WildSci: Advancing Scientific Reasoning from In-the-Wild Literature

Tengxiao Liu Deepak Nathani Zekun Li Kevin Yang William Yang Wang
University of California, Santa Barbara
tengxiao@ucsb.edu

Abstract

Recent progress in large language model (LLM) reasoning has focused on domains like mathematics and coding, where abundant high-quality data and objective evaluation metrics are readily available. In contrast, progress in LLM reasoning models remains limited in scientific domains such as medicine and materials science due to limited dataset coverage and the inherent complexity of open-ended scientific questions. To address these challenges, we introduce WildSci, a new dataset of domain-specific science questions automatically synthesized from peer-reviewed literature, covering 9 scientific disciplines and 26 subdomains. By framing complex scientific reasoning tasks in a multiple-choice format, we enable scalable training with well-defined reward signals. We further apply reinforcement learning to finetune models on these data and analyze the resulting training dynamics, including domain-specific performance changes, response behaviors, and generalization trends. Experiments on a suite of scientific benchmarks demonstrate the effectiveness of our dataset and approach. We release WildSci to enable scalable and sustainable research in scientific reasoning. [1]

1 Introduction

Advancing AI for science requires models that combine domain-specific expertise with strong reasoning capabilities to support real-world scientific discovery [Zhang et al., 2024b, Nathani et al., 2025, Gottweis et al., 2025, Prabhakar et al., 2025]. Recent advances in large language model (LLM) reasoning have achieved impressive progress in mathematical and coding domains [Yang et al., 2024b, Shao et al., 2024], especially driven by the rapid development of reinforcement learning techniques for reasoning tasks [Shao et al., 2024, Team et al., 2025a, Hu et al., 2025, Yu et al., 2025, Liu et al., 2025b, Zeng et al., 2025]. These areas offer natural advantages: they involve objective, verifiable answers and have abundant high-quality datasets available in the community. This enables the development of reasoning models with access to scalable, well-structured training and evaluation.

In contrast, scientific domains remain relatively underexplored in the context of reinforcement learning with verifiable rewards (RLVR) [Lambert et al., 2024, Su et al., 2025]. Scientific questions are often more complex and multifaceted, requiring not only logical and mathematical skills but also deep domain-specific knowledge to contextualize and analyze information accurately. Existing datasets are mostly skewed toward traditional natural sciences such as physics, chemistry, and biology, and tend to be limited in both scope and disciplinary diversity [Rein et al., 2023, Li et al., 2023, Zhang et al., 2024a, Lu et al., 2025]. As a result, many scientific areas, particularly interdisciplinary fields like materials science and medicine, remain underrepresented [Wadden et al., 2024]. Furthermore, most existing data sources are drawn from textbooks or general pretraining corpora [Yue et al., 2024b, Yuan et al., 2025], which lack the specificity of research-level content.

[1] https://huggingface.co/datasets/JustinTX/WildSci

Table 1: Comparison of datasets by source type, domain coverage, size, question length, and originality. 'New?' indicates whether questions are newly generated or parsed from existing corpora.

Dataset	Source	Domains	# Q	Avg. Len.	New?
Camel-AI Science	GPT-4 (self-generated)	Phys, Bio, Chem	60K	30 ± 15	Yes
Sci-Instruct	Textbooks, problem sets, websites	Phys, Chem, Math, Formal Proofs	254K	41 ± 33	No
SCP-116K	Educational materials	Phys, Chem, Bio	116K	62 ± 74	No
Natural Reasoning	Pretraining corpora	Multiple	2.8M	55 ± 21	Yes
WildSci	Peer-reviewed papers	Multiple (research focused)	56K	82 ± 19	Yes

To address this gap, we propose leveraging peer-reviewed scientific literature as a rich yet underutilized source for constructing domain-specific science questions in a fully automated pipeline. Unlike textbooks or problem sets, scientific papers reflect the depth, rigor, and complexity of real-world research, making them well-suited for advancing models toward research-level reasoning skills. This approach offers several key advantages: (1) it grounds questions in real-world applications and expert-validated context; and (2) it enables the creation of new questions that are unlikely to appear in pretraining corpora, helping mitigate issues of data contamination.

Another key challenge arises from the nature of science itself: many science questions are inherently open-ended and do not have a single verifiable answer. For example, explaining the observed decline in species richness in Figure 1 requires scientific judgment including interpreting evidence, reasoning about underlying mechanisms, and constructing plausible explanations. To address this, we adopt a more structured formulation by framing scientific reasoning tasks as multiple-choice questions (MCQs). MCQs are widely used in existing science benchmarks and offer a practical format for evaluation [Hendrycks et al., 2021a, Rein et al., 2023, Wang et al., 2024b, Team et al., 2025b]. This structure provides clear supervision signals, making it easier to define rewards systematically while preserving the richness of scientific reasoning. This simple setting offers a natural testbed for extending RL advances from mathematical domains to scientific reasoning tasks.

In this work, we develop a generalizable approach for creating training data grounded in real-world scientific research, and to extend RLVR reasoning to scientific domains. Our contributions are summarized as follows.

(1) We introduce a **fully automated data synthesis pipeline** that generates domain-specific questions from peer-reviewed scientific papers, followed by refinement and model voting to ensure data quality.

(2) We construct WildSci, a dataset of 56K questions **spanning 9 scientific disciplines and 26 subdomains**, providing broad and diverse coverage for scientific reasoning.

(3) We provide comprehensive analysis on how WildSci enables **effective transfer of RLVR method to scientific domains**. Models trained on WildSci show consistent improvements on multiple science benchmarks, including GPQA, SuperGPQA, and MMLU-Pro.

With new papers continuously emerging in the community, WildSci provides a sustainable data synthesis approach to support ongoing exploration of scientific reasoning. We have open-sourced the code and data for WildSci to enable scalable and sustainable research in scientific reasoning.

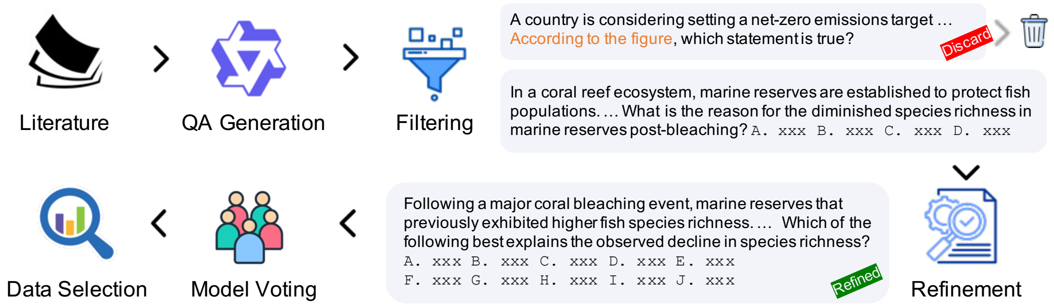
2 WildSci

2.1 Data Creation

An overview of our data creation pipeline is illustrated in Figure 1. The entire process is fully automated using large language models, with multiple stages of filtering and refinement to ensure high-quality outputs. This automation enables our pipeline to generalize seamlessly to other scientific domains with accessible research literature.

Peer-reviewed Papers We use publicly available, open-access articles from Nature Communications[2] as the data source [Li et al., 2024]. The journal categorizes its content into five major areas and 72 subdomains. We reorganize these into nine broader disciplines following the taxonomy of SuperGPQA [Team et al., 2025b]. To ensure balanced coverage, we randomly sample a subset of

[2] https://www.nature.com/ncomms/



A country is considering setting a net-zero emissions target … According to the figure, which statement is true?

In a coral reef ecosystem, marine reserves are established to protect fish populations. … What is the reason for the diminished species richness in marine reserves post-bleaching? A. xxx B. xxx C. xxx D. xxx

Following a major coral bleaching event, marine reserves that previously exhibited higher fish species richness. … Which of the following best explains the observed decline in species richness? A. xxx B. xxx C. xxx D. xxx E. xxx F. xxx G. xxx H. xxx I. xxx J. xxx

Figure 1: Overview of the data creation pipeline. Filtering is based on heuristic rules, while refinement expands the option space and rephrases questions to increase diversity.

papers from each category and generate three questions per paper based on its content. While the articles include both text and visuals, we focus exclusively on textual reasoning by using only the title, abstract, and main body, excluding all figures and tables.

QA Generation To enable a fully automated data synthesis pipeline, we employ a large language model to generate multiple-choice questions, corresponding answers and rationales directly from the paper content. The model is prompted to go through the full paper and create questions that reflect in-depth reasoning and understanding. Specifically, we instruct the model to produce context-independent questions. These are questions that can be answered without relying on figures, tables, or precise numerical details from the paper. This constraint ensures that the resulting questions emphasize knowledge-based and reasoning-intensive skills within a standalone context. The full prompt used for QA generation is provided in Appendix I.

