Investigating Iterative Self-Correction of LLMs through Reinforcement Learning and Chain-of-Thought Reasoning

Anonymous ACL submission

Abstract

Iterative self-correction has emerged as a promising inference-time approach to mitigate factual hallucinations in Large Language Models (LLMs). However, prompt-based reflection frequently suffers from multi-turn performance degradation and inconsistent outcomes across various architectures and datasets. In this work, we address this bottleneck by shifting structural and logical behavioral constraints from runtime prompt heuristics directly into the model's weights through Group Relative Policy Optimization (GRPO) reinforcement learning. We use a novel multi-dimensional approach to systematically evaluate the interplay between internalized policy mechanics and inference-time prompting constraints across various reward criteria. Evaluating our framework on extractive Question Answering (QA) benchmarks, we demonstrate that optimizing multi-faceted structural criteria via GRPO successfully insulates the model from degradation during iterative self-correction. Additionally, we uncover a compounding synergy when pairing GRPO with targeted runtime Chain-of-Thought (CoT) guidance, boosting baseline performance by 81.8% (on SQUAD 2.0) and 63.8% (on HotpotQA) under in-domain training. Finally, our analysis confirms that, rather than dataset-specific memorization, GRPO combined with CoT cultivates generalized extractive reasoning mechanics, thereby expanding the model's operational versatility across a more diverse range of downstream applications.

1 Introduction

Despite the remarkable conversational and reasoning capabilities of modern Large Language Models (LLMs), their deployment in high-stakes domains remains bottlenecked by their tendency to generate factual hallucinations (Huang et al., 2025). Hallucinated responses appear plausible, but are ungrounded and, at times, contradictory to external knowledge sources, ultimately spreading misinformation to unsuspecting individuals (Zhang et al., 2025b). To combat this vulnerability, iterative self-correction (or self-reflection) has emerged as a promising inference-time approach, wherein a model critiques and refines its initial outputs over multiple cycles (see Figure 1). Intuitively, the mechanisms underlying self-correction should succeed; literature from different domains emphasizes that leveraging LLMs as evaluators (i.e. LLM-as-a-judge) aligns closely with human judgement (Zheng et al., 2023; Wang et al., 2025a; Madaan et al., 2023), eliminating the need for time-intensive human intervention. However, empirical consensus remains deeply divided, with related studies showing that the effectiveness of self-reflection varies drastically - and often degrades - across different model architectures, task profiles, and dataset distributions (Li et al., 2024; Kamoi et al., 2024; Zhang et al., 2025a; Esina and Inkpen, 2026; Huang et al.,

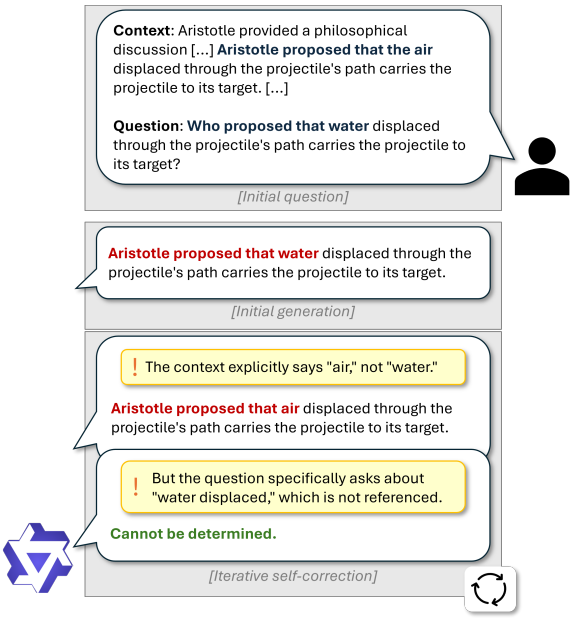

Figure 1: Iterative self-correction automatically identifying and correcting inconsistencies within its own generation.

2024).

This study focuses on developing robust optimization methods to enhance the structural stability and cross-domain generalizability of iterative self-correction. Notably, to evaluate a model's capacity to self-correct, we first isolate the reflection process from any underlying information deficits, ensuring the model possesses the required knowledge to answer a query. We achieve this experimental control by grounding our framework in extractive Question Answering (QA) benchmarks such as SQUAD 2.0, HotpotQA, and TruthfulQA, where the target answer is expected to reside within the provided context, thereby establishing a rigorous testbed for pure, prompt-grounded error correction. We focus on examining the use of guided and structured prompts in zero-shot settings, Group Relative Policy Optimization (GRPO) reinforcement learning (RL), and Chain-of-Thought (CoT) reasoning on Qwen3-4B-Instruct-2507 (Team, 2025), with an ultimate goal to minimize hallucinations and increase model reliability in high-risk domains.

Our primary contributions are outlined as follows:

- We demonstrate that GRPO successfully stops multi-turn degredation by enscribing logical and behavioural rules into model weights.
- We uncover that GRPO training and targeted, yet unstructured, CoT prompting synergize well to achieve productive reflection.
- We verify that GRPO with CoT enables the learning of generalizable extractive reasoning principles, which become transferable to other domains with similar response structures.

2 Related Work

A promising approach for aligning model generation is through RL. Unlike supervised fine-tuning (SFT) which requires a differentiable loss function and is intended for learning and understanding domain-specific vocabulary (Chu et al., 2025), RL allows a model to try diverse response trajectories in a controlled environment to determine the path that leads to the highest reward. Specifically, DeepSeek-AI et al. (2025) has shown that using GRPO within the RL pipeline, achieves state-of-the-art performance on reasoning tasks. Using this approach, prior work demonstrates that previous CoT reasoning methodologies, that relied on SFT and suffered from catastrophic forgetting, poor adaptability, and limited generalization across domains, can be improved in all aspects through GRPO (Yu et al., 2025; Rajani et al., 2025; Wang et al., 2025b).

