Iterative Self-Questioning Supervision with Semantic Calibration for Stable Reasoning Chains in Large Language Models

Yaxuan Luan *

Posted Date: 9 February 2026

doi: 10.20944/preprints202602.0653.v1

Keywords: circular reasoning supervision; self-questioning; semantic calibration; reasoning consistency

 doi:10.20944/preprints202602.0653.v1

Article

Iterative Self-Questioning Supervision with Semantic Calibration for Stable Reasoning Chains in Large Language Models

Yaxuan Luan

University of Southern California, Los Angeles, USA; yaxuanlu@alumni.usc.edu

Abstract

This study addresses the inconsistency, semantic drift, and logical breaks that large language models often exhibit in complex reasoning tasks and proposes a unified cyclic self-questioning supervision framework that integrates information flow from questioning to reflection and renewed reasoning. The framework includes four core components, namely questioning generation, reflection modeling, semantic calibration, and renewed reasoning, and forms an iterative reasoning chain that allows the model to identify potential uncertainties and adjust its reasoning path based on internal feedback in each round. The method first uses the questioning module to produce structured queries about the initial reasoning result and to extract possible logical weaknesses from the generated content. The reflection module then interprets the questioning content, locates errors, and produces internal feedback signals that guide reasoning improvement. The semantic calibration mechanism converts the reflection output into intermediate states that influence the reasoning space and provide a more stable foundation for renewed reasoning. Through multiple iterations, the framework increases the internal consistency of the reasoning chain. Systematic experiments conducted on open reasoning datasets show significant gains in accuracy, explanation consistency, semantic alignment, and self-consistency, which confirms the importance of internal reflection and semantic calibration in improving reasoning quality. Sensitivity studies on learning rate, reflection length, reasoning temperature, and parallelism further reveal how the cyclic system depends on internal feedback absorption and semantic stability. The unified framework provides an extensible path for enhancing the structural quality of reasoning chains in large models and offers an interpretable foundation for high-reliability reasoning scenarios.

Keywords: circular reasoning supervision; self-questioning; semantic calibration; reasoning consistency

I. Introduction

Large language models have demonstrated strong generative and reasoning capabilities across many complex tasks. However, their reasoning processes still show clear signs of fragility. They often lack continuous self-monitoring, do not possess stable self-checking or self-correction abilities, and struggle to maintain consistent and reliable reasoning chains in complex, ambiguous, or cross-domain scenarios [1]. As a result, these models are prone to hallucinations, logical breaks, and partial inconsistencies when dealing with multi-step reasoning, long-chain logic, conflicting task constraints, or cross-paragraph semantic integration. With large models now used in high-risk fields such as scientific analysis, knowledge services, medical interpretation, judicial assessment, and engineering design, the absence of controllable, interpretable, and verifiable reasoning procedures has become a critical barrier to their trustworthiness and real-world deployment [2].

In recent years, research on generation reliability has grown rapidly. Existing efforts include reflective generation, self-evaluation, chain-based reasoning, adversarial reinforcement, and knowledge enhancement. However, most of these approaches remain limited to single-step reflection

 doi:10.20944/preprints202602.0653.v1

or isolated post-hoc revision. They lack a unified mechanism that can operate throughout the entire reasoning process. Under current frameworks, a model typically generates an answer, performs one quality check, identifies errors, and then regenerates the content. This one-directional structure cannot ensure that feedback from the review stage truly constrains subsequent generations. It also fails to create a continuous and dynamic loop of self-supervision. More importantly, such methods do not systematically manage internal conflicts within the reasoning chain, imbalanced evidence use, semantic drift, or cross-step logical errors. They therefore struggle to meet higher-level requirements for reasoning consistency [3].

As the task scale increases, information distribution becomes more complex, and deeper reasoning is required; static model parameters or one-time reflection are no longer sufficient for stable reasoning in real scenarios. This has drawn growing attention from both academia and industry to a new demand. Large language models need a cyclic self-questioning mechanism. They must be able to generate questions, identify risk points, calibrate semantic relations, and iteratively reconstruct their own answers during reasoning. This ability is not only about correcting results. It represents a form of process supervision. It requires models to extract diagnostic signals from their own outputs and improve logical consistency and semantic reliability through repeated questioning, reflection, and renewed reasoning. Such a cyclic supervision paradigm is becoming a key pathway for next-generation models to achieve higher reliability, stronger interpretability, and more advanced intelligent behavior.

