Curriculum Learning Query Rewriting Optimization Based on Reinforcement Learning

Wenbo Tong *

International College, Hebei University, Baoding, Hebei, 071000, China

* Corresponding Author Email: 20232605096@stumail.hbu.edu.cn

Abstract. Query rewriting is the key to improving the effectiveness of information retrieval, but existing reinforcement learning (RL) based methods often face the challenges of unstable training and slow convergence when the difficulty of the query is uneven. Inspired by the human learning process of "from easy to difficult", curriculum learning (CL), as a general training strategy that can improve the speed of model generalization and convergence, provides new ideas for solving this problem. This study uses CL to provide a sample scheduling strategy for RL training, with the core being a three-level adaptive curriculum learning scheduler. The scheduler dynamically adjusts the mixing ratio of extracting samples of different difficulty levels from the dataset based on the real-time performance (reward) of the model during training, thereby achieving a progressive training strategy for RL models from easy to difficult. Experiments on the HotpotQA dataset show that compared with the random sampling baseline, this method improves the reward value, average performance and training stability by 17.4%, 24.9% and 51.9% respectively, which preliminarily verifies the effectiveness of CL in natural language processing tasks.

Keywords: Query rewriting, reinforcement learning, curriculum learning, HotpotQA.

1. Introduction

In information retrieval and retrieval enhancement generation, rewriting the user's relatively vague original query into a clearer and retrieval-friendly expression helps to narrow the gap between the input text and the real retrieval needs [1]. Although the RL-based approach shows potential in this task, its application faces two bottlenecks: On the one hand, the query difficulty in the real scene is significantly different, from simple entity queries to complex problems requiring multi-step reasoning. The traditional uniform sampling training strategy ignores this difference, resulting in insufficient learning on difficult samples and over fitting on simple samples, resulting in uneven learning [2]; On the other hand, the reward model in human feedback reinforcement learning may have problems such as false generalization and improper setting under limited preference data, which aggravates the unstable training and sparse reward [3], especially when dealing with complex queries. And the curriculum learning imitates the human progressive learning process, aiming to optimize training through carefully designed data presentation order [4]. The aim of this study is to explore the integration of curriculum learning mechanisms into a reinforcement learning training framework for query rewriting, and to systematically guide the learning process by dynamically adjusting the presentation order and proportion of training samples of different difficulty levels. The goal of the research is to construct a more robust and efficient training paradigm to overcome the limitations of traditional methods and ultimately significantly improve the rewriting quality and retrieval performance in complex query scenarios.

2. Dataset and methods

2.1. Dataset and model

This study uses HotpotQA data sets with significant differences in sample query difficulty [5]. Most of the questions in this dataset can only be answered by integrating multiple Wikipedia paragraphs. Each sample is usually equipped with a multi-segment context (common settings include several interference segments), and marked with supporting facts (which sentences are really used for

reasoning), so as to facilitate interpretability and evidence positioning. Qwen2.5-coder-1.5b-instrument was used as the baseline model.

2.2. Method framework

This research proposes a three-module framework including difficulty assessment, curriculum scheduling, and RL trainer. Through these three processes, the CL model is trained, and then compared with the baseline model without adaptive curriculum scheduling to show the performance difference of the CL model. The overall process is shown in Figure 1.

