P1: Mastering Physics Olympiads with Reinforcement Learning

Jiacheng Chen*, Qianjia Cheng*, Fangchen Yu*, Haiyuan Wan, Yuchen Zhang, Shenghe Zheng, Junchi Yao, Qingyang Zhang, Haonan He, Yun Luo, Yufeng Zhao, Futing Wang, Li Sheng, Chengxing Xie, Yuxin Zuo, Yizhuo Li, Wenxuan Zeng, Yulun Wu, Rui Huang, Dongzhan Zhou, Kai Chen, Yu Qiao, Lei Bai✉, Yu Cheng✉, Ning Ding✉, Bowen Zhou✉, Peng Ye✉†, Ganqu Cui✉†

P1 Team, Shanghai AI Laboratory

* Equal Contribution ✉ Corresponding Authors † Technical Leads

cuiganqu@pjlab.org.cn, yepeng@pjlab.org.cn P1 Tech Blog

Abstract | Recent progress in large language models (LLMs) has moved the frontier from puzzle-solving to science-grade reasoning—the kind needed to tackle problems whose answers must stand against nature, not merely fit a rubric. Physics is the sharpest test of this shift, which binds symbols to reality in a fundamental way, serving as the cornerstone of most modern technologies. In this work, we manage to advance physics research by developing large language models with exceptional physics reasoning capabilities, especially excel at solving Olympiad-level physics problems. We introduce P1, a family of open-source physics reasoning models trained entirely through reinforcement learning (RL). Among them, P1-235B-A22B is the **first open-source model with Gold-medal performance** at the latest International Physics Olympiad (IPhO 2025), and wins 12 gold medals out of 13 international/regional physics competitions in 2024/2025. P1-30B-A3B also surpasses almost all other open-source models on IPhO 2025, getting a silver medal. Further equipped with an agentic framework PhysicsMinions, P1-235B-A22B+PhysicsMinions achieves overall No.1 on IPhO 2025, and obtains the highest average score over the 13 physics competitions. Besides physics, P1 models also present great performance on other reasoning tasks like math and coding, showing the great generalibility of P1 series.

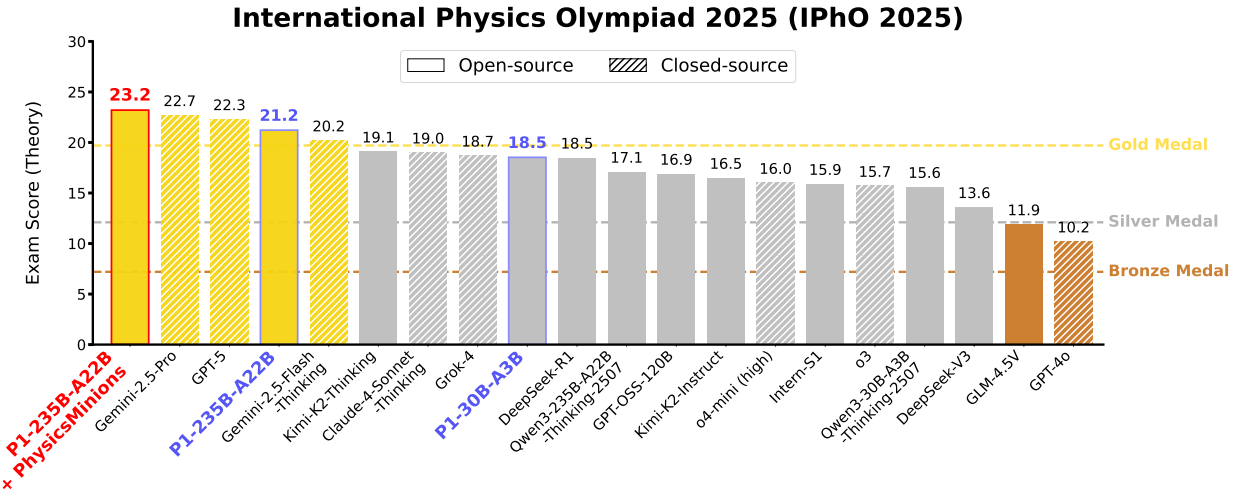

Figure 1 | Breakthrough in open-source physics reasoning: P1−235B−A22B stands as the first and only open-source model to win a gold medal at the International Physics Olympiad 2025 (IPhO 2025), placing 3rd behind Gemini-2.5-Pro and GPT-5. Even at mid-scale, P1−30B−A3B achieved silver and ranked 8th out of 35 evaluated models, outperforming almost all other open-source models. With the PhysicsMinions agent framework, P1−235B−A22B + PhysicsMinions ranks No.1 on IPhO 2025."

Contents

1. Introduction

Recent advances in Large Language Models (LLMs) (Comanici et al., 2025; Guo et al., 2025; Yang et al., 2025) have pushed the boundaries of artificial intelligence from proficiency in symbolic manipulation and general knowledge retrieval to the challenging domain of science-grade reasoning (Bai et al., 2025a; Gibney, 2025; Zhang et al., 2024). This transition marks a critical frontier: models must no longer simply align with rubrics or datasets, but must produce outputs that withstand the rigorous constraints of physical reality and natural law. Among all scientific disciplines, **physics** stands as the sharpest and most unforgiving test of this capability—it binds abstract reasoning with empirical consistency, forming the very foundation of modern science and technology.

Mastering physics requires more than factual recall or formulaic application—it demands conceptual understanding, system decomposition, and precise multi-step reasoning grounded in physical law. These skills are most rigorously tested in **Olympiad-level competitions** such as the International Physics Olympiad (IPhO) (Qiu et al., 2025; Yu et al., 2025b), where each problem condenses the essence of advanced reasoning into a single challenge requiring both analytical precision and creative insight. As such, physics Olympiads provide a high-fidelity and standardized testbed for assessing whether LLMs can exhibit genuine scientific reasoning (Yu et al., 2025a).

We view competitive problem-solving as a critical milestone toward *machine scientific discovery* (Oh et al., 2025; Zheng et al., 2025c). Before models can explore new physical frontiers, they must first demonstrate mastery of human-level reasoning within well-defined laws of nature. Advancing model performance on Olympiad problems is therefore not the end goal, but a necessary step toward building AI systems capable of assisting—or even pioneering—future physics research.

In this work, we introduce **P1**, a new family of open-source physics reasoning models that push the frontier of scientific AI. Our design integrates both *train-time* and *test-time scaling*, ensuring that the model not only acquires stronger reasoning ability through reinforcement learning (RL) (Sutton et al., 1998), but also deploys it adaptively through agentic control at inference.

- **Train-time Scaling.** The P1 models are trained purely through RL post-training (Cui et al., 2025a; Guo et al., 2025) on top of base language models. We propose a multi-stage RL framework that progressively enhances reasoning ability through adaptive learnability adjustment and stabilization mechanisms. This design supports long-term, sustained optimization and effectively mitigates common challenges such as reward sparsity, entropy collapse, and training stagnation.
- **Test-time Scaling.** During inference, we combine P1 models with the *PhysicsMinions* (Yu et al., 2025b) agent framework, which equips the model with iterative correction and self-verification abilities. This framework enables multi-turn reflection—allowing P1 to reason, critique, and refine its own solutions, much like human physicists do. Through structured test-time reasoning, the model extends its effective problem-solving depth without additional training.

We release two variants of the P1 model family and evaluate them on **HiPhO** (Yu et al., 2025a), a new benchmark aggregating the latest 13 Olympiad exams from 2024–2025. Our flagship model, `P1-235B-A22B`, achieves a milestone for the open-source community—becoming the first open model to reach **Gold-medal performance** on the IPhO 2025, earning 12 golds and 1 silver across the full HiPhO suite. The lightweight variant, `P1-30B-A3B`, attains **Silver-medal performance** on IPhO 2025, surpassing almost all prior open baselines. Combined with the PhysicsMinions agent system, P1 attains the **No.1 overall score** on both the IPhO 2025 and HiPhO leaderboards.

Beyond physics, the P1 models demonstrate remarkable **generalizability**. The 30B variant significantly outperforms its base model (Qwen3-30B-A3B-Thinking-2507) across seven math, coding, and general

Statistics	Number
Data Composition	
Total Problems	5,065
- Problems from Olympiads	4,126 (81%)
- Problems from textbooks	939 (19%)
Total Answers	6,911
Total Fields	5
Total Subfields	25
Total Answer Types	6
Data Sources	
Textbooks	10
Olympiad Types	9
- Sets Collected	199
Token Statistics	
Average Question Tokens	367
Max Question Tokens	3386
Average Solution Tokens	349
Max Solution Tokens	5519

Table 1 | Statistics of the training data.

