Transactions on Computational and Scientific Methods | Vo. 4, No. 8, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Policy-Guided Path Selection and Evaluation in Multi-Step Reasoning with Large Language Models

Ray Pan
Independent Researcher, Seattle, USA
raypan.research@gmail.com

Abstract: This paper addresses core challenges in Chain-of-Thought reasoning with large language models, including path instability, structural redundancy, and lack of strategy control. It proposes a reasoning optimization framework that integrates multi-path evaluation and policy-based scheduling. The framework consists of two main components: the Multi-Path Adaptive Evaluation (MPAE) module and the Policy-Aware Reasoning Scheduler (PARS). These components systematically improve Chain-of-Thought performance from two perspectives: structural quality modeling and behavioral decision control. MPAE encodes multiple reasoning paths into vector representations and assigns semantic scores. It constructs a learnable path quality function and uses the scores to guide path aggregation and answer generation. PARS introduces reinforcement learning to build a path selection policy network. It dynamically adjusts scheduling behavior based on reward signals. This improves the stability and consistency of reasoning outputs. Experiments are conducted on the GSM8K benchmark for mathematical reasoning. The evaluation includes multiple metrics such as accuracy, consistency, and robustness. Compared to existing Chain-of-Thought methods, the proposed framework shows clear advantages in structural selection and strategy adaptability. Ablation studies reveal the individual contributions of MPAE and PARS to overall performance. Additional experiments on path distribution and robustness confirm that the framework maintains stable reasoning under high uncertainty. The overall approach features a clear structure, controllable strategy, and adaptive path selection. It effectively enhances Chain-of-Thought reasoning and output quality in complex tasks.

Keywords: Chain-of-Thought reasoning; path selection; policy scheduling; reasoning stability

1. Introduction

Chain-of-thought (CoT) reasoning has emerged as a powerful paradigm to enhance the inference ability of large language models. It has shown great potential in complex tasks, mathematical reasoning, and logical question answering[1,2]. The core idea is to guide the model to generate a series of intermediate reasoning steps before producing a final answer. This mimics how humans solve problems by "writing drafts." It helps reduce the risk of incorrect direct answers. CoT is especially useful for tasks that require multi-step logic, causal analysis, or semantic disambiguation. However, despite its advantages, CoT still suffers from uncertainty in output, strong path dependence, and high sensitivity to prompt design. These limitations affect its stability and reliability in real-world applications[3].

In current practice, the quality of CoT outputs heavily depends on prompt design and randomness in generation. As a result, the same question may yield different reasoning paths and answers in different runs. A widely used solution is Self-Consistency. It generates multiple independent reasoning paths for the same problem and then aggregates or votes on the results to select the most reasonable answer. This approach improves stability to some extent. However, it does not solve two fundamental issues. First, semantic redundancy and logical conflicts among different paths are not well modeled. Second, the selection and scoring of paths rely on static rules rather than dynamic learning, which limits adaptability to task complexity and model behavior[4].

More broadly, CoT reasoning can be seen as a "generate-evaluate-select" process. Each reasoning path represents an "action" by the model. The final answer reflects its "strategy preference." This perspective suggests the applicability of reinforcement learning (RL)[5]. By building an RL framework, the model can receive feedback signals through repeated trials. It can then learn which reasoning paths are more effective and which intermediate steps are more valuable. RL can optimize not only answer accuracy but also semantic coherence and logical consistency. This leads to better reasoning quality and interpretability[6].

The integration of RL with language model reasoning has become a key trend in natural language processing. It has shown promise in dialogue systems, code generation, and decision-making tasks. Introducing RL into CoT can fundamentally change the reliance on static prompt templates. It enables the model to adaptively adjust reasoning structures and develop more generalizable and controllable reasoning behaviors. RL also supports task transfer. It allows the reuse of reasoning strategies across different task types. This extends CoT's applicability to complex cross-domain problems. Compared to prompt engineering or manual path design, RL offers a theoretical and practical solution for automatic optimization[7].

As intelligent systems based on large language models evolve rapidly, achieving trustworthy, stable, and interpretable reasoning has become a major challenge. A CoT framework optimized by self-consistency and reinforcement learning can overcome the uncertainty of single-path reasoning. It introduces dynamic learning mechanisms to improve overall reasoning strategies. This approach is not just a technical enhancement. It represents a systematic upgrade in modeling the cognitive capabilities of language models. It holds strong theoretical value and broad application potential.

2. Related work

2.1 Large Language Model

Large language models have become a core technology in natural language processing in recent years. Their performance gains result from the combination of massive parameter scales and high-quality training corpora[8]. Through autoregressive language modeling, these models acquire not only basic semantic understanding and text generation abilities but also capabilities in cross-task transfer, few-shot learning, and contextual awareness[9,10]. This shift from statistical language models to general-purpose intelligence engines enables strong performance across tasks such as question-answering, dialogue, summarization, translation, and code generation. In open-domain settings, their ability to handle complex tasks without task-specific fine-tuning highlights their strong generalization potential.

However, despite their natural language generation capacity, large language models still face clear limitations in logical reasoning, mathematical calculation, and multi-step planning. These weaknesses stem from the fact that reasoning skills are not explicitly optimized during training. Instead, the models learn patterns in language through data-driven objectives[11,12]. As a result, when faced with tasks that require intermediate reasoning steps, models often default to direct answer generation. This leads to irrelevant answers, logical gaps, or computational errors. To address this, researchers are exploring ways to guide models to produce reasoning chains. This approach aims to make the reasoning process more transparent and stable by showing the path to the answer rather than just the final output[13].

