R-MEL: RECOVERING CONTRASTIVE SIGNAL FROM ALL-NEGATIVE GROUPS VIA PREFIX-PRIMED REVISION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has enabled significant advances in LLM reasoning through contrastive learning from trajectory groups. However, when all sampled trajectories fail verification ("all-negative" groups), no contrastive pair exists and the group is discarded—wasting approximately 30% of training compute. We observe that failed trajectories often contain correct reasoning prefixes before diverging into errors. We propose R-MEL (Revision-Augmented Meta-Experience Learning), which recovers contrastive signal from all-negative groups by truncating failed trajectories at candidate bifurcation points and generating correct continuations. On mathematical reasoning benchmarks, R-MEL achieves 33.17 average Pass@1 and the highest Avg@8 (30.78) across all conditions, outperforming baselines on 3 of 5 benchmarks including a statistically significant +4.0 percentage point improvement on MATH-500. Analysis reveals an inverted-U pattern in revision effectiveness: intermediate-difficulty prompts show the highest success rate (15.4%), providing insight into when prefix-primed revision is most beneficial.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful paradigm for improving large language model (LLM) reasoning capabilities (DeepSeek-AI et al., 2025; Shao et al., 2024). Unlike traditional RLHF (Ouyang et al., 2022), RLVR uses programmatic verifiers to provide objective pass/fail signals, enabling scalable training without human annotation. Methods like Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and Meta-Experience Learning (MEL) (Huang et al., 2026) leverage contrastive learning from groups of sampled trajectories, using correct and incorrect pairs to compute relative advantages and extract reusable meta-experiences.

However, a critical limitation of these contrastive methods is their reliance on mixed-outcome groups—groups containing both correct and incorrect trajectories. When all sampled trajectories fail verification (an "all-negative" group), no contrastive pair exists and the group is discarded. In our experiments, approximately 30% of training groups are all-negative during early training, representing substantial wasted computation and lost learning signal. This problem is particularly acute for challenging problems where failure rates are high.

We observe that failed trajectories are not uniformly wrong—they often contain correct reasoning prefixes before diverging into errors. This "good prefix" represents partial progress that could be salvaged. If we can identify where reasoning diverges and generate a correct continuation from that point, we can recover a contrastive pair that provides meaningful learning signal.

In this paper, we propose **R-MEL** (Revision-Augmented Meta-Experience Learning), which recovers contrastive signal from all-negative groups through prefix-primed revision. When an all-negative group is detected, R-MEL truncates a seed trajectory at candidate bifurcation points and generates

[1] https://gitlab.com/fars-a/verifier-edited-mel-negatives

continuations. If a revision passes verification and maintains sufficient prefix overlap with the original failure, the pair is used for standard MEL meta-experience construction. This approach converts discarded groups into useful training data without requiring additional rollout compute.

Our contributions are:

- We identify the all-negative waste problem in RLVR contrastive methods and propose prefix-primed revision as a principled solution.
- We introduce R-MEL, which recovers contrastive signal from discarded groups by truncating failed trajectories and generating correct continuations.
- We demonstrate that R-MEL achieves 33.17 average Pass@1 on mathematical reasoning benchmarks, outperforming the MEL baseline (33.00) and achieving the highest Avg@8 (30.78) across all conditions.
- We reveal an inverted-U pattern in revision effectiveness: intermediate-difficulty prompts show the highest revision success rate (15.4%), providing insight into when prefix-primed revision is most beneficial.

2 Related Work

Reinforcement Learning with Verifiable Rewards. Reinforcement learning from human feedback (RLHF) has become a standard paradigm for aligning language models with human preferences (Ouyang et al., 2022). Recent work has extended this to domains with verifiable rewards (RLVR), where correctness can be automatically checked. DeepSeek-R1 (DeepSeek-AI et al., 2025) demonstrated that large-scale RLVR can incentivize sophisticated reasoning capabilities in LLMs. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) introduced contrastive learning from groups of sampled trajectories, using correct and incorrect pairs to compute relative advantages. Meta-Experience Learning (MEL) (Huang et al., 2026) further refines this approach by constructing meta-experiences from trajectory pairs and internalizing them into model memory. Direct Preference Optimization (DPO) (Rafailov et al., 2023) provides an alternative by directly optimizing preferences without explicit reward modeling. DAPO (Yu et al., 2025) scales these approaches to open-source systems. A common limitation across these methods is the reliance on contrastive pairs: when all sampled trajectories fail verification (all-negative groups), no learning signal is available and the group is discarded.

