KG-CoT: Chain-of-Thought Prompting of Large Language Models over Knowledge Graphs for Knowledge-Aware Question Answering

Ruilin Zhao[1], **Feng Zhao**[1*], **Long Wang**[1], **Xianzhi Wang**[2] and **Guandong Xu**[2,3]
[1]Natural Language Processing and Knowledge Graph Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[2]Data Science and Machine Intelligence Lab, University of Technology Sydney, Sydney, Australia
[3]The Education University of Hong Kong, Hong Kong, China
{ruilinzhao, zhaof, m202273788}@hust.edu.cn, xianzhi.wang@uts.edu.au, gdxu@eduhk.hk

Abstract

Large language models (LLMs) encounter challenges such as hallucination and factual errors in knowledge-intensive tasks. One the one hand, LLMs sometimes struggle to generate reliable answers based on the black-box parametric knowledge, due to the lack of responsible knowledge. Moreover, fragmented knowledge facts extracted by knowledge retrievers fail to provide explicit and coherent reasoning paths for improving LLM reasoning. To address these challenges, we propose KG-CoT, a novel knowledge-augmented paradigm that leverages a small-scale step-by-step graph reasoning model to reason over knowledge graphs (KGs) and utilizes a reasoning path generation method to generate chains of knowledge with high confidence for large-scale LLMs. Extensive experiments demonstrate that our KG-CoT significantly improves the performance of LLMs on knowledge-intensive question answering tasks, such as multi-hop, single-hop, and open-domain question answering benchmarks, without fine-tuning LLMs. Moreover, KG-CoT can reduce the number of API calls and cost and can generalize to various LLMs in a lightweight plug-and-play manner.

1 Introduction

Recently, large language models (LLMs) [OpenAI, 2023] have achieved remarkable results in various downstream natural language understanding and generating tasks with elaborate natural language prompts [Wei *et al.*, 2022]. Despite their significant ability to generate fluent and coherent natural language responses, LLMs suffer from hallucination and factual errors when performing knowledge-intensive tasks [Ji *et al.*, 2023]. The essential reason for these problems lies in the black-box nature of parametric knowledge, which makes it difficult to locate and update knowledge facts stored in the parameters [Zhu *et al.*, 2020; Cao *et al.*, 2021]. This results in two main challenges that hinder the adaptation of LLMs in knowledge-intensive tasks.

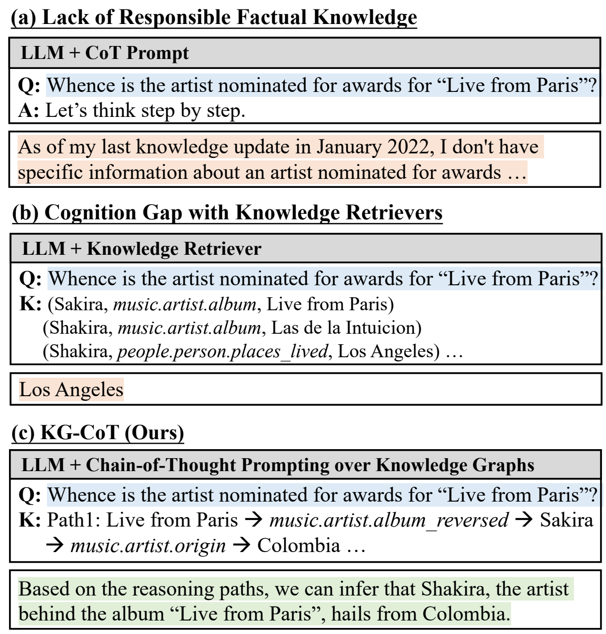

Figure 1: (a) LLMs may struggle to provide responsible answer based on the static parametric knowledge. (b) The high relevance of fragmented knowledge facts doesn't necessarily imply the usefulness for LLM reasoning. (c) Our proposed KG-CoT enables LLMs to think with KGs for knowledge-aware reasoning.

Challenge 1: Lack of Responsible Factual Knowledge. Since it is challenging to revise and expand the parametric knowledge, LLMs are hard to access the most recent updates in various domains [Wang *et al.*, 2023d]. Therefore, when encountering questions that require up-to-date or domain-specific knowledge, LLMs may struggle to provide responsible answers based on the static parametric knowledge [Chen *et al.*, 2023]. Although elaborate prompts [Wei *et al.*, 2022; Yao *et al.*, 2023] can be used to decompose complex questions into multiple steps to enhance the logical reasoning capability of LLMs, it is difficult to fully compensate for the lack of explicit factual knowledge. As a result, the benefit of elaborate prompting diminishes [Wang *et al.*, 2023a] especially in tasks where accurate and deep understanding of subject entity is crucial for generating correct response.

*Corresponding author

Challenge 2: Cognition Gap with Knowledge Retrievers. Augmenting LLMs with external knowledge graphs is a natural and promising solution for addressing the lack of knowledge described above [Bollacker *et al.*, 2007]. KGs are structured, explicit, and responsible, which can provide reliable knowledge subgraphs to explicitly enhance the knowledge-aware reasoning process of LLMs [Shi *et al.*, 2021]. However, the cognition gap in understanding and reasoning between LLMs and knowledge retrievers significantly limits the performance of LLM+KG paradigm. Knowledge retrievers prioritize knowledge facts commonly based on representation similarity [Li *et al.*, 2023], but the relevance in this context does not necessarily guarantee usefulness for specific reasoning tasks of LLMs [Sun *et al.*, 2023]. This cognition gap results in LLMs being compelled to continuously evaluate the usefulness of fragmented knowledge facts and recurrently invoke knowledge retrievers to provide adequate knowledge for reasoning [Sun *et al.*, 2023]. This leads to a significant increase in the complexity and cost of the LLM+KG paradigm.