Against this backdrop, building a unified framework for self-questioning and renewed reasoning holds significant theoretical and practical value. Theoretically, it helps characterize the internal reasoning behavior of generative models, shifting the process from a black-box output to a structured and dynamic evolution. With internal control signals produced through self-questioning, it becomes possible to better understand failure modes in reasoning and to integrate insights from logical reasoning, cognitive modeling, and semantic consistency research. Practically, a general cyclic supervision system can improve reasoning quality without requiring external labels. It enables models to maintain stable conclusion consistency, lower hallucination rates, and stronger cross-domain adaptability in complex tasks. From an application perspective, this will enhance model credibility in scientific research, medical diagnosis, legal analysis, business decision-making, and safety auditing. It also establishes a foundation for building safe, transparent, and verifiable large-model ecosystems [4].

Furthermore, a unified framework for self-questioning and renewed reasoning carries major implications for future development paradigms in large language models. It encourages a shift from a result-centered to a process-centered design philosophy. This transformation will inspire new directions in architectural design, training strategies, alignment methods, and system integration. By incorporating reflective signals, logical constraints, and sequence-level supervision into a continuous iterative structure, the model can form an evolutionary reasoning mechanism. It can gain abilities of self-correction, self-improvement, and self-stabilization. As models grow in scale and as tasks become more diverse, such cyclic self-supervision will be essential for achieving higher-level intelligent behaviors. It will also open new possibilities for models that require long-term planning, deep reasoning, and multi-stage decision-making.

II. Related Work

Research on the reasoning reliability of large language models can be divided into several major directions. These include chain-based reasoning, process supervision, reflective generation, and retrieval-augmented reasoning. Chain-based reasoning methods construct explicit multi-step logical chains that guide the model to unfold its reasoning step by step. This improves the ability to decompose complex problems and increases the transparency of intermediate steps [5]. However, such methods often rely on static and linear reasoning structures. They lack dynamic detection of incorrect steps and do not provide mechanisms for self-correction. When dealing with long reasoning chains, multi-hop logic, or cross-paragraph semantic integration, chain-based reasoning can still

 doi:10.20944/preprints202602.0653.v1

accumulate misleading intermediate steps. This may cause the conclusion to deviate from the truth. Although chain-based reasoning encourages structured thinking, it remains difficult to build a unified system that supports cyclic self-examination and renewed reasoning.

Process supervision methods aim to enhance reasoning consistency by using intermediate process signals. They introduce constraints, rewards, or evaluation mechanisms into intermediate steps to increase controllability within the reasoning process. Related studies attempt to incorporate structured information into verification steps, process annotations, or process rewards to improve the interpretability and stability of reasoning trajectories. However, these approaches rely on external annotations or additional supervision signals. They therefore struggle to remain generalizable in resource-limited, cross-task, or cross-domain settings. More importantly, process supervision strengthens constraints on the reasoning path but does not provide a mechanism that allows the model to identify defects, generate questions, and recalibrate itself based on internal signals. As a result, it is difficult for such methods to form a coherent loop from reflection to renewed reasoning [6].

Reflective generation methods encourage the model to review its own output once or several times. They attempt to improve final answers by detecting errors, identifying logical inconsistencies, and generating suggestions for refinement. These methods help reduce hallucinations and improve answer quality. However, most approaches still follow a single-round reflection or stage-wise revision paradigm. They lack continuous and dynamic cycles. When task complexity increases, when knowledge coverage widens, or when semantic conflicts appear more often, a single round of reflection cannot correct deeper errors in the reasoning chain [7]. In addition, many reflective methods treat reflection as a separate module rather than as part of an integrated reasoning framework. As a result, interaction between reflection and generation remains weak. It is difficult to form stable feedback signals that can guide multi-step reasoning.

