteinLMBench (Shen et al., 2024) for biology. To broaden coverage beyond core disciplines, results on SuperGPQA (Du et al., 2025) and MMLU-Pro (Wang et al., 2024b) are further reported at the

Table 3: Performance comparison across scientific benchmarks (Medicine, Physics, Chemistry, Biology). Best scores are in **bold**, second best are underlined.

Discipline	Benchmark	Qwen3-4B-Instruct	Qwen3-4B-Meg.Sci	Qwen2.5-7B-Instruct	Qwen3-8B-Instruct	Qwen3-8B-Meg.Sci.	**Qwen3-4B-Open-Sci**	**Qwen3-8B-Open-Sci**
Medicine	MedQA	68.30	70.12	67.68	75.33	77.32	71.82	**78.57**
	PubMedQA	72.60	76.30	75.10	75.50	**77.90**	77.00	77.30
	SuperGPQA-Med.	30.17	31.11	31.43	35.67	**39.97**	33.70	38.81
Physics	PHYSICS	23.20	21.07	14.52	25.17	25.02	29.97	**39.22**
	MMLU-Pro-Phy.	67.51	69.28	57.74	73.06	75.44	78.37	**81.99**
	SuperGPQA-Phy.	30.79	31.67	24.11	36.13	36.98	41.51	**48.22**
Chemistry	ChemBench	47.68	50.30	43.39	52.36	54.18	52.13	**56.20**
	MMLU-Pro-Chem.	68.64	70.67	54.15	72.26	75.18	78.36	**83.22**
	SuperGPQA-Chem.	26.40	27.87	23.29	30.02	33.35	36.69	**43.70**
Biology	ProteinLMBench	55.72	55.51	57.10	57.63	56.67	56.99	**59.96**
	MMLU-Pro-Bio.	76.29	78.24	70.99	82.57	81.59	82.01	**85.22**
	SuperGPQA-Bio.	32.32	34.82	26.43	35.71	37.95	34.82	**40.18**
Avg		47.91	51.41	42.80	52.36	53.97	54.22	**61.05**

sub-discipline level. In addition, to examine generalization, we also include GPQA-Diamond (Rein et al., 2024), AIME2025 (American Institute of Mathematics, 2025), and Math-500 (Hendrycks et al., 2021), which measure general reasoning and mathematical ability.

Compared Models and Scientific Reasoning Datasets. We illustrate the effectiveness of Open-Sci by training it on current state-of-the-art LLMs, including Qwen2.5 (Qwen et al., 2025), Qwen3 (Yang et al., 2025), LLama-3.1 (Grattafiori et al., 2024) series models. We also compare Open-Sci with other scientific reasoning datasets, *i.e.*, NaturalReasoning (Yuan et al., 2025), MegaScience (Fan et al., 2025), and WebInstruct (Yue et al., 2024).

Training setup. We fine-tune Llama3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen et al., 2025) and Qwen3-{4B, 8B, 14B} (Yang et al., 2025) on the Open-Sci dataset. Training is implemented with MS-Swift (Zhao et al., 2025). Unless otherwise specified, we use a learning rate of 1×10^{-5}, set the global batch size to 128, and train models with 2 epochs using a warmup ratio of 0.05. All experiments are conducted on NVIDIA A100 GPUs.

Evaluation setup. Inference is carried out using vLLM (Kwon et al., 2023), and all evaluations are performed with the OpenCompass (Contributors, 2023) framework. For tasks requiring LLM-as-a-judge, we employ Qwen2.5-72B-Instruct (Qwen et al., 2025) to ensure consistent and reliable scoring. Due to the small sample sizes of AIME2025, GPQA-Diamond, and Math-500, we adopt a repeated evaluation. We report average@4 results for AIME2025 and GPQA-Diamond, and average@2 results for Math-500. All evaluations are conducted under a zero-shot setting. For decoding, Qwen2.5 and Llama use greedy decoding, while the Qwen3 family follows the official recommended setting with the temperature fixed at 0.7.

4.2 Main Evaluation Results

Consistent improvements on scientific reasoning benchmarks. Fine-tuning on Open-Sci consistently improves performance across all scientific benchmarks. As reported in Table 3, Qwen3-8B-Open-Sci improves PHYSICS from 25.17% to **39.22%**, MMLU-Pro-Chemistry from 72.26% to **83.22%**, with comparable gains in Biology and Medicine. Specifically, on SuperGPQA-Biology, the performance improves from 35.71% to **40.18%**, and on MedQA, it increases from 75.33% to **78.57%**. Even the smaller Qwen3-4B benefits from Open-Sci (e.g., PHYSICS +**6.77%**, ChemBench +**4.45%**). These consistent advances across all four domains demonstrate that the systematically curated organization of Open-Sci provides a reliable and effective foundation for model training. Importantly, despite its broad disciplinary coverage, Open-Sci does not dilute performance in any individual domain; instead, it enables comprehensive improvements across physics, chemistry, biology, and medicine. This shows that Open-Sci supplies not only precise and high-quality data but also balanced and well-structured scientific coverage, making it a practically effective resource for advancing scientific reasoning.

Boosting performance on general reasoning and mathematics. Although Open-Sci is primarily designed for scientific reasoning, models fine-tuned on it perform well across broader domains. As il-

Table 4: Performance comparison on General and Math benchmarks. Style: **best**, second best.

Model	General			Math	
	MMLU-Pro	SuperGPQA	GPQA-Diamond	AIME2025	Math-500
Qwen3-4B-Instr.	59.78	34.05	45.96	18.33	84.70
Qwen3-4B-MegaSci	61.55	34.15	43.31	17.5	84.30
Qwen2.5-7B-Instr.	55.51	31.09	37.50	7.50	77.30
Qwen3-8B-Instr.	65.52	38.51	47.73	16.67	86.10
Qwen3-8B-MegaSci	67.08	40.81	44.57	21.67	87.00
Qwen3-4B-Open-Sci	65.96	37.57	49.87	40.00	93.20
Qwen3-8B-Open-Sci	**71.25**	**42.61**	**55.68**	**45.56**	**94.70**

lustrated in Table 4, on MMLU-Pro and SuperGPQA, Open-Sci-8B-Instruct achieves strong results ($\mathbf{71.25}\%$ and $\mathbf{42.61}\%$, respectively), and leads on GPQA-Diamond. The most notable improvement is in mathematics: Open-Sci-8B-Instruct scores $\mathbf{45.56}\%$ on AIME2025, more than double the MegaScience-trained model. We believe these cross-domain benefits arise because precise scientific problems require systematic reasoning steps, such as careful use of definitions, symbolic manipulation, and quantitative calculation, that are also fundamental to general and mathematical tasks. By training on scientifically rigorous and well-structured data, models learn transferable reasoning patterns that extend beyond individual domains. In this sense, the precise and organized nature of Open-Sci not only strengthens domain-specific scientific reasoning, but also provides a foundation that supports generalization to mathematics and broader reasoning challenges.

