

Language Models Coupled with Metacognition Can Outperform Reasoning Models

Vedant Khandelwal^{1,2}, Francesca Rossi¹, Keerthiram Murugesan¹, Erik Miehling¹, Murray Campbell¹, Karthikeyan Natesan Ramamurthy¹, Lior Horesh¹

¹IBM Research, NY, USA

²University of South Carolina, SC, USA

vedant@email.sc.edu, francesca.rossi2@ibm.com, Keerthiram.Murugesan@ibm.com, Erik.Miehling@ibm.com, mcam@us.ibm.com, knatesa@us.ibm.com, lhoresh@us.ibm.com

Abstract

Large language models (LLMs) excel in speed and adaptability across various reasoning tasks, but they often struggle when strict logic or constraint enforcement is required. In contrast, Large Reasoning Models (LRMs) are specifically designed for complex, step-by-step reasoning, although they come with significant computational costs and slower inference times. To address these trade-offs, we employ and generalize the SOFAI (Slow and Fast AI) cognitive architecture into SOFAI-LM, which coordinates a fast LLM with a slower but more powerful LRM through metacognition. The metacognitive module actively monitors the LLM’s performance and provides targeted, iterative feedback with relevant examples. This enables the LLM to progressively refine its solutions without requiring the need for additional model fine-tuning. Extensive experiments on graph coloring and code debugging problems demonstrate that our feedback-driven approach significantly enhances the problem-solving capabilities of the LLM. In many instances, it achieves performance levels that match or even exceed those of standalone LRMs while requiring considerably less time. Additionally, when the LLM and feedback mechanism alone are insufficient, we engage the LRM by providing appropriate information collected during the LLM’s feedback loop, tailored to the specific characteristics of the problem domain and leads to improved overall performance. Evaluations on two contrasting domains—graph coloring, requiring globally consistent solutions, and code debugging, demanding localized fixes—demonstrate that SOFAI-LM enables LLMs to match or outperform standalone LRMs in accuracy while maintaining significantly lower inference time.

Introduction

Reasoning tasks, from classical constraint satisfaction problems to complex debugging scenarios in software development, continue to pose significant challenges for artificial intelligence (AI) systems. Large language models (LLMs) have shown remarkable flexibility and speed, quickly generalizing across a wide array of reasoning domains (Brown et al. 2020; Kojima et al. 2022). However, LLMs often struggle with tasks requiring strict logical consistency and hard constraint satisfaction (Jiang et al. 2023; Valmeekam

et al. 2022). In contrast, large reasoning models (LRMs) offer robust step-by-step deliberative reasoning, but their effectiveness comes with substantially higher computational demands and slower inference times (Kambhampati 2024; Stechly, Valmeekam, and Kambhampati 2024).

Reconciling the trade-off between speed and reliability remains an open research problem. To tackle this issue, we draw on previous work related to architectures inspired by the dual-process cognitive theory of human decision-making (Booch et al. 2021; Fabiano et al. 2023; Ganapini et al. 2022; Lin et al. 2024), specifically the SOFAI architecture (Fabiano et al. 2025). SOFAI employs two types of problem solvers: “fast” solvers, which are generally quicker but less reliable, and “slow” solvers, which are more reliable but typically require more resources and time. Additionally, SOFAI incorporates a metacognitive component that selects the most appropriate solver for each problem instance, based on the risk profile and the available resources. This approach has been shown to achieve better overall performance than any single solver alone, particularly when the fast solver is an LLM-based planner and the slow solver is a classical symbolic planner (Fabiano et al. 2025).

In this study, we define a generalized version of the SOFAI architecture, which we call SOFAI-LM, by introducing a training-free metacognitive feedback loop for LLMs. Specifically, we implement a feedback mechanism connecting the fast solver — an LLM in our case — with the metacognitive component. This empowers the LLM to self-correct and improve its outputs iteratively, without any additional fine-tuning. If the LLM fails to converge within a fixed number of iterations, the metacognitive component invokes the slower, deliberative solver — an LRM in this case — along with the accumulated feedback and a history of past solutions for a problem.

The paper makes the following key contributions:

- We introduce a generalized version of the SOFAI architecture, called SOFAI-LM. This incorporates a training-free metacognitive governance module that features iterative S1 feedback and selective fallback.
- Through extensive experiments in graph coloring and code debugging, we demonstrate that an LLM guided by the SOFAI-LM metacognitive feedback can match or even outperform a standalone LRM in both domains.

- Additionally, we investigate the impact of the information collected in the LLM + Metacognition loop when calling the LRM. We characterize the domains where sharing this information enhances the performance of the LRM.

Overall, our findings reveal that the architecture invokes the slower, compute-intensive LRM only when necessary, making use of the information accumulated during the LLM’s feedback loop. This approach is tailored to domain-specific characteristics, leading to significant improvements in the LRM’s computational efficiency and inference time. To achieve these contributions, we address the following research questions:

- RQ1: Can a feedback-driven LLM outperform an LRM?
- RQ2: How do the type of feedback and the format of episodic memory affect a feedback-driven LLM?
- RQ3: Can the information gathered by SOFAI-LM, when used iteratively with an LLM, enhance the performance of an LRM?
- RQ4: Does SOFAI-LM perform better than its LRM counterpart?

Our experimental results from the two problem domains indicate a promising directions to some of these research questions. Additionally, they demonstrate that increased feedback improves outcomes when iterating with an LLM.

Background

The SOFAI architecture

Kahneman’s influential work Thinking, Fast and Slow popularized the distinction between System 1 (fast, intuitive thinking) and System 2 (slow, analytical thinking), providing a conceptual lens that has shaped both cognitive psychology (Kahneman 2011). In the realm of AI, (Booch et al. 2021) introduced the SOFAI (Slow and Fast AI) architecture, which combines a fast solver and a slow solver under the governance of a metacognition.

The SOFAI architecture draws inspiration from dual-process theories of human cognition, which differentiate between fast, intuitive reasoning (System 1) and slow, deliberate reasoning (System 2) (Fabiano et al. 2025). In this framework, System 1 solvers are fast and intuitive, like LLMs, that rapidly generate candidate solutions using learned heuristics. In contrast, System 2 solvers like LRM rely on slow, deliberate reasoning to ensure correctness and reliability. A key advantage in SOFAI is its metacognitive module, which monitors the performance of the solvers and dynamically selects the most suitable solver for a given problem. This selection process balances speed and accuracy, as metacognitive module considers solution quality, available resources, and past solver performance to adjust future decisions, thus promoting adaptability, reliability, and efficient resource use.

The original SOFAI architecture successfully demonstrated this hybrid strategy in applications such as automated planning and pathfinding (Booch et al. 2021; Fabiano et al. 2023; Ganapini et al. 2022; Lin et al. 2024), showing that alternating between neural intuition and symbolic rigor can

yield efficiency gains without sacrificing correctness. However, previous SOFAI-like approaches typically treated S1 and S2 as independent solvers, simply choosing between them on a per-instance basis rather than enabling iterative collaboration.

Problem domains

Reasoning over structured and unstructured domains remains a cornerstone of AI, underpinning both theoretical advancements and practical applications. To rigorously assess the capabilities of reasoning systems, we focus on two representative and challenging problem domains: graph coloring decision problem and automated program debugging.

Graph Coloring Decision Problem. The graph coloring decision problem is a canonical constraint satisfaction problem (CSP) with broad relevance in scheduling, allocation, and resource management. Given an undirected, unweighted graph $G = (V, E)$ and a positive integer k , the objective is to determine whether there exists an assignment of at most k colors to the vertices such that no two adjacent nodes share the same color:

- **Input:** An undirected, unweighted graph $G = (V, E)$ and a positive integer k .
- **Goal:** Decide if there exists a function $f : V \rightarrow \{1, 2, \dots, k\}$ such that for every edge $(u, v) \in E$, $f(u) \neq f(v)$.

Unlike the optimization variant of the problem, which aims to find minimum number of colors, the decision version focuses on satisfiability for a fixed k , directly reflecting global consistency requirements of many real-world problems.

We utilize the DIMACS representation (Johnson and Trick 1996) to express graph coloring instances, which provides a standardized format for benchmarking and visualization. Notably, verifying a candidate solution in the graph coloring decision problem is computationally easy. However, generating a correct solution or rectifying an incorrect one requires reasoning over the entire graph’s structure. *Global* consistency is crucial, as each coloring choice may influence the overall feasibility of the solution.

```

1   c Example of a graph coloring problem in DIMACS
2   p edge 5 5
3   c edges
4   e A B
5   e A C
6   e B C
7   e C D
8   e D E

```

Code Debugging. Code debugging is a fundamental yet time-consuming task in software engineering, accounting for up to half of the overall effort involved in development (McConnell 2004). While the advent of LLMs has led to significant advancements in code generation, automated debugging—the ability to identify and repair faults

in code—has remained an underexplored challenge until recently (Tian et al. 2024). In this paper, we focus on the DebugBench, a large-scale benchmark designed to systematically evaluate the debugging capabilities of LLMs (Tian et al. 2024). It comprises 4,253 instances that cover Python, Java, and C++. It includes a diverse taxonomy of bug types and programming scenarios. Each example in DebugBench is generated by deliberately introducing a bug into an otherwise correct code snippet. This ensures that only a *local* region of the code is erroneous while the remainder of the code remains correct. This benchmark design necessitates precisely identifying and repairing erroneous segments without introducing new errors, distinguishing it from problems like graph coloring, where every color assignment must be checked against the global structure of the problem.

Domain Relevance and Motivation. We believe that the two problem domains of graph coloring and code debugging encompass a range of reasoning challenges that AI systems often encounter. Graph coloring requires holistic and globally consistent reasoning over combinatorial structures, while code debugging, as defined in DebugBench, assesses a model’s ability to make precise and targeted corrections based on the local context. By evaluating both of these domains, we offer a comprehensive assessment of the ability to tackle diverse, high-impact, and practically relevant tasks.

The SOFAI-LM Architecture

Our architecture builds upon the SOFAI architecture by incorporating a feedback loop between the System 1 solver and the metacognition component. In this paper, we focus on a specific instance of this architecture, in which the System 1 (S_1) solver is a large language model (LLM) and the System 2 (S_2) solver is a logical reasoning model (LRM). We refer to this instance as SOFAI-LM. Figure 1 provides an overview of this architecture.

System 1 solver: a Large Language Model. As mentioned earlier, the System 1 (S_1) solver in SOFAI-LM is an LLM, selected for its ability to rapidly generate initial candidate solutions for a given problem instance. Let x denote the problem input and M represent the episodic memory (a set of previously encountered examples and their solutions). The S_1 solver produces a candidate solution y as follows:

$$S_1(x, M) \rightarrow y, \quad (1)$$

where y is the output generated by the LLM for input x .

Metacognitive Governance Module (MC). The metacognitive governance module (MC) of the architecture orchestrates the reasoning process by evaluating, guiding, and monitoring the outputs of the S_1 solver, and deciding when and if to invoke the S_2 solver. MC operates through the following sequential steps:

- **Evaluation:** After the S_1 solver generates a candidate solution y , MC evaluates its correctness using a problem-specific correctness function. For example, in graph coloring, the correctness function $C(y)$ can be defined as:

$$C(y) = \frac{\sum_{(u,v) \in E} \mathbb{I}[y(u) \neq y(v)]}{|E|}, \quad (2)$$

where E is the set of edges in the graph, and $y(u)$ denotes the color assigned to vertex u .

Code debugging (DebugBench). For code-debugging tasks, we interface with the official *LeetCode* API. Given a candidate patched code y , the API compiles and executes it on all hidden test cases and returns the pass ratio

$$\text{Pass}(y) = \frac{\# \text{ tests passed by } y}{\# \text{ total tests}}. \quad (3)$$

MC deems y correct when $\text{Pass}(y) = 1.0$. If $\text{Pass}(y) < 1.0$, the API also returns the last failing testcase $\langle \text{input}, \text{expected output}, \text{actual output} \rangle$.

- **Feedback Generation:** If the solution y does not meet the required correctness threshold θ , MC generates feedback $F(y)$ for S_1 . This feedback may include information about specific errors, violated constraints, or example problems relevant to problem instance.

- **Iterative Refinement and Monitoring:** MC then iteratively refine the candidate solution by incorporating the feedback and re-invoking the S_1 solver:

$$y_{t+1} = S_1(x, M \cup F(y_t)), \quad t = 0, 1, \dots, T-1, \quad (4)$$

Where T is the maximum number of allowed iterations. During this process, MC continuously monitors improvement in solution quality, e.g., by checking whether $C(y_{t+1}) > C(y_t)$ at each iteration.

- **Solver Selection:** If the candidate solution fails to achieve the correctness threshold after T iterations, or if no further improvement is observed, MC invokes the System 2 (S_2) solver for final deliberation.

System 2 solver: Large Reasoning Model. The System 2 (S_2) solver is instantiated as a large reasoning model (LRM) chosen for its logical inference capabilities, usually stronger than those of LLMs. When invoked by MC, the S_2 solver receives the original problem instance x as input, and may also receive the final unsuccessful attempt y_T or the whole feedback history, depending on the implementation. The S_2 solver then produces the final solution y^* :

$$S_2(x, y_T, H) \rightarrow y^*, \quad (5)$$

where H denotes the available history of attempts and feedback, and y^* is the output of the LRM.

Generalist and Specific modules in SOFAI-LM. The architecture’s feedback loop and metacognitive monitoring require no additional training or fine-tuning for the language models employed, enabling straightforward adaptation to diverse reasoning tasks such as graph coloring and code debugging. On the other hand, the evaluation and feedback generation modules are domain-specific, since they need to check correctness (or other properties) and generate feedback for candidate solutions in specific problem domains.

Workflow Overview. The full SOFAI-LM workflow proceeds as follows:

1. The S_1 solver (an LLM) receives a problem instance x and generates an initial candidate solution y .

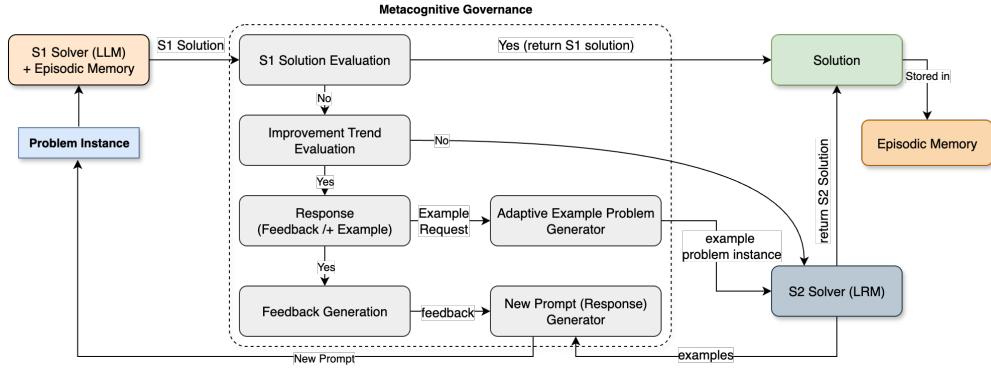


Figure 1: The SOFAI-LM architecture.

2. MC evaluates y using the correctness function $C(y)$.
3. If y does not meet the correctness threshold, MC generates feedback $F(y)$ and calls again the S_1 solver, monitoring progress for up to T iterations.
4. If no satisfactory solution is found after T iterations, or if improvement stagnates, MC invokes the S_2 solver to produce the final solution y^* .

Experimental Setting

To use SOFAI-LM to address considered research questions, we design a set of experiments spanning two domains and multiple architecture configurations. Our goal is to systematically analyze influence of feedback, episodic memory, and solver initialization on both performance and computational efficiency of employees’ LLM and LRM, as detailed below¹.

Language models

System 1 (LLM). As mentioned above, the S_1 solver in SOFAI-LM is an LLM. In particular, we use **Granite 3.3 8B** (without thinking) (IBM Research 2025) and, in separate experiments, also **Llama3.1** (Dubey et al. 2024) as S_1 .

System 2 (LRM). The S_2 solver in SOFAI-LM is an LRM. In particular, we use **DeepSeek R1 8B** (Guo et al. 2025), **Granite 3.3 8B** (with thinking) (IBM Research 2025), and **Qwen 3 8b** (Yang et al. 2025).

All models’ inference work was conducted using the Ol-lama (Marcondes et al. 2025), with decoding parameters set to `seed=12345`, `temperature=0.0`, `top_k=1`, and `top_p=1.0`, yielding greedy sampling and improved output stability across runs. This standardized configuration facilitates more consistent comparisons between LLMs’ and LRMs’ performance under matched inference conditions.

Evaluation Domains

- **Graph Coloring Decision Problem:** For this domain, we generate both solvable and unsolvable instances of graphs ranging from 5 to 25 vertices. For each size, 100 unique

¹Additional domain-specific details and prompts examples are provided in the Supplementary Material accompanying this paper.

instances are created, with edge probabilities sampled uniformly from $[0.1, 0.9]$ to ensure a broad spectrum of structural complexity and constraint tightness. This enables evaluation of LLM and LRM performance under varying degrees of combinatorial difficulty and solution feasibility.

- **DebugBench:** We use the Python and C++ subsets of DebugBench (Tian et al. 2024), a benchmark designed to assess the practical debugging abilities of LLMs and LRMs on realistic program repair tasks. The Python subset consists of diverse bug types, including logic errors, API misuse, and syntax errors, spanning both single- and multi-function programs. The C++ subset emphasizes language-specific pitfalls such as pointer misuse, memory management bugs, and object lifetime errors, providing a rigorous test of model reasoning in complex, stateful codebases. Each DebugBench instance contains one localized bug, requiring the solver to identify and repair the fault while preserving the correctness of the unaffected code.

Feedback and Episodic Memory Variants

To analyze the impact of the metacognitive module’s feedback and memory on the LLMs and LRMs, we define and test the following variants:

Feedback Types

- **Multi-Line Feedback (MLF):** Feedback containing all error information and violated constraints or failing test cases, presented in a structured, hierarchical, multi-sentence format. This organization aims to make each issue explicit and easier to parse for the LLM.
- **Single-Line Feedback (SLF):** The same content as MLF, but compressed into a single line, providing concise guidance while preserving all informational content.

Episodic Memory Variants

- **Minimal Episodic Memory (MEM):** Stores only pairs including a problem instance and its correct solution, with no historical context.
- **Extended Episodic Memory (EEM):** Stores the full sequence of LLM solution attempts and all feedback exchanges for each instance, along with the problem instances and their correct solutions.