Retrieval-augmented reasoning methods improve performance on cross-document and cross-domain tasks through external knowledge, evidence selection, and structured information integration. These methods have shown progress in knowledge grounding and factual consistency. However, retrieval modules also introduce noise, uncertainty, and conflicting evidence. This increases the sensitivity of the reasoning process to the quality of external information. Although some studies attempt to mitigate these issues through semantic filtering, conflict detection, or evidence weighting, current approaches still rely heavily on external mechanisms. They do not enable the model to develop internal capabilities for self-questioning or self-correction when facing inconsistent reasoning or conflicting evidence. Overall, existing research still lacks a unified framework that integrates chain-based reasoning, reflective mechanisms, and process supervision into a cyclic feedback loop. This leaves important room for developing more stable, more intelligent, and more self-auditing reasoning systems for large language models.

III. Proposed Framework

This study constructs a unified loop system consisting of question generation, reflective modeling, semantic calibration, and re-reasoning, enabling large language models to automatically identify potential risk points and form a sustainable self-supervised closed loop during the reasoning process. The overall architecture of the model is shown in Figure 1.

 doi:10.20944/preprints202602.0653.v1

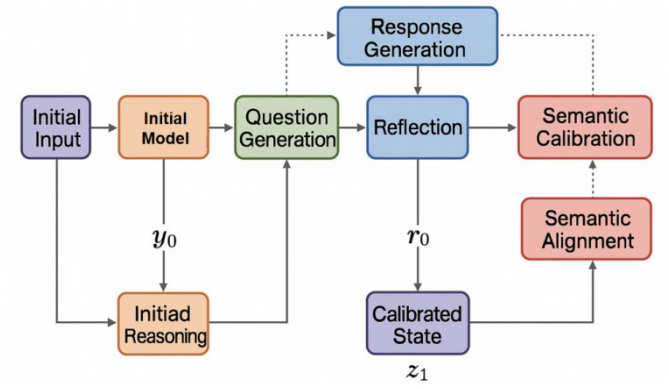

Figure 1. Overall Model Architecture.

First, in the initial inference phase, the model generates a preliminary answer y_0 based on the input sequence x, and then uses a conditional generation function:

$$y_0 = f_\theta(x) \quad (1)$$

Obtain the initial inference path. To enable the model to proactively detect potential inconsistencies, the system introduces a self-challenge mechanism. A parameterized challenge function performs a semantic scan of the current answer, generating a set of internal challenge sequences q_0, formally expressed as:

$$q_0 = g_\theta(x, y_0) \quad (2)$$

Here, q_0 represents spontaneous questions, risk point identification, and logical uncertainty identification related to the reasoning chain. The core objective of this stage is to enable the model to generate structured thinking signals for its own content, thereby establishing the first round of internal supervision for reasoning.

After obtaining the question sequence, the model enters a reflection phase, using a reflection function to interpret, respond to, and structurally reconstruct the question content. The reflection result is denoted as r_0, and given in the following form:

$$r_0 = h_\theta(x, y_0, q_0) \quad (3)$$

This process maps challenge signals into actionable reasoning feedback, including logical chain verification, semantic ambiguity identification, conflict information analysis, and potential hypothesis detection. The reflection phase is not a simple evaluation of the answer, but a generative internal feedback structure designed to provide binding semantic guidance for subsequent reasoning phases. To ensure that feedback signals truly participate in updating the reasoning chain, the system introduces a semantic consistency determination function in the reflection construction to match and align key content, defined as:

$$s_0 = Align(r_0, y_0) \quad (4)$$

Where s_0 represents the structured consistency representation between reflection and answer, used to reduce the noise impact of semantic drift and redundant feedback.

During the semantic calibration phase, the model transforms the reflection chain into fine-tuning signals that can be used for re-inference and re-parameterizes the original inference state using semantic calibration operators. The calibrated internal state z_1 is:

$$z_1 = Calibrate(r_0, y_0, s_0) \quad (5)$$

The calibration process integrates preliminary reasoning, questioning signals, and reflective feedback, enabling the model to dynamically adjust its reasoning representation space based on internal self-examination information, thereby forming a semantically more stable intermediate state. This mechanism plays a pivotal role in the recurring system, allowing reflective information to penetrate deeper into the model's reasoning structure rather than remaining at the external evaluation layer, thus substantially influencing subsequent reasoning steps.