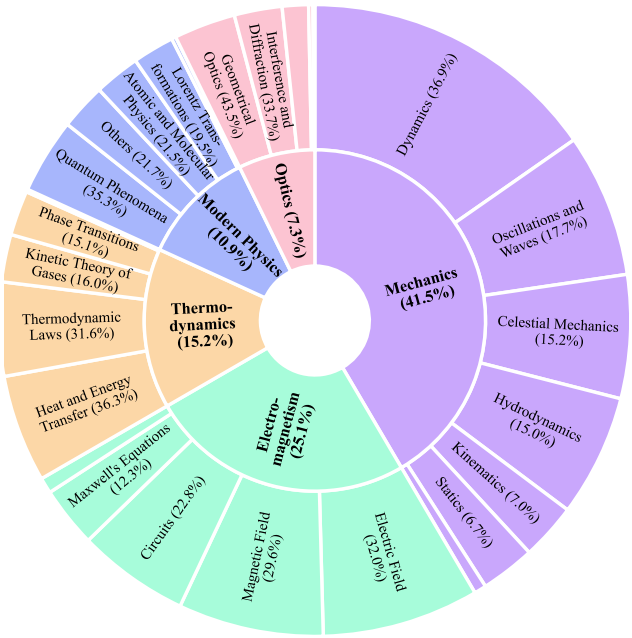

Figure 2 | Field distribution of the training data.

reasoning benchmarks, suggesting that physics post-training cultivates transferable reasoning skills rather than domain overfitting.

Contributions. This work makes the following key contributions:

- We introduce **P1**, the first family of open-source LLMs capable of achieving **Gold-medal performance** at the International Physics Olympiad level.
- We develop a **multi-stage RL post-training framework** with adaptive learnability scaling and training stabilization for sustained reasoning improvement.
- We unveil the **P1 ecosystem**—a fully open-source, end-to-end platform encompassing models, training algorithms, an evaluation benchmark, and an agentic inference framework, providing a unified foundation for advancing scientific reasoning in the open community.

Together, these advances mark a significant step toward LLMs that can engage in genuine scientific reasoning and eventually contribute to the frontier of physics research.

2. Physics Dataset

2.1. Overview

We introduce a systematically curated dataset of 5,065 Olympiad-level, text-based physics problems, designed to advance LLMs toward genuine scientific reasoning. As summarized in Table 1, the dataset combines 4,126 problems from physics Olympiads with 939 from competition textbooks, spanning 5 fields and 25 subfields (see Figure 2). Rather than pursuing broader coverage, we focus on depth and rigor—physics Olympiads uniquely couple abstract reasoning with empirical law, providing an exacting testbed for improving and evaluating model reasoning under physical constraints. Following PHYSICS (Zheng et al., 2025b) and HiPhO (Yu et al., 2025a), we further refine the data construction pipeline by integrating the strengths of human and model annotation,

achieving finer-grained extraction and higher data quality. This dataset thus bridges the gap between general reasoning corpora and science-grade problem solving, offering high-difficulty, high-fidelity supervision for post-training.

Structured Problem Formulations

[Theoretical Problem 3: Cavitation (10.0 points)]
[Introduction] Cavitation is the phenomenon of vapour bubbles or "cavities" occurring in a liquid medium due to …
[Potentially useful information] Let us say we have a closed jar containing water and air. If the air is too dry…
[Part B: Main dynamics (6.0 points)] Now we will study the detailed dynamics of a spherical bubble consisting …
B.2 A water tank under the external pressure $p_\infty^- = 101\ kPa$, containing a nucleus of radius $R_0 = 10^{-5}m$ initially in equilibrium, was exposed to vacuum, so that the system suddenly has $p_\infty = 0$. Estimate the terminal (asymptotic) value of the growth speed R', as well as the time it reaches this terminal value.

Expert-Authored Solutions

Putting $R^{"}$ and $R \to \infty$ in $\rho R R^{"} + \frac{3}{2}\rho(R')^2 + \frac{2\sigma}{R} - \frac{q_0 R_0^{3\gamma}}{R^{3\gamma}} = p_v - p_\infty$ (19) yields

$$R'(\infty) = \sqrt{\frac{2(p_v - p_\infty)}{3\rho}} = \sqrt{\frac{2p_v}{3\rho}} \approx 1.24 m/s$$

The initial acceleration can also be found from (19) as

$$R^{"}(0) = \frac{q_0 + p_v - 2\sigma/R_0}{\rho R_0} = \frac{p_\infty^-}{\rho R_0} \approx 10^7 m/s^2$$

and hence the time for the speed to be stabilized can be estimated as

$$t = \frac{R'(\infty)}{R^{"}(0)} \approx 0.1\mu s$$

Rule-Verifiable Answers

Answer: 1.24 **Unit:** m/s
Answer Type: Numerical Value

Answer: 0.1 **Unit:** μs
Answer Type: Numerical Value

Labeled Metadata

Physics Field: Mechanics
Source: APhO 2023

Figure 3 | A data sample from the training data.

As illustrated in Figure 3, each instance in the dataset follows a structured *Question–Solution–Answer* schema, enriched with metadata, providing a well-organized format to support diverse avenues of research.

- **Structured Problem Formulations.** Physical problem statements are preserved in their original form whenever possible. For excessively long problems, subdivisions are introduced to respect model context limits while retaining the logical integrity of the task.
- **Expert-Authored Solutions.** Solution processes are authored by human physics experts, providing authentic reasoning trajectories.
- **Rule-Verifiable Answers.** Verifiable final answers supply the unambiguous correctness criteria required for RLVR. Annotations on type, unit, and scoring points support reliable validation and mirror the weighted criteria of human grading.
- **Labeled Metadata.** Each instance is tagged with physics field and source, enabling analysis of domain coverage, guiding data selection strategies, and providing a basis for studying the impact of provenance on training dynamics.

2.2. Dataset Construction

Data Collection. Our data construction is guided by a central objective: enabling LLMs to internalize structured reasoning that aligns with physical laws and empirical consistency—a prerequisite for genuine scientific intelligence. We therefore prioritize problems that combine reasoning depth with rule-verifiable outcomes—conditions empirically linked to stronger reasoning gains in recent studies (Guo et al., 2025; Wen et al., 2025; Yang et al., 2025). Within physics, these properties converge most naturally in high school Olympiad problems: they demand rigorous modeling, symbolic precision, and multi-step inference, yet remain tractable without domain-specific obscurity. This has been

empirically validated in the PHYSICS (Zheng et al., 2025b), where LLMs perform worse on high school Olympiad problems than on undergraduate non-physics-major questions, highlighting the distinct reasoning challenge posed by Olympiad problems. Accordingly, we assemble two complementary sources. The first comprises ten major physics Olympiads(up to 2023)—including APhO, IPhO and others—spanning regional to international tiers and capturing a naturally graded difficulty spectrum. The second consists of ten authoritative competition textbooks, offering systematically organized examples and exercises with expert-authored solutions.

Data Annotation. Driven both by the demand for high-quality extraction and the heterogeneity of the sources, the construction process is organized as a multi-stage pipeline.

1. **PDF-to-Markdown Conversion.** Source materials in PDF format are parsed into Markdown using Optical Character Recognition (OCR) tools.
2. **Questions and Solutions Parsing.** Extraction strategies are tailored to the two source types. For textbooks, model-assisted parsing leverages structural cues (e.g., chapter boundaries and numbering) to align exercises with their solutions. Olympiad problems, characterized by lengthy statements and multiple sub-questions, are manually restructured by experts to separate shared background from sub-questions, preserving both readability and fidelity.
3. **Answer Annotation.** Answers are automatically extracted by models and decomposed into structured lists, allowing each sub-answer to be individually validated against model outputs. Units are separated into explicit fields to support standardized, rule-based scoring.
4. **Language Normalization.** Problems originating in Chinese (e.g., CPhO) are translated into English with Claude to maintain a consistent monolingual corpus.

Quality Control. To ensure reliability, we implement a multi-stage quality control pipeline, integrating model-based automated procedures and expert review. **(1) OCR Correction.** Texts parsed from complex page layouts or low-quality scans are manually validated against the original PDFs to correct OCR artifacts. **(2) Answer Cross-Validation.** Three models—Gemini-2.5-Flash, Claude-3.7-Sonnet, and GPT-4o—independently extract answers from each question–solution pair; Consensus is established when at least two models agree; items without such agreement are removed. **(3) Data Filtering.** Problems requiring diagram drawing or involving unverifiable answers (e.g., proofs or explanations) are removed. **(4) Expert Review.** Claude-3.7-Sonnet performs a comprehensive consistency audit, followed by targeted manual refinement. After these stages, the dataset contracts from 6,516 to 5,065 items, yielding an English, text-only corpus with rule-verifiable answers, well-suited for RLVR.

3. Approach

3.1. RL Formulation

We formulate the problem of solving physics Olympiad tasks as a reinforcement learning (RL) (Sutton et al., 1998) process. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ denote the underlying Markov Decision Process (MDP), where:

- $\mathcal{S}$ represents the state space, corresponding to the model context, including the problem statement and all previously generated reasoning steps.
- $\mathcal{A}$ is the action space, defined over the token vocabulary from which the model generates its next output token.

- $P(s' \mid s, a)$ is the (deterministic) transition function, which appends the newly generated token a to the state s, resulting in an updated context s'.
- $r(s, a)$ is the reward function, which evaluates the correctness and quality of the final solution trace.

The learning objective is to maximize the expected return:

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} r(s_t, a_t)\right], \tag{1}$$

where π_θ is the policy parameterized by model parameters θ, and $\tau = (s_0, a_0, \ldots, s_T)$ denotes a trajectory sampled from π_θ.

