Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37201

An Empirical Investigation of Whether Role Asymmetry Can Overcome Reasoning Blind Spots in Homogeneous Models through Adaptive Routing and Local Repair

Wenke Li

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
u202414221@hust.edu.cn

Abstract. Under the test-time compute scaling paradigm, improving the multi-step reasoning accuracy of lightweight large language models within a limited computational budget remains a critical challenge. Existing self-correction and multi-agent debate methods often suffer from rigid interaction schemes, coarse-grained correction, and homogeneous role configurations. Whether role-asymmetric constraints can enable a critic agent to overcome the reasoning biases of its generator has rarely been tested directly. This paper develops a collaborative reasoning framework combining adaptive routing with role-asymmetric local repair, treating this hypothesis as a falsifiable research question, and conducts systematic experiments on two backbone models, Qwen2.5-7B-Instruct and Qwen3-8B. Results show that the framework's disadvantage relative to simple baselines is capability-dependent: on the weaker model, it is statistically indistinguishable from standard chain-of-thought and other single-model baselines while significantly outperforming multi-agent debate; on the stronger model, it significantly underperforms most baselines. A sanity check confirms that the critic reliably identifies clearly structured arithmetic errors; however, its recall drops sharply on subtle errors requiring cross-step verification of quantitative relationships, and the net effect of repair mirrors this pattern, close to neutral on the weaker model and clearly negative on the stronger one. Ablation experiments show that the marginal effects of routing and role asymmetry reverse with model capability, turning negative on the stronger model even as removing either component improves its accuracy. These findings provide reproducible empirical evidence for systematic reasoning blind spots in self-criticism by homogeneous models, and indicate that role asymmetry implemented only through prompting cannot reliably overcome cognitive homogeneity.

Keywords: large language models, multi-agent collaborative reasoning, role asymmetry, local repair, limitations of self-criticism

1. Introduction

In recent years, large language models have shown strong capabilities on complex tasks such as mathematical reasoning and code generation. Prior studies have shown that, with model parameters

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37201

held fixed, allocating additional computation at inference time can substantially improve model performance, establishing test-time compute scaling as an increasingly important research direction in both academia and industry [1]. This paradigm is particularly appealing for lightweight models with approximately 7–8 billion parameters. Such models offer low deployment costs and fast response speeds, but they are prone to logical errors in multi-step reasoning. If their accuracy can be improved consistently through a controllable amount of additional inference computation, their practical deployment scope could be significantly expanded.

Existing efforts toward this goal can be broadly divided into three directions: prompt-based single-model methods, represented by chain-of-thought prompts and self-consistency voting [2, 3]; the iterative self-correction methods, which prompt models to reflect on and revise their own outputs [4]; and multi-agent debate method, in which multiple model instances cross-examine one another [5]. However, several recent systematic studies have questioned the latter two approaches. A critical survey by Kamoi et al. concludes that, absent external feedback, no prior work convincingly demonstrates that large language models can reliably self-correct their own reasoning outside of tasks exceptionally suited to self-correction [6]. The tradeoff between computational cost and accuracy that such comparisons must confront is itself a long-standing concern in machine learning, extending well beyond debate-based methods [7]. Several critical evaluations of multi-agent debate have reached similar conclusions [8, 9]. These limitations can be summarized in three aspects: fixed interaction rounds prevent computation from being allocated adaptively; coarse-grained correction through global rewriting can introduce new errors; and homogeneous roles cause critics and generators to share similar reasoning biases, making errors difficult to detect.

To address these issues, this paper proposes AR-ALR(Adaptive Routing and Asymmetric Local Repair), a collaborative reasoning framework based on adaptive routing and role-asymmetric local repair. The framework organizes generation, criticism, and repair into an explicit iterative state-transition architecture: under asymmetric constraints on both roles and information access, the critic progressively localizes erroneous steps; the routing module dynamically determines, according to confidence, whether to produce the output directly or enter the repair loop; and the repair agent replaces only the identified erroneous steps. Unlike prior work, this paper do not merely propose a mechanism and report its pointwise performance. Instead, this paper treats the role-asymmetry hypothesis itself as the object of empirical investigation. This paper first conducts sanity checks of the criticism mechanism to reduce potential interference from implementation-level factors, and then perform parallel experiments on two backbone models with different capability levels, Qwen2.5-7B-Instruct and Qwen3-8B, to systematically examine whether the hypothesis holds across model capabilities.

The significance of this study lies in three aspects. First, motivated by practical deployment requirements, this paper explores a feasible approach to maximizing the reasoning capability of lightweight models under a limited computational budget, providing empirical evidence for the design of low-cost reasoning systems. Second, through rigorous cross-model comparisons, this paper directly addresses recent concerns about the effectiveness of multi-agent paradigms and homogeneous-model self-criticism. The resulting evidence is informative regardless of whether it supports or challenges the original hypothesis. Third, the judgment records and failure cases collected from the critic agents provide empirical material for analyzing the consistency and biases of models acting as evaluators, thereby contributing to research on the reliability of large language model evaluation.

