Template-Theorems Graph Construction to Enhance Mathematical Reasoning Capabilities of LLM

Yarong Lan[1,2] **Yajing Xu**[1,2] **Huajun Chen**[1,2,3*]
[1]Zhejiang University
[2]ZJU-Ant Group Joint Lab of Knowledge Graph
[3]Zhejiang Key Laboratory of Big Data Intelligent Computing
{yrlan16, huajunsir}@zju.edu.cn

Abstract

Large language models (LLMs) have made significant strides in mathematical reasoning, particularly at the elementary level. However, they continue to face substantial challenges when confronted with complex, advanced mathematical problems. In contrast to humans—who can effectively draw upon prior experiences in solving similar problems and retrieve relevant knowledge and theorems from memory—LLMs often struggle to accurately identify analogous problems and to recall or apply appropriate theorems. To overcome these limitations, we introduce a novel framework for constructing a template-theorems knowledge base, leveraging the capabilities of large language models. Inspired by the associative mechanisms of human cognition, our approach abstracts real-world problems into generalized templates and establishes intricate linkages between these templates and pertinent theorems. This design enables the efficient expansion of a comprehensive knowledge base, even when starting from a limited set of seed examples. Moreover, we develop an efficient retrieval strategy that, given a new problem, systematically extracts and presents the most relevant knowledge from the knowledge base as contextual input to the LLM. Extensive experiments on multiple public mathematical datasets and models demonstrate that our approach consistently surpasses conventional methods. Comprehensive ablation studies further corroborate the effectiveness of both our knowledge base construction and retrieval modules.

1 Introduction

Mathematical reasoning is both a fundamental and challenging task in the advancement of artificial intelligence. As a core aspect of human cognition, mathematical reasoning (MR) has long been recognized as a central objective in AI research (Iuculano and Menon 2018). With exponential increases in both training data and model capacity, mainstream large language models (LLMs)—such as those developed by OpenAI(OpenAI 2024) and Google(Google 2023)—have achieved remarkable progress on mathematical reasoning benchmarks (Cobbe et al. 2021a; Wang et al. 2024b; Hendrycks et al. 2021). Nevertheless, these models continue to exhibit limitations when confronted with more complex mathematical challenges.

*Corresponding author.

Recent studies in mathematical reasoning have shown that large language models(OpenAI 2024; Dubey et al. 2024; Team 2025; DeepSeek-AI et al. 2025)acquire extensive mathematical knowledge during pre-training, both implicitly and explicitly, spanning a wide range of theorems from elementary to advanced levels. However, in actual problem-solving and reasoning tasks, these theorems are often not properly retrieved or applied(Wang et al. 2024a). This indicates that, while the models may "know" the relevant mathematics, they have yet to develop efficient mechanisms for knowledge retrieval and application. In contrast, humans

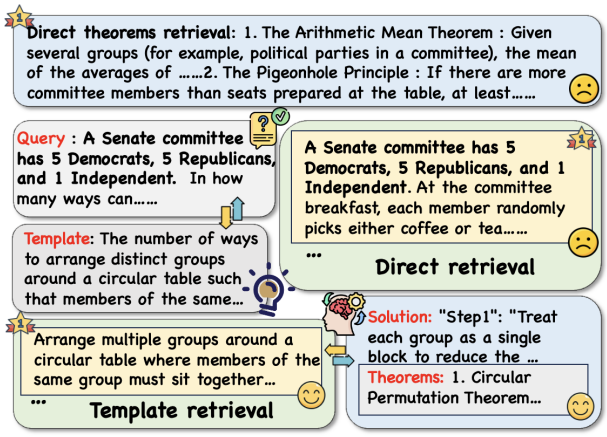

Figure 1: An illustrative comparison between retrieval using a brain-inspired template-theorems graph and retrieval using direct question or theorems with similarity.

solving mathematical problems typically search their vast mental knowledge base for similar problems they have previously encountered and recall the key theorems needed to address them(Silver, Katz, and Lesgold 1980). This cognitive process enables efficient identification of relevant knowledge while avoiding the misapplication of unrelated theorems.Inspired by this human approach, we propose to construct a graph that encapsulates the relationships between mathematical problems and theorems, emulating the memory organization of the human brain. Just as people naturally connect their experience solving problems with the pertinent theorems, this "template-theorems graph" stores a rich set

of interlinked mathematical problems and associated theorems. Serving as an additional knowledge base(Bollacker et al. 2008; Auer et al. 2007), this graph will underpin a robust mechanism for knowledge retrieval and application, thereby helping models to more effectively locate and utilize the mathematical knowledge required for complex reasoning tasks.

