Physics-Guided Policy Optimization with Self-Distillation

Ke Wang[*] KEWANGV@AMAZON.COM
Amazon, Arlington, USA

Yuning Wu[*] YUNINGWU@AMAZON.COM
Amazon, New York, USA

Haoran Liu LIUHR@AMAZON.COM
Amazon, Arlington, USA

Chaoqun Jia ENZOJIA@AMAZON.COM
Amazon, Arlington, USA

Devin Chen DEVICHEN@AMAZON.COM
Amazon, San Jose, USA

Kai Wei KAIWE@AMAZON.COM
Amazon, Seattle, USA

Abstract

Self-distilled policy optimization (SDPO) has become a popular paradigm for LLM post-training, where a model learns from its own predictions conditioned on privileged information. SDPO, however, is sensitive to how much each update step should be trusted: corrections from a self-teacher can be highly informative on some batches and misleading on others, and applying them uniformly with a fixed step size can destabilize training. Drawing inspiration from viscous-fluid dynamics and formalizing the analogy at the SDE level, we propose Physics-Guided Policy Optimization (PGPO), which introduces an information-modulated step-size multiplier derived from a mutual-information estimate between the student's predictions and the feedback-conditioned teacher. We show that this modulation preserves the order-1 weak-approximation guarantees of vanilla SGD, and incurs negligible overhead per iteration. We evaluate PGPO on the Science-QA dataset, where it outperforms SDPO on 3 of the 4 domains with gains of up to +4.5 points, while remaining stable in a setting where SDPO collapses late in training.

1. Introduction

On-policy distillation (OPD) [2, 12, 13, 16, 18] has emerged as a popular training paradigm for large language model post-training, utilizing a teacher model to generate dense, token-level guidance for supervising student model training. A recent variant, self-distilled policy optimization (SDPO) [4], has attracted community attention by treating a single model as both teacher and student: the teacher receives privileged information (e.g., correct solutions) to generate signals, while the student learns from these signals without such access. This method eliminates the need for a separate, larger teacher model and facilitates iterative self-improvement. Existing studies [5, 15, 17, 22, 23] have shown that combining SDPO with reinforcement learning with verified rewards (RLVR) [9] leads to efficient performance gains.

[*]. Equal contribution.

© K. Wang[*], Y. Wu[*], H. Liu, C. Jia, D. Chen & K. Wei.

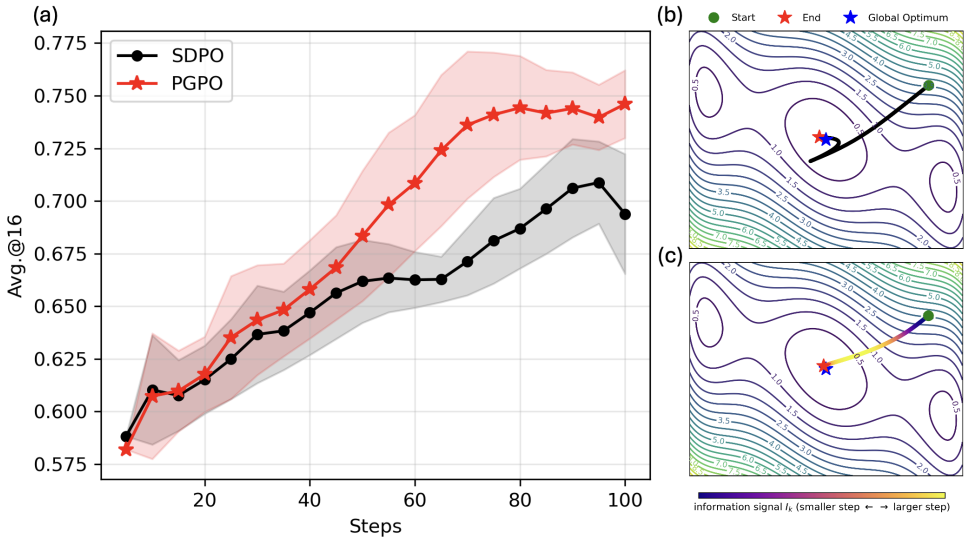

Figure 1: (a) Avg.@16 of PGPO vs. SDPO on Science Q&A (Physics). Motivating intuition: (b) SDPO applies a uniformly-scaled step; (c) PGPO scales each step by an information signal I_k to an oracle teacher (illustration only) with brighter colors marking larger steps.