DOI: 10.54254/2755-2721/2026.37201

2. Related work

2.1. Single-model reasoning baselines and test-time compute scaling

The tradeoff between computation and solution quality has been studied in artificial intelligence for decades, predating large language models. Zilberstein formalized this tradeoff through the framework of anytime algorithms, which can be interrupted at any point to yield a usable result whose quality improves with additional computation time [10]. In the context of test-time compute scaling, Snell et al. systematically investigated the relationship between test-time computation and model performance under a fixed reasoning budget, comparing strategies including iterative self-improvement and search based on process reward models [1]. Their empirical results showed that allocating computation adaptively according to problem difficulty is often more effective than simply increasing the number of pretrained parameters, which provides empirical support for test-time compute scaling. Chain-of-thought prompting established a foundational paradigm based on guiding logical deduction through intermediate steps [2]. Self-consistency decoding, by sampling multiple paths and voting, initiated the precedent of expanding the calculation during the inference period through repeated sampling [3]. Liu et al. situate both techniques within a broader survey of prompting-based methods for language models [11]. Although these approaches can improve reasoning performance, they generally rely on implicit logical deduction or indiscriminate repeated sampling and lack mechanisms for targeted error localization and correction.

2.2. Self-correction mechanisms and their inherent limitations

To introduce explicit error correction, self-correction frameworks such as Self-Refine have constructed an iterative optimization mechanism based on natural language self-feedback, prompting the model to repeatedly evaluate and rewrite its own output [4]. However, the critical survey by Kamoi et al. synthesizes evidence across the literature indicating that single-model self-correction has inherent limitations in pure logical reasoning tasks without external supervision [6]. Cognitive homogeneity and reasoning biases during generation and criticism can create blind spots in error detection, and several of the surveyed studies report that self-correction may even cause correct results to be altered. Moreover, correction through full-answer rewriting can easily disrupt previously correct steps, resulting in cascading errors and substantial Token overhead.

2.3. Multi-agent debate and the bottleneck of blind collaboration

To mitigate the reasoning biases of a single model, Du et al. proposed a multi-agent debate mechanism, enabling multiple model instances to debate and cross-examine one another to improve logical rigor [5]. However, a growing body of recent research has revealed fundamental bottlenecks in this paradigm. Wang et al. found that agents in standard multi-agent debate frequently operate under inconsistent and incomplete knowledge, forming isolated "cognitive islands" that impede convergence toward correct answers [8]. Zhang et al. further argued that the reported benefits of multi-agent debate are often overstated due to evaluation methodology, and that model heterogeneity is critical to realizing genuine gains [9]. A key underlying issue is that existing debate frameworks generally rely on homogeneous role prompts and fixed interaction rounds, while lacking fine-grained error localization and role-asymmetric constraints. Consequently, multiple instances essentially still use similar thinking patterns to mutually confirm each other. This tendency echoes a much older observation from group decision-making research: McCauley's experimental work on

DOI: 10.54254/2755-2721/2026.37201

groupthink distinguishes compliance-based from internalization-based mechanisms by which members of a cohesive group suppress independent judgment under pressure toward unanimity [12].

2.4. Budget-aware control and verification mechanisms

The computational cost caused by multi-round interaction has made budget-aware control an increasingly important trend. Zhou discusses this cost-accuracy tradeoff in the context of ensemble methods, where combining multiple models can improve accuracy at a proportional increase in computational cost, and notes that the practical benefit of such combination is easily overestimated when cost differences are overlooked [7]. In terms of verification, research on learning from erroneous examples offers a complementary perspective: Dieterich et al., in a systematic review, found that identifying and explaining errors at the level of individual steps consistently supports deeper learning than evaluating only final outcomes [13]. External feedback has also been used for verification and correction. A related idea has a longer history in software engineering, where automated program repair methods such as GenProg use the results of external test execution to guide the correction of faulty code [14]. In a survey of LLM-based autonomous agents, Wang et al. described reflection mechanisms such as Reflexion, which uses verbalized self-critique to improve later decisions [15]. Collectively, these approaches point toward a verification paradigm that is fine-grained and traceable.

In summary, existing research has established a relatively mature technical foundation for test-time scaling, multi-agent interaction, and budget control. Nevertheless, three major limitations remain: fixed interaction rounds, which constrain adaptive allocation of computation; global rewriting, which can introduce errors into previously correct reasoning; and role homogeneity, which can create blind spots in error detection. The AR-ALR framework is designed to address this research gap. However, rather than assuming the effectiveness of the proposed mechanism and merely reporting its performance, this paper further uses rigorous cross-model experiments to investigate the conditions under which the core hypothesis of role asymmetry actually holds.

