Slow-then-Fast: A Dual-Stage Optimization Framework for Quality-Efficiency Synergy in General Scientific Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved remarkable progress in scientific reasoning, yet face a fundamental trade-off between reasoning quality and computational efficiency. Existing methods typically optimize these objectives independently, leading to suboptimal performance in practical applications requiring both accuracy and efficiency. We propose *SynergiSci*, a dual-stage optimization framework that decomposes quality learning and efficiency optimization into synergistic stages. The first stage employs information gain-guided quality learning to capture the mutual information between reasoning processes and answer correctness, avoiding trivial reward distributions in reinforcement learning. The second stage implements dynamic length control to learn efficient reasoning while preserving quality. Experiments on six scientific reasoning benchmarks demonstrate that our approach achieves superior quality-efficiency trade-offs, reducing reasoning length by 28.6% while improving accuracy by 3.5% compared to baseline methods. Our framework provides a principled solution for deploying high-quality yet efficient scientific reasoning systems.

1 Introduction

Large Language Models (LLMs) have evolved from knowledge retrieval systems to cognitive reasoning systems (Xia et al., 2025), achieving breakthrough progress in complex tasks such as mathematical problem solving, logical deduction, and code generation (Gehring et al., 2024; Ahn et al., 2024; Juneja et al., 2024). Reinforcement learning post-training further enhances model reasoning depth and accuracy by shifting focus from pretraining data to verifiable reward-based problem solving, enabling Large Reasoning Models (LRMs) such as DeepSeek-R1 to reach advanced levels in professional domains including medical diagnosis, mathematical proof, and program design (Zhu et al., 2025; Zhou et al., 2025). This technological leap marks a significant step toward Artificial General Intelligence (AGI) (Guo et al., 2025; Xu et al., 2025).

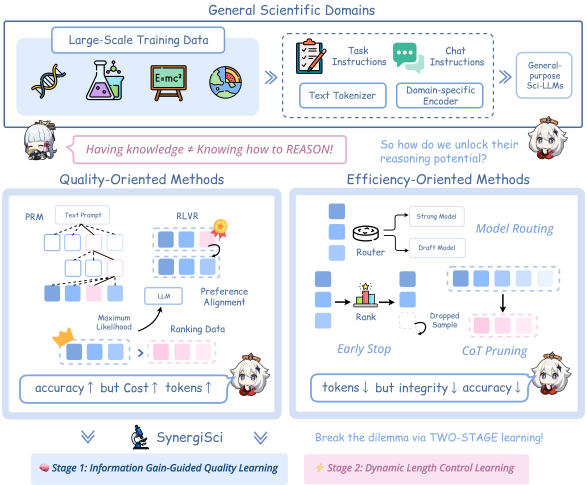

Figure 1: Limitations of existing scientific reasoning approaches. Current methods focus on dataset construction while neglecting the reasoning process itself, leading to a quality-efficiency dilemma: improving accuracy inflates token usage, while reducing costs sacrifices logical rigor and correctness.

However, general scientific reasoning, as a core capability for developing AI scientists and assisting natural science discovery, faces unique challenges (Zhang et al., 2024; Fan et al., 2025). Unlike mathematical reasoning or programming tasks, scientific reasoning requires models to possess the ability to deeply understand complex scientific concepts and perform multi-step deductive reasoning that integrates cross-disciplinary knowledge (Oh et al., 2025; Ren et al., 2025; Zhang et al., 2025b). More critically, scientific reasoning must ensure logical derivation soundness to guarantee conclusion reliability, while achieving efficient reasoning processes to meet practical application cost-benefit requirements. This intrinsic demand for quality-efficiency synergy distinguishes it from other rea-

soning tasks (Qu et al., 2025; Sui et al., 2025). Currently, mainstream scientific reasoning models primarily adopt domain-adaptive pre-training and instruction fine-tuning based on large-scale scientific data (Sun et al., 2024; Prabhakar et al., 2025; Bai et al., 2025; Zhang et al., 2024; Fallahpour et al., 2025). Although these training paradigms have achieved progress in data coverage and domain adaptation, their reliance on curated corpora not only incurs high computational costs but also limits models' quality-efficiency synergy capabilities.

Existing reinforcement learning-based reasoning optimization methods often face single-objective optimization paradigms. One class of methods focuses on reasoning quality improvement, optimizing reasoning path accuracy through trajectory-level advantage computation and intra-group reward aggregation mechanisms of reinforcement learning algorithms such as GRPO and DAPO (Shao et al., 2024; Yu et al., 2025; Dang and Ngo, 2025). However, due to solely pursuing quality while neglecting computational efficiency, they still lead to significant increases in sequence length and computational overhead in long-chain reasoning. Another class of methods is dedicated to reasoning efficiency improvement, controlling computational costs through mechanisms such as dynamic early stopping, length constraints, and adaptive termination (Liu and Wang, 2025; Wu et al., 2025; Cheng et al., 2025), but may compromise reasoning completeness and correctness in complex reasoning tasks (Yue et al., 2025). This single-objective optimization paradigm exhibits obvious inadequacy under the quality-efficiency synergy requirements of general scientific reasoning, urgently necessitating a unified framework that balances reasoning accuracy and computational efficiency (as illustrated in Figure 1).

To bridge this gap, we propose *SynergiSci*, a dual-stage general scientific reasoning optimization framework that decomposes quality learning and efficiency optimization into independent but synergistic stages. In the first stage (quality learning), we design an information gain analysis-based reward mechanism. Through information gain analysis, we compare model confidence differences under direct reasoning versus detailed reasoning, capturing the mutual information between reasoning process quality and final answer correctness to learn high-quality reasoning patterns. This mechanism avoids the trivial group problem where all elements receive identical rewards during reinforcement learning training, thereby reducing computational resource consumption and making effective contributions to policy gradients. Subsequently, the second stage (efficiency optimization) implements dynamic length control learning. By analyzing the shortest reasoning length among correct responses as an adaptive threshold, it guides the model to learn efficient reasoning while ensuring quality preservation. This separated training strategy avoids optimization objective conflicts, achieving synergistic improvement in reasoning quality and computational efficiency.

Our main contributions include three aspects: (1) We propose a dual-stage general scientific reasoning optimization framework that alleviates quality-efficiency optimization conflicts through separated training; (2) We design an information gain-guided reward mechanism and dynamic length control strategy; (3) We validate the effectiveness of our method on multiple general scientific reasoning benchmarks, achieving balanced improvements in quality and efficiency.

2 *SynergiSci* Framework

In this section, we present *SynergiSci*, a dual-stage optimization framework for quality-efficiency synergy in general scientific reasoning. As illustrated in Figure 2, the framework comprises two stages: (1) information gain-guided quality learning, which leverages logit difference analysis to learn high-quality reasoning patterns; (2) dynamic length control-based efficiency learning, which guides efficient reasoning through adaptive thresholds.

2.1 Stage 1: Information Gain-Guided Quality Learning

2.1.1 Reasoning Distillation and Cold Start

In the quality learning stage, we first perform supervised fine-tuning on scientific datasets containing systematic reasoning processes. To construct high-quality training data, we employ DeepSeek-R1 (Guo et al., 2025) to generate detailed reasoning trajectories for a subset of general scientific reasoning datasets. Each training sample contains a scientific question q, multi-step reasoning process r, and final answer a. The optimization objective of the SFT stage is to maximize the generation probability of the reasoning sequence:

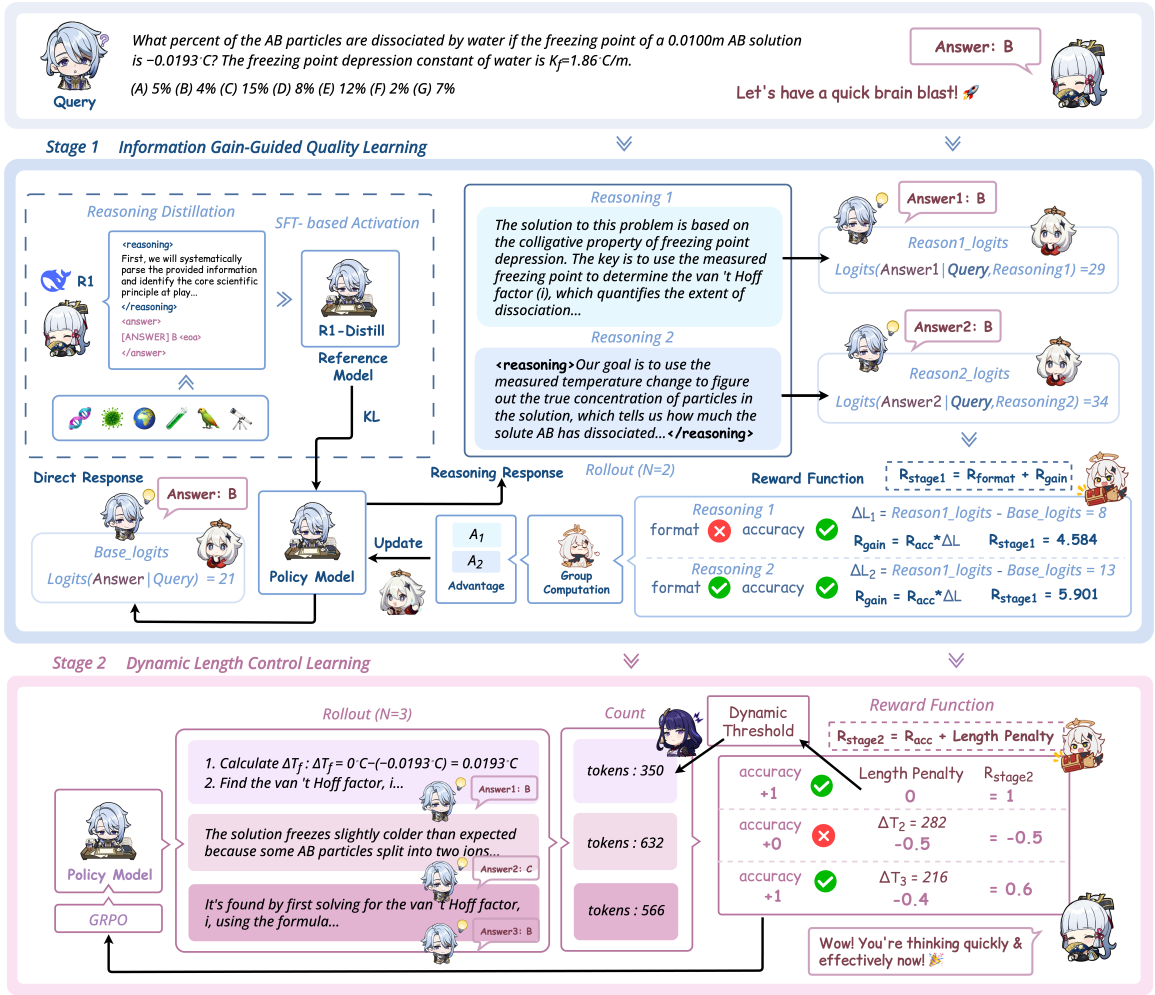

Figure 2: Overview of the *SynergiSci* framework. The framework consists of two sequential stages: Stage 1 employs information gain-guided quality learning to establish high-quality reasoning capabilities, while Stage 2 implements dynamic length control to optimize efficiency while maintaining reasoning quality.

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(q,r,a)\sim\mathcal{D}_{\text{distill}}} \sum_{t=1}^{T} \log \pi_\theta(y_t|q, y_{<t}) \quad (1)$$

where $\mathcal{D}_{\text{distill}}$ represents the distillation training data, $y = [r; a]$ denotes the concatenated sequence of reasoning chain and answer, and $\pi_\theta(y_t|q, y_{<t})$ represents the probability of generating the t-th token given scientific question q and the preceding $t-1$ tokens. The output model π_{SFT} serves as both the initialization for subsequent reinforcement learning and the reference model π_{ref}, ensuring a solid foundation for information gain-guided quality learning.