However, SDPO faces critical challenges during training. While dense token-level supervision enables rapid early gains, it becomes unstable as the student model evolves. As generated prefixes drift from the teacher's conditioning distribution, corrections become unreliable or directionally misleading. Moreover, applying corrections uniformly across all samples, including correct ones, introduces ambiguity and corrupts learned behaviors. Extreme token-level rewards further destabilize training by dominating gradient updates, ultimately leading to collapse. Behind these symptoms lies a fundamental credit assignment problem: *treating all tokens and samples uniformly fails to account for their varying informativeness and reliability.*

To alleviate these issues, several methodologies have been proposed: routing corrections based on sample correctness [10], filtering signals by model confidence [21], and anchoring updates to environment rewards [8]. In contrast to these approaches, we propose **Physics-Guided Policy Optimization** (PGPO), a method grounded in physical intuition from fluid mechanics [1]. PGPO interprets the optimization process as a particle moving through a viscous medium, where viscosity adaptively modulates step sizes based on local mutual information. This mechanism mirrors non-Newtonian fluids, where viscosity adapts to flow conditions, providing a principled alternative to uniform correction strategies that preserves informative signals while suppressing noise. By scaling each update step according to the informativeness of the current self-teacher signal, PGPO addresses the credit assignment problem and prevents the instability that plagues training.

Our contributions are as follows:

- We propose Physics-Guided Policy Optimization (PGPO), a method inspired by fluid mechanics that addresses credit assignment in self-distilled policy optimization (SDPO).
- We show that PGPO admits an order-1 weak SDE approximation and demonstrate its effectiveness in preliminary experiments.

2. Preliminaries

2.1. Problem Formulation

Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ consisting of N prompt-response pairs, where $\mathbf{x}_i$ denotes the input prompt and $\mathbf{y}_i$ the target response from offline datasets or online rollouts. We aim to optimize a large language model (LLM) parameterized by $\theta \in \mathbb{R}^d$, denoted as π_θ, by solving the empirical risk minimization problem:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta), \tag{1}$$

where $\ell_i(\theta) := \ell(\theta; \mathbf{x}_i, \mathbf{y}_i)$ represents the training objective on the i-th example. We use this notation throughout the paper for simplicity.

In self-distilled policy optimization (SDPO) [4], the student π_θ generates responses, receives feedback f_i, and distills from a self-teacher that retrospectively corrects mistakes:

$$\ell_i(\theta) = \sum_t \mathbb{D}_{\text{KL}} \left(\pi_\theta(\cdot|\mathbf{x}_i, \mathbf{y}_{<t}) \,\|\, \text{stopgrad}(\pi_\theta(\cdot|\mathbf{x}_i, f_i, \mathbf{y}_{<t})) \right). \tag{2}$$

2.2. Stochastic Gradient Descent

To minimize $\mathcal{L}(\theta)$, we can employ stochastic gradient descent (SGD) and its adaptive variants (e.g., Adam [6]). At each iteration k, SGD samples a mini-batch $\mathcal{B}_k \subset \{1, ..., N\}$ and updates with learning rate $\eta > 0$:

$$\theta_{k+1} = \theta_k - \eta \cdot \nabla \widehat{\mathcal{L}}(\theta_k), \quad \text{where} \quad \nabla \widehat{\mathcal{L}}(\theta_k) = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla_\theta \ell_i(\theta_k). \tag{3}$$

2.3. Stochastic Differential Equation

The dynamics of SGD can be interpreted as stochastic differential equations (SDEs) [11]. Rewrite the SGD update rule as:

$$\theta_{k+1} - \theta_k = -\eta \cdot \nabla \mathcal{L}(\theta_k) + \sqrt{\eta} \cdot \mathbf{v}_k, \tag{4}$$

where $\mathbf{v}_k = \sqrt{\eta} \cdot (\nabla \mathcal{L}(\theta_k) - \nabla \widehat{\mathcal{L}}(\theta_k))$ is a d-dimensional random vector representing the gradient noise. By the Central Limit Theorem [19], for sufficiently large mini-batch sizes, $\mathbf{v}_k$ is approximately Gaussian with mean zero and covariance matrix $\eta \cdot \mathbf{\Sigma}(\theta_k)$, where

$$\mathbf{\Sigma}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\nabla \mathcal{L}(\theta) - \nabla_\theta \ell_i(\theta) \right) \left(\nabla \mathcal{L}(\theta) - \nabla_\theta \ell_i(\theta) \right)^\top \tag{5}$$

captures the variance of stochastic gradients across dataset.