3. Methodology

Maximizing the reasoning accuracy of lightweight large language models under a limited computational budget is a central challenge in test-time compute scaling. Existing multi-agent debate and self-correction mechanisms suffer from two major limitations. First, homogeneous roles and symmetric prompts can amplify cognitive biases. Second, global rewriting-based correction can disrupt the steps that have already been derived correctly, leading to error cascades and significant Token overhead. To address these issues, this paper proposes a local-repair reasoning framework based on adaptive routing and role-asymmetric criticism, AR-ALR. By jointly employing routing, asymmetric criticism, and local repair, the framework aims to balance reasoning accuracy with on-demand allocation of computational resources.

3.1. Problem definition and symbolic representation

This paper formulates the reasoning task as the problem of finding an optimal answer under a computational budget constraint. For a question Q, the reasoning process generated by the model can be decomposed into a chain of thought S consisting of n sequential steps:

DOI: 10.54254/2755-2721/2026.37201

$$S = \{s_1,\ s_2,\ \ldots,\ s_n\} \tag{1}$$

In the formula (1), S represents the complete set of reasoning steps generated by the model for Q, si $(i \in \{1, \ldots, n\})$ represents the i-th step, and n denotes the total number of steps in the chain. The optimization objective is to maximize the probability that the final answer matches the reference answer under a hard budget B, defined as the upper bound on the total number of Tokens consumed by all agent interactions for a single query:

$$max_{\pi}\ P\big(A = A^{*} \big| Q, \pi\big),\quad s.t.\quad C\big(S\big) \le B \tag{2}$$

In the formula (2), A^{*} represents the reference answer, A represents the final answer produced by the system under policy π, and π represents the control policy composed of the router, critic, and repair operator. $C(S)$ represents the cumulative number of Tokens consumed to generate and repair S, including routing samples, critic verification, and local rewriting, while B denotes the predefined upper bound on the Token budget.

3.2. Overall system architecture and collaborative workflow

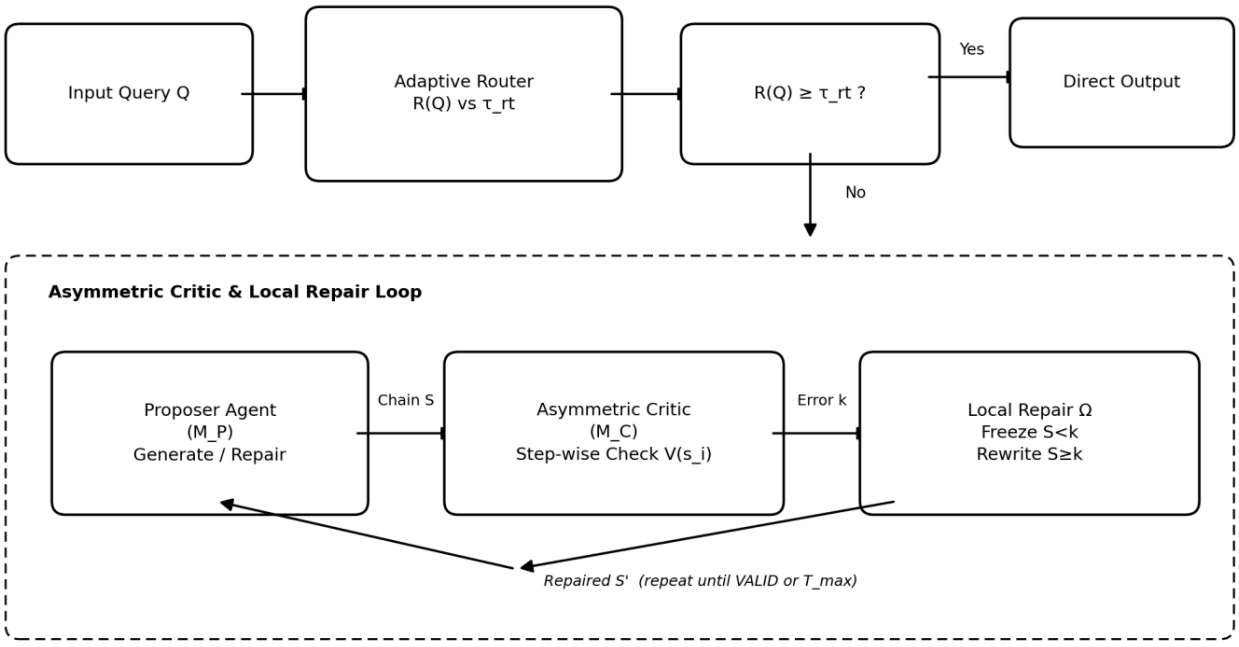


Figure 1. Overall architecture and data flow of the AR-ALR framework (photo credit: original)

The AR-ALR framework consists of an adaptive router, an asymmetric critic network, and a local repair operator, as illustrated in Figure 1. The router first assesses the difficulty of question Q, and simple questions are output directly. Otherwise, the generator produces an initial chain of thought. The critic then verifies the reasoning step by step and identifies the first erroneous step k. The repair

DOI: 10.54254/2755-2721/2026.37201

operator freezes the prefix and reconstructs only the suffix beginning at k. The repaired chain is then submitted for verification again. This process forms a closed loop of "localization–freezing–repair–reverification" until verification succeeds or the maximum number of rounds Tmax is reached.