However, while humans are adept at discerning underlying connections between disparate topics, current approaches are largely constrained to identifying such relationships through surface-level similarity retrieval(Wu et al. 2025). Traditional similarity retrieval methods (such as cosine similarity (Salton and McGill 1983)and BM25(Robertson 2009)) primarily rely on the shallow semantic similarity of question text or its encoded representations. However, in mathematical word problems, proof-based problems, and even some competition questions, the surface expression of the question can vary significantly, while their underlying structures, problem-solving strategies, or theorems applied often remain the same. Retrieval methods based solely on surface semantic similarity frequently fail to identify high-value samples that are "structurally equivalent but differently expressed."

To address the limitations of semantic-based retrieval, recent studies have introduced retrieval approaches based on structural templates—such as code snippets, mathematical expressions, or induction frameworks—to better capture problem instances with fundamentally similar structures(Yang et al. 2024). Building on this idea, we propose a template extraction method that elevates the information stored in the knowledge graph from specific questions to higher-level, more generalizable problem-solving templates. This approach enables a more effective representation of the deep connections between problems and allows the model to accurately identify and apply core mathematical knowledge and reasoning strategies, even in the face of diverse surface expressions.

We conducted comprehensive evaluations leveraging LLMs across five widely-used mathematical reasoning datasets (Cobbe et al. 2021a; Wang et al. 2024b; Hendrycks et al. 2021; Zhang et al. 2025) with different levels. Experimental results indicate that our proposed framework yields robust and consistent improvements over established baselines, thereby substantiating its efficacy in augmenting the mathematical reasoning proficiency of LLMs. Moreover, our approach achieves these gains while also reducing inference latency, as evidenced by comparative analysis of response time metrics—highlighting notable improvements in both effectiveness and computational efficiency. Through meticulous ablation studies, we further corroborate the hypothesis that LLMs encode substantial implicit mathematical knowledge (such as theorems and problem-solving strategies); however, they frequently fail to autonomously retrieve and deploy this knowledge in problem-solving contexts. Overall, our contributions are outlined as follows:

- We present a novel graph database architecture for mathematics, the first to systematically integrate problem-solving templates with foundational theorems, drawing inspiration from the structural principles of human memory. This database has undergone rigorous validation through both algorithmic and systematic manual verification, establishing high reliability.
- We propose a comprehensive pipeline encompassing both construction and enhancement phases, which leverages the capabilities of LLMs to automatically generate template-theorems graphs. This approach is applied to enhance the performance of models in tasks involving mathematical reasoning.
- Our method has demonstrated superior performance across a variety of datasets and models, thereby providing strong evidence for the universality and robustness of the proposed approach. Furthermore, experimental results indicate that the method consistently achieves high performance across diverse tasks and scenarios, highlighting its broad applicability and significant practical value.

2 Related Work

2.1 Mathematical Reasoning in LLM

In recent years, the pursuit of equipping large language models (LLMs) with advanced mathematical problem-solving capabilities has emerged as a prominent and dynamic subfield within artificial intelligence. This focus has catalyzed the proliferation of a wide array of benchmark datasets designed to evaluate LLMs on multiple dimensions of mathematical proficiency (Cobbe et al. 2021b; Hendrycks et al. 2021; Chen et al. 2023). Methodologically, existing research can be broadly categorized along two principal axes. One major line of inquiry is devoted to enhancing open-source lightweight LLMs through strategies such as supervised fine-tuning and instruction tuning, thereby augmenting their adaptability and robustness across diverse mathematical reasoning scenarios(Zhang et al. 2023; Sheng, Li, and Zeng 2025).In parallel, a substantial body of work has concentrated on improving the zero-shot and few-shot (Wei et al. 2022)reasoning abilities of proprietary, closed-source models. This is accomplished by developing sophisticated prompting paradigms—including chain-of-thought reasoning, stepwise problem decomposition, and multi-path search integration—to more effectively navigate complex mathematical solution spaces (Wang et al. 2022; Besta et al. 2023; Yao et al. 2023b). Notably, approaches predicated on closed-source LLMs often encounter intrinsic limitations arising from their reliance on single-step inference and repeated, resource-intensive API interactions, which may hinder their scalability and limit opportunities for further optimization. Concurrently, the development of rigorous evaluation protocols and comprehensive benchmark frameworks for assessing mathematical reasoning at varying levels of complexity has become a central concern, reflecting the need for systematic and nuanced model assessment within this rapidly evolving domain.