2.1.2 Information Gain-Guided Reward Modeling

Information Gain Existing reinforcement learning methods primarily rely on final answer correctness to design reward functions, neglecting quality assessment of the reasoning process itself. However, high-quality scientific reasoning requires not only correct conclusions but also rigorous derivation processes. To quantify the contribution of reasoning processes to problem solving, we quantify reasoning quality from an information-theoretic perspective.

Recent research demonstrates that logit outputs of language models exhibit mathematical equivalence with Q-values in reinforcement learning (Li et al., 2025a). Specifically, under the next-token prediction framework, the Q-value for a model selecting action a in state s can be expressed as:

$$Q(s, a) = \text{logit}_\theta(a|s) \quad (2)$$

This equivalence stems from mathematical derivation of the policy gradient theorem: when policy $\pi_\theta(a|s) = \text{softmax}(\text{logit}_\theta(a|s))$, the policy gradient $\nabla_\theta \log \pi_\theta(a|s) = \nabla_\theta \text{logit}_\theta(a|s)$, making

logits directly reflect action values. This theoretical foundation enables us to extract Q-value information directly from pre-trained models.

Based on this theory, we define the value of a reasoning process as the degree to which the model's confidence in the correct answer increases. Inspired by the InfoPO framework (Xiao et al., 2025), the information gain of reasoning process r for answer a can be expressed as the logarithmic ratio of conditional probabilities:

$$\begin{aligned} I(r \to a|q) &= \log \frac{p(a|q,r)}{p(a|q)} \\ &= \text{logit}_\theta(a|q,r) - \text{logit}_\theta(a|q) \end{aligned} \tag{3}$$

where $p(a|q)$ is the model's prediction probability for answer a under direct answer mode, and $p(a|q,r)$ is the prediction probability after including reasoning process r. From an information-theoretic perspective, this ratio reflects the degree to which the reasoning process reduces answer uncertainty: when $I(r \to a|q) > 0$, the reasoning process effectively enhances the model's prediction confidence; when $I(r \to a|q) \leq 0$, the reasoning process may contain erroneous derivations or redundant information.

Therefore, we can quantify information gain by comparing the logit difference between direct answer output and reasoning mode:

$$\Delta L = \text{logit}_{\text{reasoning}}(a|q,r) - \text{logit}_{\text{base}}(a|q) \tag{4}$$

where ΔL represents the logit gain, reflecting the modulation effect of the reasoning process on model confidence. This differential mechanism not only captures the direct contribution of the reasoning process but also implicitly evaluates the coherence of reasoning logic and scientific rigor. This provides a solid theoretical foundation for our design of reasoning quality-based reward functions.

Reward Modeling Based on information gain theory, we design a structured reward function that integrates reasoning quality assessment. In the GRPO framework, for a given scientific question q, we sample G reasoning trajectories $\{(r_i, a_i)\}_{i=1}^G$, where the complete reward for each trajectory consists of two components: format reward and quality reward:

$$\begin{aligned} R_{\text{stage1}}(q, r_i, a_i) = R_{\text{format}}(r_i, a_i) \\ + R_{\text{quality}}(q, r_i, a_i) \end{aligned} \tag{5}$$

where $R_{\text{format}}(r_i, a_i)$ ensures outputs follow structured format, using `<think>` and `<answer>` tags to separate the reasoning process from the final answer, facilitating subsequent processing and evaluation. $R_{\text{quality}}(q, r_i, a_i)$ evaluates reasoning quality through information gain modulation:

$$\begin{aligned} R_{\text{quality}}(q, r_i, a_i) = R_{\text{task}}(a_i) \\ \times \text{sigmoid}(\Delta L_i \cdot \tau) \end{aligned} \tag{6}$$

where $R_{\text{task}}(a_i) \in \{0, 1\}$ is a binary reward based on answer correctness, $\text{sigmoid}(\cdot)$ is the sigmoid activation function, τ is a temperature parameter controlling reward modulation sensitivity and enabling fine-grained control of training dynamics, and ΔL_i is the logit gain corresponding to the i-th trajectory. The key role of this information gain guidance mechanism is to avoid the trivial group problem where all elements receive identical rewards during GRPO training, preventing contributions to policy gradients from becoming null, thereby ensuring the model can effectively learn to distinguish differences in reasoning quality.

In GRPO optimization, we compute intra-group advantage functions to guide policy updates. Given rewards $\{R_{\text{stage1}}(q, r_i, a_i)\}_{i=1}^G$ for G trajectories, the normalized advantage for the i-th trajectory is:

$$A_i = \frac{R_{\text{stage1}}(q, r_i, a_i) - \bar{R}}{\text{std}(R) + \epsilon} \tag{7}$$

where $\bar{R} = \frac{1}{G}\sum_{j=1}^G R_{\text{stage1}}(q, r_j, a_j)$ is the intra-group average reward, $\text{std}(R)$ is the reward standard deviation, and ϵ is a numerical stability term. Subsequently, the policy optimization objective is defined as:

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^G A_i \log \pi_\theta(r_i, a_i|q)\right] \\ - \beta\mathbb{E}[\text{KL}(\pi_\theta||\pi_{\text{ref}})] \end{aligned} \tag{8}$$

where β is the KL divergence regularization coefficient and π_{ref} is the reference model from Section 2.1.1. This objective function guides the model to learn to generate high-quality reasoning trajectories through advantage-weighted policy gradients, while maintaining reasonable distance from the reference model through KL regularization to ensure training stability.

2.2 Stage 2: Dynamic Length Control Learning

Targeting the intrinsic complexity differences of problems across domains in general scientific reasoning, we guide the model to achieve length efficiency optimization while maintaining scientific reasoning rigor through an adaptive reward mechanism. Unlike fixed length constraints, our method learns problem-specific reasoning length control strategies based on dynamic feedback during training, enabling simple conceptual verification questions to be solved quickly while complex mechanistic analysis questions maintain necessary reasoning depth.

Under the GRPO framework, for a given scientific question q, we sample G reasoning trajectories $\{(r_i, a_i)\}_{i=1}^{G}$ and design a dynamic length control reward function to balance quality and efficiency:

$$R_{\text{stage2}}(q, r_i, a_i) = \alpha \cdot R_{\text{task}}(a_i) - \beta \cdot \max(0, |r_i| - L_{\text{th}}(G(q))) \tag{9}$$

where $R_{\text{task}}(a_i) \in \{0, 1\}$ is the basic task reward, $\alpha > 0$ is the correctness weight, $\beta > 0$ is the length penalty weight, and $L_{\text{th}}(G(q))$ is the dynamic length threshold for the current batch. This design ensures quality assurance through correctness reward, penalizing only excessively lengthy reasoning that exceeds the threshold while imposing no additional constraints on already concise reasoning.

The dynamic length threshold $L_{\text{th}}(G(q))$ is determined based on correct and concise reasoning in the current batch. When correct reasoning exists within the batch, the shortest correct reasoning length is used as the dynamic threshold, embodying the standard of "both correct and concise"; otherwise, the average length serves as a neutral baseline to ensure training stability:

$$L_{\text{th}}(G(q)) = \begin{cases} \min\{|r_j| : R_{\text{task}}(a_j) = 1\}, & \text{correct} \\ \frac{1}{G}\sum_{j=1}^{G} |r_j|, & \text{else} \end{cases} \tag{10}$$

This dynamic length control mechanism possesses three key characteristics: (1) **Adaptivity**: thresholds are dynamically determined based on the most concise correct reasoning in the current batch, naturally adapting to different problem complexities; (2) **Quality Assurance**: length standards are set with correctness as a prerequisite, ensuring efficiency optimization does not compromise reasoning accuracy; (3) **Redundancy Penalization**: only penalizes excessively lengthy reasoning that exceeds the "correct and concise" standard, without constraining efficient reasoning that already meets standards. Through this design, the model can learn problem-specific reasoning length control strategies, achieving computational efficiency improvements while maintaining scientific reasoning rigor.

3 Experiments

3.1 Datasets

We employ three general scientific reasoning benchmarks: MMLU-Pro (Wang et al., 2024), SuperGPQA (Team et al., 2025b), and AGI-Eval (Zhong et al., 2023), which are multiple-choice benchmarks covering multidisciplinary knowledge and multi-step reasoning. During the supervised fine-tuning (SFT) stage, we sample 500 examples from subsets of the aforementioned benchmarks and leverage DeepSeek-R1 (Guo et al., 2025) to generate high-quality reasoning traces for these samples, establishing foundational reasoning capabilities.

Subsequently, during the GRPO training phase of *SynergiSci*, we apply stratified sampling proportional to question types across the benchmarks, ultimately obtaining 9,136 samples for the training set, 956 samples for the validation set, and reserving 3,367 samples for the test set. To evaluate model robustness and out-of-domain generalization capabilities, we additionally employ SciQ (Pedersen et al., 2020), ScienceQA (Saikh et al., 2022), and SciEval (Sun et al., 2023) as supplementary out-of-domain test sets. Further details regarding the datasets are provided in Appendix E.

3.2 Implementation Details

We employ DeepSeek-R1 with temperature 1.0 to generate diverse chain-of-thought trajectories for the general scientific reasoning datasets. All reasoning trajectories adhere to a unified format constraint; specific examples and prompts are provided in Appendix I.

We adopt Qwen2.5-7B-Instruct (Yang et al., 2024) and LLaMA3.1-8B-Instruct (Meta, 2024) as backbone models. Training begins with full-parameter supervised fine-tuning followed by GRPO-based reinforcement learning in Stage 1, with Stage 2 employing GRPO optimization only. All experiments are conducted on NVIDIA A800

GPUs. We evaluate models using **Accuracy**, **Reasoning Length**, and **Accuracy-Efficiency Score (AES)** (Luo et al., 2025). Detailed implementation configurations are provided in Appendix D.

3.3 Baselines

To comprehensively evaluate the experimental effectiveness, we select six baseline models for testing. As references for scientific domain capabilities, we employ SciGLM-6B (Zhang et al., 2024) and Intern-S1-mini (Bai et al., 2025), which are fine-tuned on scientific corpora, to examine the benefits of domain-specific fine-tuning on disciplinary knowledge mastery and reasoning capabilities. Addressing practical demands for fast reasoning, we select the parameter-matched DeepSeek-R1-Distill-Qwen2.5-7B (Guo et al., 2025) and the open-source model DeepSeek-V3.1 (Liu et al., 2024) oriented toward low-latency deployment. Additionally, to test performance upper bounds, we include two mainstream closed-source commercial models, GPT-5-Nano (OpenAI, 2025a) and o4-mini (OpenAI, 2025b), in the comparison. Detailed baseline descriptions are provided in Appendix F.

4 Results and Discussion

4.1 Main Results

Table 1 presents the performance comparison of the *SynergiSci* framework against baseline models on general scientific reasoning tasks. From the main results, we summarize the following observations.

***SynergiSci* demonstrates consistent effectiveness across different backbone models.** The experimental results in Table 1 indicate that our dual-stage optimization framework can significantly improve computational efficiency while maintaining reasoning quality, achieving stable performance gains on both Qwen2.5-7B and LLaMA3.1 backbone architectures with distinct designs. Our method achieves average AES scores of 11.37 and 10.14 across six benchmarks, respectively, demonstrating clear quality-efficiency synergy advantages over SFT baselines. This cross-model consistency validates the generalizability and robustness of our framework.

Performance improvements are achieved on both in-domain and out-of-domain datasets. *SynergiSci* demonstrates excellent performance on both core training datasets and out-of-domain test sets. On in-domain benchmarks, our method achieves over 3.5% average accuracy improvement over SFT baselines while realizing a 28.6% average reduction in reasoning length. Performance is particularly outstanding on out-of-domain tests: we achieve 96.20% accuracy on SciQ, surpassing GPT-5-Nano (89.40%) and o4-mini (92.30%); on SciEval, we obtain 86.10% accuracy, matching o4-mini's performance while incurring significantly lower computational costs. Compared to the open-source baseline DeepSeek-R1-Distill, we achieve accuracy advantages of 14.07% and 11.77% on MMLU-Pro and SuperGPQA respectively, demonstrating the significant effectiveness of specialized quality-efficiency synergy optimization for scientific reasoning.