To derive the continuous-time limit, we identify the discrete update with the Euler-Maruyama discretization of an SDE [7]. Setting the time step $\Delta t = \eta$ and recognizing that $\mathbf{v}_k \approx \sqrt{\eta} \cdot \mathbf{\Sigma}^{1/2}(\theta_k)\, \mathbf{z}_k$, where $\mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathbf{\Sigma}^{1/2}$ denotes the matrix square root.

$$\theta_{k+1} = \theta_k - \eta \cdot \nabla \mathcal{L}(\theta_k) + \eta \cdot \mathbf{\Sigma}^{1/2}(\theta_k)\, \mathbf{z}_k. \tag{6}$$

Taking the continuous-time limit yields the SGD-SDE:

$$d\theta_t = \underbrace{-\nabla\mathcal{L}(\theta_t)}_{\text{drift}}\, dt + \sqrt{\eta} \cdot \underbrace{\mathbf{\Sigma}^{1/2}(\theta_t)}_{\text{diffusion}}\, d\mathbf{w}_t. \tag{7}$$

where $\mathbf{w}_t$ is a standard d-dimensional Brownian motion. The learning rate η sets the time scale: one unit of SDE time corresponds to $1/\eta$ SGD iterations. This SGD-SDE provides a continuous-time approximation of SGD in the *weak sense*: their statistical properties (e.g., marginal distributions, expectations) converge as $\eta \to 0$ (see Appendix A).

Algorithm 1: Physics-Guided Policy Optimization (PGPO)

Data: Dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, initial θ_0, learning rate η, sensitivity α, max multiplier $\rho_{\max}$, iterations K
Result: Optimized parameters θ_K
Initialize optimizer state;
for $k = 0, 1, \ldots, K-1$ **do**
  Sample mini-batch $\mathcal{B}_k \subset \{1, \ldots, N\}$;
  Generate responses and collect feedback for SDPO;
  Compute gradient: $\mathbf{g}_k = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla_\theta \ell_i(\theta_k)$;
  Compute information signal I_k;
  Compute physics-guided multiplier: $\rho_k = \min(\exp(\alpha \cdot I_k), \rho_{\max})$;
  Update: $\theta_{k+1} = \theta_k - \eta \cdot \rho_k \cdot \mathbf{g}_k$;
end
return θ_K;

3. Physics-Guided Policy Optimization

The SGD-SDE can be interpreted through fluid mechanics [1], where the parameter trajectory θ_t represents a particle moving through a viscous fluid. The drift term $-\nabla\mathcal{L}(\theta_t)$ acts as a pressure gradient driving the particle toward loss minima, while the diffusion term $\mathbf{\Sigma}^{1/2}(\theta_t)d\mathbf{w}_t$ captures stochastic fluctuations from mini-batch sampling, analogous to Brownian motion. This perspective suggests a natural mechanism for adaptive step-size control: modulating viscosity based on how informative the self-teacher's correction is. In physical systems, high viscosity slows motion while low viscosity permits faster movement; we let an uninformative correction keep baseline viscosity, and an informative correction reduce it.

We quantify the informativeness of the correction using the mutual information between predictions and feedback on mini-batch $\mathcal{B}_k$:

$$I_k := \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \left[H(\mathbf{y}_i | \mathbf{x}_i, \mathbf{y}_{<t}) - H(\mathbf{y}_i | \mathbf{x}_i, f_i, \mathbf{y}_{<t}) \right], \tag{8}$$

where high I_k indicates that the feedback is highly informative; the student's prediction without feedback diverges from the feedback-conditioned teacher, so the correction carries a strong, reliable signal. Conversely, low I_k indicates that the student already agrees with the teacher, making the

correction small. This motivates an adaptive viscosity that decreases (taking larger steps) when the correction is informative, and stays at baseline when it is not.
With the physical interpretation and a quantified informativeness signal in hand, we introduce Physics-guided Policy Optimization (PGPO), where a multiplier $\rho(I_k)$ modulates the gradient step:

$$\theta_{k+1} = \theta_k - \eta \cdot \rho(I_k) \cdot \nabla\widehat{\mathcal{L}}(\theta_k), \quad \text{where} \quad \rho(I_k) = \min\left(\exp\left(\alpha \cdot I_k\right), \rho_{\max}\right), \tag{9}$$

with $\alpha > 0$ controlling sensitivity and $\rho_{\max} \geq 1$ capping acceleration. This yields:

