On the Evolution of Language Models without Labels: *Majority Drives Selection, Novelty Promotes Variation*

Yujun Zhou[1,2†*], **Zhenwen Liang**[1 †], **Haolin Liu**[1,3], **Wenhao Yu**[1], **Kishan Panaganti**[1], **Linfeng Song**[1], **Dian Yu**[1], **Xiangliang Zhang**[2], **Haitao Mi**[1], **Dong Yu**[1]

[1]Tencent AI Lab, [2]University of Notre Dame, [3]University of Virginia

[†] Core contributors

Correspondence to: yzhou25@nd.edu, zhenwzliang@global.tencent.com

Abstract

Large language models (LLMs) are increasingly trained with reinforcement learning from verifiable rewards (RLVR), yet real-world deployment demands models that can self-improve without labels or external judges. Existing self-improvement approaches primarily rely on self-confirmation signals (e.g., confidence, entropy, or consistency) to generate rewards. This reliance drives models toward over-confident, majority-favored solutions, causing an entropy collapse that degrades pass@n and reasoning complexity. To address this, we propose EVOL-RL, a label-free framework that mirrors the evolutionary principle of balancing selection with variation. Concretely, EVOL-RL retains the majority-voted answer as an anchor for stability, but adds a novelty-aware reward that scores each sampled solution by how different its reasoning is from other concurrently generated responses. This *majority-for-stability* + *novelty-for-exploration* rule mirrors the variation–selection principle: *selection prevents drift, while novelty prevents collapse*. Evaluation results show that EVOL-RL consistently outperforms the majority-only baseline; e.g., training on label-free AIME24 lifts Qwen3-4B-Base AIME25 pass@1 from baseline's 4.6% to 16.4%, and pass@16 from 18.5% to 37.9%. EVOL-RL also improves out-of-domain generalization (from math reasoning to broader tasks, e.g., GPQA, MMLU-Pro, and BBEH).

1 Introduction

The reasoning capabilities of Large Language Models (LLMs) have advanced dramatically, particularly through paradigms like Reinforcement Learning with Verifiable Rewards (RLVR) [10, 7, 28]. The next frontier of intelligence lies in enabling LLMs to autonomously evolve, continuously learning from the vast, unlabeled data streams they encounter in real-world environments. This *label-free evolving* paradigm allows a model to iteratively improve itself while solving tasks, without relying on ground-truth labels or external judges, making it both practical and necessary.

The fundamental flaw in relying on internal signals is not merely that they are initially noisy or biased, but that the learning process itself actively degrades the quality of the reward signal over time [16, 31]. By rewarding conformity to its self-confirmation, the model systematically eliminates the solution diversity [12]. This creates a degenerative feedback loop: a progressively narrower and more biased policy generates an increasingly impoverished reward signal, which in turn accelerates the policy's collapse into a low-entropy state [6, 16]. Recent studies also show that training on self-generated data

[*]Work done during Yujun's Internship at Tencent AI Lab.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

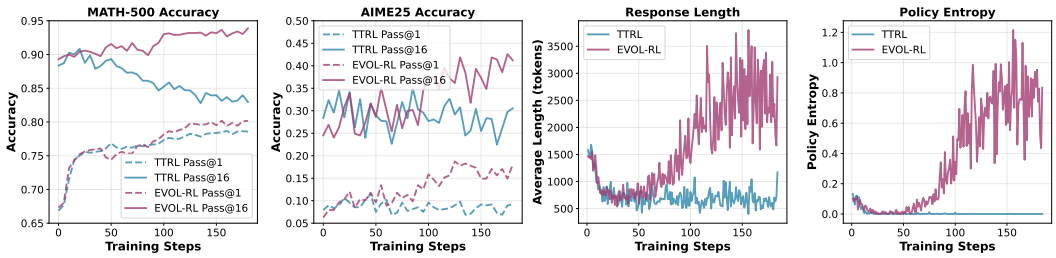

Figure 1: TTRL's entropy collapse vs. EVOL-RL's diversity preservation on Qwen3-4B-Base (trained label-free on MATH-500). Majority-only TTRL drives pass@$n > 1$ down, shortens reasoning, and collapses entropy, whereas EVOL-RL improves accuracy, sustains reasoning diversity.

can harm diversity over time [22] and eventually lead to collapse. Figure 1 illustrates this phenomenon in reasoning: under traditional Test-Time Reinforcement Learning (TTRL) [40], pass@1 may rise but pass@n drops, while response length and complexity steadily decline, indicating that the model fails to evolve.

To address this, we propose ***EVolution-Oriented and Label-free Reinforcement Learning (*EVOL-RL*)***, a simple objective that combines a stabilizing ***selection*** signal with an explicit ***variation*** incentive. Concretely, EVOL-RL retains the majority-voted answer as the anchor for stability, but adds a novelty-aware reward that scores each sampled solution by how different its reasoning is from other concurrently generated responses (semantic similarity of their reasoning traces). This majority-for-stability + novelty-for-exploration rule mirrors the variation–selection principle: *selection prevents drift; novelty prevents collapse*. As demonstrated in Figure 1, EVOL-RL successfully averts all symptoms of diversity collapse, fostering a healthy equilibrium between refining known solutions and discovering new ones. This balanced approach translates into substantial performance gains, especially in out-of-domain generalization. For instance, after training on AIME24, EVOL-RL elevates the Qwen3-4B-base model's pass@1 accuracy on the AIME25 benchmark from 4.6% (TTRL) to 16.4%, while more than doubling the pass@16 accuracy from 18.5% to 37.9%.

Contributions. (1) We diagnose why majority-only objectives shrink exploration during label-free training and formalize their link to entropy collapse on reasoning tasks. (2) We provide a new perspective on label-free learning by framing it as an evolutionary system, which allows us to connect this diversity collapse to the classic problem of premature convergence. (3) Guided by this principle, we design EVOL-RL, a practical novelty-aware reward that complements majority selection to enable stable, label-free improvement, reversing the pass@n decline and improving out-of-domain accuracy. (4) We deliver state-of-the-art results in unsupervised RL, demonstrating that EVOL-RL achieves significant generalization gains where prior methods fail, such as more than tripling pass@1 accuracy and doubling pass@16 accuracy on the challenging AIME25 benchmark.

