南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Call Me When Necessary: LLMs can Efficiently and Faithfully Reason over Structured Environments

Sitao Cheng[1]*, Ziyuan Zhuang[1]*, Yong Xu[2], Fangkai Yang[2], Chaoyun Zhang[2], Xiaoting Qin[2], Xiang Huang[1], Ling Chen[2], Qingwei Lin[2], Dongmei Zhang[2], Saravan Rajmohan[2], Qi Zhang[2]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, China

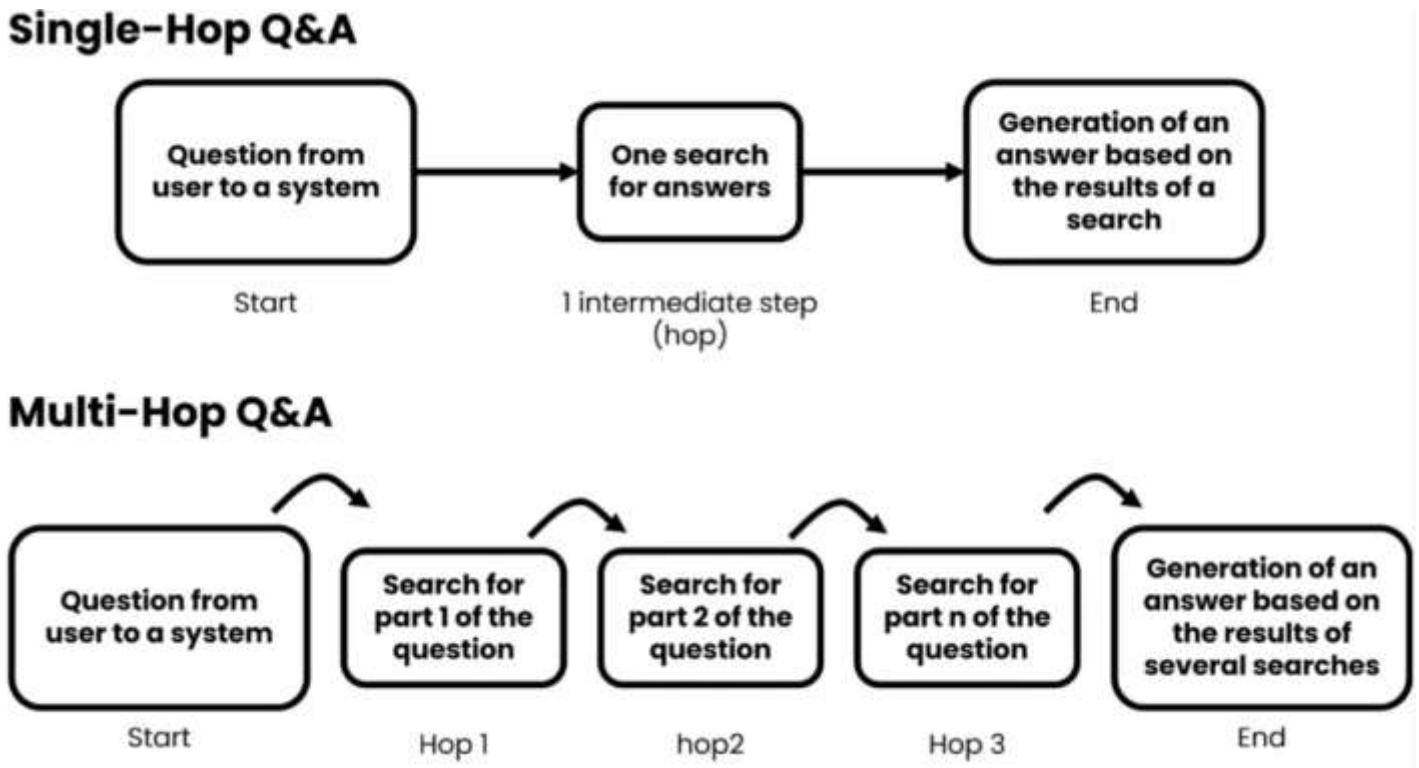
[2]Microsoft

{stcheng, ziyuan.zhuang}@smail.nju.edu.cn, {yox, fangkaiyang}@microsoft.com

ACL24

Background

Multi-Hop Question



示例：

问题：“爱因斯坦的出生地所在国家的首都是哪里？”

