

# GraphIC: A Graph-Based In-Context Example Retrieval Model for Multi-Step Reasoning

Anonymous ACL submission

## Abstract

In-context learning (ICL) enhances large language models (LLMs) by incorporating demonstration examples, yet its effectiveness heavily depends on the quality of selected examples. Current methods typically use text embeddings to measure semantic similarity, which often introduces bias in multi-step reasoning tasks. This occurs because text embeddings contain irrelevant semantic information and lack deeper reasoning structures. To address this, we propose **GraphIC**, a graph-based retrieval model that leverages reasoning-aware representation and specialized similarity metric for in-context example retrieval. GraphIC first constructs *thought graphs*—directed, node-attributed graphs that explicitly model reasoning steps and their dependencies—for candidate examples and queries. This approach filters out superficial semantics while preserving essential reasoning processes. Next, GraphIC retrieves examples using a novel similarity metric tailored for these graphs, capturing sequential reasoning patterns and asymmetry between examples. Comprehensive evaluations across mathematical reasoning, code generation, and logical reasoning tasks demonstrate that GraphIC outperforms 10 baseline methods. Our results highlight the importance of reasoning-aware retrieval in ICL, offering a robust solution for enhancing LLM performance in multi-step reasoning scenarios.<sup>1</sup>

## 1 Introduction

In-context learning (ICL) (Brown et al., 2020) allows large language models (LLMs) to adapt to new tasks by incorporating a few demonstration examples within the input prompt, without updating model parameters. However, studies reveal that ICL performance heavily depends on the quality of selected in-context examples (ICEs) (Zhao

et al., 2021; Liu et al., 2022), motivating extensive research on ICE selection strategies. Current approaches (Liu et al., 2022; Rubin et al., 2022) typically use text embeddings to measure semantic similarity between queries and candidate examples, achieving success in semantic-centric tasks like text classification and translation (Agrawal et al., 2023).

However, these text-based methods face significant limitations in multi-step mathematical and logical reasoning tasks. This is because text embedding encodes a substantial amount of shallow semantic information, which is irrelevant to the underlying reasoning processes. This extraneous information introduces bias in the selection of ICEs (An et al., 2023). As shown in Figure 1 (left), when solving a speed calculation problem, text-based methods may retrieve examples about distance/time calculations due to shallow semantic similarity, inducing incorrect reasoning paths (e.g., calculating distance instead of speed). This observation motivates our key insight: Effective ICE selection for reasoning tasks requires representations that explicitly model cognitive processes rather than textual surface forms.

Drawing from cognitive science (Friston, 2008) and graph-based reasoning works (Besta et al., 2024; Yao et al., 2023), we propose *thought graphs*—directed node-attributed graphs where nodes represent reasoning steps and edges denote step dependencies (i.e., a child step can only proceed after the parent step is completed). This representation filters irrelevant semantics while preserving essential reasoning patterns. An example of a thought graph is shown in Figure 2 (a).

Furthermore, we propose a similarity metric for thought graphs tailored to multi-step reasoning tasks. Specifically, we introduce a parameter  $W$  to model the reasoning pattern of a given process. Given two thought graphs,  $G_1$  and  $G_2$ , we first estimate the reasoning pattern parameter  $W_1$  for  $G_1$ ,

<sup>1</sup>Our code is available at <https://anonymous.4open.science/r/GraphIC-2B06/>.

### Question:

Marissa is hiking a 12-mile trail. She took 1 hour to walk the first 4 miles, then another hour to walk the next two miles. If she wants her average speed to be 4 miles per hour, what speed (in miles per hour) does she need to walk the remaining distance?

### Text-based Retrieval:

Q: Jeannie hikes the 12 miles to Mount Overlook at a pace of 4 miles per hour, and then returns at a pace of 6 miles per hour. How long did her hike take, in hours?

A: Up, took  $12 / 4 = 3$  hours.  
Down, took  $12 / 6 = 2$  hours.  
Total time was  $3 + 2 = 5$  hours.

Q: Sadie, Ariana and Sarah are running a relay race. Each part of the race is a different length and covers different terrain. It takes Sadie 2 hours to run through the forest at an average speed of 3 miles per hour. Ariana sprints across the open field on her section at 6 miles per hour for half an hour. If Sarah runs along the beach at four miles per hour and their total time for the race is four and half hours, what is the total distance of the race?

A: Sadie ran for 3 miles/hour \* 2 hour = 6 miles.  
Ariana covered a distance of 6 miles/hour \* (1/2) hour = 3 miles.  
Sarah ran for 4.5 hours - 2 hours - 0.5 hours = 2 hours.  
In this time, Sarah traveled a distance of 4 miles/hour \* 2 hours = 8 miles.  
The total distance for the race was 6 miles + 3 miles + 8 miles = 17 miles.

### Output:

12 miles / 4 mph = 3 hours  
She has already walked for  $1 + 1 = 2$  hours  
She has  $3 - 2 = 1$  hour left to walk the remaining distance  
She has 0 miles left to walk



### Graph-based Retrieval:

Q: Rachel is stuffing envelopes. She has eight hours to complete the task, and there are 1,500 envelopes. In the first hour, Rachel stuffs 135 envelopes. The second hour she stuffs 141 envelopes. How many envelopes will Rachel need to stuff per hour to finish the job?

A: Rachel has  $1500 - 135 - 141$  envelopes = 1224 envelopes remaining to stuff.  
Rachel has 8 hours - 2 hours = 6 hours left to finish the task.  
Rachel needs to stuff  $1224$  envelopes / 6 hours = 204 envelopes per hour.

Q: Allie has 9 toys, which are in total worth \$52. If we know that one toy is worth \$12, and all the other toys have the same value, how much does one of the other toys cost?

A: Allie has  $9 - 1 = 8$  toys of the same value.  
Without the value of the one \$12 toy, all 8 other toys are worth in total  $52 - 12 = \$40$ .  
That would mean, that one of the other toys is worth  $40 / 8 = \$5$ .

### Output:

Marissa has walked  $4 + 2 = 6$  miles so far.  
She has  $12 - 6 = 6$  miles left to walk.  
She wants to walk the entire trail in 12 miles / 4 mph = 3 hours.  
She has already walked for  $1 + 1 = 2$  hours.  
She has  $3 - 2 = 1$  hour left to walk the remaining 6 miles.  
She needs to walk 6 miles / 1 hour = 6 mph.



Figure 1: ICL with different ICE retrieval mechanisms. The left panel shows examples retrieved via BERT embedding (Devlin et al., 2019), while the right panel displays examples retrieved via GraphIC. Semantically related terms are highlighted in blue, and quantities or variables needing resolution are highlighted in green.

then assess its applicability to  $G_2$ . This applicability serves as the similarity measure  $s(G_1, G_2)$ . Our approach naturally captures the sequential structure of reasoning steps and accounts for the inherent asymmetry between examples, which we discuss in detail in Section 3.2.

Building upon these foundations, we introduce GraphIC, a **Graph-based In-Context Example Retrieval Model**. GraphIC achieves reasoning-aware example selection through three key phases: (1) constructing thought graphs for both the query and candidate examples, (2) calculating graph similarity using our proposed metric, and (3) retrieving the top- $k$  most relevant examples. As illustrated in Figure 1 (right), GraphIC effectively identifies examples that align with the reasoning process (such as calculations for the number of envelopes to stuff per hour or unit price), even if they lack semantic similarity to the query (e.g., speed computation). This enables the LLM to solve the problem correctly.

Through comprehensive evaluations across mathematical reasoning, code generation, and logical reasoning tasks, GraphIC demonstrates superior performance over 10 baseline methods including both training-free and training-based approaches. To sum up, our key contributions are: (1) **A representation**. We introduce a novel graph-based representation, called *thought graph*, to model multi-step reasoning processes. This representation effectively filters out irrelevant shallow semantic information while preserving the essential reasoning steps. (2) **A similarity metric**. We introduce a similarity metric for thought graphs tailored to

multi-step reasoning tasks that capture the sequential nature of the steps and the asymmetry between examples. (3) **Empirical validation**. Our experimental results indicate that GraphIC, despite being a training-free model, outperforms both training-free and training-based models across various multi-step reasoning tasks.

## 2 Related Work

Existing ICE selection techniques can be classified as either training-free or training-based, depending on whether a retriever needs to be trained.

Training-free approaches are generally divided into two types: (i) those that use heuristic criteria such as similarity (Liu et al., 2022; Hu et al., 2022), diversity (Cho et al., 2023; Zhang et al., 2022b; Levy et al., 2023; Hongjin et al., 2022; Zhang et al., 2023), complexity (Fu et al., 2022), or combinations of these (Agrawal et al., 2023; Tonglet et al., 2023; Gupta et al., 2023) to select in-context examples (ICEs); (ii) those that leverage feedback from LLMs, such as probability distributions (Wu et al., 2023; Nguyen and Wong, 2023; Li and Qiu, 2023; Yang et al., 2023), perplexity (Gonen et al., 2023), or the model’s generated output (An et al., 2023) to guide the selection process. Training-free approaches eliminate the computational and time costs of model training, but their simpler architecture often leads to lower performance than training-based methods.