To address these challenges, we propose a Chain-of-Thought prompting over Knowledge Graphs (KG-CoT), a novel knowledge-augmented framework that utilizes a step-by-step graph reasoning model to augment LLMs with responsible chains of knowledge in a plug-and-play manner. **To address the lack of responsible factual knowledge *(Challenge 1)***, we propose a step-by-step graph reasoning model to reason over KGs. Starting from the question entity, the step-by-step graph reasoning model calculates scores for relations in a KG and constructs the transition matrix for each reasoning step. By utilizing the transition matrix, the graph reasoning model can traverse various paths in the KG, hopping among relations and exploring entities with high confidence for problem solving. **To address the cognition gap between LLMs and knowledge retrievers *(Challenge 2)***, we develop a reasoning path generation method. Starting from the question entity, it retraces the step-by-step reasoning process and generates explicit reasoning paths along the transition matrix. In this way, the graph reasoning model can plug into LLMs and enable joint reasoning of LLMs over KGs.

Our main contributions are as follows:

- **Large + Small:** We propose a knowledge-augmentation paradigm for LLMs that combines large-scale LLMs with small-scale step-by-step graph reasoning models to augment LLMs with KGs without fine-tuning LLMs.
- **Responsibility:** We propose using a graph reasoning model over KGs as a enhancement of CoT prompting to generate responsible chains of knowledge for improving knowledge-aware reasoning capability of LLMs.
- **Efficiency:** Our proposed KG-CoT prompting significantly improves the performance of LLMs on several knowledge-intensive benchmarks without fine-tuning LLMs, and outperforms prior retrieval-augmented and knowledge base question answering baselines.
- **Adaptability and Generality:** Our proposed KG-CoT can be generalized to various LLM backbones (e.g., closed-source or open-source LLMs) with reduced API calls and costs in a lightweight plug-and-play manner.

2 Related Work

In this section, we introduce related LLM-based QA systems from two categories based on their utilization of knowledge.

2.1 LLM + Parametric Knowledge

As the model scale increases, the emergent ability enables LLMs to comprehend natural language instructions and activate the parametric knowledge [Petroni *et al.*, 2019] stored in their parameters for downstream NLP tasks.

Recently, Wei *et al.* first introduces the concept of chain-of-thought prompting (CoT), in which a series of intermediate reasoning steps is generated to solve complex problems through manually constructed prompts. Kojima *et al.* demonstrates the ability of LLMs to generate CoT, even in zero-shot scenarios. Consequently, Zhang *et al.*, Shao *et al.*, and Liang *et al.* leverage manually constructed CoT examples to automatically generate high-quality CoT demonstrations. Huang *et al.* fine-tunes LLMs based on their self-generated CoT examples and demonstrates the self-improvement capability of LLMs.

However, the difficulty of modifying and updating the parametric knowledge leads to LLMs utilizing outdated or incorrect implicit parametric knowledge for response generation, which strongly limits the validity and interpretability of black-box LLMs. In this case, a natural and promising solution is to augment LLMs with external world knowledge.

2.2 LLM + External Knowledge

Retrieval-augmented generation (RAG) is a natural way to augment LLMs with external knowledge [Lewis *et al.*, 2020]. This approach aims to retrieve relevant knowledge from massive knowledge bases (KBs) and directly augment LLMs with external world knowledge. Paranjape *et al.* enhances the ability of knowledge retriever to increase the probability of relevant passages being ranked among the top-10 most relevant. In addition, Ma *et al.* retrieves knowledge triplets over knowledge graphs (KGs) for question answering. Zhao *et al.* converts KGs to text descriptions to augment LMs. KGs are structured, explicit, and interpretable, since several paths from the question entity to the answer entity can be identified.

However, recent works generally utilized representation-based multi-model pre-training for augmenting LLMs with KGs [Zhao *et al.*, 2024; Ye *et al.*, 2023]. In addition to limiting the adaptability to closed-source LLMs, these methods ignore the elaborate knowledge structure and explicit reasoning paths, which can serve as explicit clues for joint reasoning with LLMs. Although Wang *et al.* has demonstrated that LLMs have preliminary graph reasoning abilities, the over-reliance on LLMs results in limited adaptability when dealing with large-scale KGs and complex multi-hop tasks.

To address these challenges, we propose the KG-CoT prompting, which includes a lightweight joint reasoning model to alleviate a portion of the reasoning burden of LLMs and perform joint reasoning over KGs. The graph reasoning model can generate explicit reasoning paths relevant to the questions, enabling LLMs to "think with KGs" for answer generation.