3.3. Adaptive routing and computational budget allocation

To avoid unnecessarily invoking multi-round interactions for simple problems, this paper introduces an adaptive routing mechanism based on sampling confidence. For a question Q, the backbone model M performs K independent decoding samples at temperature Tsample. For each sample, this paper computes the average log-likelihood of its first step and exponentiate the result to obtain the confidence score $R(Q)$:

$$R(Q) = (\frac{1}{K}) \sum_{k=1}^{K} exp[\frac{1}{|s_{1,k}|} \sum_{t=1}^{|s1,k|} log\, p(w_t|w_{<t}, Q)] \tag{3}$$

In the formula (3), K represents the number of samples, $s1, k$ represents the first step of the k-th sample, $|s1, k|$ represents its Token count, t ranges over all Token positions within that step, wt represents the t-th Token, and $p(wt|w < t, Q)$ represents the corresponding conditional probability. $R(Q)$ takes values in (0,1]; a value closer to 1 indicates higher confidence in the first step. If the first step of a given sample is empty, that sample is assigned a confidence of 0 when computing the average. The routing decision rule is as follows:

$$Action = \{Direct_Output,\ if\ R(Q) \geq \tau_{rt}\ ;\ Multi_Agent_Loop,\ if\ R(Q) < \tau_{rt}\} \tag{4}$$

In the formula (4), τrt represents the routing confidence threshold. When $R(Q) \geq \tau rt$, the system directly produces the output; otherwise, it enters the criticism–repair loop. Because this criterion relies solely on the local confidence of the first step, it may misclassify problems in which the first step is simple but the subsequent reasoning becomes substantially more difficult. An artificially high confidence in the first step may therefore cause the router to incorrectly pass the problem through and bypass repair and verification that would otherwise be required. This limitation was specifically analyzed in the experimental section.

3.4. Role-asymmetric criticism and process verification

In traditional multi-agent debates, each agent has the same role and directly outputs the answer. This easily leads to the accumulation of thinking patterns and blind conformity. The premise that differentiated roles can yield better outcomes than symmetric ones has a long history in economic theory: Akerlof showed that asymmetries in the information or role held by different parties in an interaction can fundamentally alter the resulting outcome [16]. This paper therefore imposes role-asymmetric constraints in the proposed framework: the generator agent MP has full authority to generate answers, whereas the critic agent MC is explicitly prohibited from directly generating an answer and is restricted to step-wise verification and structured diagnosis. The generator and critic in the implementation share the same backbone model weights; their roles are differentiated solely through prompt structure and output permissions, without independent training or parameter isolation. Whether this design will lead to a "self-review" type cognitive blind spot is a key issue to

DOI: 10.54254/2755-2721/2026.37201

be examined in the experimental part. The critic scans the reasoning chain S and produces a judgment $V(si)$ and a diagnosis Ci for each step si :

$$V(s_i),\ C_i = M_C(Q,\ S^{\leq i},\ C_{i-1}) \tag{5}$$

In the formula (5), $V(si) \in \{0,1\}$ represents the critic's judgment of the i-th step, where 1 indicates that the step passes verification and 0 indicates that an error is present. $S \leq i$ represents the prefix of the first i steps, $Ci-1$ represents the accumulated diagnostic context, and Ci represents the diagnostic feedback for the i-th step.

3.5. Fine-grained error localization and local repair operator

Once an error is identified, traditional Self-Refine-style approaches regenerate the entire answer, potentially disrupting previously correct steps and introducing cascading errors. To address this issue, this paper proposes a local repair operator Ω, which first identifies the index k of the earliest erroneous step:

$$k = min\{\ i \in \{1,2,\ldots,n\} \mid V(s_i) = 0\ \} \tag{6}$$

In the formula (6), k represents the smallest step index at which a deviation is detected. If no step satisfies $V(si) = 0$, setting $k = \infty$, indicating that all steps have passed verification and the loop is terminated. Once k is identified, the operator freezes the prefix $S < k$ and provides the generator only with the prefix, Q, and diagnosis Ck , from which a new suffix beginning at sk is reconstructed:

$$S_{\geq k'} = M_P(Q,\ S_{<k},\ C_k) \tag{7}$$

In the formula (7), $S \geq k'$ represents the repaired suffix newly generated by the generator based on the diagnosis. The repaired complete chain of thought is then updated by concatenation:

$$S' = S_{<k}\ \cup\ S_{\geq k'} = \{s_1,\ldots,s_{k-1},s_{k'},\ldots,s_{m'}\} \tag{8}$$

In the formula (8), the new chain S' is submitted for verification again, forming a closed loop that continues until verification succeeds or Tmax is reached. If verification still fails after Tmax rounds, the system directly uses the current repaired S' as the final output without enforcing further correction. This strategy guarantees termination within the budget, at the cost of that a small number of problems may be submitted without passing verification.