Finally, in the re-inference phase, the system uses the calibrated state z_1 as the conditional input for the new round of inference, generating an updated inference result y_1, which is defined as follows:

$$y_1 = f_\theta(x, z_1) \quad (6)$$

 doi:10.20944/preprints202602.0653.v1

This process integrates reasoning, questioning, reflection, and calibration into a recursively expandable loop, enabling the model to continuously iterate and improve internally. Through multiple rounds of self-questioning, it achieves convergence and stability of the reasoning chain. This unified loop supervision system not only provides large language models with structured self-regulation capabilities but also endows the reasoning process with interpretability, diagnosability, and sustainable evolution, thus laying the foundation for building more reliable and transparent large model reasoning mechanisms.

IV. Experimental Analysis

A. Dataset

This study uses the open dataset e-SNLI as the primary data source. The dataset contains natural language inference pairs with human-annotated explanations. It extends traditional natural language inference by explaining each sample. The input, therefore, includes a premise, a hypothesis, and the corresponding reasoning statement. This structured form of reason and inference offers a natural basis for building questioning, reflection, and renewed reasoning mechanisms. It allows the model to learn interpretable logical structures and reflection cues from human reasoning chains.

Since e-SNLI covers the three relations of entailment, contradiction, and neutrality, its explanations include rich logical markers, contrastive cues, and key semantic evidence. This makes it well-suited for training a self-questioning mechanism. The model can use these explanation signals to generate questions about its own reasoning. It can also learn from the logical patterns in the explanations to identify inconsistency, semantic drift, or insufficient evidence. The paired natural language samples in the dataset contain diverse expressions, which support cross-sentence reasoning, semantic alignment, and reflective generation.

In addition, the open license of e-SNLI enables further exploration in complex reasoning, reflection-enhanced generation, and process supervision. It also allows natural integration with the cyclic supervision system proposed in this study. With the help of the explanatory annotations, the model can form a continuous structure that spans initial reasoning, question generation, reflection construction, and semantic calibration. This makes the dataset a key resource for exploring unified mechanisms of self-questioning and renewed reasoning. The dataset has a suitable scale, a clear structure, and well-defined logical tasks. It is therefore highly appropriate for the unified cyclic reasoning supervision framework proposed in this research.

B. Experimental Results

First, this paper carries out a comparative evaluation between the proposed framework and several representative baseline methods, and the detailed performance of each approach is summarized in Table 1. This experiment provides an overall view of how the method behaves under the same task setting as existing models.

Table 1. Comparative experimental results.

Method	Acc	BLEU	BERTScore	Self-Consistency Score
Powerinfer-2[8]	78.4	21.7	0.842	63.1
Openlem [9]	80.2	24.9	0.856	67.4
Tabi [10]	82.6	26.3	0.869	70.8
Bird [11]	83.1	27.5	0.873	72.2
Ours	87.8	33.6	0.904	81.4

Compared with representative existing models, the cyclic self-questioning framework proposed in this study shows consistent and significant advantages across several key metrics. Traditional methods rely on static single-round generation or limited process supervision. They often suffer from

 doi:10.20944/preprints202602.0653.v1

broken reasoning chains or local deviations when handling complex semantic relations or cross-sentence logical judgments. The results in the table show that single-stage reasoning models generally reach an accuracy range of 78 to 83 percent. In contrast, the proposed method embeds questioning, reflection, and semantic calibration into the reasoning chain. This allows the final reasoning stage to continuously absorb self-feedback and correct early unstable factors. As a result, accuracy increases to 87.8 percent, which demonstrates the strong effect of the cyclic supervision structure on decision quality.

For explanation consistency, the differences in BLEU and BERTScore are especially clear. They indicate a structural improvement in the quality of reflective content and final reasoning explanations. Traditional models depend on a single explanation step and often fail to reduce semantic drift between the input and the internal reasoning process. The semantic calibration module in this study integrates reflective content in a structured way. It provides more precise contextual information for the renewed reasoning stage. This significantly improves the alignment between generated explanations and the true semantic chain. The higher BLEU and BERTScore values show that the cyclic feedback mechanism enhances the coherence of the entire reasoning chain and makes explanation generation more reliable and stable.