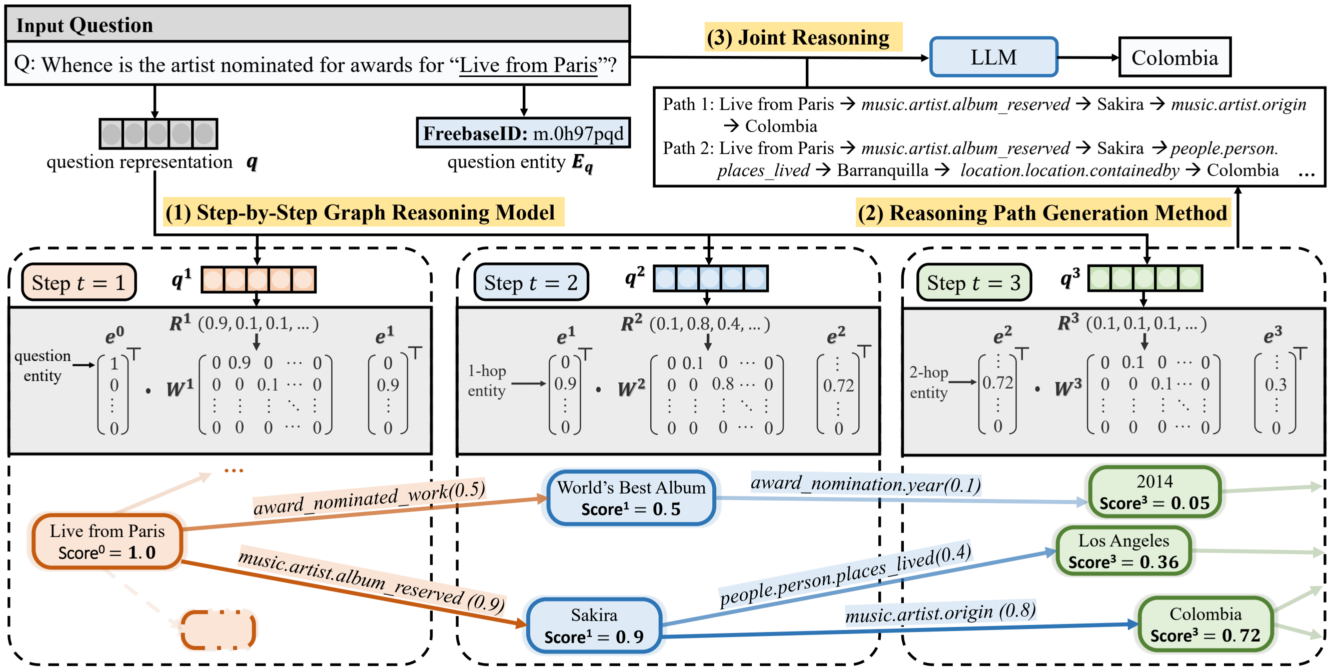

Figure 2: An overview of the KG-CoT. (1) We first propose a step-by-step graph reasoning model to reason over KGs and explore entities with high confidence in problem solving. (2) We develop a reasoning path generation method to extract reasoning paths for LLMs. (3) We concatenate the question context and reasoning paths, and utilize elaborate instructions to prompt LLMs for answer generation.

3 KG-CoT

KG-CoT augments LLMs with relevant knowledge by applying a small graph reasoning model to reason over KGs and generate reasoning paths with high confidence in LLM reasoning. First, we propose a graph reasoning model to perform step-by-step reasoning over the KGs and find candidate entities with high confidence. Then, we introduce the reasoning path generation method to generate the reasoning paths based on the step-by-step reasoning process. Finally, we leverage the reasoning paths to prompt LLMs for answer generation.

3.1 Step-by-Step Graph Reasoning Model

Prior semantic parsing based models [Li and Ji, 2022] have shown that the natural language question can be converted into its logical form, which is called a query graph. These findings suggest that complex questions can be decomposed into multiple meta-questions over the KGs, which is similar to chain-of-thought prompting [Wei *et al.*, 2022]. Inspired by this, we propose a graph reasoning model to imitate the question decomposition and step-by-step reasoning over KGs.

Initialization. Let G denote the KG, n denote the number of entities in the entity set, and m denote the number of relation in the relation set. We first initialize an entity state $\mathbf{e}^0 \in [0,1]^n$, which is a one-hot vector that indicates whether the corresponding entity is mentioned in the context of questions. For example, if only the i th entity is mentioned in the question, the $e_i^0 \in \mathbf{e}^0$ is initialized to 1 and others are set to 0. Moreover, we initialize a triplet matrix $\mathbf{M} \in [0,1]^{n\times n}$, which is a one-hot matrix that indicates the relation index $M_{ij} = k$ if it exists between the entity $i < n$ and entity $j < n$.

Relation Score Calculation. Inspired by [Shi *et al.*, 2021], we separate the graph reasoning process into T steps. At step $t < T$, we calculate scores for all relations in the KGs $\mathbf{R}^t \in [0,1]^m$. The score of each relation $r_i^t \in \mathbf{R}^t$ indicates the probability of a "hop" occurring for the current entity based on this relation. The calculation of relation score $\mathbf{R}^t$ is calculated as follows:

$$\mathbf{R}^t = \text{Sigmoid}(\text{MLP}(\mathbf{q}^t)), \quad (1)$$

where $\mathbf{q}^t$ is the question representation at step t. We consider the question representation at different steps to focus on different parts of the question context. In this way, we can implicitly decompose the question and and force the graph reasoning model to focus on different relations at different steps. At step t, the question representation $\mathbf{q}^t$ can be formulated as follows:

$$\mathbf{q}, (h_1, \ldots, h_{|q|}) = \text{Encoder}(q), \quad (2)$$

$$\mathbf{Q}^t = f^t(\mathbf{q}), \quad (3)$$

$$\boldsymbol{\alpha}^t = \text{Softmax}(\mathbf{Q}^t[h_1; \ldots; h_{|q|}]^T), \quad (4)$$

$$\mathbf{q}^t = \sum_{i=1}^{|q|} \alpha_i^t h_i, \quad (5)$$

where $\mathbf{q}$ is the question embedding and $(h_1, \ldots, h_{|q|})$ is a sequence of hidden states associated with the question. f^t is used to project the question embedding $\mathbf{q}$ to the attention query $\mathbf{Q}^t$ at step t. We calculate the attention weights $\boldsymbol{\alpha}^t$ and calculate the question representation at step t by taking the weighted sum of the hidden states.