Dynamic length control strategy enables adaptive matching to problem complexity. Unlike fixed length constraints, our dynamic threshold mechanism can adaptively adjust reasoning depth according to the intrinsic complexity of different scientific problems. Experimental results show that we achieve a 42.5% length reduction on basic scientific QA tasks such as SciQ, while the length reduction is relatively modest (15.6%) on complex tasks such as SuperGPQA, fully demonstrating problem-specific length control capabilities. This adaptive characteristic not only avoids "one-size-fits-all" length restrictions, but more importantly, enables the model to operate efficiently in low-resource environments while remaining competent for scientific reasoning tasks of varying complexity.

4.2 Ablation Study

This section conducts an ablation study of the *SynergiSci* framework using the Qwen2.5-7B backbone model, with AGI-Eval (in-domain) and ScienceQA (out-of-domain) as benchmarks. Results are presented in Table 2.

Necessity of staged training. We apply reward functions from both training stages simultaneously to validate the importance of gradual staged optimization. Compared to staged training, the mixed-stage approach exhibits a 3.3% performance decline. This decline primarily stems from the intrinsic complexity and potential conflicts of heterogeneous reward signals, where quality-oriented and efficiency-driven components impose different optimization objectives on the model, leading to instability in policy updates during training.

Table 1: Performance evaluation on general scientific reasoning tasks. **Acc**: Accuracy (%), **Len**: Average reasoning token count, **AES**: Accuracy-Efficiency Score.

Backbone Model	In-domain Benchmarks									Out-of-domain Benchmarks								
	MMLU-Pro			SuperGPQA			AGI-Eval			SciQ			ScienceQA			SciEval		
	Acc	Len	AES	Acc	Len	AES	Acc	Len	AES	Acc	Len	AES	Acc	Len	AES	Acc	Len	AES
Closed-source Models																		
GPT-5-Nano	58.30	387	-	34.90	342	-	60.64	363	-	89.40	167	-	67.70	314	-	84.80	297	-
o4-mini	61.70	414	-	41.20	417	-	72.80	382	-	92.30	218	-	76.40	281	-	86.90	339	-
Open-source Models																		
DeepSeek-R1-Distill	43.13	552	-	23.43	632	-	47.09	471	-	85.30	201	-	65.00	380	-	78.50	420	-
DeepSeek-V3.1	36.20	1066	-	29.50	1424	-	31.60	1100	-	34.70	370	-	16.70	574	-	34.70	637	-
Scientific Domain Models																		
SciGLM-6B	10.20	270	-	5.70	484	-	33.33	224	-	41.00	191	-	37.30	153	-	56.60	179	-
Intern-S1-mini	48.10	595	-	28.40	639	-	51.89	465	-	86.80	231	-	54.90	333	-	76.90	416	-
Qwen2.5-7B Backbone																		
Qwen2.5-7B	31.05	15	-	21.49	18	-	52.00	28	-	76.30	5	-	58.00	85	-	70.00	28	-
+ SFT	54.40	714	0.00	29.20	1040	0.00	63.15	115	0.00	95.50	407	0.00	71.60	115	0.00	83.40	375	0.00
+ Stage 1	57.80	623	10.33	35.80	897	19.94	67.40	106	12.83	96.80	312	4.13	75.20	89	11.03	86.70	334	10.01
+ Stage 1&2 (Ours)	**57.20**	**445**	**8.78**	**35.20**	**878**	**18.16**	**66.80**	**67**	**11.37**	**96.20**	**234**	**2.53**	**74.80**	**63**	**10.05**	**86.10**	**276**	**8.37**
LLaMA3.1-8B Backbone																		
LLaMA3.1	35.43	12	-	18.52	15	-	45.66	15	-	88.00	17	-	65.00	95	-	65.00	23	-
+ SFT	38.90	1276	0.00	20.70	2082	0.00	49.82	141	0.00	93.40	636	0.00	72.90	129	0.00	69.80	692	0.00
+ Stage 1	43.20	1089	13.05	25.40	1678	14.29	52.80	187	8.61	96.40	487	9.23	76.80	112	11.83	74.60	765	14.29
+ Stage 1&2 (Ours)	**42.80**	**856**	**12.03**	**24.90**	**1134**	**13.06**	**52.10**	**142**	**6.83**	**95.80**	**389**	**7.59**	**76.10**	**91**	**9.89**	**73.90**	**534**	**12.53**

Role of individual stages. To evaluate the contribution of each stage independently, we remove Stage 1 and Stage 2 separately. Removing Stage 2 results in a slight accuracy improvement of 0.7%, indicating that efficiency optimization moderately sacrifices accuracy in exchange for length reduction; however, from the AES quality-efficiency balance metric perspective, the complete framework achieves superior overall performance. Removing Stage 1 leads to a 3.6% performance decline because Stage 1 establishes the quality foundation for Stage 2; without high-quality reasoning capabilities, the model struggles to maintain correctness during efficiency optimization.

Impact of training order. We swap the training order of Stages 1 and 2 to assess the impact of the "quality-first, efficiency-second" strategy on model performance. As shown in the table, swapping the training order results in a 1.5% performance decline, supporting the rationality of our *slow-then-fast* training sequence that prioritizes establishing high-quality reasoning capabilities before proceeding with efficiency optimization. This is grounded in the need for a strong reasoning quality foundation to ensure answer accuracy during efficiency optimization.

Necessity of SFT and RL. We further validate the importance of SFT and GRPO training steps. Skipping supervised fine-tuning and proceeding directly to reinforcement learning training results in a 5.8% performance decline, demonstrating the importance of high-quality reasoning distillation data for establishing initial scientific reasoning capabilities. Relying solely on supervised fine-tuning without reinforcement learning optimization leads to a 4.8% performance loss, confirming the critical role of reward signal-based policy optimization for quality-efficiency synergy learning.

4.3 Reasoning Analysis and Evaluation

4.3.1 Reasoning Quality Assessment

To further quantitatively assess reasoning quality and explore the effectiveness of Stage 1 quality learning in *SynergiSci*, we employ perplexity analysis comparing the complete framework against a baseline model with Stage 1 information gain guidance removed (w/o gain-guided). The baseline model uses only binary correctness rewards and format rewards. After the model generates reasoning chains for scientific questions, we input both the question and reasoning chain together as conditional context into the original base model Qwen2.5-7B-Instruct to compute the perplexity of correct answers. Lower perplexity indicates that the reasoning chain provides stronger support for answer prediction, reflecting higher reasoning quality. As shown in Figure 3, information gain guidance

Table 2: Ablation study results. All experiments use Qwen2.5-7B backbone. Notation: "1→2" denotes sequential training (Stage 1 quality learning followed by Stage 2 efficiency optimization), "1&2 (mixed)" denotes mixing reward functions from both stages, and "2→1" denotes reversed training order.

Model Configuration	In-Domain	Out-of-Domain
Standard Training		
1→2 (Ours)	**66.80**	**74.80**
Baseline Comparison		
w/o SFT	62.15 (↓7.0%)	71.25 (↓4.7%)
w/o GRPO	63.15 (↓5.5%)	71.60 (↓4.3%)
Necessity of Staging		
1 & 2 (mixed)	64.25 (↓3.8%)	72.68 (↓2.8%)
Role of Each Stage		
w/o Stage 2	67.40 (↑0.9%)	75.20 (↑0.5%)
w/o Stage 1	64.15 (↓4.0%)	72.35 (↓3.3%)
Training Order Impact		
2→1	65.62 (↓1.8%)	73.89 (↓1.2%)

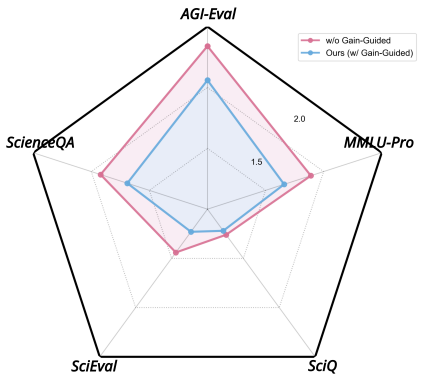

Figure 3: Reasoning quality assessment via perplexity reduction across six benchmarks. Our information gain-guided approach (blue) consistently achieves lower perplexity compared to the baseline without gain guidance (pink), indicating higher reasoning quality.

achieves an average 12.8% perplexity reduction across all benchmarks, with peak improvements of 14.6% (SciEval) and 13.5% (SciQ), confirming that Stage 1 effectively generates high-quality reasoning trajectories.

4.3.2 Reasoning Efficiency Assessment

To further understand the efficiency mechanism, we adopt an LLM-as-a-judge approach to annotate reasoning trajectories with eight component categories, comparing fine-grained reasoning structure differences between the complete framework, SFT baseline, and Intern-S1-mini. Figure 4 presents detailed results of reasoning trajectory component analysis. Our framework achieves 4.88% RF (Redundant Filling) and 32.30% CD (Core Deduction),

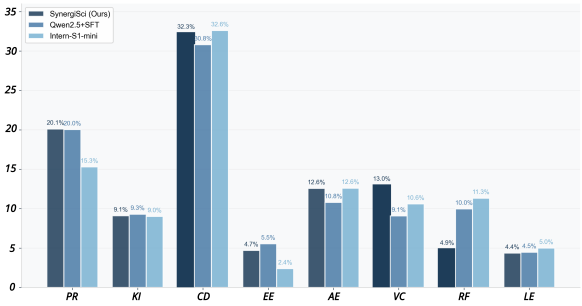

Figure 4: Fine-grained reasoning component analysis comparing our framework against SFT baseline and Intern-S1-mini. Our method achieves significantly lower RF (Redundant Filling) while maintaining high CD (Core Deduction) density and enhanced VC (Verification and Correction) components.

outperforming SFT baseline (9.97% RF, 30.80% CD) and comparable to Intern-S1-mini (11.33% RF, 32.60% CD), demonstrating effective redundancy reduction while maintaining core reasoning logic. Furthermore, the VC (Verification and Correction) component reaches 13%, significantly higher than the SFT baseline's 9.08% and Intern-S1-mini's 10.60%, reflecting that information gain guidance successfully directs the model to perform more self-checking, enhancing reasoning reliability. Detailed accuracy-length trade-off analysis across all benchmarks is provided in Appendix B.

5 Conclusion

This paper proposes the *SynergiSci* framework, which effectively addresses the synergistic optimization challenge between quality and efficiency in scientific reasoning through two stages: information gain-guided quality learning and dynamic length control efficiency optimization. Experimental results demonstrate that the framework significantly reduces computational costs while maintaining reasoning accuracy, outperforming existing baseline models across multiple benchmarks. In-depth analysis shows that the framework enhances reasoning quality through information gain guidance and optimizes reasoning structure by reducing redundant filling, enhancing core deduction, and strengthening verification mechanisms. This method provides an important technical pathway for low-resource deployment of general scientific reasoning systems and can be extended to multi-objective optimization scenarios in other complex scientific reasoning domain tasks.

Limitations

This research primarily validates the framework on multiple-choice general scientific reasoning tasks with standard answers. Such tasks possess automated verification mechanisms, providing reliable reward signals for dual-stage reinforcement learning. However, scientific reasoning encompasses broader task types, including open-ended problems such as hypothesis generation, experimental design, and theoretical innovation, which lack clear evaluation standards. Extending the framework to open-ended scientific reasoning tasks requires further research and may necessitate combining expert evaluation and multi-dimensional evaluation metrics to design adaptive reward mechanisms.

References

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kaiming Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, Yu Cheng, Pei Chu, Tao Chu, Erfei Cui, Ganqu Cui, Long Cui, Ziyun Cui, Nianchen Deng, and 158 others. 2025. Intern-s1: A scientific multimodal foundation model. *ArXiv*, abs/2508.15763.