Filtering To eliminate questions that require fine-grained recall or external references, we apply a set of heuristic and keyword-based filters. These filters remove items involving specific sections, detailed experimental results, or figure and table references (Figure 1). Following previous work [Yuan et al., 2025, Gao et al., 2024], we also apply 13-gram deduplication against GPQA, SuperGPQA, and MMLU-Pro to eliminate semantically similar questions. The deduplication overlap rate is 0.0%, suggesting that our dataset is both new and free of substantial overlap with existing resources.

Refinement The refinement stage focuses on enhancing question quality by increasing both their difficulty and diversity. Each question, along with its answer choices and rationale, is passed to LLM, which is prompted to paraphrase the question, eliminate surface-level cues, and expand the number of options (e.g., from 4 to 10 choices). This augmentation not only makes the questions more challenging but also reduces the chance of correct guesses during training. To better assess domain expert level reasoning, we explicitly instruct the model to remove well-known axioms or clues that are too obvious within a specific scientific field.

Model Voting Dataset quality can be negatively impacted by questions that are poorly constructed, lack necessary information, or are inherently unanswerable [Vendrow et al., 2025]. A practical way to assess answerability is to have models attempt to solve the questions. To validate the clarity of generated data, we apply model-based voting using an ensemble of open-source LLMs. Each model receives the question and answer choices, along with an additional fallback option: "None of the above / The question is unanswerable." This allows models to flag ambiguous or ill-formed questions while trying to approach the solutions. Based on the model voting results, we discard questions where the majority of models select the unanswerable option.

Data Selection We further categorize the questions by the level of agreement among ensemble models, using this as a proxy for clarity and difficulty: This grouping supports more controlled training setups, enabling models to learn from data with varying complexity:
All Aligned All model responses match the synthetic answer, indicating the question is clear and easily answerable.
Majority Aligned The majority vote of the model responses matches the original labels. These

questions are valid but more challenging, as models do not consistently arrive at the correct answer.
Majority Divergent The majority of responses differ from the original labels. This indicates either that the questions are challenging or that the synthetic labels may be incorrect, with the majority answers potentially representing more reasonable solutions.
All Divergent No single answer is selected by more than half of the ensemble responses. Such questions are likely highly difficult or ambiguous, resulting in diverse interpretations.

2.2 Statistics

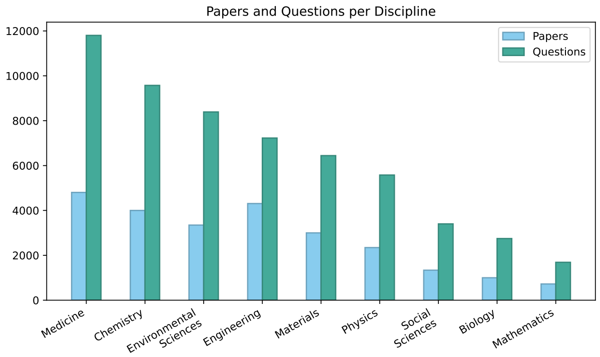


Figure 2: Comparison of the number of papers and generated questions across different disciplines.

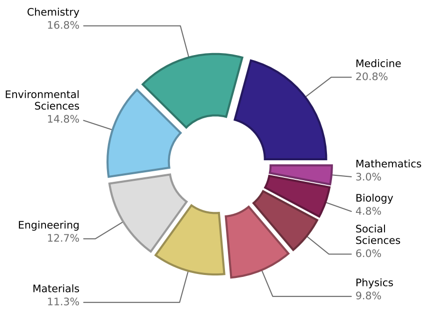


Figure 3: Distribution of questions after filtering and refinement in WildSci.

Domains The distribution of selected papers and the resulting question dataset after filtering and refinement are illustrated in Figure 2 and Figure 3. Note that many subdomains are interdisciplinary, for example, Biochemistry can fall under both Biology and Chemistry. This classification serves primarily to provide a high-level overview of the dataset composition. A comprehensive taxonomy and detailed subdomain statistics are included in Appendix D.

Length We measure question length by counting the number of words in each question, excluding the options. While word count does not directly indicate difficulty, it serves as a proxy for the complexity and richness of the question descriptions. Grounded in real-world research, questions in WildSci often include background information to provide context. Compared to other datasets in Table 1, WildSci has the longest questions, with a mean length of 81.74 words. Among different disciplines, Mathematics has the highest average word count (93.9), followed by Physics (89.8) and Engineering (88.6). In contrast, questions in Biology are the shortest on average, with 69.9 words. A detailed distribution of question lengths is provided in Appendix D.

2.3 Training

Reward Design In order to enable RLVR in science domains, WildSci allows obtaining verifiable rewards during training through simple option answer matching. We denote the original labels produced by our data creation pipeline as synthetic labels, denoted $y_{\texttt{syn}}$. For each question, there is exactly one such label generated automatically from the pipeline. In this setting, we define the reward function based solely on matching the prediction $\hat{y}$ and the synthetic label $y_{\texttt{syn}}$:

$$\mathcal{R}_{\texttt{syn}}(\hat{y}) = \begin{cases} 1.0, & \text{if } \hat{y} = y_{\texttt{syn}}, \\ 0.0, & \text{otherwise.} \end{cases} \tag{1}$$

To prevent models from memorizing option positions, we randomly shuffle the choices in each training epoch, ensuring a balanced distribution and reducing overfitting to option labels.

Training Algorithm Following previous work that advances mathematical reasoning of LLMs, we adopt Group Relative Policy Optimization (GRPO) as the training algorithm [Shao et al., 2024].

3 Experiments

3.1 Evaluation Benchmarks

To assess the scientific reasoning capabilities of language models, we evaluate them on a suite of benchmarks targeting scientific question answering. We report accuracy for each dataset by checking whether the final selected option in the model's response is correct.

WildSci-Val We construct an in-domain validation set using All Aligned questions from WildSci. Specifically, we randomly sample 100 questions from each of the 9 disciplines, resulting in a total of 900 questions. Each question includes 10 available options, with exactly one correct answer.

GPQA To mitigate potential bias introduced by answer choice order, we introduce **GPQA-Aug**, an augmented variant of the original GPQA-Diamond dataset [Rein et al., 2023]. For each of the 198 questions, we generate four versions by permuting the answer options so that the correct answer appears once in each of the four positions, resulting in a total of 792 examples.

SuperGPQA SuperGPQA [Team et al., 2025b] is a large-scale benchmark designed to evaluate knowledge and reasoning across 285 graduate-level disciplines, with 26,529 questions in total.

MMLU-Pro MMLU-Pro [Wang et al., 2024b] is a dataset of 12,032 questions that builds upon the original MMLU [Hendrycks et al., 2021a] by shifting its focus to reasoning-based questions.

3.2 Data Creation

We adopt `Qwen2.5-32B-Instruct` [Yang et al., 2024a] and `Qwen3-32B` [Shao et al., 2024] as the default model in our data generation pipeline, balancing quality and cost-effectiveness, as supported by findings in Moshkov et al. [2025]. For the model voting stage, we additionally employ `Mistral-Small-24B-Instruct-2501`, selected for their similar performance in scientific reasoning tasks. We generate 4 responses per model using temperature of 0.8, resulting in a total of 8 responses used in the voting process.

4 Results and Analysis

4.1 Improving Science Reasoning Abilities

Table 2: Performance comparison across different benchmarks. The Average column is computed across the three public benchmarks GPQA-Aug, SuperGPQA and MMLU-Pro. Maj. Aligned stands for the Majority Aligned subset of WildSci.