Policy Gradient. The policy gradient (Sutton et al., 1999) method optimizes π_θ by ascending the gradient of the expected return:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, A^\pi(s_t, a_t)\right], \tag{2}$$

where $A^\pi(s_t, a_t)$ is the advantage function estimating the relative value of action a_t in state s_t. We adopt this standard form and instantiate it via GSPO below.

Group Sequence Policy Optimization (GSPO). GSPO (Zheng et al., 2025a) elevates optimization from the token level (Shao et al., 2024; Yu et al., 2025c) to the sequence level, employing length-normalized sequence likelihood importance ratios:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{1/|y_i|} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right), \tag{3}$$

where $|y_i|$ denotes the sequence length, and the $1/|y_i|$ term implements length normalization to reduce variance. The corresponding advantage function is computed at the sequence level:

$$\hat{A}_i^{\text{GSPO}} = R_i - \frac{1}{G}\sum_{j=1}^{G} R_j, \tag{4}$$

with the objective function:

$$J^{\text{GSPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G} \min\left(s_i(\theta)\hat{A}_i^{\text{GSPO}}, \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i^{\text{GSPO}}\right)\right]. \tag{5}$$

3.2. Instantiation

Reward Design. To instantiate the RL algorithms for solving physics problems, we must concretely define the reward function. Following the Correct-or-Not design in RLVR methods (Guo et al., 2025; Zheng et al., 2025a), we employ a binary reward scheme based on answer correctness, leveraging the fact that our physics dataset contains verifiable ground-truth outcomes:

$$r = \begin{cases} 1, & \text{if the predicted answer matches the ground truth,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

However, physics problems often involve multiple sub-questions or require multiple final results (e.g., solving for both a and b). To account for this structure, we adopt a test-case-style reward aggregation similar to program evaluation, defining the final reward as:

$$R = \frac{1}{N} \sum_{i=1}^{N} r_i, \tag{7}$$

where N is the number of required sub-answers in the problem, and r_i denotes the correctness indicator for the i-th sub-answer.

Answer Extraction. We design prompts that enforce a `multi-box` output format, requiring the model to place each sub-answer sequentially inside separate `\boxed` environments, in order to simplify the answer extraction. Specifically, we apply system prompt as shown in Figure 4 into our physics dataset.

System Prompt of Multi-box Style

```
Please answer the problem adhering to the following rules:
1. Please use LaTeX format to represent the variables and formulas
   used in the solution process and results.
2. Please put the final answer(s) in \boxed{}, note that the unit
   of the answer should not be included in \boxed{}.
3. If the problem requires multiple answers, list them in order,
   each in a separate \boxed{}.
```

Figure 4 | System prompt design for P1 training.

Verifier Design. To handle the inherent complexity of physics answers, which often appear as symbolic expressions rather than single numeric values, we adopt a hybrid verification framework that integrates both rule-based and model-based components:

- *Rule-based verifier.* Inspired by DrGRPO (Liu et al., 2025b), we combine symbolic computation with rule-based checks using SymPy (Meurer et al., 2017) and math-verify (Kydlíček) heuristics. This allows robust equivalence testing of algebraic expressions, including commutativity, factorization, and simplification.
- *Model-based verifier (used only in validation).* Complementing the rule-based system, we follow the XVerify (Chen et al., 2025) paradigm and employ a large language model (Qwen3-30B-A3B-Instruct-2507) as an answer-level verifier. Given the problem statement, the extracted model prediction, and the ground truth, the verifier outputs a binary judgment (*correct* or *incorrect*), improving robustness against cases that are challenging for purely symbolic methods.

3.3. Technical Design

3.3.1. Adaptive Learnability Adjustment

Achieving sustained performance growth during post-training is a key challenge for large language models. In practice, RL fine-tuning often faces performance bottlenecks after an initial phase of rapid improvement (Cui et al., 2025b; Yu et al., 2025c). These plateaus can be attributed to entropy

collapse, sparse rewards, limited base model capacity, or imperfect data quality (Cui et al., 2025b; Zhang et al., 2025a). We collectively refer to these factors as a reduction in *learnability*. For example, entropy collapse diminishes exploration capability and thus reduces learnability; similarly, if the model capability and the dataset difficulty are mismatched, the model either fails to learn or trivially solves problems without meaningful updates, again resulting in lowered learnability. To address this issue, especially under the constraints of our relatively small physics dataset, we design adaptive mechanisms that ensure continuous learnability throughout the training process. Specifically, we propose two complementary strategies: *preliminary pass rate filtering* and *adaptive exploration space expansion*.

Preliminary Pass Rate Filtering. Data quality directly impacts learnability, and even after extensive curation, physics datasets may still contain problematic samples, such as tasks relying on unavailable diagrams or incomplete context. To mitigate this issue, we apply a filtering procedure before training, based on pass rate statistics. Concretely, we perform rollouts on the training dataset using the Qwen3-30B-A3B-Thinking model under a pass@88 setting, and exclude tasks that are either too easy (pass rate > 0.7) or too difficult (pass rate = 0). Formally, the retained dataset is:

$$\mathcal{D}_{\text{filtered}} = \{\, q \in \mathcal{D} \mid 0 < \text{pass}(q) \leq 0.7 \,\}. \tag{8}$$

This filtering serves two purposes:

1. **Removing tasks with pass = 0 or pass = 1 prevents zero-learnability cases.** As defined in GSPO (Zheng et al., 2025a), when all group samples share the same outcome, the estimated advantage ($\hat{A}^{\text{GSPO}}$) becomes zero, eliminating effective learning signals. Filtering these tasks mitigates reward sparsity.
2. **Removing overly easy tasks (pass > 0.7) prevents entropy collapse.** According to the entropy mechanism analysis (Cui et al., 2025b), when the majority of tasks can be solved with high confidence, RL updates push the policy toward low-entropy solutions, accelerating collapse. Given that modern base models already possess strong reasoning capabilities, excluding trivial tasks helps maintain diversity and prevents premature convergence.

Adaptive Exploration Space Expansion. Even with curated data, performance bottlenecks can also arise from insufficient exploration (Cui et al., 2025b; Luo et al., 2025). As the model improves, fixed rollout configurations may fail to provide adequate opportunities for exploration, leading to stagnation. To address this, we dynamically expand the exploration space in line with the model's evolving capability, thereby sustaining learnability.

We consider two complementary dimensions of expansion:

1. **Group size expansion.** In the GSPO (Zheng et al., 2025a) group-based advantage estimation framework, each training question q is evaluated with a group of G sampled responses $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)$, and the relative rewards within the group are used to construct standardized advantages. Increasing the group size G enhances the probability of generating at least one high-quality trajectory, especially for difficult problems where success rates are intrinsically low. Larger G thus provides a stronger learning signal by ensuring that rare but informative trajectories are more likely to appear within each group, which in turn amplifies effective gradient updates during training. This mechanism directly alleviates reward sparsity and strengthens the stability of group-based advantage estimation.

2. **Generation window expansion.** The maximum output length (generation window) limits the depth of reasoning that can be expressed. If too short, the model's reasoning chains for complex physics problems are truncated, producing incomplete or incorrect answers. We gradually extend the generation window as training progresses, allowing the model to explore longer and more coherent reasoning chains. This expansion improves solvability for high-complexity problems and reduces truncation-induced errors.

3.3.2. Training Stabilization Mechanism

Mitigate Train-inference Mismatch Recent studies (Liu et al., 2025a; Yao et al., 2025) have noticed that the train-inference engine difference is a key cause to instability in training. To formally understand this statement, we rewrite Eq. 2 as,

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta^{rollout}}\left[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta^{train}(a_t \mid s_t)\, A^\pi(s_t,a_t)\right], \tag{9}$$

where $\pi_\theta^{rollout}$ denote the policy used to generate trajectories during rollout, and π_θ^{train} denote the policy evaluated during gradient computation. Modern RL frameworks (Sheng et al., 2024; Zhu et al., 2025b) often adopt different engines for rollout (e.g. vllm (Kwon et al., 2023) and SGLang (Zheng et al., 2024)) and training (e.g. FSDP (Merry et al., 2021) and Megatron (Shoeybi et al., 2019)). While this design significantly improve throughout, it inadvertently introduce mismatch between $\pi_\theta^{rollout}$ and π_θ^{train} due to differences in numerical precision, computation optimization strategies, and kernel implementations, leading to biased gradient estimates and training instability.

$$\pi_\theta^{rollout}(a|s) \neq \pi_\theta^{train}(a|s) \tag{10}$$

To mitigate this mismatch and stablize training, we adopt Truncated Importance Sampling (TIS) (Yao et al., 2025), which applies importance weighting to rebalance gradients computed under the training policy π_θ^{train} using trajectories sampled from the rollout policy $\pi_\theta^{rollout}$.

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta^{rollout}}\left[\sum_{t=0}^{T}\min\left(\frac{\pi_\theta^{train}(a_t|s_t)}{\pi_\theta^{rollout}(a_t|s_t)}, C\right)\nabla_\theta \log \pi_\theta(a_t|s_t)A^\pi(s_t,a_t)\right], \tag{11}$$

where C is a truncation hyperparameter that controls the variance of the importance weights. The truncation operator $\min(\cdot, C)$ prevents excessively large weights that could destabilize training, while still correcting for the distributional shift.