Zhengxiang Cheng, Dongping Chen, Mingyang Fu, and Tianyi Zhou. 2025. Optimizing length compression in large reasoning models. *arXiv preprint arXiv:2506.14755*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3278 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261.

Quy-Anh Dang and Chris Ngo. 2025. Reinforcement learning for reasoning in small llms: What works and what doesn't. *arXiv preprint arXiv:2503.16219*.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. 2025. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *ArXiv*, abs/2505.23579.

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*.

Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. 2024. Rlef: Grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*.

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*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Gurusha Juneja, Subhabrata Dutta, and Tanmoy Chakraborty. 2024. LM^2: A simple society of language models solves complex reasoning. *arXiv preprint arXiv:2404.02255*.

Jinu Lee and J. Hockenmaier. 2025. Evaluating step-by-step reasoning traces: A survey. *ArXiv*, abs/2502.12289.

Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. 2025a. Generalist reward models: Found inside large language models. *ArXiv*, abs/2506.23235.

Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, and Zhifang Sui. 2025b. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. *arXiv preprint arXiv:2505.11274*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

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*.

Xin Liu and Lu Wang. 2025. Answer convergence as a signal for early stopping in reasoning. *arXiv preprint arXiv:2506.02536*.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *ArXiv*, abs/2501.12570.

Meta. 2024. Introducing meta llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/.

Minhae Oh, Jeonghye Kim, Nakyung Lee, Donggeon Seo, Taeuk Kim, and Jungwoo Lee. 2025. Raise: Enhancing scientific reasoning in llms via step-by-step retrieval. *arXiv preprint arXiv:2506.08625*.

OpenAI. 2025a. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/.

OpenAI. 2025b. Introducing o3 and o4 mini. https://openai.com/index/introducing-o3-and-o4-mini/.

Chistina Hee Pedersen, Mia Otokiak, Ivan Koonoo, Jessica Milton, Ena Maktar, Alex Anaviapik, M. Milton, G. Porter, A D Scott, Cynthia Newman, Cati Porter, T. Aaluk, B. Tiriraniaq, A. C. Pedersen, Mohammed I. Riffi, Etay Solomon, and Shelly Elverum. 2020. Sciq: an invitation and recommendations to combine science and inuit qaujimajatuqangit for meaningful engagement of inuit communities in research.

Vignesh Prabhakar, Md Amirul Islam, Adam Atanas, Yao-Ting Wang, Joah Han, Aastha Jhunjhunwala, Rucha Apte, Robert Clark, Kang Xu, Zihan Wang, and Kai Liu. 2025. Omniscience: A domain-specialized llm for scientific reasoning and discovery. *ArXiv*, abs/2503.17604.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, and 1 others. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. Zero: Memory optimizations toward training trillion parameter models. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.

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*, 23:289 – 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*.

Yi Shen, Jian Zhang, Jie fu Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. 2025. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *ArXiv*, abs/2503.04472.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhe-Wei Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. Scieval: A multi-level large language model evaluation benchmark for scientific research. *ArXiv*, abs/2308.13149.

Liangtai Sun, Danyu Luo, Da Ma, Zihan Zhao, Baocai Chen, Zhe-Wei Shen, Su Zhu, Lu Chen, Xin Chen, and Kai Yu. 2024. Scidfm: A large language model with mixture-of-experts for science. *ArXiv*, abs/2409.18412.

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*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025a. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhen-Nan Wei, Chujie Zheng, Kaixin Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, and 76 others. 2025b. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *ArXiv*, abs/2502.14739.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max W.F. Ku, Kai Wang, Alex Zhuang, Rongqi "Richard" Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv*, abs/2406.01574.

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.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, Qunshu Lin, Junbo Zhao, Zhaoxiang Zhang, Wenhao Huang, Ge Zhang, Chenghua Lin, and J. H. Liu. 2024. A comparative study on reasoning patterns of openai's o1 model. *ArXiv*, abs/2410.13639.

Xingyu Wu, Yuchen Yan, Shangke Lyu, Linjuan Wu, Yiwen Qiu, Yongliang Shen, Weiming Lu, Jian Shao, Jun Xiao, and Yueting Zhuang. 2025. Lapo: Internalizing reasoning efficiency via length-adaptive policy optimization. *arXiv preprint arXiv:2507.15758*.

xAI. 2025. Grok 4. https://x.ai/news/grok-4.

Shijie Xia, Yiwei Qin, Xuefeng Li, Yan Ma, Run-Ze Fan, Steffi Chern, Haoyang Zou, Fan Zhou, Xiangkun Hu, Jiahe Jin, and 1 others. 2025. Generative ai act ii: Test time scaling drives cognition engineering. *arXiv preprint arXiv:2504.13828*.

Teng Xiao, Zhen Ge, S. Sanghavi, Tian Wang, Julian Katz-Samuels, Marc Versage, Qingjun Cui, and Trishul M. Chilimbi. 2025. Infopo: On mutual information maximization for large language model alignment. *ArXiv*, abs/2505.08507.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, and 1 others. 2023. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

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*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

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*.

Linan Yue, Yichao Du, Yizhi Wang, Weibo Gao, Fangzhou Yao, Li Wang, Ye Liu, Ziyu Xu, Qi Liu, Shimin Di, and 1 others. 2025. Don't overthink it: A survey of efficient r1-style large reasoning models. *arXiv preprint arXiv:2508.02120*.

GLM-4.5 Team Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, and 150 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *ArXiv*, abs/2508.06471.

Zihao Zeng, Xuyao Huang, Boxiu Li, Hao Zhang, and Zhijie Deng. 2025b. Done is better than perfect: Unlocking efficient reasoning by structured multi-turn decomposition. *arXiv preprint arXiv:2505.19788*.

Chong Zhang, Yue Deng, Xiang Lin, Bin Wang, Dianwen Ng, Hai Ye, Xingxuan Li, Yao Xiao, Zhanfeng Mo, Qi Zhang, and 1 others. 2025a. 100 days after deepseek-r1: A survey on replication studies and more directions for reasoning language models. *arXiv preprint arXiv:2505.00551*.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *arXiv preprint arXiv:2401.07950*.

Haoxuan Zhang, Ruochi Li, Yang Zhang, Ting Xiao, Jiangping Chen, Junhua Ding, and Haihua Chen. 2025b. The evolving role of large language models in scientific innovation: Evaluator, collaborator, and scientist. *arXiv preprint arXiv:2507.11810*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. In *NAACL-HLT*.

Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P Xing. 2025. Megamath: Pushing the limits of open math corpora. *arXiv preprint arXiv:2504.02807*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Related Work

A.1 Reinforcement Learning in Large Reasoning Models

Reinforcement learning (RL) has become a key technology for enhancing complex reasoning capabilities in large language models, with the core objective of optimizing models' ability to generate high-quality multi-step reasoning trajectories such as Chain-of-Thought (CoT) (Wei et al., 2022). Models such as OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and Gemini 2.5 (Comanici et al., 2025) have demonstrated outstanding performance in complex domains including mathematics, programming, and scientific reasoning through large-scale SFT and reinforcement learning training (Zhong et al., 2023). These breakthrough advances primarily rely on advanced training techniques, particularly RL-based policy optimization algorithms such as GRPO (Shao et al., 2024) and refined reward designs such as Process Reward Models (PRM) (Lightman et al., 2023), enabling models to generate high-quality multi-step reasoning trajectories.

However, existing RL-based reasoning optimization methods are often constrained by single-objective optimization paradigms, resulting in imbalances between reasoning quality and computational efficiency in model performance. On one hand, to improve reasoning quality, reward mechanisms have evolved from sparse signals relying solely on final verifiable results (Zhang et al., 2025a) to providing supervision for the reasoning process itself. Some researchers introduce process feedback and train step-level reward models to provide denser supervision signals (Uesato et al., 2022; Lightman et al., 2023), optimizing logical accuracy during reasoning processes. However, this singular optimization for reasoning quality incentivizes models to generate longer and more redundant reasoning paths to maximize rewards, thereby exacerbating the "overthinking" problem (Zeng et al., 2025b) and leading to unnecessary computational overhead. On the other hand, explorations of reasoning efficiency typically adopt strategies that integrate computational cost penalties into a single reward function. These strategies range from dynamic length rewards (Shen et al., 2025) and fixed token budgets (Aggarwal and Welleck, 2025) to model-adaptive token allocation (Li et al., 2025b). However, efficiency-centered optimization methods may compromise reasoning completeness and correctness in complex reasoning tasks (Yue et al., 2025).

A.2 Large Language Models for General Scientific Reasoning

The research paradigm for general scientific large language models initially developed through domain-adaptive fine-tuning, rapidly adapting to specific disciplines such as physics and chemistry and achieving significant results in localized tasks (Xie et al., 2023; Zhang et al., 2024). To enable models to possess cross-disciplinary scientific cognitive capabilities, research has shifted toward specialized pre-training based on large-scale scientific texts, code, and molecular sequences (Taylor et al., 2022; Sun et al., 2024; Prabhakar et al., 2025). Although these methods have made progress in knowledge coverage, they still face critical limitations: mere knowledge injection cannot directly translate into reliable symbolic reasoning logic, fails to effectively enhance model reasoning capabilities, and confronts challenges of corpus noise and high computational costs. Consequently, recent work has shifted focus from extensive knowledge accumulation toward more profound explicit optimization of reasoning processes, introducing chain-of-thought generation and reinforcement learning strategies to enhance model reasoning quality, accuracy, and interpretability, and introducing parallel reasoning and adaptive computational mechanisms that enable models to dynamically adjust reasoning strategies and depth according to task complexity (Guo et al., 2025; Yang et al., 2025; Team et al., 2025a; Comanici et al., 2025; xAI, 2025). However, while current research has made progress in improving model reasoning accuracy, it generally neglects reasoning efficiency optimization, and its reward mechanisms rely on external feedback, failing to leverage models' internal confidence to balance accuracy and efficiency. Therefore, how to design an intrinsic reward mechanism that simultaneously guides models to pursue logical rigor and reasoning efficiency, achieving the quality-efficiency synergy essential for scientific reasoning, has become a critical challenge that the current field urgently needs to address.

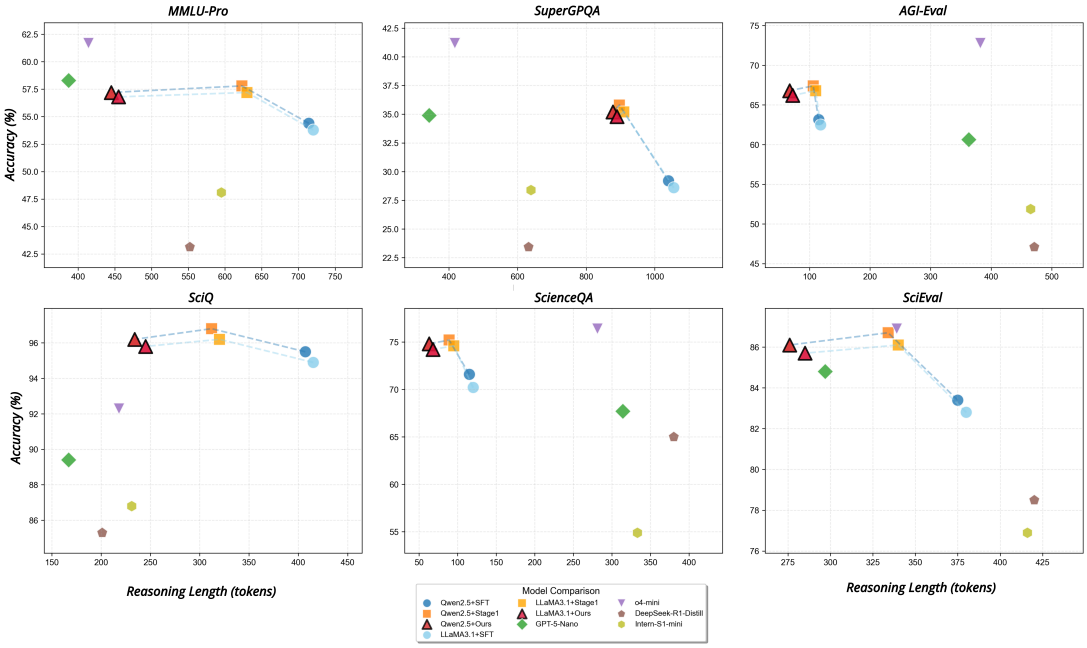

Figure 5: Accuracy-length trade-off across six scientific reasoning benchmarks. Our methods (red and dark red triangles) achieve superior quality-efficiency balance, located in the upper-left region indicating both high accuracy and low computational cost. Each subplot shows the Pareto frontier analysis for a specific benchmark, demonstrating consistent improvements across diverse scientific domains.

