ICML International Conference On Machine Learning 2025

MONASH University

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

SHANGHAI JIAO TONG UNIVERSITY

Griffith UNIVERSITY

Graph-constrained Reasoning: Faithful Reasoning on Knowledge Graphs with Large Language Models

Linhao Luo[1], Zicheng Zhao[2], Gholamreza Haffari[1], Yuan-Fang Li[1], Chen Gong[3], Shirui Pan[4]

[1]Monash University, [2]Nanjing University of Science and Technology, [3]Shanghai Jiao Tong University, [4]Griffith University

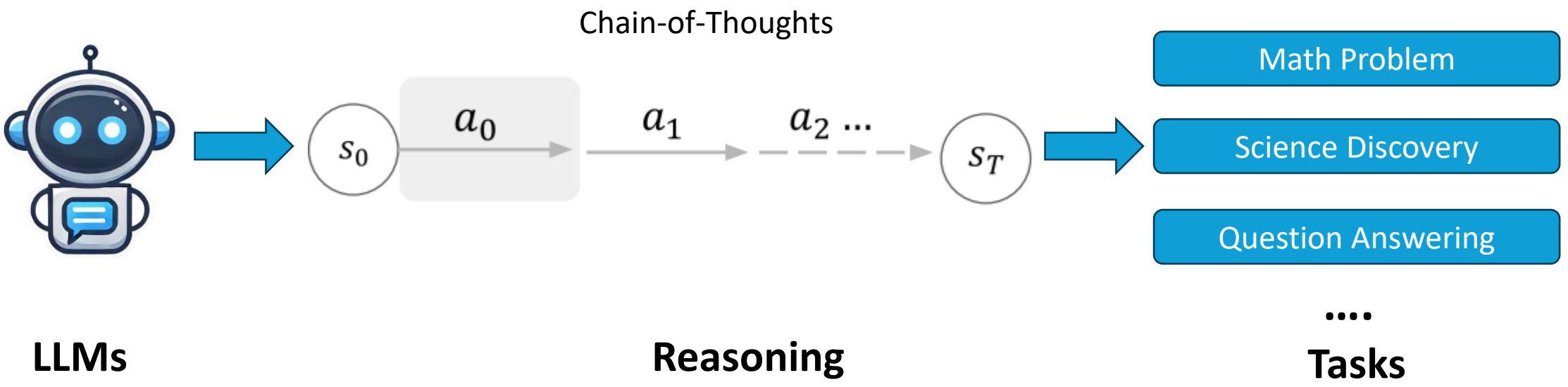
Presenter: Linhao Luo

Paper

Code

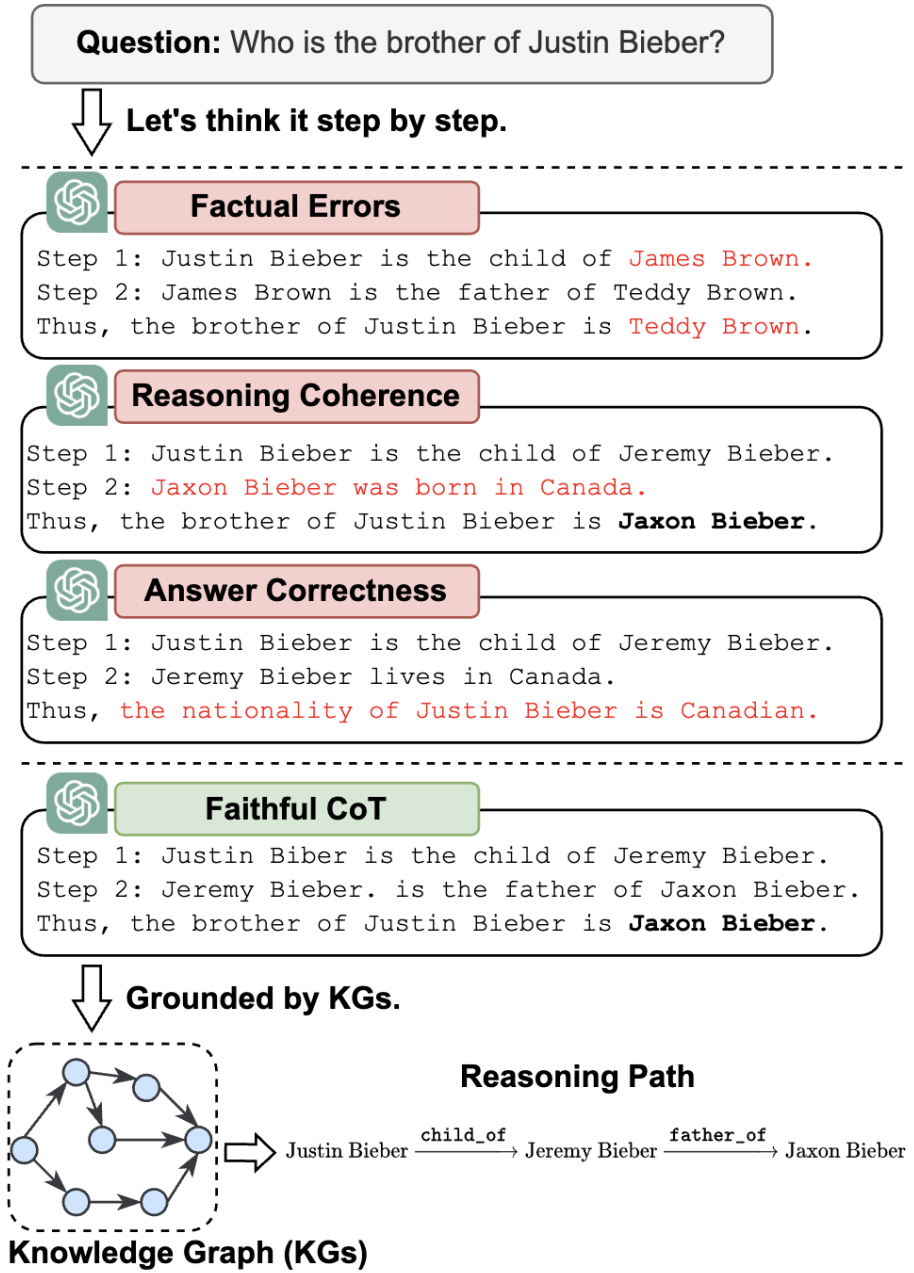
Motivations

- LLMs exhibit great reasoning ability via decoding (CoT).



Motivations

- LLMs struggle with conducting faithful reasoning due to issues of lack of knowledge and hallucination.



Examples of different reasoning errors and a faithful reasoning grounded by KG.

