# Self-Reflection and Error-Correction Mechanisms in LLMs Driven by Reinforcement Learning

Longkai Liu *

School of Management, Northeastern University, Qinhuangdao, Hebei, China

* Corresponding Author Email: liulongkai@mails.neu.edu.cn

**Abstract.** Large language models have shown very strong abilities on complex reason work, but they still have many problems, these problems include error accumulation and hallucinations. Therefore, letting these models have the ability of self-reflection and error correction has hence become a core research focus. This research puts focus on the self-reflection mechanisms which are pushed forward by reinforcement learning, and it carries out a systematic review on representative research work in this domain. The covering scope contains emergent thinking manifestation, clear teaching practice, many-circle improvement, multi-angle comment, many-intelligence cooperation, and changeable framework designs. Comparison analysis makes clear that reinforcement learning optimizes the decision-making policy through balancing gradient attribution, hence supervised fine-tuning is easy to appear behavioral collapse and distribution shift. The ability of reflection may either come forth in the process of RL training, or be explicitly constructed as a module that can be learned. The research work that will be done in future therefore needs to carry out further exploration on the interpretability of reflection, the design of reward, and the transferability across different tasks. This research offers a systematic theory frame and practical direction for the comprehension and building of self-correcting LLMs.

**Keywords:** Reinforcement Learning; Large Language Models; Self-Reflection; Gradient Attribution.

## 1. Introduction

Large language models (LLMs) have shown very outstanding abilities in recent years in many fields like mathematical logic inference, program making, and multi-step question answering [2][3]. An important restriction still exists, however, for these models have the tendency to produce false contents and mistake accumulation; An earlier mistake in reasoning often therefore causes the production of later content which looks believable but is actually wrong in facts [8]. This lack of self-correction hence restricts the secure application deployment of large language models in key use scenarios.

Traditional methods such as prompt engineering and supervised fine-tuning (SFT) provide partial answers, nevertheless they encounter difficulties. The skill of prompt making greatly relies on instructions that are made by hand, meanwhile SFT needs a great deal of marked data, and it has bad generalization ability on samples that are outside distribution [9]. As one other choice, reinforcement learning (RL) gives a frame for optimizing model action by means of reward signals, hence displaying latent possibility for bringing up internal reflection abilities [2][3][9]. The academic circle is actively exploring the mechanisms through which RL causes self-examination, the connections among different method groups, and the design rules for effective reflection training systems.

This present article at first gives definition description to the concept of self-reflection. It then carries out analysis on the core mechanisms via which reinforcement learning promotes self-examination, therefore followed by a methodology-based review of representative research works. This thesis carries out a comparative analysis at its end and also carries out a discussion on directions of future research.

## 2. The Concept of Self-Reflection

Within LLM research, self-reflection generally denotes a process where a model critiques its own outputs and implements improvements. Scholars typically analyze this concept along three dimensions.

The first aspect which is about reflecting on the generation methods well. Reflective behavior may emerge spontaneously during reinforcement learning training or can be modeled by developers through explicit training. The spontaneously emerging reflection is not an explicit training objective but a natural outgrowth from the reinforcement learning optimization process. For example, within the ReSearch framework, a certain model may, after a retrieval failure, come up with such a sentence "I made a mistake" and then proceed to revise its query accordingly [3]. The two - stage DS hypothesis provides a theoretical explanation, showing that this emergence comes from a certain kind of balance in terms of gradient attribution [12]. Conversely, explicit training considers reflection as a separate and learnable skill. For instance the ReflectEvo framework breaks reflection down into a three - stage process "validating failure→locating error→planning correction"[7]. Other methods, like SCoRe, utilize multi-round reinforcement learning to train models for rectifying their own errors [5], and SRPO introduces specific markers in multimodal tasks to generate explicit reflections [10]. There are also some other existing modes. The Reflexion method converts environmental feedback into text summaries stored in memory, thereby creating a kind of "verbal reinforcement learning" that doesn't need to update model weights [9]. The MetaAgent - Rec system employs a multi - agent design, in which there is a judgment agent that carries out meta - evaluation on the policy agent, leading to a collaborative reflection process [1]. From an architectural perspective, E-LLM that enables real-time modification at the token level can have the model insert or delete tokens when generating [11].

The second aspect, which pertains to the triggering conditions that bring about reflection. Reflection can be triggered by external signals or by the internal state of the model. The external triggering factors include things such as retrieval failures, errors detected by the verification module, or feedback from the external compiler. The internal triggering depends on the model's self-assessment of the uncertainty. Some models, when detecting that uncertainty is quite high, will learn to actively refrain from answering, or they may carry out reflections through preset multi-angle checks [4].

