Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation

Shengjie Ma[1,2,*] Chengjin Xu[1,*,†] Xuhui Jiang[1,*] Muzhi Li[3] Huaren Qu[4] Cehao Yang[1]
Jiaxin Mao[2,†] Jian Guo[1,†]

[1]IDEA Research, International Digital Economy Academy [2]GSAI, Renmin University of China [3]The Chinese University of Hong Kong
[4]The Hong Kong University of Science and Technology [*]: Equal contribution [†]: Corresponding Author

Abstract

Retrieval-augmented generation (RAG) has improved large language models (LLMs) by using knowledge retrieval to overcome knowledge deficiencies. However, current RAG methods often fall short of ensuring the depth and completeness of retrieved information, which is necessary for complex reasoning tasks. In this work, we introduce Think-on-Graph 2.0 (ToG-2), a hybrid RAG framework that iteratively retrieves information from both unstructured and structured knowledge sources in a tight-coupling manner. Specifically, ToG-2 leverages knowledge graphs (KGs) to link documents via entities, facilitating deep and knowledge-guided context retrieval. Simultaneously, it utilizes documents as entity contexts to achieve precise and efficient graph retrieval. ToG-2 alternates between graph retrieval and context retrieval to search for in-depth clues relevant to the question, enabling LLMs to generate answers. We conduct a series of well-designed experiments to highlight the following advantages of ToG-2: 1) ToG-2 tightly couples the processes of context retrieval and graph retrieval, deepening context retrieval via the KG while enabling reliable graph retrieval based on contexts; 2) it achieves deep and faithful reasoning in LLMs through an iterative knowledge retrieval process of collaboration between contexts and the KG; and 3) ToG-2 is training-free and plug-and-play compatible with various LLMs. Extensive experiments demonstrate that ToG-2 achieves overall state-of-the-art (SOTA) performance on 6 out of 7 knowledge-intensive datasets with GPT-3.5, and can elevate the performance of smaller models (e.g., LLAMA-2-13B) to the level of GPT-3.5's direct reasoning. The source code is available on `https://github.com/IDEA-FinAI/ToG-2`.

Overview

The advantage of ToG-2 can be summarized as:

- **In-depth retrieval:** ToG-2 achieves in-depth and reliable context retrieval through the guide of KGs and performs precise graph retrieval by treating documents as node contexts, achieving a tight-coupling combination of KG and text RAG, enabling deep and comprehensive retrieval processes.
- **Faithful reasoning:** ToG-2 iteratively performs a collaborative retrieval process based on KG and text, using retrieved heterogeneous knowledge as the basis for LLM reasoning and enhancing the faithfulness of LLM-generated content
- **Efficiency and Effectiveness:** a) ToG-2 is a training-free and plug-and-play framework that can be applied to various LLMs; b) ToG-2 can be executed between any associated KG and document database, while for purely document database, entities can be extracted from the documents first, and then a graph can be constructed through relation extraction or entity co-occurrence; c) ToG-2 achieves new SOTA performances on various complex knowledge reasoning datasets and can elevate the reasoning capabilities of smaller LLMs, e.g., Llama2-13B to a level comparable to direct reasoning with powerful LLMs like GPT-3.5.