Nguyen, M. V., Luo, L., Shiri, F., Phung, D., Li, Y. F., Vu, T. T., & Haffari, Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs, ACL Findings 2024

Motivations

- The correct final answer may not result from the faithful reasoning of LLMs.

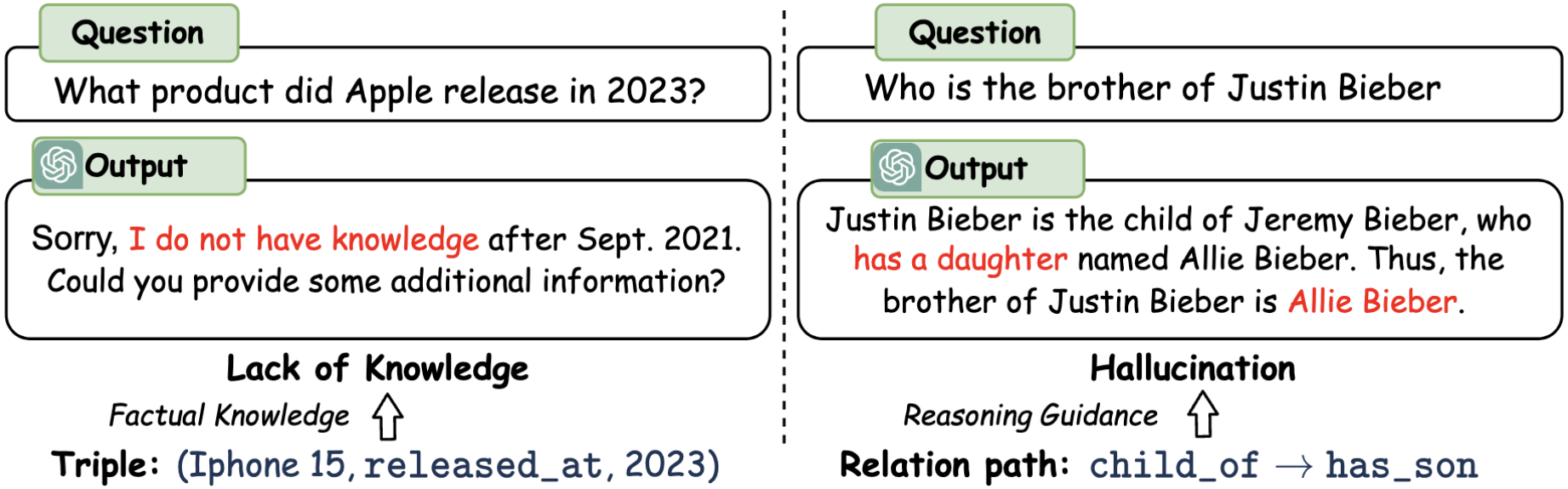
LLMs	Size	CWQ				GrailQA			
		Answer↑	Reasoning↑	Gap↓	Edit Dist.↓	Answer↑	Reasoning↑	Gap↓	Edit Dist.↓
		Fewsho CoT							
Mistral	7B	36.45	25.18	**11.27**	69.86	16.35	2.12	14.23	94.03
Qwen	7B	32.52	19.38	13.14	76.78	13.35	1.63	**11.72**	94.69
Qwen	14B	40.39	27.38	13.01	74.49	18.83	2.13	16.70	92.90
Vicuna	33B	44.50	15.92	28.58	74.60	18.26	0.95	17.31	95.39
LLaMA2	70B	49.80	33.98	15.82	62.23	22.05	2.88	19.17	92.58
ChatGPT	175B	**49.85**	**37.13**	12.72	**57.94**	**23.69**	**4.17**	19.52	**90.13**

There is a gap between answer accuracy and reasoning faithfulness.