More importantly, the self-consistency metric shows that the proposed method achieves higher stability and consistency across multiple rounds of reasoning. Traditional approaches often produce scattered or even contradictory answers in repeated inference because they lack internal correction mechanisms. The unified cyclic supervision system introduced in this study nests questioning, reflection, and calibration into a closed reasoning loop. It guides the model to converge step by step toward a consistent semantic state. Self-consistency increases from 63 to 72 percent in traditional models to 81.4 percent in the proposed approach. This shows that the model not only generates more accurate results but also maintains its reasoning trajectory more stably. It provides structural evidence for the effectiveness of cyclic self-supervision.

Furthermore, this paper investigates how different learning rate configurations influence the training behavior and final performance of the model, with the corresponding settings and outcomes organized in Table 2. This analysis helps to identify a reasonable learning rate range that balances convergence stability and optimization efficiency.

Table 2. The impact of the learning rate on experimental results.

Learning Rate	Acc	BLEU	BERTScore	Self-Consistency Score
0.0001	84.3	28.1	0.881	74.6
0.0005	86.1	30.4	0.892	78.3
0.0002	87.1	32.7	0.899	80.2
0.0003	87.8	33.6	0.904	81.4

The learning rate sensitivity experiment shows that the cyclic self-questioning framework has a clear and stable response pattern during optimization. A small learning rate, such as 0.0001, ensures stable gradient updates. However, it fails to fully absorb the high-level semantic feedback produced during questioning and reflection in the early training stage. This leads to lower performance in reasoning accuracy, explanation quality, and self-consistency. When the learning rate increases to a medium range, such as 0.0002 to 0.0003, the model can integrate questioning signals and reflective information more effectively. The semantic calibration module can adjust the reasoning chain in a more timely manner. As a result, BLEU, BERTScore, and self-consistency show clear improvements, and the semantic cycle becomes more coherent.

When the learning rate further increases to 0.0005, the model converges faster. Yet the larger update step reduces the ability to make fine-grained semantic adjustments guided by questioning and reflection. This causes slight declines in reasoning consistency and explanation stability. This trend aligns closely with the characteristics of the internal cyclic supervision mechanism. The

 doi:10.20944/preprints202602.0653.v1

framework relies on a stable semantic evolution process and injects questioning, reflection, and calibration into the reasoning structure step by step. A small learning rate limits absorption, while a large learning rate undermines the detailed structure of the reasoning chain. A medium learning rate achieves the best balance. The overall results confirm that the cyclic reasoning system requires careful tuning of the learning rate to ensure that internal reasoning feedback is absorbed fully and stably.

In addition, this paper designs a sensitivity study on the upper bound of the reflection length, and the corresponding configurations and observations are illustrated in Figure 2. By varying the maximum reflection length, the experiment explores how much reflective content is needed to effectively support multi-step reasoning without introducing excessive redundancy.

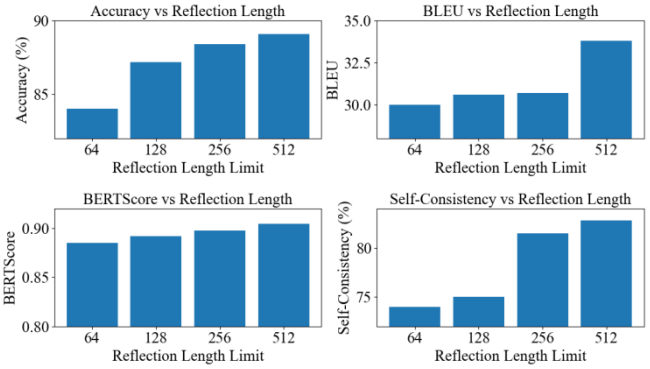

Figure 2. Reflecting on the impact of the upper limit of length on experimental results.

As the maximum reflection length is increased, the model's overall reasoning capability is consistently enhanced, indicating that richer reflective content provides more effective support for complex inference. When the reflection length is short, the model captures only limited information during the reflection stage. It cannot accurately locate errors in the intermediate reasoning chain or complete missing semantic links. This results in weaker performance in accuracy and BLEU. When the reflection length increases from 64 to 128 and 256, the model can generate more complete self-questioning content. The semantic coverage of the reflection stage expands. This allows the model to better identify weak points in its own reasoning and to form a more reliable logical path in the renewed reasoning stage.