4. Experimental results and analysis

This chapter focuses on three specific questions: whether role-asymmetric constraints improve the ability of homogeneous critic agents to detect errors generated by the generator; whether local repair

DOI: 10.54254/2755-2721/2026.37201

can bring stable accuracy gains under a limited reasoning budget; and which types of reasoning errors are most likely to cause the mechanism to fail. To avoid the conclusions being influenced by the specificity of a single model, this paper conducts parallel experiments using two backbone models with different capability levels, Qwen2.5-7B-Instruct and Qwen3-8B. Sections 4.2 to 4.4 respectively present empirical results around the above three issues.

4.1. Experimental setup

The experimental backbones are Qwen2.5-7B-Instruct and Qwen3-8B, both of which are open-source lightweight models [17, 18]. The generator and critic share the same backbone weights and are differentiated solely by their prompt structures and output permissions. Experiments are conducted on the GSM8K test set, using 500 samples for both backbone models to support direct cross-model comparison [19]. Because Fixed-Round MAD incurs a substantially higher per-query cost than the other methods (approximately 8,000–9,400 Tokens per query, roughly 8–19 times that of standard CoT) due to its fixed three-round, three-agent interaction process, its evaluation on Qwen2.5-7B-Instruct is limited to the first 150 of these 500 samples; all other methods, including AR-ALR and its ablation variants, are evaluated on the full 500-sample set for both backbone models.

The comparison baselines include standard chain-of-thought (CoT), using a single step-by-step reasoning process; chain-of-thought self-consistency (CoT-SC), using ($k = 3$) independent samples with majority voting; direct self-correction, implemented through Self-Refine-style global rewriting; and fixed-round multi-agent debate (MAD), using three homogeneous roles with a fixed three-round interaction process. The evaluation metric is exact-match accuracy:

$$Accuracy = \left(\frac{1}{N}\right) \sum_{i=1}^{N} 1\left[A_i = A_i{}^{*}\right] \tag{9}$$

In the formula (9), N represents the total number of test samples, Ai and $A_i{}^{*}$ represent the system output and reference answer, respectively, and $1[\cdot]$ represents the indicator function. This paper also measures the average Token cost:

$$AvgTokens = \left(\frac{1}{N}\right) \sum_{i=1}^{N} C(S_i) \tag{10}$$

In the formula (10), $C(Si)$ represents the total number of Tokens consumed by all agent interactions for the i-th question. Using standard CoT as the reference, this paper calculates the relative cost as follows:

$$RelativeCost = \frac{AvgTokens_{method}}{AvgTokens_{CoT}} \tag{11}$$

In the formula (11), A RelativeCost closer to 1 indicates that the computational cost approaches the lower bound associated with a single inference.

DOI: 10.54254/2755-2721/2026.37201

In addition, to examine whether the criticism mechanism itself contains implementation defects, this paper constructs several examples involving clearly structured arithmetic errors and directly evaluate the critic's error-detection capability. All errors are correctly identified. This result only shows that the criticism mechanism functions properly for explicit and isolated arithmetic errors. It reduces, but does not eliminate, the possibility that the subsequent observed failures are caused by fundamental parsing or prompt word defects. It is not sufficient to prove that there are no implementation problems in the entire pipeline.

4.2. Main experimental results

Table 1 and Figure 2 report the comparative results for the different methods on Qwen2.5-7B-Instruct, while Table 2 and Figure 3 present the corresponding results for Qwen3-8B.