Nguyen, M. V., Luo, L., Shiri, F., Phung, D., Li, Y. F., Vu, T. T., & Haffari, Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs, ACL Findings 2024

Motivations

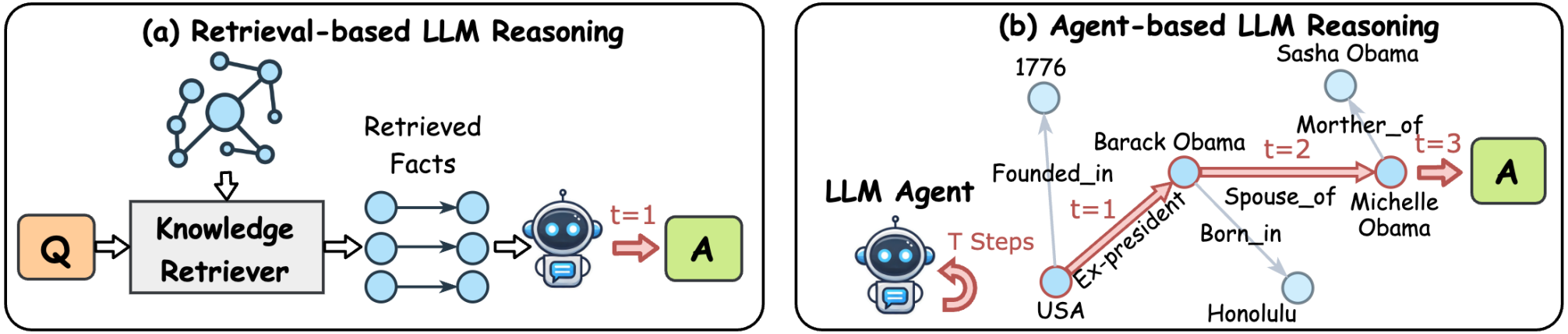
- Knowledge graphs (KGs) can be used to enhance the reasoning of LLMs.
  - KGs provide factual knowledge.
  - KGs provide structure guidance for reasoning (reasoning paths) to reduce hallucinations.



LUO, L., Li, Y. F., Haf, R., & Pan, S. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. ICLR 2024

Motivations

- Existing KG-enhanced LLM reasoning follows the retrieval-based and agent-based frameworks



Retrieval-based methods: retrieve-than-reasoning.

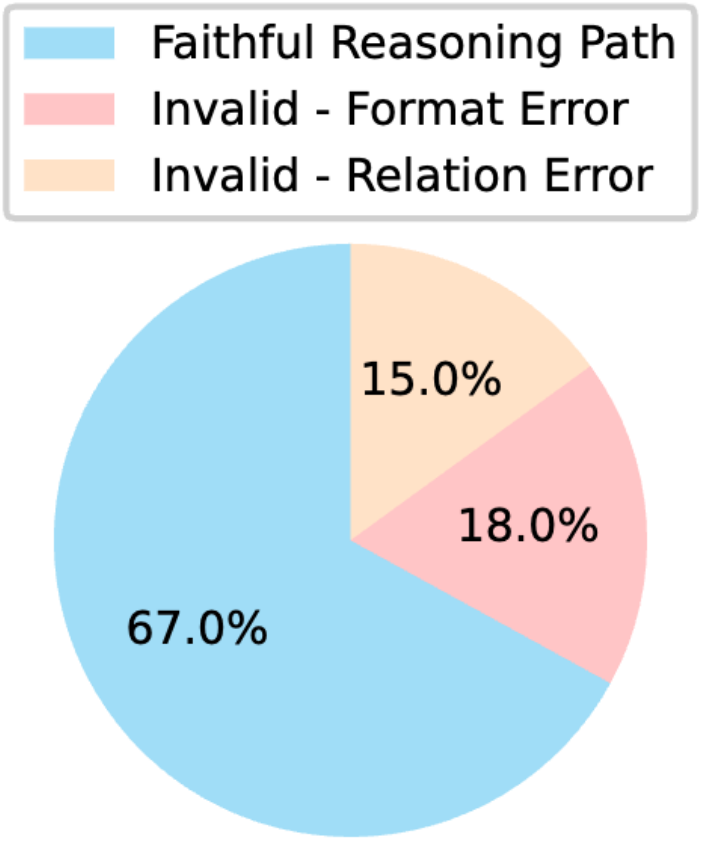
- Need additional retrievers.
- Design a retriever considering graph structure is challenging.

Agent-based methods: LLM search on graphs.

- Resource consuming (API calls)
- High-latency (time)

Motivations

- **Findings:** Existing methods (RoG) still cannot 100% ensure the faithful reasoning of LLMs.
- **Reason:** There are no constrains on the reasoning path generation. LLMs can generate paths that do not exist in the KGs.
- **Solution:** we introduce graph-constrained reasoning (GCR) to eliminate hallucinations and ensure accurate reasoning.