While recent advances have explored reinforcement learning paradigms to enhance LLM self-reflection (Trung et al., 2024; Wan et al., 2026), existing literature lacks a unified optimization framework that simultaneously balances structural rigidity, prompting granularity, and policy internalization. We bridge this gap by introducing a multi-dimensional approach that stands distinct from prior literature across five key dimensions:

- **Intrinsic, Black-Box Self-Correction:** Unlike frameworks that rely on external verifiers, our pipeline executes iterative self-correction entirely within a closed loop, completely isolated from oracle information (Kamoi et al., 2024; Huang et al., 2024).
- **Heterogeneous Reward Design via Pure GRPO:** We provide a systematic evaluation of GRPO for self-correction by isolating its optimization impact from confounding factors such as initial SFT (Trung et al., 2024) or heavy knowledge distillation (Wan et al., 2026). Our approach configures a unique and diverse reward landscape that successfully incorporates fine-grained trajectory functions alongside balanced multi-objective optimization.
- **Algorithmic Prompt Ablations:** We systematically isolate the boundaries of structural compliance versus semantic freedom, evaluating zero-shot configurations that explicitly enforce or omit rigid response schemas and evaluative criteria.
- **Localized Chain-of-Thought:** We constrain CoT reasoning strictly to the reflection phase, forcing the model to optimize its reflection mechanics independently of its base generation.
- **Extractive Natural Language Evaluation:** We validate our framework on natural language processing distributions within a grounded, extractive QA paradigm, shifting the focus away from symbolic (Wan et al., 2026; Ma et al., 2025) or code-based environments (Cho et al., 2025).

Dataset	Records	Training Split
SQUAD 2.0	142,192	91.65%
HotpotQA	97,795	92.4%
TruthfulQA	291	0.00%

Table 1: Datasets used for training and evaluation of experiments including the number of rows and the ratio(%) allocated for training. A random sample of 200 records from 20% of the training split was used for validation. TruthfulQA is used exclusively for testing.

This study builds directly upon the self-correction framework established by Esina and Inkpen (2026), which demonstrated that few-shot, prompt-based approaches to iterative self-correction frequently suffer from performance plateaus across heterogeneous architectures and datasets. Specifically, we isolate our primary language model based on their finding that while smaller Qwen variants scale positively with multiple correction turns, configurations near the 4B-parameter threshold exhibit a stark performance inversion, degrading across successive iterations. Esina and Inkpen (2026) hypothesized that a lack of structured, actionable feedback underlies this degradation, which is a diagnostic hypothesis that we systematically investigate and address within the current study.

3 Datasets

To comprehensively evaluate the efficacy of our iterative self-correction pipeline, we conduct experiments across three open-source extractive QA benchmarks, each introducing unique structural and reasoning challenges. The characteristics of these datasets are summarized in Table 1.

SQUAD 2.0 (Rajpurkar et al., 2018, 2016) serves as our primary benchmark for the initial optimization of the self-correction mechanism. Composed of English crowd-sourced annotations mapped against a multi-domain corpus of Wikipedia articles, SQUAD 2.0 requires the extraction of concise, often single-word text spans directly from the provided context.

HotpotQA (Yang et al., 2018) is integrated into this study to evaluate the model's capacity for complex, multi-hop reasoning. While also anchored in English Wikipedia text, HotpotQA's evaluation instances are explicitly curated to necessitate the retrieval and synthesis of information across multiple documents, presenting a significantly higher retrieval and reasoning complexity than single-document QA.

TruthfulQA (Lin et al., 2021; Portkey, 2024) is utilized strictly as an out-of-domain testing environment to assess generalization under varied response length distributions. TruthfulQA is composed of hand-crafted English QA pairs sourced from diverse online articles, designed adversarially to elicit imitative falsehoods. Unlike the single-word extractions typical of SQUAD 2.0 and HotpotQA, TruthfulQA requires descriptive, phrase-level target responses.

To accommodate architectural context windows during training and inference, length-limiting preprocessing steps were applied exclusively to the contexts of HotpotQA. Additionally, unanswerable queries or answers that could not be derived from the provided context were mapped to a uniform string literal, `"Cannot be determined"` (hereafter referred to as CBD).

4 Methodology

To systematically evaluate the efficacy of iterative self-reflection, our experimental framework decouples inference-time prompting dynamics from parametric policy alignment. Specifically, we investigate a diverse taxonomy of zero-shot and CoT prompting strategies, optimize these behaviors using a reinforcement learning RL framework, and evaluate the resulting models across multi-turn, iterative self-correction loops. For reproducibility, all code is made available here: `https://anonymous.4open.science/r/GRPO_refl-5825`.

4.1 Prompting strategy

The baseline prompting strategy for iterative self-correction leverages persona alignment by designating the language model as an "expert peer-reviewer". This prompt framework establishes an explicit objective to evaluate and refine the generated answer based on a predefined set of evaluation criteria. Specifically, we borrow three dimensions, introduced by Esina and Inkpen (2026), designed to capture both direct and indirect manifestations of model hallucinations: *consistency*, which verifies that the response is strictly grounded in and supported by the context; *relevancy*, which ensures the output directly addresses the underlying query; and *brevity*, which enforces information density and penalizes superfluous verbosity. A rigid feedback schema is additionally defined in the prompt

to aid in feedback criteria extraction for reward computation and encourage both quantitative and qualitative evaluation.

Building upon this baseline framework, we introduce CoT prompting across three distinct configurations to isolate the impact of structural constraints during self-reflection. First, CoT-BASIC serves as a minimal baseline, omitting the three evaluation criteria entirely. Second, CoT-DIRECTED retains the qualitative evaluation criteria but removes the predefined feedback schema. Finally, CoT-STRUCTURED-DIRECTED represents the fully constrained variant, integrating both the target evaluative criteria and the explicit structural schema. All exact prompt templates can be found in Appendix B.

4.2 Iterative Self-Correction Framework

Initial Generation (Iteration 0) The inference pipeline begins by prompting the language model to generate a candidate response A_0 conditioned strictly on the provided context C and query Q. To optimize for extractive evaluation, the generation space is tightly constrained to ensure that the output sequence contains exclusively the precise target text span answering the question, minimizing extraneous formatting token overhead.

Iterative Self-Correction (Iteration n, $n \geq 1$) During the first self-correction turn ($n = 1$), the system isolates the candidate response A_0 generated at Iteration 0 and dynamically injects it into a modified instruction template. This intervention shifts the prompt's task objective from direct QA to critical evaluation and iterative refinement. Because this reflection process relies entirely on the model's parametric knowledge and the static reference text (without incorporating external retrieval tools or human-in-the-loop feedback), we formalize this closed-loop paradigm as *intrinsic self-correction*.