Training-based methods are generally divided into two types. The first learns to select individual examples and then extends this to  $k$ -shot scenarios (Rubin et al., 2022; Xiong et al., 2024; Gupta

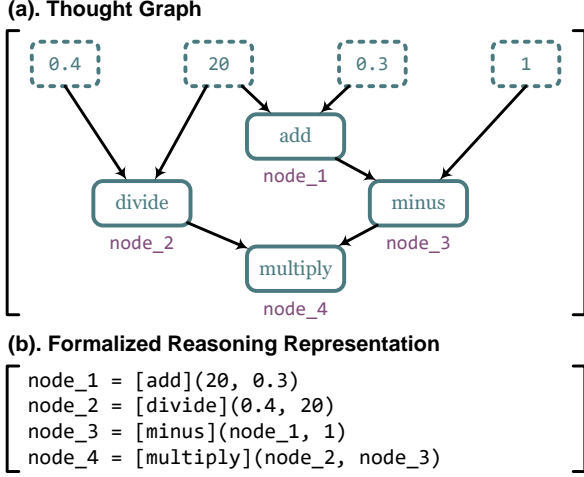


Figure 2: An example of a thought graph (a) and its corresponding FRR (b).

et al., 2024). The second models the selection of a group of examples as a whole (Ye et al., 2023; Wang et al., 2023; Zhang et al., 2022a; Scarlatos and Lan, 2023; Lu et al., 2022; Peng et al., 2023; Xu et al., 2024). Training-based approaches often deliver better performance, but they are computationally expensive and time-consuming.

Our proposed GraphIC involves using LLM-generated output to select ICEs, similar to Skill-kNN and DQ-LoRe. Although methods increase retrieval time due to invoking the LLM during the retrieval, their use of LLMs makes them suitable for more complex tasks.

### 3 The Proposed GraphIC

#### 3.1 Thought Graphs

**Introduction of Thought Graphs.** As outlined earlier, thought graphs are directed, node-attributed graphs designed to model the reasoning process explicitly. In a thought graph, each node  $i$  is attributed with an embedding vector  $x_i \in \mathbb{R}^d$  of a text, representing the key information of the current reasoning step. Directed edges between vertices represent the dependencies between steps. For example, if step B depends on the completion of step A, there is an edge from step A to step B.

As illustrated in Figure 2 (a), the text corresponding to each vertex is "divide", "add", "minus", and "multiply". Therefore, the attributes of each vertex are  $\text{Emb}(\text{"divide"})$ ,  $\text{Emb}(\text{"add"})$ ,  $\text{Emb}(\text{"minus"})$ , and  $\text{Emb}(\text{"multiply"})$ , representing four base operations in a mathematical reasoning task. Here,  $\text{Emb}(\cdot)$  represents calculating the embedding of a given text using an embedding model, such as BERT. In this graph, an edge from the vertex la-

beled "add" to the vertex labeled "multiply" indicates that addition is performed first, and its result is then used in multiplication.

Due to the inherent differences between tasks, the text used to compute embedding varies slightly. Specifically, for mathematical reasoning, code generation, and logical reasoning tasks, we use mathematical operations, code snippets, and intermediate conclusions, respectively, as input texts. This is because, in mathematical reasoning tasks, a reasoning step can be represented by "which mathematical operations were performed on which variables"; in code generation tasks, it corresponds to "what code was executed"; and in logical reasoning tasks, it reflects "which existing conclusions were used to derive new conclusions". Examples of thought graphs for the four benchmark datasets are provided in Appendix A.1.

**Construction of Thought Graphs.** For candidate examples, which include ground-truth natural language reasoning processes (i.e., chain-of-thought answers), we employ the following method to generate thought graphs. First, we use an LLM to formalize the reasoning process for each example. This formalized representation is called the Formalized Reasoning Representation (FRR). FRR filters irrelevant shallow semantics and extracts the key information from the reasoning process. Moreover, FRR is designed to be easily converted to a graph by applying rule-based parsing. The prompts used for generating FRRs and the pseudocode for parsing FRRs are shown in Appendix A.2.

Figure 2 (b) illustrates the FRR of a thought graph. Each line in the FRR follows the format  $A = [B](C)$ , where A represents the name of the vertex, denoting the vertex itself. B represents the label of the vertex, which will be used to compute its embedding. C denotes the parent vertex of the given vertex. For example, in the FRR shown in Figure 2 (b), by parsing the first line, we create node\_1, set its label to "add", and add edges from "20" and "0.3" to node\_1. The remaining lines are parsed similarly to the first one. Note that in Figure 2 (a), the vertices representing numbers are indicated by dashed lines, which signifies that these vertices do not represent specific operations and will be removed once the entire graph is constructed.

For queries, since they lack ground-truth reasoning processes, we first use an LLM to generate pseudo natural language reasoning processes. We then follow the same procedure as with candidate

examples to create thought graphs. Note that it is unnecessary to ensure the correctness of the generated reasoning process, as it is only used to aid in the retrieval of ICEs and is not part of the final answer. In fact, when a generated answer is incorrect, it is often not entirely wrong, but partially correct, which still reflects the reasoning process and can be helpful for retrieving examples. This ensures that GraphIC is robust to inaccuracies in the generated reasoning process. Even when the reasoning process is incorrect, GraphIC still maintains high accuracy. A detailed discussion and related experiments are provided in Appendix A.3.

### 3.2 Similarity Measure for Thought Graphs

In this subsection, we first introduce a simplified case to illustrate our core idea, then extend it to propose our similarity measure.

**Simplified Case.** Consider a thought graph  $G$  with vertices  $v_1, \dots, v_n$ , where each vertex  $v_i$  corresponds to a reasoning step with an attribute vector  $x_i \in \mathbb{R}^d$  (e.g., an embedding of a mathematical operation). Let  $X = [x_1, \dots, x_n] \in \mathbb{R}^{n \times d}$  denote the matrix of concatenated attributes. The adjacency matrix  $A$  of  $G$  is defined such that  $A_{ij} = 1$  if there is a directed edge from  $v_i$  to  $v_j$ , and  $A_{ij} = 0$  otherwise. For a vertex  $v_i$ , let  $\text{pa}(v_i)$  denote the set of its parent vertices (direct predecessors).

Generally, given two thought graphs  $G_1$  and  $G_2$ , the similarity  $s(G_1, G_2)$  is computed in two steps: (1). **Extract Reasoning Pattern:** Solve Equation (1) for  $G_1$  to obtain  $W_1$ , which encapsulates  $G_1$ 's reasoning pattern. (2). **Compute Applicability:** Evaluate how well the reasoning pattern  $W_1$  applies to thought graph  $G_2$ . This applicability, denoted as  $\mathcal{A}(W_1; G_2)$ , serves as the similarity measure  $s(G_1, G_2)$ . We now provide a detailed explanation of these two steps:

Firstly, we derive a matrix  $W$  that captures directional relationships in  $G$ . This is achieved by solving the following optimization problem:

$$W = \arg \max_{\substack{W \in \mathbb{R}^{d \times d}, \\ \|W\|_F = 1}} \mathcal{A}(W; G), \quad \text{where} \quad (1)$$

$$\mathcal{A}(W; G) = \sum_{i=1}^n \underbrace{\left( \sum_{j \in \text{pa}(v_i)} x_j \right)^\top}_{z_i} W x_i.$$

Here, the term  $z_i = \sum_{j \in \text{pa}(v_i)} x_j$  calculates the sum of the attributes of  $v_i$ 's parent nodes, aggregating precursor information for step  $v_i$ .  $z_i^\top W$  represents a linear transformation of this informa-

tion.  $z_i^\top W x_i$  measures the inner product similarity between transformed precursor information  $W z_i$  and the current step's attribute  $x_i$ . By maximizing  $\mathcal{A}(W; G)$  under the Frobenius norm constraint ( $\|W\|_F = 1$ ),  $W$  learns to predict the attribute of the next step based on precursor information, effectively encoding the directional relationships in  $G$ .

For instance, consider a thought graph  $G$  with two vertices:  $v_1$  with attribute  $x_1 = \text{Emb}(\text{"add"})$  and  $v_2$  with  $x_2 = \text{Emb}(\text{"multiply"})$ , connected by an edge  $v_1 \rightarrow v_2$ . In this setup,  $W$  learns to map  $x_1$  ("add") to  $x_2$  ("multiply"), indicating that "multiply" typically follows "add" in  $G$ . When this learned mapping is applied to another graph  $G'$ , a high value of  $\mathcal{A}(W; G')$  suggests that  $G'$  also contains frequent "add"  $\rightarrow$  "multiply" sequences, reflecting similar problem-solving logic. Therefore,  $\mathcal{A}(W; G')$  quantifies how well the reasoning pattern of  $G$  applies to  $G'$ . A higher  $\mathcal{A}(W; G')$  indicates an alignment between the two graphs' transitions, indicating that  $G'$  follows a similar problem-solving trajectory. This metric, therefore, provides an effective measure of the similarity in reasoning between  $G$  and  $G'$ , making it suitable for retrieving ICEs that share compatible problem-solving logic.

**Proposed Similarity Metric.** The similarity metric in GraphIC builds upon the simplified case while introducing two critical enhancements to better model reasoning dynamics and computational efficiency.