解析：
1.先回答：“爱因斯坦出生在哪里？”

答案是 **德国。**

2.再回答：“德国的首都是哪里？”

答案是 **柏林。**

3.最终得出答案：**柏林。**

Background

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

step-by-step interaction(CoT)

To faithfully reason, prior works adopt an iterative way that start from certain elements , instantiate on SEs and then gradually expand the reasoning path

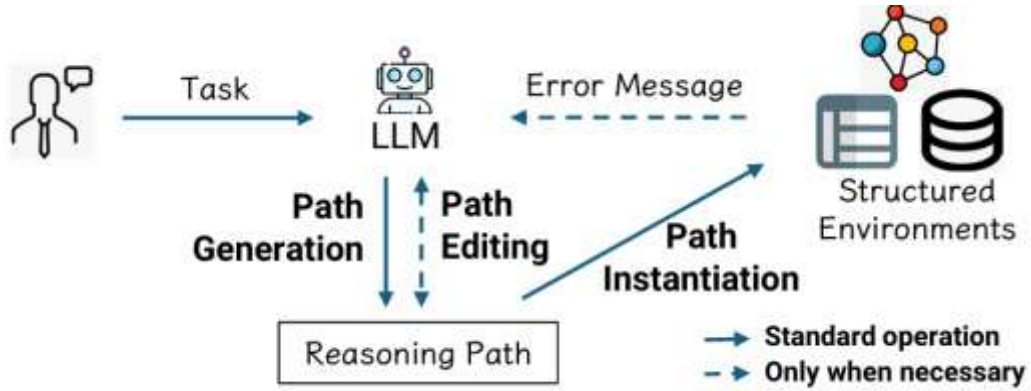
However, the reasoning efficiency is sacrificed and thus hinder the practical feasibility.

fine-tuned methods

Alternatively, fine-tuned methods inject environments into model parameters by tuning with human-labeled supervision. During inference, they recall schema patterns to build reasoning paths without interaction with SEs. This end-to-end paradigm is efficient.

However, it is never ensured that the model output can be grounded on SEs. Study shows that 50% paths of RoG (LUO et al., 2024) failed to yield faithful results.

Plan first, then Edit if Needed



Method Type	Existing Issues	Readi's Improvements
Step-by-step Reasoning	High number of LLM calls, low efficiency, and error propagation	Generate the path in one go, with editing allowed when necessary
End-to-End Fine-Tuning	Unreliability, requirement for large amounts of annotations, and poor generalization	Does not rely on training data; corrects via real-time feedback