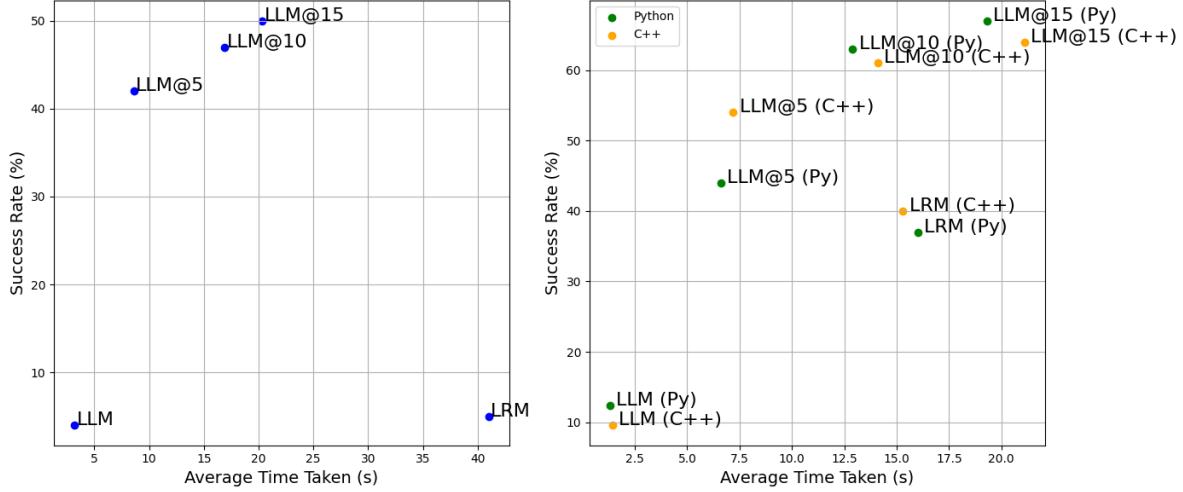


Figure 2: Each point corresponds to a configuration: **LLM**, **LLM@5**, **LLM@10**, **LLM@15**, and **LRM**. Left: Graph coloring problems (Solvable, size = 25). Right: Code debugging (Python and C++). The x-axis shows the average time per instance (in sec), while the y-axis shows the success rate (in %). The LLM is Granite3.38b (without thinking); the LRM is Deepseek R18b.

LRM prompting

When the metacognitive governance module determines that the LLM cannot solve a problem after T iterations, an LRM is invoked with one of the following prompts:

- **Problem Instance-Only (PO):** the LRM receives only the original problem statement.
- **Best Attempt (BA):** LRM is provided with the best solution produced by LLM, along with the problem instance.
- **Full History (FH):** LRM is given entire history of LLM’s attempts and MC feedback for the problem instance.

Results

Given the above experimental setting, we now describe the experiments we conducted to address the considered research questions and the results shown by the experiments.

RQ1: Can a feedback-driven LLM outperform an LRM?

To address this question, we evaluate whether the SOFAI-LLM feedback loop enables an LLM (in this case, Granite3.38b without thinking) to match or surpass the performance of an LRM (in this case, DeepSeek R18B) across the two considered problem domains: graph coloring and code debugging.

Figure 2 presents a comparative analysis of various configurations, noted as follows: LLM, LLM@5 (i.e., the LLM with ≤ 5 iterations with feedback loop), LLM@10, LLM@15, and LRM. The figure shows that increasing the number of iterations leads to a consistent increase in solved problems, with the most significant improvements observed between LLM and LLM@5. While LLM@15 achieves the highest success rate in both domains, it does so with more compute time. Notably, the LRM, though faster on average, performs substantially worse on larger graph sizes in the graph coloring domain. Overall, it is clear that iterating the

use of the LLM dominates the LRM, since it can solve many more problem instances while using much less time².

RQ2: How do the type of feedback and the format of episodic memory affect a feedback-driven LLM?

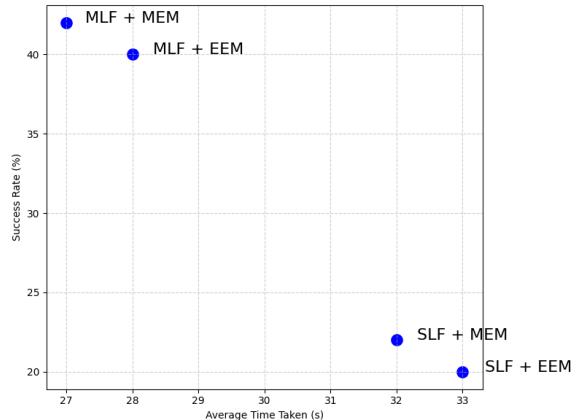


Figure 3: The x-axis shows the average time per instance (in sec), while the y-axis shows the success rate (in %) for four metacognitive configurations, for the graph coloring domain (with graphs of size 25). Feedback types: Minimal Line Feedback (MLF) and Single Line Feedback (SLF), each paired with Minimal Episodic Memory (MEM) or Extended Episodic Memory (EEM). The LLM is Granite3.38b (without thinking) and the LRM is Deepseek R18b.

To respond to this question, we conducted an experiment with four variants of the metacognitive loop in the graph coloring domain (graph size = 25): Minimal Line Feed-

²Results with other LLMs and LRMs are consistent and are included in the Supplementary Material accompanying this paper.

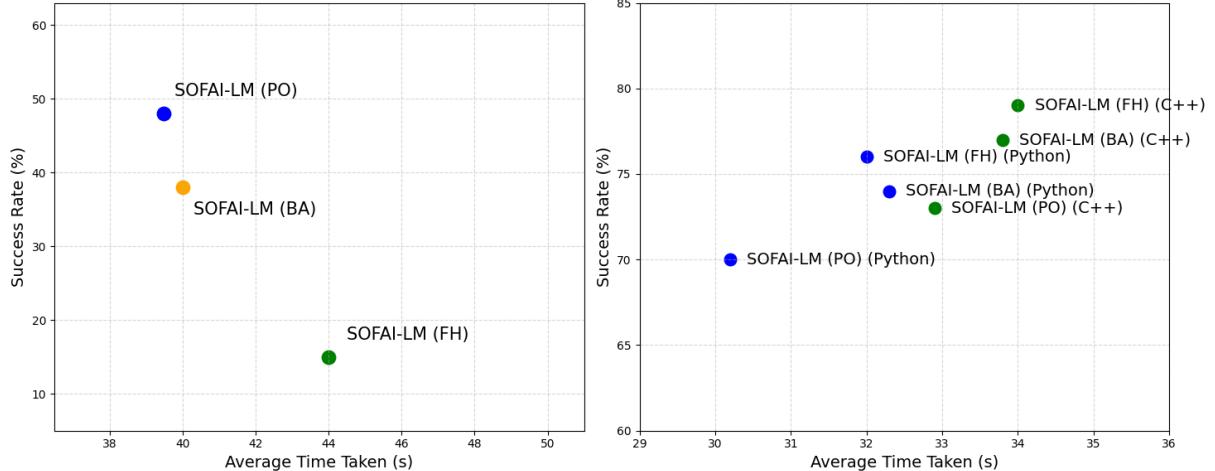


Figure 4: Left: graph coloring, with graph size 25. Right: code debugging (Python and C++). The x-axis shows the average time per instance (in sec), while the y-axis shows the success rate (in %) for SOFAI-LM using three LRM prompting strategies: PO (just the problem instance), BA (best attempt from LLM), and FH (full history of LLM@15 attempts). The LLM is Granite3.38b (without thinking) and the LRM is Deepseek R18b.

back (MLF) and Single Line Feedback (SLF), each paired with either Extended Episodic Memory (EEM) or Minimal Episodic Memory (MEM).

Figure 3 shows the success rate versus time for these four configurations. We observe that using MLF consistently outperforms SLF, achieving a higher success rate with lower compute time across both episodic memory settings. Additionally, MEM yields better performance than its EEM counterpart, suggesting that limited context stored in the episodic memory improves the model’s ability to focus on relevant corrections. The best performing configuration is **MLF + MEM**, which achieves the highest success rate (42%) at the lowest average time (27s). In contrast, SLF + EEM lags both in success rate and time, indicating that both overly concise feedback and redundant historical context can hinder reasoning performance. Summarizing, the combination of MLF with memory-limited feedback (MEM) offers the best trade-off between success rate and time. Because of this analysis, for the code debugging domain, we adopt the SLF + MEM setting across all experiments².

RQ3: Can the information gathered by SOFAI-LM, when used iteratively with an LLM, enhance the performance of an LRM?

To investigate whether information generated during LLM iterations benefits the subsequent LRM solver, we compare three prompting strategies when calling the LRM: (1) PO (just problem instance), where LRM receives no prior context from LLM; (2) BA (Best Attempt), where LRM is called with most promising partial solution produced by LLM; and (3) FH (Full History), where LRM is provided full sequence of LLM’s attempts and MC feedback.

Figure 4 shows the trade-off between success rate and time for all three LRM prompting strategies, for both the graph coloring domain (left) and the code debugging domain (right). In the graph coloring domain, using PO con-

sistently yields the highest success rate and most favorable time efficiency, with the MLF+MEM configuration. While BA and FH can leverage additional context, they often introduce noise and degrade performance, especially in more complex memory and feedback settings. On the other hand, for the code debugging domain, results are the opposite: BA and FH show improvements in success rate over PO, albeit at the cost of increased time. To justify this difference between the two domains, we conjecture that, in domains where fixing an incorrect solution requires only a local revision, such as in code debugging, providing negative examples is helpful. On the contrary, in domains where fixing an incorrect solution requires a possible revision of the whole solution, like in graph coloring, then it is not helpful, and actually detrimental, to provide negative examples and feedback on them².

RQ4: Does SOFAI-LM perform better than its LRM counterpart?

To address this question, we compare the behavior of SOFAI-LM, where the LLM and metacognitive controller drive most reasoning and selectively invoke the LRM only when needed, versus using the LRM alone. We assess both the overall success rate and computation time per instance across the graph coloring (with graph size 25) and code debugging (both in Python and C++) domains.

Figure 5 presents the results for both domains. The SOFAI-LM pipeline consistently achieves a much higher success rate compared to the LRM in both domains, while also reducing the compute time, especially in the graph coloring domain. Notably, SOFAI-LM solves over 70% of code debugging problems (Python: 70%, C++: 73%) in less time than the LRM, which solves only 37% (Python) and 40% (C++) with higher average runtimes. In the graph coloring domain, the advantage is even more pronounced: SOFAI-LM solves 42% of the problems with a much lower time,

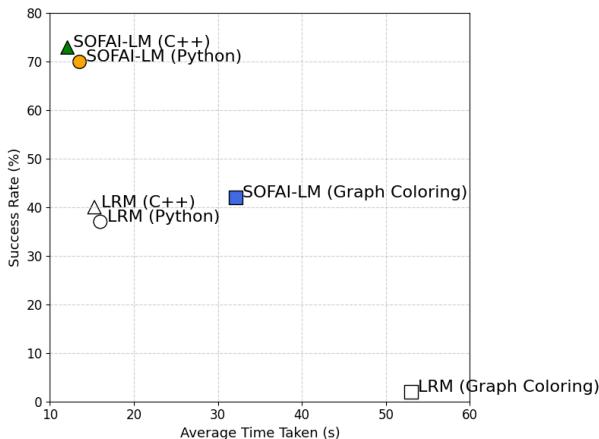


Figure 5: The x-axis shows the average time per instance (in sec), while the y-axis shows the success rate (in %) for SOFAI-LM and LLM approaches across three settings: graph coloring (size 25), DebugBench-Python, and DebugBench-C++. Each point is marked by both domain and method. The LLM is Granite3.38b (without thinking) and the LRM is Deepseek R18b.

whereas the LRM solves just 2% of the problems and takes significantly longer. Therefore, selective fallback to LRM, as orchestrated by the SOFAI-LM pipeline, delivers a superior trade-off between success rate and efficiency compared to LRM reasoning alone, across both reasoning tasks².

Related Work

LLM-based Reasoning and Constraint Satisfaction. LLMs have demonstrated emerging reasoning capabilities using prompting strategies such as chain-of-thought (CoT) reasoning (Wei 2022), zero-shot reasoning (Kojima et al. 2022), scratchpads (Nye 2021), and self-consistency (Wang 2022). Extensions such as tool augmentation (Schick, Dwivedi-Yu, and Lazaridou 2023; Paranjape and Chen 2023) and iterative self-refinement (Shinn and Cassano 2023; Madaan 2023) enable models to tackle complex reasoning without explicit training. Iterative verification approaches, such as PiVe (Han 2023) and debate-based reasoning (Li 2023), further improve solution quality. However, empirical evaluations reveal persistent weaknesses in enforcing hard constraints and guaranteeing globally correct solutions, particularly in combinatorial tasks like graph coloring (Kambhampati 2024; Valmeekam 2022; Stechly, Valmeekam, and Kambhampati 2024) and in program repair benchmarks like DebugBench (Tian et al. 2024). These limitations motivate architectures that combine the adaptability of LLMs with the reliability of structured reasoning methods.

Fine-tuning and Verifier-Augmented Models. Several efforts aim to bridge the gap between LLMs and robust reasoning through additional training or verifier integration. Reinforcement learning-based fine-tuning, as in DeepSeek-R1 (Guo et al. 2025) and ThinkGPT (Besta 2023), biases mod-

els toward stepwise reasoning behaviors, while supervised alignment methods enhance code reasoning (Chen et al. 2021). Verifier-augmented approaches explicitly incorporate symbolic or learned checkers (Lightman 2023; Paul 2023), improving accuracy on domains such as formal mathematics (Polu and Sutskever 2022) and theorem proving (Welleck 2022). Despite these gains, such models require substantial compute resources and often lose task generality, making them difficult to transfer between diverse problem classes such as graph-based CSPs and software debugging.

Neuro-Symbolic and Metacognitive Architectures. Neuro-symbolic architectures inspired by dual-process cognition (Kahneman 2011) combine fast heuristic reasoning (“System 1”) with slow deliberate reasoning (“System 2”). The SOFAI framework (Booch et al. 2021) instantiated this paradigm in AI, improving planning and pathfinding performance via instance-level solver selection (Fabiano et al. 2023; Ganapini et al. 2022; Lin et al. 2024). Related work integrates symbolic solvers or verifiers into neural pipelines, e.g., differentiable reasoning modules (Evans 2021), SAT/SMT solver integration (Amizadeh 2020), and theorem-proving hybrids (Selsam 2019). More recent approaches explore metacognition for LLMs, including self-evaluation and strategy adaptation (Shinn and Cassano 2023; Madaan 2023). However, these frameworks often rely on static solver selection or additional trained verifiers, limiting flexibility and adding engineering overhead.

Conclusions and Future Work

This paper introduced SOFAI-LM, a training-free metacognitive architecture that couples the speed and flexibility of large language models (LLMs) with the robust reasoning capabilities of large reasoning models (LRMs). This work enables LLMs to iteratively refine their outputs through targeted feedback and, when necessary, selectively invoke LRMs. Across two distinct reasoning domains—global constraint satisfaction in graph coloring and localized program repair in code debugging—SOFAI-LM consistently matches or surpasses the performance of standalone LRMs while requiring significantly less computation. A key advantage of SOFAI-LM is its model-agnostic design: any LLM can serve as the fast solver and any LRM as the slow solver, with domain-specific evaluation and feedback modules providing the only customization required. This flexibility makes SOFAI-LM applicable to a broad range of reasoning challenges beyond those explored here. Our experiments further show how differences in problem structure—global versus local fixes—affect the value of information sharing between the iterative LLM loop and LRM fallback, offering insights into how hybrid systems can be tailored to specific domains. Future work will focus on automating and optimizing the metacognitive module itself. By learning policies for solver selection, iteration depth, and feedback and evaluation strategies, we aim to create a fully self-improving reasoning framework. Such a system could dynamically adapt across tasks, further reducing computational cost while improving reliability, bringing us closer to scalable, domain-general AI reasoning.

References

- Amizadeh, S. e. a. 2020. NeuroSAT: End-to-End SAT Solver Learning. *ICLR*.
- Besta, M. e. a. 2023. ThinkGPT: Enhancing LLM Reasoning with Chain-of-Thought Fine-Tuning. *arXiv preprint arXiv:2309.02664*.
- Booch, G.; Fabiano, F.; Horesh, L.; Kate, K.; Lenchner, J.; Linck, N.; Loreggia, A.; Murugesan, K.; Mattei, N.; and Rossi, F. e. a. 2021. Thinking fast and slow in AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15042–15046.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; and Askell, A. e. a. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; and Fan, A. e. a. 2024. The Llama 3 Herd of Models. *CoRR*.
- Evans, R. e. a. 2021. Making Neural Deduction Differentiable. *Nature Machine Intelligence*.
- Fabiano, F.; Ganapini, M. B.; Loreggia, A.; Mattei, N.; Murugesan, K.; Pallagani, V.; Rossi, F.; Srivastava, B.; and Venable, K. B. 2025. Thinking Fast and Slow in Human and Machine Intelligence. *Commun. ACM*, 68(8): 72–79.
- Fabiano, F.; Pallagani, V.; Ganapini, M. B.; Horesh, L.; Loreggia, A.; Murugesan, K.; Rossi, F.; and Srivastava, B. 2023. Plan-SOFAI: A Neuro-Symbolic Planning Architecture. In *Neuro-Symbolic Learning and Reasoning in the era of Large Language Models*.
- Ganapini, M. B.; Campbell, M.; Fabiano, F.; Horesh, L.; Lenchner, J.; Loreggia, A.; Mattei, N.; Rossi, F.; Srivastava, B.; and Venable, K. B. e. a. 2022. Combining Fast and Slow Thinking for Human-like and Efficient Decisions in Constrained Environments. In *NeSy*, 171–185.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; and Bi, X. e. a. 2025. Deepseekr1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Han, J. e. a. 2023. PiVe: Prompting with iterative verification improving graph-based generative capability of LLMs. *arXiv:2305.12392*.
- IBM Research. 2025. Granite 3.3 8B Instruct Model Card. <https://huggingface.co/ibm-granite/granite-3.3-8b-instruct>. Accessed on 2025-08-01.
- Jiang, Y.; Wang, Y.; Zeng, X.; Zhong, W.; Li, L.; Mi, F.; Shang, L.; Jiang, X.; Liu, Q.; and Wang, W. 2023. FollowBench: A Multi-Level Fine-Grained Constraints Following Benchmark for Large Language Models. *arXiv preprint arXiv:2310.20410*.
- Johnson, D. S.; and Trick, M. A. 1996. *Cliques, coloring, and satisfiability: second DIMACS implementation challenge, October 11-13, 1993*, volume 26. American Mathematical Soc.
- Kahneman, D. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.
- Kambhampati, S. 2024. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1): 15–18.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.
- Li, Y. e. a. 2023. Improving Factuality via Multi-Agent Debate and Self-Consistency. *arXiv preprint arXiv:2305.14325*.
- Lightman, A. e. a. 2023. Let’s Verify Step by Step. *arXiv preprint arXiv:2305.20050*.
- Lin, B. Y.; Fu, Y.; Yang, K.; Brahman, F.; Huang, S.; Bhagavatula, C.; Ammanabrolu, P.; Choi, Y.; and Ren, X. 2024. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*, 36.
- Madaan, A. e. a. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint arXiv:2303.17651*.
- Marcondes, F. S.; Gala, A.; Magalhães, R.; Perez de Britto, F.; Durães, D.; and Novais, P. 2025. Using ollama. In *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, 23–35. Springer.
- McConnell, S. 2004. *Code complete*. Pearson Education.
- Nye, M. e. a. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv:2112.00114*.
- Paranjape, A.; and Chen, X. e. a. 2023. Hindsight Chain-of-Thought Reasoning. In *NeurIPS*.
- Paul, D. e. a. 2023. Chain-of-Verification Reduces Hallucination in LLM Reasoning. *arXiv preprint arXiv:2309.11495*.
- Polu, S.; and Sutskever, I. 2022. Minerva: Solving Quantitative Reasoning Problems with Language Models. *arXiv preprint arXiv:2206.14858*.
- Schick, T.; Dwivedi-Yu, J.; and Lazaridou, A. e. a. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*.
- Selsam, D. e. a. 2019. Learning a SAT Solver from Single-Bit Supervision. *ICML*.
- Shinn, N.; and Cassano, F. e. a. 2023. Reflexion: An Autonomous Agent with Dynamic Memory and Self-Reflection. *arXiv preprint arXiv:2303.11366*.
- Stechly, K.; Valmeeekam, K.; and Kambhampati, S. 2024. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*.
- Tian, R.; Ye, Y.; Qin, Y.; Cong, X.; Lin, Y.; Pan, Y.; Wu, Y.; Haotian, H.; Weichuan, L.; and Liu, Z. e. a. 2024. DebugBench: Evaluating Debugging Capability of Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, 4173–4198.