2.2 Knowledge Graph Construction with LLM

Since the advent of large-scale pre-trained models (LLMs), the automatic construction of knowledge graphs leveraging

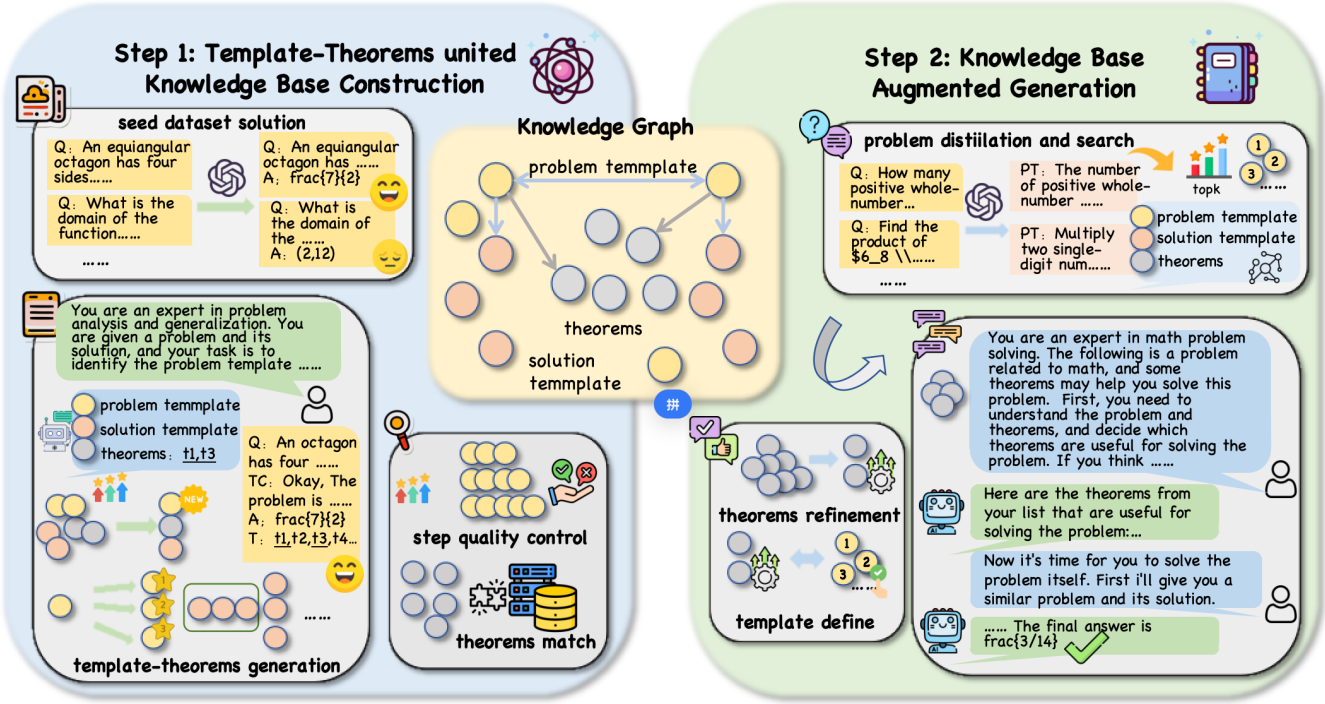

Figure 2: The overview of our approach. The pipeline consists of template-theorems united knowledge base construction and knowledge base augmented generation. On the left side we demonstrate the details of constructing a graph using seed entity. On the right side we detail the process of answering a math problem using the retrieved information.

these models has emerged as a focal point in recent research(Yao et al. 2023a; Trajanoska, Stojanov, and Trajanov 2023). Prior work in this area predominantly follows two paradigms. The first involves designing dedicated frameworks that guide LLMs to extract information from textual input and construct knowledge graphs(Han et al. 2023; Huang et al. 2024; Chen et al. 2024), a process commonly referred to as "text-to-graph generation," wherein natural language passages are transformed into corresponding semantic graphs. However, this approach is often hindered by the prevalence of noise in the source text, making noise reduction both complex and resource-intensive. The second paradigm seeks to directly generate knowledge graph structures from the internal representations of large models, thereby extracting the knowledge implicitly encoded within. Typically, such methods(Cohen et al. 2023) begin with seed entities and iteratively expand to form tree-like graph structures. While both paradigms facilitate the construction of knowledge graphs containing static information, their practical utilization in downstream applications remains limited.

To address these limitations, the present study innovatively integrates elements from both paradigms and proposes a novel two-stage generation process. Initially, we leverage a limited set of existing problem-centric datasets to automatically extract and generate brain-inspired template theorem graphs from the intrinsic knowledge of large models, tailored to meet specific design criteria. Subsequently, a systematic verification mechanism is employed to ensure the high quality and reliability of the generated graphs, effectively obviating the need for laborious noise reduction steps. This method not only enhances the efficiency and quality of knowledge graph construction, but also facilitates the effective application of resultant graphs in downstream tasks.