Readi

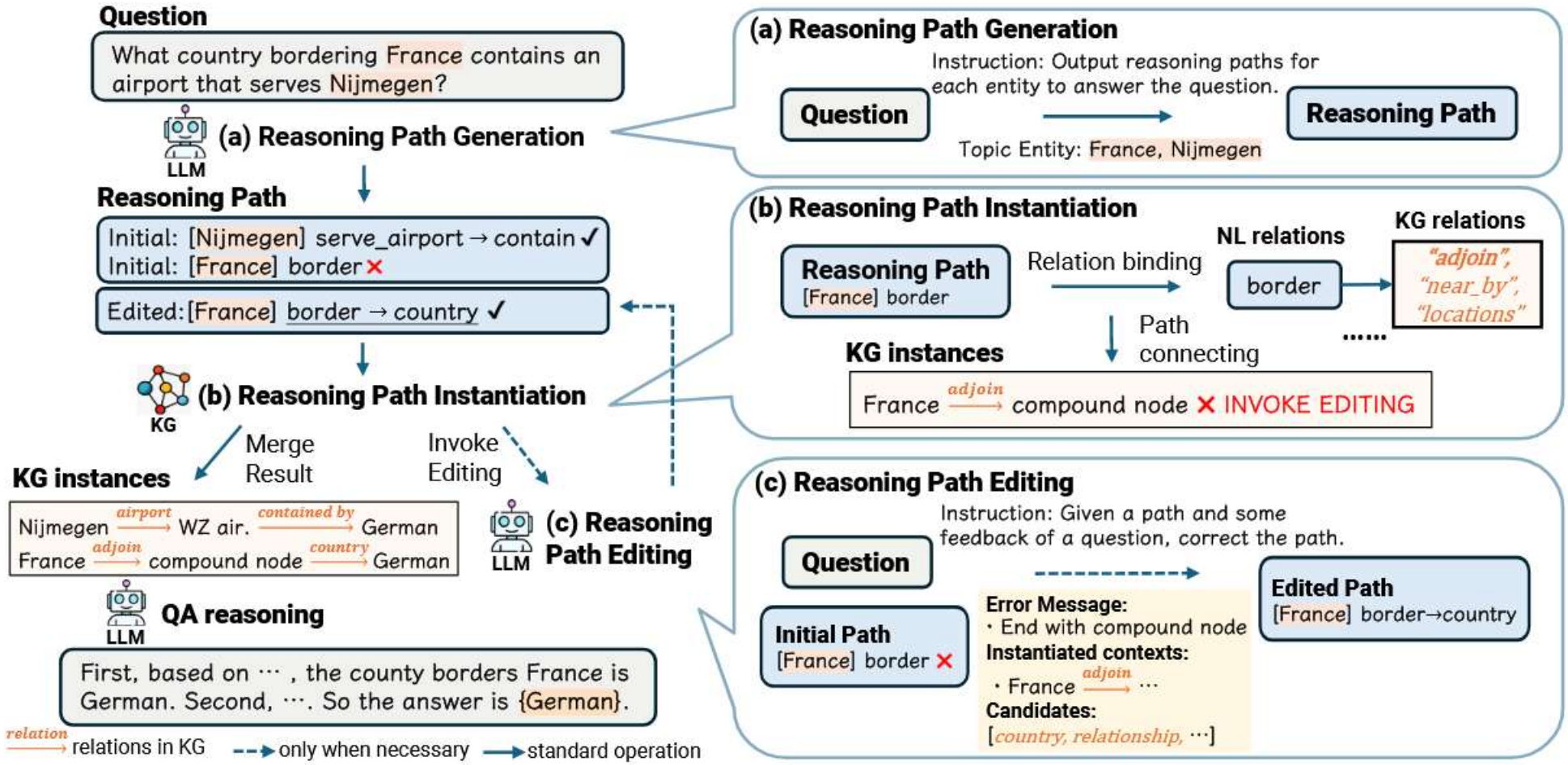


Figure 3: A running example of Readi on KGQA. An LLM initially generates an reasoning path for a question.