Valmeekam, K.; Olmo, A.; Sreedharan, S.; and Kambham-
pati, S. 2022. Large Language Models Still Can't Plan:
A Benchmark for LLMs on Planning and Reasoning about
Change. In *NeurIPS 2022 Foundation Models for Decision
Making Workshop*.

Valmeekam, K. e. a. 2022. Large language models still can't
plan: A benchmark for LLMs on planning and reasoning
about change. In *NeurIPS Foundation Models for Decision
Making Workshop*.

Wang, X. e. a. 2022. Self-consistency improves chain-of-
thought reasoning in language models. *arXiv:2203.11171*.

Wei, J. e. a. 2022. Chain-of-thought prompting elicits rea-
soning in large language models. In *NeurIPS*.

Welleck, S. e. a. 2022. NaturalProofs: Formal Theorem
Proving with LLMs. In *ICML*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.;
Yu, B.; Gao, C.; Huang, C.; and Lv, C. e. a. 2025. Qwen3
technical report. *arXiv preprint arXiv:2505.09388*.

Supplementary Material

Algorithm 1: SOFAI-LM for Graph Coloring

Algorithm 1: Feedback-Driven LLM with LRM Fallback (Graph Coloring)

Require: Undirected graph $G = (V, E)$, color limit k , max iterations T

Ensure: Coloring assignment or “NOT SOLVABLE”

```

1:  $history \leftarrow []$  {stores  $(assign, feedback)$ }
2:  $trends \leftarrow []$  {stores feedback and mistakes}
3: for  $t = 1$  to  $T$  do
4:    $prompt \leftarrow \text{GenerateGCPrompt}(G, k, history)$ 
5:    $resp \leftarrow \text{CallLLM}(prompt)$ 
6:    $assign \leftarrow \text{ParseColoring}(resp)$ 
7:    $C \leftarrow \frac{|\{(u,v) \in E : assign[u] \neq assign[v]\}|}{|E|}$ 
8:   if  $C = 1.0$  then
9:     return  $assign$  {all edges valid  $\rightarrow$  solved}
10:    end if
11:     $conflicts \leftarrow \{(u, v) \in E : assign[u] = assign[v]\}$ 
12:     $subG \leftarrow \text{InducedSubgraph}(G, \text{vertices in } conflicts)$ 
13:     $fb \leftarrow \text{FormatFeedbackMLF}(conflicts, subG)$ 
14:     $history.append(assign, fb)$ 
15:     $trends.append(fb)$ 
16:  end for
17:  {Three prompting variants:}
18: // (1) PO: problem instance only
19: 20:  $p_{PO} \leftarrow \text{GenerateLRMPromptPO}(G, k)$ 
21:  $r_{PO} \leftarrow \text{CallLRM}(p_{PO})$ 
22: // (2) BA: best LLM attempt + problem
23:  $best \leftarrow \text{argmax}(history, \text{highest } C)$ 
24:  $p_{BA} \leftarrow \text{GenerateLRMPromptBA}(G, k, best)$ 
25:  $r_{BA} \leftarrow \text{CallLRM}(p_{BA})$ 
26: // (3) FH: full LLM feedback history
27:  $p_{FH} \leftarrow \text{GenerateLRMPromptFH}(G, k, history)$ 
28:  $r_{FH} \leftarrow \text{CallLRM}(p_{FH})$ 
29: Choose final  $r_*$  based on selected variant
30:  $assign^* \leftarrow \text{ParseColoring}(r_*)$ 
31:
32: return  $assign^*$ 

```

Description (Graph Coloring):

- **Evaluation Function $C(assign)$:** Computes the fraction of edges properly colored (Eq. 2 in paper): $C = \frac{|\{(u,v) \in E : assign[u] \neq assign[v]\}|}{|E|}$. A value of 1.0 means a valid coloring.
- **Improvement Trend:** Track fb each iteration to measure progress and detect stagnation. If stagnation detected, it invokes LRM
- **Adaptive Feedback:** On conflicts, extract the induced subgraph of miscolored edges and generate multi-line feedback (MLF) listing conflicting pairs and the subgraph structure.
- **LRM Prompting Levels:**
 1. PO (Problem-Only): Only the original graph and k .

2. BA (Best Attempt): Graph, k , and the single best LLM assignment.
 3. FH (Full History): Graph, k , and the entire $(assign, feedback)$ history.
- **Fallback Selection:** Empirically, PO yields highest success for graph coloring (Fig. 4 left). The module selects the variant that maximizes solve rate vs. time.

Algorithm 2: SOFAI-LM for Code Debugging

Algorithm 2: Feedback-Driven LLM with LRM Fallback (Code Debugging)

Require: Buggy snippet C_b , description x , test suite \mathcal{T} , max iterations T

Ensure: Patched code C^*

```

1:  $history \leftarrow []$  {stores  $(code, feedback)$ }
2:  $trends \leftarrow []$  {stores feedback and mistakes}
3: for  $t = 1$  to  $T$  do
4:    $prompt \leftarrow \text{GenerateCDPrompt}(C_b, x, history)$ 
5:    $resp \leftarrow \text{CallLLM}(prompt)$ 
6:    $code \leftarrow \text{ParseCode}(resp)$ 
7:    $(passed, total) \leftarrow \text{LeetCodeAPI.Test}(code, \mathcal{T})$ 
8:    $P \leftarrow passed/total$ 
9:   if  $P = 1.0$  then
10:    return  $code$  {all tests passed  $\rightarrow$  solved}
11:   end if
12:    $fails \leftarrow \text{IdentifyFailures}(code, \mathcal{T})$ 
13:    $fb \leftarrow \text{FormatFeedbackSLF}(fails, passed, total)$ 
14:    $history.append(code, fb)$ 
15:    $trends.append(fb)$ 
16: end for
17: {Three prompting variants:}
18: // PO: only problem description
19: 20:  $p_{PO} \leftarrow \text{GenerateLRMPromptPO}(C_b, x)$ 
21:  $r_{PO} \leftarrow \text{CallLRM}(p_{PO})$ 
22: // BA: problem + last LLM code
23:  $last \leftarrow history[-1].code$ 
24:  $p_{BA} \leftarrow \text{GenerateLRMPromptBA}(C_b, x, last)$ 
25:  $r_{BA} \leftarrow \text{CallLRM}(p_{BA})$ 
26: // FH: full LLM feedback history
27:  $p_{FH} \leftarrow \text{GenerateLRMPromptFH}(C_b, x, history)$ 
28:  $r_{FH} \leftarrow \text{CallLRM}(p_{FH})$ 
29: Select final  $r_*$  based on variant trade-offs
30:  $code^* \leftarrow \text{ParseCode}(r_*)$ 
31:
32: return  $code^*$ 

```

Description (Code Debugging):

- **LeetCode API Evaluation $P(code)$:** Returns pass ratio $P = \frac{\#passed}{\#total}$; a value of 1.0 indicates full correctness.
- **Improvement Trend:** Records feedback and mistakes each iteration to monitor learning. If it identifies stagnation, LRM is invoked.
- **Feedback Generation:** Use single-line feedback (SLF) listing failing testcase IDs and pass/total for concise guidance.

- **LRM Prompting Levels:** Same three variants (PO, BA, FH). For code debugging, BA and FH often improve solve rate at cost of time (Fig. 4 right).
- **Fallback Selection:** The metacognitive module chooses the prompting level that best balances accuracy and inference time for the domain.

Sample Input Prompts

Graph Coloring ($|V| = 10, k = 4$)

Task: Graph Coloring Decision Problem (< 5 colors)
 You must assign an integer color to every vertex of the undirected graph below such that no two adjacent vertices share the same color. If no coloring exists using at most 4 colors, respond exactly with: NOT SOLVABLE
 Do not output anything else or use quotes.

Input Graph
 p edges 10 12
 c edges
 a b
 a c
 b d
 b e
 c f
 d g
 e h
 f i
 g j
 h i
 h j
 i j

Output Format
 Provide one (vertex color) pair per line, sorted lexicographically:
 (a 1)
 (b 2)
 ...
 Where 'color' is an integer in the inclusive range [1, 4].
 Return only this list, with no prose, headings, or extra punctuation.

Constraints Recap
 - Use < 5 distinct colors.
 - Vertex identifiers are case-sensitive and must match those in the input.
 - Output must be either the ordered list above or the token 'NOT SOLVABLE'.

Code Debugging ('the-kth-factor-of-n')

You are an expert programmer specializing in code debugging. Your task is to analyze and fix the provided code snippet based on the problem description. IMPORTANT: You MUST return the complete, corrected code enclosed within `<code>` and `</code>` tags. Do not include any other explanatory text in your response.

Problem Description

You are given two positive integers n and k . A factor of an integer n is defined as an integer i where $n \% i == 0$. Consider a list of all factors of n sorted in ascending order, return the k th factor in this list or return -1 if n has less than k factors.

Buggy Code

```
```  

class Solution:

 def kthFactor(self, n: int, k: int) -> int:

 j = 0

 for i in range(1, n + 1):

 if n % i == 0:

 num = i

 j += 1

 if j == k:

 break

 return num if j == k+1 else -1
```

```

Correct Code:

Figure 6: Sample input prompts for the Graph Coloring and Code Debugging domains.

Detailed Walk-through: Graph Coloring

This section provides a step-by-step illustration of the SOFAI-LM architecture's workflow on a representative graph coloring problem, with a maximum of two iterations using LLM. We demonstrate the iterative feedback loop, the generation of adaptive examples, successful and unsuccessful resolution paths by the LLM, and the final structure of the episodic memory.

Problem Instance

The architecture is tasked with solving a 4-coloring problem for a graph with 10 vertices and 12 edges. The prompt is provided to the System 1 (LLM) solver (Figure 7).

Scenario 1: Initial Attempt and Feedback Generation

The LLM processes the prompt and produces its first candidate solution (Figure 8). This solution contains several constraint violations. Specifically, it assigns the same color to adjacent vertices in the subgraph formed by nodes $\{h, i, j\}$.

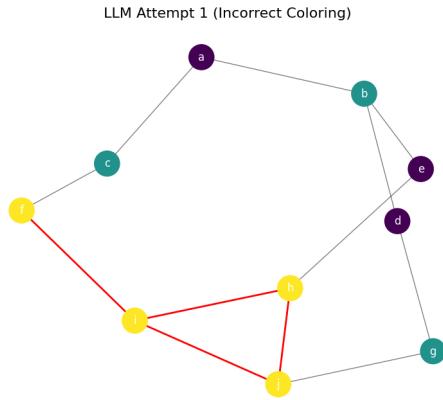


Figure 9: Visualization of the LLM's first incorrect attempt. Conflicting edges are highlighted in red.

Metacognitive Feedback The Metacognitive Governance (MC) module evaluates the solution, identifies the errors, and generates feedback in multiple formats to guide the LLM (Figure 10).

Scenario 2: Iterative Failure and LRM Fallback

This scenario illustrates the case where the LLM fails to find a solution within $T = 2$ iterations, triggering a fallback to the System 2 (LRM) solver. After receiving feedback, the LLM tries again but introduces a new error (Figure 11).

LLM Attempt 2 (Incorrect Coloring)

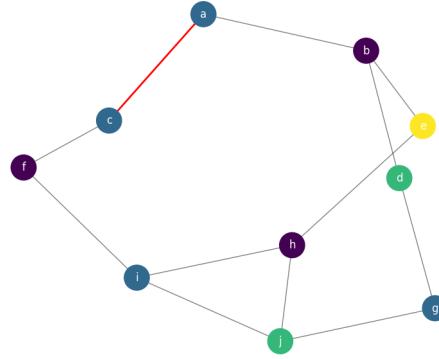


Figure 12: Visualization of the LLM's second incorrect attempt. The new conflicting edge (a,c) is highlighted.

Fallback to LRM Having failed to converge, the MC module invokes the LRM using the **Problem-Only (PO)** prompting strategy.

Scenario 3: Iterative Success

This path demonstrates a successful outcome. Using the feedback from its first attempt, the LLM generates a new, valid coloring (Figure 13).

LLM Attempt 2 (Correct Coloring)

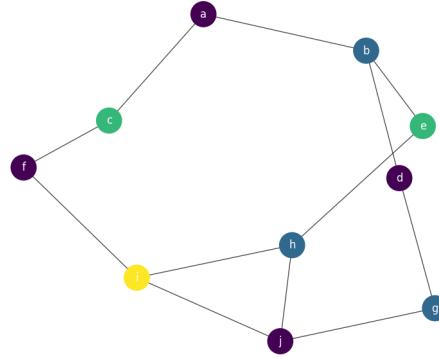


Figure 14: Visualization of the LLM's successful attempt. The coloring is valid.

Scenario 4: Episodic Memory Storage

Following the successful resolution, the MC module updates the episodic memory. The content stored depends on the chosen memory variant (Figure 15).

```
### Task: Graph Coloring Decision Problem (< 5 colors)
You must assign an integer color to every vertex of the undirected graph
below such that no two adjacent vertices share the same color. If no
coloring exists using at most 4 colors, respond exactly with: NOT SOLVABLE
Do not output anything else or use quotes.
```

```
### Input Graph
p edges 10 12
c edges
a b, a c, b d, b e, c f, d g, e h, f i, g j, h i, h j, i j
```

```
### Output Format
Provide one (vertex color) pair per line, sorted lexicographically:
```

```
(a 1)
(b 2)
```

```
...
```

```
Where 'color' is an integer in the inclusive range [1, 4].
```

```
Return only this list, with no prose, headings, or extra punctuation.
```

Figure 7: The initial prompt given to the LLM for the graph coloring task.

```
(a 1)  (b 2)  (c 2)  (d 1)  (e 1)
(f 3)  (g 2)  (h 3)  (i 3)  (j 3)
```

Figure 8: The LLM's first incorrect candidate solution.

Single-Line Feedback (SLF):

That was incorrect. The coloring is invalid:
adjacent-conflict(s) on pairs: (h,i), (h,j), (i,j).

Multi-Line Feedback (MLF):

That was incorrect. The coloring is invalid for the following reason(s):

1. adjacent-conflict: vertices h and i share color 3
2. adjacent-conflict: vertices h and j share color 3
3. adjacent-conflict: vertices i and j share color 3

Adaptive Example:

```
### Adaptive Example
```

To help you, here is a smaller, related subproblem:

Subproblem Graph:

```
p edge 3 3
e h i
e h j
e i j
```

Coloring:

```
(h 1)  (i 2)  (j 3)
```

Figure 10: Feedback generated by the MC module, including concise (SLF), structured (MLF), and adaptive example formats.

```
(a 2)  (b 1)  (c 2)  (d 3)  (e 4)
(f 1)  (g 2)  (h 1)  (i 2)  (j 3)
```

Figure 11: The LLM's second incorrect candidate solution.

```
(a 1)  (b 2)  (c 3)  (d 1)  (e 3)
(f 1)  (g 2)  (h 2)  (i 4)  (j 1)
```

Figure 13: The LLM’s second, and correct, candidate solution.

Minimal Episodic Memory (MEM)

```
{
  "problem_instance": {
    "task": "Graph Coloring Decision Problem (< 5 colors)",
    "graph": "p edges 10 12\n c edges\n a b\n ...",
    "k": 4
  },
  "correct_solution": "(a 1)\n(b 2)\n(c 3)\n(d 1)\n(e 3)\n..."
}
```

Extended Episodic Memory (EEM)

```
{
  "problem_instance": { ... },
  "interaction_history": [
    {
      "attempt": 1,
      "candidate_solution": "(a 1)\n(b 2)\n(c 2)\n...",
      "feedback_received": "That was incorrect. ..."
    },
    {
      "attempt": 2,
      "candidate_solution": "(a 1)\n(b 2)\n(c 3)\n...",
      "feedback_received": "Correct"
    }
  ],
  "correct_solution": "(a 1)\n(b 2)\n(c 3)\n..."
}
```

Figure 15: Comparison of memory storage variants. MEM stores the final solution, while EEM stores the full interaction history.

Detailed Walk-through: Code Debugging

This section details the SOFAI-LM workflow for a code debugging task from the DebugBench benchmark, with a maximum of two iterations using LLM. This domain requires localized fixes and benefits from a different feedback strategy than graph coloring.

Problem Instance

The task is to fix a bug in a Python function named `kthFactor`. The initial prompt provided to the LLM is shown in Figure 16.

Scenario 1: Iterative Success

This scenario shows the LLM successfully debugging the code after receiving targeted feedback from the testing environment.

LLM’s First Attempt (Incorrect) The LLM attempts a fix but introduces a new logical error. The loop condition is now incorrect, causing it to miss factors for some inputs. The proposed code is shown in Figure 17.

Metacognitive Feedback from Test Harness The submitted code is executed against a hidden test suite. The MC module receives the failure result and formats it as feedback (Figure 18). For code debugging, a concise feedback string (SLF) is effective.

LLM’s Second Attempt (Correct) Using the feedback, the LLM corrects its logic and produces the correct code (Figure 19). This version correctly iterates through all possible factors up to ‘n’.

This corrected code passes all test cases, and the MC module returns it as the final solution.

Scenario 2: LLM Fails & LRM is Invoked

In this scenario, the LLM fails to produce a correct solution after the maximum number of iterations.

Attempt 1 is the same as in the successful scenario (Figure 17), resulting in the feedback shown in Figure 18.

Attempt 2 (Incorrect): The LLM makes a different error, incorrectly handling the 1-based vs. 0-based indexing of ‘k’.