Model	**WildSci-Val**	**GPQA-Aug**	**SuperGPQA**	**MMLU-Pro**	**Average**
Qwen2.5-1.5B-Instruct	$46.70_{1.98}$	$23.98_{0.83}$	$18.10_{0.63}$	$31.47_{0.84}$	$24.52_{0.76}$
+ WildSci All Aligned	$\mathbf{80.48}_{0.26}$	$\mathbf{28.95}_{0.08}$	$23.85_{0.25}$	$\mathbf{42.54}_{0.28}$	$\mathbf{31.78}_{0.20}$
+ WildSci Maj. Aligned	$80.33_{0.11}$	$25.76_{0.44}$	$\mathbf{24.41}_{0.06}$	$41.95_{0.03}$	$30.71_{0.17}$
Qwen2.5-3B-Instruct	$72.45_{0.40}$	$28.03_{1.97}$	$23.21_{0.08}$	$44.18_{0.15}$	$31.80_{0.73}$
+ WildSci All Aligned	$\mathbf{85.00}_{0.40}$	$\mathbf{33.04}_{0.86}$	$\mathbf{26.39}_{0.38}$	$\mathbf{49.33}_{0.27}$	$\mathbf{36.25}_{0.50}$
+ WildSci Maj. Aligned	$84.71_{0.06}$	$30.98_{0.89}$	$26.01_{0.24}$	$48.66_{0.10}$	$35.22_{0.41}$

In this section, we apply GRPO to train models on different subsets of WildSci. Table 2 reports the main results across multiple evaluation sets. WildSci-Val shows in-domain performance, while the Average column reflects mean accuracy across three out-of-domain science benchmarks: GPQA-Aug, SuperGPQA, and MMLU-Pro. Training `Qwen2.5-1.5B-Instruct` on the All Aligned subset significantly boosts in-domain accuracy from 46.7% to 80.48%. It also improves generalization to public benchmarks, with gains of 4.97%, 5.75%, and 11.07% on GPQA-Aug, SuperGPQA, and MMLU-Pro respectively, resulting in an average increase of 7.26%. Consistent gains are also observed with 3B models, where WildSci yields an average improvement of 4.45% across science benchmarks. While training on the Majority Aligned subset also brings similar improvements, its

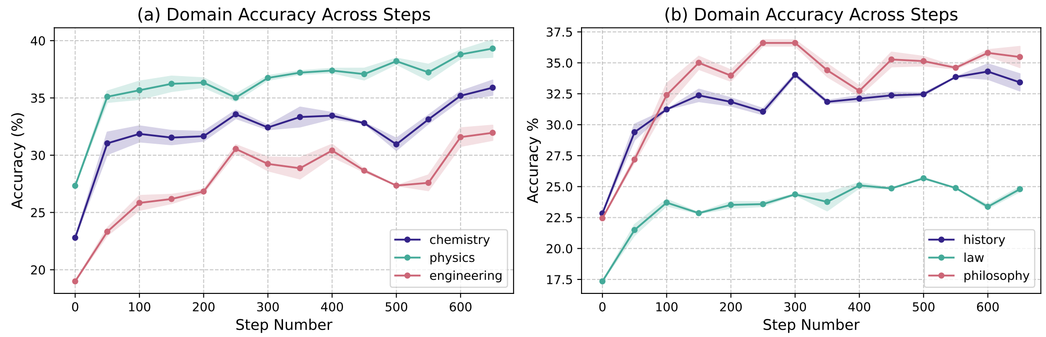


Figure 5: Domain-specific accuracy trends on MMLU-Pro during training. We report mean accuracy across three runs for each domain. Shaded regions indicate standard deviation. (a) shows steady improvements in domains with higher WildSci coverage (chemistry, physics, engineering), while (b) illustrates more variable performance in domains with lower coverage (law, history, philosophy).

performance is slightly lower than the All Aligned subset. Overall, we do not observe substantial performance differences across subsets, aligning with findings from Wang et al. [2025]. A detailed distribution and comparison of subsets is provided in Figure 6 and Appendix E.

Furthermore, we observe that **models continue to generalize even after overfitting on the validation set**. As illustrated in Figure 4, while the model's performance on in-domain valid set begins to decline sharply after step 400, its accuracy on OOD test datasets continues to improve. This pattern mirrors the post-saturation generalization phenomenon identified by Wang et al. [2025] in math domain. The presence of this generalization behavior indicates WildSci's potential as an ideal testbed for investigating RL reasoning in scientific contexts.

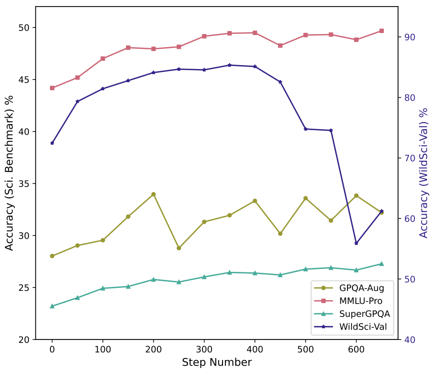


Figure 4: Performance trends on validation and test sets during training of the 3B model on WildSci All Aligned. The model exhibits continued generalization on test sets even after overfitting on the validation set.

4.2 Domain-Specific Performance Dynamics

Figure 5 illustrates the performance trends across subdomains of MMLU-Pro during training on a 1.5B model with WildSci Majority Aligned data. We observe a sharp increase in accuracy within the first 100 steps, likely due to the model's rapid adaptation to the multiple-choice answer format. After this initial phase, performance stabilizes but exhibits domain-specific dynamics. Interestingly, we find that in domains such as chemistry, physics, and engineering, where WildSci provides more coverage, accuracy continues to improve steadily throughout training (Figure 5(a)). In contrast, domains with sparser data coverage, such as law, history, and philosophy, show more fluctuation and slower gains (Figure 5(b)). The observed performance differences highlight the need for more balanced data coverage across domains to ensure uniform model improvements. A detailed breakdown of domain-level performance on SuperGPQA and MMLU-Pro can be found in Appendix F.

4.3 Analyzing Format Alignment and Reasoning Improvement

As our training data consists of multiple-choice questions, the model naturally adapts to the expected answer format and becomes familiar with the task structure. This raises the question: *are the observed improvements on science reasoning benchmarks only a result of format alignment, or do they reflect gains in reasoning ability?*

To disentangle the effect of format adaptation from actual reasoning improvement, we track the model's adherence to the expected answer format over training steps. Figure 7 presents both the proportion of responses from which a valid final answer can be reliably extracted and the corresponding accuracy trends.

Our analysis shows that even by step 5, the model achieves a high extraction success rate of 88.86%, which quickly converges to around 95% within just 20 steps. This rapid convergence suggests that **the model learns the structural template of multiple-choice answers very early during training**. Furthermore, we continue to observe steady improvements in accuracy beyond this point, indicating that the later-stage performance gains cannot be attributed solely to improved format adaptation. Instead, these results indicate **a gradual enhancement in the model's reasoning capabilities** as it attempts and learns from more diverse responses during training.

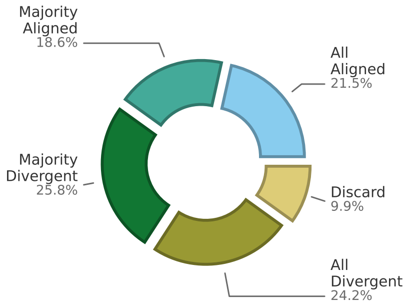


Figure 6: Proportion of different data splits within WildSci.

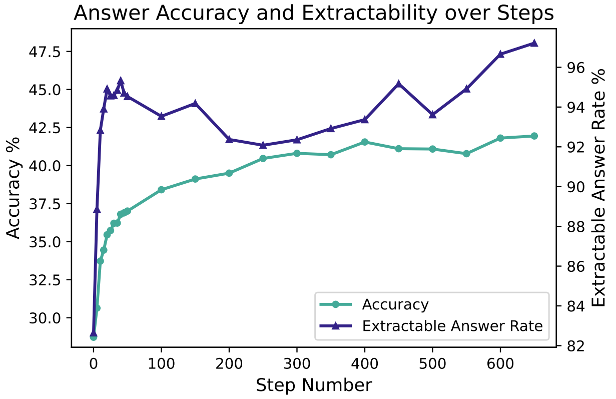


Figure 7: Performance change on subdomains from MMLU-Pro on a trained 1.5B model with WildSci Maj. Aligned.

4.4 Ablation Analysis

In this section, we conduct an ablation analysis to examine the impact of question refinement on final model performance. During the refinement phase, we enhance both the semantic diversity of questions and the breadth of the answer option space. The results are presented in Table 3. In both settings, we train the `Qwen2.5-1.5B-Instruct` model using All Aligned questions from WildSci.

Even without refinement, the model achieves an improved average performance of 28.83%, representing a 4.31% gain over the baseline average accuracy of 24.52%. However, with only four options per question in the training set, the model only achieves 66.63% on the valid set which contains 10 options per question. Notably, the expanded answer space introduced during refinement not only balances the distribution among the answer space, but also reduces the likelihood of the model arriving at the correct answer through random guessing, thereby encouraging more robust learning and leading to further performance improvements on OOD science benchmarks.

Table 3: Performance comparison before and after refinement stage. The average is computed across the three public benchmarks GPQA-Aug, SuperGPQA and MMLU-Pro. We use All Aligned data as the training data in both settings.