3 Method

Inspired by the processes underlying human mathematical reasoning, we leverage large language models (LLMs) to construct a graph-based framework that integrates high-level problem-solving templates with mathematical theorems. We further design a tailored retrieval mechanism to efficiently identify and deliver relevant contextual information from this graph. Starting with a minimal set of seed data, LLMs are employed to generate a substantially expanded knowledge base, with rigorous quality control procedures applied to ensure data reliability. The resulting structured repository serves as a robust foundation for precise retrieval and provision of problem-specific knowledge. An overview of our methed is shown in Figure 2.

3.1 Template-Theorems Graph Construction

In mathematics and other domains of complex reasoning, the formulation and presentation of questions are highly variable, and even questions of the same underlying type may appear in a multitude of different expressions. When confronted with such problems, humans are able to intuitively retrieve and align analogous experiences from their prior encounters, often transcending superficial textual differences.

In contrast, conventional external retrieval methods typically rely on semantic similarity metrics to identify the most similar questions, which tend to capture only surface-level resemblance. Consequently, the retrieved examples often diverge from those that humans would perceive as truly analogous, thus failing to provide meaningful experiential guidance and, in some cases, introducing extraneous or confusing information. A similar challenge arises in the retrieval of relevant theorems: naively selecting theorems solely on the basis of semantic similarity to the question can result in a mismatch between the theorems actually required for problem-solving and those retrieved. To bridge this gap, we emulate the associative linking structure between related questions and theorems as manifested in human cognition, and propose the construction of a template-theorems graph.

Generation Templates are defined as generalized representations of problems and their corresponding solutions, abstracted from specific instances. By leveraging these templates, it becomes possible to efficiently address a broad range of problems that fall within the scope of the template's applicability. In our framework, templates are further categorized into problem templates and solution templates. A solution template is explicitly organized into a sequence of reasoning steps, with each step encapsulating a substantive aspect of problem-solving, as opposed to a superficial partitioning based solely on punctuation or syntactic boundaries. To operationalize this approach, we introduce a **two-stage generation pipeline**.

- **Base Generation**: A seed dataset is first constructed by selecting questions with associated difficulty labels from the corpus. The difficulty label is subsequently leveraged during the second stage of the generation pipeline. In the initial generation phase, a reasoning model is employed alongside a carefully designed few shot prompts tailored to each question, enabling the model to generate an initial set of answers. In the subsequent stage, the model produces an unconstrained set of theorems that are invoked during the problem-solving process. This step serves as the foundation for template construction, as the theorems generated here are closely aligned with the specific problem and may be regarded as concrete instantiations of mathematical theorems within a particular context. Building upon this, question–answer reasoning chains are utilized as inputs to guide the generation of template questions and corresponding solutions, using a limited set of well-crafted few shot prompts to ensure conformity with predefined requirements. For each template, a series of related theorems—corresponding to the number of reasoning steps—is generated. The templates and theorems produced through the aforementioned two-stage process exhibit strong alignment with the predefined criteria.
- **Advanced Generation**: The volume of data produced during the basic generation stage is inherently limited and insufficient for capturing the full diversity of problem types and solution strategies. To address this limitation, we introduce an advanced generation stage, which comprises two complementary methods.

  The first method involves randomly sampling data across different difficulty levels from the basic generation stage and assembling them as few-shot examples. Using prompts specifically crafted for this setting, and leveraging a model distinct from that used in the basic generation stage, we directly generate templates and corresponding theorems across varying levels of difficulty, independent of any specific problem instance. This approach substantially enhances data diversity while mitigating potential biases introduced by reliance on a single model, as employing a different model helps ensure a broader and more balanced data distribution.

  The second method within the advanced generation stage focuses on modifying the core question within an existing template, while preserving the original problem context. Consequently, the associated templates and theorems are also adapted to reflect these changes. The newly generated templates in this paradigm maintain an inherent hierarchical relationship with the original templates, collectively forming a natural branching structure within the template-theorems graph.

Verification Given that large language models are prone to hallucination, it is essential to enforce rigorous quality control on the generated data. To this end, we have devised a set of validation rules. For the basic generation stage, two principal verification procedures are employed:

- Answer Validation: The model-generated answers to questions are compared against the ground-truth answers in the dataset. Templates and theorem data derived from incorrect answers are systematically filtered out, ensuring that only high-quality, accurate data are retained.
- Consistency Validation: Theorems are generated both from responses to actual questions and from corresponding template questions. However, theorems generated on the basis of template questions are generally more aligned with our definition of requisite mathematical theorems. Thus, we treat the theorems derived from template questions as the reference set and assess their consistency with the theorems generated from actual questions. A theorem is deemed valid if it meets a specified equivalence criterion; the data as a whole are retained if the proportion of valid theorems relative to all generated theorems exceeds a predefined threshold.