Fallback to LRM After this second failure, the MC module would invoke the LRM. As our results show, providing the LRM with the **Best Attempt (BA)** or **Full History (FH)** is beneficial in this domain. The LRM would receive the original problem and either the last code attempt or the full sequence of attempts and feedback to produce the final, correct solution.

You are an expert programmer specializing in code debugging. Your task is to analyze and fix the provided code snippet based on the problem description.
IMPORTANT: You MUST return the complete, corrected code enclosed within <code> and </code> tags. Do not include any other explanatory text in your response.

```
### Problem Description
You are given two positive integers n and k. A factor of an integer n is defined as an integer i where  $n \% i == 0$ . Consider a list of all factors of n sorted in ascending order, return the kth factor in this list or return -1 if n has less than k factors.

### Buggy Code
```python
class Solution:
 def kthFactor(self, n: int, k: int) -> int:
 j = 0
 for i in range(1, n + 1):
 if n % i == 0:
 num = i
 j += 1
 if j == k:
 break
 return num if j == k+1 else -1
```
```

```

Figure 16: The initial prompt for the code debugging task 'the-kth-factor-of-n'.

```
<code>
class Solution:
 def kthFactor(self, n: int, k: int) -> int:
 factors = []
 for i in range(1, n // 2 + 1): # Incorrect loop range
 if n % i == 0:
 factors.append(i)
 factors.append(n)

 if k <= len(factors):
 return factors[k - 1]
 else:
 return -1
</code>
```

Figure 17: The LLM's first, incorrect attempt to fix the buggy code.

Your code failed.  
- Failing Test Case Input: n = 4, k = 3  
- Expected Output: 4  
- Your Code's Output: -1  
- Reason: Your code's logic for finding factors is incorrect.  
Please re-examine the code and provide a new, corrected version.

Figure 18: Feedback generated after the first attempt failed a test case.

```

<code>
class Solution:
 def kthFactor(self, n: int, k: int) -> int:
 count = 0
 for i in range(1, n + 1):
 if n % i == 0:
 count += 1
 if count == k:
 return i
 return -1
</code>

```

Figure 19: The LLM’s second, successful attempt at fixing the bug.

```

<code>
class Solution:
 def kthFactor(self, n: int, k: int) -> int:
 factors = []
 for i in range(1, n + 1):
 if n % i == 0:
 factors.append(i)