Firstly, in Equation (1), the information state  $z_i$  for each node aggregates only its direct parent attributes. However, human reasoning often incorporates information from multiple prior steps and occasionally revisits initial premises. To capture this, we propose an iterative aggregation process:

$$Z^{(h+1)} = [(1 - \lambda)\tilde{A} + \lambda\tilde{B} + I]Z^{(h)}, \quad (2)$$

where  $Z^{(h)}$  represents node information states at iteration  $h$ ;  $Z^{(0)}$  is initialized with node attributes  $X$ ;  $I$  is identity matrix;  $\tilde{A} = D_A^{-\frac{1}{2}} A D_A^{-\frac{1}{2}}$ ,  $D_A = \text{diag}(\deg_{\text{out}}(v_1), \dots, \deg_{\text{out}}(v_n))$ ;  $\tilde{B} = D_B^{-\frac{1}{2}} B D_B^{-\frac{1}{2}}$ , and  $B$  is traceback matrix enabling backtracking to root nodes, defined as  $B_{ij} = 1$  if  $\deg_{\text{in}}(v_j) = 0$  and  $\deg_{\text{in}}(v_i) > 0$ , otherwise  $B_{ij} = 0$ , with  $\deg_{\text{in}}(v_j)$  representing the in-degree of node  $v_j$ . A visual example of matrix  $B$  is presented in Appendix A.4.  $\lambda$  is a balancing hyperparameter ( $0 \leq \lambda \leq 1$ ).

In each iteration,  $z_i$  is updated by combining three information sources, corresponding to the



matrices  $A$ ,  $B$ , and  $I$ . Specifically,  $AZ^{(h)}$  aggregates information from direct parent nodes, similar to the simplified case;  $BZ^{(h)}$  facilitates backtracking in the reasoning process; and  $IZ^{(h)}$  maintains current information.

Secondly, we observe that directly solving the optimization problem in Equation (1) presents challenges related to both uniqueness and computational efficiency. Specifically, the number of vertices  $n$  in the thought graph is typically much smaller than the embedding dimension  $d$  (i.e.,  $n \ll d$ ). For instance, in the GSM8K dataset, thought graphs often contain fewer than 10 vertices, whereas the embedding dimension  $d$  can reach up to 768 when using BERT. This large disparity in dimensions leads to a non-unique solution for  $W$ . Additionally, storing and computing a  $d \times d$  matrix is computationally expensive. To address both the issues of uniqueness and computational efficiency, we constrain  $W$  to be rank 1, allowing it to be expressed as  $W = \alpha\beta^\top$ . As a result,  $\mathcal{A}(W; G)$  simplifies to:

$$\mathcal{A}(W; G) = \mathcal{A}(\alpha, \beta; G) = \sum_{i=1}^n z_i^\top \alpha \beta^\top x_i. \quad (3)$$

Both theoretical analysis and empirical evidence, presented in Appendix A.5, demonstrate that the rank-1 assumption has minimal impact on solving the optimization problem.

Furthermore, with the rank-1 assumption, we can obtain the closed-form solution of Equation (1), as shown in Equation (4), which greatly reduces the computational complexity of solving the optimization problem. The proof is shown in Appendix A.6

$$W = \alpha\beta^\top, \alpha = U[0, :], \beta = V[0, :] \quad (4)$$

where  $U, \Sigma, V = \text{SVD}(X^\top Z)$ .

**Discussion.** Here, we compare proposed similarity with a widely used attributed-graph similarity to highlight two key properties of our metric: encoding directional relationships and asymmetry.

A common attributed-graph similarity metric computes the graph’s embedding through an iterative formula, and then evaluates the similarity between graphs by calculating the cosine similarity of their embeddings. The iterative formula generally takes the following form:

$$X^{(h+1)} = \tilde{A}X^{(h)}, X^{(0)} = X. \quad (5)$$

Firstly, the embedding-based approach fails to effectively capture the sequential relationships between vertices, which are crucial in multi-step reasoning tasks. In these tasks, the order of opera-

tions can significantly alter the reasoning pattern. Therefore, the embedding-based similarity is not well-suited for such tasks. In contrast, our proposed similarity accounts for the transformation of information from previous vertices to current ones, inherently capturing these sequential relationships.

Secondly, our proposed similarity is asymmetric. Specifically, for two thought graphs  $G_1$  and  $G_2$ ,  $s(G_1, G_2) \neq s(G_2, G_1)$ . This asymmetry reflects real-world scenarios. For instance, as illustrated in Appendix A.7, mathematical problem A might be a subproblem of problem B. In such a case, referencing B can help resolve A, but referencing A does not necessarily resolve B.

### 3.3 Example Retrieval

After introducing thought graphs and the proposed similarity metric, we now describe how GraphIC enhances ICL. During the preparation stage, GraphIC generates thought graphs for all candidate examples and estimates their reasoning patterns, denoted as  $W_1, \dots, W_N$ , where  $N$  is the number of candidates. Then, given a query, GraphIC creates a thought graph for the query, denoted as  $G^q$ , computes the applicability of each reasoning pattern to this thought graph,  $\mathcal{A}(W_i, G^q)$ , and then selects the top  $k$  as ICEs.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets and Evaluations.** We conduct experiments across four multi-step reasoning tasks: mathematical reasoning (GSM8K (Cobbe et al., 2021) and AQUA (Ling et al., 2017)), code generation (MBPP (Austin et al., 2021)), and logical reasoning (ProofWriter (Tafjord et al., 2021)). Model performance is measured by answer accuracy for GSM8K, AQUA, and ProofWriter, and by the pass@1 metric (Chen et al., 2021) for MBPP.

**Models and Hyper-parameters.** We use both an open-source and a closed-source model. For the open-source model, we select Llama-3.1-8B-Instruct (hereafter referred to as Llama-3), one of the most advanced 7B-level models. For the closed-source model, we choose GPT-4o-mini, balancing performance and testing costs. Unless explicitly mentioned otherwise, all evaluations use an 8-shot setting, which is the most common setting in chain-of-thought scenarios. We set the temperature to 1e-5 to minimize randomness, ensuring consistency across generations, and use 3 iterations in Equa-

tion (2) with  $\lambda$  values from  $\{0, 0.1, 0.2, 0.3\}$  (see Appendix 5 for specific details). More experiment details are in Appendix B.1.

## 4.2 Baselines

We compare GraphIC against six training-free retrieval methods spanning random, similarity-based, diversity-based, and complexity-based approaches, including: (1) **Random** randomly selects  $k$  unique ICEs from the candidate set; (2) **BM25** (Robertson et al., 2009) selects the top  $k$  examples based on BM25 scoring; (3) **BERT** (Devlin et al., 2019) is a dense retriever using cosine similarity with BERT-base-uncased embeddings; (4) **Complex-CoT** (Fu et al., 2022) selects  $k$  examples based on complexity, quantified by newline characters; (5) **Auto-CoT** (Zhang et al., 2022b) clusters candidates and selects the closests to each cluster center; and (6) **Skill-kNN** (An et al., 2023) prompts LLM to generate task-relevant skills for query and candidates, followed by dense retrieval.

We also compare with four training-based methods, which encompass both single-example and combination-example retrieval strategies, including: (1) **EPR** (Rubin et al., 2022) is trained to retrieve the single most relevant ICE, with top  $k$  examples being selected during inference; (2) **CEIL** (Ye et al., 2023) uses determinantal point processes to select ICEs balancing similarity and diversity; (3) **DQ-LoRe** (Xiong et al., 2024) uses dual queries and low-rank approximation re-ranking to identify ICEs; and (4) **GistScore** (Gupta et al., 2024) encodes task-specific information into gist tokens for selecting ICEs.

## 4.3 Main Results

Table 1 shows the performance comparison results. As a training-free method, GraphIC consistently outperforms both training-free and training-based baselines in most settings.

We find that training-based baselines generally outperform training-free baselines, as they learn explicit patterns from in-context examples. For instance, with Llama-3, training-based methods average over 67% performance, while the top training-free baseline (Complex-CoT) reaches only 66.54%. Among the training-based baselines, DQ-LoRe is the most effective, reaching 68.33%. This method leverages LLM-generated outputs during retrieval, demonstrating that incorporating LLM outputs improves performance on reasoning tasks. Among training-free baselines, Complex-CoT and BM25

perform best. Notably, both use asymmetric retrieval, reinforcing that asymmetric retrieval aligns well with real-world reasoning scenarios.

GraphIC integrates the strengths of existing methods—utilizing LLM outputs for reasoning tasks and asymmetric retrieval—while introducing a reasoning-aware representation and similarity metric. Consequently, it not only surpasses all training-free baselines by (2.57% with GPT-4o-mini and 4.29% with Llama-3), but also outperforms all training-based methods (by 1.18% using GPT-4o-mini and 2.5% using Llama-3), highlighting its effectiveness in reasoning tasks.

A closer analysis highlights GraphIC’s substantial advantages in mathematical and logical reasoning tasks compared to code generation, particularly for complex problem instances. For example, on GSM8K, GraphIC outperforms all baselines by 0.65% and 3.57% with the two LLMs. In the more challenging AQUA dataset, the performance improvements become even more pronounced, reaching 3.47% and 7.64%.

## 4.4 Ablation Study

We conduct a series of ablation studies to systematically evaluate the contribution of each component in GraphIC, focusing on three main aspects: filtering shallow semantic information, integrating the graph structure, and the proposed similarity metric.