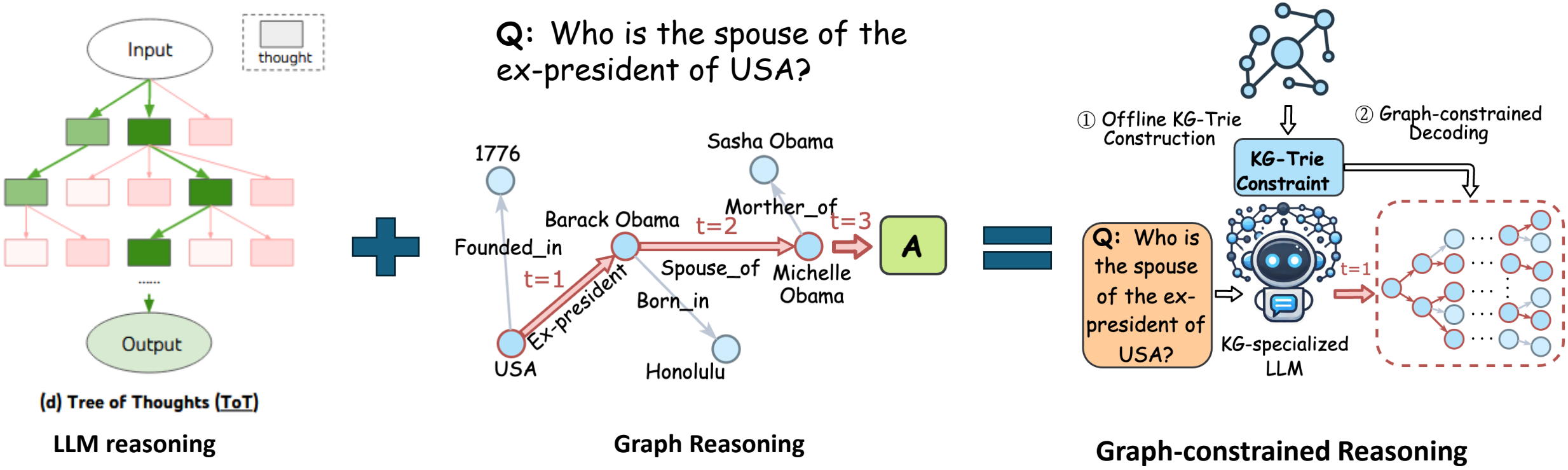


Reasoning Errors in RoG[1]

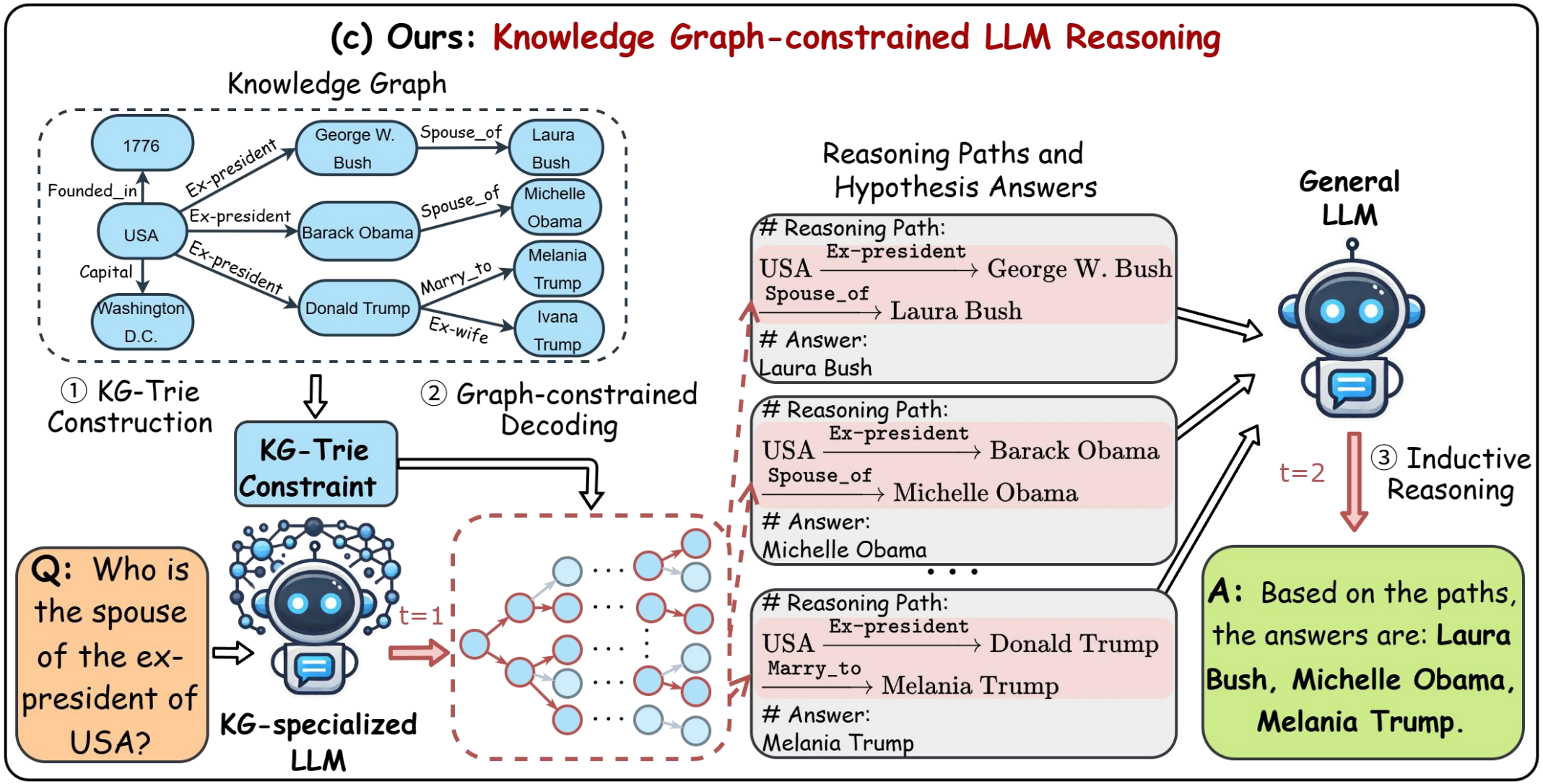
Luo Linhao, Li Yuan-Fang, Haffari Gholamreza, Pan Shirui, Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning, ICLR 2024