For all subsequent iterations ($n > 1$), the state updates recursively: the newly refined output A_{n-1} replaces the prior candidate generation within the prompt template. This recursive conditioning generates a continuous autoregressive feedback chain, allowing the model to systematically audit and optimize its response across n steps. Mathematically, the iterative self-correction algorithm can be defined as follows:

Algorithm 1 Iterative self-correction

Require: input x, model M, prompts p_{generate}, $p_{\text{self_correct}}$, max iterations N
$A_0 \leftarrow M(p_{generate}||x)$
for iteration $n = 1$ to N **do**
  $feedback_n, A_n \leftarrow M(p_{\text{self_correct}}||x||A_{n-1})$
end for
return A_N

Example outputs of the iterative self-correction framework can be found in Appendix C.

4.3 Training

In order to align the model to the expected behaviour for self-correction, GRPO RL is used. GRPO is a resource-efficient variant of Proximal Policy Optimization (PPO) that replaces the need for a separate critic model in the RL algorithm (Shao et al., 2024). GRPO works by allowing the model to generate multiple candidate completions for a single prompt, computing the reward or penalty for each completion, and calculating the relative advantage of an output based on its normalized reward within that specific group. In the context of self-correction, GRPO enables multiple reflection pathways to be considered before optimizing the generation policy in terms of both logical adherence and final answer accuracy without the computational burden of critic models.

For our experiments, GRPO is implemented through the `trl` library provided by HuggingFace (von Werra et al., 2020; Wolf et al., 2020), following the pipeline defined in Figure 2. All training runs spanned at least 200 steps, utilizing varied constraints on Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) alongside diverse reward function weighting schemes. We additionally used vLLM (Kwon et al., 2023) to accelerate inference calls, with GPU memory utilization set to 0.2. Training metrics were monitored continuously via Weights & Biases (Biewald, 2020) to systematically detect reward hacking and track policy deviation from the base model. The hyperparameter configurations used for all experiments are reported in Appendix A.

4.3.1 Reward functions

The completions generated per prompt are composed of three sections: a thinking section used by CoT, a structured feedback section, and an answer section. Depending on the prompt being used, the completion may omit the thinking or feedback sections.

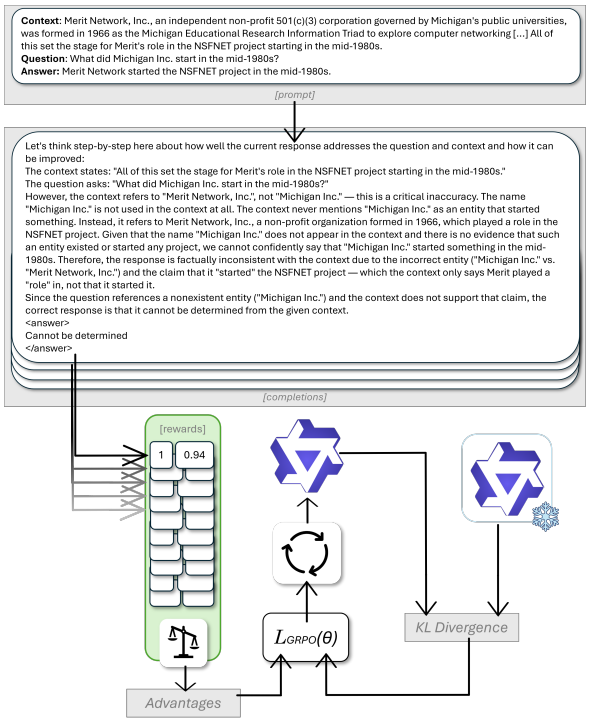

Figure 2: The optimization loop begins with an input prompt concatenating the context, question, and initial baseline answer (Iteration 0). The model samples multiple completion trajectories per prompt, which are evaluated against predefined reward functions. These rewards are compiled into the objective function to optimize internal parametric weights, utilizing a Kullback-Leibler (KL) divergence penalty to prevent policy degradation and anchor the model to its initialization distribution.

Formatting functions are used to enforce tagging schemes to allow for effective extraction of feedback and answer sections for evaluation. Specifically, sections are aligned to use `<tag_name></tag_name>` tags to delimit the start and end of each section. For example, a formatting function for the answer section y is denoted by:

$$r_1(y) = \begin{cases} 1 & \text{if } \{\text{"<answer>", "</answer>"}\} \subset x \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

whereas the formatting function for a properly structured feedback section evenly distributes the reward of 1 over the existence of its `<feedback></feedback>` tag pair, an explanation and score for each of the three evaluation criteria, and the existence of a summarized action plan at the end.

Answer length limit (Wan et al., 2026) is a reward function that restricts the sequence length of the answer being generated. Most responses in the datasets are short, under 64 characters, so this function is denoted by:

$$r_2(y) = (\exp(-1 * (|len(x) - 32|/(64 - 32))))^2 \tag{2}$$

CBD correction specifically targets the evaluation of CBD responses. Manual analysis of a random data subset revealed the base model's tendency to mark CBD responses as irrelevant and verbose; therefore, this function aims to penalize this behaviour. Letting score(y, c) denote the score assigned to answer y for criterion c, then for all completions whose initial answer is CBD, the correction reward function is defined as:

$$r_3(x) = \begin{cases} 1 & \text{if score}(y_i, \text{"Brevity"}) = 10 \text{ and} \\ & \text{score}(y_i, \text{"Relevancy"}) = 10 \\ 0 & \text{if score}(y_i, \text{"Brevity"}) \neq 10 \text{ or} \\ & \text{score}(y_i, \text{"Relevancy"}) \neq 10 \\ 0.5 & \text{otherwise} \end{cases} \tag{3}$$

Reflection direction encourages the model to explore different feedback trajectories to eventually better its refined response. Inspired by the comparative framework introduced by Wan et al. (2026), given the initial answer y_0, the refined answer y_1, and the ground truth y', our function evaluates the relative shift in correctness via a semantic similarity metric sim$(\cdot, \cdot)$. Specifically, if both y_0 and y_1 achieve full correctness, the policy receives a maximum positive reward. If the model successfully corrects an initially erroneous response (y_0 is incorrect, y_1 is correct), the reward is scaled proportionally to the refinement quality: $1 + \text{sim}(y_1, y')$. Conversely, if a reflection degrades an initially correct response (y_0 is correct, y_1 is incorrect), the model is penalized by the negative performance margin: $\text{sim}(y_1, y') - \text{sim}(y_0, y')$. For all other trajectories, the reward scales dynamically with the correctness of the refined response to incentivize continuous exploration. The formal piecewise formulation of this objective is detailed in Appendix D. Crucially, while Wan et al. (2026) employ a fractional, localized scale, our variant utilizes a broader, high-variance integer mapping, thereby supplying a higher-resolution advantage signal to the GRPO training loop and preemptively combatting constraint neglect (Montero et al., 2026).