To this end, we develop several variants of the GraphIC model: (1) **Text** relies solely on text embedding, the same as the BERT approach; (2) **FRR** retrieves examples using text embeddings of FRRs; (3) **Graph** employs Equation (5) to generate graph embeddings and use cosine similarity to retrieve examples; (4) **Graph+B** employs Equation (2) to generate graph embeddings and use cosine similarity to retrieve examples; (5) **Graph+S** excludes the backtracking mechanism from the full GraphIC model during computing  $Z$ ; and (6) **Graph+B+S** represents the full GraphIC model, integrating all components.

We conduct experiments using Llama-3 across four datasets, with the results presented in Table 2. First, we observe that employing FRR to compute text embeddings significantly improves performance (FRR versus Text), especially on GSM8K and ProofWriter, with performance increasing from 74.15 to 78.31 and from 73.75 to 82.50, respectively. This improvement is due to FRR’s ability to filter out substantial shallow semantic noise. Second, converting FRR into a graph representa-

LLM	Model	GSM8K	AQUA	MBPP	ProofWriter	Avg.
GPT-4o-mini	Random	92.90 (0.31)	71.58 (0.72)	72.76 (0.74)	64.90 (0.93)	75.54 (0.36)
	BM25	92.64	70.47	73.4	66.25	75.69
	BERT	93.02	66.93	74.2	65.25	74.85
	Complex-CoT	92.49	67.32	74.2	64.25	74.57
	Auto-CoT	92.72	69.69	73.8	62.25	74.62
	Skill-kNN	92.34	71.65	72.0	66.00	75.50
	EPR	93.02	72.04	73.8	68.50	76.84
	CEIL	92.57	72.44	73.8	69.50	77.08
	DQ-LoRe	93.32	69.69	74.6	66.50	76.03
	GistScore	93.25	69.69	72.8	67.00	75.69
	GraphIC	<b>93.48</b>	<b>73.62</b>	<b>75.2</b>	<b>70.75</b>	<b>78.26</b>
Llama-3.1 -8B-Instruct	Random	78.86 (0.87)	53.15 (1.85)	57.72 (1.06)	76.10 (2.45)	66.46 (0.84)
	BM25	77.71	46.85	62.0	77.75	66.08
	BERT	74.15	50.39	60.8	73.75	64.77
	Complex-CoT	79.30	50.00	58.6	78.25	66.54
	Auto-CoT	72.78	42.91	58.4	78.00	63.02
	Skill-kNN	77.56	50.39	60.8	74.00	65.69
	EPR	75.66	53.94	62.0	79.25	67.71
	CEIL	75.51	51.97	62.4	81.00	67.72
	DQ-LoRe	77.93	54.33	59.8	81.25	68.33
	GistScore	74.60	44.49	60.4	79.50	64.75
	GraphIC	<b>79.98</b>	<b>57.48</b>	61.6	<b>84.25</b>	<b>70.83</b>

Table 1: Main results on two LLMs and four datasets. For random retrieval, we present the mean and standard deviation derived from five independent experiments.

tion (i.e., a thought graph) further improves performance (Graph versus FRR), as seen on the AQUA dataset, where performance rises from 50.78 to 54.72. This suggests that the graph structure is better suited for capturing the reasoning process than linear text. Third, introducing the reasoning-aware similarity metric also enhances performance (Graph+B+S versus Graph), increasing from 54.72 to 57.48 on AQUA, highlighting the effectiveness of reasoning-aware similarity. Additionally, our experiments show that incorporating the traceback mechanism better simulates the thought process, leading to further performance gains (Graph+B versus Graph, and Graph+B+S versus Graph+S).

In summary, the findings emphasize that each component of GraphIC contributes significantly to improving model performance, with the most substantial gains occurring when all components are combined.

#### 4.5 Analysis

**The performance of GraphIC steadily improves as  $k$  increases.** We analyze how the performance of GraphIC and top training-based/free baselines (DQ-LoRe and Complex-CoT) changes as  $k$  increases. Specifically, we vary the number of examples  $k$  in the set  $\{1, 2, 4, 8\}$  and use Llama-3 as LLM.

Model	GSM8K	AQUA	MBPP	ProofWriter
Text	74.15	50.39	60.8	73.75
FRR	78.31	50.78	60.4	82.50
Graph	78.46	54.72	60.4	83.50
+B	78.92	56.30	61.0	83.75
+S	79.07	49.21	60.4	<b>84.25</b>
+B+S	<b>79.98</b>	<b>57.48</b>	<b>61.6</b>	<b>84.25</b>

Table 2: Ablation Study.

As shown in Figure 3, we observe that model performance generally improves with  $k$ . Notably, GraphIC consistently performs better as  $k$  increases, whereas other methods experience performance degradation. Additionally, training-based and training-free methods show different trends with  $k$  increases. Specifically, training-based methods, such as DQ-LoRe, directly optimize the probability of LLM generating correct answers in 1-shot scenarios. As a result, they tend to achieve superior performance in low-shot settings, especially in 1-shot cases. However, as  $k$  increases, the performance gains of these methods tend to slow down or even decline. In contrast, training-free methods, like Complex-CoT and GraphIC, typically underperform compared to training-based methods in low-shot settings. However, as the value of  $k$  in-

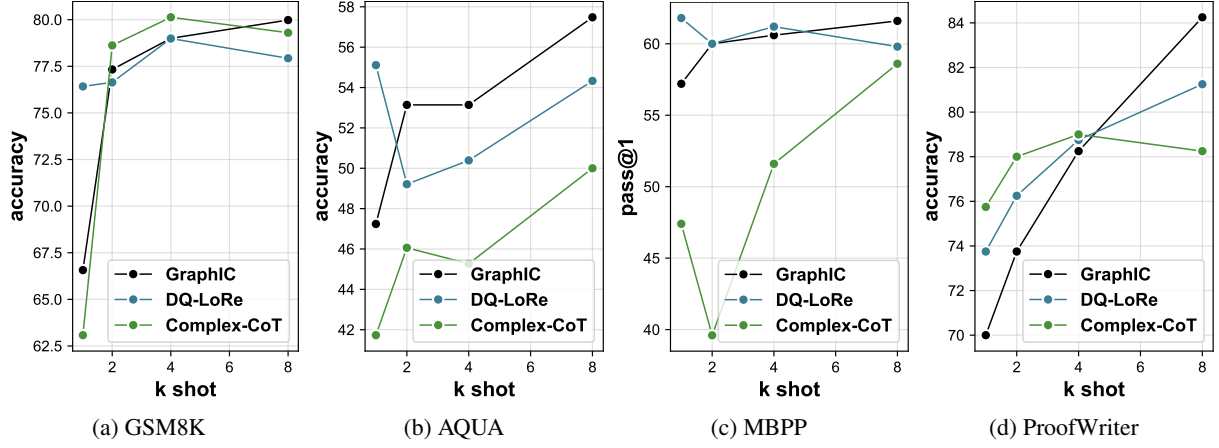


Figure 3: Performance of GraphIC and Top Training-based/Training-free Baselines (DQ-LoRe and Complex-CoT) across 1–8 Shot Settings.

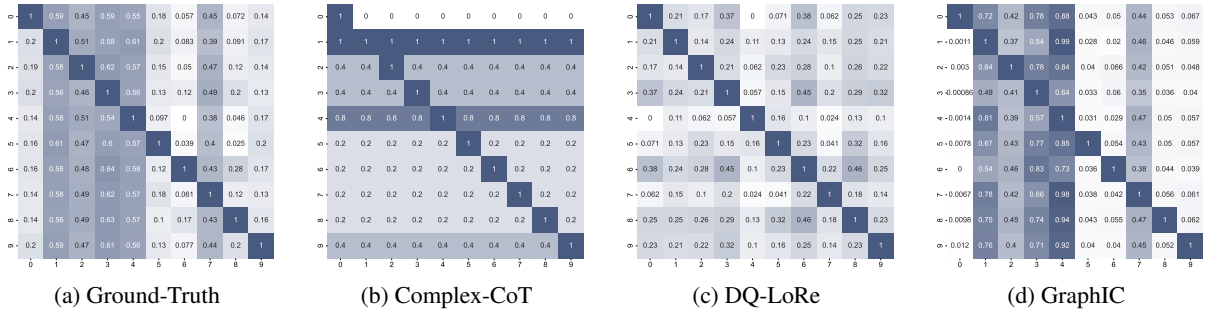


Figure 4: Ground-truth matrix and score matrices of various models. The matrix values have been linearly scaled to the range  $[0,1]$ , and the diagonal elements have been set to 1.

creases, they exhibit more stable improvement.

**Asymmetric retrieval aligns more closely with real-world scenarios.** As previously discussed, GraphIC is an asymmetric method that reflects real-world scenarios. To further validate this asymmetry, we randomly select 10 examples from GSM8K and compute their "improve matrix"  $S$ , which captures the mutual improvement relationships between examples. In this matrix,  $S_{ij}$  is defined as the probability of solving the  $j$ -th example correctly when using the  $i$ -th example as the ICE. The resulting improve matrix (Figure 4 (a)), clearly illustrates the asymmetry in real-world scenarios: while the  $i$ -th example may aid in solving the  $j$ -th example accurately, the reverse is not necessarily true.