2 Method

Our approach is illustrated in Figure 2. which uses Group Relative Policy Optimization (GRPO) [21] as its optimization algorithm, but guides it with a novel reward function that explicitly balances majority with novelty.

2.1 Reward Design: Implementing Selection and Variation

Our reward design directly implements the principles of selection and variation to counteract diversity collapse. **Selection**, via the majority vote, provides a stable signal anchored to correctness, preventing the policy from drifting. **Variation**, driven by semantic novelty, provides the necessary exploratory pressure to maintain a diverse set of reasoning strategies. A key design choice is that the novelty incentive is applied strategically to all solutions. For responses that align with the majority, rewarding novelty encourages the discovery of multiple valid reasoning paths, which directly improves pass@n performance. For minority solutions, it incentivizes exploration of the broader reasoning space, increasing the probability of finding a correct solution and avoiding convergence to common failure modes. We provide a detailed intuition about how EVOL-RL avoids collapse in Appendix **??**.

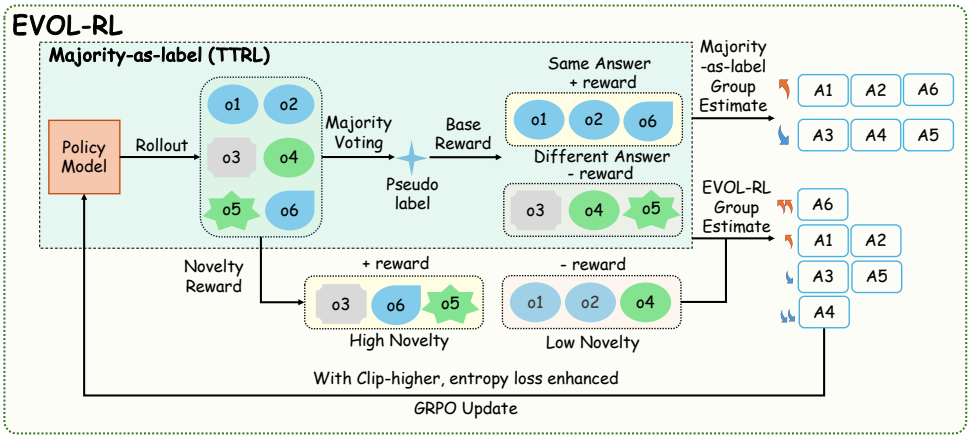

Figure 2: An overview of the EVOL-RL framework. For each prompt, the policy generates multiple responses. These are grouped by their final answer to identify the majority group. A novelty score is then computed for each response based on its semantic dissimilarity to others. Finally, a reward is assigned based on both majority (selection) and novelty (variation), guiding the policy update via GRPO. In the illustration, colors group responses by their final answer, while different marker shapes indicate semantically distinct reasoning paths.

2.2 How EVOL-RL Avoids Collapse Through an Evolutionary Analogy.

EVOL-RL avoids this failure mode by mirroring biological evolution, which balances a stabilizing **Selection** pressure with a dynamic **Variation** mechanism. The majority vote acts as our **Selection** pressure, providing a crucial anchor to correctness. By itself, however, this would lead to a uniform population of solutions vulnerable to collapse.

To prevent this, our three-part **Variation** strategy creates a robust exploratory dynamic. The **entropy regularizer** acts as a "mutation rate," constantly supplying diverse solutions. The **novelty reward** then gives a "survival bonus" to semantically different solutions. Finally, **asymmetric clipping** ensures that when a highly beneficial "mutation"—a rare, novel, and correct solution—appears, its strong learning signal is fully preserved.

This design makes a collapsed state inherently unstable: any solution deviating from a uniform cluster is by definition highly novel, receives a higher reward, and forces the learning algorithm to shift probability towards it, thus ensuring the policy remains robustly diverse.

3 Experiments

The experimental setup is outlined in Appendix C.1. In Appendix D, we provide additional experiments and analyses, including ablation studies, training dynamics, and further experiments showing that EVOL-RL components enhance supervised GRPO (RLVR). We also report generalization results on broader reasoning benchmarks such as MMLU-Pro [26], SuperGPQA [23], and BBEH [11].

3.1 Main Results

Our main results are presented in Table 1. We highlight four key findings that demonstrate the superiority of EVOL-RL over the majority-only TTRL baseline.

EVOL-RL Enhances Both Pass@1 and Pass@16 Performance. Across all experimental settings, EVOL-RL consistently and substantially improves 'pass@16' performance over TTRL, with gains frequently exceeding 20 percentage points on the most challenging benchmarks (e.g., +24.2pp on AIME24 for the 4B model). EVOL-RL also delivers more consistent and substantial improvements to pass@1 accuracy than TTRL. This demonstrates that our method strengthens not only the model's single-shot accuracy but also its ability to explore through multiple attempts.

Consistent Performance Across Scales and Data Sizes. The benefits of EVOL-RL are robust across both model scales (4B and 8B) and training data sizes, from the large MATH-TRAIN set to smaller,

Table 1: Comparison of models trained with TTRL and EVOL-RL. Each cell shows pass@1/pass@16 (averaged on 32 rollouts). Δ uses red (+) for positive and blue for negative values, showing the difference between w/EVOL-RL and w/TTRL.