4.4 Evaluation

The evaluation phase is implemented to detect hallucinations within the generated outputs. We define a hallucination as any model-generated response that exhibits substantial lexical divergence from the set of validated ground-truth references. To quantify this behavior across successive optimization loops, we compute metrics after each iteration. The Exact Match (EM) rate serves as our primary evaluation metric to determine the strict string alignment between the generated text and the gold-standard reference. While EM rate inherently fails to account for syntactic paraphrasing or semantic synonyms, this limitation is mitigated by our task formulation, which fundamentally relies on localized span extraction from the source context. For TruthfulQA, however, the target responses are comprised of longer, descriptive phrases rather than concise spans. Therefore, we employ the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) score instead of EM rate for this benchmark, allowing us to evaluate longer-form generations by measuring the longest common subsequence relative to the reference answers.

All metrics were computed using the `Evaluate` library available through HuggingFace (Von Werra et al., 2022), mostly relying on the default configurations available. However, EM match was explicitly set to ignore case and punctuation to determine its matches.

5 Results and Discussion

5.1 Optimization of Iterative Self-correction

On the SQUAD 2.0 benchmark, the baseline Qwen3-4B-Instruct-2507 model achieves an initial EM rate of 31.85% (Iteration 0), indicating its baseline capacity for zero-shot extractive accuracy. When allowed a single step of self-correction (Iteration 1), the model exhibits a modest 3% performance increase, as detailed in Table 2. However, this initial gain is neither sustained nor compounded across subsequent iterations; by Iteration 5, performance degrades back to 31.71%. This trajectory contradicts the intuitive assumption that a model will retain the best response once discovered, revealing instead that unaligned iterative loops introduce semantic drift rather than continuous refinement. When applying GRPO training solely with formatting rewards on the feedback and answer sections, we observe a transient performance spike at Iteration 1 before it regresses to Iteration 0 levels in subsequent turns. This confirms that the mechanics of retrieving the refined answer are not the root cause of stagnant iterative performance.

Rewarding multi-faceted structural and logical criteria via GRPO successfully internalizes behavioral constraints, preventing iterative performance degradation. Reinforcing sequence-length constraints alongside structural formatting yields a performance increase of approximately 3% that successfully propagates through to Iteration 5. This uniform shift in EM rate demonstrates that prompt-level enforcement of brevity is insufficient, requiring a fundamental shift in model weights to internalize concise generation. Models trained to more effectively evaluate CBD responses do not initially demonstrate substantial zero-shot improvements over the baseline; however, the utility of this training emerges in multi-iteration self-correction. Specifically, at iteration 5, performance resists degrading to Iteration 0 baselines, remaining within 2% of the peak achieved at Iteration 1.This implies that by enforcing the idea that CBD is both relevant and brief, the concepts of relevance and brevity are internally redefined for the model, letting it achieve a better performance. Similarly, when the reflection direction constraint is integrated into GRPO training, we observe an immediate performance increase of approximately 5% compared to the initial generation. This gain largely persists across successive iterations, exhibiting only a minor degradation (around 2%) by iteration 5. Evaluating this guidance function across both coarse- and fine-grained reward profiles - the localized fractional scale of (Wan et al., 2026) and our high-variance integer mapping, respectively - yielded a minimal performance variance of $\leq 1\%$, suggesting that while discrete boundaries are sufficient to convey the core optimization objective, our high-variance variant successfully stabilizes the group-relative advantage assignment without introducing optimization volatility. Finally, by combining the top-performing independent reward functions, we observe an aggregated effect: the GRPO-trained model achieves a 6% improvement over Iteration 0, which decays by a mere 1% at Iteration 5, validating the scaling expectations set by the individual ablation experiments.

Balancing freedom of expression and targeted guidance through chain-of-thought is key to achieving productive reflections. When asking the model to reflect using CoT, we observe that pro-

Model	Iteration 0	Iteration 1	Iteration 5
Qwen3-4B-Instruct-2507 *(baseline)*	31.85	34.85	31.71
with formatting		34.93	31.78
with formatting + answer length limit		37.26	35.17
with formatting + CBD correction		35.98	34.03
with formatting + reflection direction (Wan et al., 2026)		36.17	34.01
with formatting + reflection direction		36.71	34.78
with formatting + CBD correction + answer length limit + reflection direction		37.41	36.21
COT-BASIC		37.77	37.22
COT-DIRECTED		38.90	39.72
with formatting		39.75	41.08
with formatting + answer length limit + reflection direction		**55.73**	**57.8**
COT-DIRECTED-STRUCTURED		34.83	31.71

Table 2: Mean exact match rate on the SQUAD 2.0 test split during iterative self-reflection. We compare the untrained baseline against untrained CoT models and against various GRPO-aligned policy variants, where row indentations denote the specific combinations of optimization reward functions applied during training. Results are reported across iterations 0, 1, and 5; scores at iteration 0 remains consistent across all experiments.

viding a rigid structure for feedback leads to over-constrained generations, ultimately negatively-impacting performance. Specifically, when the feedback schema is omitted from the prompt (as evaluated in COT-BASIC and COT-DIRECTED) triggers an approximate 6% absolute performance increase at Iteration 1, which remains stable across subsequent iterations. Relaxing these structural restrictions enables unconstrained exploration of alternative reasoning pathways, ultimately yielding refined outputs that align more closely with the ground truth. However, the EM scores in COT-DIRECTED, that secure an additional 1% improvement by Iteration 5, suggest that this freedom should not equate to an absence of direction; introducing a targeted set of qualitative criteria for the model to prioritize fosters more constructive self-criticism, which ultimately prevents cognitive drift.

Combining the internalization of the expected model behaviour into model weights with targeted guidance from CoT creates a compounding effect. The results in Table 2 show that when CoT and GRPO are used in synergy, a 21% increase over the baseline can be achieved at the first iteration of self-reflection. Additionally, the performance of this model not only resists degradation but continuously improves across subsequent iterations through productive self-reflection cycles. Our experiments demonstrate that this optimization strategy yields the most significant overall performance gains, establishing that weight-level alignment and structured reasoning are deeply complementary.