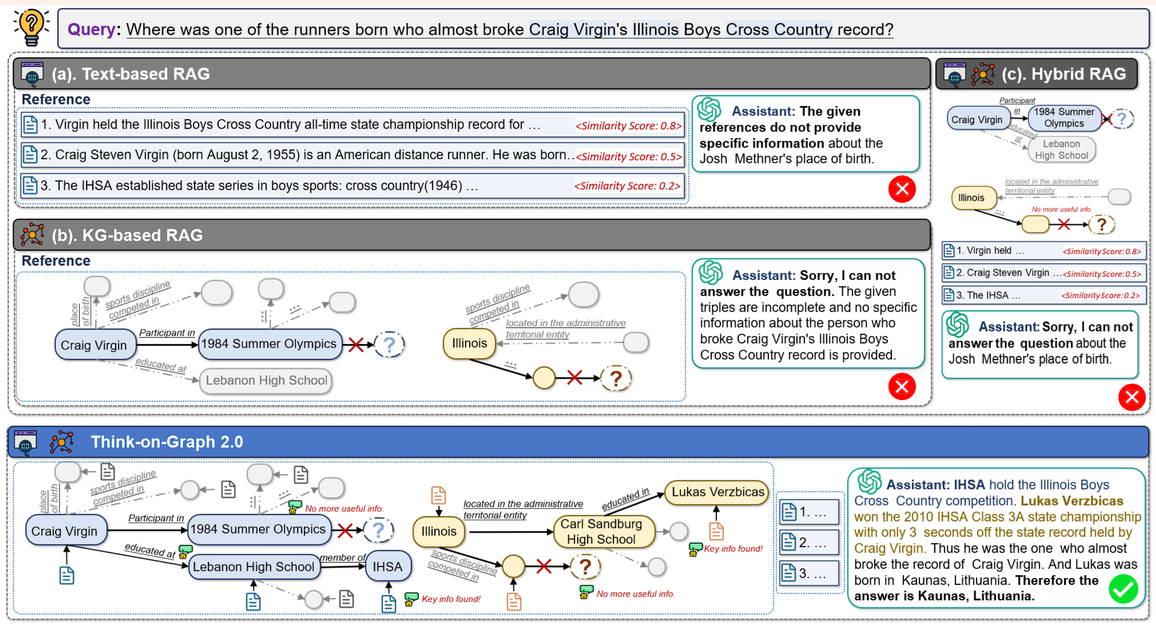

Figure 1. The illustration of differences among (a) text-based RAG, (b) KG-based RAG, (c) KG+Text (Loose-coupling) Hybrid RAG, and (d) our proposed KG×Text (Tight-coupling) RAG framework:
a: The dataset size is too large in the semantic retrieval-based RAG paradigm, resulting in low information density. Also, traditional retrieval systems struggle to capture deep connections between facts, thus failing to focus on key points in the question.
b: In the KG path inference-based LLM augmentation paradigm, the information provided by triples lacks both depth and details. Even information may be missing, due to the incompleteness of the KG.
c: The proposed $\text{ToG}_{2.0}$ combines the advantages of both approaches. The KG helps to understand deep connections between different facts and precisely narrows down the search scope, while entity context-based retrieval supplements the information missing in the knowledge graph's triple path reasoning.

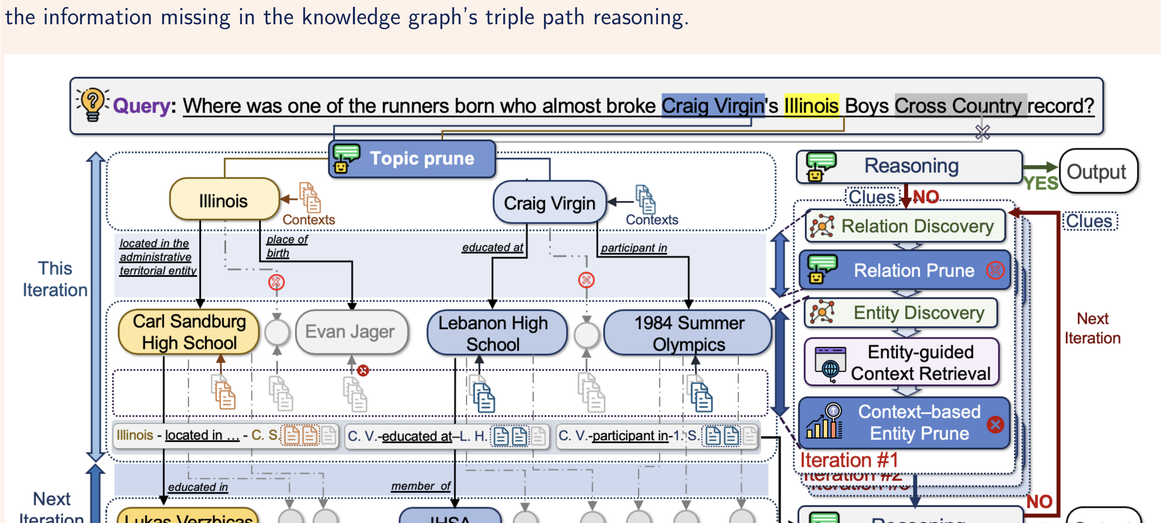

Figure 2. The workflow of ToG-2.

Methods

The iterative process will be explained in two main parts, **Context-enhanced Graph Search** and **Knowledge-guided Context Retrieval**. Formally, in the i-th iteration, the topic entities are denoted as $\mathcal{E}^i_{topic} = \{e^i_1, e^i_2, \ldots, e^i_j\}$ and their preceding triple paths are $\mathcal{P}^{i-1} = \{P^{i-1}_1, P^{i-1}_2, \ldots, P^{i-1}_j\}$, $P^{i-1}_j = \{p^0_j, p^1_j, \ldots, p^{i-1}_j\}$, where $j \in [1, W]$, W is a hyperparameter of exploration width (the max number of retained topic entities in each iteration, and p^{i-1}_j is a single triple $(e^{i-1}_j, r^{i-1}_j, e^i_j)$ containing r^{i-1}_j as the relation between e^{i-1}_j and e^i_j in the KG, which could be either direction. Note that $i = 0$ indicates the initialization phase and the P^0 is empty.

