Training Medical QA Models Based on Mixed Rewards from Multiple-Choice and Open-Ended Questions

Yue Qiu[1], Yujan Ting[2], Pei Dong[1], Terrence Chen[2], Weijing Huang[2]*
[1]Beijing United Imaging Intelligence Co., Ltd., Beijing, China
[2]United Imaging Intelligence, Boston, MA, USA
{yue.qiu, yujan.ting, pei.dong, terrence.chen, weijing.huang}@uii-ai.com

Abstract

Reinforcement learning (RL) for large language models (LLMs) typically requires clear reward signals, which are often unavailable for open-ended (OE) questions where answer evaluation is ambiguous without scalable expert labeling. We investigate whether LLMs benefit from training on mixed data with varying reward clarity. Our approach combines Multiple-choice questions (MCQs), which offer clear binary rewards, with OE questions, for which we use simpler, potentially noisy rewards such as Jaccard similarity or LLM-based evaluators. We hypothesize that MCQs can stabilize training when mixed with OE questions. Our experiments show this mixed-data approach consistently improves medical question-answering performance across model scales.

1 Introduction

Reinforcement learning (RL) has shown promise in enhancing the reasoning capabilities of large language models (LLMs) (Schulman et al., 2017; Yu et al., 2025). RL thrives on clear and consistent reward signals that provide unambiguous feedback. Multiple-choice questions (MCQs) exemplify this ideal scenario by offering binary rewards: answers are either correct or incorrect. This clarity provides a stable learning signal for the model.

However, specialized domains like medicine frequently require open-ended (OE) questions where evaluating answers involves greater complexity. For these questions, defining clear reward signals is challenging. Consider this medical scenario: “What is the first step in the management of a patient with congestive heart failure, type 2 diabetes, altered mental status, and a serum glucose level of 500 mg/dL?” If the ground truth is “IV NS,” and a model answers, “Start IV dextrose-containing fluids,” is this answer entirely wrong, partially correct, or acceptable? While human expert labeling could provide accurate assessments, this approach is expensive and not scalable for on-policy RL, where rewards are dynamically needed during training.

This limitation raises a crucial research question: Can LLMs benefit from training on data with noisy reward signals, and if so, how can such data be effectively utilized? Reward models (RMs) (Su et al., 2025) trained to mimic human expert evaluations provide numerical scores (e.g., 0 to 1), but potentially introducing additional biases, such as length bias (Bu et al., 2025). Alternatively, simpler metrics like Jaccard similarity (Jaccard, 1912) between model outputs and reference answers provide more direct, but potentially noisier, reward signals.

In this work, we investigate a novel strategy that leverages both the stability of clear reward signals from MCQs and the broader coverage of OE questions despite their inherently noisier rewards. We propose mixing MCQs and OE questions within the same training batches. The hypothesis is that the unambiguous feedback from MCQs serves as an anchor, stabilizing the training process, while the model still learns from diverse OE data. We explore various reward mechanisms for OE questions, including Jaccard similarity and LLM-based reward models. However, due to the complexity of medical terms, such as abbreviations and spelling errors, Jaccard can fail to capture semantic equivalence. For instance, “Administer intravenous normal saline” and the ground truth “IV NS” are identical in meaning but receive a score of 0 due to token mismatch. Notably, after training with our mixed-reward strategy, the model still produced the correct answer "Administer intravenous normal saline" to this question, demonstrating robustness even under noisy reward signals.

In summary, our main contribution is demonstrating that RL can benefit from an expanded dataset, even if it includes noisy rewards. We evaluate the performance of our method on several medical QA benchmarks, including MedQA-USMLE,

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8721–8729

MMLU-Pro, and CMB-Exam, to show consistent accuracy improvements. This mixed-data, mixed-reward strategy aims to effectively balance reward signal quality with data diversity, leading to more robust medical QA models.

2 Related Works

Rule-Based Rewards for Clear Feedback Rule-Based Rewards are crucial in Reinforcement Learning (RL) for tasks with clear correctness criteria, offering deterministic feedback that can reduce manual annotation and enhance model safety (Hu et al., 2023; Mu et al., 2024). DeepSeek-R1 (Guo et al., 2025) has shown its effectiveness. We use rule-based rewards for multiple-choice questions (MCQs) due to their clear binary feedback, making it a core part of our reward strategy.