Examples

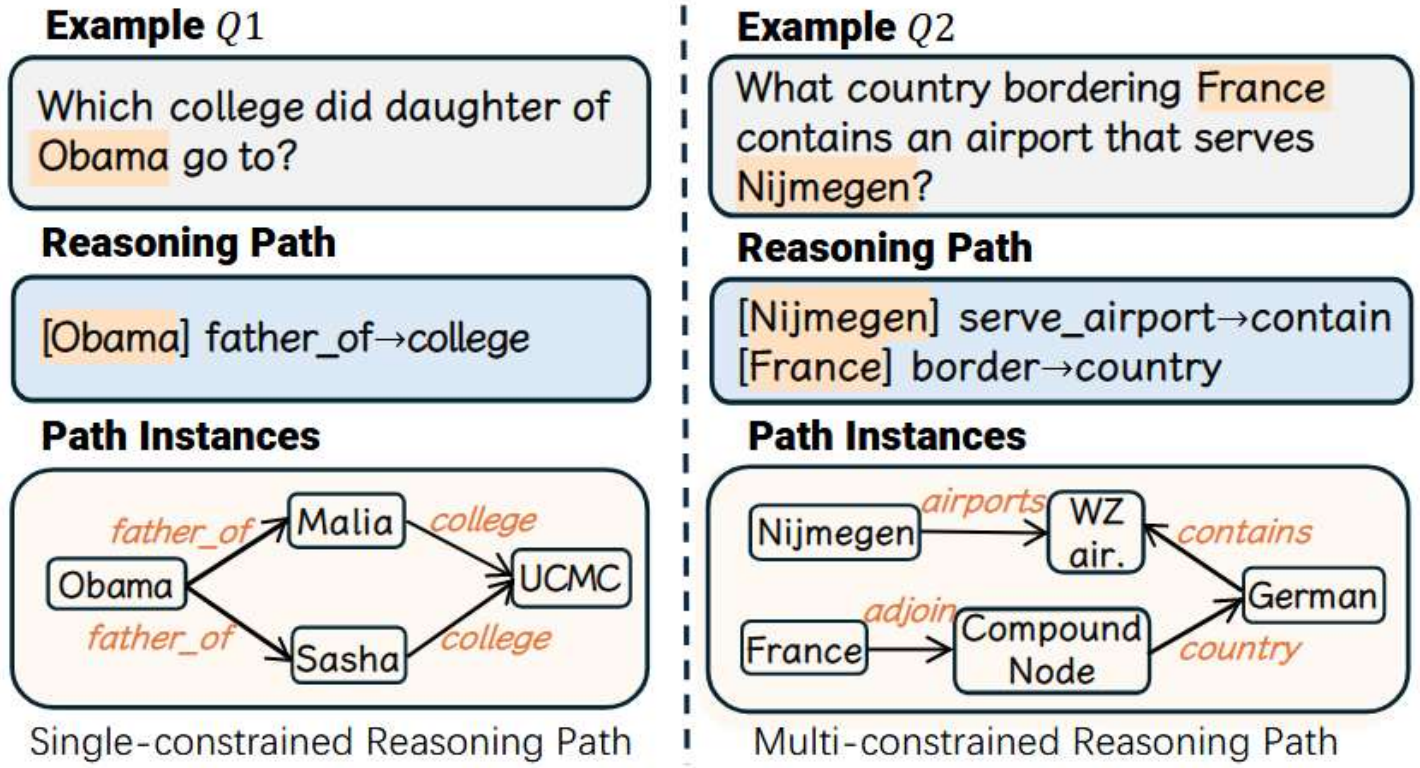


Figure 2: Examples of the question, reasoning path, and corresponding path instances on knowledge graph.

Experimental Results

Methods	WebQSP	CWQ	MQA-1H	MQA-2H	MQA-3H
Training-based Method					
EmbedKGQA (Saxena et al., 2020)	66.6	-	**97.5**	98.8	94.8
NSM (He et al., 2021)	67.7	47.6	97.1	99.9	98.9
TransferNet (Shi et al., 2021)	71.4	48.6	97.5	**100**⋆	**100**⋆
SR+NSM+E2E (Zhang et al., 2022)	69.5	49.3	-	-	-
UniKGQA (Jiang et al., 2023c)	75.1	50.7	**97.5**	99.0	99.1
ReasoningLM (Jiang et al., 2023b)	78.5	**69.0**⋆	96.5	98.3	92.7
RoG (LUO et al., 2024)	**85.7**⋆	62.6	-	-	84.8
Inference-based Method					
Davinci-003 (Ouyang et al., 2022)	48.7	-	52.1	25.3	42.5
GPT3.5 (OpenAI, 2022)	65.7	44.7	61.9	31.0	43.2
GPT4 (OpenAI, 2023)	70.7	52.1	71.8	52.5	49.2
AgentBench (Liu et al., 2024)	47.8	24.8	-	-	-
StructGPT (Jiang et al., 2023a)	69.6	-	97.1	97.3	87.0
Readi-GPT3.5	74.3	55.6	98.4	**99.9**	**99.4**
Readi-GPT4	**78.7**	**67.0**	**98.5**⋆	**99.9**	99.2

Experimental Results

Variance of Readi	Answer Coverage Rate (AC)				QA Performance (Hit@1)			
	Corrupt	*Empty*	GPT3.5	GPT4	*Corrupt*	*Empty*	GPT3.5	GPT4
w/o edit	-	-	56.7	62.7	-	-	51.0	57.2
w/ edit by GPT3.5	54.0	56.4	62.5	64.3	57.3	58.5	58.7	58.5
w/ edit by GPT4	55.6	63.9	68.6	65.8	58.2	59.9	58.1	59.3

Table 3: Answer Coverage Rate (AC) and QA Performance (Hit@1) of variance of Readi (GPT3.5 as reasoning module). Each column denotes a path generation method. *Corrupt* means a path with some randomly-sampled relations. *Empty* means empty path. w/o edit means we only leverage the initial reasoning path.

Thanks