In this context, the generation process of language models is reinterpreted as a sequence of reasoning actions. At each step, the model selects a token based on choices in a latent semantic space. For complex tasks, this often leads to multiple valid reasoning paths. This multiplicity is a strength that enables output diversity.

However, it also introduces variability and uncertainty in the quality of responses. Therefore, it is important to guide the model toward coherent and consistent reasoning paths while preserving output diversity. The Chain-of-Thought paradigm was developed to address this need. It has been effectively applied in the context of large model architectures[14].

As large language models continue to evolve, their reasoning behaviors are shifting from data imitation to explanation generation[15]. This trend requires researchers to redesign generation strategies, prompt structures, and learning objectives. Models must not only produce correct answers but also provide clear, coherent, and reasonable reasoning processes. A key challenge is how to balance controllability and flexibility in reasoning paths. Another is how to maintain fluency while ensuring logical structure. Addressing these challenges is essential for advancing model capabilities. It also reveals the reasoning potential of language models beyond language understanding. This work provides a theoretical and practical foundation for applications in high-stakes domains such as education, law, and healthcare.

2.2 Chain-of-Thought

Chain-of-thought reasoning is a paradigm designed to enhance large language models by guiding them to generate intermediate reasoning steps. Unlike traditional one-step generation strategies, it encourages models to produce a sequence of coherent reasoning steps. This makes their logical deduction, problem decomposition, and knowledge retrieval more aligned with human cognitive patterns. It improves accuracy in tasks such as mathematical problem-solving, logical question answering, and commonsense reasoning. It also provides stronger interpretability. In multi-step reasoning tasks, Chain-of-Thought helps reduce the risk of models producing seemingly correct but flawed answers. It supports the construction of clear problem-solving structures for complex tasks[16].

The effectiveness of Chain-of-Thought heavily depends on the design and structure of prompts. Traditional few-shot prompting can trigger reasoning, but it is highly sensitive to the choice of examples and has limited generalization. To address this, later studies proposed methods such as automatic prompt generation, adaptive path construction, and self-questioning. These methods activate the model's internal reasoning trace and help generate task-relevant reasoning processes. They enrich the implementation of Chain-of-Thought from different perspectives and show its potential as a general-purpose reasoning mechanism in multi-task learning. At the same time, Chain-of-Thought has become a key metric for evaluating reasoning ability[17]. The plausibility, coherence, and stability of the generated paths are now important indicators of model quality.

Despite notable progress, Chain-of-Thought still faces key challenges. First, the reasoning paths generated by the model are often unstable[18]. The same input may produce semantically different steps under different random seeds. This variability affects the reliability of final answers. Second, there is often no internal quality control for the reasoning paths. The generation process can include logical gaps, redundant information, or contradictions. In addition, most current implementations rely on static prompts. They lack dynamic optimization and cannot easily adapt to diverse tasks or changing inputs. Improving the robustness and generalization of the chain of thought remains an important research focus[19].

To address these challenges, one new approach treats Chain-of-Thought as a sequential decision-making task. Reinforcement learning is introduced to help the model learn how to select optimal reasoning paths after generating multiple candidates. This approach maintains the interpretability and diversity of Chain-of-Thought[20]. It also strengthens the model's ability to make strategic decisions in complex reasoning tasks. By combining this with self-consistency, the model can generate multiple paths, perform cross-path voting, and update its path preferences through reinforcement learning. This allows for a better balance between output quality and reasoning stability. Such a direction brings dynamic learning into the chain of thought and lays the foundation for building more advanced language intelligence systems.

3. Method

This study presents an enhanced Chain-of-Thought reasoning framework that integrates self-consistency principles with reinforcement learning-based strategies. The primary objective is to improve both the stability of reasoning paths and the overall generation quality in complex tasks involving large language models. Traditional Chain-of-Thought approaches often rely on fixed prompts and lack the flexibility to adapt to diverse input scenarios or task-specific reasoning demands. To address these limitations, the proposed framework introduces two complementary modules that enable dynamic reasoning control and intelligent path selection. This design shifts the reasoning process from static generation to a more interactive and strategically guided paradigm, which is essential for tasks that require multi-step logic and semantic coherence.

At the core of the framework are the Multi-Path Adaptive Evaluation (MPAE) and the Policy-Aware Reasoning Scheduler (PARS). MPAE evaluates multiple candidate reasoning paths by assigning scores based on semantic consistency and logical validity. This scoring mechanism supports the filtering and aggregation of high-quality paths, ensuring that the most plausible and structurally sound reasoning chains are selected. PARS complements this by introducing a path selection policy network trained through reinforcement learning. This network learns to prioritize efficient and stable reasoning trajectories in future generations. Together, these two modules form a cohesive system that enhances the model ’ s ability to adapt to variable reasoning conditions. The framework also enables the model to retain the memory of past reasoning patterns and apply them to new tasks, making the overall process more robust and cognitively aligned. The complete architecture of the system is illustrated in Figure 1.