Handling Ambiguity in Rewards of Open-Ended Questions Evaluating responses to open-ended (OE) questions is more challenging, as defining clear rules becomes difficult. Instead, reward models (e.g. medical_o1_verifier_3B (Chen et al., 2024) and RLVR (Su et al., 2025)) are trained on labeled data in the format of <question, answer, ground-truth> to give scores for answers on OE questions. They can be resource-intensive to develop and may introduce their own noise. For example, reinforcement learning only relying on medical_o1_verifier_3B suffers from reward hacking in our preliminary experiments. The examples are shown in Appendix A.1

Mixed-Data Training, Curriculum Learning, and Multi-Task RL Given the availability of both clear rewards from MCQs and graded, noisier rewards from OE questions (e.g., Jaccard similarity), we propose a mixed-data training strategy. This approach is conceptually grounded in principles from curriculum learning (Bengio et al., 2009) and multi-task reinforcement learning (Teh et al., 2017). Our mixed-training approach similarly aims to leverage diverse signal types within a unified learning process. This overall strategy is implemented using the DAPO algorithm (Yu et al., 2025), selected for its effectiveness in policy optimization.

Medical Reasoning LLMs The application of advanced LLMs to the medical domain has seen growing interest. HuatuoGPT-o1 (Chen et al., 2024) was among the first medical LLMs demonstrating complex reasoning, trained using SFT followed by PPO with a medical verifier. MED-RLVR (Zhang et al., 2025) utilizes a rule-based reward with PPO to significantly boost performance on medical MCQs. Our work builds on these efforts by exploring how a mixed-reward strategy, combining reward signals from multiple-choice questions and open-ended questions, can further enhance medical QA capabilities.

3 Methodology

3.1 Reinforcement Learning Algorithm

DAPO for Mixed-Reward Training We adopt the **D**ecouple Clip and Dynamic s**A**mpling **P**olicy **O**ptimization (DAPO) algorithm (Yu et al., 2025) for our reinforcement learning training. DAPO, a variant of GRPO (Shao et al., 2024), is selected not only for its effectiveness in policy optimization, but also for its suitability for mixed-reward training. DAPO covers several features, with the key feature of dynamic sampling particularly suitable for mixed-reward training.

Dyanmic Sampling in DAPO ensures that each training batch contains prompts that provide effective gradients (eliminating zero-gradient groups). This is particularly useful when training with a mix of MCQs (clear rewards) and OE questions (potentially noisy rewards, examples shown in Appendix A.2), as DAPO can dynamically curate batches of appropriate difficulty. This process, where the algorithm adaptively selects data for optimal learning, aligns with the philosophy of curriculum learning and proved effective in our experiments. Our preliminary experiments revealed that datasets composed exclusively of open-ended questions needed more iterations to identify useful training data and form effective learning batches (see Appendix A.3).

3.2 Train Dataset Collection

Our train dataset contains different types and different languages of medical questions, including close-end datapoints from MedQA-USMLE (Jin et al., 2021), MedMCQA (Pal et al., 2022), CMB-Exam (Wang et al., 2023), and open-end datapoints from HuaTuo medical verifiable questions (Chen et al., 2024). We use untrained LLMs to perform 16 rollouts on each question and filter out simple ones with all correct answers. After that, the remaining datapoints can be considered as the difficult and challenging ones.

Model	Dataset	MedQA-USMLE	MMLU-Pro		CMB-Exam	MCQs Avg.	HealthBench-Small
			Health	Biology			
HuatuoGPT-o1-7B		70.7%	59.78%	74.06%	80.75%	71.32%	0.5642
Deepseek-R1-Distill-Qwen-7B		36.06%	30.56%	60.25%	33.3%	40.04%	0.4116
Qwen2.5-3B-Instruct		32.60%	26.28%	52.02%	**67.71%**	44.65%	**0.6659**
	MCQA 18.6k	53.89%	44.5%	65.27%	65.10%	57.19%	0.6202
	Open-ended QA 16k	49.02%	30.44%	22.04%	59.44%	40.24%	0.5623
	Mix 34.6k	**57.50%**	**46.94%**	**67.78%**	66.82%	**59.76%**	0.576
Qwen2.5-7B-Instruct		59.62%	52.69%	70.15%	78.34%	65.2%	0.7286
	MCQA 18.6k	69.13%	57.82%	72.66%	79.53%	69.79%	**0.7385**
	Open-ended QA 16k	68.11%	56.72%	71.27%	76.97%	68.27%	0.7234
	Mix 34.6k	**71.33%**	**61.37%**	**76.85%**	**79.95%**	**72.38%**	0.6934
Qwen3-4B		71.09%	60.88%	**81.45%**	69.59%	70.75%	**0.8693**
	MCQA 18.6k	73.06%	**63.37%**	79.08%	**70.84%**	71.59%	0.8745
	Open-ended QA 16k	69.91%	56.36%	78.1%	70.8%	68.79%	0.778
	Mix 34.6k	**73.61%**	61.74%	81.31%	70.54%	**71.8%**	0.8341