3.4. Training Dynamics

Implementation. For implementation of P1 training pipeline, we adopt the slime (Zhu et al., 2025b) framework, which is a efficient LLMs post-training framework connecting Megatron with SGLang. As for base model, we use `Qwen3-30B-A3B-Thinking` and `Qwen3-235B-A22B-Thinking` as starting points of our `P1-30B-A3B` and `P1-235B-A22B`. To adaptively adjust aforementioned learnability during training, we periodically resume training from previous checkpoint with updated configurations.

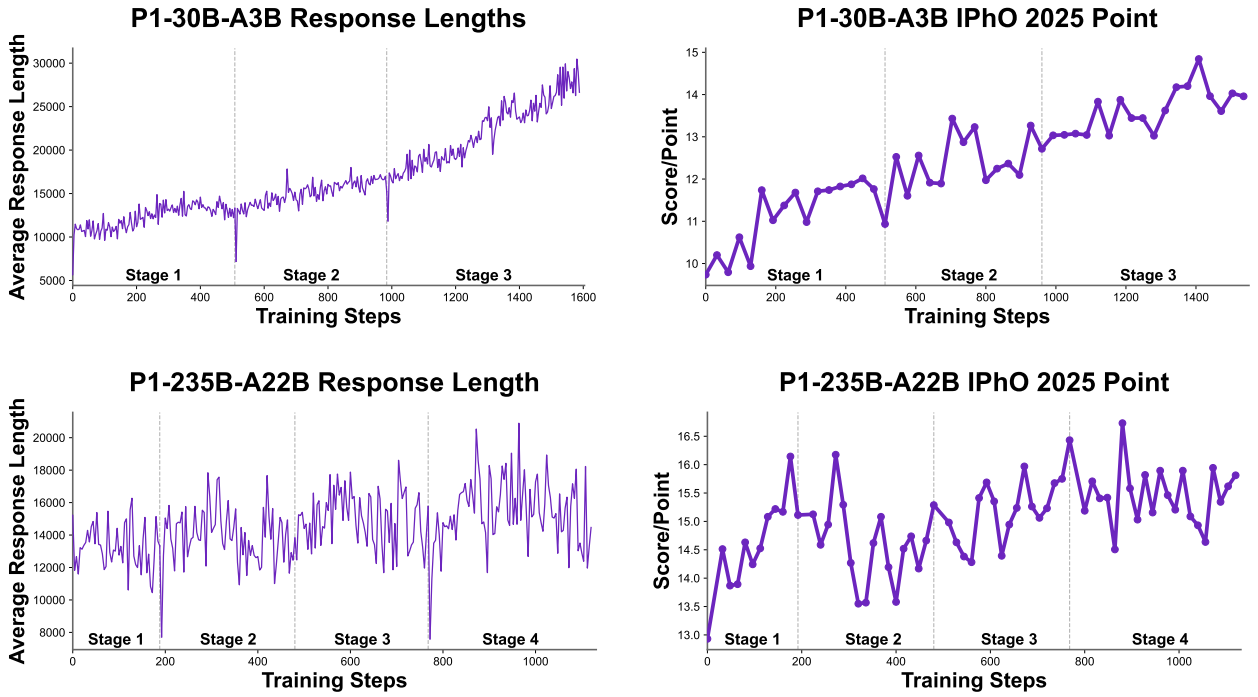

Figure 5 | Training dynamics of P1 models. The upper side shows dynamics of P1-30B and the lower side shows the P1-235B variant. The LHS present the average response length on train dataset during RL training process. The RHS present the evaluation results on IPhO 2025 on model checkpoints during training, please note that evaluation here leverage rule-based verifier and model-based verifier (Qwen3-30B-Instruct-2507).

Settings of Different Phrases. We report the settings of different stages during the post training process of P1 models. As shown in the table, we gradually expand the exploration space by increasing group size and generation window. We also apply the train-inference correction technique during the whole multi-stage process. Notely, we build up our algorithm upon GSPO to stabilize the training of MoE models. Besides, we only enable the rule-based verifier during training, the reason for which will be discussed in Section 5.2.

Table 2 | Configuration of different phrases in P1 training.

Model	**P1-30B-A3B**			**P1-235B-A22B**			
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3	Stage 4
Group size	16	32	32	16	32	32	32
Generation Window	48k	48k	64k	48k	48k	64k	80k
Lr	1×10^{-6}			1×10^{-6}			
Algorithm	GSPO w. tis			GSPO w. tis			
Verifier Choice	Rule-based Verifier Only			Rule-based Verifier Only			
Rollout Batch Size	2048			1024			
Update Batch Size	512			256			

Analysis. As shown in Figure 5, we present training dynamics of multi-stages RL training. It can be observed that the response length of P1 gradually increases during training, revealing the model's ability to reason and solve complex problems getting more deeply and requires a larger exploration

space to explore. Therefore, we gradually expand the exploration space by increasing generation window and group size by switching to different stages. Furthermore, as for validation results, we can observe that the P1 models experience an steady improvement on IPhO 2025 during training, revealing the effectiveness and stability of our training algorithm.

3.5. Agentic Augmentation

During inference, we scale up test-time effort by incorporating multi-agent framework. We directly apply P1 model into `PhysicsMinions` (Yu et al., 2025b), which is a latest agentic framework designed for complex physics reasoning. `PhysicsMinions` consists of three coevolutionary studios: the *Visual Studio*, the *Logic Studio*, and the *Review Studio*. Given a multimodal problem with diagrams or plots, the Visual Studio first observes, validates, and reflects on the input to extract structured information, which is then passed to the Logic Studio. In the Logic Studio, a solver generates an initial solution and an introspector refines it through self-improvement before passing it on. The Review Studio then applies dual-stage verification: the Physics-Verifier checks physical consistency (e.g., constants and units), while the General-Verifier conducts more detailed inspections of logic, reasoning, and calculations.

If either verification stage fails, a detailed bug report is returned to the Logic Studio, where the introspector revises the solution and resubmits it to Review Studio for verification. This process repeats until the solution passes a predefined number of consecutive verifications (CV), which is the only hyperparameter in the system. A solution that passes CV checks consecutively is accepted as the final solution; if it fails CV times consecutively, the solver regenerates a new candidate solution. This collaborative critique-and-refine cycle defines the system's coevolutionary process, with CV set to 2 by default. It's noteworthy that because P1 models are text-only LLMs, we disable the Visual Studio for adaptivity. In simple words, we instantiate the solver in the Logic Studio as well as the dual verifiers in the Review Studio using P1 models within `PhysicsMinions`.

4. Experiment

4.1. Experimental Setup

Test Dataset. To evaluate performance on challenging physics Olympiads, we constructed the HiPhO benchmark (Yu et al., 2025a), the first High School Physics Olympiad benchmark covering the 13 most recent high school physics Olympiads from 2024 to 2025. These competitions range from international to regional levels and span 7 major types: IPhO, APhO, EuPhO, NBPhO, PanPhO, PanMechanics, and F=MA. The selection was based on both their global influence and the availability of human score distributions[1].

Comparison Models. We compared against 33 representative models, including 11 closed-source and 22 open-source models, selected to reflect the current frontier of large language models:

- **Closed-source models:** GPT-5 (OpenAI, b), o3 (OpenAI, c), o4-mini (high) (OpenAI, c), o4-mini (OpenAI, c), GPT-4o (OpenAI, a), Gemini-2.5-Pro (Comanici et al., 2025), Gemini-2.5-Flash-Thinking (Comanici et al., 2025), Grok-4 (xAI), Claude-4-Sonnet-Thinking (Anthropic), Claude-4-Sonnet (Anthropic), Mistral-Medium-3 (Mistral).
- **Open-source models:** Intern-S1 (Bai et al., 2025a), InternVL3 Series (Zhu et al., 2025a), Qwen2.5-VL Series (Bai et al., 2025b), GLM-4.5V (Team et al., 2025a), DeepSeek-VL2 (Wu et al., 2024), LLaMA4-Scout-17B (Meta), Phi-4-multimodal (Abouelenin et al., 2025), GPT-OSS-120B (Agarwal

[1] CPhO, USAPhO, and APhO-2024 were excluded due to the lack of complete official contestant scores.

Table 3 | Evaluation results on the HiPhO benchmark (13 physics Olympiads from 2024–2025) using the *exam score* metric. Gold, Silver and Bronze indicate scores above the respective thresholds. Models are ranked by medal counts; **bold** is the highest score, and underline is the second highest. Here, only the theoretical parts of exams are used, hence Full Mark (Model) ≤ Full Mark (Human).