Precise data outweighs massive noisy data. As shown in Table 3 and Table 4, compared with MegaScience's $\sim$ 1.25M mixed-source corpus, Open-Sci uses only 186k samples ($\sim 14.88\%$ of MegaScience) yet surpasses it on most benchmarks. On average across scientific benchmarks, Open-Sci achieves superior performance with substantial gains also appearing on general and mathematical tasks, *e.g.*, GPQA-Diamond and AIME2025. We suggest that these advantages stem from Open-Sci's emphasis on systematic precision-oriented curation, particularly through the generative-discriminative dual refinement stage. Other datasets like MegaScience and WebInstruct often suffer from incomplete context, ambiguous formulations, overly long or repetitive answers, and imprecise formulas (see Appendix A.2 for representative side-by-side examples). Such noise reduces training data usefulness, limiting models' ability to develop robust reasoning skills. By systematically removing these issues through introspective quality assessment, Open-Sci offers more precise, scientifically rigorous data, enabling more reliable and efficient learning. This observation highlights how a much smaller corpus can decisively outperform substantially larger but noisier datasets.

Broader effectiveness across model families. We fine-tune leading open-source models from the Llama and Qwen series on Open-Sci. As shown in Table 5, we draw two main observations. First, Open-Sci fine-tuning consistently improves performance across different model families and parameter scales. For instance, Qwen3-8B achieves an average gain of $+\mathbf{9.16}\%$, with striking improvements on mathematics tasks, *e.g.*, AIME2025 ($+\mathbf{28.89}\%$). Qwen3-14B exhibits similarly large benefits, $+\mathbf{7.16}\%$ on average. We also observe the improvement in Llama-3.1-8B ($+\mathbf{3.11}\%$) and Qwen2.5-7B ($+\mathbf{1.02}\%$). This stems from the dataset's combination of precise scientific formulation and balanced disciplinary coverage, allowing models of varying architectures and scales to consistently acquire more rigorous reasoning patterns. Second, improvement extent depends on the model's baseline reasoning capacity. The Qwen3 series, with stronger reasoning priors, achieves the most substantial gains, while Llama-3.1-8B and Qwen2.5-7B show relatively smaller improvements. We suggest that this is because scientifically precise data poses more complex reasoning challenges, which weaker models may struggle to fully exploit. To further check that the gains generalize beyond the Llama and Qwen families, we report results on Gemma-3-4B in Appendix E.2, where Open-Sci yields consistent improvements across all six evaluated benchmarks.

4.3 Ablation Study

Data Efficiency. To further examine the data efficiency of Open-Sci, we conduct a comparison by sampling 200k instances from WebInstruct, NaturalReasoning, and MegaScience, matching the scale of our dataset. As shown in Table 6, models fine-tuned on Open-Sci consistently achieve

Table 5: Generalization across **Science**, **General**, and **Math** benchmarks. Numbers are accurate (%). For rows fine-tuned on our dataset (-Open-Sci), we show per-task gain Δ (Open-Sci – Baseline) in parentheses and report the mean gain **Avg.** Δ.

Model	Science				General		Math		Avg. Δ
	MedQA	PHYSICS	ChemBench	ProteinLMB.	MMLU-Pro	SuperGPQA	AIME2025	Math-500	
Llama-3.1-8B-Instr.	65.60	7.65	40.14	55.40	48.08	20.50	0.00	52.10	
Llama-3.1-8B-Open-Sci	67.99 (+2.39)	15.13 (+7.48)	40.94 (+0.80)	57.52 (+2.12)	54.91 (+6.83)	24.86 (+4.36)	0.83 (+0.83)	52.20 (+0.10)	**+3.11**
Qwen2.5-7B-Instr.	67.68	14.52	43.27	57.10	55.51	31.09	7.50	77.30	
Qwen2.5-7B-Open-Sci	71.35 (+3.67)	14.91 (+0.39)	39.98 (-3.29)	57.20 (+0.10)	59.22 (+3.71)	31.49 (+0.40)	10.00 (+2.50)	78.00 (+0.70)	**+1.02**
Qwen3-4B-Instr.	68.30	23.20	47.68	55.72	59.78	34.05	18.33	84.70	
Qwen3-4B-Open-Sci	71.82 (+3.52)	29.97 (+6.77)	52.13 (+4.45)	56.99 (+1.27)	65.96 (+6.18)	37.57 (+3.52)	40.00 (+21.67)	93.20 (+8.50)	**+6.74**
Qwen3-8B-Instr.	75.33	25.17	52.36	57.63	65.52	35.99	16.67	86.10	
Qwen3-8B-Open-Sci	78.57 (+3.24)	39.22 (+14.05)	56.20 (+3.84)	59.96 (+2.33)	71.25 (+5.73)	42.61 (+6.62)	45.56 (+28.89)	94.70 (+8.60)	**+9.16**
Qwen3-14B-Instr.	79.43	31.94	44.20	61.76	69.67	42.57	25.00	88.30	
Qwen3-14B-Open-Sci	81.88 (+2.45)	42.38 (+10.44)	45.47 (+1.27)	60.91 (-0.85)	73.54 (+3.87)	45.89 (+3.32)	55.83 (+30.83)	95.30 (+7.00)	**+7.16**

Table 6: Performance comparison across scientific, general, and mathematical benchmarks using 200k sampled data from different datasets. Best scores are in **bold**, second best are underlined.

Dataset	Science Avg.	General Avg.	Math Avg.
Ours	**62.25**	**56.51**	**70.13**
MegaScience	57.70	48.61	58.95
NaturalReason	51.60	40.66	39.05
WebInstruct	50.83	36.62	25.15

Table 7: Ablation study on different pipeline components. "w/o" denotes removing the corresponding component. Results are averaged across science, general, and mathematics benchmarks.