 if k < len(factors): # Off-by-one error
 return factors[k]
 else:
 return -1
</code>

```

Figure 20: A second, different incorrect attempt from the LLM.

## Experimental Results : Graph Coloring

### RQ1: Can a feedback-driven LLM outperform an LRM?

The following figures illustrate the performance comparison between a feedback-driven LLM and a standalone LRM across various model combinations and graph sizes for solvable graph coloring problems.

#### Model Combination 1: Granite 3.3B (w/o thinking) vs. DeepSeek R1 8B

**Description** The set of figures (21 through 24) compares the success rate and average time taken for two different approaches to solving graph coloring problems. The first approach uses Granite 3.3B as a fast LLM, showing its performance as a single-pass model (LLM) and with 5, 10, and 15 iterations of feedback (LLM@5, LLM@10, LLM@15). The second approach uses DeepSeek R1 8B as a standalone slow-thinking LRM. Across all graph sizes, the iterative LLM configurations consistently achieve a higher success rate than the standalone LRM. As the number of iterations increases from 5 to 15, the LLM's success rate improves, albeit with a corresponding increase in time. Notably, as the graph size increases, the performance of the standalone LRM diminishes rapidly, whereas the iterative LLM maintains a significantly higher success rate, highlighting the efficacy of the feedback loop.

**Summary Finding** A feedback-driven LLM consistently and substantially outperforms a standalone LRM in both success rate and efficiency for solving graph coloring problems, with the performance gap widening as graph size increases.

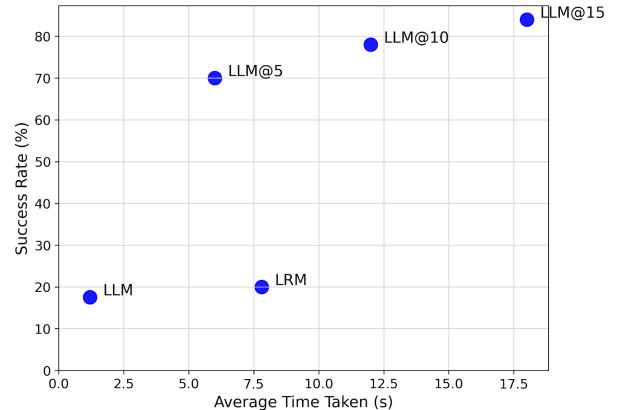


Figure 21: RQ1 - Model Combination 1, Size 5: Success rate versus average time for an iterative LLM (Granite 3.3B w/o thinking) and a standalone LRM (DeepSeek R1 8B) on solvable graph coloring problems of size 5.

We detail the performance of various model combinations across different problem sizes for the graph coloring domain.

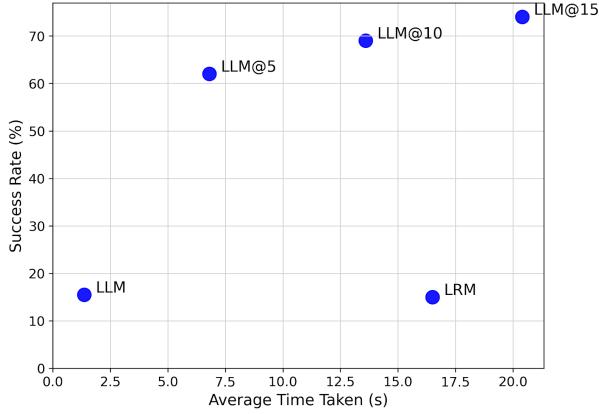


Figure 22: RQ1 - Model Combination 1, Size 10: Success rate versus average time for an iterative LLM (Granite 3.3B w/o thinking) and a standalone LRM (DeepSeek R1 8B) on solvable graph coloring problems of size 10.

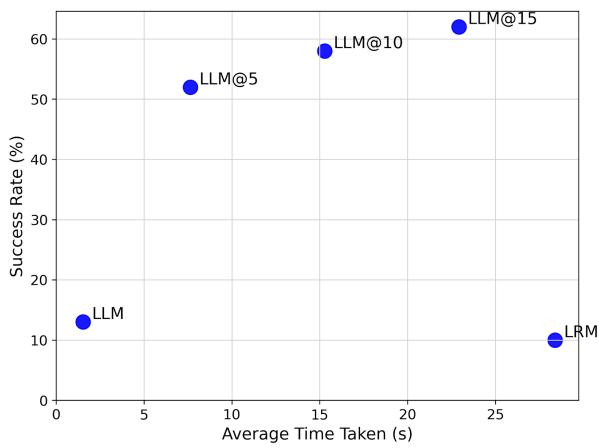


Figure 23: RQ1 - Model Combination 1, Size 15: Success rate versus average time for an iterative LLM (Granite 3.3B w/o thinking) and a standalone LRM (DeepSeek R1 8B) on solvable graph coloring problems of size 15.

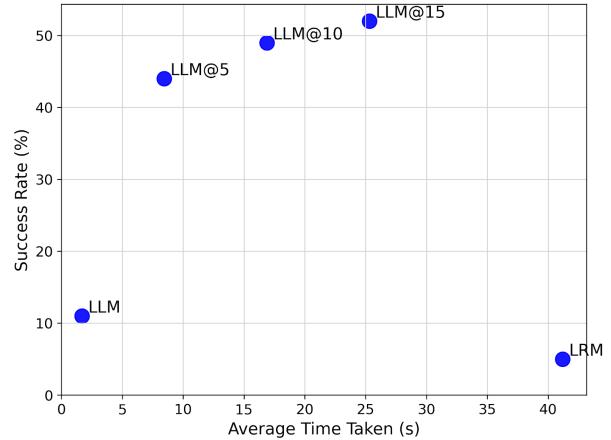


Figure 24: RQ1 - Model Combination 1, Size 20: Success rate versus average time for an iterative LLM (Granite 3.3B w/o thinking) and a standalone LRM (DeepSeek R1 8B) on solvable graph coloring problems of size 20.

## Model Combination 2: Granite 3.3B (w/o thinking) vs. Granite 3.3B (w/ thinking)

**Description** This series of plots (Figures 25 through 29) illustrates the performance trade-offs between using Granite 3.3B in an iterative feedback loop (LLM) versus its standalone reasoning mode (LRM). While the LRM is faster for smaller problems (sizes 5–15), its success rate is consistently lower than the iterative LLM configurations (LLM@5 and higher). For larger, more complex problems (sizes 20 and 25), the iterative LLM not only achieves a significantly higher success rate but also does so in less time than the standalone LRM. This demonstrates a clear advantage for the iterative, feedback-driven approach, even when comparing against the same base model in a specialized reasoning mode.

**Summary Finding** Iterative feedback enables a base LLM to achieve superior accuracy compared to the same model operating as a standalone LRM, and this performance advantage becomes more pronounced in both success rate and time efficiency as graph size increases.

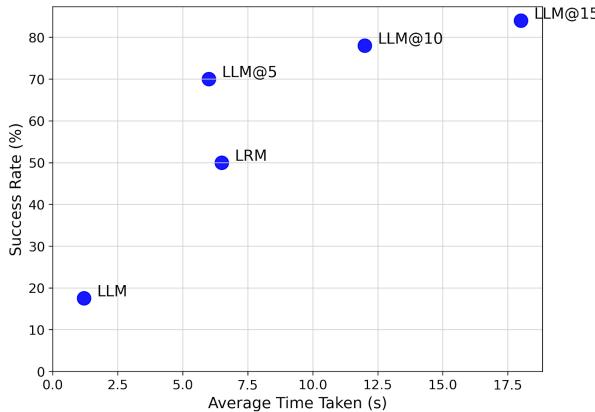


Figure 25: RQ1 - Model Combination 2, Size 5

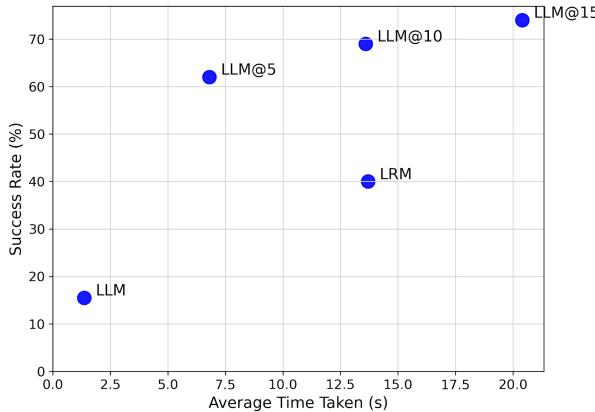


Figure 26: RQ1 - Model Combination 2, Size 10

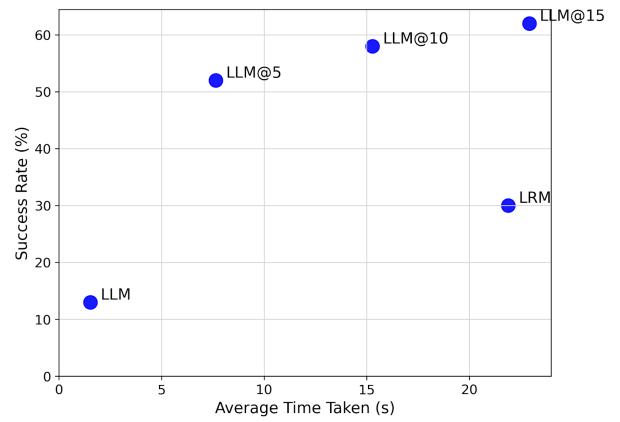


Figure 27: RQ1 - Model Combination 2, Size 15

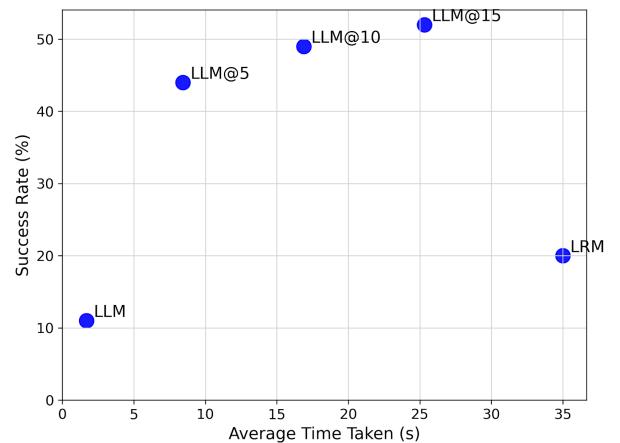


Figure 28: RQ1 - Model Combination 2, Size 20

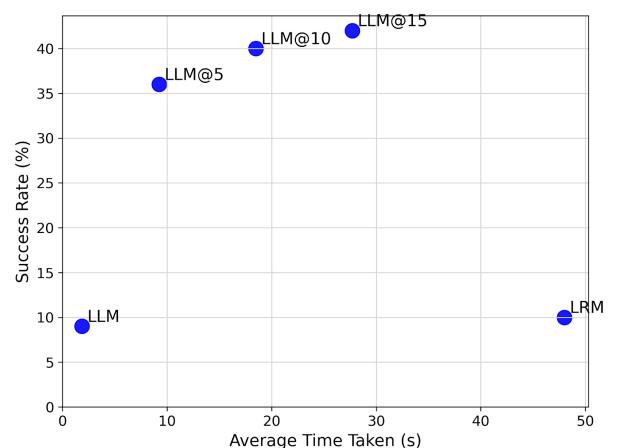


Figure 29: RQ1 - Model Combination 2, Size 25

### Model Combination 3: Granite 3.3B (w/o thinking) vs. Qwen 2.5 Pro

**Description** These figures (Figures 30 through 34) show the performance of the iterative Granite 3.3B LLM against the Qwen 2.5 Pro model acting as a standalone LRM. The trend is consistent across all graph sizes: the iterative LLM (LLM@5 and above) consistently achieves a much higher success rate than the Qwen 2.5 Pro LRM. While the LRM is faster, its low success rate makes it an unreliable standalone solver, especially as problem size increases. The feedback loop clearly empowers the Granite model to find correct solutions far more frequently, demonstrating a robust performance advantage over a capable LRM from a different model family.

**Summary Finding** An iterative LLM with feedback demonstrates a commanding performance lead in solution accuracy over a different, powerful standalone LRM, confirming that the benefit of the iterative architecture is not limited to a single model family.

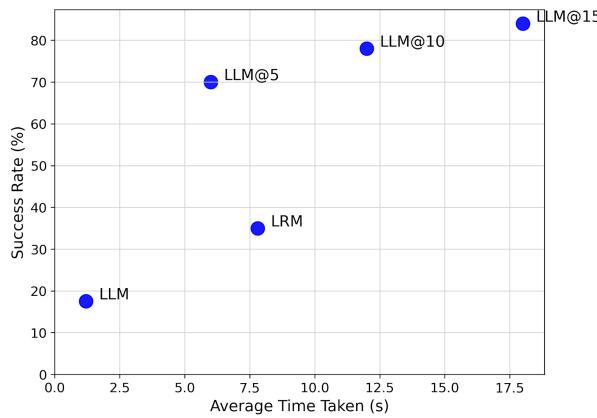


Figure 30: RQ1 - Model Combination 3, Size 5

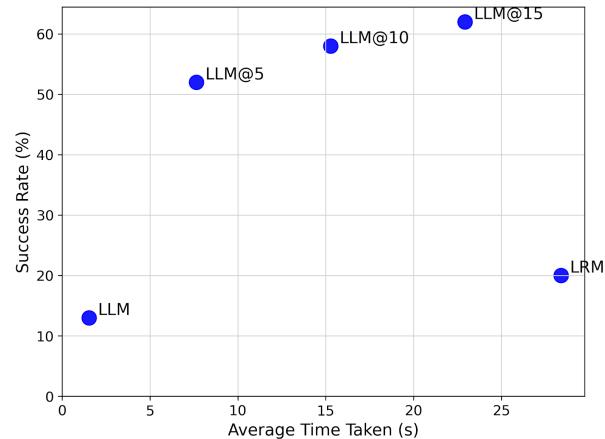


Figure 32: RQ1 - Model Combination 3, Size 15

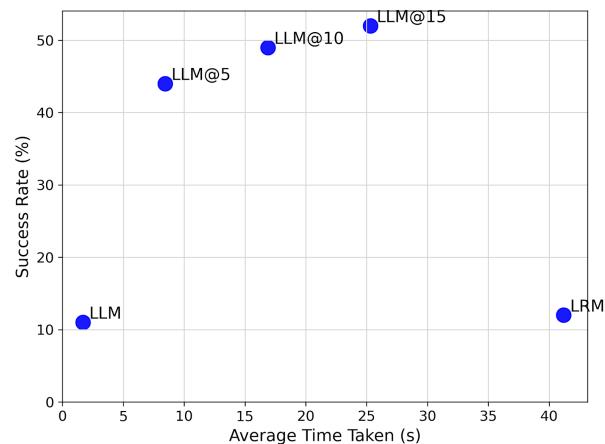


Figure 33: RQ1 - Model Combination 3, Size 20

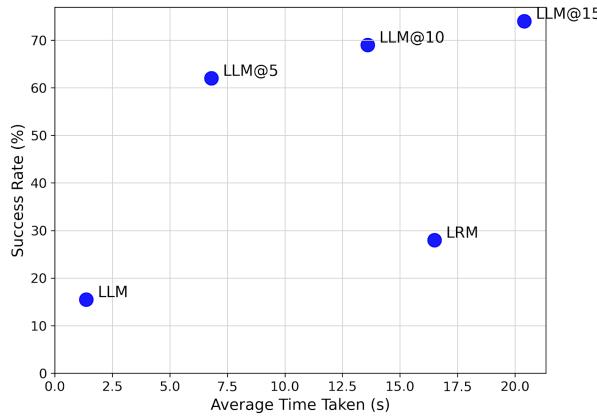


Figure 31: RQ1 - Model Combination 3, Size 10

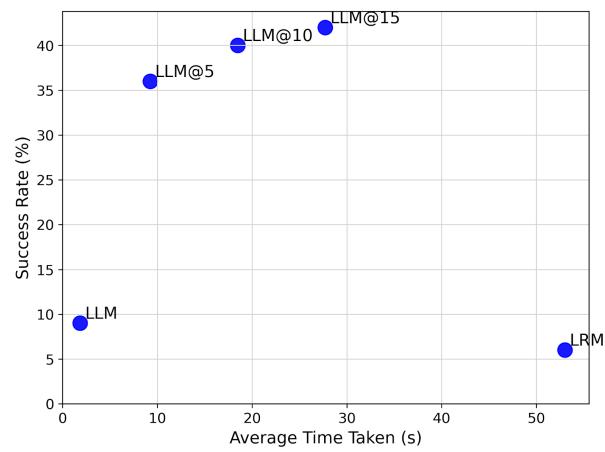


Figure 34: RQ1 - Model Combination 3, Size 25

## Model Combination 4: Llama 3.1 vs. DeepSeek R1 8B

**Description** This final set of figures (Figures 35 through 39) contrasts the performance of Llama 3.1, a powerful LLM, in a feedback loop against the DeepSeek R1 8B LRM. The results reinforce the core finding: the iterative feedback mechanism is critical for high performance. Even though Llama 3.1 requires slightly more time, it achieves a high success rate that is vastly superior to the standalone LRM, especially as graph size increases. The DeepSeek R1 8B LRM's performance degrades significantly on larger problems, while the iterative Llama 3.1 maintains strong performance, proving the general applicability of the feedback-driven method with different state-of-the-art LLMs.

**Summary Finding** The performance benefits of the iterative feedback architecture are model-agnostic, enabling another powerful LLM like Llama 3.1 to significantly outperform a standalone LRM on complex reasoning tasks.

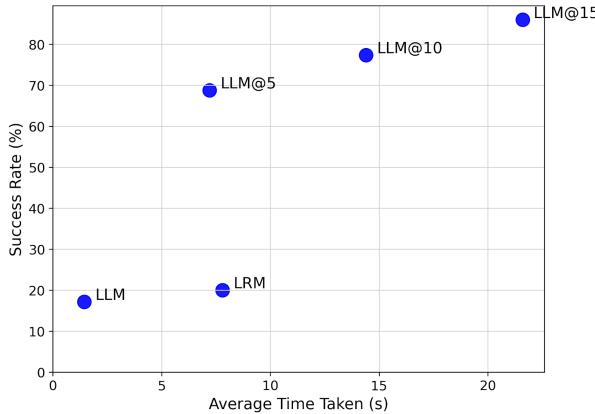


Figure 35: RQ1 - Model Combination 4, Size 5

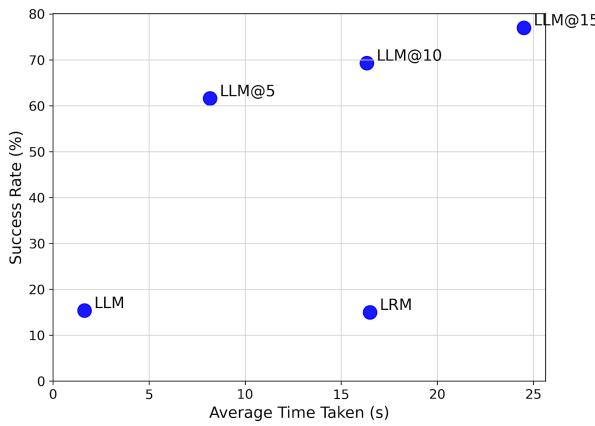


Figure 36: RQ1 - Model Combination 4, Size 10

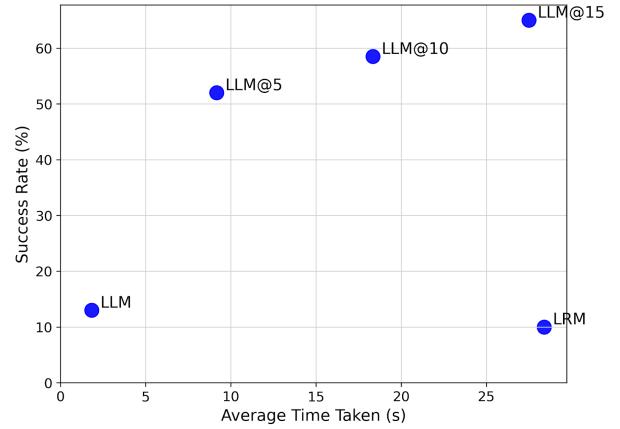


Figure 37: RQ1 - Model Combination 4, Size 15

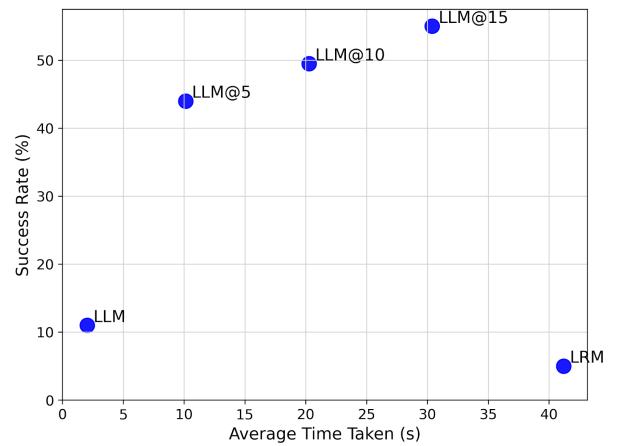


Figure 38: RQ1 - Model Combination 4, Size 20

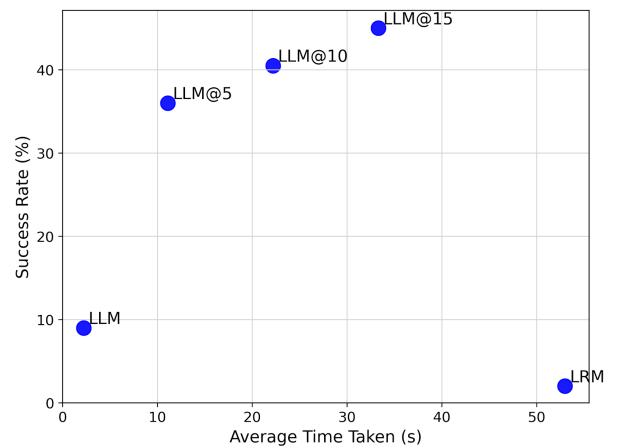


Figure 39: RQ1 - Model Combination 4, Size 25

## RQ2: How do the type of feedback and the format of episodic memory affect a feedback-driven LLM?

The following figures evaluate the impact of different feedback and memory strategies on LLM performance. The configurations are Multi-Line Feedback (MLF) and Single-Line Feedback (SLF), each paired with Minimal Episodic Memory (MEM) or Extended Episodic Memory (EEM).