- High I_k (informative correction): $\rho_k > 1 \rightarrow$ reduced viscosity $\rightarrow$ larger steps
- Low I_k (uninformative correction): $\rho_k \approx 1 \rightarrow$ baseline viscosity $\rightarrow$ standard steps

The cap $\rho_{\max}$ mirrors a physical safety constraint, limiting acceleration when the information signal spikes, while $\rho_k \geq 1$ keeps each update at least as large as the baseline SGD step.
Algorithm 1 presents the complete PGPO framework. The key innovation is the physics-guided multiplier ρ_k, which modulates step size based on the per-batch information signal I_k, with only $O(1)$ overhead per iteration, since I_k reuses the forward pass used for gradient computation. The framework extends to adaptive optimizers (e.g., Adam [6]) by modulating the final update without altering moment accumulation: for Adam with update direction $\mathbf{h}_k$, the viscosity-modulated update becomes $\theta_{k+1} = \theta_k - \eta \cdot \rho_k \cdot \mathbf{h}_k$.

3.1. Experimental Setup

Models. We use Qwen3-8B [20], the general reasoning model, as the initial checkpoint.
Training. We train on Science Q&A [3] across four domains (Chemistry, Physics, Biology, Materials Science) using undergraduate-level reasoning problems, where each domain is a separate training task. We use sensitivity parameters $\alpha \in \{0.5, 1.0, 1.5\}$, with all other hyperparameters following Hübotter et al. [4], details in Appendix B.
Evaluation. We assess in-domain generalization on held-out test splits and report avg@16. All hyperparameters follow Hübotter et al. [4], details in Appendix B.

3.2. Main Results

Table 1: Performance comparison between SDPO and PGPO across scientific domains.

Method	Chemistry	Physics	Biology	Materials
SDPO	72.57	77.05	61.62	74.14
PGPO (best α)	76.02	77.56	61.25	78.65
Improvement	+3.45	+0.51	-0.37	+4.51

PGPO outperforms SDPO in 3 out of 4 scientific domains, achieving notable improvements in Chemistry (+3.45) and Materials (+4.51), with a modest gain in Physics (+0.51). However, PGPO shows a slight decline in Biology (-0.37). The results demonstrate that physics-guided adaptive step-size control provides consistent benefits across most domains, with the largest gains in Chemistry and Materials Science. The full ablations are presented in Appendix C.

4. Conclusion

We proposed Physics-Guided Policy Optimization (PGPO), a method that addresses the credit assignment problem in self-distilled policy optimization (SDPO) through adaptive step-size modulation motivated by an analogy with fluid mechanics. We showed that PGPO is an order-1 weak approximation of a modulated-drift SDE, and demonstrated through preliminary experiments that PGPO achieves consistent improvements over SDPO across multiple scientific reasoning domains. The current instantiation modulates step size at the batch level and is sensitive to the choice of sensitivity parameter α, which we observed in the Biology setting where no single α matched SDPO; finer-grained (e.g., per-token) modulation and adaptive α schedules are natural directions for future work. These results suggest that PGPO offers an alternative approach to stabilizing self-distillation in large language model post-training.

References

[1] Yunus Cengel and John Cimbala. *Ebook: Fluid mechanics fundamentals and applications (si units)*. McGraw Hill, 2013.

[2] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[3] Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2025. URL https://arxiv.org/abs/2406.09098.

[4] Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

[5] Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms?, 2026. URL https://arxiv.org/abs/2603.24472.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[7] PE Kloden and E Platen. Numerical solution of stochastic differential equations springer. *Berlin, Germany*, 1992.

[8] Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation, 2026. URL https://arxiv.org/abs/2603.11137.

[9] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord,

Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

[10] Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing, 2026. URL https://arxiv.org/abs/2604.02288.

[11] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms, 2017. URL https://arxiv.org/abs/1511.06251.

[12] Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

[13] Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

[14] Riccardo Mannella. Numerical integration of stochastic differential equations, 1997. URL https://arxiv.org/abs/cond-mat/9709326.