Credit Assignment in Language Model RL. Effective credit assignment is crucial for RL training of language models. VinePPO (Kazemnejad et al., 2024) addresses this by using Monte Carlo rollouts to estimate token-level advantages, enabling more precise credit assignment than trajectory-level rewards. Process reward models (Lightman et al., 2023) provide step-by-step verification, allowing fine-grained supervision of reasoning chains. Setlur et al. (2024) scale automated process verifiers to improve reasoning through progress-based rewards. Khandoga et al. (2026) propose causal credit assignment methods that go beyond uniform credit distribution. These approaches share the goal of providing more informative learning signals, but focus on refining rewards for successful trajectories rather than recovering signal from failures.

Learning from Mistakes. Several recent works explore how to leverage incorrect trajectories for learning. RLMEC (Chen et al., 2024) introduces fine-grained RL with minimum editing constraints, identifying and correcting errors while preserving correct portions. EditGRPO (Zhang et al., 2025) applies post-rollout edits to improve clinical report generation. Feng et al. (2025) propose confidence reweighting to leverage negative RL groups that would otherwise be discarded. Save the Good Prefix (Liu et al., 2026) uses process-supervised RL to precisely penalize errors while preserving correct prefixes. An et al. (2023) demonstrate that learning from mistakes can improve LLM reasoning. Our work differs from these approaches by specifically targeting the all-negative waste problem in RLVR, using prefix-primed continuation to recover contrastive signal while maintaining the meta-experience learning framework.

3 METHOD

3.1 BACKGROUND: META-EXPERIENCE LEARNING

Meta-Experience Learning (MEL) (Huang et al., 2026) extends Group Relative Policy Optimization (GRPO) (Shao et al., 2024) by constructing and internalizing reusable meta-experiences from contrastive trajectory pairs. Given a prompt x, MEL samples K trajectories $\mathcal{Y} = \{y_i\}_{i=1}^{K}$ from the current policy π_θ and partitions them using a verifier into correct trajectories $\mathcal{Y}^+ = \{y_i : r(y_i) = 1\}$ and incorrect trajectories $\mathcal{Y}^- = \{y_i : r(y_i) = 0\}$.

When both $\mathcal{Y}^+$ and $\mathcal{Y}^-$ are non-empty, MEL identifies a *bifurcation point*—the first position where correct and incorrect trajectories diverge—and extracts a meta-experience consisting of a critique and heuristic. These meta-experiences are validated via replay and distilled into model weights through negative log-likelihood (NLL) loss.

The All-Negative Problem. A critical limitation of MEL is its reliance on mixed-outcome groups. When all K sampled trajectories fail verification ($|\mathcal{Y}^+| = 0$), no contrastive pair exists and the group is discarded. In our experiments with Qwen2.5-7B-Instruct on DAPO-Math-17k, approximately 30% of training groups are all-negative during early training, representing substantial wasted computation and lost learning signal.

3.2 R-MEL: PREFIX-PRIMED REVISION

We propose **R-MEL** (Revision-Augmented Meta-Experience Learning), which recovers contrastive signal from all-negative groups through prefix-primed revision. The key insight is that failed trajectories often contain correct reasoning prefixes before diverging into errors. By truncating at a candidate bifurcation point and generating a new continuation, we can potentially recover a correct trajectory that shares a meaningful prefix with the original failure.

Revision Procedure. For an all-negative group, R-MEL performs the following steps:

1. **Seed Selection**: Select the longest trajectory $y^- \in \mathcal{Y}^-$ as the seed, as longer trajectories often represent more complete reasoning attempts.
2. **Prefix-Primed Continuation**: For each truncation ratio $\tau \in \{0.7, 0.5, 0.3\}$:
   - Truncate y^- at position $\lfloor \tau \cdot |y^-| \rfloor$ to obtain prefix p_τ
   - Generate a continuation from p_τ using the current policy
   - If the resulting trajectory y^+_{rev} passes verification, proceed to filtering
3. **LCP Filtering**: Accept the revision only if the longest common prefix (LCP) ratio between y^+_{rev} and y^- satisfies $\text{LCP}(y^+_{\text{rev}}, y^-)/|y^-| \geq \tau$. This ensures the revision preserves the intended prefix rather than diverging immediately.
4. **Meta-Experience Construction**: If a valid revision is found, construct a contrastive pair (y^+_{rev}, y^-) and proceed with standard MEL meta-experience extraction.