Step-by-Step Reasoning. Based on the relation score $\mathbf{R}^t$, we first define a transition matrix $\mathbf{W}^t \in [0,1]^{n\times n}$, which is used to describe the transitions from the current entity states $\mathbf{e}^{t-1}$ to the next entity states $\mathbf{e}^t$. We leverage the triplet matrix $\mathbf{M}$ and relation score $\mathbf{R}^t$ to construct the transition matrix $\mathbf{W}^t$:

$$W_{ij}^t = \begin{cases} R_k^t & k = M_{ij}, R_k^t \in \mathbf{R}^t, M_{ij} \in \mathbf{M}, \\ 0 & \text{Otherwise}, \end{cases} \tag{6}$$

where k is the index of the relation between entities i and j, and R_k^t is the score of relation k. Finally, we can utilize the transition matrix to perform step-by-step reasoning over the KG. The step-by-step reasoning process can be formulated as follows:

$$\mathbf{e}^t = \mathbf{e}^{t-1}\mathbf{W}^t. \tag{7}$$

The current entities $\mathbf{e}^{t-1}$ "hop" along the relations within their 1-hop neighborhood and transmit to the next entity states $\mathbf{e}^t$ based on the relation score $\mathbf{R}^t$.

After T steps reasoning, we utilize the question embedding $\mathbf{q}$ to determine the weight distribution $\boldsymbol{\beta}$ for each step, and calculate the final entity scores $\bar{\mathbf{e}}$ by taking the weighted sum of the entity scores at each step.

$$\boldsymbol{\beta} = \text{Softmax}(\text{MLP}(\mathbf{q})), \tag{8}$$

$$\bar{\mathbf{e}} = \sum_{t=1}^{T} \beta_t \mathbf{e}^t, \tag{9}$$

Training. Given the one-hot vector $\mathbf{a} \in [0,1]^n$ of the golden answer, which indicates whether the corresponding entity is the answer entity. We use the L2 Euclidean distance between $\bar{\mathbf{e}}$ and $\mathbf{a}$ to optimize the step-by-step graph reasoning model:

$$\mathcal{L} = \|\bar{\mathbf{e}} - \mathbf{a}\|^2. \tag{10}$$

3.2 Reasoning Path Generation Method

During inference, once we obtain the top-k entities $\mathbf{E}^K \subseteq \mathbf{E}$ through the graph reasoning model, we utilize the transition matrices $\mathbf{W}^1$, $\mathbf{W}^2$ ... , $\mathbf{W}^T$ to generate the reasoning paths.

Initialization. During the generation of reasoning paths, we maintain two lists L_{rp} and L_{mid}, which are used to store the candidate reasoning paths and the intermediate paths.

Extraction. Starting from the question entity E_q, we first extract the corresponding row $w_{i0}^1, w_{i1}^1, \ldots, w_{i(n-1)}^1 \in \mathbf{W}^1$, $w_{ij}^1 > 0$, which indicates the relation score of transitioning from the question entity at step $t = 0$ to the entities at step $t = 1$. In this way, we can extract a set of 1-hop paths $\mathbf{P}^1$:

$$\text{p}_{ij}^1 = <\text{“E}_i, \text{Rel}_{ij}, \text{E}_j\text{”}, [w_{ij}^t]>, \tag{11}$$

where the "key" is the extracted path and the "value" is the score of relation within it. Rel_{ij} denotes the relation between the entities i and j. For each path p_{ij}^1, we first append it to the L_{rp}. If the object entity E_j is contained in the top-k answer entities $\mathbf{E}^K$, we then append the extracted path to the L_{rp}.

Then, we start from the object entities E_j of the 1-hop paths in the L_{mid} and use the $\mathbf{W}^2$ to extract 2-hop paths $\mathbf{P}^2$:

$$\text{p}_{ik}^2 = <\text{“E}_i, \text{Rel}_{ij}, \text{E}_j, \text{Rel}_{jk}, \text{E}_k\text{”}, [w_{ij}^1, w_{jk}^2]>, \tag{12}$$

Algorithm 1 Inference process of LLM + KG-CoT

Input: Input question q, retrieved knowledge subgraph G, and a large language model LLM.
Initialize Entity score $\mathbf{e}^0 \leftarrow$ extract question entity E_q from G, triplet matrix $\mathbf{M} \leftarrow$ extract triplets from G.
Output: Output answer y

for $t = 1, ..., T$ **do**
  Compute the question representation $\mathbf{q}^t$ using (2)-(5).
  Compute the relation score $\mathbf{R}^t$ using (1).
  Compute transition matrix $\mathbf{W}^t$ using (6).
  Entity score $\mathbf{e}^t \leftarrow$ step-by-step reasoning using (7).
end for
Compute final scores $\bar{\mathbf{e}}$ using (8)-(9) and select top-k entities $\mathbf{E}^K$
Initialize $L_{mid} \leftarrow$ p^0 = ¡“E_q”, [0]¿.
for $t = 1, ..., T$ **do**
  Extract t-hop paths $\mathbf{P}^t$ using $\mathbf{W}^t$ and paths in L_{mid}.
  Update intermediate path list L_{mid} with P_{qj}^t
  if Object entity $\text{E}_j \in \mathbf{E}^K$ **then**
    Update reasoning path list L_{rp} with P_{qj}^t.
  end if
end for
Select N paths for each top-k entity $\mathbf{E}^K$ from L_{rp}.
Serialize reasoning paths to textual sentence s
Output answer y = Call(LLM, q, s).
return Output answer y

---

and update the L_{rp} and L_{mid}.

By repeating the above algorithm for T steps, we can generate candidate reasoning paths from the question entities to the top-k entities.