From Chain-of-Thought (CoT) to Graph-constrained Reasoning (GCR)

- **Graph-constrained Reasoning (GCR):**
  - Incorporates KGs into the decoding process of LLMs to achieve faithful reasoning (decoding on graphs)

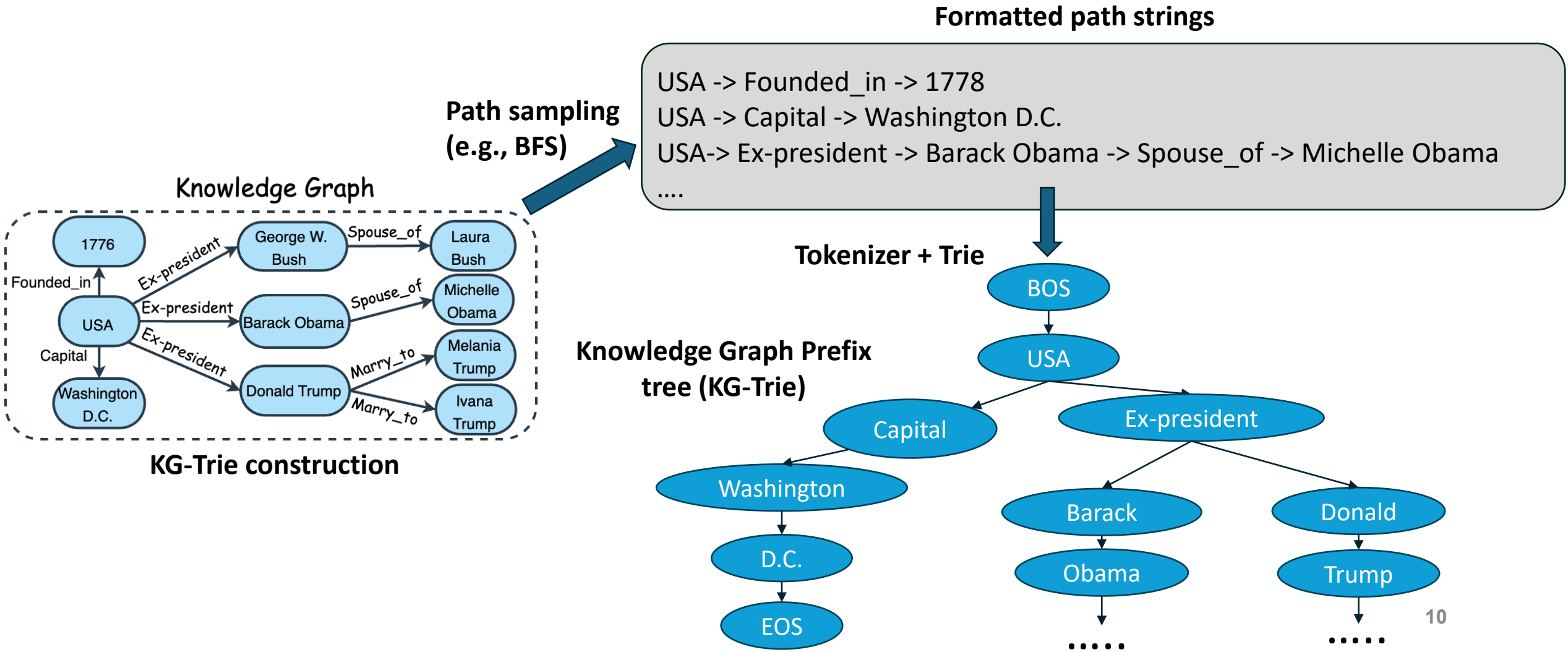


Graph-constrained Reasoning (GCR)