Variant	Science Avg	General Avg	Math Avg
Full pipeline (Ours)	62.25	56.51	70.13
w/o Question Formal.	57.87	47.22	66.22
w/o Noise-Mit.	58.15	49.12	64.99
w/o Dual Refinement	61.15	51.89	68.42

the highest average performance across scientific, general, and mathematical benchmarks under the same data budget. This demonstrates that precise and carefully curated data provides substantially greater per-sample efficiency, clearly surpassing larger-scale but noisier alternatives.

Pipeline Components. Table 7 reports ablations that remove each non-extraction stage in turn. Removing question formalization causes the largest overall drop (especially on General, -9.29), confirming that ambiguous or inconsistently structured problems reduce the effectiveness of supervision. Noise mitigation contributes comparably across the three averages: without it, trivial or ill-posed items contaminate training. Removing dual refinement has the smallest impact, yet it still degrades performance and is responsible for filtering out verbose, misaligned, or logically flawed reasoning chains. These results confirm that all three refinement stages are required, with question formalization contributing the most.

Impact of Disciplinary Diversity. To investigate whether multi-discipline training yields better generalization than single-discipline training, we conduct a controlled experiment with a fixed budget of 50k samples. We compare a model fine-tuned on Medicine-only samples (*Open-Sci-Med-50k*) against one fine-tuned on 50k samples randomly sampled across all disciplines (*Open-Sci-Mixed-50k*). As shown in Table 10, while the Medicine-specialized model achieves slightly higher performance on MedQA (76.93% vs. 75.74%), the Mixed-discipline model demonstrates superior performance on out-of-domain tasks (e.g., Physics +3.32%, Chemistry +1.11%) and general scientific benchmarks. Notably, even the Med-only model improves Physics performance over the base model (25.17% → 29.91%), suggesting that scientific reasoning skills possess a degree of transferability. The Mixed strategy nonetheless prevents domain overfitting, validating the necessity of Open-Sci's broad disciplinary coverage.

5 Conclusion

We present Open-Sci, a compact scientific reasoning dataset built by the PrecSci pipeline, which integrates source collection, question formalization, and noise mitigation with a refinement stage where a generator and a critic enforce internal consistency. Despite being only $\sim 14.88\%$ the size of comparable corpora, Open-Sci consistently improves performance on core scientific benchmarks and transfers strongly to mathematics and general reasoning. This shows that high-density data can match or surpass larger, noisier corpora, providing a recipe for efficient scientific foundation models.

Limitations

This work focuses on curating high-density scientific reasoning data for supervised fine-tuning. First, PrecSci relies on strong teacher models for both generation and critique; the architecture permits using different model families for the two roles, but the achievable data quality is still upper-bounded by the strongest available teacher. Second, although we apply LSH-based deduplication and embedding-similarity decontamination, residual overlap between our public textbook/exam sources and the pretraining corpora of evaluated base models cannot be fully ruled out; gains on multiple model families partially mitigate this concern. Finally, the cascaded filtering is intentionally conservative and may discard valuable samples along with noisy ones; recovering borderline samples is left to future work.

References

American Institute of Mathematics. 2025. Aime 2025 competition mathematical problems.

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, and 1 others. 2025. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, and 1 others. 2025. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and 1 others. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Yongchao Chen, Yueying Liu, Junwei Zhou, Yilun Hao, Jingquan Wang, Yang Zhang, and Chuchu Fan. 2025a. R1-code-interpreter: Training llms to reason with code via supervised and reinforcement learning. *arXiv preprint arXiv:2505.21668*.

Zihong Chen, Wanli Jiang, Jinzhe Li, Zhonghang Yuan, Huanjun Kong, Wanli Ouyang, and Nanqing Dong. 2025b. Graphgen: Enhancing supervised fine-tuning for llms with knowledge-driven synthetic data generation. *arXiv preprint arXiv:2505.20416*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, and 1 others. 2025. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*.

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*.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. 2025. Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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. *arXiv preprint arXiv:2103.03874*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, and 16 others. 2024. Are large language models superhuman chemists? *arXiv preprint arXiv: 2404.01475*.

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. 2023. Chenghaomou/text-dedup: Reference snapshot.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2022. Overview of bioasq 2022: the tenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International conference of the cross-language evaluation forum for European languages*, pages 337–361. Springer.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report.

A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint*.

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*.

ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, and 1 others. 2025. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. 2024. A fine-tuning dataset and benchmark for large language models for protein understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2390–2395. IEEE.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. Scidqa: A deep reading comprehension dataset over scientific papers. *arXiv preprint arXiv:2411.05338*.

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*.

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*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiheng Xi, Guanyu Li, Yutao Fan, Honglin Guo, Yufang Liu, Xiaoran Fan, Jiaqi Liu, Jingchao Ding, Wangmeng Zuo, Zhenfei Yin, and 1 others. 2025. Bmmr: A large-scale bilingual multimodal multi-discipline reasoning dataset. *arXiv preprint arXiv:2507.03483*.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, and 1 others. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought, 2025. *URL https://arxiv. org/abs/2501.04682*, 2(4).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Ilia Kulikov, Kyunghyun Cho, Dong Wang, Yuandong Tian, Jason E Weston, and 1 others. 2025. Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions. *arXiv preprint arXiv:2502.13124*.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

A Dataset Details

A.1 Diversity

Open-Sci spans four core disciplines (physics, chemistry, biology, medicine) and 47 sub-disciplines, with representative example questions for each shown in Figure 3.

Open-Sci also exhibits substantial diversity in both question formats and answer styles. As illustrated in Figures 4a and 4b, the dataset covers a wide range of assessment types, including open-ended questions (48%), calculation problems (30%), multiple-choice items (13%), and smaller shares of true/false, fill-in-the-blank, and other formats. To characterize answer diversity, we categorize each reference answer by its token length: 1–5 tokens as Factoid, 6–20 as Phrase, 21–50 as Sentence, and over 50 as Paragraph. The resulting distribution contains 14% factoid answers, 23% phrases, 18% sentences, and 45% paragraphs. Shorter responses typically align with structured formats such as multiple-choice or calculation, while longer ones arise from open-ended or analysis-style tasks requiring detailed explanations. This variety ensures that Open-Sci provides supervision signals for both concise factual queries and extended scientific argumentation, thereby promoting broader generalization across tasks.

Table 8: Detailed results on MMLU-Pro for Qwen3-8B before and after fine-tuning on Open-Sci.