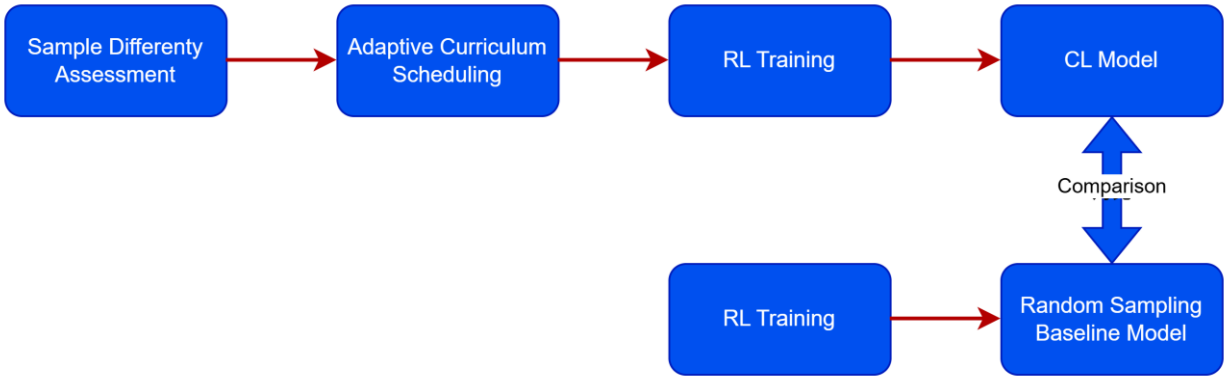

Fig. 1 Flow Chart of Curriculum Learning Performance Experiment (Picture credit: Original)

2.2.1. Difficulty assessment

In order to quantify the difficulty, the baseline model was used for zero sample evaluation, and the success rate of correctly answering the questions by the model was taken as the "difficulty score". According to the score, the query is divided into three levels: simple (success rate ≥ 0.7), medium (success rate between 0.3 and 0.7), and difficult (success rate<0.3). The final sample distribution is 30% simple, 40% medium and 30% difficult (as shown in Figure 2), which provides a basis for curriculum design.

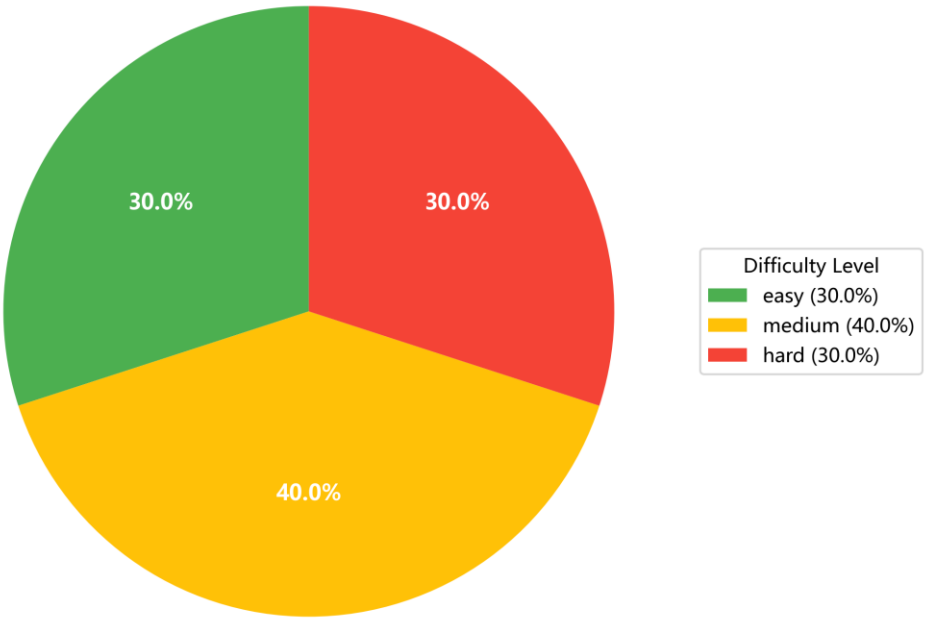

Fig. 2 Sample Difficulty Level Distribution Pie Chart (Picture credit: Original)

2.2.2. Curriculum learning scheduler

Design a three-level progressive curriculum with dynamically adjusted difficulty blending ratios, as shown in Table 1.

Table 1. Difficulty Mixing Ratio of Three Stage Samples

Stage	Simple Proportion	Medium Proportion	Difficult Proportion	Purpose
Stage 1	80%	10%	10%	laying the foundation
Stage 2	60%	30%	10%	lifting capacity
Stage 3	40%	30%	30%	adaptation to real distribution

When the average reward of the model continuously reaches the preset threshold (0.7) in the current stage, it can be considered that the model has strong performance in that stage and automatically advances to the next stage.

2.2.3. RL trainer and reward function design

Based on the Qwen2.5-Cod 1.5B model, Proximal Policy Optimization (PPO) is used to update the strategy of the LLM. For each sample, the current model samples and generates rewritten text, which is then rewarded with a scalar reward module. A clip is then applied to the ratio of the probability of the old strategy frozen at the time of sampling to limit the update step size. Add value headers as baselines to the language model during implementation and coordinate with the KL penalty to freeze the reference model relatively, in order to stabilize training. This type of method is widely used in the research and practice of reinforcement learning to enhance language models [6].