Table 1: Results of medical benchmarks. **Bold** highlights the best accuracy or score among models of the same size.

3.3 Reward Design

For the dataset consisting of multiple-choice questions, we use the following binary rule-based reward:

$$R(\hat{y}, y) = \begin{cases} 1, & \text{is\_equal}(\hat{y}, y) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\hat{y}$ is the predicted option and y is the ground-truth option. We extract predicted options using regex matching on model outputs formatted as [A-D] or [A-D]. content, counting exact matches as correct.

For the dataset with open-ended questions and short answers, we evaluate the responses by Jaccard Similarity (Jaccard, 1912) as follows:

$$R(\hat{y}, y) = \frac{\hat{y}_{\text{tokenize}} \cap y_{\text{tokenize}}}{\hat{y}_{\text{tokenize}} \cup y_{\text{tokenize}}} \tag{2}$$

where $\hat{y}$ means the predicted response and y means the ground-truth answer, and y_{tokenize} means the word set after tokenization (Bird, 2006):

$$y_{\text{tokenize}} = \text{word\_tokenize}(y) \tag{3}$$

Rule-based rewards provide clear and discrete feedback, making it easy to identify completely correct v.s. incorrect responses. However, partially correct answers can still be informative for learning. The Jaccard Similarity score, which ranges from 0 to 1, offers a softer and more nuanced reward signal. This is especially valuable for open-ended questions, where exact matches between predictions and ground truth are uncommon. Jaccard Similarity can provide continuous and fine-grained feedback, expand the reward space, and support smoother model training.

4 Experiments

4.1 Datasets

Training Data We study the impact of different data types with different reward strategies. For closed-ended MCQs in English and Chinese, we sample 9,000 and 9,600 examples from the challenging dataset described in Section 3.2, respectively. To maintain balance, 16,000 open-ended questions are sampled.

Benckmarks The experiments are performed using a widely adopted medical benchmark dataset, including MedQA-USMLE, CMB-Exam, and health and biology tracks of MMLU-Pro. We use a 4-choice concise version of MedQA-USMLE, which contains 1,273 questions. The test set of CMB-Exam consists of 11,200 multiple-choice and multiple-answer questions. In order to verify the model's generalization ability outside the training distribution, we evaluate the performance on health and biology categories in MMLU-Pro (Wang et al., 2024). The dataset contains 818 and 717 multiple-choice questions, respectively. We use accuracy as the evaluation metric. For multiple-answer questions in the CMB-Exam, only questions with exactly matching answers are counted.

To evaluate the model's performance on open-ended medical questions, we extract 100 dialogue samples from the HealthBench (Arora et al., 2025)

Model	Dataset	MedQA-USMLE	MMLU-Pro		CMB-Exam	MCQs Avg.	HealthBench-Small
			Health	Biology			
Qwen2.5-7B-Instruct	Mix 34.6k (RM)	69.68%	61.61%	74.62%	78.85%	71.19%	0.7234
	Mix 34.6k (Binary)	71.25%	61.25%	73.36%	79.44%	71.33%	0.7448

Table 2: Results of using a reward model (RM) and binarized Jaccard Similarity (Binary) on open-ended questions in the mixed dataset.