Benchmark	Base	+ Open-Sci
Math	83.94	**90.67**
Physics	73.06	**82.37**
Chemistry	72.26	**82.51**
Law	34.51	**38.24**
Engineering	52.32	**62.33**
Economics	74.64	**77.25**
Health	64.18	**65.28**
Psychology	69.67	**71.05**
Business	73.64	**77.69**
Biology	82.57	**83.54**
Philosophy	54.51	**60.32**
Computer Science	69.51	**75.12**
History	53.54	**56.96**
Other	58.87	**63.31**

A.2 Qualitative Comparison with Prior Datasets

Figure 5 illustrates concrete quality differences between Open-Sci and two representative prior datasets, WebInstruct (Yue et al., 2024) and MegaScience (Fan et al., 2025). While these datasets frequently exhibit incomplete or missing context, ambiguous formulations, overly long or repetitive answers, and imprecise or incorrect formulas, the systematic dual refinement step in PrecSci visibly reduces all of these failure modes in Open-Sci.

B Evaluation Details

The complete evaluation on MMLU-Pro is in Table 8

C Pipeline Details

We summarize the full PrecSci pipeline in Algorithm 1.

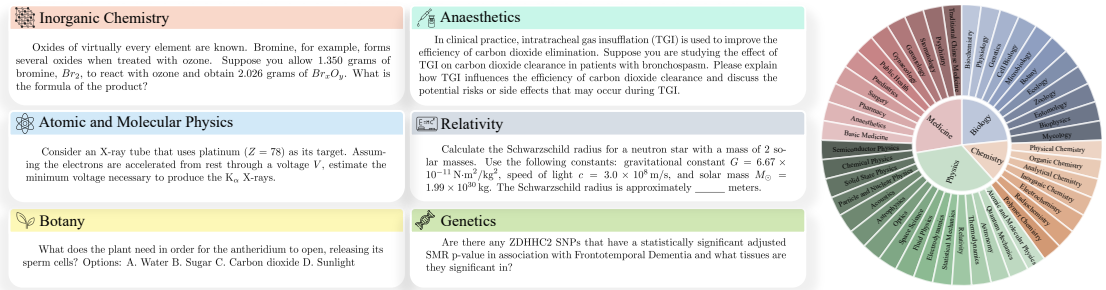
C.1 Question formalization

The prompt for question formalization is in Figures 6

C.2 Noise Mitigation

Duplicate or near-duplicate questions are removed using MinHash with locality-sensitive hashing. We adopt 3-gram shingles, a similarity threshold of 0.6, and LSH parameters of 1 band and 10 rows.

To avoid contamination with evaluation benchmarks, we encode all Open-Sci training samples and all benchmark items using the bge-large-en-v1.5 embedding model. For each training sample,



(a) Diverse question in Open-Sci (b) Diverse sub-discipline in Open-Sci

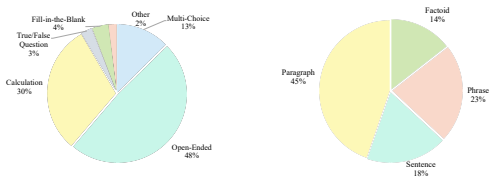
Figure 3: (a) Representative question examples across various sub-disciplines, illustrating the heterogeneity of scientific problem types. (b) Disciplinary and sub-disciplinary coverage of Open-Sci across physics, chemistry, biology, and medicine.

Table 9: Number of removed items per benchmark.

Benchmark	# Removed
ProteinLMBench	3,115
PHYSICS	4,708
MedQA	16,821
PubMedQA	191
ChemBench	6,683
MMLU-Pro	32,313
SuperGPQA	52,594
GPQA	1,741
Math	354
AIME-2025	0

we retrieve its top-5 most similar benchmark items under cosine similarity. Each retrieved pair (training item, benchmark items) is evaluated for semantic overlap using Llama-3.3-70B-Instruct with temperature set to 0 to ensure deterministic judgments. A training sample is removed if any retrieved benchmark neighbor is judged to exhibit semantic equivalence. The exact prompt used for semantic equivalence checking is provided in Figure 7. The numbers of removed items per benchmark are shown in Table 9

To further eliminate trivial, ambiguous, or ill-posed questions, we apply two-stage LLM-based filtering. Qwen2.5-7B-Instruct is used to remove overly easy items (≥ 7 out of 8 sampled responses correct). Qwen3-30B-A3B-Instruct is used to filter scientifically invalid or ill-posed items (0 out of 8 responses correct). The judgement prompt used to evaluate whether a sampled response matches the reference answer is shown in Figure 8.



(a) Distribution of question types. (b) Distribution of reference answer lengths.

Figure 4: Diversity of Open-Sci in terms of question types and reference answer lengths.

Algorithm 1 PrecSci Data Curation Pipeline

Require: Raw sources $\mathcal{S}$ (textbooks, exams, competitions); generator G; critic C; thresholds τ_e, τ_h; defect dimensions $\mathcal{F}$
Ensure: Curated dataset $\mathcal{D}_{\text{final}}$

```
D_QA ← Extract(S)                          ▷ Section 3.1
D_form ← Formalize(D_QA)                   ▷ Section 3.2
D_dedup ← LSH-Dedup(D_form)
D_decon ← Decontaminate(D_dedup)
D_clean ← {(q, a) ∈ D_decon : α_s < τ_e, α_l > τ_h}
D_syn ← ∅
for each (q, a*) ∈ D_clean do
    Sample â ~ G(q)
    if â ≡ a* then
        D_syn ← D_syn ∪ {(q, â)}
    end if
end for
D_final ← ∅
for each (q, â) ∈ D_syn do
    s_d ← C(q, â; d) for each d ∈ F
    if Σ_{d∈F} s_d = 0 then
        D_final ← D_final ∪ {(q, â)}
    end if
end for
return D_final
```

C.3 Generative-Discriminative Dual Refinement

After answer synthesis, we employ Qwen3-30B-A3B-Instruct-2507 to perform introspective quality assessment on each question-answer pair. The

WebInstruct

Q: Evaluate the integral (. . .) where d is an integer greater than 1, and the other parameters are as defined in the original question.

A: The evaluation of this integral is similar to the original integral and may also require the use of specialized mathematical software. The resulting expression may involve special functions such as elliptic integrals and arctangent functions, and its complexity may depend on the values of d and the other parameters.

MegaScience

Q: A system consists of three masses, m_1, m_2, and m_3, where m_2 is connected to m_3 through a pulley attached to m_1. There is no friction between the masses. Find the force F required so that m_2 remains at rest relative to m_3. Then, determine the accelerations of the masses when $F = 0$. Consider the effects of gravity on m_3 and the role of tension in the system.