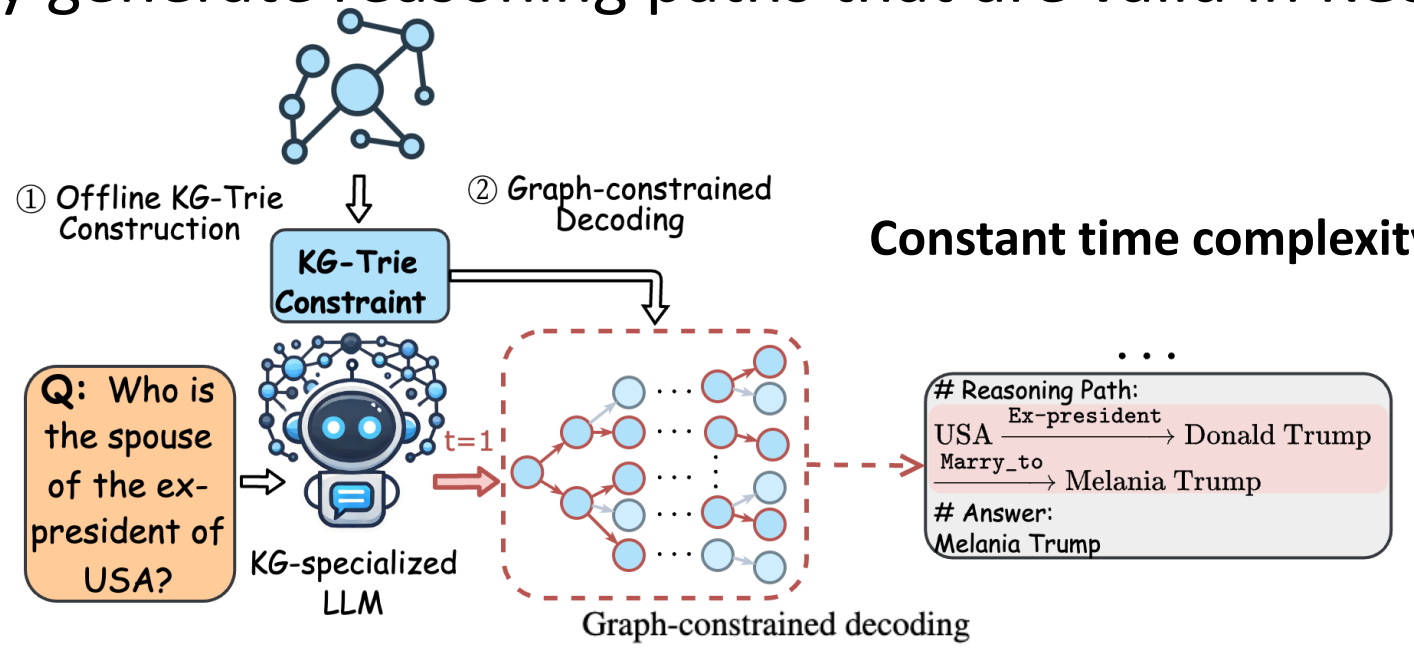
KG-Trie Construction

- **We convert KGs into KG-Tries to facilitate efficient reasoning on KGs.**



Graph-constrained decoding

- We adopt KG-Trie as constraints to guide the decoding process of LLMs and only generate reasoning paths that are valid in KGs.



Constant time complexity: $O(|\mathcal{W}_z|)$

$$P_\phi(a, w_z|q) = \underbrace{P_\phi(a|q, w_z)}_{\text{Regular decoding}} \overbrace{\prod_{i=1}^{n} P_\phi(w_{z_i}|q, w_{z_{1,\ldots,i-1}}) \mathcal{C}_\mathcal{G}(w_{z_i}|w_{z_{1,\ldots,i-1}})}^{\text{Graph-constrained decoding}}, \tag{6}$$

$$\mathcal{C}_\mathcal{G}(w_{z_i}|w_{z_{1,\ldots,i-1}}) = \begin{cases} 1, \exists \text{prefix}(w_{z_{1,\ldots,i}}, w_z), \exists w_z \in \mathcal{W}_z, \\ 0, else, \end{cases} \tag{7}$$

Graph-constrained decoding

- We finetune a lightweight KG-specialized LLMs (0.5B-7B) on the graph-constrained decoding task.