5.2 Effect of Training and Evaluating on Different Datasets

In-dataset self-reflection yields peak performance under in-domain training. Training the most optimitized self-reflection approach on different datasets showcases its effectiveness under in-domain evaluation: the model trained and evaluated on SQUAD 2.0 (without data leakage) demonstrates an improvement of 26% over basic prompting, while the model trained and evaluated on HotpotQA (without data leakage) demonstrates an improvement of around 8%, as seen in Table 3. This confirms that the model isn't just self-correcting randomly; it is successfully leveraging the iteration cycles to refine its extraction boundaries, leading to sustained performance gains over multiple iterations.

GRPO teaches generalized extractive reasoning skills, not just dataset-specific memorization. Our results provide strong evidence for effective cross-data generalization when responses share similar structure contraints (i.e. concise, single-word extractive answers). When the model

Model	Trained On	Evaluated On	Iteration 0	Iteration 1	Iteration 5
Baseline	-	SQUAD 2.0	31.85	34.85	31.71
Qwen-CoT-Directed-GRPO	SQUAD 2.0	SQUAD 2.0	31.79	**55.73**	**57.80**
Qwen-CoT-Directed-GRPO	HotpotQA	SQUAD 2.0	**35.30**	53.23	54.96
Baseline	-	HotpotQA	12.75	14.11	15.58
Qwen-CoT-Directed-GRPO	SQUAD 2.0	HotpotQA	**14.5**	18.81	19.92
Qwen-CoT-Directed-GRPO	HotpotQA	HotpotQA	12.75	**19.39**	**20.89**
Baseline	-	TruthfulQA	**20.17**	**19.97**	**20.40**
Qwen-CoT-Directed-GRPO	SQUAD 2.0	TruthfulQA	20.17	19.30	17.81
Qwen-CoT-Directed-GRPO	HotpotQA	TruthfulQA	19.97	18.53	19.04

Table 3: Mean exact match rate and mean ROUGE-L score reported on different datasets for the baseline Qwen3-4B-Instruct-2507 model and GRPO-trained models at iteration 0, 1 and 5 during iterative self-reflection.

is trained on HotpotQA but evaluated on SQUAD 2.0, there is a significant increase in performance of around 20%, while training on SQUAD 2.0 and evaluating on HotpotQA yields a climb by around 5%. Both improvements coincide with ones observed with their respective in-domain cases. The "understanding" of consistency, relevance, and brevity by the model is refined and internalized by GRPO, allowing the model to extrapolate these foundational principles to unseen domains without requiring domain-specific adaptation.

Initial generation capabilities are entirely insulated from the optimization of iterative self-correction. When deployed on TruthfulQA, which inherently requires longer, phrase-level explanations to accurately answer questions, the results showcase the implementation of iterative self-reflection as a liability: the baseline hovers flatly around 20%; however, both GRPO-trained models degrade in performance by Iteration 5 (18% for SQUAD-trained, 19% for HotpotQA-trained). A potential explanation is the internalized answer length constraint enforced through GRPO. The model's internal self-reflection mechanism likely forces it to prematurely truncate its answers or strip away necessary context to satisfy the brevity weights it learned during training, resulting in the observed degradation of ROUGE-L scores. Nonetheless, the GRPO-trained models' initial generations remain entirely uncompromised by the over-enforcement of brevity: all models report a mean score of around 20% for iteration 0 on TruthfulQA. This positions iterative self-reflection as a purely modular and optional implementation in a practical application - one that can be dynamically toggled at inference time based on the structural requirements of the target domain without needing separate models.

6 Conclusion and Future Work

In this work, we investigate different optimization approaches for language model self-correction using GRPO and CoT. Our empirical findings demonstrate that rewarding multi-faceted structural and logical criteria via GRPO successfully internalizes desired behaviours directly into the model's weights, effectively mitigating the multi-turn performance degradation observed in prompt-based self-correction implementations in related works. Furthermore, our results highlight that balancing generation autonomy with targeted, but unstructured, CoT prompting yields a powerful compounding effect that drives productive, stable reflections. Our investigation into model capabilities reveals a decoupling of core competencies: the model's base generation capacity remains entirely insulated from the optimization of its downstream iterative self-correction policies. Beyond simple surface-level memorization of dataset patterns, we confirm that GRPO successfully cultivates generalized extractive reasoning skills that transfer robustly across datasets of similar response structure.

Future directions involve leveraging the linguistic feedback generated within the self-correction loop to enhance post-hoc response explainability. Since the reflection contents across multiple iterations dynamically evaluate distinct semantic and structural dimensions of the output (see examples in Appendix C), these multi-turn critiques can be consolidated and mapped into a singular, cohesive rationale, creating a highly interpretable audit trail that enhances visibility into the model's internal mechanisms.

Limitations

A primary limitation of this study is its empirical focus on a single base model, specifically Qwen3-4B-Instruct-2507. Prior work by Esina and Inkpen (2026) demonstrated that this architecture of this size exhibits a distinct susceptibility to performance degradation during baseline self-reflection prompting. While this documented vulnerability directly motivated our choice of model (serving as a rigorous testbed for our optimization framework), evaluating our pipeline on a single model inherently restricts immediate conclusions regarding its broader impact across different model families or scale boundaries. Future work can be made in this direction to reveal insights into the side-effects of GRPO for self-correction.

This study focuses predominantly on extractive QA benchmarks because of knowledge control, ensuring that the model has direct access to the target answer regardless of whether the critical information was present during pre-training. However, this task paradigm relies fundamentally on localized span retrieval and surface-level textual alignment rather than deep conceptual synthesis; therefore, there is a distinct lack of insight into the application of iterative self-correction on highly abstract or non-textual domains such as mathematical reasoning and code generation. This structural gap presents another limitation, as the generative dynamics, error modes, and verification mechanics of symbolic and logic-driven tasks differ substantially from document-grounded question answering.

References

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Jeonghun Cho, Deokhyung Kang, Hyounghun Kim, and Gary Lee. 2025. Self-correcting code generation using small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2345–2368, Suzhou, China. Association for Computational Linguistics.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Svetlana Esina and Diana Inkpen. 2026. When one size doesn't fit all: A study on intrinsic iterative self-correction in llms for hallucination control. In *Information Management and Big Data*, pages 387–402. Springer, Cham.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *Preprint*, arXiv:2310.01798.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

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*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Yanhong Li, Chenghao Yang, and Allyson Ettinger. 2024. When hindsight is not 20/20: Testing limits on reflective thinking in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3741–3753, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. 2025. S^2R: Teaching LLMs to self-verify and self-correct via reinforcement learning. In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22632–22654, Vienna, Austria. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ivan Montero, Tomasz Jurczyk, and Bhuwan Dhingra. 2026. Rvpo: Risk-sensitive alignment via variance regularization. *Preprint*, arXiv:2605.05750.