A similar trend appears in explanation-related metrics. BERTScore improves steadily as the maximum reflection length increases. This indicates that a longer reflection space helps the model align the semantic calibration mechanism with the input more effectively. It enhances consistency within and across the reasoning chain. A short reflection length often leads to brief or fragmented reflective content and fails to address implicit relations in the reasoning process. A larger reflection length allows the model to generate more complete and better supported reflective text. This strengthens the semantic alignment between the final explanation and the input.

The self-consistency metric also increases with longer reflection lengths. This shows that a more complete reflection chain helps the model form a more stable semantic state during cyclic reasoning. A long reflection space enables the model to absorb feedback from self-questioning more fully in each reasoning round. This makes the internal semantic representation more stable. When the reflection length reaches 512, self-consistency achieves its highest level. This indicates that the flow of information across questioning, reflection, and calibration is used most effectively when the reflection capacity is large. The model maintains a high level of consistency across multiple reasoning rounds. This highlights the essential role of the cyclic self-supervision system in enhancing reasoning stability.

Finally, this paper examines the effect of different inference temperature settings on the behavior of the reasoning process, and the relevant curves and comparisons are presented in Figure 3. This part focuses on how the degree of sampling randomness influences not only accuracy but also the diversity and consistency of the generated reasoning paths.

 doi:10.20944/preprints202602.0653.v1

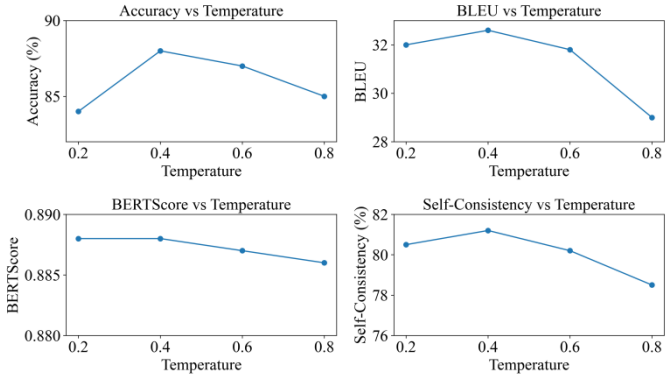

Figure 3. The effect of different inference temperature settings on experimental results.

Varying the reasoning temperature produces pronounced changes in decision stability and noticeably alters how well the model can internalize and utilize feedback within the cyclic reasoning process. At low temperature settings, the generation process becomes more deterministic. The information flow among questioning, reflection, and renewed reasoning remains stable and compact. At temperatures of 0.2 and 0.4, accuracy remains high. This indicates that the cyclic supervision structure can receive and utilize internal feedback more effectively in low randomness conditions and produce more reliable final reasoning results. When the temperature increases to 0.6 and 0.8, the randomness of generation grows. Reflective content and renewed reasoning paths become more prone to drift or local jumps, which leads to a gradual decline in accuracy.

A similar trend appears in explanation consistency metrics. BLEU and BERTScore remain high under low temperature settings. This shows that the model can generate reflective and explanatory texts with more stable semantic structures and maintain stronger coherence within the reasoning chain. As the temperature rises, the uncertainty of generation increases. The model is more likely to produce redundant or weakly related content during the reflection stage. This weakens its ability to maintain semantic alignment. At a temperature of 0.8, the evaluation metrics decrease noticeably. This indicates that high randomness disrupts the semantic calibration mechanism and weakens the correspondence between the explanation and the input evidence.

The self-consistency metric further reveals the deep influence of temperature on the cyclic reasoning framework. Self-consistency reaches its best levels under low temperature settings. This shows that the model can maintain a consistent logical state across multiple rounds of reasoning and that questioning, reflection, and renewed reasoning reinforce each other within a stable cycle. As temperature increases, the controllability of internal feedback declines. The model tends to produce divergent reasoning paths during consecutive cycles, which leads to a clear reduction in self-consistency. This shows that the cyclic supervision structure is highly sensitive to output stability and that temperature has a direct impact on whether its core mechanism can function effectively.

Overall, the results confirm that the cyclic chain of questioning, reflection, calibration, and renewed reasoning depends strongly on stability in the generation process. Moderate determinism helps ensure accurate transmission of reflective information and allows continuous optimization of the reasoning path. Excessive randomness weakens the feedback mechanism and prevents the model from maintaining a consistent semantic state across iterative cycles. The reasoning temperature experiment clearly illustrates how the cyclic self-supervision system behaves under different levels of generation control and provides important guidance for future model optimization.