Ranking. Each answer entity may correspond to multiple candidate paths in L_{rp}, and the number of hops for different paths varies. Therefore, we take the average of the scores of relations in each path as the final path score.

3.3 Joint Reasoning

For the top K candidate entity with highest confidence, we extract the path with the highest path score for each candidate entity. Thus, for each question, we utilize the step-by-step graph reasoning model (Section 3.1) and a reasoning path generation method (Section 3.2) to generate K reasoning paths with various numbers of hops and answer entities.

To maintain the chain structure, we utilize "arrows" to connect the entities and relations to construct the KG-CoT. For example, a 2-hop path p_{ik}^2:

$$\text{p}_{ik}^2 = <\text{“E}_i, \text{Rel}_{ij}, \text{E}_j, \text{Rel}_{jk}, \text{E}_k\text{”}>, \tag{13}$$

is serialized to a textual sentence, which is formulated as:

$$\text{Text}(\text{p}_{ik}^2) = \text{E}_i \rightarrow \text{Rel}_{ij} \rightarrow \text{E}_j \rightarrow \text{Rel}_{jk} \rightarrow \text{E}_k. \tag{14}$$

We serialize the K reasoning paths and concatenate them with the question context as the final input sequence. We utilize elaborate instructions to prompt LLMs to leverage these reasoning paths for answer generation.

Model	AccessKB	Multi-hopQA		Single-hop QA	Open-domain QA
		WebQSP	CompWebQ	SimpleQuestions	WebQuestions
ChatGPT + IO prompts [Patel *et al.*, 2023]	×	63.3	37.6	20.0	48.7
ChatGPT + CoT prompts [Wei *et al.*, 2022]	×	62.2	38.8	20.3	48.5
ChatGPT + SC [Wang *et al.*, 2023c]	×	61.1	45.4	18.9	50.3
Previous RA SOTA	✓	65.0^{α}	$\mathbf{70.4}^{\beta}$	85.8^{α}	56.3^{γ}
Previous KBQA SOTA	✓	76.6^{δ}	52.2^{δ}	71.1^{ϵ}	-
ChatGPT + ToG-R [Sun *et al.*, 2023]	✓	75.8	58.9	45.4	53.2
GPT-4 + ToG-R [Sun *et al.*, 2023]	✓	81.9	69.5	58.6	57.1
ChatGPT + KG-CoT (ours)	✓	**82.1**	51.6	77.8	**66.5**
GPT-4 + KG-CoT (ours)	✓	**84.9**	62.3	**86.1**	**68.0**

Table 1: Accuracy comparison with standard prompting baselines, state-of-the-art retrieval-augmented (RA) baselines (e.g., α: DiFaR^2 [Baek *et al.*, 2023], β: CBR [Das *et al.*, 2021], and γ: FiE [Kedia *et al.*, 2022], knowledge base question answering (KBQA) baselines (e.g., δ: UniKGQA [Jiang *et al.*, 2023] and ϵ: RNG [Ye *et al.*, 2022]) and recent LLM+KG baseline ToG-R [Sun *et al.*, 2023].

4 Experiments

4.1 Datasets

We evaluate KG-CoT based on 4 challenging knowledge-intensive question answering benchmarks that heavily rely on knowledge-aware reasoning with external world knowledge.

WebQSP. WebQSP is a knowledge-intensive multi-hop question answering benchmark. It contains 4,037 questions that are all 1-hop or 2-hop questions based on the Freebase. Based on previous works, we retrieve knowledge triplets within 2-hop neighborhoods of the question entities and produce a knowledge subgraph with 1,886,684 entities, 1,144 relations, and 5,780,246 knowledge triplets.

CompWebQ. CompWebQ is a multi-hop question answering benchmark. It contains 34,672 questions with many hops and constraints, which makes it challenging for LLMs to process. We utilize the retrieved knowledge subgraph of [Shi *et al.*, 2021] and utilize the original data splits for evaluation.

SimpleQuestions. SimpleQuestions is a single-hop question answering benchmark. Questions are generated based on information from Freebase, and ultimately, 108,442 questions that heavily rely on factual knowledge were generated in this study. We randomly select 1,000 questions and retrieved 1-hop neighborhood of the question entity for evaluation.

WebQuestions. WebQuestions is a challenging open-domain question answering benchmark. It contains 5,810 questions, with Freebase as the knowledge base. For each question, we retrieve the 2-hop neighborhood of the question entity and utilize the original data splits for evaluation.

4.2 Baselines

We compare with strong baselines, such as standard prompting baselines, state-of-the-art retrieval-augmented (RA) baselines and knowledge base question answering (KBQA) baselines, based on the above benchmark datasets.

Prompting Baselines. We compared with original IO prompts (IO prompts), chain-of-thought prompts (CoT prompts) and Self-Consistency (SC)

Retrieval-Augmented Baselines. We select previous SOTA of each benchmark, including direct fact retrieval DiFaR [Baek *et al.*, 2023], case-based reasoning CBR [Das *et al.*, 2021], and fusion in encoder FiE [Kedia *et al.*, 2022].

Knowledge Base Question Answering Baselines. We compared with previous state-of-the-art knowledge base question answering model on each benchmark, including UniKGQA [Jiang *et al.*, 2023] and RNG [Ye *et al.*, 2022].

LLM+KG Baseline. We also compare with recent KG-augmented baseline ToG [Das *et al.*, 2021]. Different from our motivation, it instructs LLM itself to perform retrieval, pruning and answer prediction.