Additionally, we examine the alignment between the score matrices of GraphIC and top training-free/based models with the ground-truth improve matrix. Existing ICE retrieval methods select ICEs by assigning scores to (candidate, query) pairs and comparing them. Similar to the improve matrix, we can compute a score matrix for each method, reflecting the mutual improvement relationships between examples estimated by the method. Naturally, the closer a method's score matrix aligns

with the improvement matrix, the better it captures real-world relationships. The score matrices of Complex-CoT, DQ-LoRe, and GraphIC are shown in Figure 4 (b), (c), and (d), respectively. The results clearly demonstrate that GraphIC aligns well with the ground-truth, while the other two models exhibit noticeable deviations.

**Additional Analysis.** We conducted further experiments to examine other aspects of GraphIC, including the accuracy of the generated pseudo reasoning process (Appendix B.3), comparisons of computational efficiency (Appendix B.4), the impact of the hyperparameter  $\lambda$  (Appendix B.5), and the performance on MATH, a more complex mathematical reasoning dataset (Appendix B.6).

## 5 Conclusion

We propose GraphIC, which employs a reasoning-aware representation—*thought graph* and a similarity metric to retrieve ICEs. Our experiments show that GraphIC outperforms both training-free and training-based baselines, highlighting the importance of modeling reasoning structures and providing insights for future research on graph-based solutions for complex problem-solving.



## Limitations

Our work has two limitations. First, the GraphIC approach relies on invoking a large language model (LLM) during the retrieval phase, which leads to slower response times compared to embedding-based methods. Second, the thought graph representation and similarity metric used in GraphIC are tailored for multi-step reasoning tasks and may not perform as effectively in other contexts.

## References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.

Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen, and Jian-Guang Lou. 2023. Skill-based few-shot selection for in-context learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17682–17690.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Hyunsoo Cho, Hyuhng Joon Kim, Junyeob Kim, Sang-woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk

Kim. 2023. Prompt-augmented linear probing: Scaling beyond the limit of few-shot in-context learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12709–12718.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. CodeBERT: A pre-trained model for programming and natural languages. In *The 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1536–1547.

Karl Friston. 2008. Hierarchical models in the brain. *Plos Computational Biology*, 4(11):e1000211.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shivanshu Gupta, Clemens Rosenbaum, and Ethan R Elenberg. 2024. GistScore: Learning better representations for in-context example selection with gist bottlenecks. In *Forty-First International Conference on Machine Learning*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

727	Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu,	Yingzhe Peng, Xu Yang, Haoxuan Ma, Shuo Xu,	784
728	Noah A Smith, and Mari Ostendorf. 2022. In-context	Chi Zhang, Yucheng Han, and Hanwang Zhang.	785
729	learning for few-shot dialogue state tracking. In <u>The</u>	2023. ICD-LM: Configuring vision-language in-	786
730	2022 Conference on Empirical Methods in Natural	context demonstrations by language modeling. <u>arXiv</u>	787
731	Language Processing, pages 2627–2643.	preprint arXiv:2312.10104.	788
732	Omar Khattab, Keshav Santhanam, Xiang Lisa	Stephen Robertson, Hugo Zaragoza, et al. 2009. The	789
733	Li, David Hall, Percy Liang, Christopher Potts,	probabilistic relevance framework: BM25 and be-	790
734	and Matei Zaharia. 2022. Demonstrate-search-	yond. <u>Foundations and Trends® in Information</u>	791
735	predict: Composing retrieval and language mod-	<u>Retrieval</u> , 3(4):333–389.	792
736	els for knowledge-intensive NLP. <u>arXiv preprint</u>	Ohad Rubin, Jonathan Herzig, and Jonathan Berant.	793
737	arXiv:2212.14024.	2022. Learning to retrieve prompts for in-context	794
738	Steven J Leon. 1994. Maximizing bilinear forms sub-	learning. In <u>Proceedings of the 2022 Conference</u>	795
739	ject to linear constraints. <u>Linear Algebra and its</u>	of the North American Chapter of the Association	796
740	<u>Applications</u> , 210:49–58.	for Computational Linguistics: <u>Human Language</u>	797
741	Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Di-	<u>Technologies</u> , pages 2655–2671.	798
742	verse demonstrations improve in-context composi-	Alexander Scarlatos and Andrew Lan. 2023. Ret-	799
743	tional generalization. In <u>Findings of the Association</u>	ICL: Sequential retrieval of in-context examples	800
744	for Computational Linguistics: <u>ACL 2023</u> , pages	with reinforcement learning. <u>arXiv preprint</u>	801
745	1401–1422.	arXiv:2305.14502.	802
746	Xiaonan Li and Xipeng Qiu. 2023. Finding sup-	Oyvind Tafjord, Bhavana Dalvi, and Peter Clark.	803
747	port examples for in-context learning. In <u>The</u>	2021. ProofWriter: Generating implications, proofs,	804
748	2023 Conference on Empirical Methods in Natural	and abductive statements over natural language.	805
749	Language Processing, pages 6219–6235.	In <u>Findings of the Association for Computational</u>	806
750	Hunter Lightman, Vineet Kosaraju, Yuri Burda, Har-	<u>Linguistics: ACL-IJCNLP 2021</u> , pages 3621–3634.	807
751	rison Edwards, Bowen Baker, Teddy Lee, Jan	Jonathan Tonglet, Manon Reusens, Philipp Borchert,	808
752	Leike, John Schulman, Ilya Sutskever, and Karl	and Bart Baesens. 2023. SEER: A knapsack ap-	809
753	Cobbe. 2024. Let’s verify step by step. In	proach to exemplar selection for in-context Hy-	810
754	<u>The Twelfth International Conference on Learning</u>	bridQA. In <u>The 2023 Conference on Empirical</u>	811
755	<u>Representations</u> .	<u>Methods in Natural Language Processing</u> , pages	812
756	Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blun-	13569–13583.	813
757	som. 2017. Program induction by rationale gener-	Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark	814
758	ation: Learning to solve and explain algebraic	Steyvers, and William Yang Wang. 2023. Large lan-	815
759	word problems. In <u>Findings of the Association for</u>	guage models are latent variable models: Explaining	816
760	<u>Computational Linguistics: ACL 2017</u> , pages 158–	and finding good demonstrations for in-context learn-	817
761	167.	ing. In <u>Advances in Neural Information Processing</u>	818
762	Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B	<u>Systems</u> , volume 36, pages 15614–15638.	819
763	Dolan, Lawrence Carin, and Weizhu Chen. 2022.	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	820
764	What makes good in-context examples for GPT-3? In	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	821
765	<u>Proceedings of Deep Learning Inside Out (DeeLIO</u>	et al. 2022. Chain-of-thought prompting elicits rea-	822
766	<u>2022): The 3rd Workshop on Knowledge Extraction</u>	soning in large language models. <u>Advances in neural</u>	823
767	<u>and Integration for Deep Learning Architectures</u> ,	<u>information processing systems</u> , 35:24824–24837.	824
768	pages 100–114.	Zhiyong Wu, Yaoliang Wang, Jiacheng Ye, and Ling-	825
769	Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu,	peng Kong. 2023. Self-adaptive in-context learn-	826
770	Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark,	ing: An information compression perspective for in-	827
771	and Ashwin Kalyan. 2022. Dynamic prompt learning	context example selection and ordering. In <u>Findings</u>	828
772	via policy gradient for semi-structured mathematical	of the Association for Computational Linguistics:	829
773	reasoning. In <u>The Eleventh International Conference</u>	<u>ACL 2023</u> .	830
774	<u>on Learning Representations</u> .	Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang	831
775	Tai Nguyen and Eric Wong. 2023. In-context ex-	Guo, Yichun Yin, Enze Xie, Zhicheng YANG, Qingx-	832
776	ample selection with influences. <u>arXiv preprint</u>	ing Cao, Haiming Wang, Xiongwei Han, Jing Tang,	833
777	arXiv:2302.11042.	Chengming Li, and Xiaodan Liang. 2024. DQ-LoRe:	834
778	Liangming Pan, Alon Albalak, Xinyi Wang, and	Dual queries with low rank approximation re-rank-	835
779	William Yang Wang. 2023. Logic-LM: em-	powering large language models with symbolic	836
780	powering large language models with symbolic	solvers for faithful logical reasoning. In <u>The Twelfth International</u>	837
781	solvers for faithful logical reasoning. In <u>The</u>	<u>Conference on Learning Representations</u> .	
782	2023 Conference on Empirical Methods in Natural	Weijia Xu, Andrzej Banburski, and Nebojsa Jojic. 2024.	838
783	Language Processing.	Reprompting: Automated chain-of-thought prompt	839

inference through gibbs sampling. In Forty-First International Conference on Machine Learning.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In The 2023 Conference on Empirical Methods in Natural Language Processing.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In Advances in Neural Information Processing Systems, volume 36, pages 11809–11822.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In Fortieth International Conference on Machine Learning.

Shaokun Zhang, Xiaobo Xia, Zhaoqing Wang, Ling-Hao Chen, Jiale Liu, Qingyun Wu, and Tongliang Liu. 2023. IDEAL: Influence-driven selective annotations empower in-context learners in large language models. In The Twelfth International Conference on Learning Representations.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022a. Active example selection for in-context learning. In The 2022 Conference on Empirical Methods in Natural Language Processing, pages 9134–9148.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. In The Eleventh International Conference on Learning Representations.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In Thirty-Eighth International Conference on Machine Learning, pages 12697–12706.