Table 1. Accuracy and computational cost of different methods on Qwen2.5-7B-Instruct

Method	N	Accuracy (%)	Average Tokens per Query	Relative Cost
Standard CoT	500	84.8	348.2	1.00×
CoT-SC (k=3)	500	89.4	1066.5	3.06×
Direct Self-Correction	500	85.4	978.7	2.81×
Fixed-Round MAD	150	72.7	8977.4	25.78×
AR-ALR	500	87.2	1532.9	4.40×

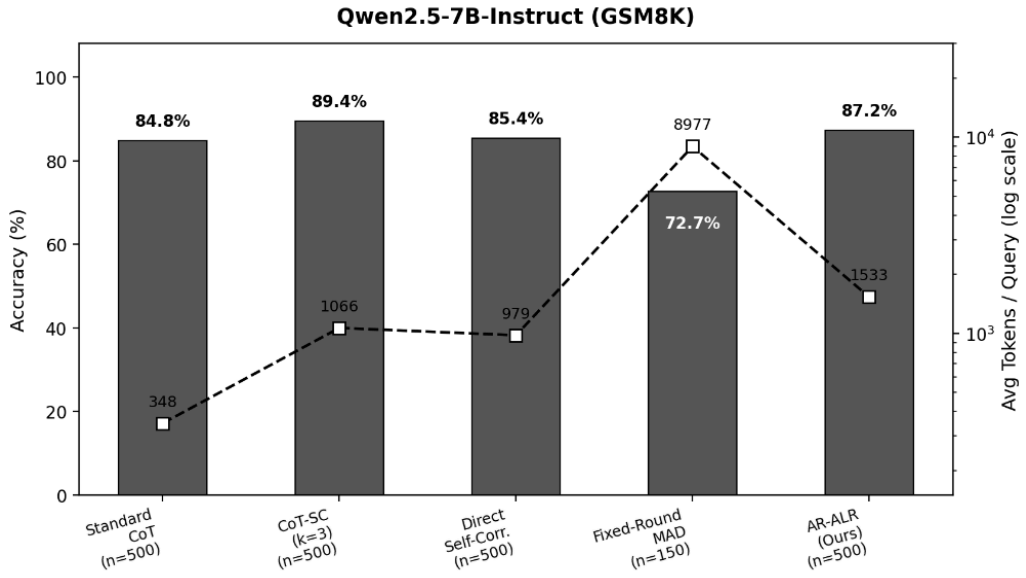


Figure 2. Comparison of accuracy and token cost across methods on Qwen2.5-7B-Instruct (photo credit: original)

DOI: 10.54254/2755-2721/2026.37201

Table 2. Accuracy and computational cost of different methods on Qwen3-8B

Method	N	Accuracy (%)	Average Tokens per Query	Relative Cost
Standard CoT	500	91.4	369.3	1.00×
CoT-SC (k=3)	500	91.6	1121.4	3.04×
Direct Self-Correction	500	91.8	1040.9	2.82×
Fixed-Round MAD	500	93.2	9424.4	25.52×
AR-ALR	500	88.2	1705.5	4.62×

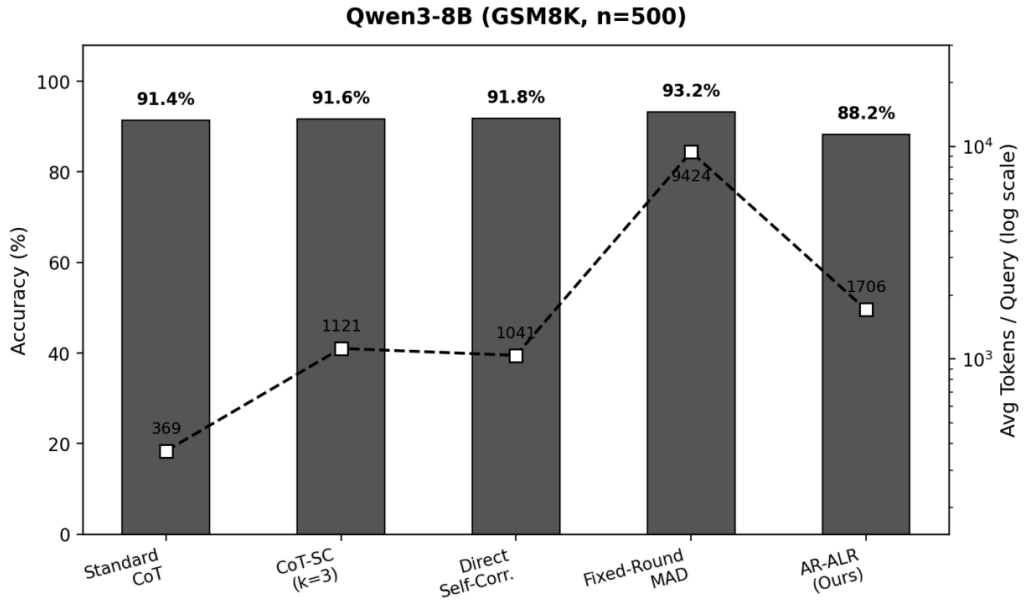


Figure 3. Comparison of accuracy and token cost across methods on Qwen3-8B (photo credit: original)