Physics Olympiad	IPhO		APhO	EuPhO		NBPhO		PanPhO		PanMechanics		F=MA		Avg.	Medal Table		
Year	2025	2024	2025	2025	2024	2025	2024	2025	2024	2025	2024	2025	2024				
Full Mark (Human)	30.0	30.0	30.0	30.0	30.0	72.0	72.0	100.0	100.0	100.0	100.0	25.0	25.0	57.2			
Full Mark (Model)	29.4	29.3	30.0	29.0	28.0	43.5	50.0	100.0	98.0	100.0	100.0	25.0	25.0	52.9			
Top-1 Score (Human)	29.2	29.4	30.0	27.0	30.0	53.2	40.8	81.0	66.5	62.0	51.0	25.0	24.0	42.2			
Top-1 Score (Compared Models)	22.7	25.9	27.9	14.9	23.4	34.1	35.9	60.3	75.4	72.1	79.0	22.8	22.4	39.8			
Gold Medal	19.7	20.8	23.3	16.5	20.4	28.6	26.5	41.5	52.0	52.0	51.0	15.0	14.0	29.3			
Silver Medal	12.1	11.1	18.7	9.8	14.2	20.1	19.4	28.5	37.5	36.0	26.0	11.0	12.0	19.7			
Bronze Medal	7.2	3.6	13.1	5.8	8.9	15.2	13.5	14.5	16.0	20.0	12.0	9.0	10.0	11.4			
P1-235B-A22B + PhysicsMinions	**23.2**	25.2	**28.0**	12.4	**23.5**	31.9	35.4	57.7	67.0	**77.5**	74.8	21.5	20.5	**38.4**	12	1	0
Gemini-2.5-Pro	22.7	**25.9**	27.9	**14.9**	21.8	32.3	**35.9**	**60.3**	64.1	69.5	70.2	**22.8**	22.0	37.7	12	1	0
P1-235B-A22B	21.2	24.7	27.4	10.8	23.0	31.8	28.4	54.7	56.7	74.7	72.9	20.9	19.4	35.9	12	1	0
Gemini-2.5-Flash-Thinking	20.2	23.9	27.4	13.2	21.9	29.0	29.3	44.6	54.9	60.5	55.9	17.8	19.1	32.1	12	1	0
GPT-5	22.3	20.2	27.0	10.3	21.7	32.9	32.8	55.9	69.8	69.4	**79.0**	22.4	**22.4**	37.4	11	2	0
o3	15.7	23.7	25.9	11.4	21.6	**34.1**	33.5	47.3	55.9	71.4	75.6	22.0	20.6	35.3	11	2	0
Grok-4	18.7	23.5	25.0	11.5	20.5	25.8	29.3	45.0	**75.4**	72.1	78.6	19.8	19.8	35.8	10	3	0
Qwen3-235B-A22B-Thinking-2507	17.1	23.0	26.2	10.9	20.4	33.6	28.1	44.7	51.8	69.1	72.9	18.5	18.9	33.5	10	3	0
DeepSeek-R1	18.5	24.6	25.4	10.8	21.4	26.3	20.5	42.2	47.4	65.4	72.5	18.3	18.5	31.7	8	5	0
Claude-4-Sonnet-Thinking	19.0	22.0	24.8	9.7	20.5	28.1	25.6	43.1	39.3	57.4	61.8	19.2	20.1	30.0	8	4	1
P1-30B-A3B	18.5	22.3	25.4	7.4	18.8	29.2	24.0	47.0	51.4	69.5	69.1	19.6	20.0	32.5	8	4	1
Qwen3-235B-A22B	17.8	23.8	26.0	9.2	21.5	28.4	31.1	42.3	49.1	63.1	44.6	18.4	17.4	30.2	8	4	1
Qwen3-32B	15.7	19.3	23.9	9.8	21.2	28.9	24.1	36.6	41.8	67.0	59.2	18.9	16.6	29.5	7	6	0
o4-mini	15.4	22.9	22.8	10.1	20.9	26.9	27.3	39.4	47.1	64.2	62.5	18.6	18.5	30.5	7	6	0
o4-mini (high)	16.0	23.7	22.9	12.0	20.1	27.4	29.8	41.4	50.9	69.1	67.3	18.6	18.8	32.2	6	7	0
Qwen3-30B-A3B-Thinking-2507	15.6	19.7	23.5	7.4	16.4	28.6	22.2	40.5	43.8	67.7	66.5	18.3	18.0	29.9	6	6	1
Kimi-K2-Thinking*	19.1	22.8	22.4	5.2	20.9	26.0	20.6	52.5	51.3	47.3	64.8	20.6	19.7	30.2	6	6	0
Kimi-K2-Instruct	16.5	19.8	24.2	11.0	16.9	26.5	26.2	35.9	41.8	65.9	58.9	16.0	18.2	29.1	5	8	0
GPT-OSS-120B	16.9	21.4	22.8	9.1	19.9	26.0	25.8	37.4	41.8	57.1	59.7	17.8	17.6	28.7	5	7	1
Intern-S1	15.9	14.2	21.7	9.0	16.6	23.0	20.5	41.1	50.3	60.4	57.4	18.4	19.5	28.3	4	8	1
Qwen3-30B-A3B	13.6	15.4	22.7	9.8	16.5	24.7	21.7	31.9	39.5	49.9	45.0	15.5	15.0	24.7	2	11	0
Claude-4-Sonnet	15.7	19.2	22.8	9.5	16.5	27.5	21.3	40.4	43.3	46.5	48.5	16.8	16.5	26.5	2	10	1
DeepSeek-V3	13.6	16.4	22.1	7.1	17.2	21.1	17.3	37.2	35.0	48.4	46.5	14.1	15.6	24.0	1	9	3
Mistral-Medium-3	14.2	14.1	19.9	8.5	12.2	20.4	19.6	30.8	28.6	32.9	36.1	13.9	14.1	20.4	1	8	4
GPT-4o	10.2	9.4	15.1	6.8	9.2	16.4	11.7	27.8	22.8	28.2	26.5	15.0	10.9	16.2	1	1	10
InternVL3-78B-Instruct	12.9	12.5	17.7	7.5	15.2	22.3	22.5	26.2	27.4	21.1	27.1	12.0	13.0	18.3	0	8	5
GLM-4.5V	11.9	4.4	16.2	8.7	14.1	19.5	14.0	18.5	16.0	47.8	39.0	13.0	13.8	18.2	0	4	9
Qwen3-8B	10.6	12.7	11.5	7.1	11.9	20.1	17.3	26.3	22.3	21.8	22.8	10.8	10.0	15.8	0	2	10
Qwen2.5-VL-72B-Instruct	10.6	7.2	13.6	6.1	8.1	13.3	11.4	26.8	18.2	24.0	28.5	13.5	9.8	14.7	0	2	7
LLaMA4-Scout-17B	9.7	9.5	13.1	5.4	10.4	22.8	12.8	26.6	24.1	35.4	34.5	8.1	6.4	16.8	0	2	7
Qwen2.5-VL-32B-Instruct	9.9	8.2	16.5	6.9	10.0	15.3	14.4	22.5	22.4	28.1	29.9	7.6	4.6	15.1	0	1	10
InternVL3-38B-Instruct	8.9	7.8	12.3	6.1	8.3	14.0	10.6	24.1	20.4	27.5	24.8	8.2	6.8	13.8	0	0	7
InternVL3-9B-Instruct	4.7	3.7	7.2	4.2	4.1	9.4	6.0	12.4	11.0	11.6	16.3	6.4	6.2	7.9	0	0	2
Qwen2.5-VL-7B-Instruct	3.5	2.5	5.7	4.4	3.6	7.3	5.5	14.7	7.6	9.8	11.5	4.4	3.5	6.5	0	0	1
Phi-4-multimodal	2.0	1.6	4.2	3.6	3.6	5.0	4.5	8.3	9.0	10.0	10.1	4.4	5.0	5.5	0	0	0
DeepSeek-VL2	1.8	0.5	2.5	3.4	3.4	5.0	3.2	5.6	4.8	7.3	6.4	5.0	3.9	4.0	0	0	0

* For `Kimi-K2-Thinking`, the inference timeout is set to 2 hours, and 9.58% of cases exceed this limit.

et al., 2025), Kimi-K2-Thinking (Team et al., 2025b), Kimi-K2-Instruct (Team et al., 2025b), DeepSeek-R1 (Guo et al., 2025), DeepSeek-V3 (Liu et al., 2024), Qwen3 Series (Yang et al., 2025).

Model Configurations. All models were evaluated under a standardized inference setup following the HiPhO protocol. The temperature was fixed at 0.6, and the maximum token limit was set as large as permitted by each model. For every problem, we conducted 8 independent inference runs and computed the average score per problem. These averages were then aggregated across problems to yield the final exam score for each Olympiad.

Evaluation Method. We follow the evaluation protocol of the HiPhO benchmark (Yu et al., 2025a) and use `Gemini-2.5-Flash` as the grader. The benchmark integrates both *answer-level* and *step-level* grading based on official marking schemes, allowing partial credit for correct intermediate reasoning. For each problem, the model's final score is defined as the maximum of its answer-level and step-level scores, consistent with human grading: a correct final answer receives full credit, while an incorrect

one can still earn partial points for valid intermediate steps. The overall exam score is obtained by summing the scores across all problems. Unlike conventional benchmarks that report accuracy, we use the *exam score* as the evaluation metric, enabling direct comparison between model performance and official medal thresholds.

4.2. Evaluation on Physics Olympiads

Excellent Single Model Performance. Evaluation results are presented in Table 3. As noted above, the P1 model series is trained purely through reinforcement learning. We first demonstrate the excellent single-model performance to evaluate the effectiveness of our training strategy.