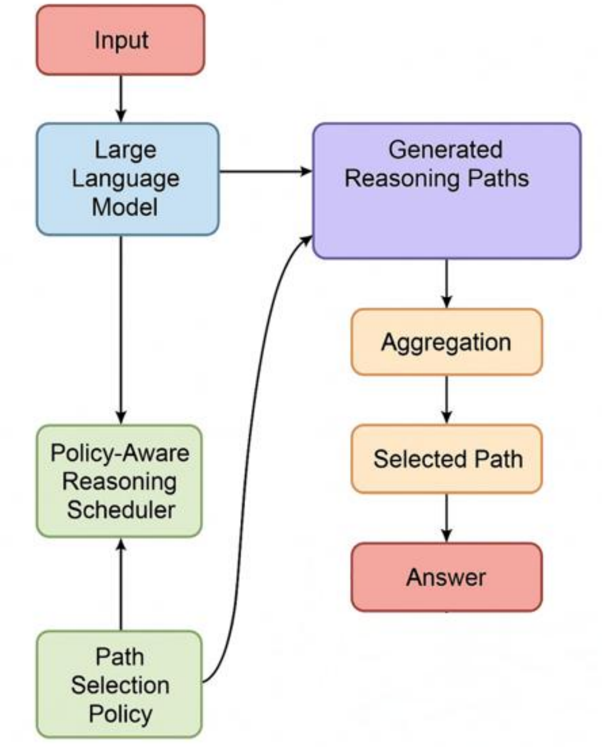

Figure 1. Overall model architecture diagram

3.1 Multi-Path Adaptive Evaluation

In Chain-of-Thought reasoning, large language models often produce multiple candidate reasoning paths in response to a single input. These paths may vary widely in semantic coherence, logical soundness, and consistency with the correct answer. Such variability presents both opportunities and challenges. On one hand, the availability of diverse paths enriches the solution space and allows for flexible exploration. On the other hand, it increases the difficulty of identifying which paths genuinely contribute to valid reasoning. Some paths may contain partially correct steps, while others may include irrelevant or logically inconsistent information. This highlights the need for a robust mechanism that can discriminate between high-quality and low-quality reasoning processes.

To address this, the Multi-Path Adaptive Evaluation (MPAE) mechanism is proposed. This module is designed to perform fine-grained assessment of multiple reasoning paths by assigning each one a score that reflects its overall plausibility and relevance to the problem. These scores are then used to rank the paths, filter out unreliable ones, and fuse the selected paths into a final answer. Unlike traditional approaches that rely on static rules or manually crafted heuristics, MPAE leverages learnable path representations and dynamic scoring strategies. This design enables the system to adapt to varying reasoning contexts and path structures, making it more scalable and generalizable. The internal structure of MPAE, including how it encodes, evaluates, and integrates paths, is illustrated in Figure 2.

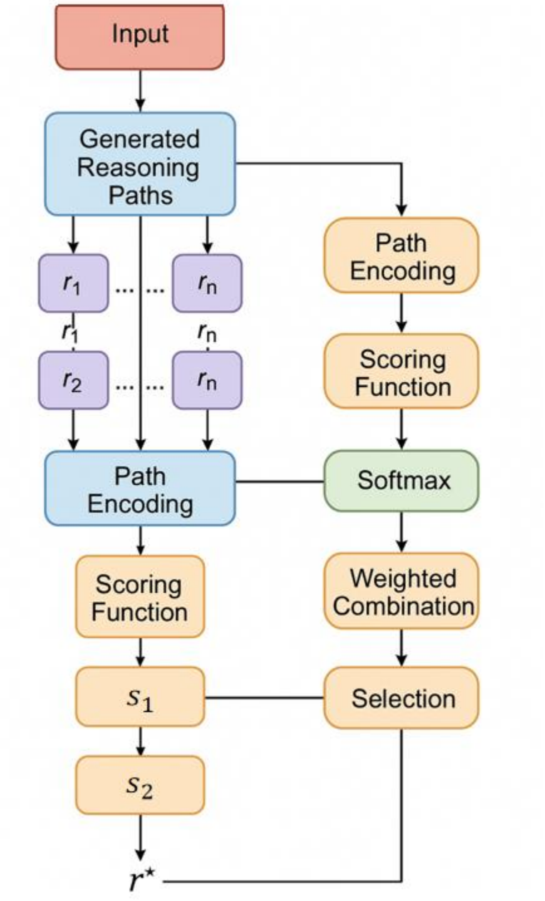

Figure 2. MPAE module architecture

Assume that the model generates N reasoning paths for input x, denoted as $\{r_1, r_2, ..., r_N\}$, and each path r_i is mapped to a vector representation $h_i \in R^d$. We define the path-scoring function as:

$$s_i = f_{score}(h_i)$$

There $f_{score}(\cdot)$ is a differentiable nonlinear mapping network, which is used to comprehensively consider the language quality, logical coherence, and contextual consistency of the path.

Next, all path scores are normalized to obtain the selection probability of each path:

$$p_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)}$$

Here, a softmax distribution is used to represent the relative importance of each path in the reasoning output, ensuring that the scores are comparable and distinguishable. The final output reasoning path r^* can be expressed as a weighted combination, or the path with the maximum score can be directly selected:

$$r^* = \arg\max_i s_i$$

In addition, the MPAE mechanism also allows sparse attention fusion between paths, that is, combining the local contents of multiple high-scoring paths at the fragment level, expressed as:

$$r^* = \sum_{i=1}^{N} \alpha_i \tau_i, where \quad \alpha_i = 1[s_i \geq \tau]$$

Where τ is the threshold and $1[\cdot]$ represents the indicator function, which is used to screen high-quality paths for final answer generation.