Relation Prune

$$\text{PROMPT}_{RP}(e^i_j, q, Edge(e^i_j)), \tag{1}$$

$$\text{PROMPT}_{RP\_cmb}(\mathcal{E}^i_{topic}, q, \{Edge(e^i_j)\}^W_{j=1}). \tag{2}$$

Context-based Entity Prune

The relevance score of z-th chunk of $c^i_{j,m}$:

$$s^i_{j,m,z} = \text{DRM}(q, [triple\_sentence(Pc^i_{j,m}) : chunk^i_{j,m,z}]). \tag{3}$$

The ranking score of a candidate entity $c^i_{j,m}$ is calculated as the exponentially decayed weighted sum of the scores of its chunks that rank in top-K:

$$\text{score}(c^i_{j,m}) = \sum_{k=1}^{K} s_k \cdot w_k \cdot \mathbb{I}\,(\text{the } k\text{-th ranked chunk is from } c^i_{j,m}), \tag{4}$$

where $w_k = e^{-\alpha \cdot k}$, s_k is the score of the k-th ranked chunk, $\mathbb{I}$ is the indicator function that equals 1 if the k-th chunk belongs to $e^i_{j,m}$, and K and α are hyperparameters.

Reasoning

$$\text{PROMPT}_{rs}(q, \mathcal{P}^i, Ctx^i, Clues^{i-1}) = \begin{cases} Ans., & \text{if the knowledge is sufficient.} \\ Clues^i, & \text{otherwise.} \end{cases} \tag{5}$$

Experiment Results

Baseline Type	Method	Datasets					
		WebQSP (EM)	AdvHotpotQA (EM)	QALD-10-en (EM)	FEVER (Acc.)	Creak (Acc.)	Zero-Shot RE (EM)
LLM-only	Direct	65.9%	23.1%	42.0%	51.8%	89.7%	27.7%
	CoT	59.9%	30.8%	42.9%	57.8%	90.1%	28.8%
	CoT-SC	61.1%	34.4%	45.3%	59.9%† (56.2%‡)	90.8%	45.4%
Text-based RAG	Vanilla RAG	67.9%	23.7%	42.4%	53.8%	89.7%	29.5%
KG-based RAG	ToG	76.2%	26.3%	50.2%	52.7%	**93.8%**	88.0%
Hybrid RAG	CoK	77.6%	35.4%‡ (34.1%†)	47.1%	**63.5%**† (58.5%‡)	90.4%	75.5%
Proposed	ToG-2	**81.1%**	**42.9%**	**54.1%**	63.1† (**59.7%**‡)	93.5%	**91.0%**

Table 1. Performance comparison of different methods with GPT-3.5-turbo across various datasets. Note that the CoK model has 6-shot and 3-shot settings. We present the best performance of CoK under different shot settings for each dataset. For AdvHotpotQA and FEVER, we use the results reported in the original paper of CoK, where † represents the 3-shot setting and ‡ represents the 6-shot setting. Bold numbers represent the highest result under parallel settings.

	Llama-3-8B		Qwen2-7B		GPT-3.5-turbo		GPT-4o	
	Direct	ToG-2	Direct	ToG-2	Direct	ToG-2	Direct	ToG-2
AdvHotpotQA	20.8	34.7 (**66.8%** ↑)	17.9	30.8 (**72.1%** ↑)	23.1	42.9 (**85.7%** ↑)	47.7	53.3 (**11.3%** ↑)
FEVER	35.5	52.9 (**49.0%** ↑)	38.6	53.1 (**38.1%** ↑)	51.8	63.1 (**21.8%** ↑)	66.2	70.1 (**5.9%** ↑)
ToG-FinQA	0	8.2	0	10.3	0	34.0	0	36.1

Table 2. Performance comparison of direct reasoning and ToG-2 with different backbone models.

KG Completeness (%)	Exploration Setting	EM (%)
100	Default	43
80	Default	41
50	Default	35
30	Default	23
30	Adjusted	29

Table 3. Impact of KG Completeness on ToG-2.0 Performance with Default ($W = 3$, $D = 3$) and Adjusted ($W = 8$, $D = 2$) Exploration Strategies. The results suggest ToG-2.0 remains resilient to moderate KG incompleteness, emphasize the model's adaptability, and provide insights into potential strategies for handling incomplete KGs.

Contact

ideaQuant WeChat
E-mails: msj@ruc.edu.cn; xuchengjin@idea.edu.cn

Note: Shengjie Ma, Chengjin Xu, and Xuhui Jiang contributed equally. Jiaxin Mao and Jian Guo are corresponding authors.