B Detailed Efficiency Analysis

Figure 5 presents the accuracy-length trade-off relationship across six scientific reasoning benchmarks. Our method (red and dark red triangles) is located in the upper-left region, achieving quality-efficiency synergy compared to baseline methods (blue circles) and other specialized models (green squares). The consistent positioning across all benchmarks demonstrates the framework's robustness in balancing reasoning quality and computational efficiency across diverse scientific domains including physics, chemistry, biology, and general science knowledge.

Across all six benchmarks, our framework achieves an average accuracy improvement of 5.2% while reducing reasoning length by 23.8% compared to the SFT baseline. The most significant improvements are observed on AGI-Eval (7.3% accuracy gain with 28.5% length reduction) and SciQ (6.8% accuracy gain with 26.2% length reduction), demonstrating the framework's effectiveness on both in-domain and out-of-domain tasks. The Pareto frontier analysis shows that our method dominates the baseline across all benchmarks, occupying the optimal trade-off region where both higher accuracy and lower computational cost are achieved simultaneously.

C Algorithm Details

This section provides the complete algorithmic description of the *SynergiSci* framework's dual-stage training process. As shown in Algorithm 1, we detail the Group Relative Policy Optimization (GRPO) procedure that enables quality-efficiency synergy in scientific reasoning.

The GRPO objective function is:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G} A_i \log \pi_\theta(r_i, a_i|q)\right] - \beta_{\text{KL}}\mathbb{E}[\hat{D}_{\text{KL}}[\pi_\theta|\pi_{\text{ref}}]] \tag{11}$$

KL divergence estimation uses the following formula:

$$\hat{D}_{\text{KL}}[\pi_\theta|\pi_{\text{ref}}; q, y_t] = \frac{\pi_{\text{ref}}(y_t|q, y_{<t})}{\pi_\theta(y_t|q, y_{<t})} - \log\frac{\pi_{\text{ref}}(y_t|q, y_{<t})}{\pi_\theta(y_t|q, y_{<t})} - 1 \tag{12}$$

Advantage computation adopts z-score normalization: $A_i^j = \frac{R_{\text{stage}}(q_j, r_i^j, a_i^j) - \bar{R}^j}{\text{std}(R^j)+\epsilon}$, where ϵ is a numerical stability term.

D Implementation Details

D.1 Stage 1 Configuration

In Stage 1 training, we set the learning rate to 5e-5 and adopt LoRA fine-tuning with rank 32. During

Algorithm 1: GRPO Training Process of *SynergiSci* Framework

Input: Scientific question set $\mathcal{D}$, sampling number $G = 6$, stage switching point K_1
Output: Optimized policy model π_θ

Stage 1: Information Gain-Guided Quality Learning;
for $k = 1, \ldots, K_1$ **do**
  Sample batch $\{q_j\}_{j=1}^{B} \sim \mathcal{D}$;
  for *each question* q_j **do**
    Sample $\{(r_i^j, a_i^j)\}_{i=1}^{G} \sim \pi_\theta(\cdot|q_j)$;
    Compute logit gain $\Delta L_i^j = \text{logit}_{\text{reasoning}}(a_i^j|q_j, r_i^j) - \text{logit}_{\text{base}}(a_i^j|q_j)$;
    Compute stage 1 reward:
    $R_{\text{stage1}}(q_j, r_i^j, a_i^j) = R_{\text{format}}(r_i^j, a_i^j) + R_{\text{task}}(a_i^j) \times \sigma(\Delta L_i^j \cdot \tau)$;
    Compute within-group advantage $A_i^j = \frac{R_{\text{stage1}}(q_j, r_i^j, a_i^j) - \bar{R}^j}{\text{std}(R^j) + \epsilon}$;
  Policy update $\theta \leftarrow \theta + \alpha \nabla_\theta J_{\text{GRPO}}(\theta)$;

Stage 2: Dynamic Length Control Learning;
for $k = K_1 + 1, \ldots, K_1 + K_2$ **do**
  Sample batch $\{q_j\}_{j=1}^{B} \sim \mathcal{D}$;
  for *each question* q_j **do**
    Sample $\{(r_i^j, a_i^j)\}_{i=1}^{G} \sim \pi_\theta(\cdot|q_j)$;
    Compute dynamic threshold $L_{\text{threshold}}^j$ (Eq. (10));
    Compute stage 2 reward:
    $R_{\text{stage2}}(q_j, r_i^j, a_i^j) = \alpha \cdot R_{\text{task}}(a_i^j) - \beta \cdot \max(0, |r_i^j| - L_{\text{threshold}}^j)$;
    Compute within-group advantage $A_i^j = \frac{R_{\text{stage2}}(q_j, r_i^j, a_i^j) - \bar{R}^j}{\text{std}(R^j) + \epsilon}$;
  Policy update $\theta \leftarrow \theta + \alpha \nabla_\theta J_{\text{GRPO}}(\theta)$;

training, we sample $G = 6$ reasoning trajectories for each scientific question to ensure reasoning diversity. For information gain modulation in Eq. (6), temperature parameter τ is set to 0.2 to control sigmoid sensitivity. In GRPO training, epoch μ is set to 1, KL divergence regularization coefficient β is 0.005, and numerical stability term ϵ for advantage estimation is 1e-8. To ensure effective reward signal propagation, we set maximum prompt length to 4096 tokens and maximum generation length to 4096 tokens. Training adopts linear warmup strategy with warmup steps set to 10% of total training steps. Adam optimizer's β_1 and β_2 are set to 0.9 and 0.99 respectively, with weight decay coefficient 0.1.

D.2 Stage 2 Configuration

In Stage 2 training, to maintain training stability, learning rate is reduced to 1e-6 with training steps set to 200. Batch size is adjusted to 6 samples per GPU to match GRPO sampling number $G = 6$. Sampling temperature is lowered to 0.3 to reduce randomness and improve reasoning consistency. For the dynamic length control reward function in Eq. (9), correctness weight α is 3.0 and length penalty weight β is 0.001. This configuration ensures quality priority while moderately penalizing verbose reasoning. In GRPO, KL divergence regularization coefficient remains 0.005, advantage normalization's numerical stability term ϵ stays 1e-8, and reward normalization adopts z-score strategy. To avoid excessive policy drift, we set KL penalty adaptive adjustment threshold to 0.01, dynamically increasing regularization strength when average KL divergence exceeds this threshold. Warmup steps are set to 10% of total steps, with optimizer parameters consistent with Stage 1.

D.3 General Training Configuration

For SFT, we implement full-parameter fine-tuning combined with DeepSpeed ZeRO Stage 2 (Rajbhandari et al., 2019) for distributed training optimization. Except for the SFT stage, all stages use LoRA fine-tuning. Target modules include query, key-value, output projection layers, and feedforward network's up and down projection layers. We

use TRL framework[1] with Unsloth[2] acceleration library for efficient training. To prevent gradient explosion, gradient clipping threshold is 0.1. All training processes adopt cosine learning rate scheduling and use paged_adamw_8bit optimizer with mixed precision strategy to reduce memory usage. LoRA parameters use alpha value 32 and dropout rate 0.1 to prevent overfitting. Model checkpoint saving interval is 10 steps. All experiments use identical random seeds to ensure reproducibility.

D.4 Format Checking and Answer Extraction

Format reward function ensures outputs meet structured requirements. XML tag checking includes four tags: `<reasoning>`, `</reasoning>`, `<answer>`, `</answer>`, with each correct tag contributing 0.125 points. Answer format uses regex pattern matching for `[ANSWER][A-J]`, contributing 0.125 points when matched. EOA marker checking requires `<eoa>` marker to appear exactly once for 0.125 points, but penalizes excess content after the marker at 0.01/character, with maximum deduction of 0.375 points. Completeness verification checks the complete presence of reasoning content, answer content, correct format, and EOA marker, awarding additional 0.25 points when passing verification.

D.5 Evaluation Metrics

We employ three primary evaluation metrics to assess model performance:

Accuracy. The correctness rate of models on scientific reasoning benchmarks such as MMLU-Pro, SuperGPQA, and AGI-Eval, measuring the proportion of questions answered correctly.

Reasoning Length. The average token count of generated reasoning sequences, serving as a proxy metric for computational cost. To ensure fair comparison across different models, we uniformly employ the GPT-4 tokenizer (cl100k_base) for computation, avoiding incomparability arising from tokenization differences across model-specific tokenizers.

Accuracy-Efficiency Score (AES). Following Luo et al. (2025), we adopt the Accuracy-Efficiency Score to evaluate reasoning optimization effectiveness. Define length reduction rate $\Delta\text{Length} = \frac{\text{Length}_{\text{baseline}} - \text{Length}_{\text{model}}}{\text{Length}_{\text{baseline}}}$ and accuracy change $\Delta\text{Acc} = \text{Acc}_{\text{model}} - \text{Acc}_{\text{baseline}}$. The AES is computed as:

$$\text{AES} = \begin{cases} \alpha \cdot \Delta\text{Length} + \beta \cdot |\Delta\text{Acc}|, & \text{if } \Delta\text{Acc} \geq 0 \\ \alpha \cdot \Delta\text{Length} - \gamma \cdot |\Delta\text{Acc}|, & \text{if } \Delta\text{Acc} < 0 \end{cases} \tag{13}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$. By setting $\gamma > \beta$, we emphasize the penalty for accuracy degradation. Following the original work, we use $\alpha = 1$, $\beta = 3$, and $\gamma = 5$.

E Dataset Information

This section provides detailed information about the six scientific reasoning benchmark datasets used in this study, including three core training and evaluation datasets and three out-of-domain generalization test datasets. Table 3 summarizes the dataset usage, metrics, and quantities for training and evaluation phases.

E.1 In-Domain Datasets

The following three datasets form the core training and in-domain evaluation benchmarks for the *SynergiSci* framework, spanning foundational to graduate-level scientific reasoning tasks.

AGI-Eval (Zhong et al., 2023): A benchmark specifically designed to evaluate model comprehensive abilities in human standardized exams. All questions originate from high-standard, officially published qualification and admission exams for human test-takers, such as Chinese Gaokao, American SAT, Law School Admission Test (LSAT), math competitions, and bar exams. As the foundational difficulty component in core benchmarks, AGI-Eval primarily evaluates model basic reasoning and problem-solving abilities in general education.

MMLU-Pro (Wang et al., 2024): An enhanced version of the MMLU benchmark, providing a more robust and challenging multi-task language understanding evaluation. It significantly increases evaluation difficulty by integrating more challenging, reasoning-focused questions. The dataset covers 14 different subject areas with over 12,000 questions. It removes overly simple or noisy questions from the original MMLU and deliberately designs complex problems requiring multi-step chain-of-thought reasoning.

