2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Multi-Stage Hallucination Mitigation and Structured Output Generation via Guided Inference and Model Synergy

Wang Ruan[1], Tengda Qi[2], Jun He[1,3(✉)], Bo Sun[1,3(✉)], and Guomin Zheng[2]

[1] School of Artificial Intelligence, Beijing Normal University, Beijing, China
[2] School of Chinese Language and Literature, Beijing Normal University, Beijing, China
[3] College of Education for the Future, Beijing Normal University, Zhuhai, Guangdong, China
{hejun, tosunbo}@bnu.edu.cn

Abstract. In the pursuit of advanced artificial intelligence capabilities, the challenges posed by Large Language Models (LLMs) cannot be overlooked. LLMs, despite their capacity for multi-step logical reasoning through CoT prompting, often encounter issues such as hallucinations that undermine the accuracy of results and inefficiency in processing. Recognizing the complementary strengths of small Language models, we introduce the MS-HM Synergy (Multi-Stage Hallucination Mitigation Synergy) framework. This novel framework, centered around guided inference and model synergy, comprises three essential stages. Firstly, Guided Inference utilizes LLMs for initial reasoning, tapping into their language understanding. Secondly, Hallucination Detection acts as a safeguard, meticulously identifying and eliminating unreliable outputs. Lastly, Result Standardization ensures the generation of coherent and structured outputs. Methodologically, LLMs are tasked with complex reasoning, while small Language models play a crucial verification role. Empirical results on benchmarks like MMLU, MATH, and LogiQA exhibit substantial performance improvements. The MS-HM Synergy not only effectively mitigates hallucinations for enhanced reliability but also boosts efficiency and flexibility, heralding a new era of leveraging combined model strengths to overcome LLM limitations.

Keywords: Large Language Models (LLMs), Multi-step logical reasoning, Hallucinations, Small Language models, Model synergy.

1 Introduction

In recent years, the development of artificial intelligence (AI) has seen significant advancements, particularly in the realm of complex reasoning tasks. Reliable AI systems capable of multi-step logical reasoning are increasingly important across various domains, including medical diagnosis, education, and decision-making processes in high-stakes environments [4,6,15,32]. Large Language Models (LLMs) have emerged as powerful tools in this context, demonstrating impressive capabilities in generating coherent and contextually relevant responses. Their ability to perform multi-step reasoning through techniques such as Chain-of-Thought (CoT) prompting has been widely

recognized [32]. However, LLMs also face notable limitations, such as the tendency to produce hallucinations—outputs that are plausible but factually incorrect—and their resource-intensive nature, which can hinder scalability and efficiency [13,27,35].

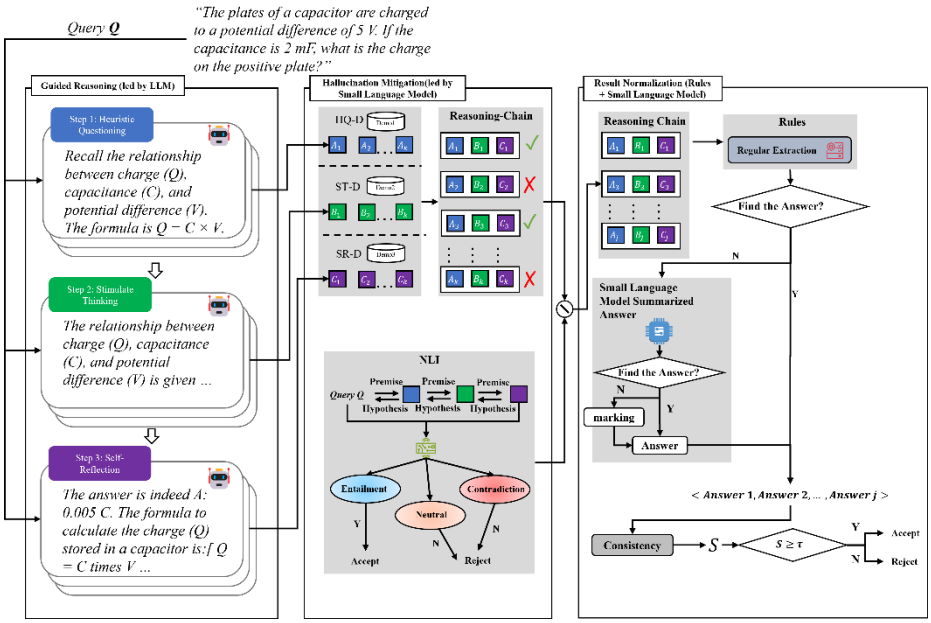

Fig. 1. Detailed overview of our proposed MS-HM Synergy (Multi-Stage Hallucination Mitigation Synergy) method.