Setting	**WildSci-Val**	**GPQA-Aug**	**SuperGPQA**	**MMLU-Pro**	**Average**
WildSci	**80.48**$_{0.26}$	**28.95**$_{0.08}$	**23.85**$_{0.25}$	**42.54**$_{0.28}$	**31.78**$_{0.20}$
WildSci w/o refinement	66.63$_{0.06}$	27.94$_{0.32}$	20.44$_{0.04}$	38.11$_{0.37}$	28.83$_{0.24}$

4.5 Mixed Training with MATH

To assess the transferability between mathematical and scientific reasoning, we conduct experiments using the MATH dataset [Hendrycks et al., 2021b] for both training and evaluation of math reasoning abilities. As shown in Table 4, model trained solely on MATH achieves improved performance on math-specific tasks but fails to generalize to science reasoning benchmarks. In contrast, training solely on WildSci also leads to improvements in math reasoning, indicating its broader effectiveness. We then perform mixed training using both MATH and WildSci, denoted as "MATH+WildSci". The resulting model demonstrates improved performance on science benchmarks while preserving its math reasoning ability. Moreover, it also improves accuracy on GPQA-Aug, highlighting the effect of combining science and math reasoning data. **These results suggest that WildSci complements existing reasoning datasets and helps enhance model generalization across diverse reasoning domains**.

Table 4: Performance comparison when training 1.5B model with MATH. The average is computed across the three public benchmarks GPQA-Aug, SuperGPQA and MMLU-Pro. We use Maj. Aligned subset as the training data for WildSci.

Setting	**WildSci-Val**	**GPQA-Aug**	**SuperGPQA**	**MMLU-Pro**	**Sci. Avg.**	**MATH**
Qwen2.5-1.5B-Instruct	$46.70_{1.98}$	$23.98_{0.83}$	$18.10_{0.63}$	$31.47_{0.84}$	$24.52_{0.76}$	$55.47_{0.83}$
WildSci	$80.33_{0.11}$	$25.76_{0.44}$	$24.41_{0.06}$	$41.95_{0.03}$	$30.71_{0.17}$	$57.00_{0.60}$
MATH	$36.96_{1.18}$	$20.54_{0.44}$	$17.47_{0.05}$	$28.86_{0.05}$	$22.29_{0.18}$	$58.92_{0.87}$
MATH + WildSci	$78.41_{0.34}$	$29.76_{0.45}$	$24.25_{0.12}$	$42.36_{0.09}$	$32.12_{0.22}$	$58.73_{0.99}$

4.6 Scientific Reasoning Type Analysis

To characterize the types of reasoning required in WildSci, two authors manually reviewed 200 randomly sampled questions and developed a classification scheme. Each question was categorized into one of the following three scientific reasoning types: (1) Mathematical calculations and derivations: Questions that require performing numerical computations or symbolic derivations. (2) Model design, method analysis, or conceptual understanding: Questions that involve understanding scientific models, evaluating methodologies, or grasping abstract concepts. (3) Causal reasoning and mechanism inference: Questions that ask about the underlying causes, mechanisms, or consequences of a phenomenon.

To extend this categorization to the full dataset, we use `Qwen2.5-32B-Instruct` to automatically classify all questions in WildSci. The resulting distribution is as follows: 40.00% of the questions involve numerical calculation, 37.59% require causal inference, and 22.41% focus on model analysis or conceptual understanding. While mathematical reasoning is common in other datasets, WildSci emphasizes higher-order scientific reasoning skills such as causal inference and model-based analysis. These types of abilities are essential in research-oriented problem solving, reflecting the complexity and diversity of scientific reasoning in our dataset.

5 Limitations

While involving domain-specific knowledge in the questions, some numerical questions in our dataset are relatively simple. Although complex mathematical reasoning is not the primary focus of this work, combining scientific knowledge with deeper quantitative reasoning is a promising future direction. Another limitation lies in the use of a multiple-choice question (MCQ) format. Despite our efforts to carefully design questions and diversify answer choices to reduce superficial patterns, the format inherently introduces the risk of models exploiting spurious heuristics. Evaluating and verifying open-ended research questions such as causal reasoning and analysis remains an open challenge in advancing scientific reasoning abilities of LLMs.

6 Conclusion

We present WildSci, a dataset of verifiable scientific reasoning questions automatically generated from peer-reviewed papers. Our pipeline synthesizes questions across 9 disciplines and 26 subdomains, using model voting for quality control and pairing each question with 10 answer options. RLVR training on WildSci leads to improved performance on multiple science reasoning benchmarks. Training dynamics further confirm the effectiveness and generalizability of our data. As scientific literature continues to grow, WildSci offers a sustainable approach to converting real world research articles into valuable data for advancing scientific reasoning in language models.

References

S. N. Akter, S. Prabhumoye, M. Novikov, S. Han, Y. Lin, E. Bakhturi, E. Nyberg, Y. Choi, M. Patwary, M. Shoeybi, and B. Catanzaro. Nemotron-crossthink: Scaling self-learning beyond math reasoning. 2025. URL https://api.semanticscholar.org/CorpusID:277955461.

A. Albalak, D. Phung, N. Lile, R. Rafailov, K. Gandhi, L. Castricato, A. Singh, C. Blagden, V. Xiang, D. Mahan, and N. Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *CoRR*, abs/2502.17387, 2025. doi: 10.48550/ARXIV.2502.17387. URL https://doi.org/10.48550/arXiv.2502.17387.

H. Cai, X. Cai, J. Chang, S. Li, L. Yao, C. Wang, Z. Gao, H. Wang, Y. Li, M. Lin, S. Yang, J. Wang, Y. Yin, Y. Li, L. Zhang, and G. Ke. Sciassess: Benchmarking LLM proficiency in scientific literature analysis. *CoRR*, abs/2403.01976, 2024. doi: 10.48550/ARXIV.2403.01976. URL https://doi.org/10.48550/arXiv.2403.01976.

G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. S. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, L. Marris, S. Petulla, C. Gaffney, A. Aharoni, N. Lintz, T. C. Pais, H. Jacobsson, I. Szpektor, N. Jiang, K. Haridasan, A. Omran, N. Saunshi, D. Bahri, G. Mishra, E. Chu, T. Boyd, B. Hekman, A. Parisi, C. Zhang, K. Kawintiranon, T. Bedrax-Weiss, O. Wang, Y. Xu, O. Purkiss, U. Mendlovic, I. Deutel, N. Nguyen, A. Langley, F. Korn, L. Rossazza, A. Ramé, S. Waghmare, H. Miller, N. Byrd, A. Sheshan, R. H. S. Bhardwaj, P. Janus, T. Rissa, D. Horgan, S. Silver, A. Wahid, S. Brin, Y. Raimond, K. Kloboves, C. Wang, N. B. Gundavarapu, I. Shumailov, B. Wang, M. Pajarskas, J. Heyward, M. Nikoltchev, M. Kula, H. Zhou, Z. Garrett, S. Kafle, S. Arik, A. Goel, M. Yang, J. Park, K. Kojima, P. Mahmoudieh, K. Kavukcuoglu, G. Chen, D. Fritz, A. Bulyenov, S. Roy, D. Paparas, H. Shemtov, B. Chen, R. Strudel, D. Reitter, A. Roy, A. Vlasov, C. Ryu, C. Leichner, H. Yang, Z. Mariet, D. Vnukov, T. Sohn, A. Stuart, W. Liang, M. Chen, P. Rawlani, C. Koh, J. Co-Reyes, G. Lai, P. Banzal, D. Vytiniotis, J. Mei, and M. Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, A. Huang, S. Zhang, K. Chen, Z. Yin, Z. Shen, J. Ge, and V. Ng. Lawbench: Benchmarking legal knowledge of large language models. In Y. Al-Onaizan, M. Bansal, and Y. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7933–7962. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.452.

K. Feng, K. Ding, W. Wang, X. Zhuang, Z. Wang, M. Qin, Y. Zhao, J. Yao, Q. Zhang, and H. Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *CoRR*, abs/2406.09098, 2024. doi: 10.48550/ARXIV.2406.09098. URL https://doi.org/10.48550/arXiv.2406.09098.

B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu, and B. Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=yaqPf0KAlN.

L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

J. Gottweis, W. Weng, A. N. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, K. Saab, D. Popovici, J. Blum, F. Zhang, K. Chou, A. Hassidim, B. Gokturk, A. Vahdat, P. Kohli, Y. Matias, A. Carroll, K. Kulkarni, N. Tomasev, Y. Guan, V. Dhillon, E. D. Vaishnav, B. Lee, T. R. D. Costa, J. R. Penadés, G. Peltz, Y. Xu, A. Pawlosky, A. Karthikesalingam, and V. Natarajan. Towards an AI co-scientist. *CoRR*, abs/2502.18864, 2025. doi: 10.48550/ARXIV.2502.18864. URL https://doi.org/10.48550/arXiv.2502.18864.