## A Supplementary Explanation of GraphIC

### A.1 Examples of Thought Graphs

Figure 5 shows examples of thought graphs from the four datasets we considered. For clarity, the label of each graph corresponds to the text associated with its attribute.

### A.2 Formalized Reasoning Representation

The prompt examples below are used to generate formalized reasoning representations for the four datasets being considered. For the test question, since no answer is provided, we will remove the section of the prompt highlighted in red. This will allow the LLM to generate both the answer and

the formalized reasoning representation simultaneously, from which we can then extract the formalized reasoning representation.

#### GSM8K:

```
Translate the given calculations into code form. Each line of
code MUST follow the format specified below:
output_variable = [description of operation](input_variable_1,
..., input_variable_n)

Q: You can buy 4 apples or 1 watermelon for the same price.
You bought 36 fruits evenly split between oranges,
apples and watermelons, and the price of 1 orange is $0
.50. How much does 1 apple cost if your total bill was
$66?

A: If 36 fruits were evenly split between 3 types of fruits,
then I bought 36/3 = <<36/3=12>>12 units of each
fruit
If 1 orange costs $0.50 then 12 oranges will cost $0.50 * 12 =
$<<0.5*12=6>>6
If my total bill was $66 and I spent $6 on oranges then I
spent $66 - $6 = $<<66-6=60>>60 on the other 2 fruit
types.
Assuming the price of watermelon is W, and knowing that you
can buy 4 apples for the same price and that the price
of one apple is A, then 1W=4A
If we know we bought 12 watermelons and 12 apples for $60,
then we know that $60 = 12W + 12A
Knowing that 1W=4A, then we can convert the above to $60 =
12(4A) + 12A
$60 = 48A + 12A
$60 = <<60=60>>60A
Then we know the price of one apple (A) is $60/60= $
<<60/60=1>>1
#### 1

Code:
total_fruits = 36
types_of_fruits = 3
price_per_orange = 0.50
total_oranges = 12
total_bill = 66
equivalent_apples_for_watermelon = 4
total_apples_and_watermelons = 12
fruits_per_type = [divide](total_fruits, types_of_fruits)
cost_of_oranges = [multiply](total_oranges, price_per_orange)
remaining_budget = [minus](total_bill, cost_of_oranges)
price_per_apple = [construct and solve an equation](
total_apples_and_watermelons,
equivalent_apples_for_watermelon, remaining_budget)

...

Q: {{question}}

A: {{answer}}

Code:
```

#### AQUA:

```
Translate the given calculations into code form. Each line of
code MUST follow the format specified below:
output_variable = [description of operation](input_variable_1,
..., input_variable_n)

Q: In a group of 6 boys and 4 girls, four children are to be
selected. In how many different ways can they be
selected such that at least one boy should be there?
Options: A)209, B)210, C)211, D)213, E)215

A: To determine the number of ways to select 4 children from a
group of 6 boys and 4 girls such that at least one boy
is included, we will use the method of complement
counting.

First, let's calculate the total number of ways to select 4
children from 10 children (6 boys + 4 girls):

\[\frac{10!}{4!(10-4)!} = \frac{10!}{4! \times 6!} = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = 210\]
```

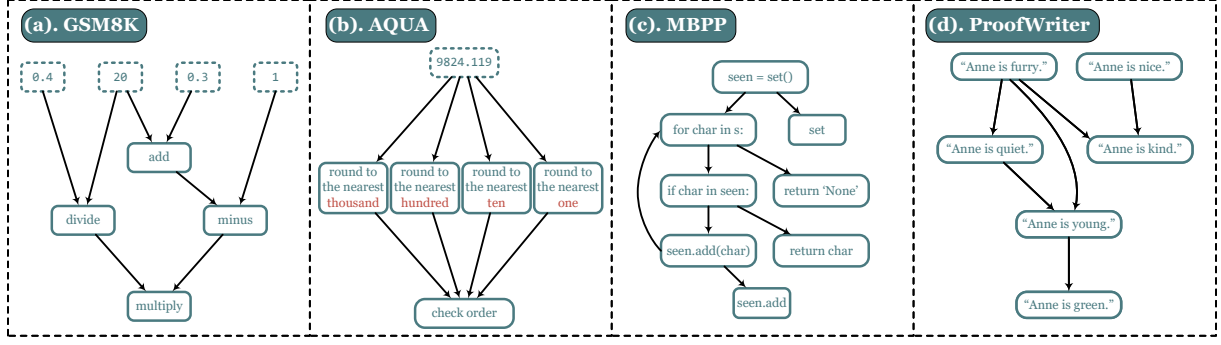


Figure 5: Examples of thought graphs for the four datasets we consider. The labels of vertices represent the text for computing embeddings. Dashed-line vertices are included for clarity but are not part of the actual thought graph.

```

Next, we calculate the number of ways to select 4 children
with no boys, i.e., all girls. Since there are only 4
girls, and we need to select all 4 of them:

\[
\binom{4}{4} = 1
\]

Now, subtract the number of ways to select all girls from the
total number of ways to select 4 children to find the
number of ways that include at least one boy:

\[
\binom{10}{4} - \binom{4}{4} = 210 - 1 = 209
\]

Thus, the number of ways to select 4 children with at least
one boy is:

\[
\boxed{209}
\]

#### A

Code:
total_children = 10
children_to_select = 4
boys = 6
girls = 4
total_ways_to_select = [combination](total_children,
    children_to_select)
all_girls_selection = [combination](girls, children_to_select)
ways_with_at_least_one_boy = [subtract](total_ways_to_select,
    all_girls_selection)

...

Q: {{question}}
A: {{answer}}
Code:

```

### MBPP:

```

Text: Write a function to generate a square matrix filled with
elements from 1 to n raised to the power of 2 in spiral
order.

Test Cases:
assert generate_matrix(3)==[[1, 2, 3], [8, 9, 4], [7, 6, 5]]
assert generate_matrix(2)==[[1,2],[4,3]]
assert generate_matrix(7)==[[1, 2, 3, 4, 5, 6, 7], [24, 25, 26,
    27, 28, 29, 8], [23, 40, 41, 42, 43, 30, 9], [22, 39,
    48, 49, 44, 31, 10], [21, 38, 47, 46, 45, 32, 11], [20,
    37, 36, 35, 34, 33, 12], [19, 18, 17, 16, 15, 14, 13]]

Code:
def generate_matrix(n):
    if n<=0:
        return []
    matrix=[row[:] for row in [[0]*n*n]
    row_st=0
    row_ed=n-1

```

```

col_st=0
col_ed=n-1
current=1
while (True):
    if current>n*n:
        break
    for c in range (col_st, col_ed+1):
        matrix[row_st][c]=current
        current+=1
    row_st+=1
    for r in range (row_st, row_ed+1):
        matrix[r][col_ed]=current
        current+=1
    col_ed-=1
    for c in range (col_ed, col_st-1, -1):
        matrix[row_ed][c]=current
        current+=1
    row_ed-=1
    for r in range (row_ed, row_st-1, -1):
        matrix[r][col_st]=current
        current+=1
    col_st+=1
return matrix

...

Text: {{question}}

Test Cases:
{{test_list}}

Code:

```

### ProofWriter:

```

Formalize the following proof process based on the given
information. Where Each line of formalized proof MUST
follow the format specified below.
output_triple = [rule used](input_triple_1, ... ,
    input_triple_n)

Q: Triples:
1. Anne is not big.
2. Anne is cold.
3. Anne is red.
4. Dave is green.
5. Dave is rough.
6. Erin is green.
7. Erin is kind.
8. Erin is rough.
9. Fiona is green.
10. Fiona is not nice.
Rules:
1. If Erin is cold then Erin is rough.
2. If something is rough then it is nice.
3. All green, big things are kind.
4. If Dave is kind then Dave is cold.
5. If something is green and not rough then it is big.
6. All nice, rough things are big.
7. If Dave is cold and Dave is nice then Dave is red.
Based on the above information, is the following statement
true or false? Dave is red.
A) True B) False

```



Proof: To determine if the statement "Dave is red" is true or false, let's evaluate the given information step by step:

From the triples:

1. Dave is green.
2. Dave is rough.

From the rules:

1. If something is rough then it is nice. (Dave is rough, so Dave is nice.)
2. If Dave is cold and Dave is nice then Dave is red. (However, there's no information saying Dave is cold, so we can't use this rule directly.)
3. All nice, rough things are big. (Dave is nice and rough, so Dave is big.)

Now, we need to verify whether Dave being cold can be derived:

1. If Dave is kind then Dave is cold. (To use this, we need to verify if Dave is kind.)
  - All green, big things are kind. (Dave is green and big, so Dave is kind.)
  - Hence, Dave is kind, which according to rule 4, makes Dave cold.

Since now we know:

- Dave is cold.
- Dave is nice.

We can use rule 7:

- If Dave is cold and Dave is nice then Dave is red.

Since both conditions are satisfied, Dave is red.

The statement "Dave is red" is **True** (A).