On the other hand, smaller models, such as those based on BERT and its variants, offer complementary strengths. These models are often more efficient in terms of computational resources and can provide precise outputs for specific tasks. However, they typically have a narrower reasoning scope and may struggle with complex, multi-step problems that require deeper contextual understanding [26]. This dichotomy between LLMs and smaller models highlights the need for a framework that can effectively combine their respective strengths while mitigating their weaknesses.

Existing solutions to address the limitations of LLMs, such as self-correction mechanisms and ensemble methods, have shown promise but often lack structured collaboration between different model types. These approaches typically rely on a single model's capabilities or simple aggregations of multiple models' outputs, without fully leveraging the unique strengths of each model type in a coordinated manner [17,23,33]. As a result, there remains a significant gap in effectively integrating the creativity and reasoning depth of LLMs with the precision and efficiency of smaller models.

To bridge this gap, we propose the MS-HM Synergy (Multi-Stage Hallucination Mitigation Synergy) framework, a novel approach designed to enhance reasoning performance through phased collaboration between LLMs and small models. This framework consists of three essential stages:

- **Guided Inference:** In this stage, the LLM takes the lead in decomposing complex tasks through heuristic questioning, deliberation, and reflection. By breaking down the problem into manageable components, the LLM generates initial reasoning steps and potential solutions, leveraging its advanced language understanding and reasoning capabilities.
- **Hallucination Detection:** Following the initial reasoning, small models, such as BERT-NLI, are employed to meticulously identify and flag contradictions or inconsistencies in the LLM's outputs. This stage acts as a safeguard, ensuring that unreliable or hallucinated outputs are detected and addressed before finalizing the solution.
- **Result Standardization:** The final stage involves the hybrid use of rule-based and model-driven methods to format and standardize the output. This ensures that the generated results are coherent, structured, and aligned with predefined criteria, enhancing the overall reliability and usability of the framework.

By integrating these three stages, the MS-HM Synergy framework aims to effectively mitigate hallucinations, enhance reliability, and improve efficiency in complex reasoning tasks. Our methodology leverages the strengths of both LLMs and small models, creating a Synergistic environment where each model type contributes to the overall reasoning process. Empirical evaluations on established benchmarks, including MMLU [9], MATH [10], and LogiQA [20], demonstrate substantial performance improvements compared to existing state-of-the-art methods, such as traditional CoT and Socratic questioning [24]. The results highlight the framework's ability to enhance accuracy, manage complex problem decomposition, and provide a more robust and flexible solution for AI-driven reasoning tasks.

The main contributions of our paper are as follows:

1. **A novel phased collaboration paradigm between LLMs and small models:** We introduce the MS-HM Synergy framework, a pioneering approach that enables structured collaboration between Large Language Models (LLMs) and small models through a multi-stage process. This framework leverages the strengths of both model types, combining the reasoning depth and creativity of LLMs with the precision and efficiency of small models to enhance overall performance.
2. **NLI-based hallucination detection using task-specific small models:** Our framework incorporates Natural Language Inference (NLI)-based hallucination detection, utilizing task-specific small models such as BERT-NLI. This stage meticulously identifies and flags contradictions or inconsistencies in the LLM's outputs, effectively mitigating hallucinations and ensuring the reliability of the generated results.
3. **Hybrid standardization combining rules and model verification:** We propose a hybrid approach for result standardization, integrating rule-based formatting with model-driven verification. This ensures that the final outputs are coherent, structured, and aligned with predefined criteria, enhancing the usability and consistency of the framework across diverse applications.
4. **Empirical validation across reasoning, math, and coding tasks:** Our method demonstrates significant performance improvements across a wide range of tasks, including reasoning tasks (MMLU, LogiQA), mathematical problem-solving (MATH), and coding tasks. These empirical results validate the effectiveness of the

MS-HM Synergy framework in enhancing accuracy, efficiency, and flexibility compared to traditional methods.

2 Related Work

Synergistic Model Systems:

Model collaboration has been an area of active research in the field of artificial intelligence, with several prior works attempting to combine the strengths of different models. For instance, LLM cascades [36] have been proposed, where multiple LLMs are chained together in a sequential manner. The idea is to pass the output of one LLM as the input to the next, hoping to achieve more refined results. Knowledge distillation [8,11,29] is another approach, which aims to transfer the knowledge from a larger, more complex model (usually an LLM) to a smaller, more efficient one. This can potentially reduce the computational cost while maintaining a certain level of performance.