========== Prompt Input ==========

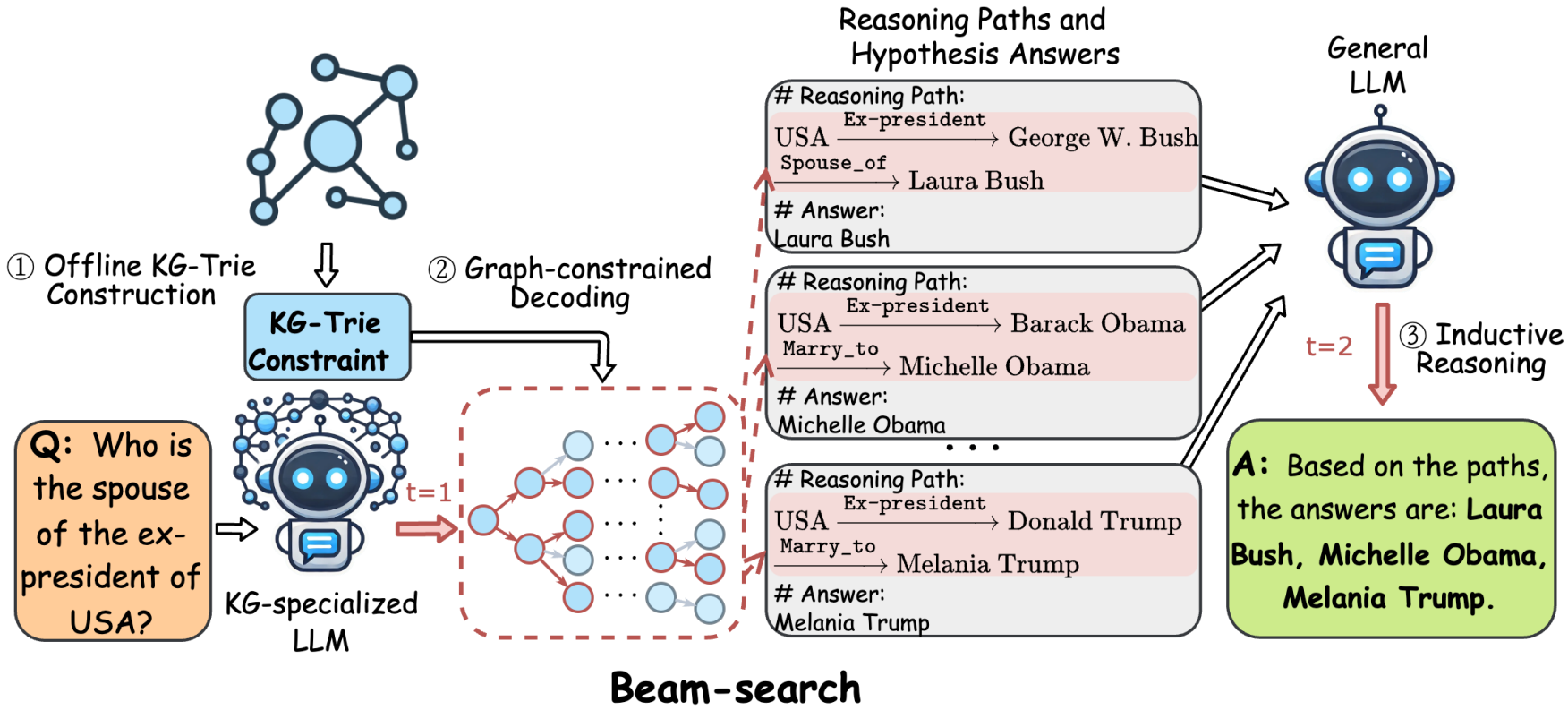
Please generate some reasoning paths in the KG starting from the topic entities to answer the question.
Question: what is the name of justin bieber brother?

========== LLM Output ==========

Reasoning Path: <PATH> Justin Bieber → people.person.parents → Jeremy Bieber → people.person.children → Jaxon Bieber </PATH>
Answer: Jaxon Bieber

Graph Inductive Reasoning

- The graph-constrained decoding can be paired with beam-search LLM generation to explore *K* reasoning paths in a single LLM call, which are then input into a powerful general LLM (e.g., ChatGPT) to derive final answers.



Results

Table 1. Performance comparison with different baselines on the two KGQA datasets.

Types	Methods	WebQSP		CWQ	
		Hit	F1	Hit	F1
LLM Reasoning	Qwen2-0.5B (Yang et al., 2024a)	26.2	17.2	12.5	11.0
	Qwen2-1.5B (Yang et al., 2024a)	41.3	28.0	18.5	15.7
	Qwen2-7B (Yang et al., 2024a)	50.8	35.5	25.3	21.6
	Llama-2-7B (Touvron et al., 2023)	56.4	36.5	28.4	21.4
	Llama-3.1-8B (Meta, 2024)	55.5	34.8	28.1	22.4
	GPT-4o-mini (OpenAI, 2024a)	63.8	40.5	63.8	40.5
	ChatGPT (OpenAI, 2022)	59.3	43.5	34.7	30.2
	ChatGPT+Few-shot (Brown et al., 2020)	68.5	38.1	38.5	28.0
	ChatGPT+CoT (Wei et al., 2022)	73.5	38.5	47.5	31.0
	ChatGPT+Self-Consistency (Wang et al., 2024b)	83.5	63.4	56.0	48.1
Graph Reasoning	GraftNet (Sun et al., 2018)	66.7	62.4	36.8	32.7
	NSM (He et al., 2021)	68.7	62.8	47.6	42.4
	SR+NSM (Zhang et al., 2022)	68.9	64.1	50.2	47.1
	ReaRev (Mavromatis & Karypis, 2022)	76.4	70.9	52.9	47.8
	UniKGQA (Jiang et al., 2022)	77.2	72.2	51.2	49.1
KG+LLM	KD-CoT (Wang et al., 2023)	68.6	52.5	55.7	-
	EWEK-QA (Dehghan et al., 2024)	71.3	-	52.5	-
	ToG (ChatGPT) (Sun et al., 2024)	76.2	-	57.6	-
	ToG (GPT-4) (Sun et al., 2024)	82.6	-	68.5	-
	EffiQA (Dong et al., 2024)	82.9	-	69.5	
	RoG (Llama-2-7B) (Luo et al., 2024)	85.7	70.8	62.6	56.2
	GNN-RAG (Mavromatis & Karypis, 2024)	85.7	71.3	66.8	59.4
	GNN-RAG+RA (Mavromatis & Karypis, 2024)	90.7	73.5	68.7	60.4
	GCR (Llama-3.1-8B + ChatGPT)	**92.6**	73.2	72.7	60.9
	GCR (Llama-3.1-8B + GPT-4o-mini)	92.2	**74.1**	**75.8**	**61.7**

KGQA Performance

Table 2. Efficiency and performance comparison of different methods on WebQSP.