Figure 1 illustrates the R-MEL framework. The revision budget B controls the total number of continuation attempts across all truncation ratios (default $B = 4$).

3.3 DESIGN CHOICES

Truncation Ratios. We use a progressive schedule $\tau \in \{0.7, 0.5, 0.3\}$, starting with conservative truncation (preserving 70% of the seed) and progressively becoming more aggressive. This allows the model to first attempt minimal corrections before exploring more substantial revisions. The schedule is tried in order, stopping at the first successful revision.

Revision Budget. The budget $B = 4$ represents a trade-off between compute cost and coverage. Our ablation study shows that $B = 1$ captures approximately 35% of the rescued groups (99 vs. $\sim$283 for $B = 4$), with diminishing returns from additional attempts. The first attempt at $\tau = 0.7$ is most likely to succeed when the error is localized near the end of the trajectory.

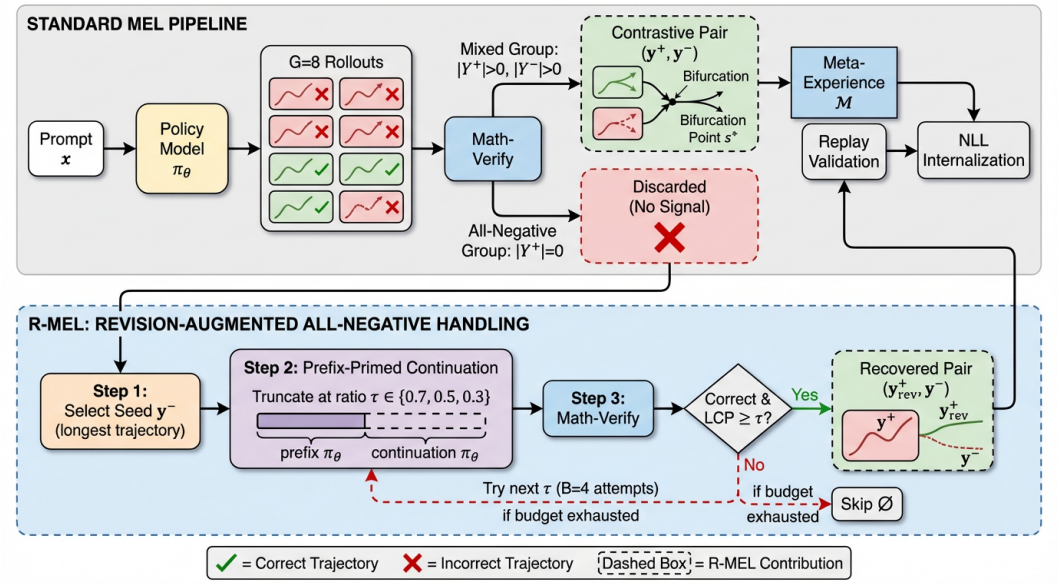

Figure 1: Overview of R-MEL framework. **Top**: Standard MEL pipeline where all-negative groups ($|\mathcal{Y}^+| = 0$) are discarded. **Bottom**: R-MEL extension that recovers contrastive signal via prefix-primed revision. Given a seed trajectory y^-, R-MEL truncates at multiple ratios $\tau \in \{0.7, 0.5, 0.3\}$ and generates continuations. If a revision passes verification and maintains LCP $\geq \tau$ with the seed, the pair (y^+_{rev}, y^-) is used for meta-experience construction.

Seed Selection. We select the longest trajectory as the seed, as it typically represents the most complete reasoning attempt. Alternative strategies (e.g., highest log-probability) showed similar performance in preliminary experiments.

4 Experiments

4.1 Experimental Setup

Model and Training. We use Qwen2.5-7B-Instruct (Yang et al., 2024) as our base model. Training is conducted on 8×A100-80GB GPUs using the VERL framework with AdamW optimizer (learning rate 1×10^{-6}, batch size 128). We train for 100 steps with $K = 8$ rollouts per prompt, temperature 1.0, and clip ratio 0.2. For R-MEL, we use revision budget $B = 4$ and prefix constraint $\tau = 0.3$.