$$C(t) : max(sim(f(t), [f(T_i)_{i=0}^{n}]) > k \tag{1}$$

$$t_{t \in C(t)} / t_{total} > \alpha \tag{2}$$

For the data generated through the final mixed two-stage process, we employ two rules to ensure data quality:

- Step Quality Control: We posit that solution templates for more challenging problems should inherently involve a greater number of reasoning steps. Accordingly, for each designated difficulty level, we filter out data samples whose number of steps falls outside the prescribed range.

$$\overline{S} - \frac{1}{n}(S_{max} - S_{min}) < S < \overline{S} + \frac{1}{n}(S_{max} - S_{min}) \tag{3}$$

- Theorem Matching: We compare the generated theorems against an authoritative theorem library, applying a verification procedure analogous to the aforementioned consistency check, but with an appropriately adjusted threshold. This step ensures that the generated theorems are logically consistent with established mathematical standards.

3.2 Graph Retrieval-Augmented Generation

Building upon the foundation of a template-theorem library that emulates human cognitive structures, it is also imperative to simulate the human retrieval process when confronted with novel problems. Accordingly, following the construction of the initial template-theorem graph, we design an enhanced retrieval and generation procedure aimed at efficiently identifying and leveraging the most relevant information. This process ensures that the retrieved knowledge is optimally integrated, thereby maximizing the model's problem-solving efficacy.

Problem Distillation As previously noted, mathematical questions exhibit significant variability in their formulations. Relying solely on direct question-based similarity retrieval is insufficient to bridge the gap between retrieved questions and the specific requirements of new queries. To address this limitation, we introduce a question distillation step, employing carefully designed prompts to extract a broader set of representative problem templates from the questions to be answered. This approach facilitates more effective alignment with the template-theorem graph during retrieval. Notably, this distillation process is conducted offline using a well-chosen model prior to engaging in actual question answering.

Retrieval and Refinement In alignment with the approach whereby entities in a knowledge graph are linked through textual similarity, we utilize the distilled template questions to query the template-theorem graph, thereby retrieving the top- k most relevant questions along with their associated solution templates and theorems. To minimize the introduction of irrelevant information, the retrieval process is further constrained not only by the parameter k, but also by a matching similarity threshold m. The specific retrieval and return mechanism follows the procedural steps detailed in the subsequent algorithm.

Upon retrieval, the first step is to aggregate all relevant theorem information, as distinct questions may reference identical theorems. All candidate theorems are presented as input to the model, which, guided by tailored prompts, integrates them into a comprehensive theorem table. The second step involves reorganizing the retrieved templates—ranked by similarity—according to the consolidated theorem table. Early experimental results indicate that, in cases where the final selected theorems and templates are not fully compatible, their combined use may underperform compared to using either component in isolation. To address this, the template definition step re-selects the most relevant template based on the established theorem table.

Ultimately, the finalized theorem table and the most relevant template are provided as contextual input to the downstream model tasked with question answering, thereby enhancing its performance.

Algorithm 1: Similarity-based Top-k Retrieval

Input: Query q, candidate set $C = \{c_1, c_2, ..., c_n\}$, similarity function $sim(\cdot, \cdot)$, top number k, similarity threshold m
Output: At most k candidates with highest similarity (all with $sim \geq m$)

```
Initialize empty list S
for each c in C do
    Compute similarity s = sim(q, c)
    Add (c, s) to list S
end for
Sort S by similarity s in descending order
Initialize R ← empty list
for each (c, s) in the first k elements of S do
    if s ≥ m then
        Add c to R
    else
        break
    end if
end for
return R
```

4 Experimental Settings

4.1 Datasets

We conduct evaluations on a total of five publicly available datasets in the field of mathematical reasoning, encompassing both in-distribution (IND) datasets—MATH, and out-of-distribution (OOD) datasets, including MMLU-pro-Math(Wang et al. 2024b), AMC10, AMC12(Zhang et al. 2025), AIME24(aim 2024). The inclusion of OOD datasets allows us to rigorously assess the model's ability to generalize to unfamiliar scenarios. These datasets collectively span a wide spectrum of mathematical topics and difficulty levels, ranging from primary to college-level mathematics.

Compared to prior studies, our work places greater emphasis on datasets at the college level and beyond, as these remain particularly challenging and are of significant interest for advancing research in mathematical reasoning. Furthermore, the selected datasets comprise both open-ended and multiple-choice questions, enabling a thorough examination of the model's performance across various question formats. This comprehensive suite of evaluations ensures a robust assessment of model capabilities in mathematical reasoning over diverse domains and levels of complexity.