This paper also presents an experiment on the sensitivity of inference parallelism settings to accuracy metrics, aiming to examine how different levels of parallel reasoning influence the stability and reliability of the model's decision process. By adjusting the degree of parallelism during the inference stage, the experiment evaluates how expanded or reduced branching paths interact with the cyclic self-supervision mechanism. This design helps reveal whether the internal feedback loop—consisting of questioning, reflection, and semantic calibration—can remain effective under varying

 doi:10.20944/preprints202602.0653.v1

inference configurations. The experimental results corresponding to this analysis are shown in Figure 4.

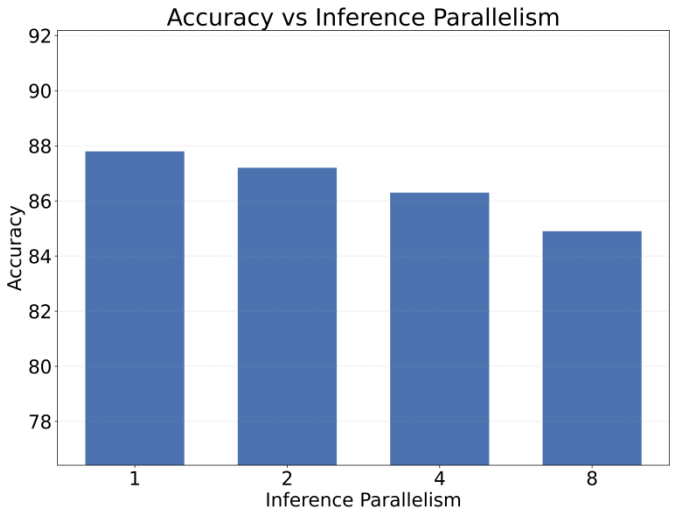

Figure 4. Sensitivity experiment of inference parallelism setting to accuracy metrics.

Adjusting the degree of inference parallelism leads to distinct differences in accuracy, and this influence is tightly coupled with the structural properties of the cyclic self-questioning framework itself. At low levels of parallelism, the model can fully absorb feedback from the questioning and reflection stages in each reasoning round. This allows the paths of semantic calibration and renewed reasoning to remain aligned and produces a stable and compact reasoning chain. The high consistency of this information flow promotes gradual convergence of the internal reasoning state. As a result, the model maintains high accuracy when the parallelism is set to 1 or 2.

When the parallelism increases to 4, the diversity of generated content expands. Different reasoning branches begin to diverge more in their semantic structure. The self-questioning mechanism can still correct local reasoning steps. However, the noise brought by multiple branches interferes with the ability of the reflection stage to capture the most important evidence. The calibration step cannot fully cover all reasoning paths. This leads to a slight decrease in accuracy. The results suggest that moderate parallelism can provide useful redundancy, but too many branches weaken the focused feedback effect of the cyclic supervision mechanism.

When the parallelism increases to 8, the number of reasoning branches grows sharply. The model faces greater uncertainty in the renewed reasoning stage, and the internal semantic alignment becomes harder to maintain across branches. Conflicts among a large number of parallel reasoning results increase the burden on semantic calibration. The internal representation space struggles to form a unified semantic center during iteration. This causes accuracy to drop more noticeably. High parallelism enlarges the exploration space, but it reduces the effectiveness of the reflection chain within the cyclic reasoning system.

Overall, the results further confirm the structural dependency of the cyclic self-questioning framework on reasoning stability. Moderate determinism and a limited number of generated branches help maintain a robust semantic feedback loop and allow the model to make full use of the internal supervision signals provided by questioning and reflection. Excessive parallelism disrupts the consistency of the internal iterative chain and makes it difficult for the reasoning process to maintain a coherent direction under noise. This highlights the need to tune inference parameters carefully around the cyclic supervision mechanism to ensure that the model remains in a reasoning state conducive to convergence.