Training Dataset	Model	MATH	AIME24	AIME25	AMC	GPQA
Qwen3-4B-Base						
–	Base Model	67.4/89.6	10.0/32.4	5.5/30.0	39.3/75.2	34.4/85.6
MATH-TRAIN	w/TTRL	75.4/86.9	12.1/23.2	6.8/28.6	42.5/75.2	36.5/81.4
	w/EVOL-RL	80.0/93.3	20.7/47.6	17.5/39.9	51.4/80.3	37.2/88.7
	Δ	+4.6/+6.4	+8.6/+24.4	+10.7/+11.3	+8.9/+5.1	+0.7/+7.3
MATH-500	w/TTRL	79.3/83.2	10.0/28.0	7.2/29.9	47.6/72.0	36.2/75.9
	w/EVOL-RL	79.8/93.8	19.0/43.2	16.1/41.9	50.3/82.2	38.8/89.1
	Δ	+0.5/+10.6	+9.0/+15.2	+8.9/+12.0	+2.7/+10.2	+2.6/+13.2
AIME24	w/TTRL	73.8/84.5	16.7/16.7	4.6/18.5	43.6/65.8	35.1/73.5
	w/EVOL-RL	79.6/93.6	20.6/40.9	17.1/42.0	49.9/80.9	38.0/87.8
	Δ	+5.8/+9.1	+3.9/+24.2	+12.5/+23.5	+6.3/+15.1	+2.9/+14.3
Qwen3-8B-Base						
–	Base Model	63.6/91.5	12.0/39.4	8.2/30.8	38.7/77.6	34.9/88.0
MATH-TRAIN	w/TTRL	81.1/91.1	16.7/37.6	15.6/35.9	53.6/74.0	38.1/77.1
	w/EVOL-RL	83.6/94.1	26.0/51.7	21.6/43.1	55.5/86.1	43.5/88.1
	Δ	+2.5/+3.0	+9.3/+14.1	+6.0/+7.2	+1.9/+12.1	+5.4/+11.0
MATH-500	w/TTRL	85.7/91.9	17.7/40.1	16.5/34.3	51.1/79.1	43.5/84.0
	w/EVOL-RL	84.7/95.1	24.1/49.5	20.2/44.4	58.8/86.0	43.9/92.2
	Δ	-1.0/+3.2	+6.4/+9.4	+3.7/+10.1	+7.7/+6.9	+0.4/+8.2
AIME24	w/TTRL	76.8/86.2	20.0/20.0	11.4/25.4	49.5/69.1	38.3/74.7
	w/EVOL-RL	83.1/94.2	25.4/38.1	16.5/34.7	54.4/85.8	45.2/90.0
	Δ	+6.3/+8.0	+5.4/+18.1	+5.1/+9.3	+4.9/+16.7	+6.9/+15.3

specialized sets like AIME24. This suggests our method is a fundamental improvement that scales effectively with both model capacity and data volume.

Strong Cross-Task Generalization Within Mathematics. EVOL-RL demonstrates powerful generalization, learning abstract reasoning skills that transfer effectively across different mathematical domains. A compelling example is seen with the 4B model: when trained exclusively on the smaller MATH-500 dataset, its pass@16 performance on the difficult AIME24 benchmark (43.2%) is nearly identical to the performance achieved when training on AIME24 directly (40.9%), confirming that EVOL-RL learns fundamental skills rather than simply overfitting. This effect is further amplified by scale; for the 8B model, the model trained on MATH-500 surpasses the performance of its AIME24-trained counterpart on the AIME24 benchmark by over 11 percentage points in pass@16. This strong cross-task transfer confirms that EVOL-RL fosters the development of fundamental and transferable reasoning abilities.

Generalization Beyond Mathematics. The advantages of EVOL-RL extend beyond mathematics. On the GPQA benchmark, TTRL's performance consistently degrades below the base model, whereas EVOL-RL reliably improves it, achieving gains of +7 to +15 pp in pass@16 over TTRL. This shows our method fosters a more general reasoning ability that transfers effectively across domains.

3.2 Training Dynamics: How EVOL-RL Escapes Entropy Collapse

To understand the reasons for EVOL-RL's better performance, we analyze its training dynamics in comparison to TTRL in a label-free setting, as shown in Figure 3. An analysis of the training dynamics for the 8B models is presented in Appendix D.2.

Stage 1: Initial Collapse Under Majority Signal. Across all three training settings, a consistent initial dynamic unfolds: both EVOL-RL and TTRL show a sharp drop in policy entropy and average

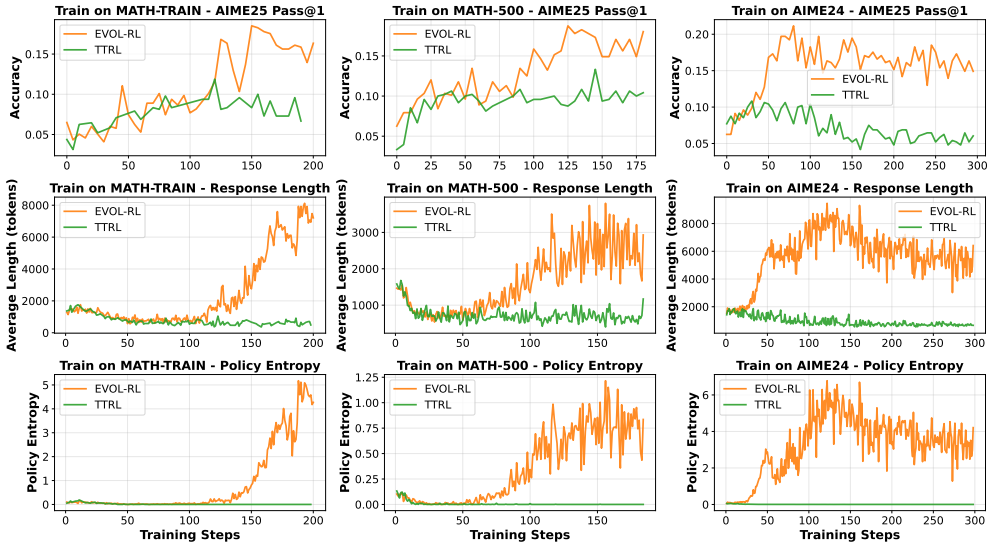

Figure 3: Training dynamics for EVOL-RL and TTRL. **Left:** models trained on *MATH-TRAIN*. **Middle:** models trained on *MATH-500*. **Right:** models trained on *AIME24*. Each panel plots, over training steps, (i) Pass@1 on *AIME25*, (ii) average response length on the training set, and (iii) policy entropy on the training set.

response length. This initial phase demonstrates the powerful homogenizing effect of the majority-driven reward, which quickly pushes both models toward short, high-frequency response templates. For TTRL, this collapsed state proves to be permanent; it remains trapped in this low-entropy, low-complexity state for the duration of the training run, regardless of the dataset's scale or difficulty.