However, a critical drawback of these existing methods is the lack of phased interaction. Most of them treat the models as parallel agents rather than sequential collaborators. In LLM cascades, the communication between each stage is often simplistic and lacks the fine-grained coordination required for complex tasks. In knowledge distillation, the focus is mainly on knowledge transfer rather than on how the models can work together in a structured, phased manner to solve problems. This lack of phased collaboration limits the effectiveness of these methods in handling complex reasoning scenarios.

Hallucination Mitigation:

Several techniques have been proposed to mitigate hallucinations in LLMs. Self-consistency checks [30], as explored by some researchers, involve having the LLM generate multiple outputs and then checking for consistency among them. If there are discrepancies, the model attempts to reconcile them. Retrieval-augmented generation [16,28] is another approach, where the LLM retrieves relevant information from an external database or knowledge source and incorporates it into its generation process. This is supposed to ground the output in factual knowledge and reduce the likelihood of hallucinations. Confidence calibration [18,19] also plays a role, where the LLM estimates the confidence of its own outputs and adjusts its generation based on that confidence.

Nevertheless, these methods have their own limitations. They rely heavily on LLM introspection, which inherits the bias of the model itself. Since the LLM is evaluating its own work, it may not be objective enough to detect and correct all hallucinations. In contrast, our proposed approach uses external Natural Language Inference (NLI) models. These NLI models are trained to identify contradictions and inaccuracies, adding an extra layer of objectivity to the hallucination detection process. They can analyze the LLM output from an independent perspective, providing a more reliable means of mitigating hallucinations.

Output Standardization:

Regarding output standardization, both rule-based systems and model-based approaches have been explored. Rule-based systems, such as regex templates [14,21],

offer a deterministic way to format the output. They define a set of patterns and rules that the generated text must conform to. For example, in generating reports, regex templates can ensure that the date, time, and other relevant information are presented in a specific format. However, the major drawback of rule-based systems is their lack of adaptability. They are rigid and cannot easily handle variations in the input or changes in the required output format.

Model-based approaches, like using T5 for text simplification [25], rely on the power of language models to transform the output. T5 can take a complex text and simplify it according to certain criteria. But these pure model methods are computationally heavy. They require significant computational resources to train and operate. Our proposed hybrid approach combines the advantages of both. It uses a combination of rule-based formatting and model verification. The rules provide the basic structure and consistency, while the model verification ensures that the output meets the specific requirements of the application. This hybrid approach balances the need for adaptability and computational efficiency.

Positioning:

Our proposed MS-HM Synergy (Multi-Stage Hallucination Mitigation Synergy) framework uniquely integrates NLI validation, structured reasoning chains, and hybrid standardization into a unified pipeline. Unlike other methods that focus on only one aspect, such as just hallucination mitigation or just output formatting, our framework combines all these elements to address multiple challenges simultaneously. It provides a comprehensive solution for complex reasoning tasks, ensuring reliable and structured outputs. This integration is what sets our framework apart from the existing literature and positions it as a significant contribution to the field of artificial intelligence.

3 Method

3.1 Guided Inference (LLM-Dominated)

Heuristic Questioning: Let the input question be Q. The generation of multiple-round results $A_1, A_2, \dots, A_k$of heuristic questioning is modeled as:

$$P(A_1, A_2, \dots, A_k \mid Q) = \prod_{i=1}^{k} P(A_i \mid A_1, \dots, A_{i-1}, Q) \tag{1}$$

Stimulate Thinking: The generation of multiple-round results $B_1, B_2, \dots, B_k$ of stimulate thinking, conditional on the results $A_1, A_2, \dots, A_k$ of heuristic questioning and the input question Q, is modeled as:

$$\begin{aligned} &P(B_1, B_2, \dots, B_k \mid A_1, A_2, \dots, A_k, Q) \\ &= \prod_{i=1}^{k} P(B_i \mid B_1, \dots, B_{i-1}, A_1, \dots, A_k, Q) \end{aligned} \tag{2}$$

Self-Reflection: The generation of multiple-round results $C_1, C_2, \dots, C_k$ of self-reflection, conditional on the results $B_1, B_2, \dots, B_k$ of stimulate thinking, $A_1, A_2, \dots, A_k$of heuristic questioning, and the input question Q, is modeled as:

$$P(C_1, C_2, \ldots, C_k \mid B_1, B_2, \ldots, B_k, A_1, A_2, \ldots, A_k, Q) = \prod_{i=1}^{k} P(C_i \mid C_1, \ldots, C_{i-1}, B_1, \ldots, B_k, A_1, \ldots, A_k, Q) \quad (3)$$

3.2 Hallucination Mitigation

NLI - Based Natural Language Inference: In the Hallucination Detection phase, which is dominated by small language models, we employ Natural Language Inference (NLI) to filter out erroneous or contradictory information in the content generated by the Large Language Model (LLM). Specifically, we use a small Language model such as BERT for NLI tasks [3].