A. Gulati, B. Miranda, E. Chen, E. Xia, K. Fronsdal, B. de Moraes Dumont, and S. Koyejo. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning. *39th International Conference on Machine Learning (ICML 2025)*, 2025. Preprint available at: https://openreview.net/pdf?id=YXnwlZe0yf, ICML paper: https://openreview.net/forum?id=kqj2Cn3Sxr.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021b. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html`.

J. Hu, Y. Zhang, Q. Han, D. Jiang, X. Zhang, and H. Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *CoRR*, abs/2503.24290, 2025. doi: 10.48550/ARXIV.2503.24290. URL `https://doi.org/10.48550/arXiv.2503.24290`.

D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, and H. Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024. doi: 10.48550/ARXIV.2411.15124. URL `https://doi.org/10.48550/arXiv.2411.15124`.

G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: communicative agents for "mind" exploration of large language model society. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html`.

Z. Li, X. Yang, K. Choi, W. Zhu, R. Hsieh, H. Kim, J. H. Lim, S. Ji, B. Lee, X. Yan, L. R. Petzold, S. D. Wilson, W. Lim, and W. Y. Wang. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. *CoRR*, abs/2407.04903, 2024. doi: 10.48550/ARXIV.2407.04903. URL `https://doi.org/10.48550/arXiv.2407.04903`.

T. Liu, Q. Guo, X. Hu, Y. Zhang, X. Qiu, and Z. Zhang. RLET: A reinforcement learning based approach for explainable QA with entailment trees. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7177–7189. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.483. URL `https://doi.org/10.18653/v1/2022.emnlp-main.483`.

Y. Liu, Z. Yang, T. Xie, J. Ni, B. Gao, Y. Li, S. Tang, W. Ouyang, E. Cambria, and D. Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *CoRR*, abs/2503.21248, 2025a. doi: 10.48550/ARXIV.2503.21248. URL `https://doi.org/10.48550/arXiv.2503.21248`.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025b. doi: 10.48550/ARXIV.2503.20783. URL `https://doi.org/10.48550/arXiv.2503.20783`.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

D. Lu, X. Tan, R. Xu, T. Yao, C. Qu, W. Chu, Y. Xu, and Y. Qi. SCP-116K: A high-quality problem-solution dataset and a generalized pipeline for automated extraction in the higher education science domain. *CoRR*, abs/2501.15587, 2025. doi: 10.48550/ARXIV.2501.15587. URL https://doi.org/10.48550/arXiv.2501.15587.

X. Ma, Q. Liu, D. Jiang, G. Zhang, Z. Ma, and W. Chen. General-reasoner: Advancing llm reasoning across all domains. https://github.com/TIGER-AI-Lab/General-Reasoner/blob/main/General_Reasoner.pdf, 2025.

L. McInnes and J. Healy. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018. URL http://arxiv.org/abs/1802.03426.

I. Moshkov, D. Hanley, I. Sorokin, S. Toshniwal, C. Henkel, B. D. Schifferer, W. Du, and I. Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmath-reasoning dataset. 2025. URL https://api.semanticscholar.org/CorpusID:278000515.

D. Nathani, L. Madaan, N. Roberts, N. Bashlykov, A. Menon, V. Moens, A. Budhiraja, D. Magka, V. Vorotilov, G. Chaurasia, D. Hupkes, R. S. Cabral, T. Shavrina, J. N. Foerster, Y. Bachrach, W. Y. Wang, and R. Raileanu. Mlgym: A new framework and benchmark for advancing AI research agents. *CoRR*, abs/2502.14499, 2025. doi: 10.48550/ARXIV.2502.14499. URL https://doi.org/10.48550/arXiv.2502.14499.

A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. Flores, G. H. Chen, T. J. Pollard, J. C. Ho, and T. Naumann, editors, *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

V. Prabhakar, M. A. Islam, A. Atanas, Y. Wang, J. Han, A. Jhunjhunwala, R. Apte, R. Clark, K. Xu, Z. Wang, and K. Liu. Omniscience: A domain-specialized LLM for scientific reasoning and discovery. *CoRR*, abs/2503.17604, 2025. doi: 10.48550/ARXIV.2503.17604. URL https://doi.org/10.48550/arXiv.2503.17604.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Y. Su, D. Yu, L. Song, J. Li, H. Mi, Z. Tu, M. Zhang, and D. Yu. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *CoRR*, abs/2503.23829, 2025. doi: 10.48550/ARXIV.2503.23829. URL https://doi.org/10.48550/arXiv.2503.23829.

K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, C. Tang, C. Wang, D. Zhang, E. Yuan, E. Lu, F. Tang, F. Sung, G. Wei, G. Lai, H. Guo, H. Zhu, H. Ding, H. Hu, H. Yang, H. Zhang, H. Yao, H. Zhao, H. Lu, H. Li, H. Yu, H. Gao, H. Zheng, H. Yuan, J. Chen, J. Guo, J. Su, J. Wang, J. Zhao, J. Zhang, J. Liu, J. Yan, J. Wu, L. Shi, L. Ye, L. Yu, M. Dong, N. Zhang, N. Ma, Q. Pan, Q. Gong, S. Liu, S. Ma, S. Wei, S. Cao, S. Huang, T. Jiang,

W. Gao, W. Xiong, W. He, W. Huang, W. Wu, W. He, X. Wei, X. Jia, X. Wu, X. Xu, X. Zu, X. Zhou, X. Pan, Y. Charles, Y. Li, Y. Hu, Y. Liu, Y. Chen, Y. Wang, Y. Liu, Y. Qin, Y. Liu, Y. Yang, Y. Bao, Y. Du, Y. Wu, Y. Wang, Z. Zhou, Z. Wang, Z. Li, Z. Zhu, Z. Zhang, Z. Wang, Z. Yang, Z. Huang, Z. Huang, Z. Xu, and Z. Yang. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599, 2025a. doi: 10.48550/ARXIV.2501.12599. URL `https://doi.org/10.48550/arXiv.2501.12599`.

M. Team, X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, C. Zheng, K. Deng, S. Jia, S. Jiang, Y. Liao, R. Li, Q. Li, S. Li, Y. Li, Y. Li, D. Ma, Y. Ni, H. Que, Q. Wang, Z. Wen, S. Wu, T. Xing, M. Xu, Z. Yang, Z. M. Wang, J. Zhou, Y. Bai, X. Bu, C. Cai, L. Chen, Y. Chen, C. Cheng, T. Cheng, K. Ding, S. Huang, Y. Huang, Y. Li, Y. Li, Z. Li, T. Liang, C. Lin, H. Lin, Y. Ma, T. Pang, Z. Peng, Z. Peng, Q. Qi, S. Qiu, X. Qu, S. Quan, Y. Tan, Z. Wang, C. Wang, H. Wang, Y. Wang, Y. Wang, J. Xu, K. Yang, R. Yuan, Y. Yue, T. Zhan, C. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Y. Zhao, X. Zheng, C. Zhong, Y. Gao, Z. Li, D. Liu, Q. Liu, T. Liu, S. Ni, J. Peng, Y. Qin, W. Su, G. Wang, S. Wang, J. Yang, M. Yang, M. Cao, X. Yue, Z. Zhang, W. Zhou, J. Liu, Q. Lin, W. Huang, and G. Zhang. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *CoRR*, abs/2502.14739, 2025b. doi: 10.48550/ARXIV.2502.14739. URL `https://doi.org/10.48550/arXiv.2502.14739`.

S. Toshniwal, W. Du, I. Moshkov, B. Kisacanin, A. Ayrapetyan, and I. Gitman. Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=mTCbq2QssD`.

J. Vendrow, E. Vendrow, S. Beery, and A. Madry. Do large language model benchmarks test reliability? *CoRR*, abs/2502.03461, 2025. doi: 10.48550/ARXIV.2502.03461. URL `https://doi.org/10.48550/arXiv.2502.03461`.

D. Wadden, K. Shi, J. Morrison, A. Naik, S. Singh, N. Barzilay, K. Lo, T. Hope, L. Soldaini, S. Z. Shen, D. Downey, H. Hajishirzi, and A. Cohan. Sciriff: A resource to enhance language model instruction-following over scientific literature. *CoRR*, abs/2406.07835, 2024. doi: 10.48550/ARXIV.2406.07835. URL `https://doi.org/10.48550/arXiv.2406.07835`.

X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a. URL `https://openreview.net/forum?id=bq1JEgioLr`.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL `http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html`.

Y. Wang, Q. Yang, Z. Zeng, L. Ren, L. Liu, B. Peng, H. Cheng, X. He, K. Wang, J. Gao, W. Chen, S. Wang, S. S. Du, and Y. Shen. Reinforcement learning for reasoning in large language models with one training example. 2025. URL `https://api.semanticscholar.org/CorpusID:278171513`.