- **P1-235B-A22B** ranks alongside Gemini-2.5-Pro and Gemini-2.5-Flash-Thinking at the top of the medal table, earning 12 gold and 1 silver medal, and surpassing major closed-source models such as GPT-5 (11 gold), Grok-4 (10 gold), and Claude-4-Sonnet-Thinking (8 gold). Impressively, P1-235B-A22B scores 21.2 / 30 at the latest International Physics Olympiad (IPhO 2025), ranks Top 3 globally—behind only Gemini-2.5-Pro (22.7) and GPT-5 (22.3)—becoming the first and only open-source model to achieve gold-medal performance on IPhO 2025.
- **P1-30B-A3B** earned 8 gold, 4 silver, and 1 bronze medal, ranking third among existing open-source models—just behind the much larger Qwen3-235B-A22B-Thinking-2507 and DeepSeek-R1. It surpasses comparable models such as Qwen3-32B (7 gold, 6 silver) and Qwen3-30B-A3B-Thinking-2507 (6 gold, 6 silver, 1 bronze), demonstrating strong performance-to-scale efficiency.

Agentic Boost. With the combination with the PhysicsMinions system, P1's average performance improves from 35.9 to 38.4, reaching the overall Top-1 position across all models and outperforming leading closed-source models such as Gemini-2.5-Pro (37.7) and GPT-5 (37.4). P1-235B-A22B + PhysicsMinions also achieved new state-of-the-art results across four physics Olympiads. The joint configuration outperformed the best compared models on IPhO 2025 (23.2 vs. 22.7), slightly surpassed the record on APhO 2025 (28.0 vs. 27.9) and EuPhO 2024 (23.5 vs. 23.4), and secured a decisive advantage on PanMechanics 2025 (77.5 vs. 72.1). These results highlight how the integration of P1 with the multi-agent PhysicsMinions framework substantially enhances reasoning and problem-solving performance, showcasing the "model + system" paradigm for complex scientific reasoning.

Gold-Level Performance on CPhO 2025. We further evaluate `P1-235B-A22B` on the newly held 2025 Chinese Physics Olympiad (CPhO 2025), one of the most challenging physics Olympiads worldwide, renowned for its long multi-step reasoning problems. On the theoretical exam, `P1-235B-A22B` obtains a score of 227 / 320, assessed by human experts strictly following the official marking scheme. This score substantially exceeds the top-1 human medalist's 199, as shown in Table 4. This demonstrates that our open-source model can attain gold-medal performance on CPhO 2025, marking an important milestone for open-source physics reasoning and showing that it can already match and even exceed elite human performance on some of the most challenging physics Olympiads.

Table 4 | Performance comparison between P1-235B-A22B and top-1 human medalist on CPhO 2025.

Theoretical Problem	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total Score
Full Mark	45	45	45	45	45	50	45	320
Top-1 Human Medalist	43	14	32	26	40	29	15	199
P1-235B-A22B	35	25	39	21	31	39	37	227

5. Discussion

5.1. Generalizability of P1

We conduct post-training on a specialized dataset to enhance physics problem-solving abilities. Beyond domain-specific improvements, we further investigate: *Does the P1 series model preserve, or even enhance, its general reasoning ability across mathematics, STEM, and coding domains?* To examine this, we compare the P1 series models with their respective base models across diverse benchmarks: six mathematics datasets (AIME24, AIME25, HMMT (Balunović et al., 2025), IMO-AnswerBench (Luong et al., 2025), AMOBench (An et al., 2025), BeyondAIME (ByteDance-Seed, 2025)), two STEM-oriented evaluations (GPQA (Rein et al., 2024), HLE (Phan et al., 2025)), one coding benchmark (LiveCodeBench (Jain et al., 2024)), and a general reasoning task (LiveBench (White et al., 2024)). Figure 6 summarizes the results.

Notably, the P1 models demonstrate consistent advantages over their base counterparts: P1-30B-A3B outperforms Qwen3-30B-A3B-Thinking-2507 across all metrics, while P1-235B-A22B achieves superior performance to Qwen3-235B-A22B-Thinking-2507 in most categories (AIME24, AIME25, GPQA, HLE, IMO-AnswerBench, AMOBench, and overall average). This pattern highlights that P1 models not only maintain but enhance general reasoning abilities beyond the target domain—even on more challenging mathematics benchmarks (e.g., IMO-AnswerBench, AMOBench) that test advanced problem-solving skills.

These results suggest that domain-focused post-training can induce transferable improvements in general reasoning. We hypothesize two contributing factors. First, the optimization process refines reasoning trajectories in a way that transcends domain boundaries, enabling strategies beneficial for mathematics (including advanced subsets like IMO-style problems), STEM, and coding tasks to emerge. Second, the training dataset, while specialized, shares structural similarities with other domains—such as rigorous symbolic manipulation, multi-step logical deduction, and abstract problem modeling—that underpin general reasoning. Thus, the P1 models generalize effectively to neighboring domains and advanced sub-domains of mathematics, providing evidence that domain-focused post-training can simultaneously act as a general reasoning amplifier.

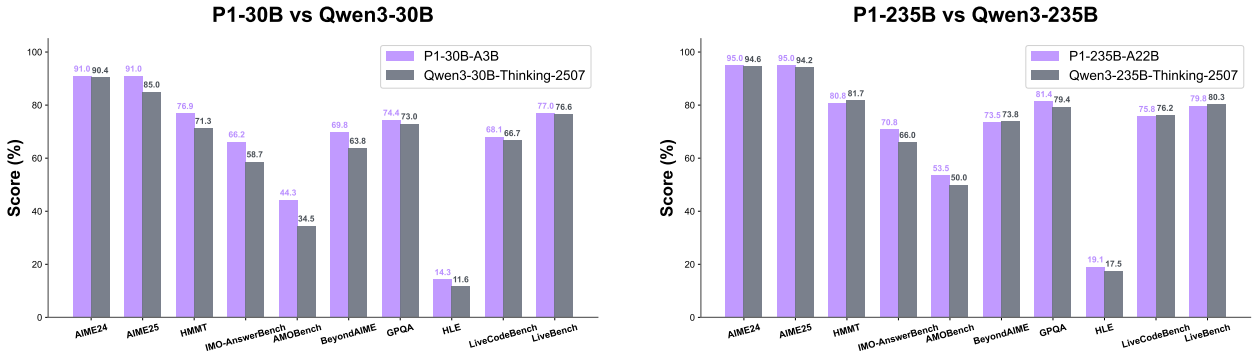

Figure 6 | Performance comparison between P1 models and base models on math and code datasets. P1 models present great general reasoning ability across mathematics, STEM, and coding domains.

5.2. Rule-based vs Model-based Verifier for Training

As stated in Section 3.4, due to the difficulty of implementing a comprehensive and fully correct verification mechanism purely based on rules, we design a hybrid verifier that integrates both rule-

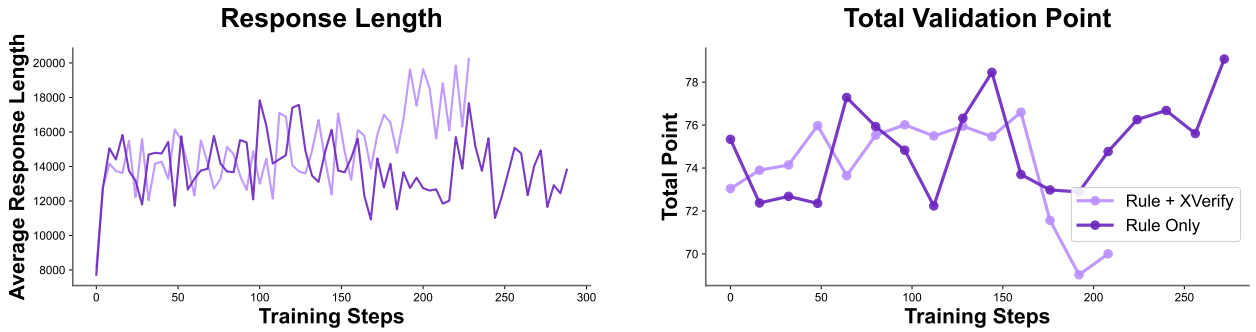

Figure 7 | Training dynamics differences when using an XVerify-like judge during training (P1-235B-A22B). The LHS shows the average response length on the training dataset. The RHS shows the total validation points across 5 representative competitions. We can observe that the inclusion of the XVerify judge even exerts negative effects on the post-training process, as indicated by a consistent increase in response length alongside degraded validation performance.

based and model-based reasoning to achieve broader coverage. However, we found that applying a model-based verifier directly in the post-training process can be risky. Figure 7 compares the training dynamics with and without a model-based verifier. We observe that the variant employing the model-based verifier exhibits an explosive increase in response length, yet its validation performance deteriorates compared to the setup using only a rule-based verifier.

We attribute this phenomenon to two primary causes. (1) The model-based verifier is susceptible to being *hacked* by the policy model. Since the verifier itself is a language model, it may develop unintended biases, favoring atypical response patterns such as overly verbose or stylistically peculiar answers, which can be exploited by the policy model to obtain artificially high rewards. (2) The negative impact of *false positives* (i.e., incorrectly rewarding wrong answers) is substantially more detrimental than that of *false negatives* (i.e., missing some correct responses). In our observations, while the rule-based verifier fails to recognize certain valid answers, the model-based verifier expands the recognition scope to include more potentially correct samples—but at the cost of introducing incorrect judgments.