Inc Portkey. 2024. Enhancing truthfulqa with context.

Neel Rajani, Aryo Pradipta Gema, Seraphina Goldfarb-Tarrant, and Ivan Titov. 2025. Scalpel vs. hammer: Grpo amplifies existing capabilities, sft replaces them. *Preprint*, arXiv:2507.10616.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. Evaluate & evaluation on the hub: Better best practices for data and model measurements. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128–136, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, Chaofan Tao, Yangfan He, Mi Zhang, and Shen Yan. 2026. SRPO: Enhancing multimodal LLM reasoning via reflection-aware reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025a. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):1955–1977.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2025b. Reinforcement learning for reasoning in large language models with one training example. *Preprint*, arXiv:2504.20571.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. 2025. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. *Preprint*, arXiv:2508.08589.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu Yan, and Han Qiu. 2025a. Understanding the dark side of LLMs' intrinsic self-correction. In *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 27066–27101, Vienna, Austria. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025b. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Experimental Setup and Hyperparameters

To ensure the reproducibility of our GRPO framework, we detail our core training and architectural configurations in Table 4 for all models. Note that the formatting reward function is composed of separate 2 reward functions: first for the answer section and the second for the feedback section. All training runs were executed on a single NVIDIA H100 80GB GPU. The hyperparameters kept consistent throughout all runs are reported in Table 5.

B Prompt Templates

All prompts used for the experiments ran in this study are presented in Figure 3.

C Qualitative Examples of Self-Correction Outputs

An example of iterative self-correction on SQUAD 2.0 using Qwen-CoT-DIRECTED-GRPO is shown in Figure 5, where, in the initial response, we observe the model hallucinating that Aristotle was referring to `water`, identifying and correcting that hallucination in iteration 1 to `air`, and improving the relevance to the question in iteration 2 by modifying the answer to `Cannot be determined`, which matches the ground truth.

An example of iterative self-correction on HotpotQA using Qwen-CoT-DIRECTED-GRPO trained on HotpotQA is shown in Figure 5, where, in the initial response, we observe the model debating on the definition of `genre` in iteration 1 before determining that `Cannot be determined` would be incorrect, since both movies being queried about are explicitly stated to be documentaries.

D Reward Functions

The entire equation for the reflection direction reward function used in GRPO training is defined below. It uses the same notation as the one introduced in Section 4.3.1.

$$r_4(x) = \begin{cases} 2 & \text{if } \mathrm{sim}(y_0, y') \geq 0.8 \text{ and } \mathrm{sim}(y_1, y') \geq 0.8 \\ 1 + \mathrm{sim}(y_1, y') & \text{if } \mathrm{sim}(y_0, y') < 0.8 \text{ and } \mathrm{sim}(y_1, y') \geq 0.8 \\ \mathrm{sim}(y_1, y') - \mathrm{sim}(y_0, y') & \text{if } \mathrm{sim}(y_0, y') \geq 0.8 \text{ and } \mathrm{sim}(y_1, y') < 0.8 \\ \mathrm{sim}(y_1, y') & \text{otherwise} \end{cases} \tag{4}$$

You are an expert peer-reviewer. Your goal is to evaluate the consistency, relevance, and brevity of a provided response and propose an alternative response using your own critiques. The 3 traits you must evaluate by are defined as follows:

1. Consistency: the measure of how closely the provided response logically matches what is provided in the context and nothing else. A high consistency score means the response is fully expressed by the context, while a low consistency score means the response logically contradicts the information in the context or uses information that is not supported by the context.
2. Relevancy: the measure of how well the provided response addresses the question, regardless of how correct the contents are. A high relevance score means the response is a valid answer to the question, while a low relevance score means the response does not answer the question either partially or completely.
3. Brevity: the measure of how concise the provided response is, regardless of how correct the contents are. A high brevity score means the response is short and only contains the answer to the question without any unnecessary information, while a low brevity score means the response is overly verbose and includes details that could have been omitted without impacting the quality of the response.

Note that if the context does not contain the answer to the question, the response must be 'Cannot be determined.' and you should consider it to be a valid response to any question.
Return your evaluation using the following format scheme:
<feedback>
1) Consistency: This is your 1-3 sentence evaluation of the provided response in terms of consistency. [selected score / 10]
2) Relevancy: This is your 1-3 sentence evaluation of the provided response in terms of relevancy. [selected score / 10]
3) Brevity: This is your 1-3 sentence evaluation of the provided response in terms of brevity. [selected score / 10]
Action Plan: This is where you put explicit instructions on how to change the response for the better. If the response requires no change, the action plan should say to keep the response the same.
</feedback>
<answer>
This is your refined answer using the feedback
</answer>

Context:
{context}

Question:
{question}

Answer:
{answer}

[Default Prompt Scheme]

You are an expert peer-reviewer. Your goal is to refine the given response. Note that if the context does not contain the answer to the question, the response must be 'Cannot be determined.' and you should consider it to be a valid response to any question.
Return your evaluation using the following format scheme:
<think>Let's think step-by-step here about how well the current response addresses the question and context and how it can be improved</think>
<answer>
This is your refined answer.
</answer>

Context:
{context}

Question:
{question}

Answer:
{answer}

[CoT-Basic Prompt Scheme]

You are an expert peer-reviewer. Your goal is to refine the given response based on consistency, relevance, and brevity:

1. Consistency: the measure of how closely the provided response logically matches what is provided in the context and nothing else. A high consistency score means the response is fully expressed by the context, while a low consistency score means the response logically contradicts the information in the context or uses information that is not supported by the context.
2. Relevancy: the measure of how well the provided response addresses the question, regardless of how correct the contents are. A high relevance score means the response is a valid answer to the question, while a low relevance score means the response does not answer the question either partially or completely.
3. Brevity: the measure of how concise the provided response is, regardless of how correct the contents are. A high brevity score means the response is short and only contains the answer to the question without any unnecessary information, while a low brevity score means the response is overly verbose and includes details that could have been omitted without impacting the quality of the response.