The input to the NLI process is structured as follows: we consider the content generated by the LLM as the ``hypothesis," and either the original query Q or a relevant snippet from a trusted knowledge base as the ``premise." The small model then determines the semantic relationship between the hypothesis and the premise, which can be classified as entailment, contradiction, or neutrality.

For example, in the context of our guided inference process, we have multiple - round generation results $A_1, A_2, \ldots, A_k$ from heuristic questioning, $B_1, B_2, \ldots, B_k$ from stimulate thinking, and $C_1, C_2, \ldots, C_k$ from self - reflection. These are spliced into Reasoning - Chains, where each chain is of the form (A_i, B_i, C_i) for i $= 1, \ldots,$ k.

We use NLI to check the consistency between different stages of the reasoning chain. The NLI process validates the relationships as follows: Query Q serves as the premise for A_i (where A_i is the hypothesis of Q), $A_i$serves as the premise for B_i(where B_i is the hypothesis of A_i), and B_i serves as the premise for C_i (where C_i is the hypothesis of B_i).

We define a scoring function for each reasoning chain (A_i, B_i, C_i) as follows:

$$\text{Score}(A_i, B_i, C_i) = \begin{cases} 1 & \text{if NLI(Premise: } Q\text{,Hypothesis: } A_i\text{)=Entailment} \\ & \text{and NLI(Premise: } A_i\text{,Hypothesis: } B_i\text{)=Entailment} \\ & \text{and NLI(Premise: } B_i\text{,Hypothesis: } C_i\text{)=Entailment} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Correction Strategy: Once the scores for each reasoning chain are calculated, we apply a filtering mechanism. We retain only those reasoning chains for which the product of the scores across all chains is equal to $1, i.e., \prod_{i=1}^{k} \text{Score}(A_i, B_i, C_i) = 1$. This ensures that only the reasoning chains that exhibit consistent entailment relationships from the query through all stages of the reasoning process are kept.

For reasoning chains that do not meet this criterion, if there is a contradiction detected (i.e., when NLI indicates a contradiction between the premise and hypothesis at any stage), we remove the entire reasoning chain. In cases where the relationship is marked as neutral, we also remove the corresponding part of the reasoning chain, as it does not provide a clear and consistent logical progression. This way, we effectively mitigate hallucinations in the LLM - generated content by leveraging the NLI capabilities of the small language model.

3.3 Result Standardization (Rule-Based + Small Language Model Synergistic

In the Result Standardization phase, our goal is to transform the corrected reasoning results from the previous stages into a target format that meets the requirements of downstream tasks. This is achieved through a synergy of rule - based methods and small language models.

Regular Extraction:

- **Structured Extraction:** We begin by using regular expressions and keyword matching techniques to extract the target information from the reasoning chains that have passed the hallucination detection phase.
- **Format Constraints:** In addition to information extraction, we enforce output templates to ensure that the results are presented in a consistent and usable format. This could involve adhering to JSON fields for structured data representation or using fixed option lists in classification tasks.

Small Language Model Summarized Answer:

- **Classification Verification:** After the rule - based extraction, we employ task - fine - tuned small models, such as BERT classifiers [7], to re - verify the extracted results. This step is crucial to confirm that the extracted classification labels or other key information is consistent with the overall text content.
- **Exception Handling:** In cases where the rule - based extraction fails to find an answer, we leverage the small language model to generate a simplified answer. For instance, if the input text is a long and complex description, and the regular extraction methods cannot identify a clear classification, the small model can summarize the text into a single classification label. If the small model also fails to find an answer, we mark the result as such, indicating that further investigation or human intervention might be required. Of course, the marked answers are considered wrong answers in the experiment.

Once the above steps are completed, we use a Consistency mechanism to set a threshold (denoted as τ). This threshold is used to evaluate the reliability of the generated answers. If the consistency score S of the answers meets or exceeds the threshold ($S \geq \tau$), the answers are accepted; otherwise, they are rejected. This final check helps to ensure that only high - quality and reliable results are delivered for downstream use.