The third dimension that has determined the functional scope of reflection. Reflection's role is not only in correcting wrong answers but also shows a broader functional level. Besides correcting mistakes, when the model is uncertain, reflection can guide it to retain the answer or to simplify the correct reasoning process efficiently for the sake of high efficiency. This functional diversity shows that reflection is not a single ability but a range of behaviors that involve from error identification to optimization and metacognitive judgments. This kind of hierarchical perspective that has quite significant advantages has the possibility of optimizing those relevant behaviors under a unified objective function, eliminating the need for separate modeling of each sub-capability.

## 3. Core Mechanisms for Fostering Introspection with Reinforcement Learning

Standard supervised fine-tuning (SFT) maximizes the likelihood of training data, but it suffers from two major drawbacks in self-correction tasks [6]. The first is distributional shift: during training, the model sees corrected examples, but at test time it must fix its own mistakes — leading to a mismatch between training and inference distributions. The second is behavioral collapse: instead of learning a genuine correction strategy, the model takes a shortcut by producing a strong initial answer and skipping subsequent revisions.

Reinforcement Learning (RL) handles distributional shifts by using on-policy sampling to directly optimize the policy. More importantly, RL’s reward signal can encourage "progress" rather than just final correctness. For instance, the SCoRe framework introduces a progress reward that specifically rewards moving from an incorrect state to a correct one [5]. Similarly, SRPO defines a reflection

effectiveness reward that gives positive feedback only when reflection clearly improves the answer quality [10].

The Two Stage Decision Sampling (DS) Hypothesis offers a theoretical explanation based on gradient attribution [12]. It decomposes an LLM's policy into a sampling policy (π_sample) for generating content and a decision policy (π_d) for deciding when to stop or retry. Theoretical analysis shows that RL's advantage function provides balanced gradient updates to both policies. In contrast, the KL divergence penalty in SFT introduces an imbalance due to length weighting: the sampling policy is heavily regularized while the decision policy receives almost no optimization. This imbalance explains why RL can learn the decision policy essential for reflection, whereas SFT typically yields models that mimic the textual form of reflection without acquiring its functional capability.

Existing RL methods for reflection use various technical components. GRPO is widely adopted because it is computationally efficient and eliminates the need for a separate value network [1] [5]. Research like the SCoRE study employs REINFORCE class algorithms for multi round optimization, while DPO is used for learning from preference data. Reward structures have evolved from simple outcome based rewards to specialized ones such as reflection token rewards, progress rewards, and composite rewards that combine effectiveness with length penalties. Reflection representation has also progressed from free text responses to more structured formats, including special tokens and multi stage templates. Table 1 provides a comparative summary of key elements in representative RL based reflection methods.

**Table 1.** Comparison of Key Elements in Representative RL Methods.

| Method | Algorithm | Reward Design | Reflection Representation | Special Mechanism |
|---|---|---|---|---|
| ReSearch (Chen et al., 2025) | GRPO | Result Correctness + Format | Text within <think> tags | Retrieved Result Masking |
| Reflect, Retry, Reward (Bensal et al., 2025) | GRPO | Reflection Tokens Only | Self-Evaluation Text | Reward Only Reflection Tokens |
| SCoRe (Kumar et al., 2024) | REINFORCE | Final Correctness + Progress Reward | Multi-turn Responses | Two-Stage Training |
| SRPO (Wan et al., 2025) | GRPO | Task Reward + Reflection Effectiveness + Length | <reflect> Tag | Multimodal + Reflection Reward |
| Reflexion (Shinn et al., 2023) | None(prompt) | None (Verbal Feedback) | Verbal Summary | Episodic Memory |

What makes RL-driven reflection work is its online optimization and balanced gradient attribution—these allow the decision-making policy to be properly trained. Also, well-designed reward functions push the model toward genuine correction behaviors instead of just copying patterns.

## 4. Categorization and Analysis of Reflection

Single-turn and verbalized reflection represent the early directions in this field. ReSearch first noticed reflective behavior emerging during RL training on multi-hop QA tasks. When retrieval failed, the model would say something like "I made a mistake" and then correct its query. This behavior wasn't explicitly taught—it came purely from RL optimization. The key idea was to embed retrieval operations into the reasoning chain and mask retrieval results when computing the loss, so gradients only affected the thought process and query generation. Reflexion introduced "verbal reinforcement learning": it turns environmental feedback (binary or scalar) into natural language summaries stored in episodic memory. This lets the model learn from past experiences in later attempts without updating weights, showing that language itself can act as a policy gradient. The Two Stage DS Hypothesis then provided a systematic theoretical explanation from a gradient attribution view. By splitting the policy into sampling and decision components, it showed that RL's surrogate reward yields balanced gradients, whereas SFT's KL penalty creates an imbalance that sidelines the decision policy. This theory not only explains why reflection emerges but also guides future method design.