As shown in Table 1, on Qwen2.5-7B-Instruct, AR-ALR (87.2%) is statistically indistinguishable from Standard CoT (84.8%, p=0.35), CoT-SC (89.4%, p=0.30), Direct Self-Correction (85.4%, p=0.56), and the w/o Adaptive Router variant (85.0%, p=0.30) under the McNemar test, while significantly outperforming Fixed-Round MAD (72.7%, p=0.002). In contrast, Table 2 shows that on Qwen3-8B, AR-ALR (88.2%) performs significantly worse than CoT-SC, Direct Self-Correction, and MAD (all $p<0.05$). Taken together, the two sets of results reveal a capability-dependent pattern: on the weaker model, AR-ALR holds its own against simple single-model baselines and clearly outperforms traditional multi-agent debate; on the stronger model, this advantage disappears and reverses into a significant disadvantage relative to most baselines, including MAD. Rather than scaling up with model capability as initially expected, the benefit of the proposed mechanism instead erodes as the backbone model becomes stronger.

4.3. Attribution analysis

To answer the question of whether role-asymmetric constraints genuinely improve error-detection capability, this paper decomposes the performance before and after local repair is triggered into

DOI: 10.54254/2755-2721/2026.37201

routing and repair contributions. This paper further conducts ablation experiments to directly evaluate the marginal effects of routing and role asymmetry. Table 3 compares the full AR-ALR framework with two variants: "without adaptive routing," which forces all problems to enter the repair loop, and "without role asymmetry," in which the critic is reduced to a homogeneous role that can directly provide an ansor.

Table 3. Ablation study of routing and role-asymmetric constraints (accuracy, %)

Method Variant	Qwen2.5-7B Accuracy (%)	Qwen3-8B Accuracy (%)
AR-ALR (Full)	87.2	88.2
w/o Adaptive Router	85.0	89.6
w/o Asymmetric Critic	83.4	92.0

Table 3 shows that the marginal effects of routing and role-asymmetric constraints reverse with backbone model capability. On Qwen2.5-7B-Instruct, removing routing (85.0%) or role asymmetry (83.4%) reduces accuracy relative to the full framework (87.2%), indicating that both components contribute positively on the weaker model. On Qwen3-8B, however, removing routing (89.6%) or role asymmetry (92.0%) instead increases accuracy relative to the full framework (88.2%), with removing role asymmetry producing the largest improvement. This result provides concrete and quantifiable counterfactual evidence regarding whether role asymmetry can overcome the reasoning biases of homogeneous critics: role asymmetry implemented at the prompting level exhibits a positive effect only on the weaker model, whereas on the stronger model, the existence of this constraint actually hinders the overall performance.

A further decomposition of the repair stage itself reveals a pattern that tracks the same capability-dependent trend. On Qwen2.5-7B-Instruct, among the 288 problems that enter the local repair loop, 3 of the 50 problems with initially incorrect answers are successfully repaired (6.0%), while 2 of the 238 problems with initially correct answers are incorrectly modified (0.8%), yielding a small net-positive effect of one additional correct answer overall. On Qwen3-8B, by contrast, 9 of the 45 problems with initially incorrect answers are successfully repaired (20.0%), whereas 23 of the 447 problems with initially correct answers are incorrectly modified (5.1%), yielding a net effect of fourteen fewer correct answers overall. The sign and magnitude of this net effect mirror the accuracy comparisons in Section 4.2: on the weaker model, where the net effect of repair is close to neutral, AR-ALR remains statistically indistinguishable from the simpler baselines; on the stronger model, where the net effect turns clearly negative, AR-ALR falls significantly behind. This pattern also explains the observation in Table 3 that removing repair-related constraints improves accuracy on Qwen3-8B but reduces it on Qwen2.5-7B-Instruct: the component being removed is not uniformly harmful, but rather one whose net contribution itself depends on backbone model capability.

4.4. Error case analysis

Error-case analysis shows that all failure cases are incorrectly judged as valid during the criticism stage and are concentrated on subtle errors that require cross-step verification of quantitative relationships, rather than isolated single-step arithmetic errors. For example, in a mood-statistics problem, the generator correctly counted 12 good-mood days among the 30 days in the first step, but in the final step it counted the 2 good-mood days occurring in the last 3 days a second time, producing 14 instead of the reference answer of 2. The critic failed to identify this cross-step double counting during verification. In another interest calculation problem, the generator incorrectly

DOI: 10.54254/2755-2721/2026.37201

treated simple interest as compound interest, and the critic likewise failed to detect the error. These observations suggest that critics built on a homogeneous backbone model tend to follow the generator's existing narrative logic and verify the reasoning locally, rather than actively cross-checking quantitative relationships across steps. This provides empirical support at the mechanism level for the existing judgment that self-criticism by homogeneous models is subject to systematic reasoning blind spots [6]. The specific failure pattern observed here—correctly executing individual steps while failing to detect an inconsistency in how quantities carry across steps—parallels a distinction long established in research on procedural errors in mathematical problem solving, where Brown and Burton showed that many errors reflect systematic, rule-governed misapplications of otherwise-correct local procedures rather than random mistakes [20].