4 Experiment Setups

We utilized GPT-3.5 Turbo [22] and Qwen-14B-chat [1] as the large language models. We evaluated the Reflection and Heuristic Questioning method across various complex reasoning tasks, including physics and chemistry tasks from the Massive Multitask Language Understanding (MMLU) dataset [9], mathematical tasks from the MATH dataset [10], logical reasoning tasks based on LogiQA [20], and classroom dialogue coding tasks (The classroom dialogue dataset consists of 175 teacher-student interactions, categorized into five distinct dialogue types.). Several state-of-the-art prompting methods were adopted as baselines, including the Standard Prompting, where the LLM directly answers questions, and the Chain-of-Thought (CoT) [32], which guides the LLM to first generate reasoning steps before providing the final answer. The Self-Consistency Chain-of-Thought (SC-CoT) method [31] further enhanced the reasoning process by running CoT multiple times and selecting the most consistent answers through marginalization.In addition, there are the Tree-of-Thoughts (ToT) [34], the Graph-of-Thoughts (GoT) [2] and the SOCRATIC QUESTIONING [24] method.

To ensure a fair comparison, we use Exact Match (EM) to evaluate the accuracy of all language tasks, following previous studies [5,12]. All questions in MMLU Physics,

MMLU Chemistry, and LogiQA are multiple-choice, with answers always represented by a single letter (e.g., “A”, “B”, or “C”). For the convenience of parsing the model's final outputs, we use “The answer is:” as a prefix for answers (e.g., A, B, C, or D) in the context examples for all methods. When parsing the outputs, we first extract the answer following “The answer is:” using a template-based approach.

5 Results

5.1 Quantitative Results

Table 1 resents the quantitative results of accuracy in language reasoning tasks using the Qwen-14BChat model. Our method demonstrates a significant improvement over previous state-of-the-art methods, with absolute advantages of 0.63%, 8.08%, 6.89%, 12.59%, and 2.29% in the Logic benchmarks, Physics, Chemistry, MATH, and Classroom dialogue encoding classifications, respectively. The Tree-of-Thoughts (ToT) and Socratic Questioning methodologies require the preconstruction of specific data structures, which inherently influence their outcomes based on the nature of the structures utilized. Consequently, these approaches are not directly compared with the MS-HM Synergy method when evaluated under the Qwen-14BChat model.

Table 1. Accuracy (%) using Exact Match with the Qwen-14BChat model. The best performance is highlighted in bold, and the second-best performance is underlined.

Qwen-14BChat	LogiQA	MMLU-physics	MMLU-chemistry	MATH(DA)	Classroom dialogue	Avg
Standard-Prompting	58.39	63.83	54.19	38.15	24.00	47.71
CoT	58.52	62.13	54.68	39.63	26.85	48.36
SC-CoT	58.78	62.55	56.16	41.48	25.14	48.82
MS-HM Synergy	**59.41**	**71.91**	**63.05**	**54.07**	**29.14**	**55.52**

Table 2. Accuracy (%) using Exact Match with the GPT-3.5-turbo model. The best performance is highlighted in bold, and the second-best performance is underscored.

GPT-3.5-turbo	LogiQA	MMLU-physics	MMLU-chemistry	MATH(DA)	Classroom dialogue	Avg
Standard-Prompting	54.67	65.11	53.20	7.00	17.14	39.42
CoT	48.33	67.66	57.14	7.33	20.00	40.09
SC-CoT	49.00	68.51	59.33	7.00	21.71	41.11
ToT	22.22	40.00	26.60	0.00	\	41.11
SOCRATIC QUESTIONING (2-Turns)	**59.33**	71.49	63.55	7.67	\	22.21
SOCRATIC QUESTIONING (3-Turns)	58.00	69.39	63.55	11.67	\	50.65
MS-HM Synergy	53.24	**72.77**	**64.03**	**53.33**	**24.57**	**53.59**

Table 2 presents the quantitative results of accuracy in language reasoning tasks using the GPT-3.5-turbo model. Among them, the comparative experimental data of MS-HM Synergy in Table 2 comes from the article [24]. Our method demonstrates superior performance over previous state-of-the-art approaches, with improvements of 1.28%, 0.48%, 41.66%, and 2.86% in the categories of Physics, Chemistry, MATH, and Classroom Dialogue Encoding, respectively. This effectively highlights the advantages of our approach.

It is important to note that the Classroom Dialogue dataset is in Chinese, whereas the Socratic Questioning method primarily targets English-language datasets. Therefore, implementing the Socratic Questioning method across different languages presents significant challenges and may not fully align with the original spirit of the

method. As a result, the experimental results for the Socratic Questioning method on the Classroom Dialogue dataset are not available in Table 2.