A (shortened): "We have a system of three masses . . ." [long derivations omitted]

$F = (m_1 + m_2 + m_3)\,\frac{m_2}{m_3}\,g \quad \ldots$
$F = (m_1 + m_2 + m_3)\,g\,\frac{m_2}{m_3} \quad \ldots$
$F = (m_1 + m_2 + m_3)\,g\,\frac{m_3}{m_2} \quad \ldots$
$F = (m_1 + m_2 + m_3)\,g\,\frac{m_3}{m_3} \quad \ldots$

A more detailed diagram or description would be needed for a precise solution.

Incomplete /Missing info | Ambiguous | Overly long | Repetitive | Imprecise /Wrong formula

Open-Sci

Q: A small block of mass m is placed at rest on a wedge of mass M, which is placed on a horizontal table. Assume all contact surfaces are smooth, and the angle between the wedge and the horizontal plane is θ. In order for the block to remain at rest relative to the wedge (i.e., not sliding down), find the magnitude of the horizontal force F applied to the wedge.

A (shortened):

Since all surfaces are smooth, the block experiences gravity mg and the normal force N only. To keep the block at rest relative to the wedge, choose the ground frame and require zero acceleration along the plane (. . .). Resolve N:

$$\begin{aligned} &\text{Vertical:} && N\cos\theta = mg \quad (1)\\ &\text{Horizontal:} && N\sin\theta = ma \quad (2)\end{aligned}$$

From (1) $N = \dfrac{mg}{\cos\theta}$, substitute into (2) to get

$$a = g\tan\theta\,.$$

Apply Newton's second law to the wedge+block system (. . .):

$$F = (M+m)a = (M+m)g\tan\theta\,.$$

$$\boxed{F = (M+m)g\tan\theta}$$

Figure 5: Quality comparison of QA pairs across WebInstruct, MegaScience, and Open-Sci.

Prompt for question formalization:

```
### INPUT
Here is the markdown content that you need to analyze, correct, and filter:
Question: {question}
### GLOBAL TASKS
1. **Strip embedded answers**
- If the correct answer or solution steps appear in the stem, remove them entirely.
2. **Fix formatting & LaTeX**
- Scan all math expressions, symbols, superscripts/subscripts, fractions, etc., and correct any mis-parses (e.g.,
  "x2" → "x^2", missing backslashes).
3. **Normalize markdown**
- Ensure headings, lists, code blocks and inline math are valid Markdown/LaTeX.
4. **Maintain Language**
- Ensure the output question is in the same language as input_question

### PER-TYPE RULES
Current Question Type: {question_type}

#### 1. Multiple-Choice
- Remove any duplicate option lists.
- Ensure exactly one set of options labeled **A.**, **B.**, **C.**, **D.**, etc.
- If labels are missing or inconsistent, insert or renumber them sequentially.
#### 2. True/False
- If presented as a declarative statement, rephrase to:
> "Determine whether the following statement is true or false:
> _<original statement>_"
- Remove any embedded "True" or "False" answers.
#### 3. Short-Answer
- Confirm the question ends with a clear prompt (e.g. "What is…?", "Explain why…?").
- Remove any parenthetical answer hints.
#### 4. Fill-in-the-Blank
- Represent blanks consistently as `____` (at least 3 underscores).
- Ensure each blank corresponds to exactly one missing answer.
- Remove any answers shown inline.
- If there are multiple blanks, number them:
> "The capital of France is (1) ____, and its currency is (2) ____."
#### 5. Calculation
- Strip out worked-out solutions and final answers.
- Verify units are present and correctly formatted (e.g., "m", "kg", "$").
- Confirm numeric values in the stem are clear (e.g., no stray commas or spaces).
- If a formula is required, ensure it's in LaTeX math delimiters.
### OUTPUT
Return exactly this, in Markdown, without surrounding triple-tick fences:
**Refined Question**
<your cleaned, corrected question markdown here directly without redudant information such as Question
Type>
**Rationale**
<Briefly note what you removed or fixed>
```

Figure 6: Prompt for question formalization.

Prompt for decontamination

```
I will now give you two questions: Original question and Candidate question. Please help me
determine if the following two questions are the same.

Original question:
`<ORIGINAL_PROBLEM>`

Candidate question:
`<CANDIDATE_PROBLEM>`

Disregard the names and minor changes in word order that appear within.
If their question prompts are very similar and, without considering the solution process, they produce
the same answer, we consider them to be the same question.

Output Format:
Analysis: [Provide a detailed analysis evaluating the similarity between these questions]
Decision: [YES/NO]
```

Figure 7: Prompt for decontamination.

model evaluates four defect dimensions, each dimension, a defect score from 0 (perfect) to 3 (severe) is assigned. Only samples achieving zero defects across all dimensions are retained, ensuring the final dataset meets strict internal consistency criteria. This stage reduced the dataset from 195,681 to 185,816 instances. The prompt is in Figure 9.

To synthesize structured reasoning chains, the generator must first extract a concise final answer a^* from the raw scientific reference (textbook solution, exam key, or competition answer). Because the format of reference answers differs across disciplines (e.g., chemical equations, IUPAC names, balanced reactions, or multi-part medical conclusions), we use discipline-specific extraction prompts: Figure 10 (biology), Figure 11 (chemistry), Figure 12 (medicine), and Figure 13 (physics).

C.4 Cross-model Critic Validation

To verify that the gains from the dual refinement stage do not depend on using the same model as generator and critic, we run two cross-model checks.

Critic replacement. We sample 200 question-answer pairs and have both Qwen3-30B (our default critic) and DeepSeek-V3, drawn from a different model family, score them along the four objective defect dimensions. The two critics agree on more than 92% of judgements. The disagreement is asymmetric: among the 200 items, 55 are rejected by Qwen3-30B but accepted by DeepSeek-V3, while only 3 items go in the reverse direction. This rules out a self-preference bias for Qwen; if