Formalized Proof:

```
"Dave is nice." = [rule2]("Dave is rough.")
"Dave is big." = [rule6]("Dave is nice.", "Dave is rough.")
"Dave is kind." = [rule3]("Dave is green.", "Dave is big.")
"Dave is cold." = [rule4]("Dave is kind.")
"Dave is red." = [rule7]("Dave is cold.", "Dave is nice.")
...
```

Proof: `{{answer}}`

Formalized Proof:

The pseudo-code of a parser that transforms formalized reasoning representations into a thought graph is provided in Algorithm 1. *Inputs*, *Output*, and *OperationName* are extracted following the pattern outlined below.

```
Output = [OperationName](input_1, ..., input_n)
```

### A.3 Robustness to the Correctness of Pseudo Reasoning Process

As mentioned in Section 3.1, ensuring the correctness of the generated pseudo-reasoning process is not required. However, we have the following reasons for ensuring that GraphIC maintains high performance even when the generated process is incorrect.

1. Firstly, even if the generated answers are not correct, it does not necessarily mean that the reasoning steps are entirely incorrect; rather, many are "almost correct," involving only minor errors (Wei et al., 2022). This is also helpful for retrieving examples with similar reasoning processes.

### Algorithm 1 Parsing Formalized Reasoning Representation

**Require:** formalized reasoning representation *FRR*

**Ensure:** Corresponding graph  $G(V, E)$

```
1: NodeSet  $\leftarrow \emptyset$ 
2: EdgeSet  $\leftarrow \emptyset$ 
3: line  $\leftarrow$  first line of FRR
4: while line  $\neq$  NULL do
5:   Extract Inputs, Output, and OperationName from line
6:   for each input in Inputs do
7:     if input  $\notin$  NodeSet then
8:       Add input to V
9:     end if
10:  end for
11:  if Output  $\notin$  V then
12:    Add Output to V, labeled as OperationName
13:  end if
14:  for each input in Inputs do
15:    Add directed edge from input to Output to E
16:  end for
17:  line  $\leftarrow$  next line of FRR
18: end while
19:  $G = G(V, E)$ 
```

2. To further illustrate this fact, we selected a subset from each dataset, containing all queries associated with incorrect pseudo answers. We evaluated GraphIC on these four subsets and compared its performance with top training-based (DQ-LoRe) and training-free (Complex-CoT) baselines. The results, shown in Table 3, reveal that even when GraphIC uses incorrect thought graphs to retrieve in-context examples, it still achieves a significant performance advantage. Note that the performance in Table 3 are substantially lower than those in Table 1. This is because the queries that lead to incorrect thought graphs are typically the most difficult ones.

3. Similar approaches have been applied in other ICE retrieval models and show good performance, such as Skill-kNN, which utilizes an LLM to generate the skills required to solve a problem, and DQ-LoRe, which uses an LLM to generate chain-of-thought reasoning answers.

Model	GSM8K	AQUA	MBPP	ProofWriter
Complex-CoT	38.58	28.68	4.06	54.02
DQ-LoRe	40.83	<b>33.33</b>	<b>16.75</b>	65.51
GraphIC	<b>43.08</b>	<b>33.33</b>	15.23	<b>70.11</b>

Table 3: Performance of GraphIC and top training-free/based baseline on the subset where GraphIC uses incorrect answers for creating thought graphs.

#### A.4 The Matrix $B$

As shown in Figure 6, matrix  $A$  represents the adjacency matrix of a thought graph, corresponding to the black edges in the graph. In this graph, vertices 1 and 2 have an in-degree of 0, indicating the starting points of reasoning, while vertices 3 and 4 have non-zero in-degrees, representing intermediate steps or results of reasoning. Matrix  $B$  indicates the edges from vertices with non-zero in-degrees to those with in-degree 0, corresponding to the edges from vertices 3 and 4 to vertices 1 and 2 in the graph (marked in green). Since vertices 3 and 4 are intermediate steps or results of reasoning, and vertices 1 and 2 represent the starting points, these edges represent the retracing process.

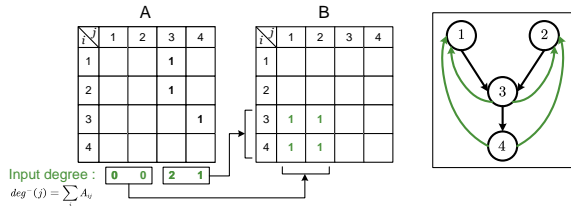


Figure 6: The definition of matrix  $B$  and its corresponding edges in a graph.

#### A.5 The Loss of Setting the Rank of Matrix $W$ to 1

First, we provide the optimal value and the optimal solution of the optimization problem defined by Equation (1), under the assumption that no constraints are imposed on  $W$ .

**Theorem 1** Consider the following optimization problem:

$$\max_{\substack{W \in \mathbb{R}^{d \times d} \\ \|W\|_F = 1}} \sum_{i=1}^n z_i^\top W x_i. \quad (6)$$

The optimal value of this problem is  $\left(\sum_{i=1}^d \sigma_i^2\right)^{1/2}$ , and the optimal solution is  $W^* = VYU^\top$ , where  $Y = \text{diag}(y_{11}, y_{22}, \dots, y_{dd})$  with  $y_{ii} = \frac{\sigma_i}{(\sum_{j=1}^d \sigma_j^2)^{1/2}}$ .

First, the objective function  $\sum_{i=1}^n z_i^\top W x_i$  can be written as  $\text{tr}(ZW^\top X^\top)$ .

Then, we rewrite the objective function using the cyclic property of the trace:

$$\text{tr}(ZW^\top X^\top) = \text{tr}(W^\top X^\top Z).$$

Next, perform the Singular Value Decomposition (SVD) of  $X^\top Z$ :

$$X^\top Z = U\Sigma V^\top,$$

where  $U$  and  $V$  are orthogonal matrices, and  $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_d)$  is the diagonal matrix of singular values.

Substituting this decomposition into the trace expression, we obtain:

$$\text{tr}(ZW^\top X^\top) = \text{tr}(W^\top U\Sigma V^\top).$$

Using the cyclic property of the trace again, we get:

$$\text{tr}(ZW^\top X^\top) = \text{tr}\left((V^\top W^\top U)\Sigma\right).$$

Define  $Y = V^\top W^\top U$ . Then the objective becomes:

$$\text{tr}(ZW^\top X^\top) = \text{tr}(Y\Sigma).$$

Since  $U$  and  $V$  are orthogonal, they preserve the Frobenius norm, i.e.,  $\|Y\|_F = \|W\|_F = 1$ . Thus, we are now tasked with maximizing  $\text{tr}(Y\Sigma)$  subject to  $\|Y\|_F = 1$ .

We know that the trace  $\text{tr}(Y\Sigma)$  is maximized when  $Y$  is diagonal, i.e., when  $Y = \text{diag}(y_{11}, y_{22}, \dots, y_{dd})$ . This follows because the off-diagonal elements of  $Y$  do not contribute to  $\text{tr}(Y\Sigma)$ , but they affect the Frobenius norm of  $Y$ . By setting these off-diagonal elements to zero and redistributing the weight to the diagonal elements, we achieve a higher value of  $\text{tr}(Y\Sigma)$ .

Thus, the optimization problem reduces to:

$$\text{tr}(ZW^\top X^\top) = \sum_{i=1}^d \sigma_i y_{ii}, \text{ s.t. } \sum_{i=1}^d y_{ii}^2 = 1.$$

To find the optimal  $y_{ii}$ , we apply Cauchy-Schwarz inequality:

$$\sum_{i=1}^d \sigma_i y_{ii} \leq \left(\sum_{i=1}^d \sigma_i^2\right)^{1/2} \left(\sum_{i=1}^d y_{ii}^2\right)^{1/2}.$$

Since  $\sum_{i=1}^d y_{ii}^2 = 1$ , this simplifies to:

$$\sum_{i=1}^d \sigma_i y_{ii} \leq \left(\sum_{i=1}^d \sigma_i^2\right)^{1/2}.$$

The maximum value of  $\sum_{i=1}^d \sigma_i y_{ii}$  is achieved when  $y_{ii} = \frac{\sigma_i}{(\sum_{j=1}^d \sigma_j^2)^{1/2}}$ . Thus, the optimal value

of the objective is  $\left(\sum_{i=1}^d \sigma_i^2\right)^{1/2}$ , and the corresponding optimal solution is  $W^* = VYU^\top$ , where

$$Y = \text{diag} \left( \frac{\sigma_1}{(\sum_{i=1}^d \sigma_i^2)^{1/2}}, \dots, \frac{\sigma_d}{(\sum_{i=1}^d \sigma_i^2)^{1/2}} \right).$$

Based on the proof above, we know that the optimal value of the optimization problem defined by Equation (1) is  $(\sum_{i=1}^d \sigma_i^2)^{1/2}$ . When a rank-1 constraint is added to  $W$ , the optimal value of the problem becomes  $\sigma_1$ . Therefore, we are interested in quantifying the loss introduced by the "rank-1 assumption," which can be assessed by the ratio of their optimal values,  $r = \sigma_1 / (\sum_{i=1}^d \sigma_i^2)^{1/2}$ . We computed the value of  $r$  on four datasets, and the results are shown in Table 4. The table demonstrates that the loss caused by the "rank-1 assumption" is less than 0.1%, implying that it does not result in a significant loss of precision.