4.2 Baseline and Metrics

We selected several representative models to demonstrate the generalizability of our approach, including popular GPT series(OpenAI 2024)models such as GPT-4o and GPT-4.1, which differ in their mathematical reasoning capabilities, as well as state-of-the-art reasoning models like DeepSeek-R1(DeepSeek-AI et al. 2025). To evaluate the effectiveness of our method, we compared its performance against zero-shot and few-shot chain-of-thought (CoT) baselines(Wei et al. 2022) (using 8 examples, consistent with the original

Model	Method	in-domain	out-domain			
		MATH	MMLU-Pro-Math	AMC10	AMC12	AIME24
GPT-4.1	0-shot	89.2	90.97	73.27	85.50	43.2
	few-shot	89.4	90.97	75.12	85.50	46.5
	Ours	**91.4**	**91.78**	**77.88**	**86.23**	**48.4**
DeepSeek-R1	0-shot	96.4	94.15	85.25	94.2	70.0
	few-shot	96.4	94.23	84.79	94.93	72.3
	Ours	**97.4**	**94.52**	**85.71**	**96.38**	**76.2**
GPT-4o	0-shot	73.4	79.27	55.80	50.72	13.20
	few-shot	74.8	78.31	57.14	52.17	13.20
	Ours	**76.4**	**79.79**	**58.52**	**54.35**	**16.15**

Table 1: A comparison of various in-context learning strategies across different benchmarks using GPT series and reasoning models. The seed problems are sourced from the math training set, making MATH500 an in-domain benchmark, while all others are considered out-of-domain. The best results for each benchmark are highlighted in bold.

works) across all model types. To ensure fairness, all experimental prompts were kept consistent apart from the information specific to each method.

In line with prior research, we adopt accuracy as the primary evaluation metric. Given that our datasets include both multiple-choice and open-ended questions, relying solely on regular expression matching for answer evaluation may fail to capture all correct responses. Therefore, we employ a dual-verification strategy: both regular expressions and model-based judgments are used to assess answer correctness. An answer is considered correct if either method validates it. For efficiency, we use GPT-4o-mini as the evaluation model throughout our experiments.

4.3 Hyperparameter Settings

For the base generation stage, we employ the deepseek-r1 model to efficiently acquire chain-of-thought data. For the advanced generation stage, we utilize the gpt-4.1 model to further enhance data diversity and set the temperature to 0.3 and top-p to 0.8. For inference, we set the temperature to 0.9 and top-p to 0.95 for GPT series.

5 Results and Analysis

5.1 Overall Performance

Table 1 presents the experimental results on several public datasets. Overall, the context information retrieved and provided by our method—utilizing the constructed template theorem graph—consistently yields accuracy improvements across various evaluation datasets and different models spanning a range of mathematical difficulty levels, when compared to traditional zero-shot and few-shot CoT methods. These empirical findings demonstrate the effectiveness of both our proposed template theorem graph and the associated retrieval approach. Notably, the observed performance gains are evident not only on the seed datasets used for constructing the template theorem graph, but also on entirely unseen datasets that were not involved in the construction process. This indicates that the template theorem graph encapsulates key experiences and fundamental theorems necessary for solving general mathematical problems, and that our method enables accurate identification and effective utilization of relevant theorems.

5.2 Ablation Study

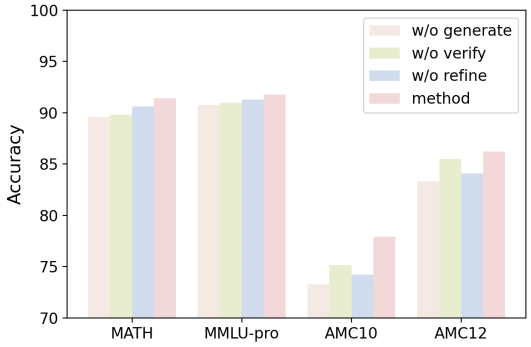

Figure 3: Ablation study for our method in the four benchmarks with GPT-4.1.

To further investigate the contribution of each component in our method to enhancing the model's mathematical capabilities, we conducted a series of ablation experiments on the overall framework. Specifically, the variant that omits the advance generation step and performs only base generation is denoted as w/o generate; the variant without the verification step is denoted as w/o verify; the variant without refining the theorems retrieved is denoted as w/o refine; and the full method is denoted as method.

As illustrated in Figure 3, the omission of any individual component within our methodology results in a notable degradation of overall performance, underscoring the indispensable nature of each procedural step to the method's efficacy. Particularly, the absence of the advance generation stage exerts a pronounced detrimental effect, most evident on out-domain datasets, thereby emphasizing the pivotal role

of enhanced data diversity during this stage in bolstering model capability. Moreover, it is noteworthy that the model employed in the advance generation phase is identical to the inference model (GPT-4.1) utilized in the ablation experiments. This further substantiates that the model inherently possesses the requisite theorems for problem-solving, and that our approach is instrumental in eliciting and effectively leveraging such embedded knowledge.