### LLM: Granite 3.3B (without thinking)

**Description** This set of figures (Figures 40 through 43) evaluates the impact of different feedback and memory strategies on the performance of the Granite 3.3B LLM. A clear and consistent trend emerges: configurations using Multi-Line Feedback (MLF) consistently outperform those using Single-Line Feedback (SLF) in success rate, often in less time. Within both groups, Minimal Episodic Memory (MEM) generally provides a slight edge over Extended Episodic Memory (EEM). The most effective configuration across all problem sizes is MLF+MEM.

**Summary Finding** For the Granite 3.3B model, detailed Multi-Line Feedback combined with a concise Minimal Episodic Memory (MLF+MEM) provides the optimal balance of high success rate and low computational time.

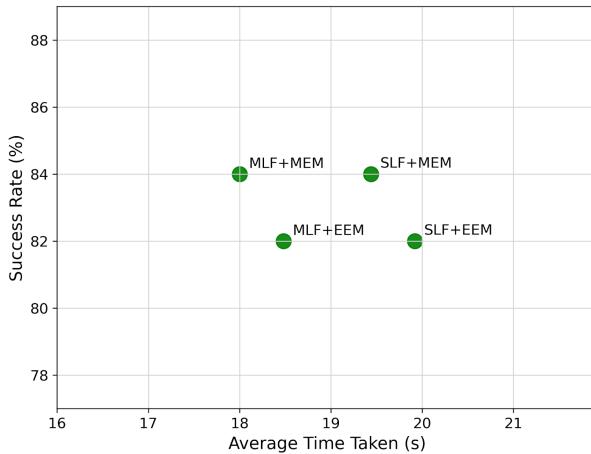


Figure 40: RQ2 - LLM: Granite 3.3B, Size 5

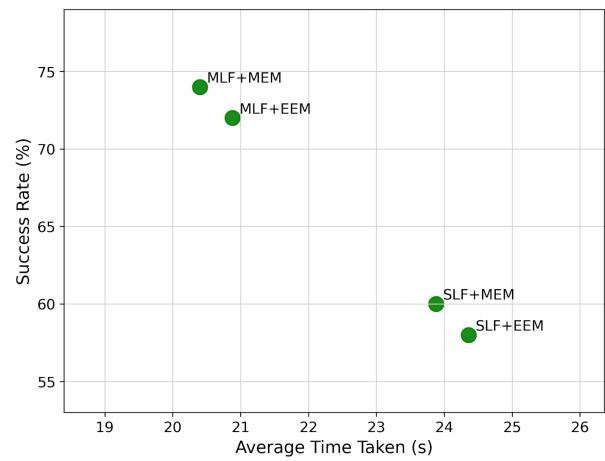


Figure 41: RQ2 - LLM: Granite 3.3B, Size 10

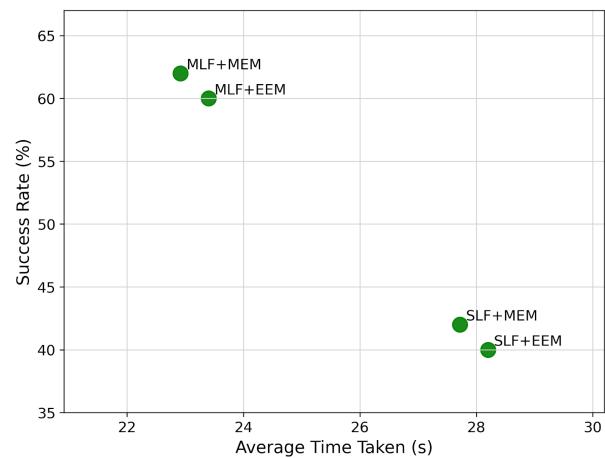


Figure 42: RQ2 - LLM: Granite 3.3B, Size 15

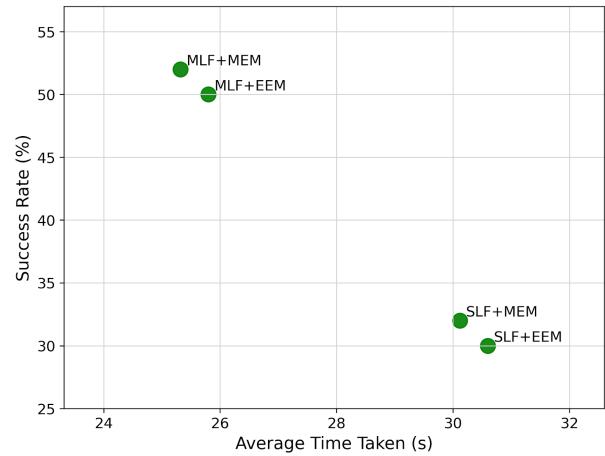


Figure 43: RQ2 - LLM: Granite 3.3B, Size 20

## LLM: Llama 3.1

**Description** This series of plots (Figures 44 through 48) shows the performance of the four metacognitive configurations for the Llama 3.1 LLM. The results demonstrate that the relative effectiveness of the different feedback and memory strategies is consistent across different LLMs. As with the Granite model, configurations with Multi-Line Feedback (MLF) achieve higher success rates than their Single-Line Feedback (SLF) counterparts. The MLF+MEM combination again emerges as the most efficient.

**Summary Finding** The principle that detailed, multi-line feedback with minimal episodic memory (MLF+MEM) provides the best performance holds true for the Llama 3.1 model, indicating that the observed trends for feedback and memory types are robust across different underlying LLMs.

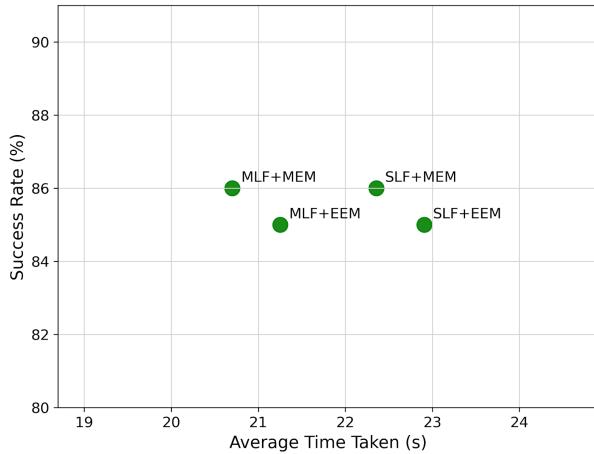


Figure 44: RQ2 - LLM: Llama 3.1, Size 5

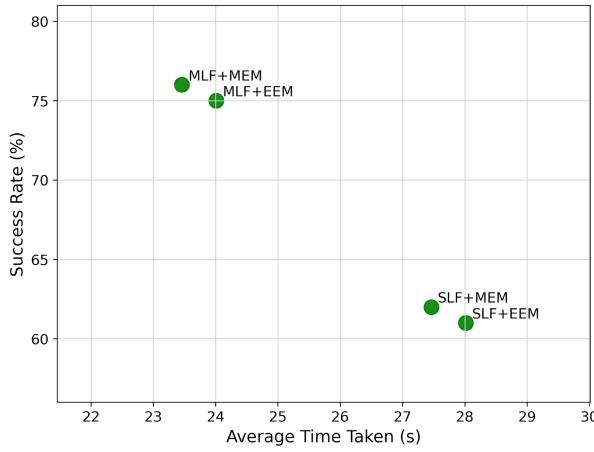


Figure 45: RQ2 - LLM: Llama 3.1, Size 10

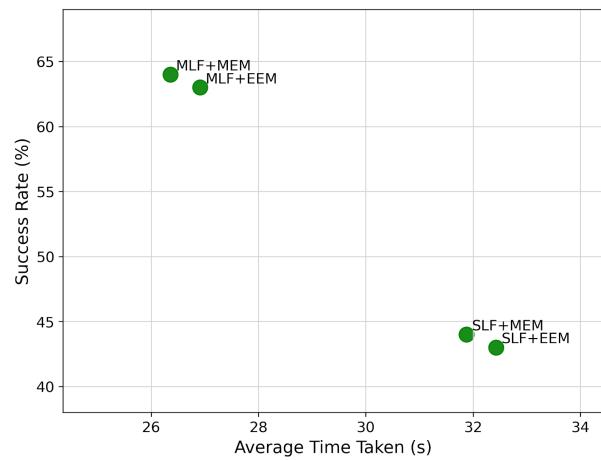


Figure 46: RQ2 - LLM: Llama 3.1, Size 15

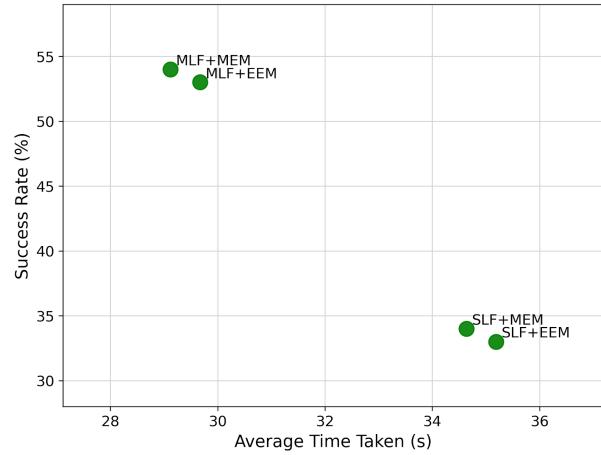


Figure 47: RQ2 - LLM: Llama 3.1, Size 20

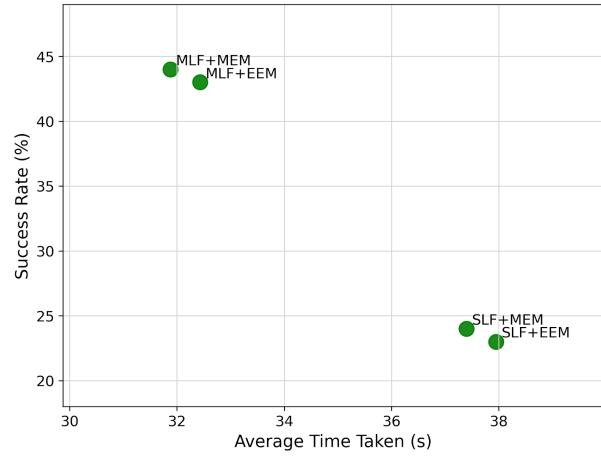


Figure 48: RQ2 - LLM: Llama 3.1, Size 25

### RQ3: Can the information gathered by SOFAI-LM, when used iteratively with an LLM, enhance the performance of an LRM?

The following figures analyze the effectiveness of three different LRM prompting strategies within the SOFAI-LM framework: Problem-Only (PO), Best Attempt (BA), and Full History (FH).

#### Pipeline: Granite 3.3B → DeepSeek R1 8B

**Description** This series of plots (Figures 49 through 52) analyzes the LRM prompting strategies where Granite 3.3B is the LLM and DeepSeek R1 8B is the LRM. The ‘Problem-Only (PO)’ strategy consistently delivers the highest success rate in the shortest amount of time. Providing the LRM with the ‘Best Attempt (BA)’ or the ‘Full History (FH)’ leads to a degradation in performance, suggesting that for a globally constrained problem like graph coloring, the incorrect partial solutions act as noise.

**Summary Finding** For graph coloring tasks, providing the LRM with only the original problem instance (PO) is the most effective strategy, as additional context from the LLM’s attempts degrades both success rate and efficiency.

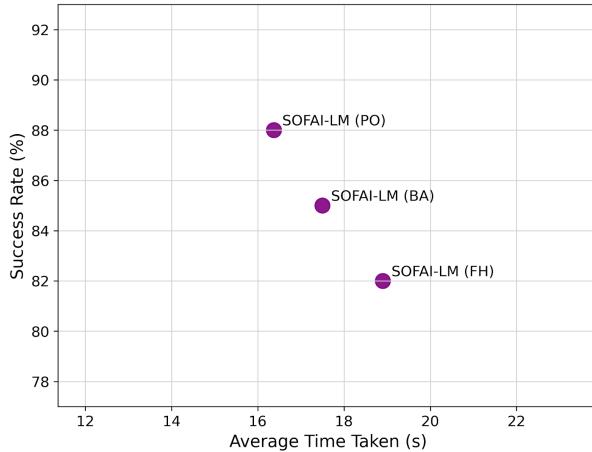


Figure 49: RQ3 - Pipeline: Granite 3.3B → DeepSeek R1 8B, Size 5

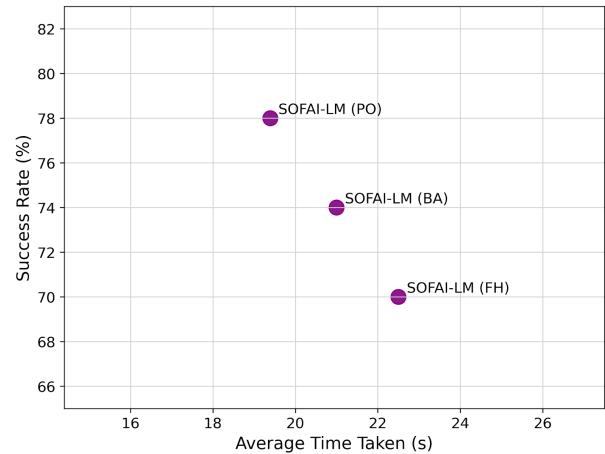


Figure 50: RQ3 - Pipeline: Granite 3.3B → DeepSeek R1 8B, Size 10

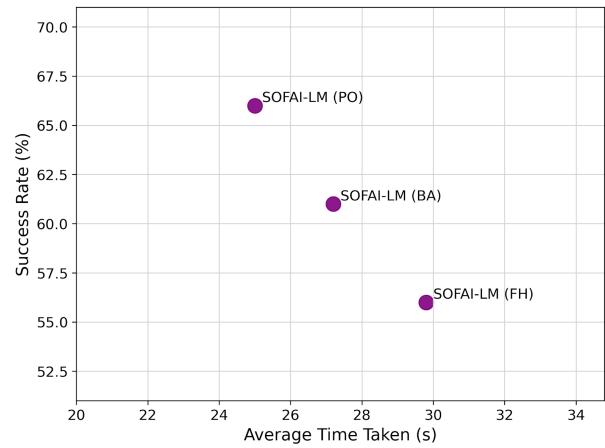


Figure 51: RQ3 - Pipeline: Granite 3.3B → DeepSeek R1 8B, Size 15

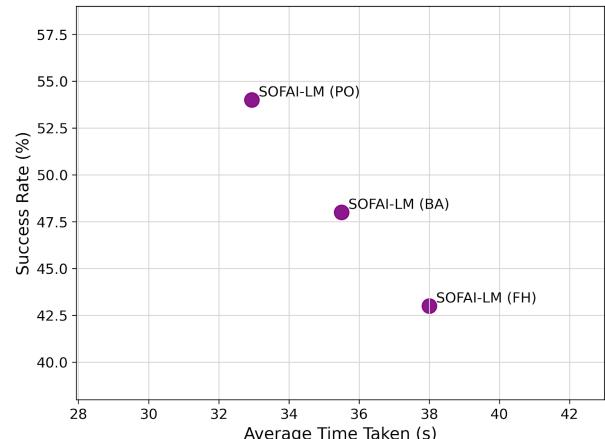


Figure 52: RQ3 - Pipeline: Granite 3.3B → DeepSeek R1 8B, Size 20

### Pipeline: Granite 3.3B → Granite 3.3B

**Description** This set of plots (Figures 53 through 57) examines the three prompting strategies when using Granite 3.3B for both the LLM and LRM roles. The results confirm the trend observed previously: the ‘Problem-Only’ (PO) strategy is superior, achieving the highest success rate. This demonstrates that even when the LLM and LRM are the same underlying model, providing historical context of failed attempts hinders the model’s ability to find a globally consistent solution.

**Summary Finding** When using a homogeneous model setup (Granite to Granite), a “clean slate” ‘Problem-Only’ prompt for the LRM is still the most effective strategy, reinforcing that context from failed attempts is detrimental for this class of problem.

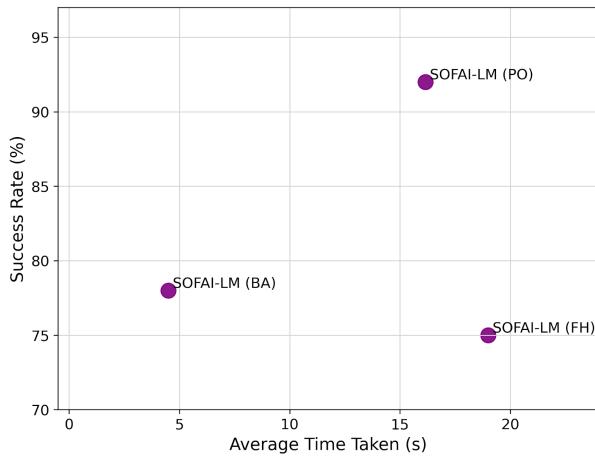


Figure 53: RQ3 - Pipeline: Granite 3.3B → Granite 3.3B, Size 5

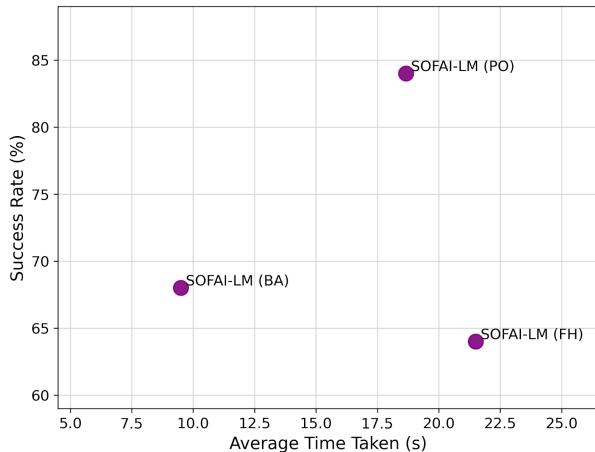


Figure 54: RQ3 - Pipeline: Granite 3.3B → Granite 3.3B, Size 10

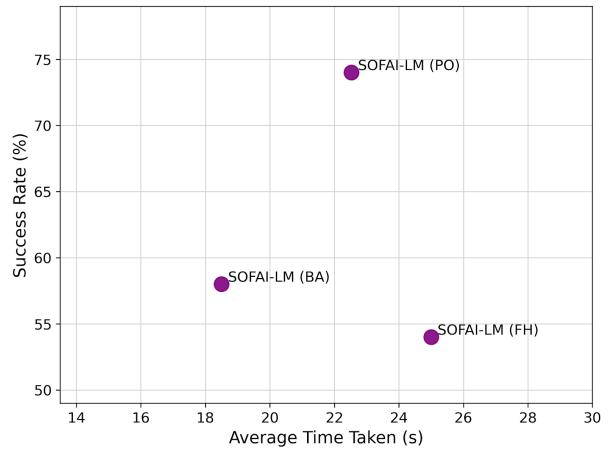


Figure 55: RQ3 - Pipeline: Granite 3.3B → Granite 3.3B, Size 15

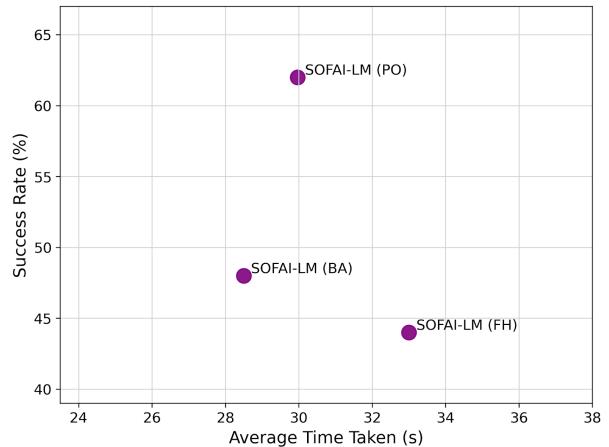


Figure 56: RQ3 - Pipeline: Granite 3.3B → Granite 3.3B, Size 20

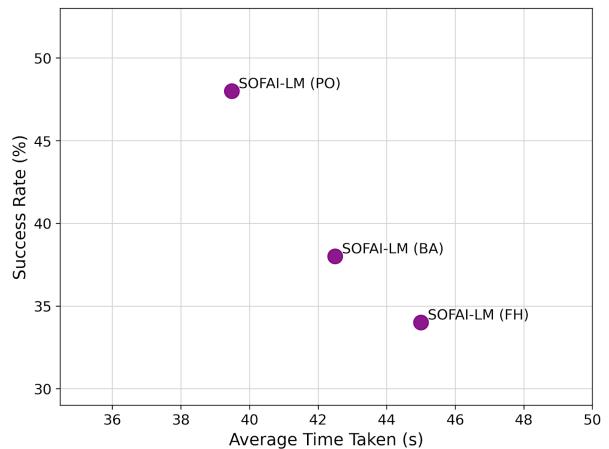


Figure 57: RQ3 - Pipeline: Granite 3.3B → Granite 3.3B, Size 25

### Pipeline: Granite 3.3B → Qwen 2.5 Pro

**Description** This series of plots (Figures 58 through 62) shows the performance of the three prompting strategies when pairing the Granite 3.3B LLM with the Qwen 2.5 Pro LRM. The results are consistent with the previous findings. The ‘Problem-Only (PO)’ strategy provides the best balance of success rate and time efficiency. Passing the ‘Best Attempt (BA)’ or ‘Full History (FH)’ to the LRM results in a clear decline in performance.

**Summary Finding** The superiority of the ‘Problem-Only’ prompting strategy for graph coloring is robust across different S2 models, as the Granite → Qwen pipeline shows the same performance degradation with added context as other model combinations.

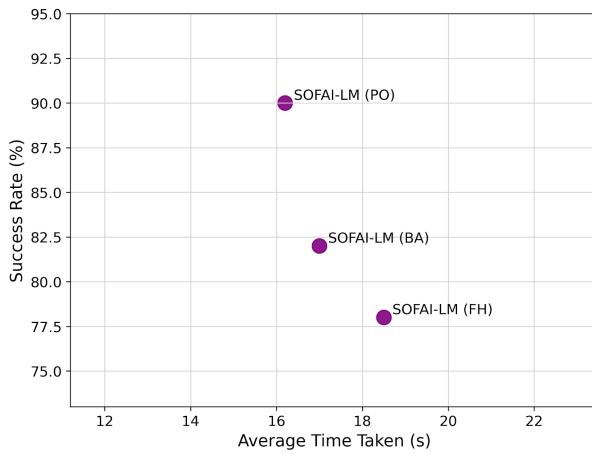


Figure 58: RQ3 - Pipeline: Granite 3.3B → Qwen 2.5 Pro, Size 5

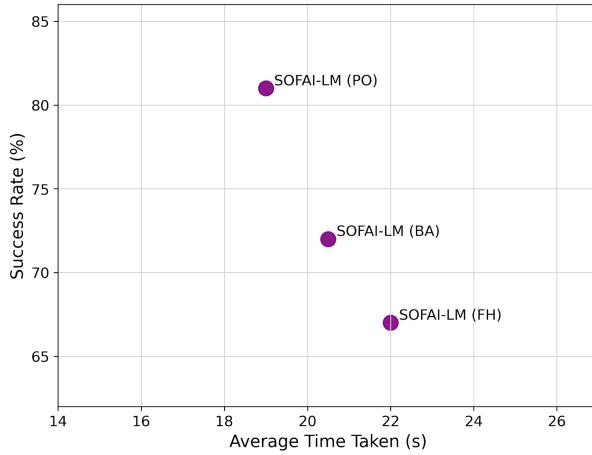


Figure 59: RQ3 - Pipeline: Granite 3.3B → Qwen 2.5 Pro, Size 10

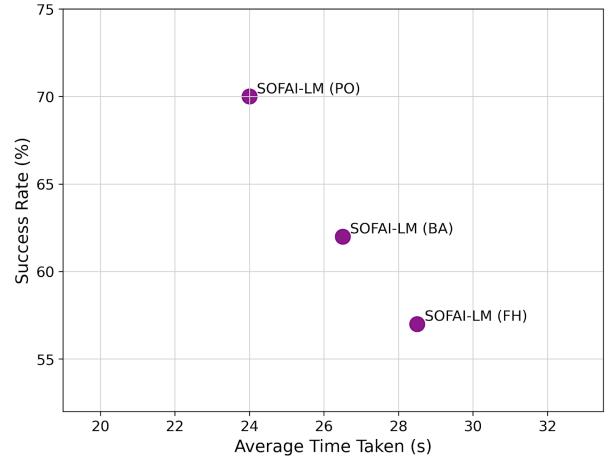


Figure 60: RQ3 - Pipeline: Granite 3.3B → Qwen 2.5 Pro, Size 15

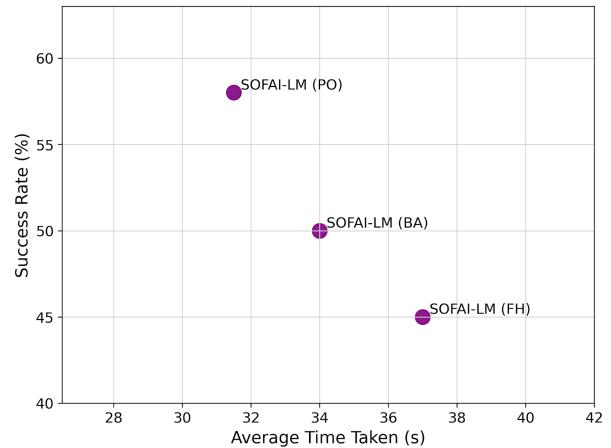


Figure 61: RQ3 - Pipeline: Granite 3.3B → Qwen 2.5 Pro, Size 20

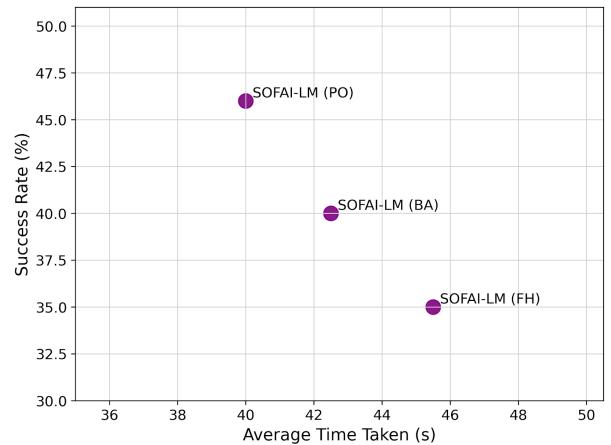


Figure 62: RQ3 - Pipeline: Granite 3.3B → Qwen 2.5 Pro, Size 25

### Pipeline: Llama 3.1 → DeepSeek R1 8B

**Description** This final set of plots for RQ3 (Figures 63 through 67) shows the performance of the prompting strategies when using Llama 3.1 as the initial LLM. The results are unequivocally consistent with all previous findings for the graph coloring domain. The ‘Problem-Only (PO)’ strategy provides the best results, achieving the highest success rates for the most reasonable time cost.

**Summary Finding** The finding that providing historical context harms LRM performance on graph coloring is robust and holds true even when a different powerful LLM, Llama 3.1, is used, confirming the generality of this key insight for the SOFAI-LM architecture in this domain.

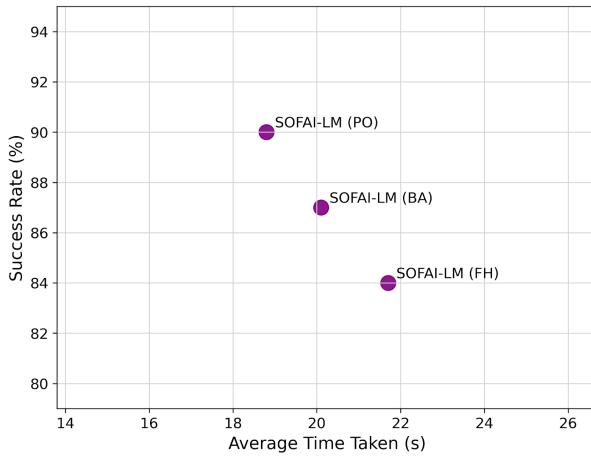


Figure 63: RQ3 - Pipeline: Llama 3.1 → DeepSeek R1 8B, Size 5

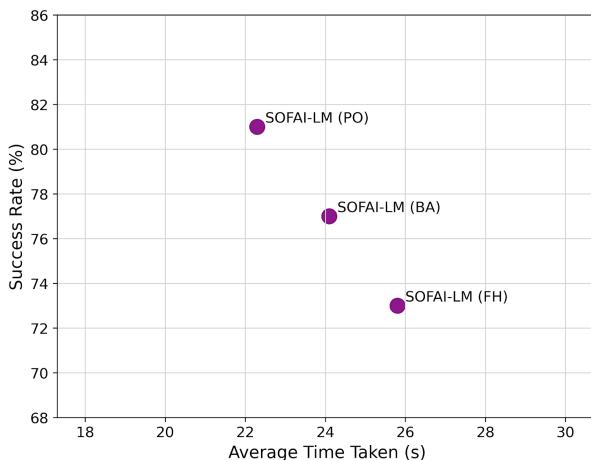


Figure 64: RQ3 - Pipeline: Llama 3.1 → DeepSeek R1 8B, Size 10