Stage 2: The Evolving Point and Coordinated Recovery. Following the initial collapse, the training dynamics reveal a crucial divergence centered around a distinct **"evolving point"**. Before this point, EVOL-RL's trajectory is nearly indistinguishable from TTRL's; both models exhibit similar performance values and trends, dominated by the majority signal. However, a clear inflection point consistently emerges for EVOL-RL, after which its performance rapidly improves. While the exact timing of this "evolving point" varies across datasets, its appearance is a robust feature of our method. After this "evolving point", EVOL-RL enters a recovery phase characterized by a sustained and coordinated rise across all key metrics: policy entropy breaks away from near-zero values, average response length increases, and out-of-domain accuracy steadily climbs. This coordinated recovery allows the model to reach a new, significantly higher performance plateau where it eventually stabilizes, demonstrating its ability to break free from the majority trap.

EVOL-RL's ability to escape the collapsed state comes from the synergy of its three core components. The entropy regularizer ensures a continuous supply of diverse rollouts, preventing the initial search space from becoming completely uniform. The asymmetric clipping preserves the full gradient signal from the rare but high-value "majority-and-novel" samples that are crucial in the early training phase. Finally, the novelty reward acts as a selection pressure, consistently re-ranking credit within the majority group to favor these distinct solutions over their near-duplicate peers.

4 Conclusion

In this work, we diagnose the entropy collapse, a critical failure mode in LLM evolving where majority-only rewards suppress solution diversity and harm generalization. To solve this, we propose EVOL-RL, a framework that balances the stability of majority-vote selection with an explicit variation incentive that rewards semantic novelty. Our experiments demonstrate that EVOL-RL successfully prevents collapse by maintaining policy entropy and reasoning complexity, which translates into substantial performance gains on both in-domain and out-of-domain benchmarks. By anchoring learning to a stable majority signal while simultaneously encouraging exploration, EVOL-RL offers a robust and practical methodology for enabling LLMs to continuously and autonomously evolve without external labels.

References

[1] Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

[2] John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*, 2025.

[3] 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*, 2025.

[4] Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, et al. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*, 2025.

[5] Runpeng Dai, Tong Zheng, Run Yang, Kaixian Yu, and Hongtu Zhu. R1-re: Cross-domain relation extraction with rlvr. *arXiv preprint arXiv:2507.04642*, 2025.

[6] Mucong Ding, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Chenghao Deng, Alec Koppel, Mengdi Wang, Dinesh Manocha, Amrit Singh Bedi, and Furong Huang. SAIL: Self-improving efficient online alignment of large language models, 2025. URL https://openreview.net/forum?id=02kZwCo0C3.

[7] 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.

[8] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[9] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2025.

[10] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[11] Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. Big-bench extra hard, 2025. URL https://arxiv.org/abs/2502.19187.

[12] Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. Aligning large language models by on-policy self-judgment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11442–11459, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.617. URL https://aclanthology.org/2024.acl-long.617/.

[13] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9, 2024.

[14] Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025.

[15] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025.

[16] Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. Internal consistency and self-feedback in large language models: A survey, 2024. URL https://arxiv.org/abs/2407.14507.

[17] Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, ShaoGuo Liu, and TingTing Gao. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356*, 2025.

[18] Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.

[19] 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.

[20] Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train? *arXiv preprint arXiv:2505.21444*, 2025.

[21] 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.

[22] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024.

[23] P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

[24] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[25] Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, Yujun Zhou, and Xiangliang Zhang. Adareasoner: Adaptive reasoning enables more flexible thinking. *arXiv preprint arXiv:2505.17312*, 2025.

[26] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

[27] Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, et al. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*, 2025.

[28] 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.

[29] 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*, 2025.

[30] Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

[31] Yuwei Zeng, Yao Mu, and Lin Shao. Learning reward for robot skills using large language models via self-alignment. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Z19JQ6WFtJ.

[32] Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*, 2025.

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

[34] Yulai Zhao, Haolin Liu, Dian Yu, SY Kung, Haitao Mi, and Dong Yu. One token to fool llm-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025.

[35] Tong Zheng, Lichang Chen, Simeng Han, R Thomas McCoy, and Heng Huang. Learning to reason via mixture-of-thought for logical reasoning. *arXiv preprint arXiv:2505.15817*, 2025.

[36] Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Xinyu Yang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, et al. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*, 2025.

[37] Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*, 2025.

[38] Yujun Zhou, Yufei Han, Haomin Zhuang, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. Defending jailbreak prompts via in-context adversarial game. *arXiv preprint arXiv:2402.13148*, 2024.

[39] Yujun Zhou, Jiayi Ye, Zipeng Ling, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, Xiangqi Wang, et al. Dissecting logical reasoning in llms: A fine-grained evaluation and supervision study. *arXiv preprint arXiv:2506.04810*, 2025.

[40] 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. *arXiv preprint arXiv:2504.16084*, 2025.

A Related Work

Enhancing Reasoning in Large Language Models. Significant progress in LLM reasoning has been driven by RLVR [10, 7, 28, 29, 27, 5], which fine-tunes models using RL on tasks where an automated verifier can confirm the correctness of the final answer, such as mathematics and coding [30, 24, 25, 3, 9, 4, 36, 39, 35]. While highly effective, the reliance of RLVR on external verifiers restricts its applicability to domains with deterministic, easily checkable solutions [34, 37, 38]. Our work contributes to the effort of improving reasoning in more general domains where such verifiers are unavailable.