Benchmarks. We evaluate on five mathematical reasoning benchmarks: AIME24 and AIME25 (30 competition problems each), AMC23 (40 problems), MATH-500 (Hendrycks et al., 2021) (500 problems), and OlympiadBench (He et al., 2024) (Olympiad-level problems). Answer verification uses Math-Verify.

Baselines. We compare three conditions: (A) **MEL**: Standard meta-experience learning that discards all-negative groups; (B) **MEL+Extra**: MEL with additional rollouts for all-negative groups (compute-matched control); (C) **R-MEL**: Our method with prefix-primed revision.

Metrics. We report Pass@1 (greedy decoding, temperature 0) and Avg@8 (mean accuracy over 8 samples at temperature 0.6). Best checkpoint is selected based on average Pass@1 across benchmarks.

Table 1: Main results on mathematical reasoning benchmarks. Best results in **bold**, second-best underlined. R-MEL achieves the highest Avg@8 (30.78) and outperforms MEL+Extra on 3/5 benchmarks for Pass@1.

Method	Pass@1					
	AIME24	AIME25	AMC23	MATH-500	OlympiadBench	Avg
MEL	10.00	**16.67**	50.00	53.80	34.52	33.00
MEL+Extra	**16.67**	13.33	50.00	53.80	32.74	**33.31**
R-MEL	13.33	6.67	**52.50**	**57.80**	**35.56**	33.17
Method	Avg@8					
	AIME24	AIME25	AMC23	MATH-500	OlympiadBench	Avg
MEL	8.75	7.50	41.56	49.75	31.98	27.91
MEL+Extra	8.33	**8.33**	43.44	47.45	31.52	27.81
R-MEL	**10.42**	**8.33**	**46.88**	**54.10**	**34.15**	**30.78**

4.2 Main Results

Table 1 presents the main experimental results. R-MEL achieves 33.17 average Pass@1, outperforming the MEL baseline (33.00) by +0.17 points and demonstrating that prefix-primed revision can recover useful training signal from all-negative groups. While MEL+Extra achieves the highest average Pass@1 (33.31), R-MEL outperforms it on 3 out of 5 benchmarks: AMC23 (52.50 vs. 50.00), MATH-500 (57.80 vs. 53.80), and OlympiadBench (35.56 vs. 32.74). The MATH-500 improvement of +4.0 percentage points is statistically significant (95% CI [1.00, 7.20] via bootstrap).

Notably, R-MEL achieves the highest Avg@8 score (30.78) across all conditions, compared to MEL (27.91) and MEL+Extra (27.81). This suggests that the revision mechanism improves not only peak accuracy but also the diversity and coverage of the learned policy. The consistent advantage on Avg@8 indicates that revision-derived meta-experiences help the model explore a broader range of correct solution strategies.

4.3 Ablation Study

We ablate two key design choices in R-MEL: the prefix constraint (τ) and the revision budget (B). Table 2 shows the results.

Prefix Constraint. Removing the prefix constraint ($\tau = 0$, accepting any correct revision) degrades average Pass@1 from 33.17 to 32.69 (-0.48). Without the constraint, 33.9% of accepted revisions have LCP ratio below 0.5, meaning the revised trajectory diverges substantially from the seed. These low-overlap pairs provide weaker contrastive signal because the "positive" trajectory shares little context with the "negative" one, making bifurcation-point detection less meaningful.

Revision Budget. Reducing the revision budget from $B = 4$ to $B = 1$ decreases average Pass@1 from 33.17 to 32.80 (-0.37). With $B = 1$, only the first truncation ratio ($\tau = 0.7$) is attempted, capturing approximately 35% of the rescued groups (99 vs. $\sim$283 for $B = 4$). The diminishing returns suggest that most revision benefit comes from the first, most conservative attempt, with additional attempts at more aggressive truncation ratios providing smaller marginal gains.

4.4 Analysis

Training Dynamics. Figure 2 shows the training dynamics across all three conditions. The MATH-500 validation curves (left panel) demonstrate that R-MEL achieves the highest final performance (57.8%), with steady improvement throughout training. All conditions show similar reward improvement trajectories, starting near -0.9 and improving toward -0.3 to -0.5.