X. Xu, Q. Xu, T. Xiao, T. Chen, Y. Yan, J. Zhang, S. Diao, C. Yang, and Y. Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *CoRR*, abs/2502.00334, 2025. doi: 10.48550/ARXIV.2502.00334. URL `https://doi.org/10.48550/arXiv.2502.00334`.

A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang,

Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024b. doi: 10.48550/ARXIV.2409.12122. URL https://doi.org/10.48550/arXiv.2409.12122.

B. Yang, Q. Yang, and R. Liu. Utmath: Math evaluation with unit test via reasoning-to-coding thoughts. *CoRR*, abs/2411.07240, 2024c. doi: 10.48550/ARXIV.2411.07240. URL https://doi.org/10.48550/arXiv.2411.07240.

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, W. Dai, Y. Song, X. Wei, H. Zhou, J. Liu, W. Ma, Y. Zhang, L. Yan, M. Qiao, Y. Wu, and M. Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025. doi: 10.48550/ARXIV.2503.14476. URL https://doi.org/10.48550/arXiv.2503.14476.

W. Yuan, J. Yu, S. Jiang, K. Padthe, Y. Li, D. Wang, I. Kulikov, K. Cho, Y. Tian, J. E. Weston, and X. Li. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions. *CoRR*, abs/2502.13124, 2025. doi: 10.48550/ARXIV.2502.13124. URL https://doi.org/10.48550/arXiv.2502.13124.

X. Yue, X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su, and W. Chen. Mammoth: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=yLClGs770I.

X. Yue, T. Zheng, G. Zhang, and W. Chen. Mammoth2: Scaling instructions from the web. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html.

M. Zaki, Jayadeva, Mausam, and N. M. A. Krishnan. Mascqa: A question answering dataset for investigating materials science knowledge of large language models. *CoRR*, abs/2308.09115, 2023. doi: 10.48550/ARXIV.2308.09115. URL https://doi.org/10.48550/arXiv.2308.09115.

W. Zeng, Y. Huang, Q. Liu, W. Liu, K. He, Z. Ma, and J. He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *CoRR*, abs/2503.18892, 2025. doi: 10.48550/ARXIV.2503.18892. URL https://doi.org/10.48550/arXiv.2503.18892.

D. Zhang, Z. Hu, S. Zhoubian, Z. Du, K. Yang, Z. Wang, Y. Yue, Y. Dong, and J. Tang. Sciinstruct: a self-reflective instruction annotated dataset for training scientific language models. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/02ee6b7295f720407b56c457b34c54d5-Abstract-Datasets_and_Benchmarks_Track.html.

Y. Zhang, X. Chen, B. Jin, S. Wang, S. Ji, W. Wang, and J. Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In Y. Al-Onaizan, M. Bansal, and Y. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural*

Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pages 8783–8817. Association for Computational Linguistics, 2024b. URL https://aclanthology.org/2024.emnlp-main.498.

Q. Zhao, Y. Huang, T. Lv, L. Cui, Q. Sun, S. Mao, X. Zhang, Y. Xin, Q. Yin, S. Li, and F. Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. *CoRR*, abs/2412.15194, 2024. doi: 10.48550/ARXIV.2412.15194. URL https://doi.org/10.48550/arXiv.2412.15194.

F. Zhou, Z. Wang, N. Ranjan, Z. Cheng, L. Tang, G. He, Z. Liu, and E. P. Xing. Megamath: Pushing the limits of open math corpora. 2025. URL https://api.semanticscholar.org/CorpusID:277510325.

A Experiment Settings

We train our model using the GRPO algorithm implemented in the verl [Sheng et al., 2024] framework. We set a maximum response length of up to 8192 tokens and applying left-side truncation. During rollout, we sample 8 responses per prompt with a temperature of 1.0. The model is trained with a learning rate of 5×10^{-7} and is updated with AdamW optimizer [Loshchilov and Hutter, 2017]. We mostly adopt the training settings from Zeng et al. [2025] and do not perform additional hyperparameter tuning. All experiments are conducted on a server equipped with 8 NVIDIA A100 GPUs, each with 40GB of memory. Training the 1.5B model for 700 steps takes one day using 4 GPUs, while the 3B model requires two days on all 8 GPUs for 700 steps. We select checkpoints for evaluation based on the in-domain validation set WildSci-Val and the final training checkpoint at step 700. During the evaluation phase, we set the temperature to 0.0 using the vLLM library to eliminate randomness in model predictions [Kwon et al., 2023]. To show potential variance in inference, we report the mean and standard deviation over three independent runs.

B Related Work

B.1 Science Reasoning Data

Domain-specific datasets have significantly advanced model reasoning, especially in mathematics [Yu et al., 2024, Yue et al., 2024a, Toshniwal et al., 2025, Albalak et al., 2025, Zhou et al., 2025]. Recent work has extended coverage to core natural sciences like physics, chemistry, and biology [Li et al., 2023, Lu et al., 2025]. SciInstruct [Zhang et al., 2024a] collects questions from textbooks, exams, and problem sets, while Yue et al. [2024b], Yuan et al. [2025], Ma et al. [2025] synthesize questions from web text and pretraining corpora. In contrast, WildSci constructs questions from peer-reviewed research papers, targeting in-the-wild scientific problems and enabling generalization to long-tail domains through an automatic data synthesis pipeline.

B.2 Reinforcement Learning in Reasoning

Reinforcement learning has advanced reasoning in areas like mathematics and coding, where simple rule-based rewards have proven effective [Team et al., 2025a, Hu et al., 2025, Yu et al., 2025, Liu et al., 2025b, Zeng et al., 2025, Shao et al., 2024, Liu et al., 2022]. In science domains, open-ended questions can make rule-based verification challenging. For efficiency and simplicity, we reformat open science research questions as MCQs. While concurrent work validates the use of MCQs for RL training [Akter et al., 2025], WildSci specifically focuses on grounding its questions within real-world scientific literature. This approach not only ensures domain specificity but also provides a convenient training signal and facilitates generalization across diverse domains.

B.3 Science Reasoning Evaluation

Recent studies have developed benchmarks to address the limited evaluation of LLMs in the science domain [Team et al., 2025b, Fei et al., 2024, Pal et al., 2022, Jin et al., 2020, Zhao et al., 2024, Yang et al., 2024c, Gao et al., 2025, Gulati et al., 2025]. SciBench [Wang et al., 2024a] provides open-ended, free-response collegiate-level problems that require multi-step reasoning abilities, including advanced mathematical derivation and understanding of scientific concepts, in chemistry, physics, and math. Other benchmarks such as SciKnowEval [Feng et al., 2024] and SciAccess [Cai et al., 2024] evaluate LLMs in memorization, comprehension and reasoning across various science domains. Domain-specialized datasets like MaScQA [Zaki et al., 2023] and UGPhysics [Xu et al., 2025] further evaluate models within a specific subject. ResearchBench [Liu et al., 2025a] provides a benchmark to test LLMs on generating innovative scientific hypotheses, introducing a framework to assess inspiration retrieval, hypothesis composition, and hypothesis ranking. Although these are effective metrics to evaluate scientific proficiency, only a few domains are actively involved, mainly chemistry, biology, materials science, and physics. WildSci encompasses a wide range of scientific disciplines to support a comprehensive dataset. Through a multidisciplinary and multiple-choice approach, we envision WildSci to be an impactful and scalable resource for scientific research and model development.

C Data Quality Analysis

C.1 Validation of Synthetic Annotations

To evaluate the reliability of the synthetic labels in WildSci, we conduct a validation study using two strong commercial models, Gemini-2.5-Pro and Gemini-2.5-Flash [Comanici et al., 2025]. We randomly sample 500 questions from both the All-Aligned and Majority-Aligned subsets and independently prompt each model to generate answers. We then compare their responses against our synthetic labels to measure answer agreement.

On the All-Aligned subset, Flash and Pro agree with our synthetic labels in 95.0% and 96.0% of cases, respectively. The two Gemini models agree with each other in 94.0% of cases, and among these agreements, 98.9% match our synthetic labels. On the Majority-Aligned subset, Flash and Pro achieve 78.6% and 80.2% agreement with our labels, while their mutual agreement rate is 80.6%, with 88.8 of those matches aligning with our labels.

These results demonstrate that **WildSci's filtering and model-voting procedure produces annotations that are highly consistent with those from substantially stronger models**. Combined with the observed performance gains when training open source models, this analysis supports that the synthetic data quality in WildSci is sufficiently high for effective model training.