Label-Free Adaptation and Self-Improvement. To overcome the limitations of verifiers and adapt to new data distributions, researchers have focused on label-free learning methods that generate reward signals without ground-truth labels. These approaches primarily fall into two categories. One line of research derives rewards from the model's intrinsic confidence, training the model to become more "certain" by rewarding low-entropy or self-consistent outputs [18, 1, 33, 32, 20, 2, 14, 15]. The other prominent paradigm, which our work directly addresses, bootstraps supervision from majority. Test-Time Reinforcement Learning (TTRL) exemplifies this by using the majority-voted answer from multiple samples as a pseudo-label for RL updates [40]. While empirically powerful, we identify a critical flaw in the majority-driven approach: it suppresses solution diversity and actively punishes correct but non-mainstream reasoning, leading to the entropy collapse we describe. While ETTRL adjusts exploration within the original self-consistency framing [17], we are the first to pin down the majority trap and redesign the learning target to couple majority with population-level diversity.

B Details of the Method

B.1 Optimization with GRPO

GRPO is a policy-gradient algorithm designed for fine-tuning LLMs without a separate value function. Its central idea is to evaluate each sampled response relative to a group of its peers generated for the same prompt. This relative evaluation is then used to update the policy with a PPO-style clipped objective, regularized by a KL penalty to ensure stable learning.

For a given prompt $\mathbf{q}$, a policy LLM $\pi_{\theta_{\text{old}}}$ generates a group of G complete responses $\{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$. Each response $\mathbf{o}_i$ receives a scalar reward r_i. Rewards within the group are normalized with a z-score to obtain a response-level advantage:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)},$$

The policy is optimized with a clipped surrogate objective:

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{\frac{\pi_\theta(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}\hat{A}_{i,t},\ \text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})},\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t}\right\} \tag{1}$$

B.2 Detailed Reward Design

Our reward design directly implements the principles of selection and variation to counteract diversity collapse. Selection, based on correctness via majority vote, provides a stable signal to prevent the policy from drifting. Variation, driven by semantic novelty, provides the exploratory pressure needed to maintain a diverse set of reasoning strategies.

A key design choice is that the novelty incentive is applied strategically to all solutions—both those that agree with the majority and those that do not. For majority-aligned solutions, rewarding novelty encourages the model to discover multiple valid reasoning paths to the correct answer, directly fighting the decline in pass@n performance. For minority solutions, rewarding novelty is crucial for escaping local optima. It discourages policy collapse into a few high-frequency failure modes and instead incentivizes exploration of the broader reasoning space, which is essential for increasing the probability of discovering a previously inaccessible, correct solution path. This integration transforms the learning process: it not only mitigates diversity collapse in the current task but also aligns with

the goals of continual learning. By preserving multiple reasoning modes while anchoring to a correct solution, EVOL-RL avoids forgetting potentially useful strategies and retains knowledge diversity for future tasks. Thus, training under EVOL-RL becomes not only an optimization for present performance but also a proactive investment in future adaptability.

Reward Formulation. For each prompt, the policy samples G responses $\{o_i\}_{i=1}^{G}$. Each response is scored on three criteria:

1. Validity: The response must provide a numeric final answer in a `\boxed{·}` format. Responses that fail this check are deemed invalid.

2. Majority (Selection): A binary label $y_i \in \{+1, -1\}$ is assigned based on whether a response's answer matches the majority-voted answer from the valid responses. This serves as our selection signal.

3. Novelty (Variation): We compute embeddings for the reasoning part of each response to form a cosine similarity matrix. For each response o_i, we calculate its mean similarity $\bar{s}_i$ to other responses in the same group (i.e., either majority or minority) and its maximum similarity m_i to any other response in the entire batch. The mean similarity is calculated on an intra-group basis because the majority and minority solutions are often semantically distant; a global mean would be dominated by this gap, obscuring the finer-grained variations among peer solutions within the majority group. The novelty score is:

$$u_i \;=\; 1 - \big(\alpha\,\bar{s}_i + (1-\alpha)\,m_i\big), \qquad \alpha \in \ (\text{default } 0.5).$$

This score is designed to penalize two distinct forms of redundancy: a high $\bar{s}_i$ indicates conformity to the group's semantic average, while a high m_i flags near-duplication of another specific response. The score promotes both local and global diversity. Finally, we min-max normalize the scores $\{u_i\}$ separately within the majority and minority groups to get $\tilde{u}_i$. This intra-group normalization is crucial, as it ensures that novelty is measured relative to one's direct peers, allowing for a fair comparison of diversity within each group.

Final Reward Mapping. We map the majority label and normalized novelty score into non-overlapping reward bands. This ensures that the selection signal from the majority vote is always prioritized, while novelty refines the reward within each group:

$$r_i \;=\; \begin{cases} -1, & \text{if invalid;} \\ 0.5 + 0.5\,\tilde{u}_i \in [0.5, 1], & \text{if } y_i = +1 \text{ (Majority: higher novelty earns higher reward);} \\ -1 + 0.5\,\tilde{u}_i \in [-1, -0.5], & \text{if } y_i = -1 \text{ (Minority: higher novelty mitigates penalty).} \end{cases}$$

Critically, this structure guarantees that any majority solution, regardless of its novelty, receives a higher reward than any minority solution. This maintains a strong pressure towards correctness. More details about the reward implementation are presented in Appendix C.5

Supporting Mechanisms. To further reinforce this reward design, we employ two complementary mechanisms. First, within the GRPO objective (Eq. 1), we use an asymmetric clipping range ($\epsilon_{\text{high}} > \epsilon_{\text{low}}$) [29]. This allows promising and novel solutions with high advantages to receive larger gradient updates, preventing them from being prematurely clipped. Second, we add a token-level entropy regularizer to maintain diversity during the initial generation process:

$$\mathcal{L}_{\text{ent}}(\theta) = -\lambda_{\text{ent}}\,\mathbb{E}_{o\sim\pi_\theta}\left[\frac{1}{|o|}\sum_{t=1}^{|o|} H(\pi_\theta(\cdot \mid o_{<t}, x))\right], \qquad H(p) = -\sum_{v} p(v)\log p(v). \tag{2}$$

The total objective, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \mathcal{L}_{\text{ent}}$, thus directs learning toward semantically distinct, high-quality responses while maintaining a diverse population of solutions.

C Implementation Details

This section provides additional details on the implementation of our reward formulation and supporting mechanisms.