The meta-experience quality, measured by p_{keep} (fraction of meta-experiences retained after replay validation), remains comparable across conditions: MEL (7.6%), MEL+Extra (7.4%), and R-MEL

Table 2: Ablation study on R-MEL components. Removing the prefix constraint ($\tau = 0$) or reducing revision budget ($B = 1$) both degrade performance, validating the design choices.

Method	Config	Avg Pass@1	Δ
MEL	–	33.00	-0.17
MEL+Extra	–	**33.31**	$+0.14$
R-MEL (full)	$\tau = 0.3, B = 4$	33.17	–
R-MEL w/o prefix	$\tau = 0, B = 4$	32.69	-0.48
R-MEL ($B = 1$)	$\tau = 0.3, B = 1$	32.80	-0.37

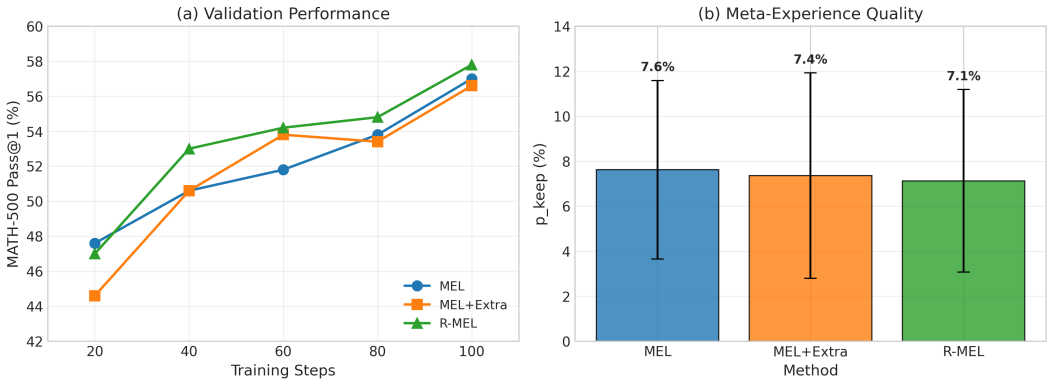

Figure 2: Training dynamics comparison. **(a)** MATH-500 Pass@1 validation performance over training steps. R-MEL (green) achieves the highest final performance (57.8%). **(b)** Meta-experience quality measured by p_{keep} (fraction of meta-experiences retained after replay validation). All methods show comparable p_{keep} ($\sim$7.1–7.6%), indicating revision-derived pairs do not degrade validation quality.

(7.1%). This indicates that revision-derived meta-experiences pass validation at rates similar to naturally-occurring ones, validating that the revision mechanism does not degrade training signal quality. The revision success rate in R-MEL ranges from 2% to 20% throughout training, with a mean of approximately 5.5% of all-negative groups successfully rescued per step.

Difficulty Pattern. Figure 3 reveals an inverted-U pattern in revision success rates with respect to prompt difficulty. Using all-negative frequency as a difficulty proxy (higher frequency indicates harder prompts), we observe that intermediate-difficulty prompts (2/3 all-negative frequency) show the highest revision success rate (15.4%), compared to easy prompts (1/3 all-negative frequency, 7.7%) and hard prompts (3/3 all-negative frequency, 3.3%).

This pattern has an intuitive interpretation: easy prompts rarely produce all-negative groups, so there are fewer opportunities for revision. Hard prompts produce many all-negative groups, but the model struggles to find correct continuations even with revision. Intermediate-difficulty prompts represent the "sweet spot" where the model can find correct continuations but struggles with full solutions from scratch. This insight suggests potential for adaptive revision strategies that allocate more budget to intermediate-difficulty prompts.

5 Conclusion

We presented R-MEL, a method for recovering contrastive signal from all-negative groups in RLVR training through prefix-primed revision. By truncating failed trajectories at candidate bifurcation points and generating correct continuations, R-MEL converts discarded training data into useful meta-experiences. Our experiments demonstrate that R-MEL achieves the highest Avg@8 (30.78) across all conditions and outperforms baselines on 3 of 5 benchmarks, with a statistically significant +4.0 percentage point improvement on MATH-500.