The reward function is applicable to samples of different difficulty levels, taking into account five dimensions: semantic similarity, keyword coverage, query length suitability, specificity, and fluency, as shown in Table 2. The reward sizes for simple, medium, and difficult samples are also differentiated. As shown in Table 3.

Table 2. Five-Dimension Settings of Reward Function

Dimension	Meaning	Specific Indicator	Weight
Semantic Similarity	Does the rewritten query still refer to the same thing in meaning as the original query	F1 of the two sets of keywords after removing the stop word	0.3
Keyword coverage rate	Whether to keep the important information words in the original query	The retention rate of the original query keywords is mainly, and a small amount of overlap with the answer keywords can be mixed when the words can be extracted	0.3
Query length suitability	Whether the rewritten length falls within a reasonable range	Ideal interval 6 – 22 words: Full Score; Too short or too long obviously reduces the score	0.2
Specificity	Whether it is specific enough to facilitate retrieval	Number of digits + predefined hits of specific English words	0.1
Fluency	Whether it is like a natural and available retrieval method	Points will be deducted for repeated punctuation, extra space before punctuation, and adjacent repeated words	0.1

Table 3. Reward Settings for Samples of Different Difficulty Levels

Difficulty Level	Reward ratio	explanation
easy	×1.0	Multiply the weighted score of the same subitem by the coefficient and truncate to [0,1]
medium	×1.2	
difficult	×1.5	

In order to evaluate the effectiveness of curriculum learning, the baseline method uses the same model and reward function, but the training samples are evenly and randomly sampled from the whole data set, without considering the difficulty difference. This comparative design ensures the fairness of the comparison and enables this study to separate the independent effects of curriculum learning strategies.

3. Experimental results and analysis

The experimental results show that under the current model and reward function settings, the CL method training model has significantly improved the training performance and stability compared with the random sampling baseline model. As shown in Table 4.

Table 4. Effect of CL method training model

Reward value increased	Relative performance increased	Stability increased
+17.4%	+24.9%	+51.9%

3.1. Training reward comparison

As shown in Figure 3, in the 10 training epochs, the reward curve of the CL method is always higher than the random sampling baseline. CL received high rewards at an early stage, indicating that starting with simple samples helps to quickly establish the foundation. The performance remained stable after the third and sixth rounds (stage transition point), and the final reward exceeded the baseline by 17.4%.