[15] Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning, 2026. URL https://arxiv.org/abs/2601.19897.

[16] Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models, 2026. URL https://arxiv.org/abs/2604.00626.

[17] Yuda Song, Lili Chen, Fahim Tajwar, Remi Munos, Deepak Pathak, J. Andrew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback, 2026. URL https://arxiv.org/abs/2602.02482.

[18] Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang,

Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026. URL https://arxiv.org/abs/2601.02780.

[19] Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[21] Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026. URL https://arxiv.org/abs/2604.03128.

[22] Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models, 2026. URL https://arxiv.org/abs/2602.12275.

[23] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

Appendix A. Weak Convergence Guarantees

We now show that PGPO admits an order-1 weak approximation by an SDE with a modulated drift. We emphasize that this is a consistency-style result rather than a convergence-rate guarantee.

Definition 1 (Weak Approximation) *[14] Let $0 < \eta < 1$, $T > 0$, and set $N = \lfloor T/\eta \rfloor$. Let $\mathcal{G}$ denote the set of functions with polynomial growth, i.e., $g \in \mathcal{G}$ if there exist constants $K, \kappa > 0$ such that $|g(\mathbf{x})| \leq K(1 + \|\mathbf{x}\|^{\kappa})$. We say that the SDE*

$$d\theta_t = \mathbf{b}(\theta_t)\, dt + \sigma(\theta_t)\, d\mathbf{w}_t, \quad \theta_0 = \theta_{init} \tag{10}$$

*is an **order** α **weak approximation** to SGD if for every $g \in \mathcal{G}$, there exists $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$,*

$$\left|\mathbb{E}[g(\theta_{k\eta})] - \mathbb{E}[g(\theta_k^{SGD})]\right| \leq C\eta^{\alpha}. \tag{11}$$

Theorem 2 (Weak Convergence of PGPO) *Let $T > 0$ and define the gradient covariance $\mathbf{\Sigma}(\theta)$ as in Eq. (5). Assume $\mathcal{L}$ and $\ell_i(\theta)$ are Lipschitz continuous with at most linear growth, and possess sufficiently smooth derivatives in $\mathcal{G}$. Further assume the physics-guided multiplier $\rho(I)$ is bounded and Lipschitz continuous in I. Then:*
***(i) Vanilla SGD:** [11] The SDE $d\theta_t = -\nabla\mathcal{L}(\theta_t)\, dt + \mathbf{\Sigma}^{1/2}(\theta_t)\, d\mathbf{w}_t$ is an order-1 weak approximation of SGD, where $t = k \cdot \eta$ is iteration time.*
***(ii) PGPO:** The modified SDE $d\theta_t = -\rho(I_t) \cdot \nabla\mathcal{L}(\theta_t)\, dt + \mathbf{\Sigma}^{1/2}(\theta_t)\, d\mathbf{w}_t$ is an order-1 weak approximation of PGPO, where I_t is the information measure at time t and the diffusion term remains unchanged.*
***Proof sketch for (ii).** Since ρ is bounded and Lipschitz in I, the modified drift $-\rho(I_t)\nabla\mathcal{L}(\theta_t)$ inherits the Lipschitz and linear-growth conditions of $-\nabla\mathcal{L}$. The result follows by applying [11] to the modified drift, with the diffusion coefficient unchanged.*

Theorem 2 states that the discrete PGPO updates are tracked, in the weak sense, by the modulated-drift SDE as $\eta \to 0$; it does not by itself imply that PGPO converges faster or to a better stationary point than SDPO. We use it solely to confirm that information-based step-size modulation does not break the SDE picture motivating the design. We also note that the form $\rho(I_k) = \min(\exp(\alpha I_k), \rho_{\max})$ is a design choice: any bounded, Lipschitz function of I_k would yield the same conclusion, and we treat α and $\rho_{\max}$ as ordinary hyperparameters.