5. Conclusion

This paper focuses on multi-step reasoning problem of lightweight large language models under a constrained computational budget, proposing an AR-ALR collaborative reasoning framework that integrates adaptive routing, role-asymmetric criticism, and local repair. The central hypothesis—whether role-asymmetric constraints can enable critic agents to overcome the reasoning biases of generator agents—is taken as a falsifiable research question. Systematic experiments are conducted on two backbone models with different capability levels, Qwen2.5-7B-Instruct and Qwen3-8B. Results show that the disadvantage of the proposed framework relative to simple baselines is capability-dependent: on the weaker model, it is statistically indistinguishable from standard chain-of-thought and other single-model baselines while significantly outperforming fixed-round multi-agent debate; on the stronger model, it significantly underperforms most baselines. Its primary contribution therefore lies not in "proposing a superior method," but in providing a clear and quantifiable empirical test of the underlying hypothesis. Sanity checks of the criticism mechanism show that the implementation can reliably identify clearly structured arithmetic errors, reducing the likelihood that the observed failures are attributable to basic implementation defects. However, when confronted with subtle errors in real problems that require cross-step verification of quantitative relationships, the critic exhibits substantially lower error-detection recall, and the net effect of local repair mirrors the same capability-dependent pattern: close to neutral on the weaker model and clearly negative on the stronger one. Ablation experiments further show that the marginal effects of routing and role-asymmetric constraints reverse with model capability: both components contribute positively on the weaker model, whereas removing either component instead improves accuracy on the stronger model, with the removal of role asymmetry producing the largest improvement. These findings provide reproducible cross-model empirical evidence for systematic reasoning blind spots in self-criticism by homogeneous models. They also indicate that role asymmetry implemented solely at the prompting level is insufficient to reliably overcome cognitive homogeneity and that its effectiveness depends on model capability. This study nevertheless has several limitations, including the restriction of the evaluation benchmark to GSM8K and the smaller evaluation scale used for the Fixed-Round MAD baseline due to its substantially higher per-query cost. Future work could investigate heterogeneous backbone models as critics, train independent process-level evaluation models, and examine the generalizability of these findings across more diverse benchmarks.

DOI: 10.54254/2755-2721/2026.37201

References

[1] Snell, C., Lee, J., Xu, K., & Kumar, A. (2025). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. https://doi.org/10.48550/arXiv.2408.03314

[2] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 24824–24837).

[3] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. arXiv. https://doi.org/10.48550/arXiv.2203.11171

[4] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. arXiv. https://doi.org/10.48550/arXiv.2303.17651

[5] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, 11733–11763.

[6] Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (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. https://doi.org/10.1162/tacl_a_00713

[7] Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. Chapman and Hall/CRC.

[8] Wang, H., Du, X., Yu, W., Chen, Q., Zhu, K., Chu, Z., Yan, L., & Guan, Y. (2025). Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing, 618*, 129063. https://doi.org/10.1016/j.neucom.2024.129063

[9] Zhang, H., Cui, Z., Chen, J., Wang, X., Zhang, Q., Wang, Z., Wu, D., & Hu, S. (2025). Stop overvaluing multi-agent debate—We must rethink evaluation and embrace model heterogeneity. arXiv. https://doi.org/10.48550/arXiv.2502.08788

[10] Zilberstein, S. (1996). Using anytime algorithms in intelligent systems. *AI Magazine, 17*(3), 73–83. https://doi.org/10.1609/aimag.v17i3.1232

[11] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), Article 195. https://doi.org/10.1145/3560815

[12] McCauley, C. (1989). The nature of social influence in groupthink: Compliance and internalization. *Journal of Personality and Social Psychology, 57*(2), 250–260. https://doi.org/10.1037/0022-3514.57.2.250

[13] Dieterich, S., Rumann, S., & Rodemer, M. (2025). Conditions for effective learning from erroneous examples: A systematic review. *Educational Psychology Review, 37*(4), Article 94. https://doi.org/10.1007/s10648-025-10071-x

[14] Le Goues, C., Nguyen, T., Forrest, S., & Weimer, W. (2012). GenProg: A generic method for automatic software repair. *IEEE Transactions on Software Engineering, 38*(1), 54–72. https://doi.org/10.1109/TSE.2011.104

[15] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

[16] Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488–500. https://doi.org/10.2307/1879431

[17] Qwen Team. (2024). Qwen2.5 technical report. arXiv. https://doi.org/10.48550/arXiv.2412.15115

[18] Qwen Team. (2025). Qwen3 technical report. arXiv. https://doi.org/10.48550/arXiv.2505.09388

[19] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv. https://doi.org/10.48550/arXiv.2110.14168

[20] Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*(2), 155–192. https://doi.org/10.1207/s15516709cog0202_4