Prompt for judgement

Please as a grading expert, judge whether the final answers given by the candidates below are consistent with the standard answers, that is, whether the candidates answered correctly.
Here are some evaluation criteria:
1. Please refer to the given standard answer. You don't need to re-generate the answer to the question because the standard answer has been given. You only need to judge whether the candidate's answer is consistent with the standard answer according to the form of the question. Don't try to answer the original question. You can assume that the standard answer is definitely correct.
2. Because the candidate's answer may be different from the standard answer in the form of expression, before making a judgment, please understand the question and the standard answer first, and then judge whether the candidate's answer is correct, but be careful not to try to answer the original question.
3. Some answers may contain multiple items, such as multiple-choice questions, multiple-select questions, fill-in-the-blank questions, etc. As long as the answer is the same as the standard answer, it is enough. For multiple-select questions and multiple-blank fill-in-the-blank questions, the candidate needs to answer all the corresponding options or blanks correctly to be considered correct.
4. Some answers may be expressed in different ways, such as some answers may be a mathematical expression, some answers may be a textual description, as long as the meaning expressed is the same. And some formulas are expressed in different ways, but they are equivalent and correct.
Please judge whether the following answers are consistent with the standard answer based on the above criteria. Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
Just return the letters "A" or "B", with no text around it.
Here is your task. Simply reply with either CORRECT, INCORRECT. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
<Original Question Begin>:
{question}
<Original Question End>
<Gold Target Begin>:
{label}
<Gold Target End>
<Predicted Answer Begin>:
{prediction}
<Predicted End>
Judging the correctness of candidates' answers:

Figure 8: Prompt for judgement.

anything, the Qwen3-30B critic is the stricter of the two, and the retained dataset is therefore the more conservative one.

Generator replacement. We further re-run the synthesis stage with Llama-3.1-70B-Instruct in place of Qwen3-30B as the generator, while keeping the rest of the pipeline unchanged, and fine-tune Llama-3.1-8B on the resulting variant. The per-benchmark numbers, summarized in Table 11, show that the Llama-generated variant still improves over the base Llama-3.1-8B on MMLU-Pro (+4.48), PHYSICS (+4.02) and MedQA (+1.01), confirming that the gains are not contingent on a specific generator family.

Taken together, the dual refinement stage spans three model families (Qwen, DeepSeek, Llama) across the generator and critic roles, which mitigates the risk of single-model bias in the final dataset.

D Cases

More cases of Open-Sci are shown in Figure 14.

E More Ablation Study

E.1 Performance under different teacher model.

We further distil the same data from Llama-3.1-70B-Instruct and fine-tune Llama-3.1-8B; the results are reported in Table 11. The Llama-distilled variant occasionally edges out on ChemBench, but training directly on Open-Sci still yields the strongest results on most benchmarks, suggesting that different teacher models emphasise different aspects of scientific reasoning and that a hybrid distillation strategy may further improve Llama-family students.

Prompt for discrimination

Role
You are a rigorous Data Quality Analyst. Your task is to evaluate Question-Answer (QA) pairs and score them on 4 specific defect dimensions.
Scoring Scale (0 - 3)
For each dimension, assign a **Defect Score** from 0 to 3 based on the severity of the issue:
- **0 (None):** The issue is completely absent. Perfect quality.
- **1 (Minor):** Slight issue, but the data is still usable/intelligible.
- **2 (Moderate):** Noticeable issue that degrades quality. Requires editing.
- **3 (Severe):** Critical failure (e.g., infinite loops, impossible to answer). The data is unusable.
Dimensions to Evaluate
1. **Incomplete Context (Context_Defect)**
- **Score 3:** Question asks about "the table below", "the provided text", or undefined variables (e.g. "What is x?") without any context provided.
- **Score 0:** Question is self-contained and fully understandable on its own.
2. **Ambiguity (Ambiguity_Defect)**
- **Score 3:** Question/Answer is logically incoherent, grammatically broken, or has multiple conflicting interpretations.
- **Score 0:** Phrasing is precise, clear, and unambiguous.
3. **Repetition & Verbosity (Repetition_Defect)**
- **Score 3:** **Pathological Looping** (repeating the same sentence 3+ times), mechanical repetition, or getting stuck in a logic loop.
- **Score 2:** Excessive fluff, redundant preambles, or saying the same thing twice in slightly different words.
- **Score 0:** Concise, direct, and to the point.
4. **Imprecise Formulas (Formula_Defect)**
- **Score 3:** LaTeX is broken/unrendered, or formulas are scientifically wrong (missing variables, wrong units).
- **Score 0:** Formulas are correctly formatted (e.g., valid LaTeX) and scientifically precise.
Output Format
Output ONLY a JSON object with the scores and a brief reasoning.
{
  "scores": {
    "Context_Defect": 0,
    "Ambiguity_Defect": 0,
    "Repetition_Defect": 3,
    "Formula_Defect": 0
},
  "total_defect_score": 3,
  "reasoning": "The answer enters an infinite loop repeating 'the value is 5' multiple times."
}
# Task
Analyze the following QA pair:
Question: {{INSERT_QUESTION}}
Answer: {{INSERT_ANSWER}}

Figure 9: Prompt for quality-aware selection. The model evaluates each QA pair across four defect dimensions (Context, Ambiguity, Repetition, Formula) with scores from 0-3, retaining only samples with zero total defects.

Table 10: Performance under different disciplinary coverage. **Med-50k**: 50k Medicine-only samples; **Mixed-50k**: 50k samples uniformly mixed across all disciplines.

Benchmark	Qwen3-8B	Med-50k	Mixed-50k
ChemBench	52.36	52.53	**53.64**
ProteinLMBench	57.63	58.37	**58.79**
MedQA	75.33	**76.93**	75.74
PHYSICS	25.17	29.91	**33.23**
MMLU-Pro	65.52	68.83	**69.06**

E.2 Multi-family Transfer to Gemma-3

Beyond the Llama and Qwen model families reported in Table 5, we evaluate whether the gains transfer to a third, more recent family. We fine-tune Gemma-3-4B-it on Open-Sci and evaluate it on six benchmarks. The results in Table 12 show consistent gains across all six, indicating that the

Prompt for reference answer extraction(Biology):