Through the above mechanism, MPAE not only improves the flexibility of reasoning path selection but also provides a stable and learnable input signal for subsequent policy schedulers (such as reinforcement learning modules). It avoids the risk of overfitting a single path and provides a structural guarantee for the overall improvement of reasoning quality. As an important part of the framework proposed in this study, this mechanism introduces a new paradigm of structured, multi-path fusion and quality assessment for chain thinking reasoning.

3.2 Policy-Aware Reasoning Scheduler

In multi-path reasoning tasks, the complexity and variability of input scenarios often exceed the capacity of static scoring functions. These functions, while useful for basic ranking, lack the flexibility to adapt to dynamic reasoning environments where different inputs may require fundamentally different path exploration strategies. Static mechanisms treat each reasoning path in isolation and make selection decisions based on fixed criteria, which may not generalize well across tasks or input types. To address this critical limitation, this study introduces the Policy-Aware Reasoning Scheduler (PARS), a module specifically designed to bring strategy-aware decision-making into the reasoning process.

PARS leverages reinforcement learning to train a path selection policy that dynamically guides the model in generating and prioritizing reasoning paths. This policy is designed to learn and model complex strategic behaviors, such as deciding the optimal timing for path sampling, identifying which candidate paths should be explored further, and determining how attention should be distributed among available options. Unlike traditional Chain-of-Thought reasoning pipelines that treat generation and evaluation as separate phases, PARS integrates them into a unified scheduling framework. This integration allows the model to build a deeper understanding of path distribution patterns and develop internal strategies that promote more stable and consistent reasoning outputs. The architectural design of PARS, including its interaction with other modules and its policy learning mechanism, is presented in Figure 3.

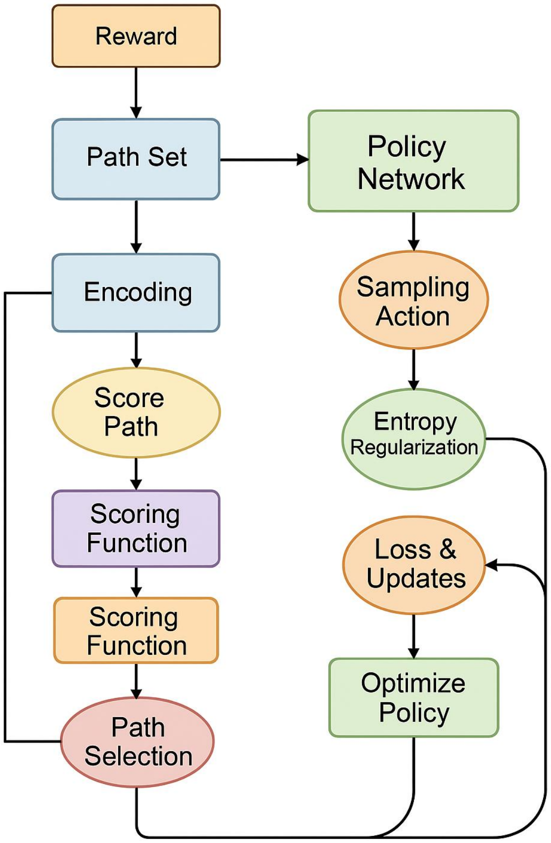

Figure 3. PARS Module Architecture

Suppose the input of the model at time step t is x_t, and the corresponding path set is $R_t = \{r_{t,1}, \ldots, t_{t,N}\}$. We define the policy network as $\pi_\theta(a_t \mid s_t)$, where s_t is the state vector, encoding the distribution information of the input x_t and the generated path, and a_t is the selected reasoning action (such as path index or path combination strategy). The policy network selects the reasoning path by sampling or greedy method:

$$a_t \sim \pi_\theta(a_t \mid s_t)$$

After the path is selected, the system obtains a reward signal r_t, which is used to measure the logical consistency brought by the choice and the rationality of the final answer. We optimize the strategy parameters based on the expected cumulative reward:

$$J(\theta) = E_{\pi_\theta}[\sum_{t=1}^{T} \gamma^t r_t]$$

Where $\gamma \in (0,1]$ is the discount factor. To improve the stability of the strategy, the policy gradient method is used to optimize θ:

$$\nabla_\theta J(\theta) = E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \widetilde{A}_t]$$

Where $\widetilde{A}_t$ is the advantage function, which can be constructed by the difference between the current path score and the baseline score.

In addition, to further constrain the semantic diversity and strategy exploration ability of path selection, we introduce an entropy regularization term in the loss function to enhance the strategy exploration ability:

$$L = -J(\theta) + \lambda \cdot H(\pi_\theta)$$

Where $H(\pi_\theta)$ represents the entropy term of the strategy, and λ is the weight hyperparameter, which is used to control the smoothness and diversity of the strategy distribution.

By introducing the PARS module, the model can achieve adaptive path scheduling and strategy learning when facing multi-path reasoning tasks, thereby actively mining potential high-quality reasoning solutions from the path space. This strategy-aware mechanism not only improves the selectivity and flexibility of the reasoning process but also lays a higher-quality structural foundation for subsequent path fusion and answer generation. PARS plays a strategic decision-making role in the entire reasoning architecture, promoting the evolution of chain thinking reasoning from static behavior to dynamic scheduling.