Model	Dataset	HuaTuo Medical Verifiable Testset	Mean Tokens per Answer
Qwen2.5-7B-Instruct		49.4%	70.20
	MCQA 18.6k	54%	50.25
	Open-ended QA 16k (Binary)	50.1%	6.64
	Mix 34.6k (Binary)	54.4%	6.64
	Mix 34.6k (RM)	60.1%	368.24

Table 3: Supplementary evaluation on HuaTuo medical verifiable dataset with different reward strategies for open-ended questions

dataset and construct a subset named HealthBench-Small. Following the official OpenAI evaluation protocol, we employ Qwen3-32B (Yang et al., 2025) as the evaluation model with score as the evaluation metric.

We use temperature 1.0 for evaluation and training parameters in Appendix A.4.

4.2 Main Results

The evaluation results in Table 1 are divided into four parts based on rows. The first part is represented by models such as HuatuoGPT-o1-7B (Chen et al., 2024) and Deepseek-R1-Distill-Qwen-7B (Guo et al., 2025). The long responses generated by Deepseek-R1-Distill-Qwen-7B make it difficult to extract valid options within an 8,192-token response length, which affects evaluation performance. The second and third parts present results across three types of datasets: an English-Chinese multiple-choice dataset, an open-ended verifiable QA dataset, and a mixed dataset combining both. For both Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2024), training on the mixed dataset leads to better performance than using only multiple-choice or only open-ended data on every benchmark. We also conduct an experiment on one of the newly open-sourced Qwen3 series models. Qwen3-4B outperforms the larger Qwen2.5-7B on most of the benchmarks in the last part of the table, indicating the promising potential of the Qwen3 series. The HealthBench dataset consists of doctor-patient conversations, but our training set lacks data in this specific dialogue format. This discrepancy may result in the observed decline in performance on the HealthBench benchmark.

We expand our study to compare different reward strategies in Table 2. In reinforcement learning, rewards for open-ended questions are typically computed using a reward model. We employed an open-source Reward Model(RM) named Qwen2.5-7B-RLVR (Su et al., 2025), instead of Jaccard Similarity, to compute rewards for open-ended questions. The reward model judges the response and generates a "YES" or "NO" output, with "YES" counted as 1 and "NO" as 0. Our experiments found comparable performance between Jaccard similarity and RM on MCQ evaluation. However, Jaccard similarity offers computational efficiency without requiring domain-specific training. And training with a reward model increases computational demands and runtime. In the second line of the table, we apply a threshold to the Jaccard Similarity score to convert it into a binary value. Scores below 0.6 are mapped to 0 and others are mapped to 1. The evaluation results show that this hard binarization of rewards performs similarly to using Jaccard Similarity.

4.3 Supplementary Open-Ended Evaluation

To better understand the impact of our mixed-reward approach on open-ended questions, we conducted supplementary experiments using 1,000 held-out questions from the HuaTuo medical verifiable dataset (Chen et al., 2024). Table 3 presents

Model	Dataset	MedQA-USMLE	MMLU-Pro		CMB-Exam	MCQs Avg.	HealthBench-Small
			Health	Biology			
Qwen2.5-7B-Instruct	MedQA 9k + CMB 9.6k	69.13%	57.82%	72.66%	79.53%	69.79%	0.7385
	MedQA-OE 9k + CMB 9.6k	60.49%	54.88%	71.55%	79.25%	66.54%	0.7205
	MedQA-OE 9k + CMB-OE 9.6k	58.92%	53.67%	68.48%	71.35%	63.11%	0.6597

Table 4: Ablation study results of using different combinations of MCQs and open-end (OE) questions.

these results, with GPT-4o serving as the evaluation model.

Our results demonstrate that the mixed training strategy significantly outperforms training exclusively on open-ended questions, a 4.3% absolute improvement. This confirms that incorporating MCQs with clear binary rewards provides stabilizing signals that enhance learning even for open-ended tasks.

The reward model baseline (Mix 34.6k RM) achieves higher scores but generates substantially longer responses (368.24 tokens v.s. 6.64 tokens for Jaccard similarity). This difference in response length reveals a critical evaluation artifact: GPT-4o-based evaluation exhibits length bias, favoring verbose responses regardless of content quality (Zheng et al., 2023). The excessive verbosity from RM-trained models suggests reward hacking, where models learn to exploit evaluation biases rather than improve answer quality.

The performance decline on HealthBench-Small warrants explanation. Our training data consists of direct QA pairs with short, factual responses, while HealthBench contains multi-turn doctor-patient dialogues requiring empathetic, conversational responses. This format mismatch, compounded by our Jaccard similarity reward favoring concise answers, explains the reduced performance on dialogue-based evaluation. This limitation highlights the challenge of generalizing across diverse medical communication formats.