C.1 Experimental Setup

Benchmarks. To test our method at scale, we use the large, standard **MATH training set (MATH-TRAIN)** [8]. We also follow the TTRL [40] by training on two much smaller test sets: the general-purpose **MATH-500** and the competition-level **AIME24** [13]. This comprehensive setup allows us to validate EVOL-RL's versatility across both large-scale and specialized training conditions. Critically, during all training runs, we use only the problem statements, without any ground-truth labels or solutions. For evaluation, we assess the performance of our trained models on a diverse set of five benchmarks to measure both in-domain and out-of-domain generalization. The evaluation suite includes **AIME24**, **AIME25**, **MATH500**, **AMC** [13], and **GPQA-Diamond (GPQA)** [19]. Detailed training configuration can be found in Appendix C.

C.2 Training Configuration.

We conduct our experiments on two recent open-source base models: **Qwen3-4B-Base** and **Qwen3-8B-Base**. Our training process is implemented using the GRPO algorithm. We adopt a setup similar to that of TTRL for generating training signals. For each problem instance, we first perform a rollout phase where the policy generates 64 candidate responses. A majority label is then determined by performing a majority vote on the final answers extracted from these 64 samples. Subsequently, a random subset of 32 of these responses is used to form a batch for a single model update step. To ensure that the model has sufficient capacity for complex, multi-step reasoning, we set the maximum response length to 12,288 tokens during generation. To guide the model's reasoning process, we utilize the system prompt from SimpleRL-Zoo [30]. Implementation details are discussed in Appendix C.

C.3 System Prompt

For all experiments, we used the following system prompt to guide the model's generation format, ensuring that it produces a step-by-step reasoning process and a clearly marked final answer [30]:

System Prompt

```
Please reason step by step, and put your final answer within
\boxed{}.
```

C.4 Answer and Reasoning Extraction

To implement the scoring criteria described in the main text, we apply the following extraction procedure for each generated response o_i:

- **Final Answer Extraction (for Validity):** We parse the response to find the content within the final occurrence of the `\boxed{·}` command. A response is deemed "valid" only if this command is present and its content contains at least one numeric digit. This extracted numeric string is used for the majority vote.

C.5 Novelty Score Calculation Details

The novelty score u_i relies on computing semantic similarity between the reasoning parts of the generated responses.

Embedding Model. We use the **Qwen3-4B-Embedding** model to generate dense vector representations for the extracted reasoning parts. Each vector is L2-normalized before similarity computation.

Cosine Similarity Matrix. For a group of G responses with corresponding L2-normalized embedding vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_G\}$, the cosine similarity matrix $\mathbf{S} \in \mathbb{R}^{G \times G}$ is computed as $\mathbf{S} = \mathbf{V}\mathbf{V}^T$, where $\mathbf{V}$ is the matrix whose rows are the vectors $\mathbf{v}_i$. The element S_{ij} represents the cosine similarity between the reasoning of response o_i and o_j.

Intra-Group Min-Max Normalization. To obtain the normalized novelty score $\tilde{u}_i \in [0, 1]$ from the raw scores $\{u_k\}$ within a specific group (e.g., the majority group), we apply standard min-max normalization:

$$\tilde{u}_i = \frac{u_i - \min(\{u_k\})}{\max(\{u_k\}) - \min(\{u_k\}) + \epsilon_{\text{norm}}}$$

where ϵ_{norm} is a small constant (e.g., 10^{-8}) to prevent division by zero in cases where all novelty scores in the group are identical.

C.6 Hyperparameter Settings

For our label-free experiments, we largely follow the settings established by TTRL to ensure a fair comparison. The general hyperparameters are detailed in Table 2, and the settings specific to our EVOL-RL method are listed in Table 3.

Table 2: General hyperparameters for label-free training, following TTRL.

Hyperparameter	Value
Train Batch Size	8
PPO Mini-Batch Size	1 (effective size of 32)
PPO Micro-Batch Size	2
Rollouts for Majority Vote	64
Rollouts Used for Training	32
Generation Temperature	1.0
Validation Temperature	0.6
Learning Rate	5e-7
Use KL Loss	True
KL Loss Coefficient	0.001

Table 3: Key hyperparameters specific to the EVOL-RL framework.

Hyperparameter	Value
Asymmetric Clipping High (ϵ_{high})	0.28
Entropy Regularizer Coefficient (λ_{ent})	0.003
Novelty Score Mixing Coefficient (α)	0.5

C.7 Computational Resources

All experiments reported in this paper were conducted on a single server equipped with 8x NVIDIA H20 GPUs.

D Additional Experimental Results

D.1 Ablation Study

Setup. We conduct an ablation study on EVOL-RL-trained models on Qwen3-4B-Base. EVOL-RL introduces three key modifications compared to the TTRL baseline: (i) the novelty-aware reward function, (ii) a rollout entropy regularizer to encourage exploration, and (iii) an asymmetric PPO clipping window (higher "ClipHigh") to better preserve learning signals from high-reward samples. We systematically remove these components one at a time ("-Novelty Reward", "-Ent", "-ClipHigh") or in combination. The results are reported in Table 4.

The Critical Role of Novelty on Easier Datasets. The importance of the novelty reward is most evident when the model is trained on the MATH-500 dataset. Removing it causes the largest performance degradation in pass@16, especially on the more difficult, out-of-domain AIME24/25. This is because on a dataset with lower complexity, a majority-only approach can quickly cause the

Table 4: Performance of Qwen3-4B-Base with EVOL-RL and its ablations on five benchmarks. Each cell reports pass@1/pass@16 accuracy.