4. Experimental Results

4.1 Dataset

The main dataset used in this study is GSM8K. It is a high-quality natural language reasoning dataset focused on elementary school mathematics problems. GSM8K is widely used to evaluate the Chain-of-Thought abilities of large language models. It contains approximately 8,500 math problems, each presented in natural language. Every problem includes a detailed reasoning process and a final answer. The questions cover basic arithmetic, unit conversion, multiplicative relations, and other common logical structures.

The dataset is designed to test a model's ability to understand problem context, build multi-step reasoning processes, and generate structured explanations. Each question requires the model not only to produce the correct answer but also to present a reasonable sequence of intermediate steps. Therefore, GSM8K aligns naturally with Chain-of-Thought reasoning and is especially suitable for evaluating multi-path generation and path selection.

In addition, GSM8K has a clear format and standardized annotations. It supports various reasoning paradigms, including manual prompt construction, automatic path sampling, and self-consistency mechanisms. Its moderate difficulty and broad coverage make GSM8K a standard benchmark for studies on Chain-of-Thought optimization methods such as MPAE and PARS. It provides a stable and reliable foundation for evaluation in this research.

4.2 Experimental Setup

In this study, we build a Chain-of-Thought optimization framework based on the pre-trained language model ChatGLM. The goal is to explore its behavior and performance in multi-path reasoning tasks. ChatGLM has strong capabilities in Chinese language understanding and generation. It supports long-text input and complex logical structure modeling. This makes it a suitable foundation for our framework. To enable path scoring, policy learning, and dynamic reasoning scheduling, we integrate the MPAE and PARS modules into ChatGLM. These modules allow the model to perform effective path selection and strategy optimization across multiple reasoning trajectories.

All experiments are conducted on a single high-performance GPU. We use FP16 precision for both inference and fine-tuning. Path generation is configured with a fixed temperature and maximum length. The policy network and path scoring function are implemented using standard Transformer submodules. We use the Adam optimizer for model updates. All experiments are performed on the GSM8K dataset. We follow consistent data splits and preprocessing procedures. Table 1 lists the key parameters used in the experimental setup.

Table 1: Detailed Experimental Setup

Parameters	**Setting Value**
Basic Model	ChatGLM
Dataset	GSM8K
Path Generation Temperature	0.7
Maximum number of paths	8
Encoder hidden dimension	768
Optimizer	Adam
Learning Rate	3e-5
Batch size	16
Epochs	200
Inference accuracy	FP16

4.3 Experimental Results

1) Comparative experimental results

This paper first gives the comparative experimental results, as shown in Table 2.

Table 2: Comparative experimental results

Method	**ACC**	**CoT Consistency**	**Path Robustness**
Standard CoT[21]	74.6	68.2	65.5
Self-Consistency[22]	80.2	75.4	72.6
ReAct[23]	82.7	78.1	75.9
Ours	85.9	83.2	81.5

As shown in the results in Table 2, the proposed method outperforms mainstream Chain-of-Thought frameworks across all evaluation metrics. This confirms the effectiveness of introducing Multi-Path Adaptive Evaluation (MPAE) and the Policy-Aware Reasoning Scheduler (PARS). In particular, the method achieves an accuracy (ACC) of 85.9 percent. This is more than 11 percentage points higher than the standard CoT method. The result shows that more refined path scoring and selection strategies can guide the model to produce more reliable answers.

The proposed method also shows clear advantages in CoT Consistency. This metric measures the logical alignment among different reasoning paths and reflects the stability of the reasoning process. Traditional CoT methods often suffer from semantic drift due to path variability. MPAE reduces this inconsistency by modeling path quality and applying normalized scoring. In addition, the policy network favors sampling of high-quality paths. This further improves structural alignment across paths and leads to greater consistency.

Path Robustness measures performance stability under multiple path sampling conditions. The results show that the proposed method reaches 81.5 percent in this metric. This is significantly better than the baseline methods. The result indicates that PARS not only selects the current best path but also learns to optimize path strategies over time. With reinforcement learning, the model develops a dynamic scheduling policy that supports stable decisions under uncertain path distributions. This improves the model ’ s adaptability to diverse inputs and task variations.

Overall, this study uses MPAE for fine-grained path-level quality modeling and introduces PARS to optimize reasoning strategies. These two components work together to improve path generation, selection, and execution stability systematically. The results clearly show that traditional static generation and voting approaches face limitations in complex reasoning tasks. In contrast, the proposed framework provides a new solution that enhances controllability and interpretability for large language models.

2) *Ablation Experiment Results*

This paper further gives the results of ablation experiments, and the experimental results are shown in Table 3.

Table 3: Ablation Experiment Results

Method	ACC	CoT Consistency	Path Robustness
Baseline	76.4	70.1	67.3
+MPAE	81.0	78.6	75.2
+PARS	79.3	74.5	73.1
Ours	85.9	83.2	81.5

As shown in the ablation results in Table 3, the two core modules proposed in this study, Multi-Path Adaptive Evaluation (MPAE) and the Policy-Aware Reasoning Scheduler (PARS), play a key role in improving final performance. The baseline model, without any path optimization mechanism, performs poorly across all metrics. This indicates that standard Chain-of-Thought prompting alone cannot effectively handle path uncertainty and reasoning consistency. It also highlights the importance of path selection and reasoning scheduling in complex tasks.