Types	Methods	Hit	Avg. Runtime (s)	Avg. # LLM Calls	Avg. # LLM Tokens
Retrieval-based	S-Bert	66.9	0.87	1	293
	BGE	72.7	1.05	1	357
	OpenAI-Emb.	79.0	1.77	1	330
	GNN-RAG	85.7	1.52	1	414
	RoG	85.7	2.60	2	521
Agent-based	ToG	75.1	16.14	11.6	7,069
	EffiQA	82.9	-	7.3	-
Ours	GCR	**92.6**	3.60	2	231

Efficiency and performance comparison

- GCR achieves state-of-the-art performance
- GCR balances well between efficiency and effectiveness.

Results

Different KG-specialized LLMs

Different general LLMs

Table 4. Comparison of different LLMs used in GCR.

Components	Learning Types	Variants	Hit	F1
KG-specialized LLM	Zero-shot	Llama-3.1-8B	28.25	10.32
		Llama-3.1-70B	38.53	12.53
	Few-shot	Llama-3.1-8B	33.24	11.19
		Llama-3.1-70B	41.13	13.14
	Fine-tuned	Qwen2-0.5B	87.48	60.03
		Qwen2-1.5B	89.21	62.97
		Qwen2-7B	92.31	72.74
		Llama-2-7B	92.55	**73.23**
		Llama-3.1-8B	**92.74**	73.14
General LLM	Zero-shot	Qwen-2-7B	86.32	67.59
		Llama-3.1-8B	90.24	71.19
		Llama-3.1-70B	89.85	71.47
		ChatGPT	**92.55**	73.23
		GPT-4o-mini	92.23	**74.05**

- Lightweight LLMs after fine-tuning enable strong graph reasoning ability.
- Larger LLMs are better in both graph-constrained decoding and inductive reasoning.

Results

Table 5: Examples of the faithful reasoning conducted by GCR. Red denotes the incorrect reasoning paths and answers, while **bold** denotes the correct paths and answers.

Case 1: Incorrect answers and hallucinated reasoning paths without constraints.	
Question	Who is niall ferguson 's wife?
Answer	Ayaan Hirsi Ali
GCR w/o constraint	# Reasoning Path: Niall Ferguson → people.person.children → Mabel Rose Ferguson → people.person.parents → Alyssa Mastromonaco #Answer: Alyssa Mastromonaco
GCR	# Reasoning Path: Niall Ferguson → people.person.children → Thomas Ferguson → people.person.parents → **Ayaan Hirsi Ali** #Answer: **Ayaan Hirsi Ali**
Case 2: Correct answers but hallucinated reasoning paths without constraints.	
Question	Where is jamarcus russell from?
Answer	Mobile
GCR w/o constraint	# Reasoning Path: JaMarcus Russell → people.person.place_of_birth → Tampa #Answer: **Mobile, Alabama**
GCR	# Reasoning Path: JaMarcus Russell → people.person.place_of_birth → **Mobile** #Answer: **Mobile**

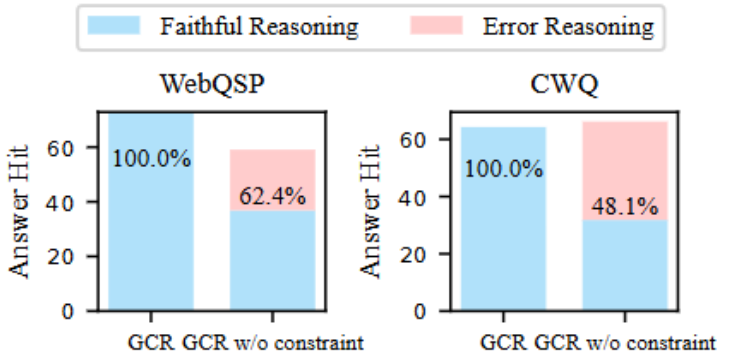


Figure 5: Analysis of performance and reasoning errors in GCR.

Faithful LLM reasoning with graph-constrained decoding

- Graph-constrained decoding can reduce the reasoning complexity and reach better performance in generating meaningful reasoning paths.
- Graph-constrained decoding can **eliminate** the hallucination in reasoning.
- The correct final answer may not result from faithful reasoning of LLMs.

Results

Table 6. Zero-shot transferability to other KGQA datasets.

Model	FreebaseQA	CSQA	MedQA
ChatGPT	85	79	64
GCR (ChatGPT)	**92**	**85**	**66**
GPT-4o-mini	89	91	75
GCR (GPT-4o-mini)	**94**	**94**	**79**

Zero-shot generalizability of GCR (Accuracy)

- Commonsense question answering (CSQA)
  - KG: Commonsense knowledge graphs
- Medical Question Answering (MedQA)
  - KG: Medical knowledge graphs

- GCR performs well with commonsense KGs due to the inclusion of commonsense knowledge in LLMs.
- GCR get limited improvement in domain-specific KGs like medical KGs, which might require further finetuning.

Thanks for your listening!

Paper

Code