Training Dataset	Model	MATH	AIME24	AIME25	AMC	GPQA
MATH-500	**w/EVOL-RL**	**79.8/93.8**	**19.0/43.2**	**16.1/41.9**	**50.3/82.2**	**38.8/89.1**
	-ClipHigh	75.1/91.8	12.2/31.8	11.4/31.3	42.7/73.9	32.3/81.8
	-Ent	79.5/93.4	18.3/38.5	14.7/34.3	48.3/78.6	38.6/87.0
	-ClipHigh-Ent	76.3/92.6	12.8/38.8	12.5/37.4	46.2/77.4	35.6/88.8
	-Novelty Reward	79.3/88.7	12.1/27.0	11.1/34.8	47.6/73.3	37.9/81.4
AIME24	**w/EVOL-RL**	**79.6/93.6**	**20.6/40.9**	**17.1/42.0**	**49.9/80.9**	**38.0/87.8**
	-ClipHigh	74.1/89.4	14.1/26.7	8.1/31.1	44.6/73.2	35.3/81.5
	-Ent	66.7/89.8	10.0/31.4	6.6/27.8	38.7/74.2	34.0/86.2
	-ClipHigh-Ent	75.3/89.0	16.6/26.9	9.2/32.2	45.8/71.2	37.1/82.0
	-Novelty Reward	79.4/93.0	17.7/35.6	15.9/37.4	48.8/79.6	37.9/87.1

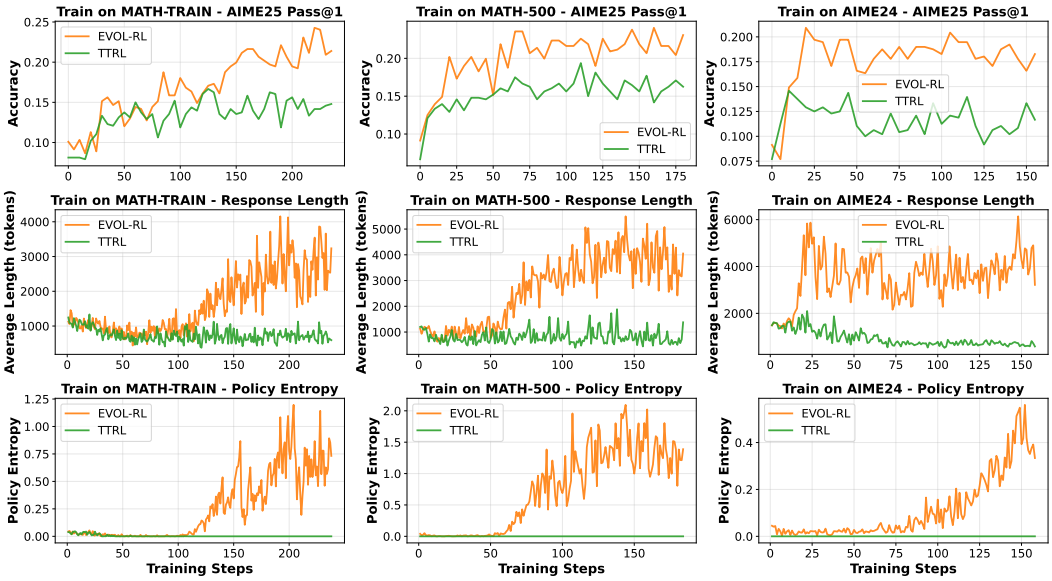

Figure 4: Training dynamics for EVOL-RL and TTRL on Qwen3-8B-Base model. **Left:** models trained on *MATH-TRAIN*. **Middle:** models trained on *MATH-500*. **Right:** models trained on *AIME24*. Each panel plots, over training steps, (i) Pass@1 on *AIME25*, (ii) average response length on the training set, and (iii) policy entropy on the training set.

model to lock into a single, repetitive reasoning template. Our novelty reward prevents this template lock-in and promotes generalizable skills.

Exploration Mechanisms as Critical Enablers on Harder Tasks. On more challenging datasets like AIME24, where the inherent problem difficulty naturally induces a higher baseline of exploration, the other two components become more critical. In this setting, removing the entropy regularizer or the asymmetric clipping consistently lowers pass@16 performance on AIME-style problems. These mechanisms act as crucial enablers for the novelty reward: the entropy regularizer ensures a rich and continuous supply of varied reasoning paths for the novelty selector to act upon, while the higher clipping threshold preserves the full learning signal from rare but high-value solutions.

D.2 Training Dynamics of 8B Models

The training dynamics of the 8B models, presented in Figure 4, largely mirror the patterns observed with the 4B models, confirming that the core mechanisms of EVOL-RL are robust to scale.

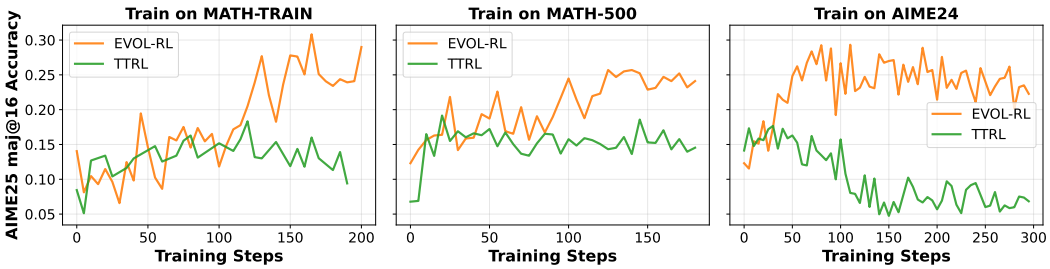

Figure 5: Training dynamics of the majority-vote accuracy (maj@16) for EVOL-RL and TTRL. Each panel plots the accuracy of the consensus answer derived from 16 rollouts over the course of training. The training datasets are: (**Left**) MATH-TRAIN, (**Middle**) MATH-500, and (**Right**) AIME24.

Across all three training datasets (MATH-TRAIN, MATH-500, and AIME24), we observe the same two-stage process. In Stage 1, both TTRL and EVOL-RL experience an initial drop in policy entropy and response length due to the strong initial pressure of the majority-vote signal. TTRL becomes permanently trapped in this low-entropy, low-complexity state.

In Stage 2, EVOL-RL consistently diverges at an "evolving point." Its policy entropy begins a sustained recovery, followed by a coordinated increase in average response length and out-of-domain accuracy on AIME25. This confirms that even at a larger scale, EVOL-RL successfully prevents entropy collapse and fosters a positive feedback loop where exploration, reasoning complexity, and performance reinforce one another, while the consensus-only TTRL approach stagnates.