C.2 Question Redundancy

We assessed question similarity using SentenceTransformer [Reimers and Gurevych, 2019] and computed cosine similarity between question pairs within the same domain. We only consider question pairs with similarity ≥ 0.9, and we found that highly similar pairs account for 2.7% in "All-aligned" and 2.3% in "Majority-aligned", indicating low redundancy. While these pairs often share surface-level phrasing (e.g., from the same paper), manual inspection shows they frequently assess distinct concepts. For example, the following pair has a similarity score of 0.902 but asks fundamentally different things.

To assess potential redundancy in the WildSci dataset, we measured pairwise question similarity within each domain using the SentenceTransformer model and computed cosine similarity between question embeddings. We considered question pairs with similarity scores ≥ 0.9 as highly similar. Such pairs constitute only 2.7% of the All-Aligned subset and 2.3% of the Majority-Aligned subset, indicating low redundancy across questions. Manual inspection further reveals that these high-similarity pairs often share surface-level phrasing (e.g., derived from the same paper section) yet still assess distinct scientific concepts or reasoning steps.

To illustrate this, Table 5 presents a representative example of two highly similar questions (cosine similarity = 0.902). Despite their surface resemblance, the two queries probe distinct reasoning objectives—quantitative estimation versus selectivity comparison—highlighting that lexical overlap does not necessarily imply conceptual redundancy.

Table 5: Example of a high-similarity question pair (cosine similarity = 0.902) that differs in reasoning objective.

Q1	A hybrid absorption–adsorption system is being evaluated for CO_2 capture using a slurry composed of ZIF-8 suspended in a 2-methylimidazole-based glycol solution. *What is the total amount of CO_2 (in moles) that can be captured by 1 liter of the slurry under ideal behavior assumptions?*
Q2	A hybrid absorption–adsorption system for CO_2 capture employs a slurry of ZIF-8 in a 2-methylimidazole-based glycol solution. *What is the ratio of the amount of CO_2 captured to the amount of N_2 captured? Assume that the selectivity is defined as the ratio of the partition coefficients of the two gases.*

Overall, this analysis confirms that WildSci maintains **high question diversity** despite its large scale, with minimal duplication and strong coverage of distinct reasoning patterns even among lexically similar items.

C.3 Validity and Difficulty

To further evaluate data quality control, we conducted an additional analysis using Gemini-2.5-Pro [Comanici et al., 2025] to assess the validity and difficulty of questions across the four WildSci subsets. For each subset, we randomly sampled 500 examples and prompted the model to independently rate question clarity and difficulty.

Validity Evaluation For validity, the model was instructed to classify each question as either *good and clear* or *unanswerable*. As summarized in Table 6, 96.6% of questions in the All-Aligned (A.A.) subset and 89.4% in the Majority-Aligned (M.A.) subset were rated as good and clear, compared to 87.4% and 71.2% for the Majority-Divergent (M.D.) and All-Divergent (A.D.) subsets, respectively.

These results indicate that the vast majority of WildSci questions are coherent and answerable, particularly in the more reliable subsets. This finding aligns with the design intuition behind our model-voting phase and helps explain why the All-Aligned and Majority-Aligned subsets lead to stronger downstream training performance.

Table 6: Validity ratings across the four WildSci subsets. A large majority of questions, especially in the aligned subsets, were rated as clear and answerable.

Subset	All-Aligned	Majority-Aligned	Majority-Divergent	All-Divergent
Good & Clear (%)	96.6	89.4	87.4	71.2
Unanswerable (%)	3.4	10.6	12.6	28.8

Difficulty Evaluation We further asked Gemini-2.5-Pro to rate each question's difficulty on a five-level scale, from **Level 1 (Trivial)** to **Level 5 (Expert)**. As shown in Table 7, a substantial portion of questions in the All-Aligned (40.8%) and Majority-Aligned (59.0%) subsets were classified as Level 4–5, requiring undergraduate- or graduate-level domain expertise.

Overall, this analysis demonstrates that the question filtering and model-voting stages effectively select for both clarity and appropriate difficulty. The All-Aligned and Majority-Aligned subsets, in particular, comprise questions that are not only valid and well-posed but also intellectually challenging, supporting effective model training in various science reasoning domains.

Table 7: Distribution of difficulty levels across subsets. Levels 4–5 indicate field-specific undergraduate or graduate-level reasoning.

Subset	L1	L2	L3	L4	L5
A.A.	0.8	28.6	29.8	38.4	2.4
M.A.	0.0	12.8	28.2	52.4	6.6
M.D.	0.2	11.8	27.6	55.6	4.8
A.D.	0.0	6.6	22.8	59.8	10.8

Table 8: Difficulty rubric used in the difficulty evaluation.

Level	Description
5	Very Challenging: Expert-level (graduate/PhD) reasoning required.
4	Difficult: Field-specific undergraduate-level understanding.
3	Moderate: General background or non-specialist undergraduate-level knowledge.
2	Easy: General science or high-school-level knowledge.
1	Simple: Common-sense or minimal prior knowledge suffices.

D Statistics of Subdomains

We present the distribution of questions across 26 subdomains in Figure 8, following the categorization used by Nature Communications. Interdisciplinary areas such as Materials Science and Biochemistry

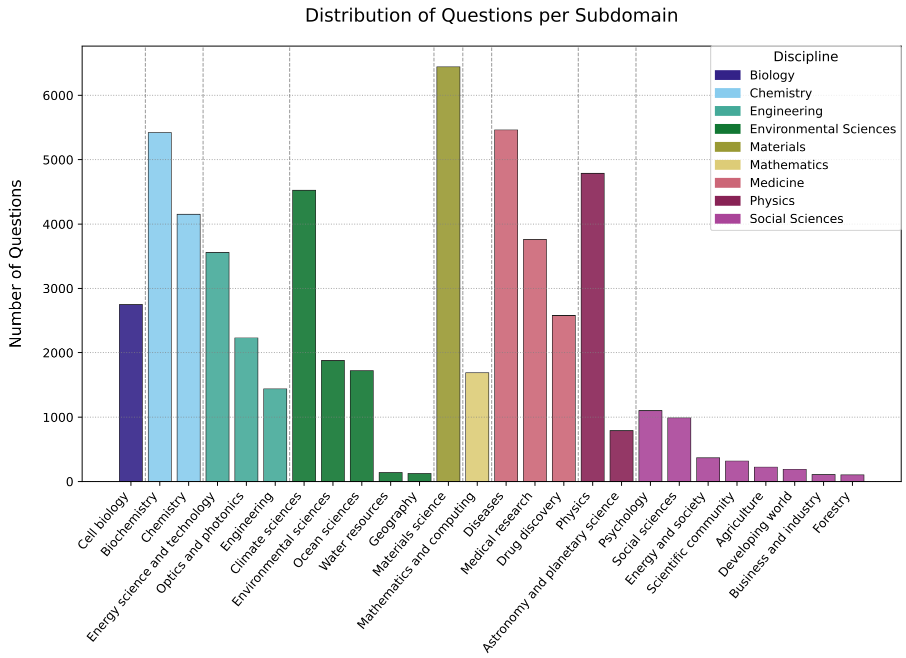


Figure 8: Question distribution across subdomains in WildSci.

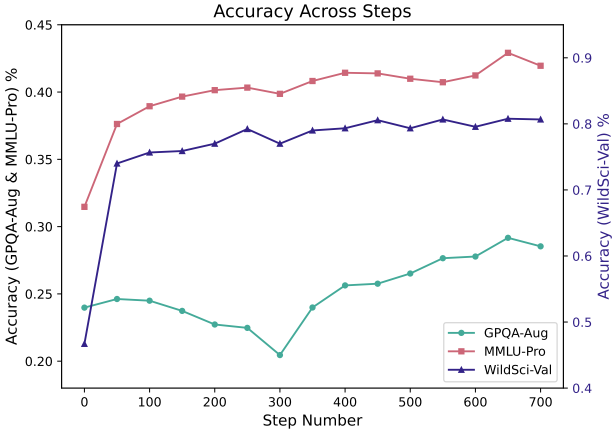


Figure 9: Performance trend on validation and OOD evaluation sets.

are more prevalent in the journal, leading to a higher number of papers, and consequently more questions from these domains. In contrast, Social Sciences represent a smaller proportion of the journal's content, so we include all available papers from this area in WildSci. The dataset is constructed from papers published prior to April 2024. As new publications become available, our data pipeline can be extended to incorporate the latest research, enabling continuous expansion of both dataset size and knowledge coverage in WildSci.