Dataset	GSM8K	AQUA	MBPP	ProofWriter
$r$	0.99978	0.99991	0.99963	0.99921

Table 4: The  $r$  value on four datasets.

## A.6 Closed Form Solution of Equation (1) with Rank-1 Assumption

Under the rank-1 assumption, the optimization problem can be expressed as

$$\max_{\substack{\alpha, \beta \in \mathbb{R}^d, \\ \|\alpha\|_2 = \|\beta\|_2 = 1}} \sum_{i=1}^n z_i^\top \alpha \beta^\top x_i.$$

This can be rewritten using the trace operator as

$$\text{tr}(Z\beta\alpha^\top X^\top) = \text{tr}(\alpha^\top X^\top Z\beta) = \alpha^\top X^\top Z\beta.$$

Thus, the problem becomes a bilinear maximization:

$$\max_{\substack{\alpha, \beta \in \mathbb{R}^d, \\ \|\alpha\|_2 = \|\beta\|_2 = 1}} \alpha^\top X^\top Z\beta.$$

The closed-form solution to this problem, as derived by (Leon, 1994), is:

$$\alpha = U[0, :], \quad \beta = V[0, :],$$

where  $U, \Sigma, V$  are the singular value decomposition (SVD) of  $X^\top Z$ .

## A.7 The Asymmetry

We provide an example to illustrate the asymmetry. As shown in the example below, solving Question B includes solving Question A, which involves further calculations to determine how long it will take to reach the minimum age required by the company for employment. Therefore, in this case, referencing B can help resolve A, but referencing A does not necessarily resolve B.

Question A:  
In 3 years, Jayden will be half of Ernesto's age. If Ernesto is 11 years old, how many years old is Jayden now?

Answer A:  
Ernesto = 11 + 3 = 14  
Jayden = 14/2 = 7 in 3 years  
Now = 7 - 3 = 4  
Jayden is 4 years old.

Question B:  
The minimum age required to be employed at a company is 25 years. Dara aspires to work for the company and will be half the age of Jane in six years. If Jane is currently working for the company and is 28 years old, how long is it before Dara reaches the minimum age required by the company to be employed?

Answer B:  
In six years, Jane will be 28+6 = 34 years old.  
Dara will be half the age of Jane in six years, meaning she will be 34/2 = 17 years old in six years.  
Currently, Dara is 17-6 = 11 years old.  
Dara has to wait for 25-11 = 14 more years to reach the company's minimum age of employment.

## B Supplementary Materials of Experiments

### B.1 Further Experiment Details

**Dataset Preprocessing.** For both GSM8K and MBPP, we utilize the original datasets without further preprocessing. For AQUA and ProofWriter, we refine the original dataset to improve the experimental setup, as described below.

Given the substantial size of the AQUA dataset, which incurs significant retrieval overhead during testing, we followed the methodology outlined in DQ-LoRe (Xiong et al., 2024), using a 1,000-sample subset for efficient evaluation.

For the ProofWriter dataset, we refined the subset selected by Logic-LM (Pan et al., 2023), excluding instances labeled as "Unknown," as these samples lacked explicit reasoning chains. Furthermore, because the original training set did not provide reasoning in natural language, we leveraged the GPT-4o-mini model to generate reasoning sequences for the training set, discarding any generated outputs deemed incorrect. We evaluate the correctness of the reasoning process by the correctness of the final result, which is a commonly used approach (Lightman et al., 2024; Xiong et al., 2024; Khattab et al., 2022). This process resulted in a refined training set of 1,358 examples with their Chains of Thought and 400 test samples from the original ProofWriter dataset.

**Thought Graph Construction and Embedding Model.** For GSM8K, AQUA, and ProofWriter, we prompt the LLM to create a FRR for thought graph construction, using vertex features from a

BERT model. For MBPP, we use the `staticfg`<sup>2</sup> module to parse Python code and generate the control flow graph, embedding each vertex’s features with CodeBERT (Feng et al., 2020).

## B.2 Values of $\lambda$

We select hyper parameter  $\lambda$  values from  $\{0, 0.1, 0.2, 0.3\}$ , and report the  $\lambda$  values chosen on various datasets and LLMs in Table 5.

Engine	GSM8K	AQUA	MBPP	ProofWriter
GPT-4o-mini	0.2	0.2	0.1	0.1
Llama-3	0.3	0.2	0.2	0.0
GPT-3.5-Turbo	0.3	/	/	/

Table 5:  $\lambda$  values chosen on various datasets and LLMs.

## B.3 Correctness of LLM Generated Answers for Creating Thought Graphs

To analyze the consistency between LLM-generated answers and real solutions, we tested the accuracy of these answers used to generate the thought graphs. The results are shown in the Table 6. From Table 6, it can be seen that these answers used to generate the thought graph already have relatively high accuracy, which ensures their consistency with the real solution. Furthermore, the table demonstrates that using the thought graph to retrieve examples can further improve accuracy, especially in mathematical reasoning and logical reasoning tasks. We use Llama-3.1-8B-Instruct for testing.

Accuracy	GSM8K	AQUA	MBPP	ProofWriter
Pseudo Answer	76.42	49.21	60.6	78.25
Final Answers	79.98	57.48	61.6	84.25
Improvement	+3.56	+8.27	+1.00	+6.00

Table 6: Correctness of LLM generated answers for creating thought graphs and final answers.

## B.4 Comparison of Computation Time with Other Similar Baselines.

Our method belongs to the category of methods that use the generated output of LLMs to select in-context examples, which also includes methods such as Skill-kNN and DQ-LoRe. These approaches involve using the LLM during the retrieval phase, resulting in longer retrieval times compared to other baselines. However, by leveraging the power of LLMs, they are suitable for complex tasks. The computation times for the three

models are presented in Table 7. Specifically, the retrieval time for the GraphIC model is similar to that of DQ-LoRe, and slightly higher than Skill-kNN. Despite this, GraphIC significantly outperforms Skill-kNN in terms of performance. Moreover, compared to DQ-LoRe, which has the same retrieval time, GraphIC not only delivers superior performance but also greatly reduces both the prepare time and training time required by DQ-LoRe.

Here, "prepare time" refers to the time spent generating the necessary outputs for retrieval, such as generating the required skills for all candidate examples in Skill-kNN. For our evaluation, we used the GSM8K dataset with the LLM configured as Llama-3.1-8B-Instruct.

Time	Skill-kNN	DQ-LoRe	GraphIC
prepare time	0.7h	20h	1.5h
training time	-	16h	-
retrieve time	0.3s	0.4s	0.4s

Table 7: Prepare time, training time, and retrieve time of GraphIC and other similar baselines.

## B.5 The Effects of $\lambda$

We analyzed the effect of  $\lambda$  values ranging from  $[0.0, 0.9]$  on the results across the four datasets utilized in our study. The corresponding results are presented in Figure 7, confirming the robustness of our method to the choice of lambda.

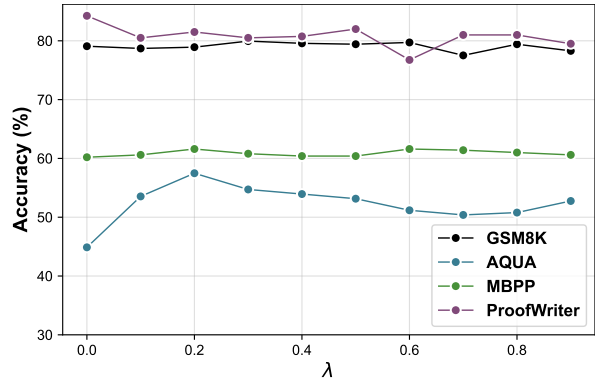


Figure 7: The performance of GraphIC on different  $\lambda$  across four datasets.

## B.6 The Performance on MATH Dataset

We conducted a comparison of GraphIC with the top-performing training-based and training-free baselines (DQ-LoRe and Complex-CoT) on MATH (Hendrycks et al., 2021) dataset. The results are presented in Table 8: the GraphIC model has consistently achieved optimal performance. We ran

<sup>2</sup><https://github.com/coetaur0/staticfg>



domly selected 500 samples from the training and testing data categorized as "level 5" difficulty in MATH, which were used as the candidate set and test set.

Model	Complex-CoT	DQ-LoRe	GraphIC
Llama-3	18.8	20.4	<b>21.8</b>
GPT-4o-mini	58.8	59.6	<b>61.2</b>

Table 8: The Performance on MATH Dataset

### B.7 The Performance on GSM8K Using GPT-3.5-Turbo

We observe that the GPT-4o-mini model’s performance on the GSM8K dataset is relatively invariant to the selection of ICEs. So, we further perform an additional experiment on the GSM8K dataset using the GPT-3.5-Turbo model. As presented in Table 9, GraphIC achieves superior performance across all metrics.

Random	BM25	BERT	Complex-CoT	Auto-CoT	Skill-kNN	EPR	CEIL	DQ-LoRe	GistScore	GraphIC
80.76(0.55)	<u>82.10</u>	80.89	81.65	82.03	81.50	81.65	81.72	<u>82.10</u>	81.72	<b>82.79</b>

Table 9: Results obtained using GPT-3.5-Turbo as the LLM on the GSM8K dataset.