After introducing the MPAE module, the model shows significant improvements in both accuracy and consistency. In particular, CoT Consistency increases from 70.1 percent to 78.6 percent. This result shows that scoring and normalized selection of reasoning paths help the model focus on logically coherent and semantically valid sequences. It also reduces interference from redundant or incorrect reasoning branches. At the same time, Path Robustness improves, demonstrating MPAE’ s ability to model stability in complex path spaces. This provides structured quality control during generation.

When only the PARS module is added, the model also achieves clear performance gains. Path Robustness increases from 67.3 percent to 73.1 percent. This shows that even without explicit path scoring, the policy scheduler can learn improved path selection strategies through reinforcement learning. It avoids random sampling and reduces the influence of suboptimal paths. PARS provides an adaptive mechanism that helps the model dynamically adjust its strategy during long reasoning processes. This strengthens behavioral consistency and generation stability.

With both MPAE and PARS integrated, the model achieves the best results across all three metrics. This demonstrates the complementary and synergistic effects of the two modules. MPAE offers fine-grained quality scoring at the path level. PARS implements decision-level strategy scheduling. Together, they form a reasoning framework that is structured, controllable, and stable in output. This design improves reasoning performance and reflects a shift in Chain-of-Thought optimization from pure generation toward a combination of strategy and structure.

3) Robustness testing under high noise path injection

This paper also presents a robustness test under high noise path injection, and the experimental results are shown in Figure 4.

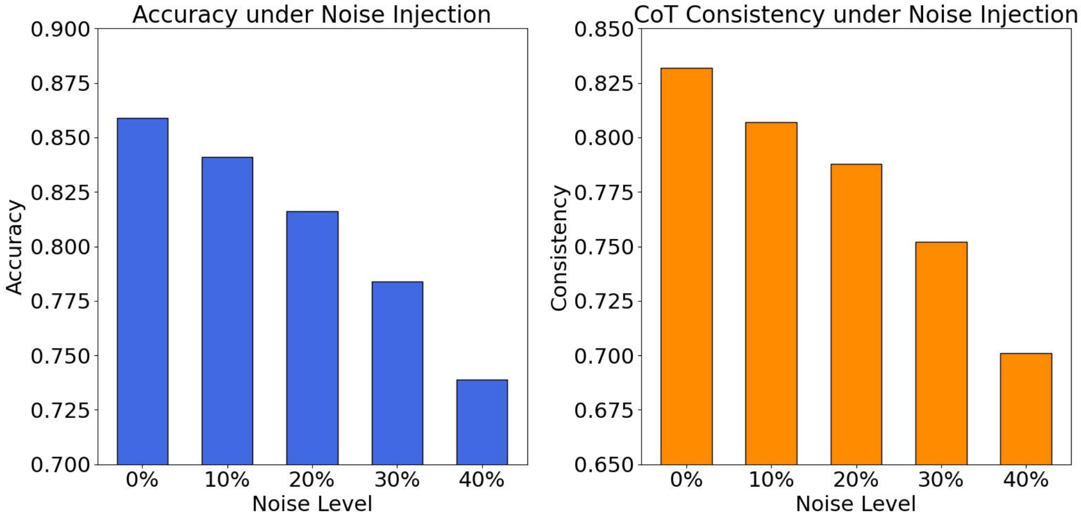

Figure 4. Robustness testing under high noise path injection

Figure 4 shows the robustness performance of the proposed reasoning framework under different levels of noise path injection. The results are presented in terms of accuracy and Chain-of-Thought consistency. As the proportion of injected noise increases, the overall model performance decreases. However, the drop is steady and the curves remain smooth. This indicates strong resistance to interference. Even when 40 percent of the paths are corrupted with high noise, the model maintains an accuracy above 0.74. This shows that the system does not collapse under severe path contamination.

This result highlights the important role of the MPAE module in path quality evaluation. By modeling multiple paths at the vector level and applying score normalization, the model can still identify high-quality reasoning chains when the path set is affected by noise. This mechanism prevents noisy paths from dominating the output. It provides structural support for reasoning stability in open-ended path spaces.

In addition, the CoT Consistency curve declines slightly faster than the accuracy curve. This suggests that while the final answers are somewhat tolerant to noise, the intermediate reasoning structures are more vulnerable. The result further confirms that language models alone cannot build robust reasoning chains. Strategy-level intervention is necessary. The PARS module learns a path scheduling policy. It increases the model's ability to actively select paths and supports consistency in reasoning structure.

In summary, this experiment demonstrates the robustness of the proposed model under extreme conditions. It also emphasizes the importance of combining multi-path modeling with strategic scheduling. In real-world applications where uncertainty or misleading paths may occur, this mechanism ensures structural clarity and behavioral consistency in the output of large language models. It shows strong potential for practical deployment.

4) *Analysis of the impact of different inference temperature settings on path distribution*

This paper also gives an analysis of the impact of different inference temperature settings on path distribution, and the experimental results are shown in Figure 5.