Multi turn reflection and training deepening aim to improve reflection quality through multiple interactions. Reflect, Retry, Reward proposed a novel idea: reward only the reflection tokens, not the correct answers. This forces the model to focus on learning effective reflection itself. Experiments showed that small models trained in this way could outperform models ten times larger on function calling and math tasks. SCoRe systematically analyzed why SFT fails at self correction—distribution shift and behavioral collapse—and proposed a two stage RL solution. Stage one keeps the initial response close to the base model while optimizing the second response. Stage two jointly optimizes both responses and adds a progress reward. This led to significant gains on MATH and HumanEval. ReflectEvo looked into how small language models (SLMs) can bootstrap their own learning using self generated reflection data. They designed a three stage reflection template—verification → localization → planning—to produce high quality data, then fine tuned the model with SFT+DPO. As a result, Llama 3 8B ended up outperforming 70B models on BIG Bench. RLRF proposed a complete reflection learning loop: first, a fine grained feedback model evaluates responses with multi dimensional scoring; second, the model searches for good corrections via self reflection; finally, it fine tunes with DPO. This loop can run over and over, steadily improving performance on factuality and math reasoning.

Moving beyond single dimension reflection, researchers have pushed into multi dimensional reflection and capability extension. PR CoT argues that one dimensional reflection isn't enough to catch all error types, so they introduced multi perspective reflection from four angles: logical consistency, information completeness, ethical bias, and alternative solutions. Experiments show that this works much better than single dimensional reflection on complex tasks like ethical decision making. LLMs can also learn self restraint—extending reflection to "knowing what one knows." The model learns to refuse answering when it's uncertain. By designing a utility function and using iterative search to generate preference data, the model significantly cuts down hallucinations in tasks like biography generation. SRPO extends reflection into the multimodal domain, using reflection effectiveness rewards and length rewards to encourage concise, useful reflections in visual language reasoning. It beats existing methods by a large margin on benchmarks like MathVista.

Other work brings in valuable ideas too. MetaAgent Rec introduces a multi agent architecture: a policy agent (handles recommendations), a critique agent (does meta evaluation), a user simulator, and an LLM meta controller work together to let recommendation strategies evolve on their own. E LLM supports reflection at the architectural level by letting the model insert or delete tokens during generation—turning reflection into editable operations. Training these editing actions with RL lets the model correct outputs in real time. Taken together, these studies trace how reflection capability research has evolved: from observing phenomena to explaining them theoretically, from single approaches to multi dimensional integration, from text only to multimodal settings, and from solo agents to multi agent collaboration. This evolution reflects the field's ongoing push to make LLMs think more like humans.

## 5. Comparison and Analysis

Regarding the method of reflection generation, emergent reflection (e.g., ReSearch) is natural and requires no annotation but offers lower controllability; explicitly trained reflection (e.g., SCoRe, SRPO) is highly controllable but relies on high-quality data; verbalized reflection (Reflexion) is usable without training, but its effectiveness is limited by the base model's capability; multi-agent collaborative reflection (MetaAgent-Rec) achieves functional separation but incurs significant communication overhead; hardware-supported reflection (E-LLM) is the most flexible but has high implementation complexity. In terms of reward design, research has evolved from trajectory-level rewards to reflection token rewards, then to progress rewards, and finally to effectiveness and length rewards. Later designs more precisely guide the model to focus on the quality of reflection itself, not just the final outcome. This evolution reflects a deeper understanding of the essence of "reflection" – it should not be a by-product of generating correct answers but an independently evaluable and

optimizable capability. Concerning reflection dimensions and scenario adaptability, single-dimensional reflection (e.g., ReSearch, Reflexion) suits tasks with clear structures; multi-dimensional reflection (e.g., PR-CoT) shows clear advantages in complex scenarios like ethical decision-making; multimodal reflection (e.g., SRPO) extends to visual-language reasoning; and multi-agent reflection (e.g., MetaAgent-Rec) is more suitable for long-term dynamic environments. Different dimensions have their applicable scenarios; future research might explore adaptive dimension selection mechanisms. Regarding the unity of theoretical explanation and empirical validation, the Two-Stage DS Hypothesis provides theoretical support for RL's superiority over SFT. It's predicted that the decision-making policy is better optimized in RL has been empirically validated in works like SCoRe and SRPO. This closed loop of theory and experiment effectively deepens the understanding of reflection mechanisms.