D.3 Analysis of the Majority Vote Signal

To further investigate the differences between EVOL-RL and TTRL, we analyze the quality of the training signal itself by tracking the accuracy of the majority vote (maj@16) over the course of training, as shown in Figure 5. This analysis reveals how the self-generated pseudo-labels evolve under each method.

A highly consistent pattern emerges across all three training datasets. TTRL initially improves the maj@16 accuracy over the base model, but it quickly converges to a performance plateau. For the remainder of the training, its maj@16 accuracy remains largely unchanged, indicating that the consensus-only approach rapidly finds a local optimum for the consensus answer and becomes locked in, unable to discover better solutions.

In contrast, EVOL-RL exhibits a markedly different dynamic. While its initial trajectory often mirrors that of TTRL, reflecting the early stabilizing influence of the consensus signal, a clear divergence occurs. Consistent with the inflection point observed in our main training dynamics analysis, EVOL-RL's maj@16 accuracy breaks away from the TTRL plateau and begins a second, sustained ascent. It reliably climbs to and stabilizes at a significantly higher level of accuracy. This demonstrates that EVOL-RL's exploration mechanisms not only improve the final policy but also progressively refine the quality of the pseudo-labels used for training, allowing the model to escape suboptimal consensus and continuously improve its understanding of the task.

D.4 EVOL-RL Components Also Strengthen Supervised GRPO (RLVR)

Setup. We apply EVOL-RL's three exploration-enhancing ingredients to a standard supervised GRPO baseline trained on *MATH* training set [8] with a ground-truth verifier (RLVR) for two epochs. Figure 6 reports the results.

The primary finding is that the three components are still synergistic, with their full combination yielding the most significant and consistent performance improvements. This complete configuration, `GRPO+ClipHigh+Ent+Novelty`, boosts pass@16 accuracy by 7% to 12% on the challenging out-of-domain AIME24 and AIME25 benchmarks. Crucially, these gains are achieved while also improving pass@1 accuracy, demonstrating that the mechanisms enhance multi-path reliability without sacrificing single-shot performance. This robust improvement extends across all evaluation benchmarks, including the cross-domain GPQA task, demonstrating the great potential of variation reward in a broader context.

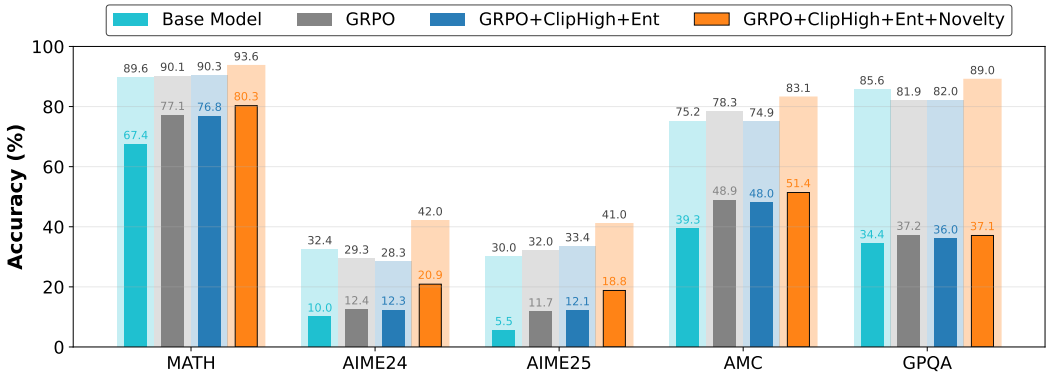

Figure 6: Performance of EVOL-RL's exploration-enhancing components when applied to a standard supervised GRPO baseline. The Qwen3-4B-Base model is trained on the MATH trainig set [8] with a ground-truth verifier (RLVR).

Table 5: Generalization performance of the Qwen3-8B-Base model on broader reasoning benchmarks after label-free training on MATH-TRAIN.

Model	**MMLU-Pro**		**SuperGPQA**		**BBEH**	
	Pass@1	Pass@4	Pass@1	Pass@4	Pass@1	Pass@4
Qwen3-8B-Base	47.3	74.5	26.5	54.1	10.4	24.0
w/TTRL	53.4	73.9	29.7	53.3	**12.1**	24.1
w/EVOL-RL	**55.3**	**78.5**	**30.2**	**57.0**	11.5	**24.9**

D.5 Generalization to Broader Reasoning Benchmarks

To assess whether the reasoning skills enhanced by our method on mathematical data are fundamental and transferable, we evaluate our models on a suite of broader, non-mathematical reasoning benchmarks. After training the Qwen3-8B-Base model on the MATH-TRAIN dataset in a label-free setting, we measure its performance on **MMLU-Pro** [26], **SuperGPQA** [23], and **BBEH** [11]. The results, presented in Table 5, demonstrate that EVOL-RL fosters a more generalizable reasoning ability compared to TTRL.

A contrasting pattern emerges between the two methods. While TTRL shows clear improvements over the base model on pass@1 accuracy, its effect on pass@4 is less consistent, falling slightly below the base model's performance on SuperGPQA and BBEH. This pattern is consistent with our findings on the mathematical reasoning tasks, where the narrow focus of the consensus-only objective can hurt multi-path reliability. In contrast, EVOL-RL demonstrates a more robustly positive transfer of skills, improving upon both the base model and TTRL across pass@1 and pass@4 metrics. For example, on MMLU-Pro, EVOL-RL achieves a pass@4 score of 78.5%, a clear improvement over TTRL's 73.9%. This indicates that our principle of encouraging diverse reasoning helps the model learn more fundamental skills that generalize effectively beyond mathematics.

E Use of Large Language Models in Preparation

We acknowledge the use of Large Language Models (LLMs) as assistants in the preparation of this manuscript. Their role included refining phrasing and improving the clarity of the text, as well as assisting with programming tasks such as code generation and debugging for our experiments. The authors critically reviewed, edited, and verified all LLM-generated content for accuracy and appropriateness, and take full responsibility for the final content of this paper.