4.3 Implementation Setting

We train the step-by-step graph reasoning model with RAdam optimizer at a learning rate of 1e-3 for 50 epochs. For the LLM, we leverage the OpenAI API to call ChatGPT and GPT-4 for evaluation. We select the "gpt-3.5-turbo" and "gpt-4" as our LLM backbones and utilize the default setting of the OpenAI API. For each question, we generate 1 KG-CoT for each top-10 candidate entity (e.g., Hit@10_Path1) and establish instructions to prompt the LLMs to directly generate answer entity for evaluation. Our code and data is available at https://github.com/HUSTNLP-codes/KG-CoT.

4.4 Main Results

As shown in Table 1, our proposed KG-CoT achieved state-of-the-art performance on 3 knowledge-intensive question answering benchmarks, including WebQSP, SimpleQuestions, and WebQuestions. Moreover, KG-CoT significantly enhances the performance of LLMs on CompWebQ, the challenging knowledge-intensive multi-hop question answering benchmark, compared to LLM baselines relying on standard CoT prompting.

On the WebQSP benchmark, our proposed KG-CoT outperforms recent LLM+KG baseline ToG [Sun *et al.*, 2023] for both LLM settings. Moreover, KG-CoT with the ChatGPT backbone even outperforms ToG with the GPT-4 backbone, demonstrating the effectiveness of our proposed "Large+Small" paradigm for LLMs.

Method	KB	WebQSP	CWQ
	Baselines		
GFC [Xie *et al.*, 2022]	✓	76.8	50.4
UniKGQA [Jiang *et al.*, 2023]	✓	76.6	52.2
ChatGPT+ToG-R [Sun *et al.*, 2023]	✓	75.8	58.9
GPT-4+ToG-R [Sun *et al.*, 2023]	✓	81.9	69.5
	Llama2-7B		
CoT	×	46.1	27.6
KG-CoT	✓	**72.4**	**46.7**
Gain		**(+26.3)**	**(+19.1)**
	Llama2-13B		
CoT	×	47.2	29.9
KG-CoT	✓	**74.6**	**50.0**
Gain		**(+27.4)**	**(+20.1)**
	ChatGPT		
CoT	×	62.2	38.8
KG-CoT	✓	**82.1**	**51.6**
Gain		**(+19.9)**	**(+12.8)**
	GPT-4		
CoT	×	67.3	46.0
KG-CoT	✓	**84.9**	**62.3**
Gain		**(+17.6)**	**(+16.3)**

Table 2: Accuracy comparison on different LLM backbones. We conduct experiments on open-sourced LLMs (e.g., Llama2-7B and -13B) and closed-sourced LLMs (e.g., ChatGPT and GPT-4).

For the single-hop and open-domain question answering benchmarks, our proposed KG-CoT also achieves competitive performances compared to previous state-of-the-art baselines. Notably, for the simple yet knowledge-intensive benchmark, LLMs that rely solely on the parametric knowledge struggle to generate the correct answer even with CoT prompting. On the one hand, the results demonstrate the effectiveness of our proposed KG-CoT in augmenting LLMs with explicit reasoning paths. On the other hand, these results align with previous findings of LLMs [Wang *et al.*, 2023b], indicating that the effectiveness of these standard prompting methods (e.g., CoT and SC) diminishes for complex problems that require extensive factual knowledge.

On the CompWebQ benchmark, our proposed KG-CoT with ChatGPT yielded at 37.2% improvement over standard prompting baselines. We observe that the performance trend of KG-CoT varies compared to those of the other QA benchmarks. This difference is attributed to our proposed graph reasoning model performing less favorably for this challenging multi-hop question answering benchmark, consequently resulting in moderate improvements on CompWebQ.

4.5 Comparison with Different LLM Backbones

To further investigate the generality of our proposed KG-CoT, we evaluate KG-CoT on different LLM backbones, such as open-source LLMs (e.g., Llama-7B and Llama-13B) and closed-source LLMs (e.g., ChatGPT and GPT-4). As shown in Table 2, our proposed KG-CoT yield significant improvements across all the LLM backbones. With the increasing intelligence of LLMs, the performance with KG-CoT consistently improves. When Llama2-13B, ChatGPT, and GPT-4 are used as the backbones, LLM+KG-CoT outperforms the existing state-of-the-art KGQA baselines.

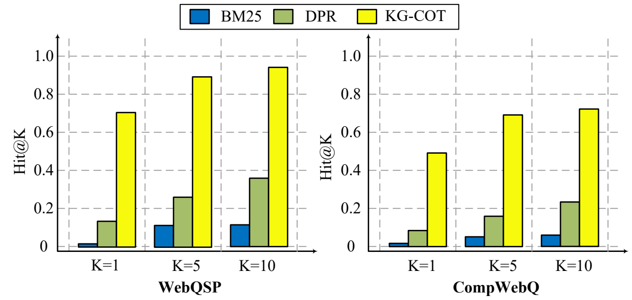

Figure 3: Performance comparison with other retriever. KG-CoT excels in locating the answer entity within the top-ranked k entities.

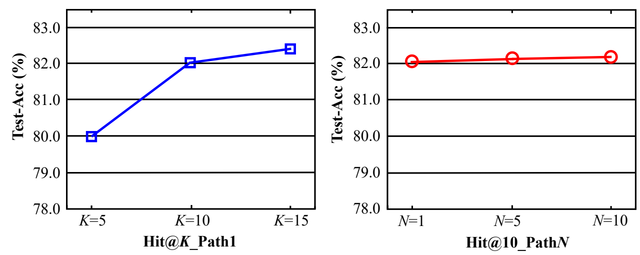

Figure 4: Effects of the number of candidate entity (e.g., Hit@K) and the path number corresponding to each entity (e.g., Hit@*K*Path*N*) on the performance of LLMs.