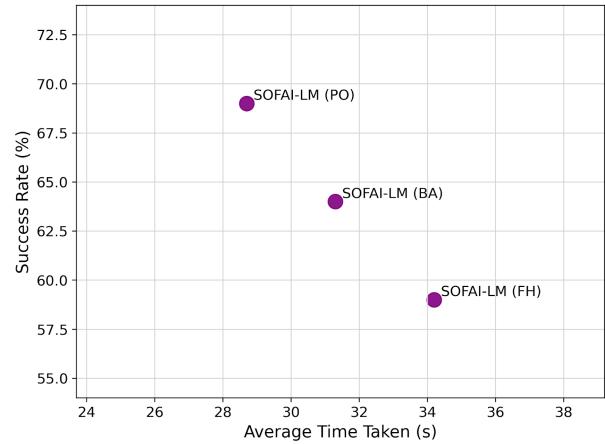


Figure 65: RQ3 - Pipeline: Llama 3.1 → DeepSeek R1 8B, Size 15

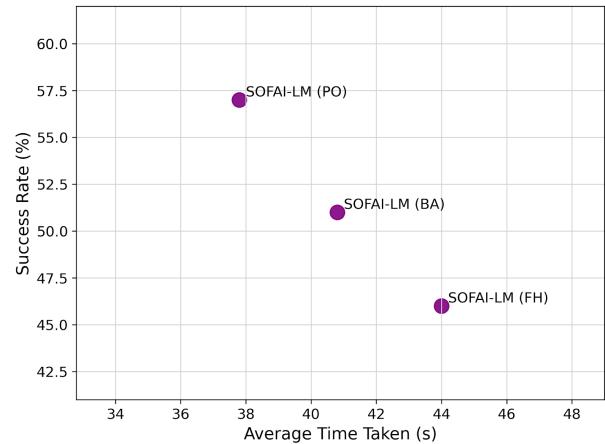


Figure 66: RQ3 - Pipeline: Llama 3.1 → DeepSeek R1 8B, Size 20

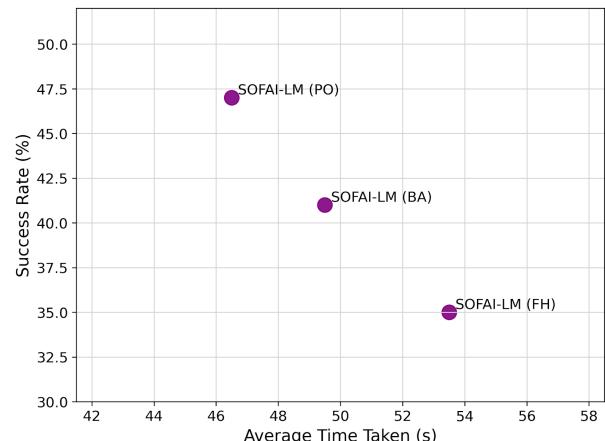


Figure 67: RQ3 - Pipeline: Llama 3.1 → DeepSeek R1 8B, Size 25

#### RQ4: Does SOFAI-LM perform better than its LRM counterpart?

The following figures provide a direct comparison between the complete, optimized SOFAI-LM architecture and using the LRM as a standalone solver.

##### Combination 1: SOFAI-LM (Granite → DeepSeek) vs. LRM (DeepSeek)

**Description** This series of plots (Figures 68 through 71) provides a direct comparison between the integrated SOFAI-LM architecture and a standalone LRM (DeepSeek R1 8B). The results decisively show the superiority of the SOFAI-LM approach. Across every graph size, SOFAI-LM achieves a dramatically higher success rate. For larger problems (size 20 and 25), SOFAI-LM is not only substantially more accurate but also faster than the standalone LRM.

**Summary Finding** The SOFAI-LM architecture significantly outperforms a standalone LRM, delivering vastly superior success rates and better time efficiency, especially as graph size increases.

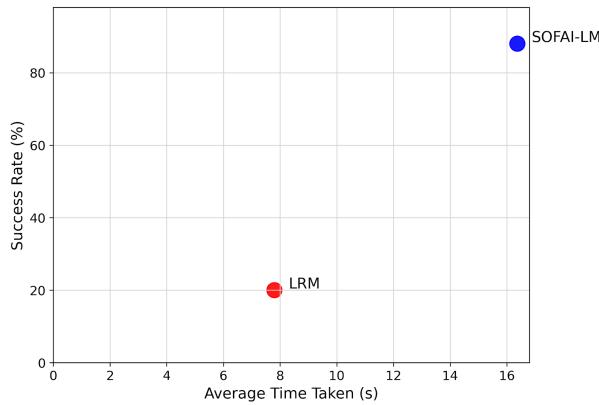


Figure 68: RQ4 - Combination 1: SOFAI-LM (Granite → DeepSeek) vs. LRM (DeepSeek), Size 5

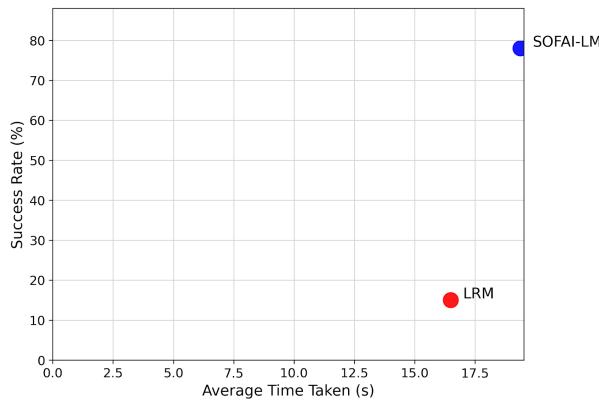


Figure 69: RQ4 - Combination 1: SOFAI-LM (Granite → DeepSeek) vs. LRM (DeepSeek), Size 10

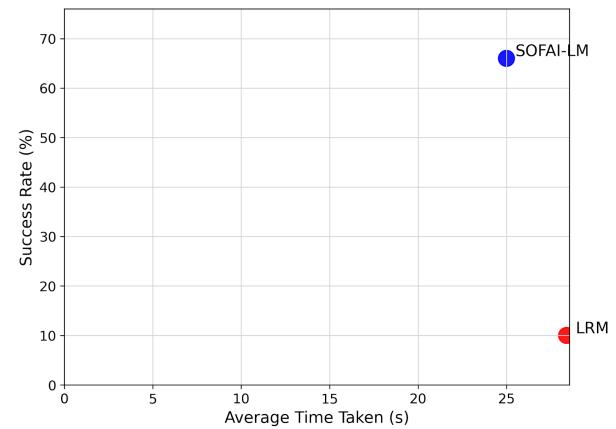


Figure 70: RQ4 - Combination 1: SOFAI-LM (Granite → DeepSeek) vs. LRM (DeepSeek), Size 15

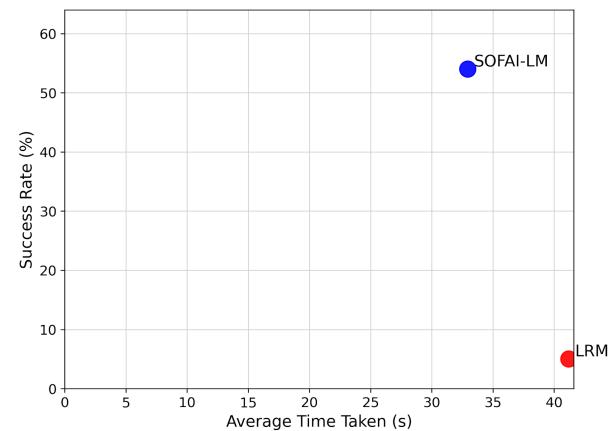


Figure 71: RQ4 - Combination 1: SOFAI-LM (Granite → DeepSeek) vs. LRM (DeepSeek), Size 20

## Combination 2: SOFAI-LM (Granite → Granite) vs. LRM (Granite)

**Description** This set of figures (Figures 72 through 76) compares the performance of the full SOFAI-LM pipeline against its LRM component, using Granite 3.3B for both roles. The results reinforce the value of the SOFAI-LM architecture even in a homogeneous model setting. The complete SOFAI-LM system consistently achieves a much higher success rate than using the Granite model as a standalone LRM. This demonstrates that the architectural design itself—using an iterative feedback loop with selective fallback—is the key driver of performance.

**Summary Finding** The SOFAI-LM architecture provides a substantial performance benefit over a standalone LRM, even when the same base model is used for both the fast and slow solver roles.

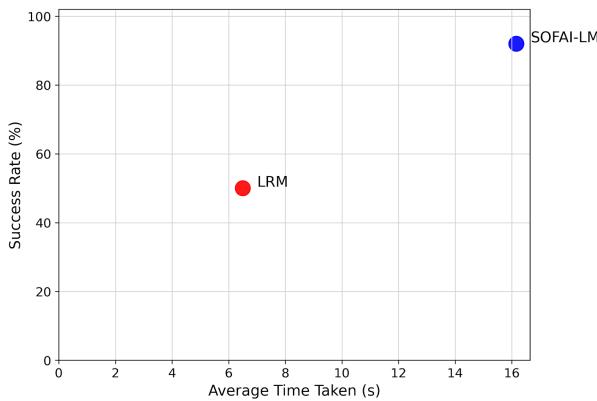


Figure 72: RQ4 - Combination 2: SOFAI-LM (Granite → Granite) vs. LRM (Granite), Size 5

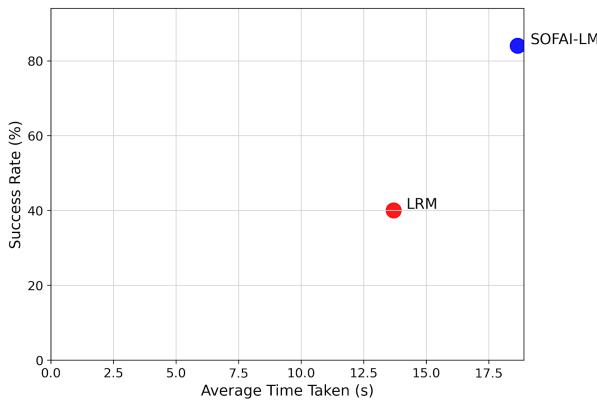


Figure 73: RQ4 - Combination 2: SOFAI-LM (Granite → Granite) vs. LRM (Granite), Size 10

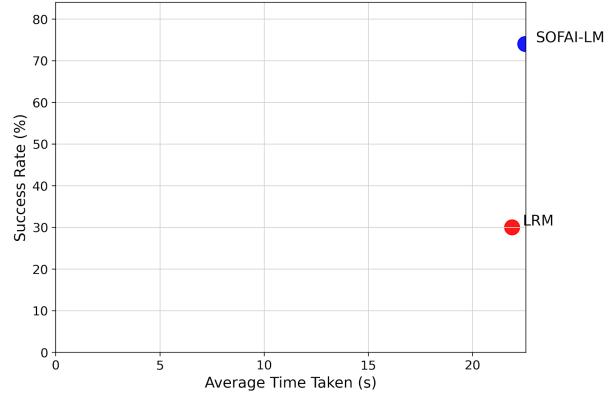


Figure 74: RQ4 - Combination 2: SOFAI-LM (Granite → Granite) vs. LRM (Granite), Size 15

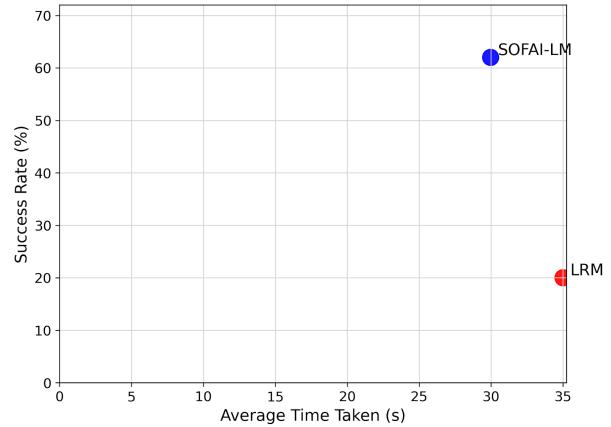


Figure 75: RQ4 - Combination 2: SOFAI-LM (Granite → Granite) vs. LRM (Granite), Size 20

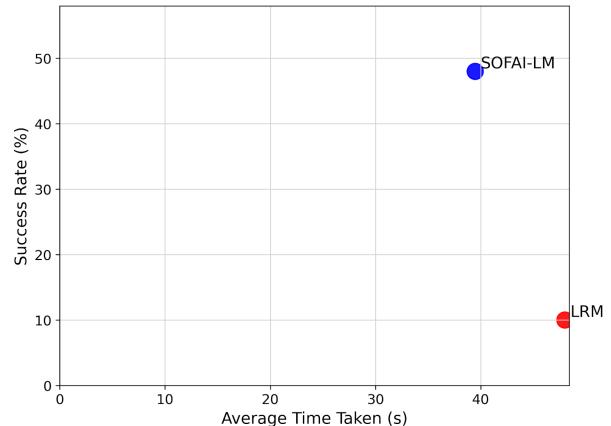


Figure 76: RQ4 - Combination 2: SOFAI-LM (Granite → Granite) vs. LRM (Granite), Size 25

### Combination 3: SOFAI-LM (Granite → Qwen) vs. LRM (Qwen)

**Description** This set of figures (Figures 77 through 81) compares the full SOFAI-LM pipeline, using Granite 3.3B as the LLM and Qwen 2.5 Pro as the LRM, against the standalone Qwen LRM. The results robustly confirm the paper's central thesis. The integrated SOFAI-LM system consistently and significantly outperforms the standalone LRM on success rate. As the problem complexity scales up, SOFAI-LM also becomes more time-efficient.

**Summary Finding** The SOFAI-LM architecture provides a decisive advantage in both accuracy and efficiency over a powerful standalone LRM like Qwen 2.5 Pro, confirming the generalizability of the approach.

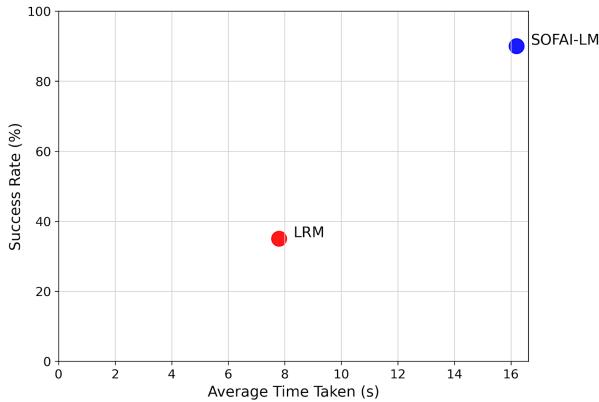


Figure 77: RQ4 - Combination 3: SOFAI-LM (Granite → Qwen) vs. LRM (Qwen), Size 5

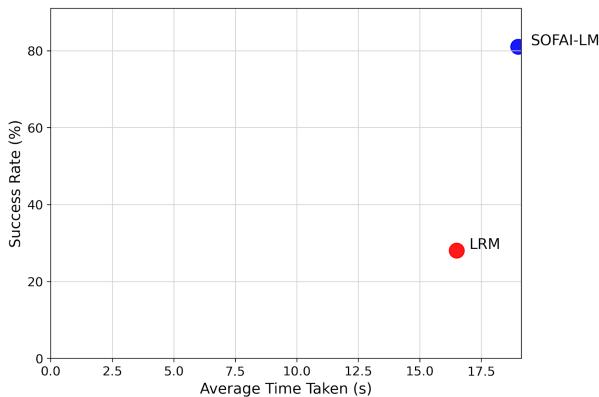


Figure 78: RQ4 - Combination 3: SOFAI-LM (Granite → Qwen) vs. LRM (Qwen), Size 10

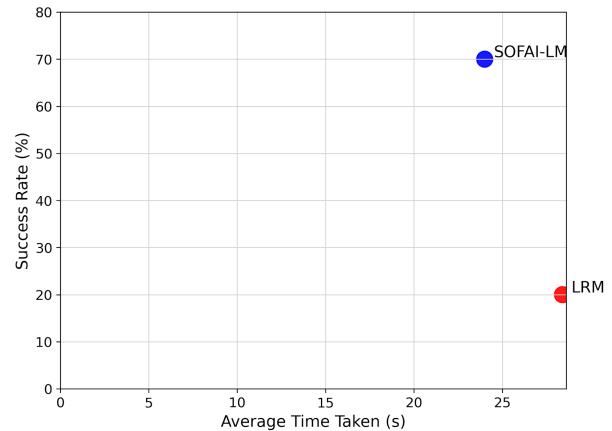


Figure 79: RQ4 - Combination 3: SOFAI-LM (Granite → Qwen) vs. LRM (Qwen), Size 15

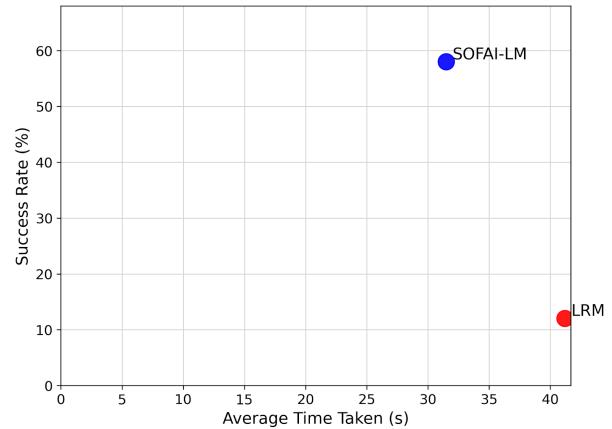


Figure 80: RQ4 - Combination 3: SOFAI-LM (Granite → Qwen) vs. LRM (Qwen), Size 20

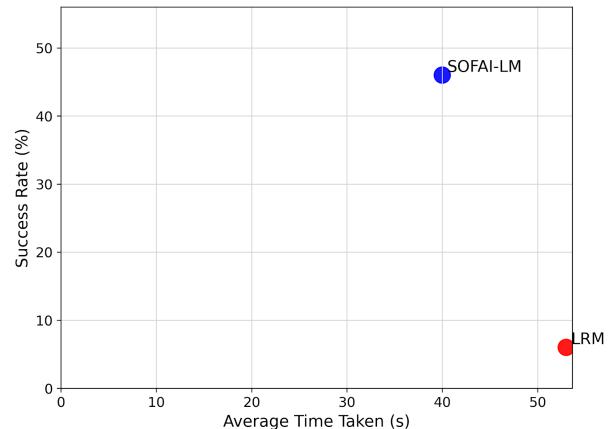


Figure 81: RQ4 - Combination 3: SOFAI-LM (Granite → Qwen) vs. LRM (Qwen), Size 25

#### Combination 4: SOFAI-LM (Llama → DeepSeek) vs. LRM (DeepSeek)

**Description** This final set of plots for RQ4 (Figures 82 through 86) uses Llama 3.1 as the S1 solver and DeepSeek R1 8B as the S2 solver. The results compellingly demonstrate the power of the SOFAI-LM architecture. The integrated system achieves a vastly superior success rate compared to the standalone DeepSeek R1 8B LRM across all problem complexities. This confirms that the architectural benefits are not dependent on a specific S1 model.

**Summary Finding** The SOFAI-LM framework effectively leverages the strengths of a powerful LLM like Llama 3.1, leading to a system that is overwhelmingly more accurate and ultimately more efficient than its standalone LRM component.

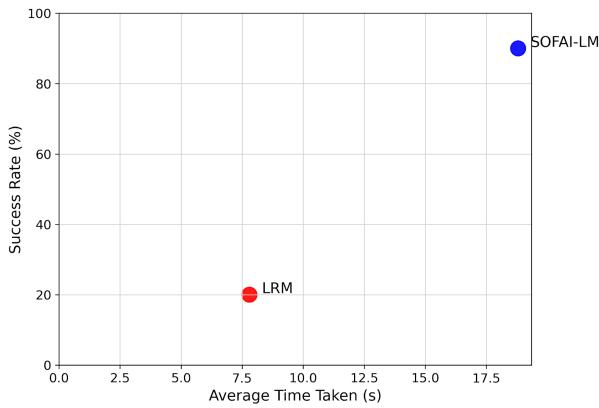


Figure 82: RQ4 - Combination 4: SOFAI-LM (Llama → DeepSeek) vs. LRM (DeepSeek), Size 5

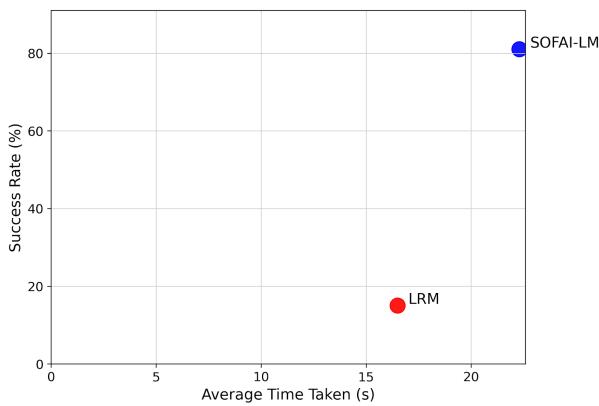


Figure 83: RQ4 - Combination 4: SOFAI-LM (Llama → DeepSeek) vs. LRM (DeepSeek), Size 10

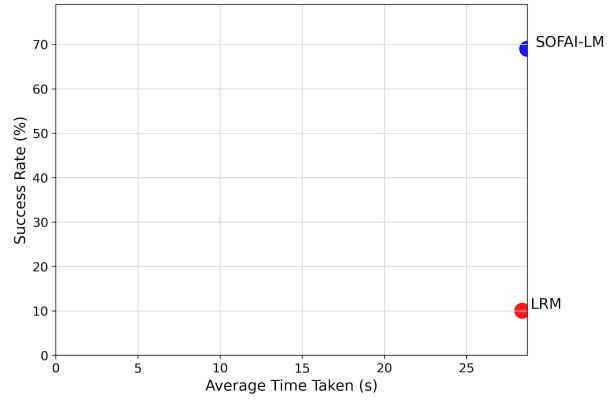


Figure 84: RQ4 - Combination 4: SOFAI-LM (Llama → DeepSeek) vs. LRM (DeepSeek), Size 15

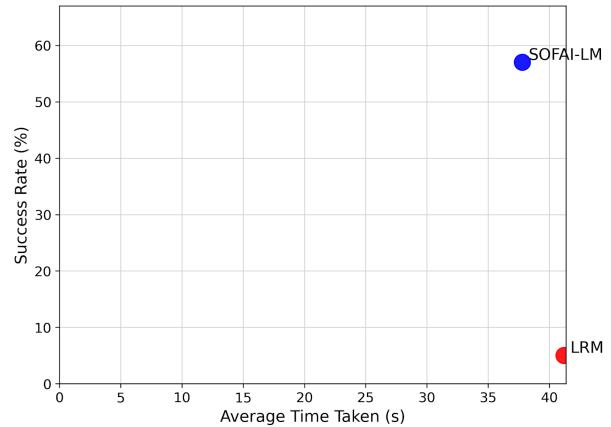


Figure 85: RQ4 - Combination 4: SOFAI-LM (Llama → DeepSeek) vs. LRM (DeepSeek), Size 20

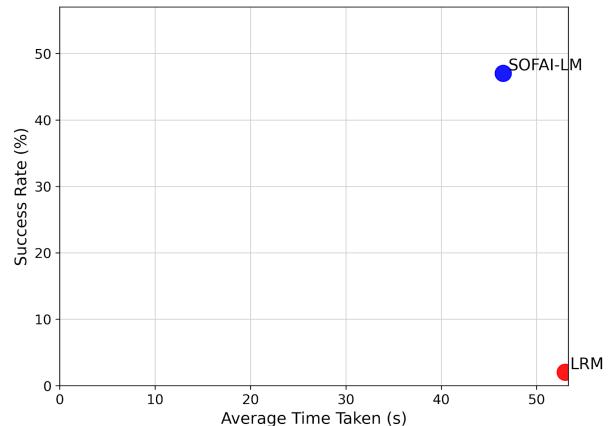


Figure 86: RQ4 - Combination 4: SOFAI-LM (Llama → DeepSeek) vs. LRM (DeepSeek), Size 25

## Experimental Results : Code Debugging

We validate our core claims across several combinations of Large Language Models (LLMs) and Large Reasoning Models (LRMs) on the code debugging domain (DebugBench), demonstrating the robustness and generalizability of the SOFAI-LM architecture.

### RQ1: Can a feedback-driven LLM outperform an LRM?

The main paper demonstrated that a feedback-driven LLM (Granite 3.3 8B) could outperform a standalone LRM (DeepSeek R1 8B). Here, we confirm this finding across three additional model combinations.

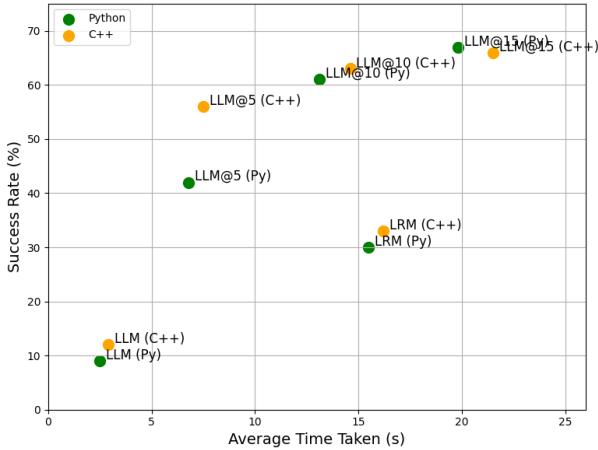


Figure 87: Comparison of success rate versus average time for a feedback-driven LLM and a standalone LRM on the DebugBench dataset (Python and C++). The configurations shown are a single-pass LLM, the LLM with 5, 10, and 15 feedback iterations (LLM@5, LLM@10, LLM@15), and the LRM. For this figure, the LLM (S1 solver) is Granite 3.3 8B (without thinking) and the LRM (S2 solver) is Granite 3.3 8B (with thinking).