This trade-off can be understood in terms of **verification precision** and **verification recall**. The rule-based verifier typically offers high precision but limited recall, while the model-based verifier increases recall at the expense of precision. During reinforcement learning, this imbalance can easily destabilize optimization: a few high-reward false positives can dominate the learning signal, leading the policy toward degenerate solution patterns.

We emphasize that raising this issue is not to deny the value of model-based verifiers. Both verification precision and recall are critical for successful post-training. The key lies in ensuring sufficiently high precision before attempting to expand recall. Only when the model-based verifier achieves a stable and reliable judgment boundary can its inclusion provide net positive benefits. Therefore, this phenomenon highlights an urgent need for developing *more robust and calibrated model-based verifiers*, capable of maintaining both correctness and coverage under the dynamics of RL-based post-training.

6. Conclusion

We release the first family of open-source models capable of mastering Olympiad-level physics problems, P1. P1 models achieve great performance on physics reasoning tasks, even reaching goal

medal performance on the latest Olympiad competitions. These achievements are made possible by the combination of train-time scaling via RL post-training and test-time scaling via agentic framework. Besides physics, P1 models also present great performance on other reasoning tasks like math and coding, showing the great generalibility of P1 series. Looking forward, P1's success in mastering Olympiad-level physics represents a key milestone toward LLMs that can assist or pioneer real-world physics research: if models can rigorously solve well-defined problems grounded in natural laws, they may eventually contribute to exploring uncharted scientific frontiers.

7. Acknowlegement

This work is supported by Shanghai AI Laboratory. We would like to extend our special thanks to the developers and maintainers of the following open-source projects, which have been critical to the implementation of this work. This includes Qwen3 (Yang et al., 2025), which provided the foundational base models for our research; slime (Zhu et al., 2025b), whose innovative framework enabled efficient reinforcement learning in our training pipeline; and verl (Sheng et al., 2024), which offered a versatile reinforcement learning framework to support model training. We also thank sglang (Zheng et al., 2024) for its efficient infrastructure for LLM serving and inference, and Megatron-LM (Shoeybi et al., 2019) for providing the large-scale model training framework.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Shengnan An, Xunliang Cai, Xuezhi Cao, Xiaoyu Li, Yehao Lin, Junlin Liu, Xinxuan Lv, Dan Ma, Xuanlin Wang, Ziwen Wang, and Shuang Zhou. Amo-bench: Large language models still struggle in high school math competitions, 2025. URL https://arxiv.org/abs/2510.26768.

Anthropic. Claude 3.7 sonnet system card. URL https://www.anthropic.com/news/visible-extended-thinking.

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, Yu Cheng, Yu Cheng, Pei Chu, Tao Chu, Erfei Cui, Ganqu Cui, Long Cui, Ziyun Cui, Nianchen Deng, Ning Ding, Nanqin Dong, Peijie Dong, Shihan Dou, Sinan Du, Haodong Duan, Caihua Fan, Ben Gao, Changjiang Gao, Jianfei Gao, Songyang Gao, Yang Gao, Zhangwei Gao, Jiaye Ge, Qiming Ge, Lixin Gu, Yuzhe Gu, Aijia Guo, Qipeng Guo, Xu Guo, Conghui He, Junjun He, Yili Hong, Siyuan Hou, Caiyu Hu, Hanglei Hu, Jucheng Hu, Ming Hu, Zhouqi Hua, Haian Huang, Junhao Huang, Xu Huang, Zixian Huang, Zhe Jiang, Lingkai Kong, Linyang Li, Peiji Li, Pengze Li, Shuaibin Li, Tianbin Li, Wei Li, Yuqiang Li, Dahua Lin, Junyao Lin, Tianyi Lin, Zhishan Lin, Hongwei Liu, Jiangning Liu, Jiyao Liu, Junnan Liu, Kai Liu, Kaiwen Liu, Kuikun Liu, Shichun Liu, Shudong Liu, Wei Liu, Xinyao Liu, Yuhong Liu, Zhan Liu, Yinquan Lu, Haijun Lv, Hongxia Lv, Huijie Lv, Qidang Lv, Ying Lv, Chengqi Lyu, Chenglong Ma, Jianpeng Ma, Ren Ma, Runmin Ma, Runyuan Ma, Xinzhu Ma, Yichuan Ma, Zihan Ma, Sixuan Mi, Junzhi Ning, Wenchang Ning, Xinle Pang, Jiahui Peng, Runyu Peng, Yu Qiao, Jiantao Qiu, Xiaoye Qu, Yuan Qu, Yuchen

Ren, Fukai Shang, Wenqi Shao, Junhao Shen, Shuaike Shen, Chunfeng Song, Demin Song, Diping Song, Chenlin Su, Weijie Su, Weigao Sun, Yu Sun, Qian Tan, Cheng Tang, Huanze Tang, Kexian Tang, Shixiang Tang, Jian Tong, Aoran Wang, Bin Wang, Dong Wang, Lintao Wang, Rui Wang, Weiyun Wang, Wenhai Wang, Yi Wang, Ziyi Wang, Ling-I Wu, Wen Wu, Yue Wu, Zijian Wu, Linchen Xiao, Shuhao Xing, Chao Xu, Huihui Xu, Jun Xu, Ruiliang Xu, Wanghan Xu, GanLin Yang, Yuming Yang, Haochen Ye, Jin Ye, Shenglong Ye, Jia Yu, Jiashuo Yu, Jing Yu, Fei Yuan, Bo Zhang, Chao Zhang, Chen Zhang, Hongjie Zhang, Jin Zhang, Qiaosheng Zhang, Qiuyinzhe Zhang, Songyang Zhang, Taolin Zhang, Wenlong Zhang, Wenwei Zhang, Yechen Zhang, Ziyang Zhang, Haiteng Zhao, Qian Zhao, Xiangyu Zhao, Xiangyu Zhao, Bowen Zhou, Dongzhan Zhou, Peiheng Zhou, Yuhao Zhou, Yunhua Zhou, Dongsheng Zhu, Lin Zhu, and Yicheng Zou. Intern-s1: A scientific multimodal foundation model, 2025a. URL https://arxiv.org/abs/2508.15763.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Ding Chen, Qingchen Yu, Pengyuan Wang, Wentao Zhang, Bo Tang, Feiyu Xiong, Xinchi Li, Minchuan Yang, and Zhiyu Li. xverify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025b.

Elizabeth Gibney. Deepmind unveils 'spectacular'general-purpose science ai. *Nature*, 641(8064): 827–828, 2025.

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

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model

serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Hynek Kydlíček. Math-Verify: Math Verification Library. URL https://github.com/huggingface/math-verify.

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

Jiacai Liu, Yingru Li, Yuqian Fu, Jiawei Wang, Qian Liu, and Yu Shen. When speed kills stability: Demystifying rl collapse from the training-inference mismatch. *Notion Blog*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005b 2025. Notion Blog.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Clara Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu H. Trinh, Quoc V. Le, and Junehyuk Jung. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1794/.

Andrew Merry, Samyam Rajbhandari, Mohammad Shoeybi, Ramesh Puri, Paul Fung, Anima Anandkumar, and Bryan Catanzaro. Fully sharded data parallel: Reducing memory usage for large model training. *PyTorch Developer Blog*, 2021. URL https://pytorch.org/blog/introducing-pytorch-fully-sharded-data-parallel-api/.

Meta. Llama 4 system card. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/.

Aaron Meurer, Christopher P Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K Moore, Sartaj Singh, et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.

Mistral. Mistral-medium-3 system card. URL https://mistral.ai/news/mistral-medium-3.

Junhyuk Oh, Greg Farquhar, Iurii Kemaev, Dan A Calian, Matteo Hessel, Luisa Zintgraf, Satinder Singh, Hado van Hasselt, and David Silver. Discovering state-of-the-art reinforcement learning algorithms. *Nature*, pages 1–2, 2025.

OpenAI. Gpt-4o system card, a. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. Gpt-5 system card, b. URL https://openai.com/index/gpt-5-system-card/.

OpenAI. Openai o3 and o4-mini system card, c. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Jiahao Qiu, Jingzhe Shi, Xinzhe Juan, Zelin Zhao, Jiayi Geng, Shilong Liu, Hongru Wang, Sanfeng Wu, and Mengdi Wang. Physics supernova: Ai agent matches elite gold medalists at ipho 2025. *arXiv preprint arXiv:2509.01659*, 2025.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025a. URL https://arxiv.org/abs/2507.01006.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025b.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *International Conference on Representation Learning*, 2020.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4, 2024.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

xAI. Grok 4 system card. URL https://x.ai/grok.

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

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient rl framework secretly brings you off-policy rl training, August 2025. URL https://fengyao.notion.site/off-policy-rl.

Fangchen Yu, Haiyuan Wan, Qianjia Cheng, Yuchen Zhang, Jiacheng Chen, Fujun Han, Yulun Wu, Junchi Yao, Ruilizhen Hu, Ning Ding, Yu Cheng, Tao Chen, Lei Bai, Dongzhan Zhou, Yun Luo, Ganqu Cui, and Peng Ye. Hipho: How far are (m)llms from humans in the latest high school physics olympiad benchmark? *arXiv preprint arXiv:2509.07894*, 2025a.