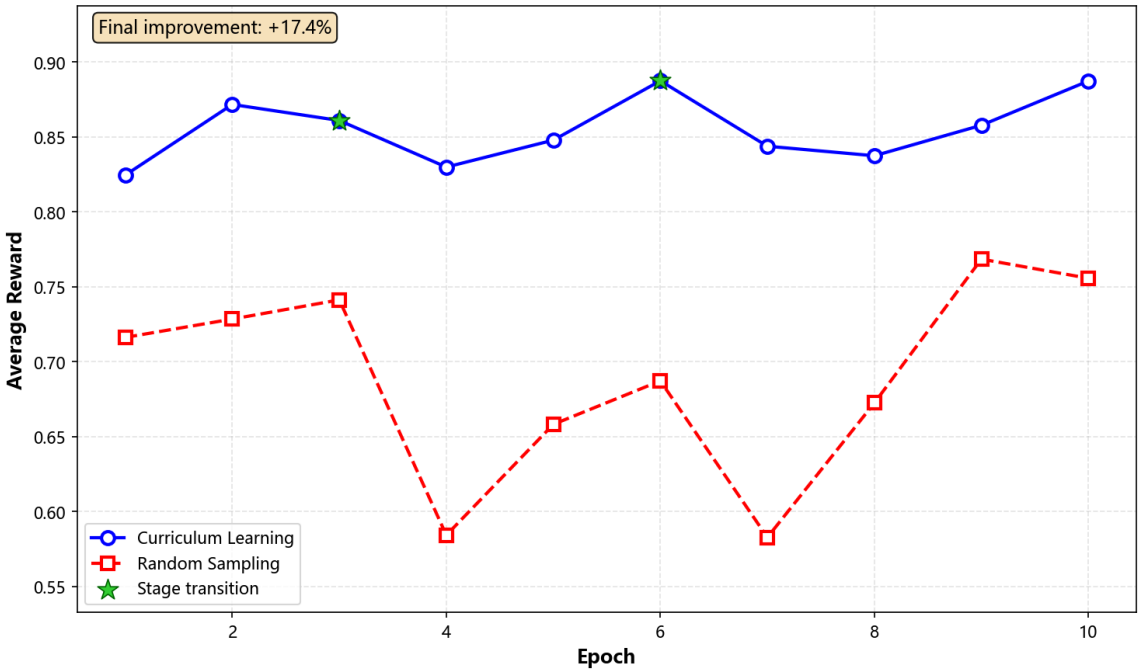

Fig. 3 Comparison of Training Rewards (Picture credit: Original)

3.2. Relative performance improvement

As shown in Figure 4, the relative performance improvement of CL in each round is more than 10%, with an average improvement of 24.9%. The performance of the fourth epoch (+42.1%) and the seventh epoch (+44.8%) was significantly improved, and the corresponding stages were advanced, showing the effectiveness of adaptive adjustment.

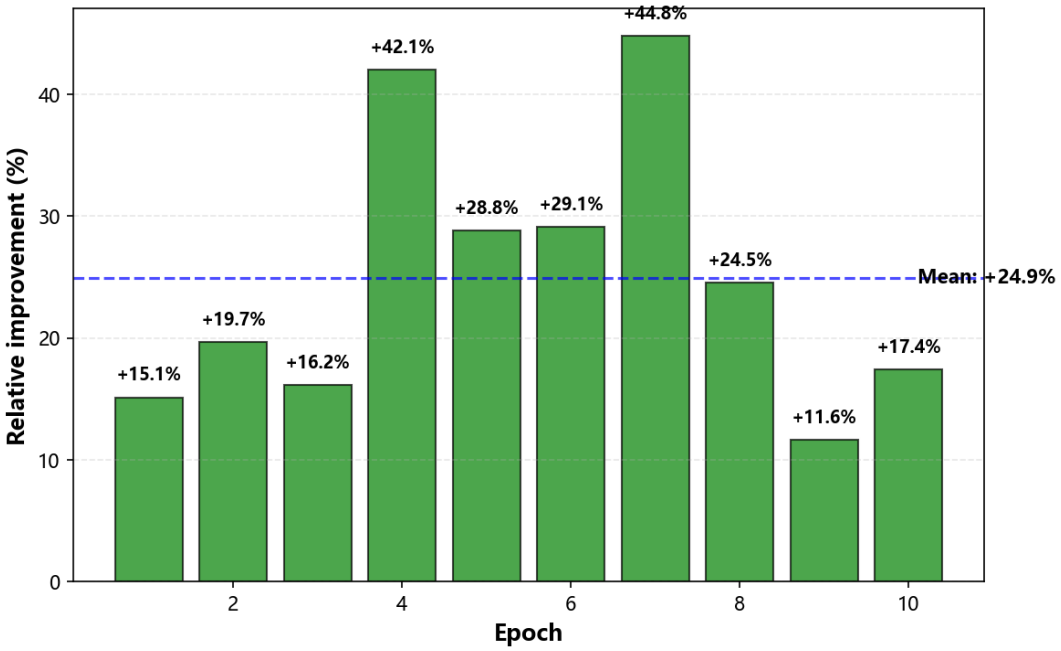

Fig. 4 Relative Performance Improvement Effect (Picture credit: Original)

3.3. Training stability improvement

As shown in Figure 5, the reward standard deviation of CL (solid blue line) is generally lower than the baseline (dashed red line), indicating a more stable training process. In the end, the stability of CL improved by 51.9% compared to the baseline, indicating that the CL strategy helps alleviate the instability of RL training.