We conducted a detailed analysis of the performance of the MS-HM Synergy method across different rounds of dialogue. By comparing the accuracy of two-round and three-round dialogues, we found that increasing the number of rounds generally enhances the model's performance. For instance, in the Qwen-14BChat model, the average accuracy increased from 53.84% to 55.40% when the dialogue rounds were increased from two to three. Similarly, in the GPT-3.5-turbo model, the average accuracy improved from 50.20% to 53.30%. These results indicate that multi-turn dialogues help models better understand contextual information, thereby improving the accuracy of their responses.

However, for certain complex tasks, such as Classroom dialogue, increasing the number of rounds did not lead to significant performance improvements and even resulted in slight decreases in some cases. This may be attributed to the more intricate nature of the dialogue content in these tasks, which poses higher demands on the model's contextual understanding and reasoning capabilities. Therefore, our experimental results suggest that while increasing the number of dialogue rounds generally boosts model performance, further optimization of dialogue strategies and model design is still required for specific tasks to better adapt to complex dialogue scenarios.

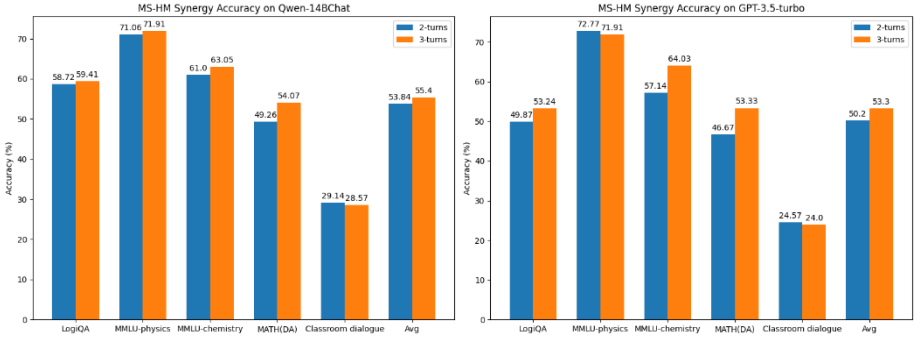

Fig. 2. Comparison of MS-HM Synergy Accuracy Across Different Turns on Qwen14BChat and GPT-3.5-turbo

5.2 Qualitative Result

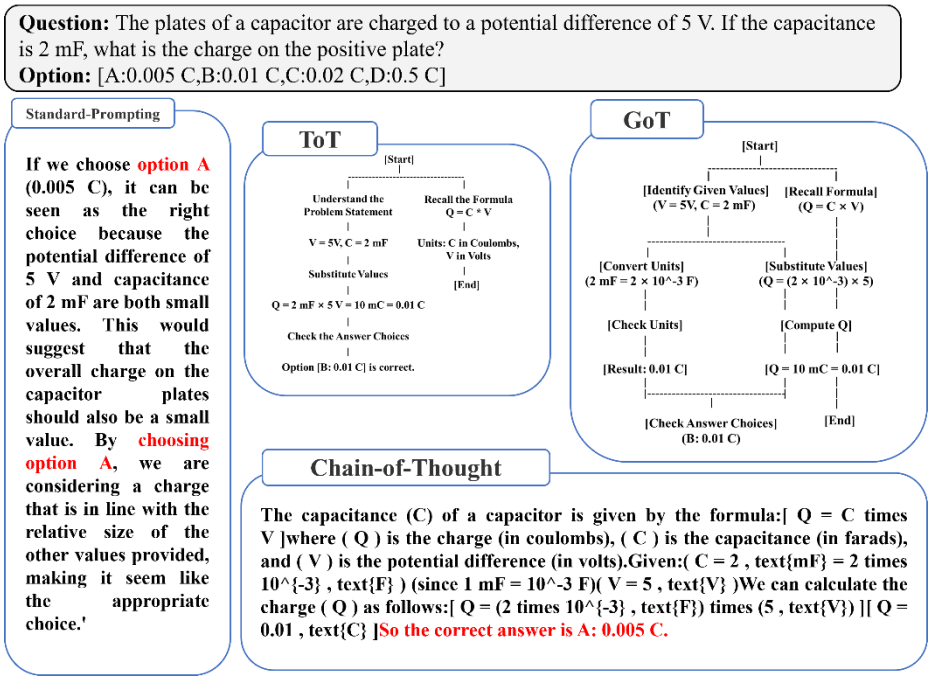

Fig. 3. Qualitative results of Standard-Prompting, CoT, ToT, and GoT on the Physics task. The correct answer of this example is B.