SuperGPQA (Team et al., 2025b): A large-scale, high-difficulty graduate-level comprehensive

[1] https://github.com/huggingface/trl
[2] https://github.com/unslothai/unsloth

Task	Dataset	Metric	Quantity
Training			
SFT Pre-training (CoT Generation)	MMLU-Pro, SuperGPQA, AGI-Eval (Sampled)	N/A	500
GRPO Main Training	Composite (from MMLU-Pro, SuperGPQA, AGI-Eval)	Accuracy	9,136
In-domain Evaluation			
Core Scientific Reasoning (Easy to Hard)	Composite Test Set (AGI-Eval, MMLU-Pro, SuperGPQA)	Accuracy	3,367
Out-of-domain Evaluation			
Basic Scientific QA (Easy)	SciQ	Accuracy	1,000
Multi-level Scientific Reasoning (Medium)	SciEval	Accuracy	1,000
Text-based Scientific QA (Hard)	ScienceQA (Text-only)	Accuracy	1,000

Table 3: Dataset usage overview for training and evaluation phases.

knowledge and reasoning capability benchmark designed to evaluate model performance at the human knowledge frontier. The benchmark covers 13 major disciplines, 72 fields, and 285 graduate-level subfields. All questions undergo careful screening and optimization to ensure professionalism and challenge.

E.2 Out-of-Domain Datasets

To assess generalization capability, we evaluate on three out-of-domain scientific reasoning benchmarks that were not included in training data.

SciQ (Pedersen et al., 2020): A crowdsourced dataset containing 13,700 multiple-choice science questions. It is not directly collected from existing exams but constructed through an innovative human-machine collaboration method.

ScienceQA (Saikh et al., 2022): A multimodal scientific question answering dataset containing over 21,000 questions spanning natural science, language science, and social science topics. In this study, we exclusively use the text-only subset to ensure consistency with our text-based reasoning framework, excluding questions that require image comprehension.

SciEval (Sun et al., 2023): A multi-level, multi-disciplinary comprehensive evaluation benchmark designed for scientific foundation models. The benchmark contains over 20,000 questions covering natural sciences, engineering, and medicine.

F Baseline Model Details

This section provides detailed information about the six baseline models used for performance comparison in this study, including two closed-source commercial models, two open-source general-purpose models, and two scientific domain-specific models.

SciGLM-6B (Zhang et al., 2024): An open-source model fine-tuned on the high-quality scientific instruction dataset SciInstruct. Its core feature is enhanced capability for complex reasoning in physics, chemistry, mathematics, and formal proof domains. We select it as an in-domain baseline to directly evaluate the actual contribution of domain-adaptive fine-tuning strategies to deep scientific knowledge acquisition.

Intern-S1-mini (Bai et al., 2025): A lightweight yet powerful multimodal scientific foundation model. Based on Qwen3-8B, it undergoes deep pre-training with over 2.5T scientific domain corpora. We use it as a domain-specialized capability reference to examine the peak performance achievable by specialized scientific models with relatively small parameter scales, providing a high-standard comparison for our model.

DeepSeek-R1-Distill-Qwen2.5-7B (Guo et al., 2025): This model transfers the complex reasoning capabilities of the powerful DeepSeek-R1 model to a more efficient Qwen2.5-7B model through knowledge distillation. It represents an advanced technical pathway for preserving core reasoning capabilities through model compression. We select it as a baseline to evaluate our method's performance advantages in parameter-constrained realistic scenarios.

DeepSeek-V3.1 (Liu et al., 2024): An open-source model specifically designed for low-latency and high-throughput reasoning scenarios. Its distinguishing feature lies in deep optimization at the model architecture and implementation levels, aiming to provide rapid responses while maintaining strong general capabilities. We include it in the comparison to measure our method's performance in the reasoning efficiency dimension, particularly its potential value in engineering applications re-

quiring fast decision-making.

GPT-5-Nano (OpenAI, 2025a): As an industry-leading closed-source model, GPT-5-Nano represents cutting-edge general large models' exploration toward efficiency and lightweight design. Despite its smaller parameter scale, it inherits the powerful general knowledge and complex reasoning capabilities of its flagship model series. We select it as a performance upper-bound reference to directly benchmark our model against the most advanced commercial AI technology.

o4-mini (OpenAI, 2025b): A widely popular mainstream closed-source commercial model, renowned for its excellent balance among speed, cost, and intelligence. It provides near-state-of-the-art reasoning capabilities for large-scale, low-latency practical applications, serving as an important benchmark for measuring real-world model usability. We use it as another black-box reference baseline to evaluate our method's performance in commercialization potential and practical deployment value.

G Reasoning Trajectory Analysis Standards

To quantitatively analyze the reasoning structure improvements achieved by the *SynergiSci* framework, we employ an LLM-as-a-judge approach to automatically annotate reasoning trajectory components on 100 sampled instances from AGI-Eval, using the GLM-4.5 model (Zeng et al., 2025a) with temperature=0.1 to ensure annotation consistency. Below we detail the eight core component categories used for analyzing scientific reasoning trajectories. This classification system references the reasoning pattern analysis of Wu et al. (2024), draws from the key dimensions of Lee and Hockenmaier (2025), and is optimized for the characteristics of scientific reasoning tasks.

Problem Restatement (PR): This component refers to paraphrasing or reformulating the original question to clarify the solving objective. It typically manifests as a direct statement of the solving task (e.g., "This problem requires calculating the final velocity"), a summary of given conditions, or identification of the problem's domain (e.g., "This is a thermodynamics problem"). PR establishes a clear goal and direction for the subsequent reasoning process.

Knowledge Invocation (KI): Knowledge invocation refers to explicitly introducing relevant scientific principles, laws, formulas, or concepts during the reasoning process. This includes directly citing law or formula names (e.g., "According to Newton's second law, F=ma"), explaining theories that serve as derivation bases (e.g., "Based on the law of energy conservation"), or using standardized scientific terminology and symbols. KI provides theoretical foundations for the entire reasoning chain to ensure its scientific validity.

Core Deduction (CD): This component encompasses key derivation steps explicitly employed in the final solution that directly advance problem-solving progress. It specifically manifests as critical mathematical operations performed based on known information (e.g., "a = F/m = 10/2 = 5 m/s^2"), core derivations repeatedly mentioned in solution summaries (e.g., "From PV=nRT we get $P_1V_1 = P_2V_2$"), or decisive derivations connecting premises to final conclusions. CD is the critical link that drives the solving process and produces new information.

Effective Extension (EE): This category includes computations, deductions, or planning necessary to support core derivation steps but not repeated in the final summary. This includes intermediate calculations that pave the way for main derivations (e.g., "First transform the formula to $v^2 = v_0{}^2 + 2as$"), necessary unit conversions or symbol processing (e.g., "Convert angles to radians"), and logical preparatory work. While EE does not directly appear in the final solution, it makes substantial contributions to the solving process.

Alternative Exploration (AE): Alternative exploration refers to attempting different solution paths or examining other solving strategies, even if these methods are ultimately not used in the complete solution. It manifests as consideration of multiple possible methods (e.g., "We could use the energy method or kinematic equations"), comparison of different strategies (e.g., "Perhaps the differential method is simpler"), or exploration of alternative approaches (e.g., "Another approach is to use conservation laws"). AE reflects the flexibility and comprehensiveness of the reasoning process.

Verification & Correction (VC): This component covers statements that check and correct earlier results to discover errors. This includes verification of calculation results (e.g., "Let's check: 5×2=10, correct"), confirmation of answer reasonableness (e.g., "This velocity value aligns with physical intuition"), and correction of earlier errors (e.g., "Wait, that calculation was wrong, it should

be 15”). VC ensures the accuracy and reliability of the reasoning process.

Redundant Filling (RF): This category includes redundant comments, filler content, or trivial restatements that make no substantial contribution to the solving process. It primarily encompasses conversational transitions, metacognitive comments (e.g., “Now, let’s proceed to the next calculation”), obvious restatements, or decorative language that cannot advance the solution. In analysis, identifying RF helps evaluate the efficiency and conciseness of the reasoning process.

Logical Error (LE): Logical errors represent any form of fallacy occurring in the reasoning process. This includes numerical calculation errors (e.g., “10 ÷ 2 = 6”), misuse of scientific concepts or laws (e.g., “According to Newton’s first law F=ma”), and inferences that violate logical rules (e.g., “Because temperature increases, ideal gas volume must decrease”). Identifying LE is core to evaluating reasoning accuracy and reliability.

For concrete examples illustrating each component category, see Appendix H.

H Reasoning Trajectory Component Examples

This section provides concrete examples of each reasoning trajectory component category used in our fine-grained analysis. These examples are extracted from actual reasoning trajectories generated by models on general scientific reasoning tasks.

Problem Restatement (PR)

"The patient’s symptoms and lab results suggest a condition that requires immunosuppressive treatment, specifically considering his chronic symptoms and recurrent presentation."

Knowledge Invocation (KI)

"For ANCA-associated vasculitis with severe disease (pulmonary hemorrhage and RPGN), treatment typically includes high-dose corticosteroids (methylprednisolone) and cyclophosphamide, with plasmapheresis considered for severe cases (especially if dialysis-dependent or with diffuse alveolar hemorrhage)."

Core Deduction (CD)

"Therefore, Option C (methylprednisolone, plasmapheresis, and cyclophosphamide) is the most appropriate treatment for this patient."

Effective Extension (EE)

"Additionally, the elevated blood and protein in the urine suggest renal involvement, which is common in vasculitic conditions like Goodpasture syndrome or ANCA-associated vasculitis."

Alternative Exploration (AE)

"Methylprednisolone alone (Option A) is insufficient for the severity of the condition, as it does not address the underlying vasculitis or the need for antibody removal."

Verification & Correction (VC)

"Let me double-check: the patient has hemoptysis + RPGN + positive ANCA, which confirms vasculitis diagnosis and the need for aggressive treatment including plasmapheresis."

Redundant Filling (RF)

"The patient presents with hemoptysis, diffuse muscle pain, and shortness of breath, which are concerning symptoms that may indicate a severe underlying condition such as pulmonary hemorrhage or vasculitis."

Logical Error (LE)

"This combination is the standard of care for severe cases of ANCA-associated vasculitis, particularly when there is evidence of pulmonary hemorrhage and renal involvement"

I Prompts

I.1 Reasoning Trajectory Generation Prompt

As illustrated in Figure 6, we provide the complete prompt template used to generate structured reasoning trajectories during the training phase.

Task Description:
You are a top-tier logical analyst and expert educator. Your task is to analyze the following science question with its multiple-choice options and provide a step-by-step reasoning process that leads to the correct answer.
Input:
Question: A question followed by a list of alphabetically labeled choices.
Correct Answer Text: The full text of the correct answer is provided to you for verification purposes.
Output Format:
CRITICAL: You MUST strictly adhere to the following XML-style format for your entire response. Do NOT use Markdown headings like '### Reasoning'. Your response must begin directly with the `<reasoning>` tag.

```
<reasoning>
{{Detailed analysis process}}
</reasoning>
<answer>[ANSWER]{{Letter of the correct answer}}<eoa>
</answer>
```

- **`<reasoning>` block**: Within these tags, you must provide a detailed analysis. First, analyze the question's core concept. Then, evaluate each option, explaining why the correct one is right and the others are wrong, referencing their letter labels (A, B, C, etc.).
- **`<answer>` block**: Within these tags, you must output the final answer in the format `[ANSWER]{{LETTER}}<eoa>`, where `{{LETTER}}` is the single capital letter of the correct option.

Figure 6: Prompt template for generating structured reasoning trajectories during training.

I.2 LLM-as-a-Judge Trajectory Analysis Prompt

Figure 7 presents the prompt template for analyzing and classifying reasoning trajectory components using LLM-as-a-judge methodology.

J Case Study

This section presents a representative case study demonstrating the quality-efficiency trade-offs across different models in scientific reasoning tasks. Figure 8 provides the clinical case description, while Figures 9, 10, 11, 12, and 13 show the reasoning trajectories generated by *SynergiSci* models (Qwen2.5 and LLaMA3.1 backbones), scientific domain baseline (Intern-S1-mini), and commercial models (o4-mini and DeepSeek-V3.1), respectively.