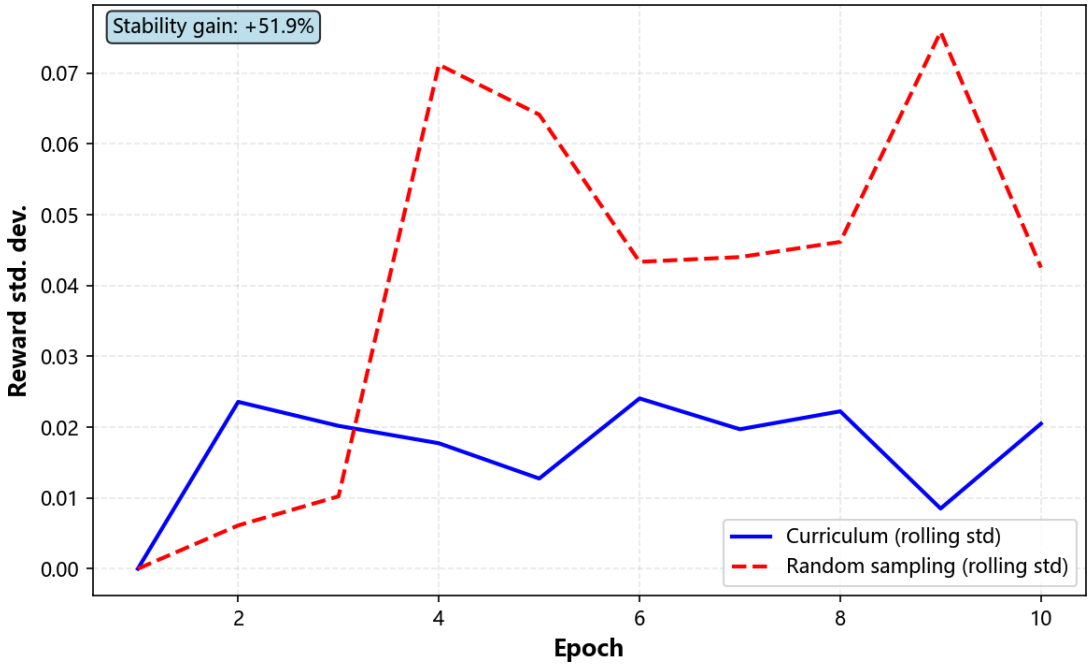

Fig. 5 Comparison of Training Stability (Picture credit: Original)

3.4. Reasons for the effectiveness of cl

The experimental results of this study confirm the effectiveness of curriculum learning in query rewriting tasks from multiple dimensions. This article believes that there are three main reasons for its advantages:

Firstly, progressive learning: The strategy of starting with simple samples reduces the initial learning difficulty, allowing the model to quickly establish basic mapping relationships. Avoiding the phenomenon of "learning shock" caused by facing too many difficult samples during the initial training stage of the model. Second, adaptive adjustment: the stage promotion mechanism based on performance threshold ensures that the curriculum progress matches the actual ability of the model. When the model is stable under the current difficulty distribution, more challenging samples are automatically introduced. This dynamic adjustment avoids the mismatch problem that may occur in fixed curricula. Xi et al. further pointed out that pure phased training may have performance fluctuations during stage transition, and mixing the starting points of different exploration difficulties can help smooth the transition and stabilize the training [7]. Thirdly, difficulty control: Compared with random sampling, the curriculum learning explicitly considers the differences in sample difficulty, allowing the model to obtain more balanced learning opportunities on different difficulty queries.

3.5. The timing for stage transition

The transition of the curriculum stage requires precise timing. Premature transition may lead to insufficient model preparation, while late transition may result in low training efficiency. As shown in Figure 4, the performance of the fourth round of training has significantly improved (42.1%), which corresponds to the transition from stage 1 to stage 2, indicating that the timing of the transition is crucial. Parashar et al.'s study also pointed out that fading out of simple tasks at the appropriate time is important to prevent overfitting [8].

3.6. Value of stability

The improvement of training stability has important practical significance: it can reduce the cost of parameter tuning and minimize repeated experiments. Usually, it also means that the learned strategies are more robust and have stronger generalization ability, and lay the foundation for the model to continuously adapt to new data distributions. The theoretical analysis of ADARFT shows that the learning signal is strongest when the success rate of the model approaches a specific target [9], which indirectly supports the rationality of setting a threshold of 0.7 in this study.