The qualitative results of the MS - HM Synergy method compared to baseline methods in the physics task are presented here. It can be observed that the MS - HM Synergy method effectively generates prompts containing the necessary information to solve the original problem. In the Guided Inference stage, the Large Language Model (LLM) taps into its language understanding to break down the problem and generate relevant prompts. Then, through the Hallucination Detection stage, unreliable information in these prompts is identified and eliminated, ensuring the accuracy of the information used for reasoning. Finally, the Result Standardization stage structures the output, making it coherent and easy to understand. By selectively utilizing these refined prompts, the MS - HM Synergy method provides a reasonable explanation and arrives at the correct final answer through reflective reasoning.

In contrast, Standard - Prompting, Chain - of - Thought (CoT), Tree - of - Thoughts (ToT), and Graph - of - Thoughts (GoT) methods yielded incorrect answers due to suboptimal reasoning paths, as shown in Fig. 3.

The Tree - of - Thoughts (ToT) method breaks problem - solving into linear, structured steps. It attempts to ensure thoroughness in reasoning, but it comes across as overly rigid and inefficient for simple tasks like the capacitor problem. Its repetitive, step - by - step nature lacks the flexibility needed for straightforward problems. In the context of our overall goal of leveraging the combined strengths of models, ToT fails

to adapt well to different problem complexities and doesn't scale well for problems where direct solutions are faster. Moreover, it does not have an effective mechanism to detect and correct potential hallucinations in the reasoning process.

Table 3. Qualitative results of MS-HM Synergy on the Physics task. The correct answer of this example is B.

	Reasoning-Chain 1	Reasoning-Chain 2	Reasoning-Chain 3
Guided Reasoning	**Step 1:** Recall $Q = C \times V$. Given $C = 2\,\text{mF}$, $V = 5\,\text{V}$. Convert 2 mF to $0.002\,\text{F}$. Then $Q = 0.002\,\text{F} \times 5\,\text{V} = 0.01\,\text{C}$. **Step 2:** By $Q = C \times V$, with $C = 0.002\,\text{F}$, $V = 5\,\text{V}$, we get $Q = 0.01\,\text{C}$, so answer is B. **Step 3:** Used correct formula, converted units right, got $Q = 0.01\,\text{C}$. Answer B is correct.	**Step 1:** Need to find charge, formula is $Q = C \times V$. Given $C = 2\,\text{mF}$, $V = 5\,\text{V}$, convert C to $0.002\,\text{F}$. **Step 2:** Substitute into $Q = C \times V$, $C = 0.002\,\text{F}$, $V = 5\,\text{V}$, get $Q = 0.01\,\text{C}$, correct option is B. **Step 3:** Recognized formula, did unit conversion right, found $Q = 0.01\,\text{C}$. B is the answer.	**Step 1:** Recall $Q = C \times V$. Given $C = 2\,\text{mF}(= 0.002\,\text{F})$, $V = 5\,\text{V}$. But wrongly calculate $Q = 0.002\,\text{F} \times 10\,\text{V} = 0.02\,\text{C}$. **Step 2:** Should be $Q = C \times V$, with correct $V = 5\,\text{V}$ gives $Q = 0.01\,\text{C}$, but wrong calc led to think C: 0.02 C is right. **Step 3:** Made error in using $V = 10\,\text{V}$ instead of 5 V, correct is C: 0.02 C.
Hallucination Mitigation	Accept	Accept	Reject
Result Normalization	B: 0.01 C	B: 0.01 C	C: 0.02 C

The Graph - of - Thoughts (GoT) method organizes problem - solving non - linearly, enabling flexible exploration of interconnected steps. While it can be effective for complex problems, it overcomplicates simple tasks by introducing unnecessary nodes, backtracking, and cognitive load. This makes it visually and procedurally inefficient for direct, formula - based solutions. Similar to ToT, GoT also lacks a comprehensive approach to handling hallucinations and standardizing the output in a way that meets the requirements of downstream tasks.

The MS - HM Synergy method offers a significant advantage in multi - step logical reasoning by overcoming key limitations of traditional Chain - of - Thought (CoT) techniques. Unlike sequential or predefined structured approaches that are prone to error propagation and challenges in decomposing complex problems, MS - HM Synergy combines the power of LLMs for complex reasoning with the precision of small language models for verification. Through the three - stage process of Guided Inference,

Hallucination Detection, and Result Standardization, it achieves more accurate and flexible problem - solving, as shown in Table 3.