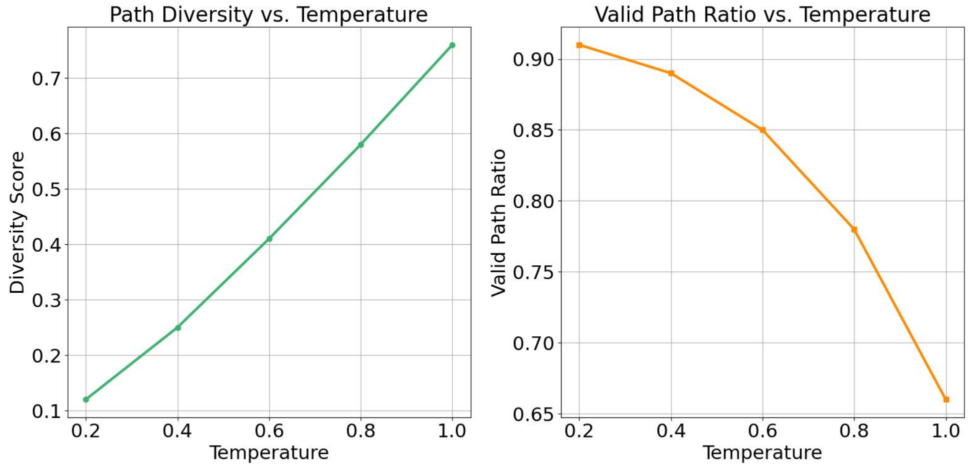

Figure 5. Analysis of the impact of different inference temperature settings on path distribution

Figure 5 illustrates the impact of different reasoning temperature settings on the structure of path distributions. The analysis is based on two dimensions: Path Diversity and Valid Path Ratio. The results show that temperature is a critical hyperparameter for controlling the characteristics of generated paths. It not only shapes the structure of the path space but also affects the stability and robustness of Chain-of-Thought reasoning. Higher temperatures lead the model to generate more diverse paths during sampling. This increases the expressive range but also introduces potential noise and uncertainty.

In the left panel, we observe a continuous rise in path diversity as the temperature increases from 0.2 to 1.0. The growth follows a clear linear trend. This indicates that higher temperatures encourage the model to break from fixed patterns and produce more structurally varied reasoning paths. Such diversity is beneficial for exploring alternative reasoning logic. For frameworks relying on multi-path evaluation and policy scheduling, this forms a necessary foundation for enhanced exploration and candidate space construction.

However, the right panel shows that the Valid Path Ratio decreases as temperature increases. The decline becomes sharper after the temperature exceeds 0.6. This suggests that while diversity improves, the proportion of logically consistent and task-relevant paths decreases. As a result, the useful information density in the path space drops. This presents challenges for Chain-of-Thought-based systems. Under high temperatures, MPAE must have stronger filtering capabilities, and PARS must manage increased uncertainty during decision-making.

Therefore, setting a proper temperature is key to building high-quality reasoning processes. Low temperatures yield concentrated paths but may lack coverage. High temperatures produce richer paths but include more noise. The proposed MPAE and PARS modules are designed to address this trade-off. They act from the perspectives of path selection and policy learning. Together, they ensure a dynamic balance between path diversity and effectiveness in the final reasoning output.

5) *Path fusion strategy (maximum/weighted/average) comparison experiment*

Finally, this paper presents a comparative experiment of path fusion strategies (maximum/weighted/average), and the experimental results are shown in Figure 6.

Figure 6 presents the comparative results of three path aggregation strategies—maximum, weighted, and average—evaluated across accuracy, reasoning consistency, and robustness. The experiment shows that the choice of fusion method has a significant impact on final reasoning quality. This is especially true in multi-path generation settings, where behavioral differences among strategies directly affect output stability and interpretability.

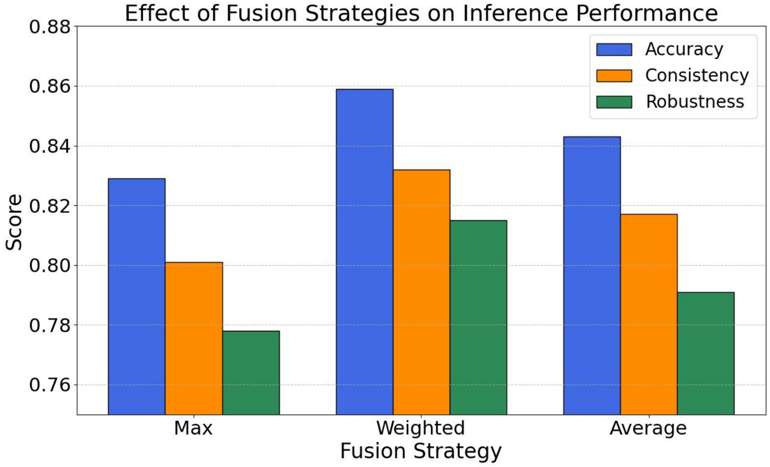

Figure 6. Path fusion strategy (maximum/weighted/average) comparison experiment

The figure shows that the weighted fusion strategy performs best on all three metrics. It achieves an accuracy of 0.859, a consistency score of 0.832, and a robustness of 0.815. These results indicate that the weighting mechanism provided by the MPAE module helps suppress suboptimal paths and highlight high-quality reasoning chains. This enables more effective path integration without sacrificing diversity. In contrast, the maximum strategy, though simple and direct, tends to rely too heavily on a single path. If that path contains noise or errors, it may negatively affect the final output.

The average fusion strategy shows intermediate performance. While it theoretically balances all paths, it fails to consider differences in path quality. As a result, low-quality paths may be incorporated into the final decision, which weakens the contribution of stronger reasoning chains. Although this method shows a certain level of stability, it falls short in high-performance settings.