Given these evaluation complexities, we prioritize MCQ benchmarks as more objective measures of medical knowledge. The clear correctness criteria and binary evaluation eliminate the confounding factors present in open-ended evaluation, such as length bias and stylistic preferences. Nevertheless, our supplementary results confirm that the mixed-reward approach benefits both question types, validating our central hypothesis that combining varying reward signals enhances overall model capability.

4.4 Ablation Study

To better understand the impact of question types on performance, we conduct an ablation study by systematically varying the composition of our training data. We convert subsets of the English (MedQA) and Chinese (CMB) MCQs training data described in Section 4.1 into open-ended formats, and explore different combinations of these datasets. We use rule-based rewards for MCQs and Jaccard similarity for OE questions. Table 4 presents these results. While training exclusively with MCQs (first row) achieves the highest performance, the mixed MCQ/OE approach significantly outperforms training on purely open-ended questions (second and third rows). This suggests that MCQs training not only enhances accuracy, but also contributes to more stable model behavior.

5 Conclusion

In this paper, we demonstrate the effectiveness of a mixed training approach combining English-Chinese multiple-choice questions and open-ended QA data, using rule-based rewards and Jaccard similarity with DAPO for reinforcement learning. This strategy consistently outperforms single-dataset approaches on most of the benchmarks across both 3B and 7B models. The performance of the newly released Qwen3-4B model aligns with the above conclusion while surpassing the larger Qwen2.5-7B, indicating that the new model is more efficient and powerful. Our approach strikes an optimal balance by leveraging the training stability of MCQs while still exposing the model to the diverse reasoning patterns essential for open-ended medical questions. This suggests a promising direction for future development.

Limitations

Due to constraints in computational resources and time, we were unable to experiment with larger or more recent model families. Our current findings are therefore limited to a single model line, and further validation on diverse architectures would strengthen the generality of the conclusions. It also remains worth exploring how the performance gains from the mixed dataset scale with model size, and where the upper bound of this approach may lie.

In addition, while our study targets multilingual and multi-type medical question answering, including English–Chinese multiple-choice and open-ended verifiable QA, real-world medical applications involve more varied formats. Notably, the HealthBench medical dialogue data differ structurally from our chosen QA settings, which may introduce a format mismatch that limits direct applicability. Beyond these design choices, our reward modeling emphasizes robustness but does not exhaustively address reward noise. Future work could investigate alternative noise characterization methods, especially for long-sequence tasks such as clinical report generation or multi-turn dialogues, where the stability of rewards is particularly critical.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang. 2025. Beyond excess and deficiency: Adaptive length bias mitigation in reward models for rlhf. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3091–3098. Association for Computational Linguistics.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Bin Hu, Chenyang Zhao, Pu Zhang, Zihao Zhou, Yuanhang Yang, Zenglin Xu, and Bin Liu. 2023. Enabling intelligent interactions between an agent and an llm: A reinforcement learning approach. *arXiv preprint arXiv:2306.03604*.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv e-prints*, pages arXiv–2503.

Yee Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. 2017. Distral: Robust multitask reinforcement learning. *Advances in neural information processing systems*, 30.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and 1 others. 2023. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. 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*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

A.1 Reward Hacking in 3B Verifier

We identified a concerning pattern where models achieve near-perfect RM scores (0.98+) by simply repeating the question without providing any actual answer. This represents a fundamental failure where the reward signal completely misaligns with actual utility. Below are reward hacking examples in the reward model medical_o1_verifier_3B:

> **Question**: A 24-year-old male developed a hyperpigmented patch on his right upper chest four years ago, which later showed thick hair growth. What is the diagnosis for this condition?
> **Predicted Answer**: A 24-year-old male developed a hyperpigmented patch on his right upper chest four years ago, which later showed thick hair growth. What is the diagnosis for this condition?
> **Ground Truth Answer**: Becker's nevus
> **Reward Model Score**: 0.9844