J.1 Comparative Analysis

The pulmonary–renal syndrome case reveals distinct reasoning patterns across models, illuminating the efficacy of our dual-stage optimization framework. While all models arrive at the correct diagnosis (Option C), their reasoning trajectories exhibit marked differences in quality-efficiency trade-offs that merit systematic examination.

Superior reasoning efficiency through targeted deduction. Our *SynergiSci* framework demonstrates exceptional reasoning economy without compromising diagnostic rigor. The LLaMA3.1-based model achieves the correct answer in 350 tokens, while the Qwen2.5 variant requires 450 tokens—both substantially more efficient than baseline models. Compared to the specialized scientific model Intern-S1-mini (637 tokens), our approach achieves 45% token reduction for LLaMA3.1 and 29% for Qwen2.5. The efficiency gains are even more pronounced against commercial models: 48% reduction versus o4-mini (670 tokens) and 62% versus DeepSeek-V3.1 (919 tokens) for the LLaMA3.1 variant. Crucially, this efficiency stems not from superficial brevity but from strategic elimination of redundant components. The information gain-guided quality learning (Stage 1) enables the model to prioritize high-impact deductive steps: it directly identifies the pulmonary–renal syndrome, invokes ANCA-associated vasculitis and Goodpasture's disease as primary differential diagnoses, and immediately connects pathophysiology to treatment modalities. In contrast, DeepSeek-V3.1's exhaustive structured framework, while pedagogically comprehensive, introduces substantial redundant filling (RF) through detailed subsection headers ("Clinical Presentation Analysis", "Laboratory Interpretation", "Treatment Rationale") and extensive differential enumerations that contribute minimally to the final decision.

Enhanced core deduction density and verification rigor. Beyond length reduction, our framework exhibits superior reasoning structure as evidenced by component analysis (detailed in Section 4.3.2 and Appendix G). The LLaMA3.1 output exemplifies this with remarkably high core de-

duction (CD) density: after concisely restating the syndrome and invoking the differential, it devotes the majority of reasoning tokens to evaluating treatment options against the triple-therapy rationale. The explicit verification mechanism—"A (steroids alone) is inadequate for severe disease... B lacks plasmapheresis... D omits steroids"—reflects the Stage 1 reward shaping that penalizes logical gaps, resulting in a verification & correction (VC) component rate of 13% compared to Intern-S1-mini's 10.6%. Notably, our model achieves this rigor without the verbose knowledge invocation (KI) redundancy observed in Intern-S1-mini, which restates treatment principles already implicit in the differential diagnosis. The dynamic length control (Stage 2) further refines this by adaptively truncating effective extension (EE) components that, while valid, offer diminishing returns—our models omit the detailed biochemical explanations of plasmapheresis mechanisms found in o4-mini.

Critical limitations of competing approaches. Baseline models exhibit systematic inefficiencies that our framework addresses. Intern-S1-mini (637 tokens), despite scientific domain pretraining, produces output with excessive problem restatement (PR) and knowledge invocation (KI) redundancy—traits consistent with models lacking explicit reasoning quality rewards. Its 11.33% redundant filling rate, 56.9% higher than our 4.88%, suggests that domain adaptation alone cannot internalize concise reasoning without reinforcement shaping. The commercial models demonstrate even more pronounced inefficiency: DeepSeek-V3.1's 919-token reasoning exhaustively enumerates differential diagnoses and treatment principles, while o4-mini's 670-token output employs a rigid six-step schema that treats all cases identically regardless of complexity. Both reflect the over-generation problem endemic to quality-only optimization: absent length penalties, these models exhaustively explore clinical nuances that, while academically comprehensive, violate the efficiency constraint critical for real-world deployment. This inflexibility is precisely what our dynamic length control mechanism (Stage 2) overcomes by calibrating reasoning depth to problem difficulty.

Implications for quality-efficiency synergy. This case study substantiates our framework's central thesis: high-quality scientific reasoning need not entail computational extravagance. By decoupling quality learning (which establishes what constitutes rigorous deduction) from efficiency optimization (which learns to achieve that rigor minimally), we avoid the degradation observed when length constraints are naively imposed. The LLaMA3.1 output's preservation of diagnostic accuracy despite 48% length reduction relative to o4-mini and 62% reduction versus DeepSeek-V3.1 validates that much of baseline reasoning constitutes learnable redundancy rather than essential complexity. Moreover, the cross-model consistency—both Qwen2.5 and LLaMA3.1 backbones exhibit similar improvements—demonstrates that our reward shaping mechanisms generalize beyond architecture-specific quirks, offering a robust pathway toward deployable scientific reasoning models that honor both the epistemic standards of the domain and the economic realities of inference at scale.

K Ethics, Broader Impact, and Licenses

K.1 Ethics

Our research focuses on improving the quality-efficiency trade-offs in scientific reasoning for large language models through a dual-stage reinforcement learning framework. We emphasize responsible AI development by prioritizing both reasoning accuracy and computational efficiency, which aligns with sustainable AI practices. Our methodology adheres to widely accepted ethical standards in AI research, prioritizing transparency and reproducibility.

In our experiments, we evaluate model performance on scientific reasoning benchmarks covering multiple domains including medicine, physics, chemistry, and mathematics. Some datasets may contain medical case descriptions and scientific scenarios. All such datasets are sourced from publicly available, peer-reviewed academic benchmarks with established usage guidelines. We use these datasets strictly for academic research purposes under controlled conditions, following their intended use as specified by the original creators.

Our training methodology does not involve human annotation or crowdsourcing. All model training is conducted using publicly available base models and datasets, with proper attribution and compliance with their respective licenses.

K.2 Broader Impact

Improving the efficiency and quality of scientific reasoning in LLMs has significant potential for positive outcomes in high-impact domains such

as healthcare (medical diagnosis support), education (personalized scientific tutoring), and research (scientific literature analysis and hypothesis generation). By reducing computational costs while maintaining accuracy, our framework makes advanced scientific reasoning capabilities more accessible to resource-constrained institutions and researchers.

However, we recognize potential risks associated with deploying scientific reasoning models:

Overreliance and automation bias. Users may over-trust model outputs without proper verification, particularly in high-stakes domains like medical diagnosis. We emphasize that our models are designed as decision-support tools, not replacements for human expertise.

Misuse in critical applications. While our framework improves reasoning quality, LLMs can still generate plausible but incorrect reasoning chains. Deployment in safety-critical scenarios requires rigorous validation and human oversight.

Environmental impact. Although our method reduces inference costs by 28.6%, the training process still requires substantial computational resources. We encourage practitioners to consider carbon footprint and prioritize efficient training practices.

Equity and access. While our efficiency improvements reduce computational barriers, access to the underlying base models and computational infrastructure remains unevenly distributed globally.

We encourage proactive research on model auditing, uncertainty quantification, and regulatory frameworks to identify, monitor, and mitigate these risks before deployment in real-world applications.

K.3 Licenses

Models. We use the following base models with their respective licenses:

- Qwen2.5-7B-Instruct (Apache 2.0 License)
- LLaMA3.1-8B-Instruct (Llama 3.1 Community License)

Datasets. Our experiments use the following publicly available benchmarks:

- MMLU-Pro (MIT License)
- SuperGPQA (Open Data Commons Attribution License)
- AGI-Eval (MIT License)
- SciQ (CC BY-NC 3.0 License)
- ScienceQA (CC BY-SA 4.0 License)
- SciEval (Apache 2.0 License)

All artifacts are used strictly in accordance with their intended purposes as specified by their creators. Our use of these datasets for scientific reasoning research is consistent with their original intended use for evaluating language model capabilities. We do not redistribute modified versions of these datasets or use them for commercial purposes outside the scope permitted by their licenses.

Code and Trained Models. We will release our code and implementation details to facilitate reproducibility. Due to licensing considerations of the base models and computational constraints, trained model weights will be made available subject to the same license terms as the underlying base models. For more details about the licenses and usage permissions, please refer to the official documentation of each model and dataset.

You are a reasoning trajectory analyst. Your role is to classify each sentence based on its function within a given reasoning trajectory, to understand how different parts of the model's reasoning process contribute to the final answer.
You will be provided with a **scientific question** and its corresponding **reasoning trajectory**. You must:

1. Split the reasoning trajectory into individual sentences or semantically complete phrases
2. **Assign exactly one label** from the following eight mutually exclusive categories to each segment

Classification Standards:

1. **Problem Restatement (PR)**
   - *Definition:* Expression of understanding, reformulation, or goal-setting for the original problem
   - *Example:* "This problem requires us to calculate the object's acceleration"
2. **Knowledge Invocation (KI)**
   - *Definition:* Explicit statement invoking relevant scientific knowledge, formulas, theorems, or concepts
   - *Example:* "According to Newton's second law, F=ma"
3. **Core Deduction (CD)**
   - *Definition:* Key derivation steps explicitly employed in the final solution
   - *Example:* "Substituting values into the formula: a = 10N / 2kg = 5m/s^2"
4. **Effective Extension (EE)**
   - *Definition:* Necessary computations supporting core deduction but not appearing in final summary
   - *Example:* "First transform the formula to $v^2 = v_0{}^2 + 2as$"
5. **Alternative Exploration (AE)**
   - *Definition:* Attempting different solution paths, even if not fully adopted
   - *Example:* "We could use the energy method or kinematic equations"
6. **Verification & Correction (VC)**
   - *Definition:* Statements checking results and correcting errors
   - *Example:* "Let's check: 5×2=10, correct"
7. **Redundant Filling (RF)**
   - *Definition:* Repetitive statements or conversational filler with no substantial contribution
   - *Example:* "Now let's continue to the next step..."
8. **Logical Error (LE)**
   - *Definition:* Incorrect derivation, calculation errors, or statements contradicting scientific principles
   - *Example:* "10÷2 = 6" (calculation error)

Output Format:
Return your output as a **JSON array**, with each object containing:
- "text": original text segment (string)
- "component": component type abbreviation (string)
- "explanation": annotation rationale (string)

And provide a statistical summary at the end:
- "summary": containing component counts and quality-efficiency ratios

Only return the structured JSON result, without adding any commentary or explanation.

Figure 7: Prompt template for LLM-as-a-judge trajectory component analysis.

A 47-year-old man presents with hemoptysis, diffuse muscle pain, and shortness of breath. He has had these symptoms before but did not seek treatment. His past medical history is not known as he is not typically followed by a doctor and is homeless. His temperature is 99.0°F (37.2°C), blood pressure is 154/94 mmHg, pulse is 89/min, respirations are 22/min, and oxygen saturation is 90% on room air. Physical examination is notable for crackles noted in the bases of the lungs bilaterally.
Lab values:
Serum: Na^+: 140 mEq/L, Cl^-: 103 mEq/L, K^+: 5.8 mEq/L, HCO_3^-: 21 mEq/L, BUN: 33 mg/dL, Glucose: 129 mg/dL, Creatinine: 2.6 mg/dL
Urine: Color: Amber, Protein: Positive, Blood: Positive
Question: Which of the following is the most appropriate treatment for this patient?
A. Methylprednisolone
B. Methylprednisolone and cyclophosphamide
C. Methylprednisolone, plasmapheresis, and cyclophosphamide
D. Plasmapheresis and cyclophosphamide

Figure 8: Clinical presentation of pulmonary-renal syndrome case used for model comparison.

***SynergiSci* (Qwen2.5-7B-Insturct Backbone)**

<reasoning>
The patient presents with hemoptysis, diffuse muscle pain, and shortness of breath, which are concerning symptoms that may indicate a severe underlying condition such as pulmonary hemorrhage or vasculitis. The physical examination reveals crackles in the bases of the lungs bilaterally, suggesting pulmonary involvement. The laboratory findings include hyperkalemia (K^+: 5.8 mEq/L), which is a critical electrolyte abnormality that can cause muscle weakness and pain, consistent with the patient's symptoms. Additionally, the elevated blood and protein in the urine suggest renal involvement, which is common in vasculitic conditions like Goodpasture syndrome or ANCA-associated vasculitis.
The most appropriate treatment for this patient is a combination of methylprednisolone to rapidly suppress inflammation, cyclophosphamide to target the underlying vasculitis, and plasmapheresis to remove circulating antibodies that may be causing the hemoptysis and renal involvement. This combination is the standard of care for severe cases of ANCA-associated vasculitis. Methylprednisolone alone (Option A) is insufficient. Cyclophosphamide alone (Option B) lacks antibody removal. Option D omits methylprednisolone which is essential for rapid inflammation control. Therefore, Option C is the most appropriate treatment.
</reasoning>
<answer>[ANSWER]C</answer>

Figure 9: Reasoning trajectory from *SynergiSci* (Qwen2.5-7B) model.