This experiment further confirms the importance of path scoring and weighted selection in the MPAE design. By modeling path distributions and enabling structured fusion, the weighted strategy improves accuracy, coherence, and robustness simultaneously. It does so without introducing extra computation during generation. This dual-level optimization of structural quality and strategic expression makes the proposed reasoning framework more adaptive and better suited for practical use.

5. Conclusion

This paper addresses the stability and controllability challenges of large language models in multi-path Chain-of-Thought reasoning. It proposes an optimization framework that integrates structural evaluation and strategic scheduling. The framework consists of two core modules. The Multi-Path Adaptive Evaluation (MPAE) module identifies and focuses on high-quality reasoning paths. The Policy-Aware Reasoning Scheduler (PARS) uses reinforcement learning to optimize path selection strategies. These two components work together to introduce a new paradigm of dynamic structural optimization and strategy-driven reasoning while preserving the model's original language generation capability. Experimental results show significant improvements in reasoning accuracy, consistency, and robustness, confirming the effectiveness of structured reasoning mechanisms for complex task modeling. By introducing MPAE, this study effectively mitigates the issues of path redundancy and quality instability found in traditional Chain-of-Thought methods. The path-scoring mechanism improves the model’s ability to identify promising paths. It also builds a learnable space for path quality representation, providing a basis for structural optimization during reasoning. The PARS module further drives the reasoning process from passive sampling to active scheduling. It enables the model to explore and optimize paths adaptively. This architectural shift from language generation to strategic reasoning enhances decision stability and interpretability in real-world applications.

The proposed framework contributes not only at the methodological level but also offers practical value. In domains such as education, healthcare, and financial decision-making, where transparency and stability of reasoning paths are critical, the framework improves reliable reasoning performance and reduces uncertainty in outputs. For tasks requiring multi-step planning and high fault tolerance, such as automated process control and intelligent question answering the proposed structure shows good transferability and broad applicability. It supports the transition of large language models from text-generation tools to cognitive systems with reasoning capabilities. Looking ahead, this method offers several directions for future work. One direction is to explore finer-grained path structure modeling, such as using causal graphs or multimodal information, to improve path discrimination in high-dimensional task spaces. Another is to enhance the scheduling module by incorporating external environment signals, making it adaptive across tasks and domains. In addition, integrating this framework with retrieval-augmented mechanisms or symbolic reasoning systems could offer further opportunities. These advances may drive continuous progress toward general-purpose reasoning in language models.

References

[1] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

[2] Lyu Q, Havaldar S, Stein A, et al. Faithful chain-of-thought reasoning[C]//The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023). 2023.

[3] Feng G, Zhang B, Gu Y, et al. Towards revealing the mystery behind chain of thought: a theoretical perspective[J]. Advances in Neural Information Processing Systems, 2023, 36: 70757-70798.

[4] Zhang X, Du C, Pang T, et al. Chain of preference optimization: Improving chain-of-thought reasoning in llms[J]. Advances in Neural Information Processing Systems, 2024, 37: 333-356.

[5] Xia Y, Wang R, Liu X, et al. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms[J]. arXiv preprint arXiv:2404.15676, 2024.

[6] Miao J, Thongprayoon C, Suppadungsuk S, et al. Chain of thought utilization in large language models and application in nephrology[J]. Medicina, 2024, 60(1): 148.

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

[8] Madaan A, Yazdanbakhsh A. Text and patterns: For effective chain of thought, it takes two to tango[J]. arXiv preprint arXiv:2209.07686, 2022.

[9] Yu, Z., He, L., Wu, Z., Dai, X., & Chen, J. (2023). Towards better chain-of-thought prompting strategies: A survey. arXiv preprint arXiv:2310.04959.

[10] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023, January). React: Synergizing reasoning and acting in language models. In International Conference on Learning Representations (ICLR).

[11] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36, 8634-8652.

[12] Huang, X., Zhang, L. L., Cheng, K. T., Yang, F., & Yang, M. (2023). Fewer is more: Boosting LLM reasoning with reinforced context pruning. arXiv preprint arXiv:2312.08901.

[13] Diao S, Wang P, Lin Y, et al. Active prompting with chain-of-thought for large language models[J]. arXiv preprint arXiv:2302.12246, 2023.

[14] Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., ... & Liu, T. (2023). Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. arXiv preprint arXiv:2309.15402.

[15]Yao Y, Li Z, Zhao H. Beyond chain-of-thought, effective graph-of-thought reasoning in language models[J]. arXiv preprint arXiv:2305.16582, 2023.

[16]Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

[17]Ranaldi L, Freitas A. Aligning large and small language models via chain-of-thought reasoning[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 1812-1827.

[18]Suzgun M, Scales N, Schärli N, et al. Challenging big-bench tasks and whether chain-of-thought can solve them[J]. arXiv preprint arXiv:2210.09261, 2022.

[19]Cheng X, Li J, Zhao W X, et al. Chainlm: Empowering large language models with improved chain-of-thought prompting[J]. arXiv preprint arXiv:2403.14312, 2024.

[20]Zheng G, Yang B, Tang J, et al. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 5168-5191.

[21]Zhou Z, Tao R, Zhu J, et al. Can Language Models Perform Robust Reasoning in Chain-of-thought Prompting with Noisy Rationales?[J]. Advances in Neural Information Processing Systems, 2024, 37: 123846-123910.

[22]Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models[J]. arXiv preprint arXiv:2203.11171, 2022.

[23]Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models[C]//International Conference on Learning Representations (ICLR). 2023.