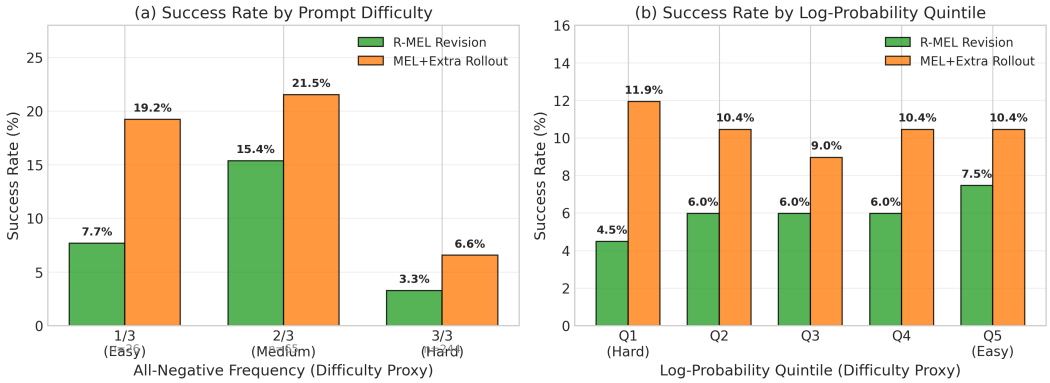

Figure 3: Revision success rate by prompt difficulty. **(a)** Success rate by all-negative frequency (difficulty proxy). R-MEL revision shows an inverted-U pattern: highest success (15.4%) at intermediate difficulty (2/3 all-neg frequency), lower at easy (7.7%) and hard (3.3%) prompts. **(b)** Success rate by log-probability quintile shows similar patterns. MEL+Extra rollout consistently outperforms R-MEL revision across all difficulty levels.

Ablation studies validate two key design choices: the prefix constraint contributes +0.48 to average Pass@1 by ensuring meaningful overlap between revised and original trajectories, and the revision budget of $B = 4$ captures substantially more rescued groups than $B = 1$. Analysis reveals an inverted-U pattern in revision effectiveness, with intermediate-difficulty prompts showing the highest success rate (15.4%), suggesting potential for adaptive revision strategies that allocate budget based on prompt difficulty. Future work may explore such adaptive approaches and extend prefix-primed revision to other domains beyond mathematical reasoning.

REFERENCES

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Learning from mistakes makes llm better reasoner. *ArXiv*, abs/2310.20689, 2023.

Zhipeng Chen, Kun Zhou, Wayne Xin Zhao, Junchen Wan, Fuzheng Zhang, Di Zhang, and Ji-Rong Wen. Improving large language models via fine-grained reinforcement learning with minimum editing constraint. *ArXiv*, abs/2401.06081, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, R. Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, A. Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, C. Deng, Chenyu Zhang, C. Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, JingChang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. Cai, J. Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, K. Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, T. Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, W. Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, W. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, X. Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Y. Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Y. Ou, Yuduan

Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Y. Zha, Yuting Yan, Z. Ren, Z. Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025.

Yunzhen Feng, Parag Jain, Anthony Hartshorn, Yaqi Duan, and Julia Kempe. Don't waste mistakes: Leveraging negative rl-groups via confidence reweighting, 2025. URL `https://arxiv.org/abs/2510.08696`.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Z. Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021.

Shiting Huang, Zecheng Li, Yu Zeng, Qingnan Ren, Zhen Fang, Qisheng Su, Kou Shi, Lin Chen, Zehui Chen, and Feng Zhao. Internalizing meta-experience into memory for guided reinforcement learning in large language models. 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. 2024.

M. Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. Beyond uniform credit: Causal credit assignment for policy optimization. 2026.

H. Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. *ArXiv*, abs/2305.20050, 2023.

Haolin Liu, Dian Yu, Sidi Lu, Yujun Zhou, Rui Liu, Zhenwen Liang, Haitao Mi, Chen-Yu Wei, and Dong Yu. Save the good prefix: Precise error penalization via process-supervised rl to enhance llm reasoning. 2026.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, M. Simens, Amanda Askell, Peter Welinder, P. Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

Rafael Rafailov, Archit Sharma, E. Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023.

Amrith Rajagopal Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. *ArXiv*, abs/2410.08146, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang,

Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Honglin Yu, Weinan Dai, Yuxuan Song, Xiang Wei, Haodong Zhou, Jingjing Liu, Wei Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yong-Xu Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025.

Kai Zhang, Christopher Malon, Lichao Sun, and Martin Renqiang Min. Editgrpo: Reinforcement learning with post -rollout edits for clinically accurate chest x-ray report generation. *ArXiv*, abs/2509.22812, 2025.

A Appendix

APPENDIX TEXT