5.3 Template and Theorems United

Just as humans often draw upon prior experience and background theorems when solving mathematical problems, models similarly require access to both types of knowledge to effectively tackle challenging tasks. Template knowledge provides procedural guidance and familiar patterns, while theorem knowledge supplies the fundamental principles needed for correct reasoning. The integration of these complementary sources of information is essential for comprehensive problem-solving, mirroring the cognitive strategies employed by human experts.

To empirically validate the necessity and effectiveness of both types of knowledge, we conducted experiments on the GPT-4.1 model under three different settings: providing only template knowledge, only theorem knowledge, and both in combination. The results in Table 2 consistently demonstrate that jointly supplying template and theorem knowledge leads to substantial improvements in model performance. This finding strongly supports our theoretical expectation that leveraging both forms of knowledge enables more robust and accurate mathematical reasoning.

	original	+Tem	+The	+Both
MATH	89.2	90.4	91.2	**91.4**
MMLU-pro-Math	90.97	90.97	91.27	**91.78**
AMC10	73.27	75.58	76.50	**77.88**
AMC12	85.50	85.50	86.23	**86.23**

Table 2: Performance comparison across various datasets with different settings: providing only template knowledge, only theorem knowledge, and both in combination.

5.4 Efficiency Analysis

Figure 4 presents a comparative analysis of the cost time across different datasets using two methods: few-shot and ours, evaluated on both GPT-4.1 and GPT-4o. Overall, our method consistently demonstrates a reduction in computational time compared to the few-shot baseline on all evaluated datasets.

Specifically, for both GPT-4.1 and GPT-4o, the "ours" approach achieves lower cost time across nearly all tasks. For instance, on GPT-4.1, the reduction in cost time is particularly evident for MATH, MMLU-Pro, and AMC12. Similarly, on GPT-4o, our method leads to a noticeable decrease in cost time, showing its robustness and efficiency across different model backbones. These results indicate that our approach is not only more effective, but also more efficient in terms of resource consumption, making it practical for deployment on various real-world mathematical evaluation tasks.

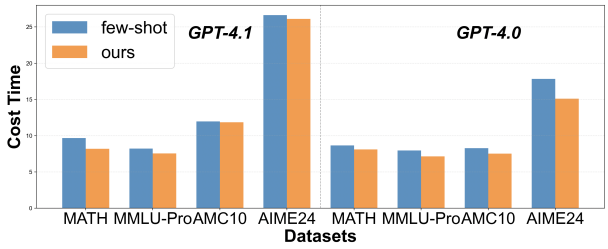

Figure 4: Efficiency analysis comparison between our method and the few-shot CoT baseline on five benchmarks with two models from the GPT series.

5.5 Feasibility and Generalization

The feasibility of our method stems from its ability to effectively bridge the gap between the experiential knowledge actually required and that returned by traditional direct retrieval methods. Furthermore, leveraging "experience" as an intermediate layer for theorem retrieval yields better results than retrieving theorems directly based on questions. While our approach is demonstrated in the context of mathematics, it also holds promise for extension to other reasoning-oriented disciplines. Here, reasoning disciplines refer to fields such as mathematics, physics, and programming, which emphasize logical reasoning processes, whereas disciplines like history, geography, and medicine—which prioritize breadth of knowledge—are categorized as knowledge disciplines. In future system development and research, we intend to incorporate additional reasoning disciplines to further validate and extend the applicability and effectiveness of our method.

6 Conclusion

In this paper, we systematically design a human-inspired graph structure that incorporates both templates and relevant theorems, and propose an automated graph construction pipeline powered by large language models, enabling the expansion of a large-scale knowledge graph from a relatively small set of problems. Additionally, we develop an efficient retrieval mechanism to effectively access relevant information within the constructed graph. Extensive experiments conducted on several widely used benchmarks demonstrate that our proposed method substantially improves the mathematical reasoning performance of large language models. Moreover, our results confirm that, although closed-source large language models possess an abundance of latent knowledge, they often lack a means to effectively leverage it. We also provide a thorough analysis of the scalability and efficacy of our approach. Taken together, our work offers new insights and perspectives for advancing mathematical reasoning and knowledge graph construction with large language models.

Acknowledgements

This work is founded by National Natural Science Foundation of China (NSFC62306276), and Fundamental Research Funds for the Central Universities (226-2023-00138). This work was supported by Ant Group. We would also like to express our sincere gratitude to Wen Zhang for her insightful comments and suggestions.