6 Conclusion

The MS-HM Synergy framework addresses the challenges of hallucinations and inefficiencies in Large Language Models (LLMs) by integrating guided inference, hallucination detection, and result standardization. This approach leverages the strengths of both large and small language models, with LLMs handling complex reasoning and small models verifying reliability. Empirical results on benchmarks like MMLU, MATH, and LogiQA demonstrate significant improvements in accuracy, efficiency, and flexibility. The MS-HM Synergy effectively mitigates hallucinations and enhances overall performance, showcasing the potential of model synergy to overcome LLM limitations and drive advancements in AI.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 62177006, 62077009), and Guangdong Provincial Natural Science Foundation (Grant No. 2025A1515010136, 2022A1515011541); and partially supported by the Guangdong Province Undergraduate Course Teaching and Research Office Construction Project (Grant No. jx2022303). Besides, this work is supported by the Interdisciplinary Intelligence Super Computer Center of Beijing Normal University at Zhuhai.

Disclosure of Interests. None.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al.: Graph of thoughts: Solving elaborate problems with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17682-17690 (2024)
3. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
4. Carmichael, Z.: Explainable ai for high-stakes decision-making. University of Notre Dame (2024)
5. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research 24(240), 1-113 (2023)
6. Creswell, A., Shanahan, M.: Faithful reasoning using large language models. ArXiv abs/2208.14271 (2022)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171-4186 (2019)

8. Gu, Y., Dong, L., Wei, F., Huang, M.: Knowledge distillation of large language models. ArXiv abs/2306.08543 (2023)
9. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
10. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874 (2021)
11. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. ArXiv abs/1503.02531 (2015)
12. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
13. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems 43(2), 1-55 (2025)
14. Kale, M., Rastogi, A.: Template guided text generation for task-oriented dialogue. arXiv preprint arXiv:2004.15006 (2020)
15. Kaur, D., Uslu, S., Durresi, A.: Trustworthy ai explanations as an interface in medical diagnostic systems. In: International Conference on Network-Based Information Systems. pp. 119-130. Springer (2022)
16. Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)
17. Li, J., Wang, X., Zhu, S., Kuo, C.W., Xu, L., Chen, F., Jain, J., Shi, H., Wen, L.: Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. Advances in Neural Information Processing Systems 37, 131224-131246 (2024)
18. Lin, Q., Zhou, L., Yang, Z., Cai, Y.: Label-confidence-aware uncertainty estimation in natural language generation. CoRR abs/2412.07255 (2024)
19. Lin, Z., Trivedi, S., Sun, J.: Generating with confidence: Uncertainty quantification for black-box large language models. arXiv preprint arXiv:2305.19187 (2023)
20. Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., Zhang, Y.: Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. arXiv preprint arXiv:2007.08124 (2020)
21. Mille, S., Sabry, M., Belz, A.: Dcu-nlg-small at the gem24 data-to-text task: Rule-based generation and post-processing with t5-base. In: Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges. pp. 84-91 (2024)
22. OpenAI, T.: Chatgpt: Optimizing language models for dialogue. openai (2022)
23. Pan, L., Saxon, M.S., Xu, W., Nathani, D., Wang, X., Wang, W.Y.: Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. ArXiv abs/2308.03188 (2023)
24. Qi, J., Xu, Z., Shen, Y., Liu, M., Jin, D., Wang, Q., Huang, L.: The art of socratic questioning: Recursive thinking with large language models. arXiv preprint arXiv:2305.14999 (2023)

25. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 140:1-140:67 (2020)
26. Ranieri, A., Caputo, D.J., Verderame, L., Merlo, A., Caviglione, L.: Deep adversarial learning on google home devices. J. Internet Serv. Inf. Secur. 11, 33-43 (2021)
27. Rillig, M.C., Ågerstrand, M., Bi, M., Gould, K.A., Sauerland, U.: Risks and benefits of large language models for the environment. Environmental science & technology 57(9), 3464-3466 (2023)
28. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.: REPLUG: retrieval-augmented black-box language models. In: Duh, K., Gómez-Adorno, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024. pp. 8371-8384. Association for Computational Linguistics (2024)
29. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for bert model compression. In: Conference on Empirical Methods in Natural Language Processing (2019)
30. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023)
31. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (2023)
32. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35, 24824-24837 (2022)
33. Yang, H., Li, M., Zhou, H., Xiao, Y., Fang, Q., Zhang, R.: One llm is not enough: Harnessing the power of ensemble learning for medical question answering. medRxiv (2023)
34. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems 36 (2024)
35. Ye, H., Liu, T., Zhang, A., Hua, W., Jia, W.: Cognitive mirage: A review of hallucinations in large language models. arXiv preprint arXiv:2309.06794 (2023)
36. Yue, M., Zhao, J., Zhang, M., Du, L., Yao, Z.: Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. ArXiv abs/2310.03094 (2023)