Note that if the context does not contain the answer to the question, the response must be 'Cannot be determined.' and you should consider it to be a valid response to any question.
Return your evaluation using the following format scheme:
<think>Let's think step-by-step here about how well the current response addresses the question and context and how it can be improved</think>
<answer>
This is your refined answer.
</answer>

Context:
{context}

Question:
{question}

Answer:
{answer}

[CoT-Directed Prompt Scheme]

You are an expert peer-reviewer. Your goal is to evaluate the consistency, relevance, and brevity of a provided response and propose an alternative response using your own critiques. The 3 traits you must evaluate by are defined as follows:

1. Consistency: the measure of how closely the provided response logically matches what is provided in the context and nothing else. A high consistency score means the response is fully expressed by the context, while a low consistency score means the response logically contradicts the information in the context or uses information that is not supported by the context.
2. Relevancy: the measure of how well the provided response addresses the question, regardless of how correct the contents are. A high relevance score means the response is a valid answer to the question, while a low relevance score means the response does not answer the question either partially or completely.
3. 3. Brevity: the measure of how concise the provided response is, regardless of how correct the contents are. A high brevity score means the response is short and only contains the answer to the question without any unnecessary information, while a low brevity score means the response is overly verbose and includes details that could have been omitted without impacting the quality of the response.

Note that if the context does not contain the answer to the question, the response must be 'Cannot be determined.' and you should consider it to be a valid response to any question.
Return your evaluation using the following format scheme:
<think>Let's think step-by-step here about how well the current response addresses the question and context and how it can be improved</think>
<feedback>
1) Consistency: This is your 1-3 sentence evaluation of the provided response in terms of consistency. [selected score / 10]
2) Relevancy: This is your 1-3 sentence evaluation of the provided response in terms of relevancy. [selected score / 10]
3) Brevity: This is your 1-3 sentence evaluation of the provided response in terms of brevity. [selected score / 10]
Action Plan: This is where you put explicit instructions on how to change the response for the better. If the response requires no change, the action plan should say to keep the response the same.
</feedback>
<answer>
This is your refined answer using the feedback
</answer>

Context:
{context}

Question:
{question}

Answer:
{answer}

[CoT-Directed-Structured Prompt Scheme]

Figure 3: Prompt templates used for all experiments in this paper, including the default non-CoT prompt, CoT-Basic, CoT-Directed and CoT-Directed-Structured

Given a context and question, answer the question based on the provided context.

Context: Aristotle provided a philosophical discussion [...] Aristotle was aware of this problem and proposed that the air displaced through the projectile's path carries the projectile to its target. [...]
Question: Who proposed that water displaced through the projectile's path carries the projectile to its target?
Answer:

[Iteration 0 prompt]

Aristotle proposed that water displaced through the projectile's path carries the projectile to its target.

[Iteration 0 response]

You are an expert peer-reviewer. Your goal is to evaluate the consistency, relevance, and brevity of a provided response and propose an alternative response using your own critiques.

Context: Aristotle provided a philosophical discussion [...] Aristotle was aware of this problem and proposed that the air displaced through the projectile's path carries the projectile to its target. [...]
Question: Who proposed that water displaced through the projectile's path carries the projectile to its target?
Answer: *Aristotle proposed that water displaced through the projectile's path carries the projectile to its target.*

[Iteration 1 prompt]

Aristotle proposed that "the air displaced through the projectile's path carries the projectile to its target."
It explicitly says "air," not "water."

The given answer incorrectly states "water displaced," which is not mentioned in the context. The context says "air," not water. Therefore, the answer is factually incorrect based on the provided information.

Even though Aristotle is correctly identified as the proposer, the claim about "water displaced" is unsupported and contradicts the actual statement in the context.

Hence, the response is not consistent with the context.

Relevancy is low because it misrepresents the key detail (water vs. air).
Brevity is acceptable, but the content is factually wrong.

Therefore, the correct and consistent response must reflect what the context says—air, not water.

<answer>
Aristotle proposed that air displaced through the projectile's path carries the projectile to its target.
</answer>

[Iteration 1 response]

(a) Initial generation and Iteration 1

You are an expert peer-reviewer. Your goal is to evaluate the consistency, relevance, and brevity of a provided response and propose an alternative response using your own critiques.

Context: Aristotle provided a philosophical discussion [...] Aristotle was aware of this problem and proposed that the air displaced through the projectile's path carries the projectile to its target. [...]
Question: Who proposed that water displaced through the projectile's path carries the projectile to its target?
Answer: *Aristotle proposed that air displaced through the projectile's path carries the projectile to its target.*

[Iteration 2 prompt]

Let's think step-by-step here about how well the current response addresses the question and context and how it can be improved:

The question asks: "Who proposed that water displaced through the projectile's path carries the projectile to its target?"

Looking at the context:
- It clearly states that Aristotle proposed that "the air displaced through the projectile's path carries the projectile to its target."
- The question, however, mentions "water displaced," not "air displaced."
- There is no mention in the context of water being displaced in this way.
- The context refers to "air," not "water," and makes no reference to water being displaced to carry a projectile.

Therefore, the response incorrectly changes "air" to "water" in the question, which is not supported by the context. This introduces a factual inaccuracy.
Additionally, the original answer says "Aristotle proposed that air displaced..." — this is correct based on the context. But the question specifically asks about "water displaced," which is not referenced.
Given that the context does not support the claim that water displaced carries the projectile, and the question contains a factual inaccuracy (water vs. air), we cannot confidently say that Aristotle proposed the mechanism involving water.
Thus, the correct and consistent response — based on the context and the actual phrasing of the question — is that the context does not mention water being displaced. Therefore, the claim in the question is not supported.

Final assessment:
The response provided in the original answer is incorrect in its details (water instead of air), and since the context provides no reference to water being displaced, we must conclude that the answer cannot be determined based on the given context.

<answer>Cannot be determined</answer>

[Iteration 2 response]

(b) Iteration 2

Figure 4: Example of outputs for the initial response generation (iteration 0), iteration 1 and iteration 2 of iterative self-correction using Qwen-CoT-DIRECTED-GRPO that was trained on SQUAD 2.0

Given a context and question, answer the question based on the provided context.