> **Question**: In a patient suspected of being diagnosed with Rabies, a sample of corneal smear was taken. Which investigation can be performed directly on the corneal smear to detect the presence of rabies virus antigen?
> **Predicted Answer**: In a patient suspected of being diagnosed with Rabies, a sample of corneal smear was taken. Which investigation can be performed directly on the corneal smear to detect the presence of rabies virus antigen?
> **Ground Truth Answer**: Immunofluorescence test
> **Reward Model Score**: 0.9648

A.2 Noisy Rewards

Reward signals produced by Jaccard similarity on open-ended dataset are considered "noisy" because semantically correct and well-reasoned responses can sometimes receive low or zero rewards due to surface-level mismatches. For instance, "Peutz-Jeghers syndrome" receives only 0.3333 similarity score compared to the ground truth "Peutz-Jegher syndrome," despite being correct. Similarly, "Penicillamine" receives a 0.0 score against "Pencillamine." Here are some cases:

A good case:

> **Question**: Analyze the transition of a curve from Blue to Red. What will happen to the Sensitivity and Specificity as a result of this change?
> **Predicted Answer**: Sensitivity and Specificity will both increase.
> **Ground Truth Answer**: Both Sensitivity and Specificity increase.
> **Jaccard Similarity Score**: 0.8333

Bad cases:

> **Question**: What is the most probable diagnosis for a female patient who presents with pigmentation of the lips and oral mucosa along with intestinal polyps, and has a family history of the same condition?
> **Predicted Answer**: Peutz-Jeghers syndrome
> **Ground Truth Answer**: Peutz-Jegher syndrome
> **Jaccard Similarity Score**: 0.3333

> **Question**: What is the appropriate treatment for a 52-year-old man presenting with jaundice, extrapyramidal symptoms, and a finding consistent with Kayser-Fleischer rings on ophthalmic examination?
> **Predicted Answer**: Penicillamine
> **Ground Truth Answer**: Pencillamine
> **Jaccard Similarity Score**: 0.0

We demonstrate that despite this inherent noise in open-ended question rewards, combining them strategically with clean binary rewards from MCQs can still improve overall performance. This addresses our core research question of whether and how LLMs can effectively learn from imperfect reward data.

A.3 Number of Generated Batches

In DAPO, dynamic sampling filters out data where all scores within a group are either 1 or 0. New samples keep generating until the number of valid data points reaches the training batch size. Figure 1 shows the number of generations required when training with different datasets on Qwen2.5-7B-Instruct. The horizontal axis represents the normalized number of training steps, corresponding to the training progress. The number of generated batches increases in experiments using MCQs datasets. This indicates many groups of data in a batch consist entirely of the same labels (all scores of 1 or 0), which reduces the data utilization. Training with open-ended datasets results in fewer generated batches, suggesting that the reward scores within each group are more diverse and less deterministic. Experiments conducted with mixed datasets demonstrate improved data utilization and enhance batch learning efficiency.

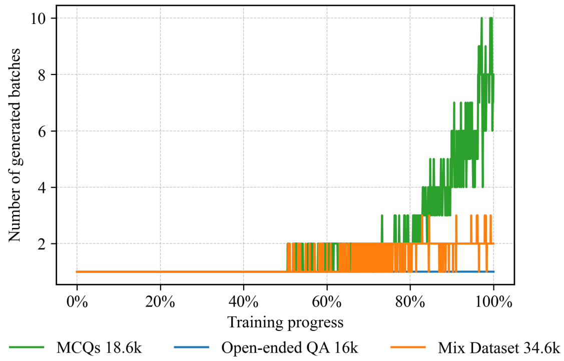

Figure 1: Number of generated batches with training progress

A.4 Parameter Setting

The parameter settings used in our train and evaluation are in Table 5. The inference engine employed is vLLM (Kwon et al., 2023) and the training framework is verl (Sheng et al., 2024).

Table 5: Training Parameters

Parameter	Value
use_kl_loss	False
kl_loss_coef	0.0
filter_groups_metric	score
clip_ratio_low	0.2
clip_ratio_high	0.28
clip_ratio_c	10.0
lr	1e−6
n_resp_per_prompt	16
weight_decay	0.1
offload	True
param_offload	True
optimizer_offload	True
gpu_memory_utilization	0.5
train_prompt_bsz	32
gen_prompt_bsz	96
max_response_length	1024
temperature	1.0
top_p	1.0
do_sample	True
enable_overlong_buffer	True
overlong_buffer_len	64
overlong_penalty_factor	1.0