Task Description
You are tasked with extracting the final reference answer from a detailed biological solution that contains both reasoning steps and the final conclusion.
Biological answers may take the form of:
- A single word/phrase (e.g., a term like chloroplasts, IL-2, mutation breeding).
- A letter corresponding to a multiple-choice option.
- A full-sentence or multi-sentence explanation (when the final answer includes description of a function, mechanism, or role).
Your goal is to output the final, standalone reference answer in its complete form, without losing essential biological context.
Input Format
You will receive:
1. A biological question (Question)
2. A detailed answer that includes reasoning steps and the final conclusion (Detailed Answer)
Output Requirements
- Extract only the final reference answer, without intermediate reasoning.
- Ensure the extracted answer is complete and self-contained (include the full descriptive conclusion if present, not just a fragment).
- If the answer contains both a choice label (e.g., "B.") and the explanatory phrase, extract both together (e.g., "B. Mutation breeding").
- If the answer is a descriptive explanation (e.g., about a cytokine's function), extract the entire explanatory block, not just the first sentence.
- Do not add new explanations or modify the original meaning.
- If multiple possible answers are given, select the one explicitly marked as final, preferred, or correct.
Special Notes for Biology
- Many biological answers require context (e.g., naming a molecule and stating its function). Always keep the final explanatory part intact.
- For multiple-choice questions, do not drop the option letter if it is part of the final answer.
- For open-ended questions (e.g., mechanisms, processes), keep the full concluding explanation. Do not shorten to a single keyword unless the original final answer is only that keyword.
Example 1(Multiple-choice)
Question:
在线粒体中，主要通过哪种结构进行氧化磷酸化反应以生成ATP?
A. 基质
B. 外膜
C. 内膜
D. 膜间隙
Detailed Answer:
…reasoning steps…
所以综合起来，答案应该是选项C，内膜。
Reference Answer:
C. 内膜
Example 2(Open-ended)
Question:
在研究中发现CD2+细胞释放出一种因子能促进自身生长和延长存活期。这种因子最可能是什么？其主要功能和免疫作用是什么？
Detailed Answer:
…reasoning steps…
1.这种因子最有可能是白细胞介素-2（Interleukin-2, IL-2）。
2.IL-2的主要功能及免疫作用：•促进T细胞增殖、分化和存活 •驱动效应和记忆T细胞形成 •调节免疫耐受并协同B细胞和NK细胞功能
Reference Answer:
1.白细胞介素-2（Interleukin-2, IL-2）
2.功能及免疫作用：促进T细胞增殖、分化和存活；驱动效应与记忆T细胞形成；调节免疫耐受；协同B细胞和NK细胞功能。
Instructions
1.Read the question carefully to understand what is being asked.
2.From the detailed answer, locate the final conclusion.
3.Extract the conclusion in full (including option labels or full descriptive text).
4.Output only the reference answer, clearly formatted.
Question:
`<PROBLEM>`
Detailed Answer:
`<ANSWER>`
Now process and return the result.

Figure 10: Prompt for reference answer extraction in biology.

Prompt for reference answer extraction(Medicine):

Task Description
You are tasked with extracting the final reference answer from a detailed medical solution that contains both reasoning/discussion steps and the final conclusion. The reference answer should be precisely the definitive conclusion, phrased exactly as it would appear in a standard medical solution key.
Input Format
You will receive:
1. A medical question
2. A detailed answer that includes reasoning steps and the final answer
Output Requirements
- Extract ONLY the final reference answer, without reasoning or intermediate discussion.
- Ensure the reference answer is complete, clinically accurate, and able to stand alone.
- Keep the answer type consistent with the question:
  - For multiple choice questions: return the final selected option (e.g., "D. 受害者的血型")
  - For short-answer/fill-in-the-blank: return the concise term or phrase (e.g., "雌激素")
  - For essay/analysis: return the final summarized conclusion (e.g., "CNETs在形态、分子机制、临床进程及预后等方面均与实性pNETs存在显著差异，符合肿瘤分类中的独立类型标准。")
- Do not shorten the final answer if it requires full context to be medically correct (e.g., mechanisms, multi-part responses).
- Do not include reasoning steps, explanations, or "因此/所以/结论是"类提示语。
- If multiple answers are presented, extract the one explicitly marked as final, correct, or preferred.
Extraction Priorities for Medical Content
1.Direct final statement (often at the end of the solution).
2.Option label if the answer is multiple choice (e.g., "A. M. tuberculosis").
3.Key medical term/phrase if fill-in-the-blank or short question.
4.Complete concluding paragraph if analytical/essay type.
Example
Question:
在法医病理学中，分析损伤模式对于确定死亡原因至关重要。当检查伤口时，法医病理学家需要考虑多种因素。下列哪一项不是考虑的因素？
A. 使用的凶器类型
B. 伤口的角度
C. 施力方向
D. 受害者的血型
Detailed Answer:
... derivation ...
所以正确答案是D. 受害者的血型。
Reference Answer:
D. 受害者的血型
Instructions
1.Carefully read the question to determine the expected format (MCQ, short answer, or essay).
2.Analyze the detailed answer and locate the final conclusion or chosen option.
3.Extract only the final reference answer in its medically correct and complete form.
4.Return it cleanly formatted, with no extra commentary.
Question:
`<PROBLEM>`
Detailed Answer:
`<ANSWER>`
Now process and return the result.

Figure 12: Prompt for reference answer extraction in medicine.

Prompt for reference answer extraction(Chemistry):

Task Description
You are tasked with extracting the final reference answer from a detailed chemistry solution that contains both reasoning steps and the final conclusion.
Chemistry answers often differ from math answers, and the final reference answer may take one of several forms:
- A short explanatory statement (for conceptual/causal questions)
- True/False judgments (often listed part by part)
- Numerical results with correct units and significant figures (e.g., grams, mol, %, K values)
- Balanced chemical equations or ionic equations
- Standardized symbolic expressions (e.g., bold text, \(\boxed{…}\), or clearly listed results)
Input Format
You will receive:
1. A chemistry question (Question)
2. A detailed answer that includes reasoning steps and the final conclusion (Detailed Answer)
Output Requirements
- Extract only the final reference answer, without reasoning steps.
- Ensure completeness: if the answer contains multiple sub-answers (e.g., parts a–d), extract all of them.
- Ensure stand-alone clarity: the extracted answer must be self-contained and usable as a standard solution.
- Preserve chemistry formatting:
  - Keep chemical equations as written (e.g., 2 NH3 ⇌ NH4+ + NH2-).
  - Keep numerical results with correct units and significant figures.
  - For multiple-choice or True/False questions, list answers clearly part by part.
- Remove redundancy: exclude repeated reasoning or irrelevant background.
- If explicit markers exist (e.g., "Final Answer", "Thus", \(\boxed{}\)), prioritize extracting from them.
- If no explicit marker exists, extract from the summary statement at the end.
Example 1
Question:
Aniline (conjugate acid pKa 4.63) is a considerably stronger base than diphenylamine (pKa 0.79). Account for these marked differences.
Detailed Answer:
…reasoning steps…
Final Answer: \boxed{\text{Aniline is a stronger base than diphenylamine due to less resonance delocalization of the lone pair on nitrogen.}}
Reference Answer:
Aniline is a stronger base than diphenylamine due to less resonance delocalization of the lone pair on nitrogen.
Example 2
Question:
Label the following statements true or false: a. … b. … c. … d. …
Detailed Answer:
…reasoning steps…
Final Answer: a. \boxed{\text{False}}
b. \boxed{\text{False}}
c. \boxed{\text{True}}
d. \boxed{\text{True}}
Reference Answer: a. False b. False c. True d. True
Instructions
1.Read the question carefully to determine its type (numerical calculation / judgment / explanation / chemical equation).
2.Locate the final conclusion in the detailed answer (usually marked as "Final Answer" or at the end).
3.Extract the complete, self-contained final answer.
4.Format it clearly so it can be used directly as a reference answer.
Question:
`<PROBLEM>`
Detailed Answer:
`<ANSWER>`
Now process and return the result.