Appendix B. Experiment Details

Method	SDPO	PGPO
General		
Model	Qwen/Qwen3-8B	Qwen/Qwen3-8B
Thinking	False	False
Data		
Max. prompt length	2048	2048
Max. response length	8192	8192
Batching		
Question batch size	32	32
Mini batch size	32	32
Number of rollouts	8	8
Rollout		
Inference engine	vllm	vllm
Temperature	1.0	1.0
Validation		
Number of rollouts	16	16
Temperature	0.6	0.6
Top-p	0.95	0.95
Training		
Optimizer	AdamW	AdamW
Learning rate	1×10^{-5} (constant)	1×10^{-5} (constant)
Warmup steps	10	10
Weight decay	0.01	0.01
Gradient Clip Norm	1.0	1.0
Sensitivity Parameter	/	$\{0.5, 1.0, 1.5\}$

Table 2: Configuration Comparison

Setting	Prompt Template
Unguided	{question} Please reason step by step, and put your final answer within \boxed{}.
Solution-guided	{question} Please reason step by step, and put your final answer within \boxed{}. Correct solution: {solution} Correctly solve the original question.

Table 3: Prompts for unguided and solution-guided generation following Hübotter et al. [4].

Appendix C. Ablation Studies

Method	Chemistry	Physics	Biology	Materials
SDPO	72.57	77.05	61.62	74.14
PGPO with $\alpha = 0.5$	71.88	75.51	52.00	75.27
PGPO with $\alpha = 1.0$	76.02	76.93	56.50	77.73
PGPO with $\alpha = 1.5$	70.94	77.56	61.25	78.65

Table 4: Performance comparison between SDPO and PGPO across scientific domains with different sensitivity parameter α settings.

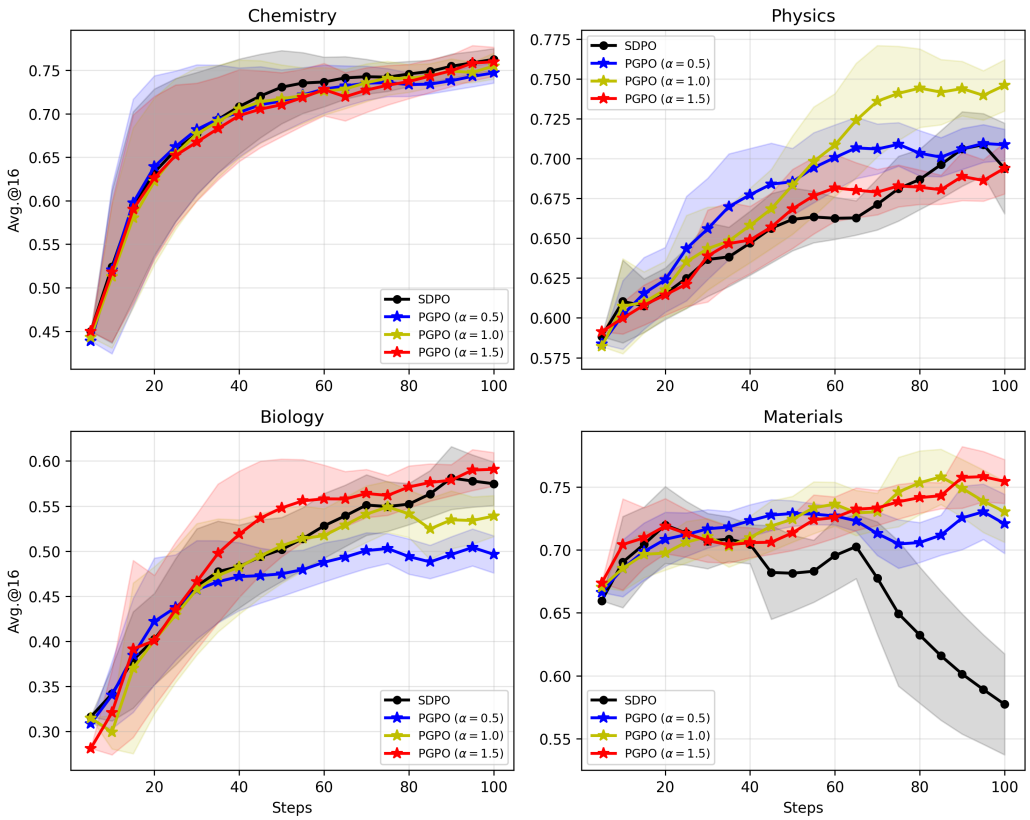

Figure 2: Full ablation study of Physics-Guided Policy Optimization (PGPO) vs. Self-Distillation Policy Optimization (SDPO). Avg.@16 comparison across Science Q&A tasks: Chemistry, Physics, Biology, and Materials.