4.6 Performance of Locating Answer Entity

To validate the effectiveness of the reasoning paths, we conducted comparative experiments with other retrieval methods (e.g., BM25 and DPR) to evaluate whether the retrieved top-ranked knowledge facts contain the golden answer entity. As shown in Figure 3, the experimental results indicate that our proposed KG-CoT achieves a significant high hit rate (e.g., Hit@1) in locating reasoning paths to the answer entity. Specifically, when considering the top-10 candidate entities (e.g., K=10), KG-CoT yields a coverage rate of up to 91.4% for the WebQSP benchmark, surpassing that of conventional knowledge retrievers. Moreover, for the complex multi-hop question answering benchmark CompWebQ, the hit rate for answers reaches 72.5%, reflecting a substantial improvement in LLM reasoning.

4.7 Hyper-parameter Sensitivity Analysis

In KG-CoT, two parameters that can influence the performance of LLMs are the number of candidate answer entities K and the number of paths extracted for each entity N.

As shown in the left part of Figure 4, we conduct experiments on the WebQSP benchmark to investigate the influence of the candidate answer entity number K. We observed that increasing the number of candidate answer entities from the top-5 to the top-10 yields in a discernible improvement in the performance of LLMs. This enhancement can be attributed to an increased coverage of correct answer entities

Question	What is the source of the river which flows under the Dona Ana Bridge?
Reasoning Paths	**Path 1:** Dona Ana Bridge → *transportation.bridge.body_of_water_spanned* → Zambezi River → *geography.river.origin* → **Mwinilunga**. **Path 2:** Dona Ana Bridge → *geography.body_of_water.bridges_rev* → Zambezi River → *geography.river.origin* → **Mwinilunga**.
Outputs	**IO**: Willow Creek. **CoT**: The Dona Ana Bridge spans the Rio Grande, a significant river in the southwestern United States. The ultimate source of the Rio Grande is in the San Juan Mountains of Colorado. More precisely, the river originates from the snowmelt in the San Juan Mountains. The specific location where the Rio Grande begins is often cited as the Willow Creek, which is in the San Juan Mountains of Colorado. **KG-CoT**: Based on the reasoning paths, the Dona Ana Bridge spans a body of water, specifically the Zambezi River, which originates in the geographical location of Mwinilunga.
Ground Truth	Mwinilunga.

Table 3: Case study for KG-CoT. LLMs can accurate answer knowledge-intensive questions and provide knowledge-aware explanations.

Method	KB	#API Call	Cost Per Call($)	Total Cost($)
GPT-3.5-turbo				
CoT	×	2	0.0001	0.30
ToG	✓	11.2	∼0.0007	∼13
KG-CoT (Ours)	✓	1	0.0006	0.92
GPT-4				
CoT	×	2	0.003	9.25
ToG	✓	11.2	∼0.025	∼400
KG-CoT (Ours)	✓	1	0.020	30.82

Table 4: The number of API calls and cost of the OpenAI API for WebQSP. We show the number of API call per question, as well as the average cost per call and total cost for the WebQSP benchmark.

along the reasoning paths, consequently reducing misguidance caused by the absence of correct reasoning paths. However, when extending the candidate answers from the top 10 to the top 15, we found minimal changes in the performance of LLMs. One one hand, this lack of improvement is attributed to the marginal increase in answer coverage. Moreover, the lower confidence associated with lower ranking reasoning paths contributes marginally to the reasoning process of LLMs.

As shown in the right part of Figure 4, we investigate the impact of the number of reasoning paths corresponding to each candidate answer entity. We observed that increasing the number of reasoning paths has minimal effect on LLM reasoning. This is attributed to the fact that our proposed step-by-step reasoning model already provides the reasoning path with high confidence, which significantly contributes to the LLM reasoning. The inclusion of low-confidence reasoning paths leads to little improvement in LLMs.

4.8 Case Study

In Table 3, our further investigation reveals how KG-CoT enhances the reasoning capability of LLMs by providing accurate factual knowledge and interpretable reasoning paths. For the question: "What is the source of the river which flows under the Dona Ana Bridge?", original prompting methods are influenced by hallucination problems, resulting in an erroneous answer "Willow Creek". Instead, KG-CoT links the question entity to the Freebase and leverages our proposed step-by-step reasoning model to extract reasoning paths with high confidence, enabling LLMs to utilize the responsible and interpretable reasoning paths to generate the correct answer.

4.9 Adaptability

As shown in Table 4, we analyze the advantages of KG-CoT in practical application from two perspectives.

Bandwidth Occupancy. Since we utilize the "Large + Small" paradigm, we only need to extract reasoning paths from the small-scale graph reasoning model and perform joint reasoning with LLMs. This eliminates the necessity of LLMs generating CoT prompts or acting as retrievers to filter triplets and determine the next-hop entity (i.e., ToG [Sun *et al.*, 2023]). On the one hand, KG-CoT reduces the number of API calls to 1 per question, achieving more efficient knowledge enhancement. On the other hand, it diminishes the bandwidth occupancy of LLMs, allowing them to allocate more bandwidth to handle requests from other users.

Inference Cost. In contrast to previous LLM+KG baseline ToG [Sun *et al.*, 2023] which requires an average of 11.2 API calls, KG-CoT can significantly reduce the cost of API calls. Furthermore, our proposed graph reasoning model focuses on "relations" within the KGs, eliminating the need for retraining the model when countering emerging entities.