3.7. Limitations and future directions

Although this study has achieved positive results, there are still some limitations that point to future research directions:

Firstly, there is a limitation on the sample size. The current experiment only uses 5000 samples, which can verify the effectiveness of the method, but a larger dataset may reveal more patterns. In the future, it is necessary to validate the scalability of methods on larger datasets. Secondly, difficulty assessment. The difficulty assessment based on the zero-sample accuracy of a specific model may be affected by the ability bias of the model itself [10]. In the future, difficulty standards for multi-model evaluation or manual annotation can be explored. Thirdly, design the reward function. Although the current reward function considers multiple dimensions, the weight setting is based on experience. Adaptive weight adjustment or learning based reward design may be a future direction for improvement. In addition, this experiment only set a fixed context length. In the future, difficulty curricula can be combined with context/sequence length curricula, drawing on Song et al.'s phased scaling of context in curriculum RL, to further stabilize long context strategy learning while controlling training costs [11].

4. Conclusion

This study proposes a reinforcement learning framework that integrates curriculum learning and proximal strategy optimization for query rewriting tasks: Through a three-level adaptive curriculum scheduler, the mixing ratio of samples with different difficulty levels is dynamically adjusted in stages during the training process, and combined with a performance threshold based stage transition mechanism, the model first establishes a stable rewriting strategy on easier samples, and then gradually transitions to more difficult and closer to real retrieval scenarios, thereby alleviating the problem of training instability under sparse rewards and high variance gradient conditions. Experiments on the HotpotQA dataset have shown that compared to the baseline using uniform random sampling under the same model and reward function design, this method not only achieves better results in final rewriting quality and cross-round average performance, but also has a smoother and less fluctuating training curve, reflecting better training stability and reproducibility. This indicates that the curriculum-based sample arrangement and threshold push can effectively match the "current model capability" and "training data difficulty", providing a feasible reference for query rewriting in retrieval enhancement scenarios.

References

[1] Ma, X., Gong, Y., He, P., et al. (2023) Query rewriting in retrieval-augmented large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 5303-5315.

[2] Peng, W., Li, G., Jiang, Y., et al. (2024) Large language model based long-tail query rewriting in taobao search. Companion Proceedings of the ACM Web Conference 2024, 20-28.

[3] Chaudhari, S., Aggarwal, P., Murahari, V., et al. (2024) RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. arXiv preprint arXiv:2404.08555. Retrieved from https://arxiv.org/abs/2404.08555

[4] Wang, X., Zhou, Y., Chen, H., et al. (2024) Curriculum learning: Theories, approaches, applications, tools, and future directions in the era of large language models. Companion Proceedings of the ACM Web Conference 2024, 1306-1310.

[5] Yang, Z., Qi, P., Zhang, S., et al. (2018) HotpotQA: A dataset for diverse, explainable multi-hop question answering. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2369-2380.

[6] Wang, S., Zhang, S., Zhang, J., et al. (2024) Reinforcement learning enhanced llms: A survey. arXiv preprint arXiv:2412.10400. Retrieved from https://arxiv.org/abs/2412.10400

[7] Xi, Z., Chen, W., Hong, B., et al. (2024) Training large language models for reasoning through reverse curriculum reinforcement learning. arXiv preprint arXiv:2402.05808. Retrieved from https://arxiv.org/abs/2402.05808

[8] Parashar, S., Gui, S., Li, X., et al. (2025) Curriculum reinforcement learning from easy to hard tasks improves LLM reasoning. arXiv preprint arXiv:2506.06632. Retrieved from https://arxiv.org/abs/2506.06632

[9] Shi, T., Wu, Y., Song, L., et al. (2025) Efficient reinforcement finetuning via adaptive curriculum learning. arXiv preprint arXiv:2504.05520. Retrieved from https://arxiv.org/abs/2504.05520

[10] Zhou, H., Huang, H., Zhao, Z., et al. (2026) Lost in benchmarks? rethinking large language model benchmarking with item response theory. Proceedings of the AAAI Conference on Artificial Intelligence, 40(41), 35085-35093.

[11] Song, M., Zheng, M., Li, Z., et al. (2025) Fastcurl: Curriculum reinforcement learning with stage-wise context scaling for efficient training r1-like reasoning models. arXiv preprint arXiv:2503.17287. Retrieved from https://arxiv.org/abs/2503.17287