## 6. Challenges and Future Directions

Even with all the progress so far, RL-driven self-reflection still has plenty of challenges. On the methodological side, most current reflections are written in free-text form, which makes it hard to measure their quality. Future work should look into more structured representations for reflection and combine them with interpretability tools to better understand the cause-and-effect between reflection and correction. Also, existing reward designs tend to be task-specific—like for math or coding—and still lack a general reward framework. Designing reflection reward signals that can work across different tasks remains a tough nut to crack. When it comes to training stability and reliability, people urgently need more stable algorithms for multi-turn training. On top of that, the reflection process itself can introduce new mistakes — for example, wrongly flagging a correct answer as wrong. That means people need defensive mechanisms to keep reflection safe and reliable. As for practicality and scalability, most of today's reflection methods have only been shown to work on the tasks they were trained on. Whether they can transfer to entirely new domains still needs to be checked. Meta-learning might offer a way to solve this transfer problem. Another issue is that methods involving multi-turn RL or multi-agent collaboration are computationally expensive. Future work should combine model compression, efficient sampling, and similar techniques to make them more practical.

## 7. Conclusions

This paper systematically reviews multiple representative works on reinforcement learning-driven self-reflection and error-correction mechanisms in LLMs. The analysis finds that RL fosters reflection by optimizing the decision-making policy through balanced gradient attribution, whereas SFT struggles to achieve the same effect due to distribution shift and behavioral collapse. The capacity for reflection can either emerge during RL training or be explicitly modeled as a learnable module. The methods for generating reflection, along with reward designs and representation forms, have evolved significantly with deepening research. Future studies need continuous exploration in areas such as interpretability, reward generalization, multi-turn stability, safety, and transferability. Enhancing reflective capability is a crucial step in the evolution of LLMs from "passive responders" to "autonomous agents", potentially paving the way for more reliable and human-like AI systems.

## References


[1] Alamdari, P. M., & Zare, G. (2025). Self-Reflective Multi-Agent Reinforcement Architecture for Autonomous Recommendation Policy Evolution.

[2] Bensal, S., Jamil, U., Bryant, C., Russak, M., Kamble, K., Mozolevskyi, D., ... & AlShikh, W. (2025). Reflect, retry, reward: Self-improving llms via reinforcement learning. arXiv preprint arXiv:2505.24726.

[3] Chen, M., Sun, L., Li, T., Sun, H., Zhou, Y., Zhu, C., ... & Chen, W. (2025). Learning to reason with search for llms via reinforcement learning. arXiv preprint arXiv:2503.19470.

[4] Costa, M., Soarez, A. R., Kim, D., & Ferreira, C. (2026). Enhancing Self-Correction in Large Language Models through Multi-Perspective Reflection. arXiv preprint arXiv:2601.07780.

[5] Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., ... & Faust, A. (2024). Training language models to self-correct via reinforcement learning. arXiv preprint arXiv:2409.12917.

[6] Lee, K., Hwang, D., Park, S., Jang, Y., & Lee, M. (2024). Reinforcement learning from reflective feedback (rlrf): Aligning and improving llms via fine-grained self-reflection. arXiv preprint arXiv:2403.14238.

[7] Li, J., Dong, X., Liu, Y., Yang, Z., Wang, Q., Wang, X., ... & Zheng, Z. (2025, July). Reflectevo: Improving meta introspection of small llms by learning self-reflection. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 16948-16966).

[8] Piché, A., Milios, A., Bahdanau, D., & Pal, C. (2024). Llms can learn self-restraint through iterative self-reflection. arXiv preprint arXiv:2405.13022.

[9] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in neural information processing systems, 36, 8634-8652.

[10] Wan, Z., Dou, Z., Liu, C., Zhang, Y., Cui, D., Zhao, Q., ... & Yan, S. (2025). Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. arXiv preprint arXiv:2506.01713.

[11] Wu, Y., Xu, G., & Zou, D. Self-reflection like Humans, Editable-LLM (E-LLM) is All You Need.

[12] Zhao, Z., Zha, Y., Zhang, H., & Xu, X. (2026). The Two-Stage Decision-Sampling Hypothesis: Understanding the Emergence of Self-Reflection in RL-Trained LLMs. arXiv preprint arXiv:2601.01580.