References

2024. AIME 2024 Dataset. https://huggingface.co/datasets/aime/aime2024. Dataset from the American Invitational Mathematics Examination (AIME) 2024. Accessed: 2024-06-09.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web (ISWC 2007)*, 722–735. Springer.

Besta, M.; Blach, N.; Kubíek, A.; Gerstenberger, R.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Podstawski, M.; Niewiadomski, H.; Nyczyk, P.; and Hoefler, T. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *AAAI Conference on Artificial Intelligence*.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 1247–1250. ACM.

Chen, H.; Shen, X.; Lv, Q.; Wang, J.; Ni, X.; and Ye, J. 2024. SAC-KG: Exploiting Large Language Models as Skilled Automatic Constructors for Domain Knowledge Graph. In *Annual Meeting of the Association for Computational Linguistics*.

Chen, W.; Yin, M.; Ku, M.; Lu, P.; Wan, Y.; Ma, X.; Xu, J.; Wang, X.; and Xia, T. 2023. Mecheng Governance: A Theoretical Framework for Heritage City Community Management. In *Proceedings of the 2023 International Conference on Digital Governance and Cultural Heritage*, 123–135. Hangzhou, China: IEEE.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021a. Training Verifiers to Solve Math Word Problems. *ArXiv*, abs/2110.14168.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cohen, R.; Geva, M.; Berant, J.; and Globerson, A. 2023. Crawling The Internal Knowledge-Base of Language Models. In *Findings*.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.-M.; Zhang, R.; Xu, R.; Zhu, Q.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *ArXiv*, abs/2501.12948.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; et al. 2024. The Llama 3 Herd of Models. *ArXiv*, abs/2407.21783.

Google. 2023. PaLM 2 Technical Report. Technical report, Google. ArXiv:2305.10403.

Han, J.; Collier, N.; Buntine, W. L.; and Shareghi, E. 2023. PiVe: Prompting with Iterative Verification Improving Graph-based Generative Capability of LLMs. In *Annual Meeting of the Association for Computational Linguistics*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.

Huang, H.; Chen, C.; Sheng, Z.; Li, Y.; and Zhang, W. 2024. Can LLMs be Good Graph Judge for Knowledge Graph Construction?

Iuculano, T.; and Menon, V. 2018. Development of Mathematical Reasoning. In Wixted, J. T., ed., *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition*, volume 4, 183–222. Hoboken, NJ: Wiley.

OpenAI. 2024. GPT-4 Technical Report. Technical report, OpenAI. ArXiv:2303.08774.

Robertson, S. E. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Salton, G.; and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sheng, Y.; Li, L.; and Zeng, D. D. 2025. Learning Theorem Rationale for Improving the Mathematical Reasoning Capability of Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15151–15159.

Silver, S.; Katz, E.; and Lesgold, A. 1980. Decision making in complex environments. *Cognitive Science*, 4(3): 367–384.

Team, Q. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

Trajanoska, M.; Stojanov, R.; and Trajanov, D. 2023. Enhancing Knowledge Graph Construction Using Large Language Models. *ArXiv*, abs/2305.04676.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024a. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-first International Conference on Machine Learning (ICML)*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E. H.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ArXiv*, abs/2203.11171.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; Li, T.; Ku, M. W.; Wang, K.; Zhuang, A.; Fan, R. R.; Yue, X.; and Chen, W. 2024b. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *ArXiv*, abs/2406.01574.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E. H.; Xia, F.; Le, Q.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*, abs/2201.11903.

Wu, W.; Jing, Y.; Wang, Y.; Hu, W.; and Tao, D. 2025. Graph-Augmented Reasoning: Evolving Step-by-Step Knowledge Graph Retrieval for LLM Reasoning. *ArXiv*, abs/2503.01642.

Yang, L.; Yu, Z.; Zhang, T.; Cao, S.; Xu, M.; Zhang, W.; Gonzalez, J. E.; and Cui, B. 2024. Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models. *ArXiv*, abs/2406.04271.

Yao, L.; Peng, J.; Mao, C.; and Luo, Y. 2023a. Exploring Large Language Models for Knowledge Graph Completion. *ArXiv*, abs/2308.13916.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023b. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *ArXiv*, abs/2305.10601.

Zhang, B.; Liu, Y.; wen Dong, X.; Zang, Y.; Zhang, P.; Duan, H.; Cao, Y.; Lin, D.; and Wang, J. 2025. BoostStep: Boosting mathematical capability of Large Language Models via improved single-step reasoning. *ArXiv*, abs/2501.03226.

Zhang, Y.; Liu, C.; Wang, R.; Zhao, T.; Li, M.; and Zhou, B. 2023. Mecheng: A Lightweight Framework for Community Governance in Historical Towns. In *NeurIPS 2023 Workshop on Digital Governance in Cultural Heritage*. New Orleans, USA.