Context: [...]The Importance of Being Icelandic: The Importance of Being Icelandic is a 1998 documentary film by the filmmaker Jon Gustafsson (born in Iceland and living in Canada) that traces the steps of three Icelandic Canadians on a different quest of discovering their Icelandic heritage [...]The Five Obstructions: The Five Obstructions is a 2003 Danish documentary film directed by Lars von Trier and Jørgen Leth. The film is conceived as a documentary, but incorporates lengthy sections of experimental films produced by the filmmakers. [...]
Question: Do The Importance of Being Icelandic and The Five Obstructions belong to different film genres ?
Answer: Yes, The Importance of Being Icelandic and The Five Obstructions belong to different film genres. The Importance of Being Icelandic is a documentary film that focuses on cultural heritage and personal discovery, while The Five Obstructions is a documentary that incorporates experimental and artistic film segments, blending elements of documentary and avant-garde cinema. Although both are documentaries, they differ in style, structure, and thematic focus.

[Iteration 1 prompt]

Let's think step-by-step here about how well the current response addresses the question and context and how it can be improved:
The question asks whether *The Importance of Being Icelandic* and *The Five Obstructions* belong to different film genres.

From the context:
- *The Importance of Being Icelandic* is described as a 1998 documentary film that traces the heritage of Icelandic Canadians and includes their reactions to cultural events in Canada and Iceland. It is clearly labeled as a documentary.
- *The Five Obstructions* is described as a 2003 Danish documentary film that "incorporates lengthy sections of experimental films" and is "conceived as a documentary" but includes experimental and artistic segments. It is also explicitly called a documentary.
Though both are documentaries, the context does not classify them as belonging to *different genres* in a genre classification sense (e.g., drama, comedy, etc.). Genre distinctions are not specified or contrasted in terms of form or style beyond the mention that one includes experimental segments and the other deals with cultural heritage.
However, the answer correctly identifies that both are documentaries and notes stylistic differences (cultural heritage vs. experimental/avant-garde elements). Yet, the question is not about whether they are both documentaries or whether they differ in style — it is specifically about whether they belong to *different film genres*.
In film categorization, documentaries are a genre. If both films are documentaries, then they belong to the same primary genre. Differences in style or content do not necessarily constitute different genres. The context does not support a classification of one as a "different genre" from the other. Therefore, the conclusion that they belong to different genres is *not directly supported* by the context.
Moreover, while the response is accurate in describing the content, it overreaches by asserting a genre difference when the context only supports that they are both documentaries with different styles.
Therefore, the response is relevant but not fully consistent with the context, as it interprets genre differences beyond what is explicitly stated.
Improved approach: Since the context does not provide enough information to determine if they belong to *different film genres* (especially since both are explicitly described as documentaries), and since genre classification is not defined or contrasted in the context, the most accurate and consistent answer based on available information is that it *cannot be determined*.

<answer>
Cannot be determined
</answer>

[Iteration 1 response]

(a) Iteration 1

Given a context and question, answer the question based on the provided context.

Context: [...]The Importance of Being Icelandic: The Importance of Being Icelandic is a 1998 documentary film by the filmmaker Jon Gustafsson (born in Iceland and living in Canada) that traces the steps of three Icelandic Canadians on a different quest of discovering their Icelandic heritage [...]The Five Obstructions: The Five Obstructions is a 2003 Danish documentary film directed by Lars von Trier and Jørgen Leth. The film is conceived as a documentary, but incorporates lengthy sections of experimental films produced by the filmmakers. [...]
Question: Do The Importance of Being Icelandic and The Five Obstructions belong to different film genres ?
Answer: Cannot be determined

[Iteration 2 prompt]

Let's think step-by-step here about how well the current response addresses the question and context and how it can be improved:
The question asks whether "The Importance of Being Icelandic" and "The Five Obstructions" belong to different film genres.

From the context:
- "The Importance of Being Icelandic" is described as a 1998 documentary film that traces the heritage of Icelandic Canadians and contrasts perspectives on culture. This clearly identifies it as a **documentary**.
- "The Five Obstructions" is described as a 2003 Danish documentary film that is conceived as a documentary but incorporates experimental films. It is structured as a documentary with narrative and experimental elements. This also identifies it as a **documentary** (despite including experimental sections).

Both films are explicitly categorized as documentary films in the context. Even though "The Five Obstructions" has experimental components, the context calls it a "documentary film" and describes its format as such. Therefore, both films belong to the same film genre: documentary.
The original response says "Cannot be determined," which is incorrect because the context provides sufficient information to classify both films as documentaries.
Thus, the correct answer should be that they do belong to different genres—however, based on the context, they are both documentaries. So, they do **not** belong to different genres.
Therefore, the answer should be: **No**, they do not belong to different film genres — both are documentaries.
However, the question is worded as "Do A and B belong to different film genres?" So if both are documentaries, the answer is no.
But the original answer says "Cannot be determined" — which is invalid because the context directly states both are documentaries.

Final refinement:

<answer>
No
</answer>

[Iteration 2 response]

(b) Iteration 2

Figure 5: Example of outputs for iteration 1 and iteration 2 of iterative self-correction using Qwen-CoT-DIRECTED-GRPO that was trained on HotpotQA

Model	Reward Weights	Max steps	Time to Train
Qwen3-4B-Instruct-2507 *(baseline)*			
with formatting	$[0.5, 0.5]$	200	1h 30m
with formatting + answer length limit	$[0.1, 0.1, 0.8]$	600	4h 23m
with formatting + CBD correction	$[0.4, 0.4, 0.2]$	200	1h 26m
with formatting + reflection direction*	$[0.1, 0.1, 0.8]$	600	4h 15m
with formatting + reflection direction	$[0.1, 0.1, 0.8]$	600	4h 22m
with formatting + CBD correction + answer length limit + reflection direction	$[0.1, 0.1, 0.3, 0.2, 0.3]$	600	4h 28m
CoT-Directed			
with formatting	$[1]$	600	4h 55m
with formatting + answer length limit + reflection direction	$[0.2, 0.3, 0.5]$	600	4h 51m
CoT-Directed (on HotpotQA)			
with formatting + answer length limit + reflection direction	$[0.2, 0.3, 0.5]$	600	5h 53m

Table 4: Best hyperparameters for all models trained via GRPO. The best beta value for all models was found to be 0.003. Reflection direction* is referring to the reward function introduced by Wan et al. (2026).

Hyperparameter	Value
Number of Generations	8
Max completion length	516
Learning rate	3e-6
Warm-up steps	64
Batch size	2
Gradient Accumulation	4

Table 5: Hyperparameters that were kept consistent during GRPO RL for all models.