**Combination 1: LLM = Granite 3.3 8B (without thinking), LRM = Granite 3.3 8B (with thinking).** Figure 87 illustrates the performance trade-off when both the S1 and S2 solvers are from the same model family. The results clearly show that the feedback-driven LLM with 5 iterations (LLM@5) surpasses the performance of its more deliberative “with thinking” counterpart used as a standalone LRM. Increasing the number of iterations to 10 and 15 further widens this performance gap, establishing that the iterative feedback mechanism is more effective than relying on the LRM’s built-in reasoning capabilities alone.

**Combination 2: LLM = Granite 3.3 8B (without thinking), LRM = Qwen 3 8B.** In Figure 88, we pair the Granite 3.3 8B LLM with the Qwen 3 8B LRM. The trend remains consistent: the iterative LLM configurations (LLM@5, LLM@10, and LLM@15) form a clear Pareto frontier, significantly outperforming the standalone Qwen 3 8B LRM in both success rate and average time. This result

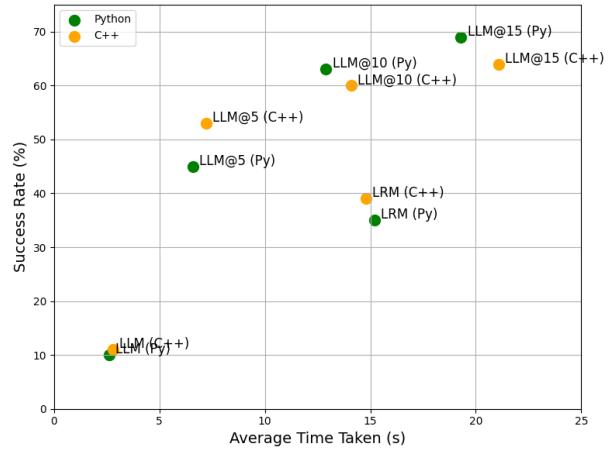


Figure 88: Comparison of success rate versus average time for a feedback-driven LLM and a standalone LRM on the DebugBench dataset (Python and C++ subsets). The configurations shown are a single-pass LLM, the LLM with 5, 10, and 15 feedback iterations (LLM@5, LLM@10, LLM@15), and the LRM. For this figure, the LLM (S1 solver) is Granite 3.3 8B (without thinking) and the LRM (S2 solver) is Qwen 3 8B.

underscores that the advantage of the SOFAI-LM feedback loop is not limited to a specific LRM but holds against other capable reasoning models.

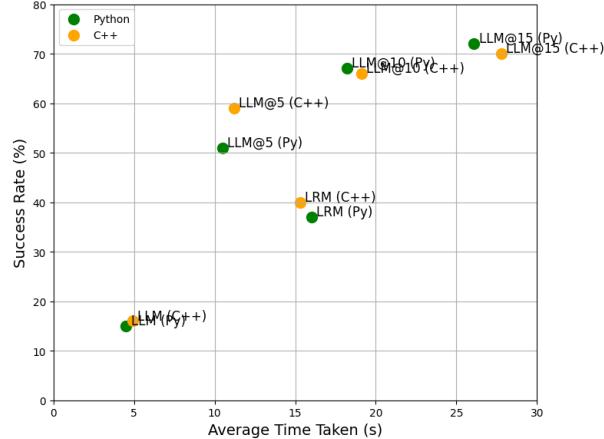


Figure 89: Comparison of success rate versus average time for a feedback-driven LLM and a standalone LRM on the DebugBench dataset (Python and C++ subsets). The configurations shown are a single-pass LLM, the LLM with 5, 10, and 15 feedback iterations (LLM@5, LLM@10, LLM@15), and the LRM. For this figure, the LLM (S1 solver) is Llama 3.1 8B and the LRM (S2 solver) is DeepSeek R1 8B.

**Combination 3: LLM = Llama 3.1 8B, LRM = DeepSeek R1 8B.** Figure 89 validates our findings with a different base LLM, Llama 3.1 8B, against the powerful DeepSeek R1 8B LRM. Despite Llama 3.1 being slightly slower per it-

eration than Granite, the feedback-driven approach remains vastly superior. The LLM@5 configuration again proves more effective and efficient than the standalone LRM, confirming that the principles of our architecture are model-agnostic and apply to different state-of-the-art LLMs.

### RQ3: Can the information gathered by SOFAI-LM, when used iteratively with an LLM, enhance the performance of an LRM?

This section explores the impact of different LRM prompting strategies: Problem-Only (PO), Best Attempt (BA), and Full History (FH), within the SOFAI-LM pipeline for the code debugging domain.

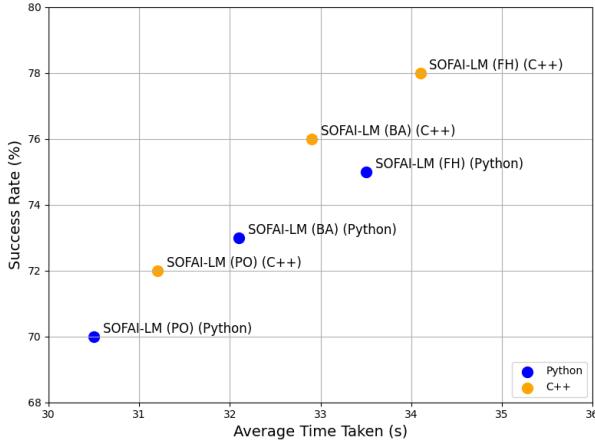


Figure 90: Comparison of SOFAI-LM performance using three different LRM prompting strategies: Problem-Only (PO), Best Attempt (BA), and Full History (FH). The plot shows success rate versus average time on the DebugBench dataset. For this figure, the LLM (S1 solver) is Granite 3.3 8B (without thinking) and the LRM (S2 solver) is Granite 3.3 8B (with thinking).

**Combination 1: LLM = Granite 3.3 8B (without thinking), LRM = Granite 3.3 8B (with thinking).** As shown in Figure 90, providing the LRM with more context from the LLM’s attempts is beneficial. For both Python and C++, the success rate increases progressively from PO to BA and again to FH. This confirms that for code debugging, where fixes are often localized, the history of failed attempts provides valuable negative constraints that help the LRM converge on a correct solution.

**Combination 2: LLM = Granite 3.3 8B (without thinking), LRM = Qwen 3 8B.** The results in Figure 91 reinforce this finding with the Qwen 3 8B LRM. The Full History (FH) strategy yields the highest success rate. Notably, for the C++ subset, providing more context via BA and FH also slightly reduces the average inference time, suggesting the information helps the LRM find a solution more efficiently, not just more accurately.

**Combination 3: LLM = Llama 3.1 8B, LRM = DeepSeek R1 8B.** Figure 92 shows a distinct stepwise improvement in success rate as more context is provided to the DeepSeek R1 8B LRM. The hierarchy of  $FH > BA > PO$  is clear, demonstrating that even when using a powerful LLM and LRM, enriching the LRM’s prompt with the full interaction history is the optimal strategy for maximizing performance on code debugging tasks.

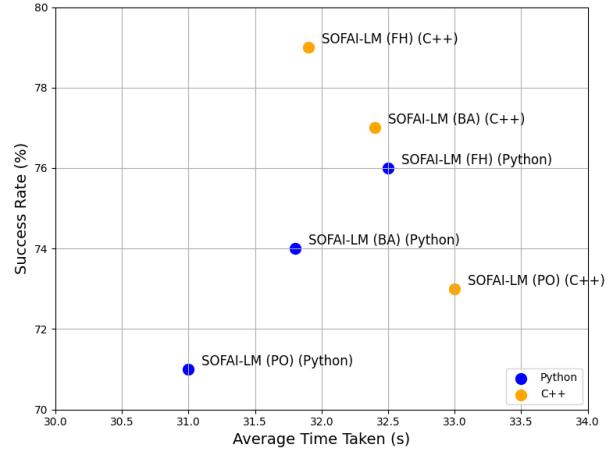


Figure 91: Comparison of SOFAI-LM performance using three different LRM prompting strategies: Problem-Only (PO), Best Attempt (BA), and Full History (FH). The plot shows success rate versus average time on the DebugBench dataset. For this figure, the LLM (S1 solver) is Granite 3.3 8B (without thinking) and the LRM (S2 solver) is Qwen 3 8B.

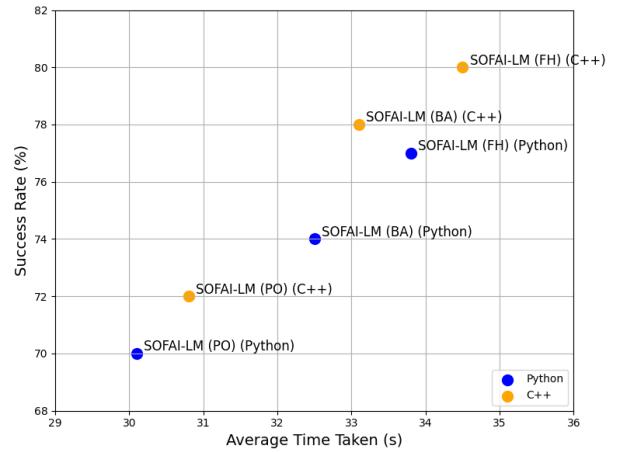


Figure 92: Comparison of SOFAI-LM performance using three different LRM prompting strategies: Problem-Only (PO), Best Attempt (BA), and Full History (FH). The plot shows success rate versus average time on the DebugBench dataset. For this figure, the LLM (S1 solver) is Llama 3.1 8B and the LRM (S2 solver) is DeepSeek R1 8B.

## RQ4: Does SOFAI-LM perform better than its LRM counterpart?

We present a direct comparison between the complete SOFAI-LM pipeline (using the best-performing FH strategy) and the standalone LRM for each model combination.

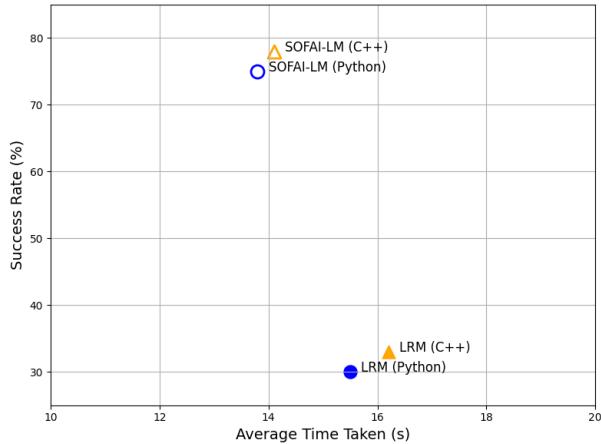


Figure 93: Overall performance comparison between the standalone LRM and the complete SOFAI-LM architecture on the DebugBench dataset. The plot shows success rate versus average time. For this figure, the LLM (S1 solver) is Granite 3.3 8B (without thinking) and the LRM (S2 solver) is Granite 3.3 8B (with thinking).

**Combination 1: LLM = Granite 3.3 8B (without thinking), LRM = Granite 3.3 8B (with thinking).** Figure 93 shows the dramatic overall improvement provided by the SOFAI-LM architecture. The pipeline achieves a success rate more than double that of the standalone LRM while also being faster on average. This highlights the synergistic benefit of the architecture, which is far superior to using either of its components in isolation.

**Combination 2: LLM = Granite 3.3 8B (without thinking), LRM = Qwen 3 8B.** The comparison in Figure 94 confirms the superiority of the SOFAI-LM framework. The integrated system significantly outperforms the standalone Qwen 3 8B LRM, achieving a much higher success rate in less time. This demonstrates the framework’s ability to effectively orchestrate different models to achieve a result better than the sum of its parts.

**Combination 3: LLM = Llama 3.1 8B, LRM = DeepSeek R1 8B.** As seen in Figure 95, the SOFAI-LM architecture’s advantage holds even when using a stronger pair of models. The pipeline, which intelligently leverages the fast Llama 3.1 8B LLM and falls back to the powerful DeepSeek R1 8B LRM only when necessary, again doubles the success rate of the standalone LRM while maintaining a slight edge in time efficiency. This result robustly confirms the value of our metacognitive approach.

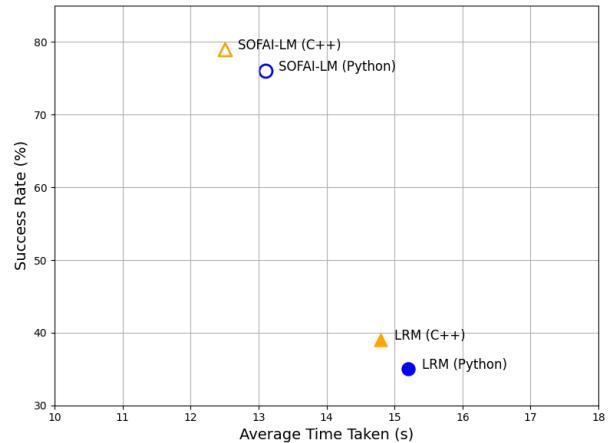


Figure 94: Overall performance comparison between the standalone LRM and the complete SOFAI-LM architecture on the DebugBench dataset. The plot shows success rate versus average time. For this figure, the LLM (S1 solver) is Granite 3.3 8B (without thinking) and the LRM (S2 solver) is Qwen 3 8B.

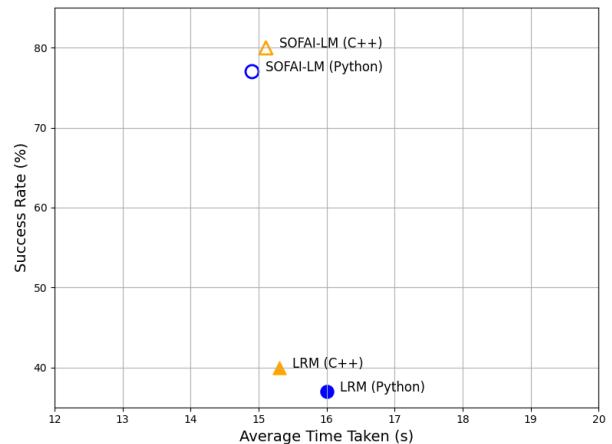


Figure 95: Overall performance comparison between the a standalone LRM and the complete SOFAI-LM architecture on the DebugBench dataset. The plot shows success rate versus average time. For this figure, the LLM (S1 solver) is Llama 3.1 8B and the LRM (S2 solver) is DeepSeek R1 8B.