5 Conclusion

In this work, we propose a novel chain-of-thought prompting over knowledge graphs (KG-CoT), which utilizes a lightweight step-by-step graph reasoning model to augment LLMs with responsible factual knowledge and explicit reasoning paths in a plug-and-play manner. This "Large + Small" paradigm alleviates the burden of LLM reasoning and enables joint reasoning with external world knowledge. Extensive experiments on 4 knowledge-intensive question answering benchmarks demonstrate the effectiveness of our proposed KG-CoT and can provide explicit reasoning paths for improving interpretability. We show that KG-CoT can utilize less bandwidth and reduce inference costs to enhance the capability of various LLMs for knowledge-aware reasoning.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2023YFF0905503, National Natural Science Foundation of China under Grants No.62072203, No.62072257 and the Australian Research Council Under Grants DP22010371, LE220100078.

References

[Baek *et al.*, 2023] Jinheon Baek, Alham Fikri Aji, Jens Lehmann, and Sung Ju Hwang. Direct fact retrieval from knowledge graphs without entity linking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10038–10055, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Bollacker *et al.*, 2007] Kurt D. Bollacker, Robert P. Cook, and Patrick Tufts. Freebase: A shared database of structured general human knowledge. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pages 1962–1963, Vancouver, British Columbia, Canada, July 2007. AAAI Press.

[Cao *et al.*, 2021] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Chen *et al.*, 2023] Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si, and Yin Zhang. Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering. In *Findings of the Association for Computational Linguistics*, pages 13207–13224, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Das *et al.*, 2021] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Huang *et al.*, 2023] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore, December 2023. Association for Computational Linguistics.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, March 2023.

[Jiang *et al.*, 2023] Jinhao Jiang, Kun Zhou, Xin Zhao, and Ji-Rong Wen. Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda, May 2023. OpenReview.net.

[Kedia *et al.*, 2022] Akhil Kedia, Mohd Abbas Zaidi, and Haejun Lee. Fie: Building a global probability space by leveraging early fusion in encoder for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4246–4260, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, pages 22199–22213, New Orleans, LA, USA, November 2022. Curran Associates, Inc.

[Lewis *et al.*, 2020] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474, Virtual Event, December 2020. Curran Associates, Inc.

[Li and Ji, 2022] Mingchen Li and Jonathan Shihao Ji. Semantic structure based query graph prediction for question answering over knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1569–1579, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[Li *et al.*, 2023] Shiyang Li, Yifan Gao, Haoming Jiang, Qingyu Yin, Zheng Li, Xifeng Yan, Chao Zhang, and Bing Yin. Graph reasoning for question answering with triplet retrieval. In *Findings of the Association for Computational Linguistics*, pages 3366–3375, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Liang *et al.*, 2023] Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4329–4343, Singapore, December 2023. Association for Computational Linguistics.

[Ma *et al.*, 2022] Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. Open-domain question answering via chain of reasoning over heterogeneous knowledge. In *Findings of the Association for Computational Linguistics*, pages 5360–5374, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[OpenAI, 2023] OpenAI. GPT-4 technical report. *arXiv preprint*, arXiv:2303.08774, March 2023.

[Paranjape *et al.*, 2022] Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D.

Manning. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. In *Proceedings of the Tenth International Conference on Learning Representations*, Virtual Event, April 2022. OpenReview.net.

[Patel *et al.*, 2023] Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. Bidirectional language models are also few-shot learners. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda, May 2023. OpenReview.net.

[Petroni *et al.*, 2019] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Shao *et al.*, 2023] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *Proceedings of the Fortieth International Conference on Machine Learning*, pages 30706–30775, Honolulu, Hawaii, USA, July 2023. Proceedings of Machine Learning Research.

[Shi *et al.*, 2021] Jiaxin Shi, Shulin Cao, Lei Hou, Juanzi Li, and Hanwang Zhang. Transfernet: An effective and transparent framework for multi-hop question answering over relation graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4149–4158, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Sun *et al.*, 2023] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *arXiv preprint*, arXiv:2307.07697, July 2023.

[Wang *et al.*, 2023a] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *arXiv preprint*, arXiv:2305.10037, May 2023.

[Wang *et al.*, 2023b] Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *arXiv preprint*, arXiv:2305.10037, May 2023.

[Wang *et al.*, 2023c] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda, May 2023. OpenReview.net.

[Wang *et al.*, 2023d] Yubo Wang, Xueguang Ma, and Wenhu Chen. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint*, arXiv:2309.02233, September 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837, New Orleans, LA, USA, November 2022. Curran Associates, Inc.

[Xie *et al.*, 2022] Minghui Xie, Chuzhan Hao, and Peng Zhang. A sequential flow control framework for multi-hop knowledge base question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8450–8460, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Yao *et al.*, 2023] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint*, arXiv:2305.10601, May 2023.

[Ye *et al.*, 2022] Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. RNG-KBQA: generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6032–6043, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Ye *et al.*, 2023] Qichen Ye, Bowen Cao, Nuo Chen, Weiyuan Xu, and Yuexian Zou. Fits: Fine-grained two-stage training for knowledge-aware question answering. *arXiv preprint*, arXiv:2302.11799, February 2023.

[Zhang *et al.*, 2023] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda, May 2023. OpenReview.net.

[Zhao *et al.*, 2023] Feng Zhao, Hongzhi Zou, and Cheng Yan. Structure-aware knowledge graph-to-text generation with planning selection and similarity distinction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8693–8703, Singapore, December 2023. Association for Computational Linguistics.

[Zhao *et al.*, 2024] Ruilin Zhao, Feng Zhao, Liang Hu, and Guandong Xu. Graph reasoning transformers for knowledge-aware question answering. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, pages 19652–19660, Vancouver, Canada, February 2024. AAAI Press.

[Zhu *et al.*, 2020] Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. Modifying memories in transformer models. *arXiv preprint*, arXiv:2012.00363, December 2020.