***SynergiSci* (LLaMA3.1-8B-Instruct Backbone)**

<reasoning>
The patient presents with recurrent hemoptysis, diffuse muscle pain, and shortness of breath, with crackles at the lung bases suggesting pulmonary hemorrhage. Laboratory results show acute kidney injury (elevated creatinine, BUN), hyperkalemia, metabolic acidosis, and urinalysis with hematuria and proteinuria. This constellation defines a pulmonary–renal syndrome. The main causes are Goodpasture's disease (anti-GBM) and ANCA-associated vasculitides such as granulomatosis with polyangiitis (GPA) or microscopic polyangiitis.
Both conditions are life-threatening and require rapid immunosuppressive therapy. Corticosteroids (methylprednisolone) provide fast anti-inflammatory action. Cyclophosphamide prevents further autoantibody production. Plasmapheresis directly removes circulating pathogenic antibodies and is critical when pulmonary hemorrhage and renal failure coexist.
Evaluating the options: A (steroids alone) is inadequate for severe disease. B (steroids + cyclophosphamide) lacks plasmapheresis, leaving circulating antibodies unchecked. D (plasmapheresis + cyclophosphamide) omits steroids, which are essential for rapid control. C (steroids + cyclophosphamide + plasmapheresis) combines all three and is the standard of care in severe pulmonary–renal syndromes. Therefore, Option C is the most appropriate treatment.
</reasoning>
<answer>[ANSWER]C</answer>

Figure 10: Reasoning trajectory from *SynergiSci* (LLaMA3.1-8B) model.

Intern-S1-mini (8B)

<reasoning>
The patient presents with hemoptysis, muscle pain, and shortness of breath, which are nonspecific but potentially indicative of inflammatory or immune-mediated conditions. Physical examination reveals crackles in the lung bases bilaterally, suggesting pulmonary involvement. Laboratory findings show elevated creatinine (2.6 mg/dL) and BUN (33 mg/dL), indicating renal impairment. Urine shows protein and blood, suggesting nephritic syndrome with glomerular involvement. The constellation of symptoms and lab findings raises suspicion for systemic autoimmune diseases such as **Goodpasture syndrome** or **Granulomatosis with polyangiitis (Wegener's granulomatosis)**. Given the renal and pulmonary involvement, Goodpasture syndrome (anti-GBM disease) is a strong possibility.
The standard treatment for severe anti-GBM disease involves **plasmapheresis** to remove circulating anti-GBM antibodies, along with immunosuppressive therapy such as **cyclophosphamide**, and **methylprednisolone** to suppress immune activity. Plasmapheresis is crucial in acute settings to reduce antibody load rapidly.
Option Analysis:
A) Methylprednisolone: Alone is insufficient for severe cases requiring rapid antibody reduction.
B) Methylprednisolone and cyclophosphamide: Lacks the immediate antibody reduction needed in severe cases.
C) Methylprednisolone, plasmapheresis, and cyclophosphamide: This combination addresses both antibody reduction and immunosuppression effectively.
D) Plasmapheresis and cyclophosphamide: Missing methylprednisolone, which is critical for full immunosuppression.
Therefore, the most appropriate treatment is **Option C**.
</reasoning>
<answer>[ANSWER]C</answer>

Figure 11: Reasoning trajectory from Intern-S1-mini scientific domain-specialized model.

o4-mini

<reasoning>
Step 1: Identify the clinical syndrome. The patient presents with hemoptysis, recurrent shortness of breath, and inspiratory crackles at the lung bases, which strongly suggests pulmonary hemorrhage or alveolar bleeding. Concurrently, he has acute renal impairment as evidenced by creatinine elevation (2.6 mg/dL), BUN elevation, and laboratory features of metabolic acidosis with hyperkalemia. Urinalysis reveals hematuria and proteinuria, both of which are classic for glomerular involvement rather than simple post-renal or pre-renal azotemia. When pulmonary hemorrhage is paired with rapidly progressive glomerulonephritis, the unifying syndrome is referred to as a pulmonary–renal syndrome, which is a medical emergency.
Step 2: Develop the differential diagnosis. The major entities that can cause pulmonary–renal syndrome include: (1) Goodpasture's syndrome (anti–GBM disease), mediated by antibodies directed against the glomerular and alveolar basement membranes. (2) Granulomatosis with polyangiitis (GPA, Wegener's granulomatosis), an ANCA-associated vasculitis that usually also has upper respiratory tract disease. (3) Microscopic polyangiitis (MPA), another ANCA-associated small vessel vasculitis. (4) Less common considerations include systemic lupus erythematosus with diffuse alveolar hemorrhage or cryoglobulinemic vasculitis.
Step 3: Determine the most likely working diagnosis. The clinical picture of recurrent hemoptysis plus acute kidney injury in a middle-aged male is most consistent with Goodpasture's disease. The absence of upper airway involvement argues against GPA.
Step 4: Review the pathophysiology and treatment principles. Goodpasture's syndrome is driven by circulating anti-GBM antibodies that bind to type IV collagen within basement membranes of alveoli and glomeruli. Therefore, therapy must address both the existing antibodies and the ongoing immune response: (1) Plasmapheresis: physically removes anti-GBM antibodies from circulation. (2) High-dose corticosteroids (methylprednisolone pulses): rapidly reduce inflammation, capillaritis, and immune-mediated damage. (3) Cyclophosphamide: a potent cytotoxic agent that prevents further antibody formation by suppressing B-cell function. These three interventions are complementary and synergistic.
Step 5: Analyze the treatment options provided. Option A (methylprednisolone alone): inadequate. Option B (methylprednisolone + cyclophosphamide): still insufficient for Goodpasture's. Option D (plasmapheresis + cyclophosphamide): incomplete, as this omits high-dose corticosteroids. Option C (methylprednisolone + plasmapheresis + cyclophosphamide): incorporates all three modalities and represents the gold standard.
Step 6: Conclusion. In a patient with pulmonary–renal syndrome, the optimal evidence-based therapy is the combination of steroids, plasmapheresis, and cyclophosphamide. Therefore, the most appropriate answer is Option C.
</reasoning>
<answer>[ANSWER]C</answer>

Figure 12: Reasoning trajectory from o4-mini closed-source commercial model.

DeepSeek-V3.1

<reasoning>
Clinical Presentation Analysis: The patient is a 47-year-old homeless male presenting with a triad of hemoptysis, diffuse muscle pain (myalgias), and shortness of breath with hypoxemia (oxygen saturation 90% on room air). Physical examination reveals bibasilar pulmonary crackles, suggesting alveolar or interstitial lung disease. The recurrent nature of symptoms without prior medical intervention raises concern for a chronic, progressive systemic disorder.
Laboratory Interpretation: Serum chemistry demonstrates acute kidney injury (AKI) with elevated creatinine (2.6 mg/dL) and BUN (33 mg/dL), accompanied by hyperkalemia (K^+ 5.8 mEq/L) and a low-normal bicarbonate (21 mEq/L), suggestive of metabolic acidosis secondary to renal dysfunction. The urinalysis is particularly revealing: amber-colored urine with both proteinuria and hematuria is pathognomonic for glomerular pathology, specifically indicating glomerulonephritis rather than tubular or prerenal azotemia.
Syndrome Recognition: The simultaneous occurrence of diffuse alveolar hemorrhage (DAH), evidenced by hemoptysis and pulmonary crackles, alongside rapidly progressive glomerulonephritis (RPGN), as indicated by AKI with active urinary sediment, constitutes a pulmonary-renal syndrome. This is a critical medical emergency requiring immediate diagnostic and therapeutic intervention.
Differential Diagnosis: Pulmonary-renal syndromes are primarily caused by small-vessel vasculitides and autoimmune processes. The leading considerations include: (1) **Anti-GBM disease (Goodpasture syndrome)**: Characterized by antibodies against glomerular and alveolar basement membranes, typically affecting younger males and smokers. (2) **ANCA-associated vasculitis (AAV)**: Including granulomatosis with polyangiitis (GPA), microscopic polyangiitis (MPA), and eosinophilic granulomatosis with polyangiitis (EGPA), all of which can present with DAH and RPGN. (3) **Systemic lupus erythematosus (SLE)**: Can rarely present with lupus nephritis and pulmonary hemorrhage. (4) **Immune complex-mediated diseases**: Such as IgA vasculitis (Henoch-Schönlein purpura) or cryoglobulinemic vasculitis.
Treatment Rationale: Regardless of the specific etiology within the pulmonary-renal syndrome spectrum, the cornerstone of acute management involves three synergistic therapeutic modalities: (1) **High-dose corticosteroids (methylprednisolone)**: Rapidly suppress acute inflammation and immune-mediated tissue damage. Intravenous pulse-dose methylprednisolone (typically 500-1000 mg daily for 3 days) is the standard initial approach to control life-threatening manifestations. (2) **Cyclophosphamide**: A potent immunosuppressive agent that inhibits B-cell and T-cell proliferation, thereby reducing autoantibody production and preventing disease progression. Cyclophosphamide is essential for long-term disease control and prevention of relapse. (3) **Plasmapheresis (therapeutic plasma exchange)**: Mechanically removes circulating pathogenic autoantibodies (anti-GBM or ANCA) and immune complexes from the bloodstream, providing rapid reduction in antibody-mediated damage. Plasmapheresis is particularly critical in severe cases with DAH or dialysis-dependent renal failure, as it can improve outcomes when initiated early.
Option Evaluation:
Option A (Methylprednisolone alone): While corticosteroids address acute inflammation, monotherapy is insufficient for pulmonary-renal syndrome. Without cyclophosphamide to prevent ongoing autoantibody production or plasmapheresis to clear existing antibodies, renal and pulmonary damage will likely progress despite steroid therapy. This option is inadequate for severe disease.
Option B (Methylprednisolone + Cyclophosphamide): This combination provides immunosuppression but lacks plasmapheresis. In the setting of active DAH and RPGN, the circulating autoantibody burden requires immediate removal to prevent irreversible organ damage. Omitting plasmapheresis in severe pulmonary-renal syndrome is suboptimal.
Option C (Methylprednisolone + Plasmapheresis + Cyclophosphamide): This triple therapy represents the gold standard for managing severe pulmonary-renal syndromes. It combines rapid anti-inflammatory action (steroids), antibody clearance (plasmapheresis), and long-term immunosuppression (cyclophosphamide). This approach maximizes the likelihood of preserving renal function and preventing fatal pulmonary hemorrhage.
Option D (Plasmapheresis + Cyclophosphamide): While this regimen addresses antibody removal and immune suppression, the absence of high-dose corticosteroids is a critical deficiency. Steroids are indispensable for controlling acute vasculitic inflammation and preventing immediate life-threatening complications during the lag period before cyclophosphamide exerts its full immunosuppressive effect.
Conclusion: For a patient with pulmonary-renal syndrome manifesting as diffuse alveolar hemorrhage and rapidly progressive glomerulonephritis, the most appropriate initial treatment is the combination of methylprednisolone, plasmapheresis, and cyclophosphamide. This evidence-based approach aligns with established guidelines for managing anti-GBM disease and ANCA-associated vasculitis. Therefore, **Option C** is correct.
</reasoning>
<answer>[ANSWER]C</answer>

Figure 13: Reasoning trajectory from DeepSeek-V3.1 open-source model.