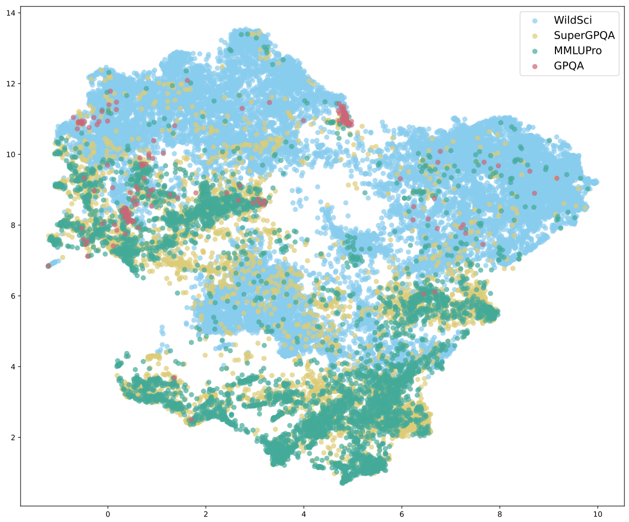


Figure 10: UMAP illustration.

E Different Data Splits

Table 9 compares model performance when trained on different combinations of WildSci subsets. We observe no substantial performance differences across the splits. We hypothesize that this is due to the exploratory and generalization-driven nature of reinforcement learning, which emphasizes domain alignment over data amount. We leave the exploration of how data difficulty and diversity affect RL-based reasoning as a promising direction for future work.

Table 9: Performance comparison across different benchmarks. The average is computed across the three public benchmarks GPQA-Aug, SuperGPQA and MMLU-Pro. A.A. means All Aligned, M.A. means Majority Aligned, M.D. means Majority Divergent.

Model	WildSci-Val	GPQA-Aug	SuperGPQA	MMLU-Pro	Average
Qwen2.5-1.5B-Instruct	$46.70_{1.98}$	$23.98_{0.83}$	$18.10_{0.63}$	$31.47_{0.84}$	$24.52_{0.76}$
WildSci A.A.	$\mathbf{80.48}_{0.26}$	$\mathbf{28.95}_{0.08}$	$23.85_{0.25}$	$\mathbf{42.54}_{0.28}$	$\mathbf{31.78}_{0.20}$
WildSci M.A.	$80.33_{0.11}$	$25.76_{0.44}$	$\mathbf{24.41}_{0.06}$	$41.95_{0.03}$	$30.71_{0.17}$
WildSci A.A.+M.A.+M.D.	$80.00_{0.59}$	$27.48_{0.19}$	$23.94_{0.01}$	$41.55_{0.17}$	$30.99_{0.12}$

F Domain-Specific Performance

We provide a breakdown of subdomain performance in the Table 10 and Table 11. In the model names, 'A.A.' refers to the All Aligned subset, and 'M.A.' refers to the Majority Aligned subset.

Table 10: Subdomain performance on SuperGPQA.

Model	Overall	Agro	Eco	Edu	Eng	Hist	Law	Lite	Mar	Medi	Mili	Phil	Sci	Socio
1.5B-A.A.	24.12	25.98	28.98	30.37	23.47	20.62	29.42	23.21	28.94	26.39	29.76	29.40	22.66	25.87
1.5B-M.A.	24.46	24.95	28.29	32.43	23.21	21.95	30.95	23.27	29.34	25.88	32.68	29.11	23.61	27.97
3B-A.A.	26.39	24.33	30.59	33.26	25.82	21.81	30.94	23.21	33.33	30.02	37.56	33.43	24.88	30.07
3B-M.A.	26.25	27.01	28.52	32.65	25.82	18.99	32.47	23.23	33.33	29.33	36.10	30.55	25.00	29.37

G Post-saturation Generalization

Beyond the 3B model, we observe similar trends with the 1.5B model. Figure 9 illustrates performance during training on the Majority Aligned subset. While validation accuracy fluctuates after step

Table 11: Subdomain performance on MMLU-Pro.

Model	Overall	Bio	Busi	CS	Econ	Health	Math	Other	Psyc	phys	Engin	Hist	law	Philo
1.5B-A.A.	42.92	64.85	50.82	46.59	54.03	38.02	57.44	35.93	53.26	40.49	29.10	34.38	22.89	36.87
1.5B-M.A.	41.98	64.99	48.29	40.73	53.44	36.80	55.07	35.07	53.26	39.49	30.44	32.55	24.80	35.27
3B-A.A.	49.48	70.71	58.56	50.73	60.43	48.04	63.43	46.54	58.77	49.04	36.02	37.01	27.16	39.08
3B-M.A.	48.73	70.57	56.78	47.32	58.06	48.17	62.47	43.83	57.90	49.73	37.26	36.75	27.79	37.88

400, accuracy on GPQA-Aug and MMLU-Pro continues to improve, indicating post-saturation generalization. We leave investigation of the underlying causes to future work.

H Distribution Illustration

To compare the coverage of WildSci with existing science benchmarks, we embed the question texts (excluding answer options) using a pretrained sentence embedding model and apply UMAP [McInnes and Healy, 2018] for dimensionality reduction. Figure 10 presents a 2D visualization of questions from different datasets. We randomly sample 5,000 questions from SuperGPQA, MMLU-Pro respectively and 20,000 questions from WildSci. In the figure, SuperGPQA and MMLU-Pro share similar distributions, WildSci occupies regions underrepresented by these benchmarks. The partial mismatch in distribution suggests that existing benchmarks serve as effective out-of-distribution (OOD) evaluations for WildSci.

I Prompts

In this section, we show the prompts we use in our data creation pipeline. During QA generation, we randomly sample two QAs from the GPQA-Extended set, excluding those in the Diamond subset, to serve as JSON examples.

QA Generation

You are an expert research scientist in [[DOMAIN]]. Your task is to extract or create three challenging and difficult, self-contained question–answer pairs from the provided academic paper. The QAs will be used as exam questions for PhD students and must be clear, extremely challenging, and context-independent, i.e., understandable on its own without referring back to the original paper.

Each QA pair must include:

1. Question:
- A clear, difficult, and standalone question.
- The question must include sufficient background information or context so that one can fully understand and attempt it without referring to the original paper.
- Define any abbreviations, notation, or domain-specific terminology used.
- DO NOT use phrases like "according to the paper" or "the proposed method."
- The question should be complex enough to require deep understanding of the subject.
- It should engage **advanced reasoning**, such as: Conceptual analysis, Theoretical or mathematical derivations, Methodological design, Causal reasoning or hypothesis testing, etc.

2. Options:
- Provide four answer choices.
- Only one option should be correct.
- The three incorrect options should be plausible but clearly wrong upon careful reasoning, ideally derived by subtly altering the logic or assumptions behind the correct answer.

3. Answer:
The correct option letter.

4. Rationale:
- A detailed explanation of why the correct answer is correct and why each incorrect option is wrong.
- DO NOT reference the original paper in any part of the rationale.
- If calculations are required, include the full step-by-step process.

Important:
- Questions must be self-contained, including any necessary context or definitions.
- Do not reference the original paper in any part of the question, options, or rationale.
- Aim for PhD-level difficulty, testing understanding of key technical ideas.
- Ensure that only one option is unambiguously correct.

Format your QA pairs in the following JSON format, here are examples: [[JSON example]]

Refinement

You are an expert research scientist in [[DOMAIN]].
Your task is to refine each provided multiple-choice question to increase its difficulty and test deeper reasoning appropriate for PhD-level understanding. Each item includes a question, answer options, and a solution. Use the provided answers and solution for reference, they may be incorrect.

Refinement Goals:
1. Expand Options with Subtle Variations
Rewrite the answer options to include 10 choices labeled A–J.
All options should seem plausible to someone with partial knowledge, but only one should be fully correct. Introduce subtle numerical or conceptual variations among options.
2. Remove Surface-Level Hints
Eliminate obvious formulas, definitions, or axioms from the question.
Assume the solver must recall or derive them independently. You may include these concepts in the solution rationale, but not in the question.
3. Increase Reasoning Depth
Replace direct or few-step problems with multi-step (>3) or causal reasoning.
Use reasoning chains (e.g., X leads to Y, which leads to Z, which explains W) or require intermediate inferences without giving all variables.
4. Rephrase to Introduce Diversity
Use varied question formats: causal, hypothetical, comparative, inferential, conditional, etc.
Maintain clarity and scientific rigor while diversifying expression.

Important:
- Each refined question should challenge deep understanding of core technical concepts, appropriate for advanced graduate-level assessment.
- The refined question must remain solvable using information provided or generally assumed background knowledge in the domain. Avoid ambiguity or underspecified problems.
- Ensure that only one option is unambiguously correct.
- If the original question is poorly written, ambiguous, or lacks depth, you may create a new question based on the same underlying concept or topic reflected in the provided QA.

Please only output the refined QA in the following JSON format: [[JSON example]]