Fangchen Yu, Junchi Yao, Ziyi Wang, Haiyuan Wan, Youling Huang, Bo Zhang, Shuyue Hu, Dongzhan Zhou, Ning Ding, Ganqu Cui, Lei Bai, Wanli Ouyang, and Peng Ye. Physicsminions: Winning gold medals in the latest physics olympiads with a coevolutionary multimodal multi-agent system, 2025b. URL https://arxiv.org/abs/2509.24855.

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

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Advances in neural information processing systems*, 2025b.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*, 2024.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025a.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025b.

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

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

Shenghe Zheng, Qianjia Cheng, Junchi Yao, Mengsong Wu, Haonan He, Ning Ding, Yu Cheng, Shuyue Hu, Lei Bai, Dongzhan Zhou, et al. Scaling physical reasoning with the physics dataset. *arXiv preprint arXiv:2506.00022*, 2025b.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pages 1–11, 2025c.

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

Zilin Zhu, Chengxing Xie, Xin Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. https://github.com/THUDM/slime, 2025b. GitHub repository. Corresponding author: Xin Lv.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *Conference on Neural Information Processing Systems*, 2025.

A. Appendix

A.1. Test-time Reinforcement Learning

Test-time training (TTT) (Sun et al., 2020; Wang et al., 2020) has received increasing research interest. These approaches adapt model parameters at test time by exploiting the structure and distributional properties of incoming test data. Recent works explore self-supervised RL with various internal reward signals including majority voting (Zuo et al., 2025), semantic coherence (Zhang et al., 2025b), token-level confidence (Zhao et al., 2025b) and self-play (Zhao et al., 2025a). By updating models at test time using RL, test-time reinforcement learning (TTRL) (Zuo et al., 2025) emerges as a promising way to enables the model to explore and improve its performance on unlabeled test set without explicit supervision. In this work, we adopt TTRL with P1-30B-A3B for further improvement due to its representativeness and effectiveness. Specifically, we merge all unlabeled test data of HiPhO, sampling 32 responses per sample. The majority voting consensus (cons@32) is then used as pseudo-labels to fine-tune the P1-30B-A3B model checkpoint for 64 steps with GSPO. We employ a batch size of 256, a mini-batch size of 32, and a maximum response length of 48K. Before training, we filter out multiple-choice questions via regular expression matching. In our preliminary experiments, we observed that multiple-choice questions tend to inflate the majority consensus ratio, which often leads to reward hacking. The validation results (mean@16) are shown in Table 5.

Table 5 | Performance of our P1-30B-A3B model combined with TTRL on the HiPhO benchmark (note that results presented here are evaluated only in answer level, leveraging Qwen3-30B-A3B-Instruct as judge).

Physics Olympiad	**IPhO**		**APhO**	**EuPhO**		**NBPhO**		**PanPhO**		**PanMechanics**		**F=MA**		**Avg.**
Year	2025	2024	2025	2025	2024	2025	2024	2025	2024	2025	2024	2025	2024	
P1-30B-A3B	12.0	14.4	23.1	0.5	12.1	24.5	13.8	41.7	49.5	58.1	73.1	17.9	17.8	27.6
P1-30B-A3B + TTRL	13.8	15.4	23.6	0.3	14.1	24.4	12.8	42.9	50.4	61.1	75.2	18.0	17.8	28.4

A.2. Case Study

This case study examines a complex problem from the 2025 International Physics Olympiad (IPhO), which investigates the physical principles of Cox's 18th-century timepiece—an ingenious device that harnesses atmospheric pressure fluctuations to generate energy. The problem requires determining optimal parameters to maximize energy dissipation through friction, involving multi-step reasoning that combines mechanical analysis, constraint formulation, and calculus-based optimization.

The P1-235B-A22B model achieved a perfect score (1.0/1.0) on this problem, demonstrating strong capabilities in:

- **Physical intuition:** Correctly identifying the critical force balance constraint at the stop position
- **Mathematical modeling:** Establishing proper relationships between system parameters
- **Optimization techniques:** Successfully applying constrained optimization to find extrema

This performance showcases P1's proficiency in handling competition-level physics problems that demand deep physical understanding and rigorous analytical reasoning.

IPhO 2025 Question 3-Part 3-C4

Background:
In 1765, British clockmaker James Cox invented a clock whose only source of energy is the fluctuations in atmospheric pressure. Cox's clock used two vessels containing mercury. Changes in atmospheric pressure caused mercury to move between the vessels, and the two vessels to move relative to each other. This movement acted as an energy source for the actual clock.

Simplified Model:
The simplified model consists of:

- A cylindrical bottom cistern containing a mercury bath
- A two-part barometric tube (completely emptied of air) dipped into the bath
- The cistern and tube are each suspended by cables through ideal pulleys, attached to a mass M that can slide on a horizontal surface
- Total mercury volume: $V_\ell = 5$ L

The system experiences solid friction between mass M and the horizontal support with magnitude F_s. Two stops limit the displacement: $-X \le x \le X$ (where $X > 0$).

Key parameters are defined as:

$$\xi = \frac{S_\mathrm{b} + S_\mathrm{c}}{S_\mathrm{b} S_\mathrm{c}} \frac{F_\mathrm{s}}{A} \quad \text{and} \quad \lambda = \frac{2\rho g X}{A} \tag{12}$$

where $S_t \ll S_b, S_c$ is assumed.

Question:
Considering $S_\mathrm{b} \simeq S_\mathrm{c}$ and $S_\mathrm{t} \ll S_\mathrm{b}$, determine the expressions for:

1. The optimal friction force $F_\mathrm{s}^\star$ as a function of ρ, g, S_c and A
2. The optimal displacement $X^\star$ as a function of ρ, g, S_c and A
3. The corresponding maximum energy $W^\star$
4. Calculate the numerical value in mJ with $A = 5 \times 10^2$ Pa

Given data: $S_c = 210\ \mathrm{cm}^2$, $A = 5 \times 10^2$ Pa, $g = 9.8\ \mathrm{m{\cdot}s^{-2}}$, $\rho_\mathrm{Hg} = 13534\ \mathrm{kg/m^3}$

P1 Solution Overview

To determine the optimal parameters for maximizing the energy dissipated by solid friction in Cox's timepiece, we analyze the system under the assumption that $S_b \approx S_c$ and $S_t \ll S_b, S_c$. The energy dissipated per period τ_1 is given by:

$$W = 4F_s X \tag{13}$$

This expression arises from the fact that the mass M moves between $-X$ and X, covering a total distance of $4X$ per cycle. The constraint on the system comes from the pressure fluctuation $P_1(t)$ with amplitude A, which must be sufficient to overcome both the restoring force due to mercury displacement and the friction force.

(1) Optimal Friction Force F_s^*

The maximum force provided by atmospheric pressure fluctuations is S_cA, where S_c is the cross-sectional area of the cistern. This force must balance the sum of the restoring force and the friction force at the stop position $x = X$:

$$S_cA = 2S_c\rho gX + F_s \tag{14}$$

Solving for F_s, we get:

$$F_s = S_c(A - 2\rho gX) \tag{15}$$

Substituting into the energy expression $W = 4F_sX$, we obtain:

$$W = 4S_c(A - 2\rho gX)X \tag{16}$$

To maximize W, take the derivative with respect to X and set it to zero:

$$\frac{dW}{dX} = 4S_c(A - 4\rho gX) = 0 \quad \Rightarrow \quad X^* = \frac{A}{4\rho g} \tag{17}$$

Substitute back to find F_s^*:

$$F_s^* = S_c\left(A - 2\rho g \cdot \frac{A}{4\rho g}\right) = S_c\left(A - \frac{A}{2}\right) = \frac{S_cA}{2} \tag{18}$$

Answer: $\boxed{F_s^* = \dfrac{S_cA}{2}}$

(2) Optimal Displacement X^*

As derived in the previous step, when we maximize the work function $W = 4S_c(A - 2\rho gX)X$, we obtain:

$$X^* = \frac{A}{4\rho g} \tag{19}$$

Answer: $\boxed{X^* = \dfrac{A}{4\rho g}}$

(3) Maximum Energy W^*

Using the expressions for F_s^* and X^*:

$$W^* = 4F_s^*X^* = 4 \cdot \frac{S_cA}{2} \cdot \frac{A}{4\rho g} = \frac{S_cA^2}{2\rho g} \tag{20}$$

Answer: $\boxed{W^* = \dfrac{S_cA^2}{2\rho g}}$

(4) Numerical Value of W^*

Given:

- $S_c = 210\ \text{cm}^2 = 0.021\ \text{m}^2$
- $A = 500\ \text{Pa}$
- $\rho = 13534\ \text{kg/m}^3$ (density of mercury)
- $g = 9.8\ \text{m/s}^2$

Compute:

$$W^* = \frac{0.021 \times (500)^2}{2 \times 13534 \times 9.8} = \frac{5250}{265266.4} \approx 0.01979\ \text{J} = 19.79\ \text{mJ} \tag{21}$$

Rounded to two significant figures: $W^* \approx 20\ \text{mJ}$

Answer: $\boxed{W^* \approx 19.8\ \text{mJ}}$