Figure 11: Prompt for reference answer extraction in chemistry.

Prompt for reference answer extraction(Physics):

Task Description
You are tasked with extracting the final reference answer(s) from a detailed physics solution that contains both reasoning steps and final results. Physics solutions often include long derivations, intermediate formulas, and unit conversions. Your goal is to extract the **final, definitive answer(s)** that can serve as the standard solution.
Input Format
You will receive:
1. A physics problem (the question)
2. A detailed solution (including reasoning, formulas, intermediate steps, and the final answer)
Output Requirements
- Extract ONLY the final reference answer(s), not the reasoning or derivations.
- Ensure the reference answer includes:
  - The **numerical value(s) with correct unit(s)** when present.
  - The **final expression or conclusion** if the solution ends with a formula or statement.
  - **All sub-answers** if the problem is divided into multiple parts (e.g., (a), (b), (c)).
- Do NOT add explanations, derivations, or restatements of the problem.
- Do NOT omit essential parts of the final answer (e.g., units, subscripts, choice letters in multiple-choice).
- If multiple possible answers are given, choose the one explicitly marked as **final** (e.g., after "The final answer is:").
- If the answer is conceptual (True/False or qualitative), extract it exactly as written.
Example 1 (Numerical)
Question:
What is the area of a circle with radius 5 cm?
Detailed Answer:
... derivation ...
Therefore, the area of the circle with radius 5 cm is 78.54 cm^2.
The final answer is: 78.54 cm^2
Reference Answer:
78.54 cm^2
Example 2 (Physics Formula)
Question:
Write the secular equation for eigenfrequencies of small oscillations.
Detailed Answer:
... derivation ...
The final answer is: $\det(K - \omega^2 M) = 0$
Reference Answer:
$\det(K - \omega^2 M) = 0$
Example 3 (Multiple parts)
Question:
(a) Compute velocity … (b) State direction …
Detailed Answer:
... derivation ...
The final answers are: (a) 4.0 m/s^2 (b) leftward
Reference Answer:
(a) 4.0 m/s^2 (b) leftward
Instructions
1. Read the question carefully.
2. Identify the **explicit final answer(s)** in the detailed solution.
3. Extract them completely (including numbers, units, or words).
4. If multiple parts exist, present each part clearly and separately.
5. Do not copy any intermediate steps or reasoning—only the final results.
Question:
`<PROBLEM>`
Detailed Answer:
`<ANSWER>`
Now process and return the result.

Figure 13: Prompt for reference answer extraction in physics.

Solid State Physics

Explain how the correlation strength in strongly correlated electron systems affects the quantum oscillations observed in magnetization and resistance under high magnetic fields.

Astrophysics

A cataclysmic variable star system exhibits a periodic brightness variation during its superoutburst phase. This phenomenon, known as a superhump, is caused by the precession of an eccentric accretion disk. Given that the orbital period of the binary system is $P_{\mathrm{orb}} = 0.15$ days and the superhump period is observed to be $P_{\mathrm{sh}} = 0.153$ days, calculate the fractional change in the superhump period $\Delta P_{\mathrm{sh}}/P_{\mathrm{sh}}$. Furthermore, discuss how this fractional change can provide insights into the mass ratio $q = M_2/M_1$ between the secondary star (M_2) and the white dwarf primary (M_1).

Organic Chemistry

An alkene which is least reactive towards electrophilic addition, among the-following is (a) $H_2C = CH - Cl$ (b) $(CH_3)_2\,C = CH_2$ (c) $(CH_3)_2\,C = C\,(CH_3)_2$ (d) $ClCH_2CH = CH_2$

Physical Chemistry

Assume you dissolve 0.303 g of benzoic acid in enough water to make 100 mL of solution and then titrate the solution with 0.178 M NaOH. What was the pH of the original benzoic acid solution?

Pharmacy

Which of the following is a TRUE statement?Options: A. Because of the pharmacological processes of ethanol in the body, blood alcohol levels can decrease faster than they can rise. B. Because of the pharmacological processes of ethanol in the body, blood alcohol levels can rise faster than they can decrease. C. None of the answers are correct. D. Blood alcohol levels are not related to the pharmacological processes involving ethanol. E. Ethanol never leaves the body, but is used up as energy.

Zoology

Primates can be grouped into two main categories. Which of the following correctly pairs the group with the location? Options: A. Old World: Asia, Africa and New World: South America B. Old World: Africa and New World: South America only C. New World: South America only D. Old World: Africa only E. Old World: Asia only

Figure 14: More cases of Open-Sci.

Table 11: Evaluation results on Llama-3.1-8B. **Distill-70B**: data distilled directly from Llama-3.1-70B-Instruct; **+ Open-Sci**: fine-tuning on Open-Sci.

Benchmark	Base	Distill-70B	+ Open-Sci
ChemBench	40.14	**50.41**	40.94
ProteinLMBench	55.40	54.66	**57.52**
MedQA	65.60	66.61	**67.99**
PHYSICS	7.65	11.67	**15.13**
MMLU-Pro	48.08	52.56	**54.91**

Table 12: Performance of Gemma-3-4B before and after fine-tuning on Open-Sci.

Benchmark	Gemma-3-4B	+ Open-Sci	Δ
MMLU-Pro	42.35	**48.94**	+6.59
MedQA	49.87	**56.00**	+6.13
ChemBench	37.62	**39.30**	+1.68
ProteinLMBench	52.40	**53.80**	+1.40
PHYSICS	9.79	**11.07**	+1.28
PubMedQA	72.60	**73.40**	+0.80

dataset's effectiveness does not depend on a Qwen- or Llama-specific pretraining recipe.