V. Conclusion

 doi:10.20944/preprints202602.0653.v1

This study proposes a cyclic self-questioning supervision framework for large language models. The framework integrates questioning, reflection, semantic calibration, and renewed reasoning into an iterative reasoning process and imposes structured constraints on generative reasoning. Unlike traditional approaches that rely on single-step thinking or static chain-based reasoning, the proposed framework generates internal feedback signals throughout the reasoning process. These signals guide the model to identify potential risks and gradually correct its reasoning path. This leads to significant improvements in logical consistency and semantic stability. The experimental results confirm the advantages of this mechanism across accuracy, explanation consistency, and self-consistency. The findings provide a solid foundation for building more reliable reasoning structures for large models.

The proposed cyclic supervision framework shows potential value in many application scenarios. For tasks that require strict logical consistency, such as legal reasoning, medical analysis, financial auditing, and safety-related decision-making, the framework provides a traceable and decomposable reasoning chain. It enables higher trustworthiness when the model processes high-risk information. For tasks that involve multi-step reasoning or complex semantic integration, such as long text question answering, knowledge graph reasoning, and cross-document information fusion, the method reduces semantic drift and improves stability across paragraph-level reasoning. By forming a cycle that evolves from thinking to questioning, reflection, and renewed reasoning, the model gains stronger robustness and interpretability. This supports the deployment of high-reliability intelligent systems.

In addition, the roles of reflection and questioning within the cycle not only improve output quality but also provide an initial form of self-monitoring in the reasoning behavior of large models. This capability is important for building autonomous intelligent agents. It allows the model to use internal feedback to adjust its behavior in complex environments and to maintain greater consistency and rationality during task execution. The semantic calibration mechanism strengthens model stability when processing information from multiple sources. This makes the framework suitable for cross-modal, multi-domain, and knowledge-intensive tasks. As large models move into broader, more sensitive, and more complex scenarios, the proposed framework becomes an important path for improving safety, reliability, and controllability.

Future work can explore the adaptability of the cyclic self-supervision framework in larger models, more diverse reasoning tasks, and multilingual settings. Integrating the self-questioning mechanism with external knowledge bases, retrieval augmented methods, or structured reasoning systems may lead to a more comprehensive reasoning ecosystem. Further research is needed on designing finer-grained reflection structures, more stable semantic calibration models, and more efficient cyclic control strategies. As the depth of reasoning demanded by applications continues to increase, cyclic reasoning may become a new paradigm for large models. The results of this study provide an extensible theoretical and technical foundation for that development.

References

1. Mu J, Zhang Q, Wang Z, et al. Self-Reflective Generation at Test Time[J]. arXiv preprint arXiv:2510.02919, 2025.
2. Xu Y, Cheng Y, Ying H, et al. SSPO: Self-traced Step-wise Preference Optimization for Process Supervision and Reasoning Compression[J]. arXiv preprint arXiv:2508.12604, 2025.
3. Sun Z, Shen Y, Zhou Q, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision[J]. Advances in Neural Information Processing Systems, 2023, 36: 2511-2565.
4. Jeong M, Sohn J, Sung M, et al. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models[J]. Bioinformatics, 2024, 40(Supplement_1): i119-i129.
5. Jang H, Jang Y, Lee S, et al. Self-Training Large Language Models with Confident Reasoning[J]. arXiv preprint arXiv:2505.17454, 2025.
6. Chen L, Prabhudesai M, Fragkiadaki K, et al. Self-questioning language models[J]. arXiv preprint arXiv:2508.03682, 2025.

 doi:10.20944/preprints202602.0653.v1

7. Weng Y, Zhu M, He S, et al. Large language models are reasoners with self-verification[J]. arXiv preprint arXiv:2212.09561, 2022, 2.
8. Xue Z, Song Y, Mi Z, et al. Powerinfer-2: Fast large language model inference on a smartphone[J]. arXiv preprint arXiv:2406.06282, 2024.
9. Mehta S, Sekhavat M H, Cao Q, et al. Openelm: An efficient language model family with open training and inference framework[J]. arXiv preprint arXiv:2404.14619, 2024.
10. Wang Y, Chen K, Tan H, et al. Tabi: An efficient multi-level inference system for large language models[C]//Proceedings of the Eighteenth European Conference on Computer Systems. 2023: 233-248.
11. Feng Y, Zhou B, Lin W, et al. Bird: A trustworthy bayesian inference framework for large language models[J]. arXiv preprint arXiv:2404.12494, 2024.