

Knowledge Editing in LLM for Multi-hop Question-Answering

동계 세미나 2025.2.27
이재욱

KE for Multi-hop QA

Knowledge Editing은 edited/updated facts를 LLM의 답변 생성에 반영하는 방법

하지만, 수정된 지식을 Multi-hop reasoning에 올바르게 반영하도록 하는 것은 더 어려움

[Question]:
What is the capital city of the country where the Next Summer Olympics will be held?

- [Updated Information]:
1. The capital city of United States is New York.
 2. The capital city of France is Marseille.



KE for Multi-hop QA

Method	MQUAKE-CF-3K						MQUAKE-T			
	1 edited		100 edited		All edited		1 edited		All edited	
	Acc.	Hop-Acc	Acc.	Hop-Acc	Acc.	Hop-Acc	Acc.	Hop-Acc	Acc.	Hop-Acc
LLaMa-2										
Size: 7B										
FT _{COT}	22.3	-	2.13	-	OOM	-	47.32	-	3.75	-
FT	28.2	7.3	2.37	0.03	OOM	OOM	56.48	33.89	1.02	0.37
ROME _{COT}	11.17	-	2.87	-	2.77	-	28.96	-	14.4	-
ROME	13.13	5.37	3.5	0.03	3.63	0.1	24.89	17.99	1.71	0.32
MEMIT _{COT}	11.83	-	9.23	-	5.57	-	36.88	-	31.58	-
MEMIT	14.97	6.43	9.4	2.47	2.3	0.37	30.89	23.98	25.21	20.13
MeLLo	33.57	9.9	20.0	10.07	17.33	9.9	97.7	0.21	62.58	3.96
PokeMQA (Ours)	44.13	30.6	37.33	27.83	32.83	23.87	75.43	60.44	74.36	60.22
Vicuna										
Size: 7B										
MeLLo	22.7	7.03	12.83	6.77	10.9	6.7	42.24	1.12	19.86	1.28
PokeMQA (Ours)	45.83	34.8	38.77	31.23	31.63	25.3	74.57	55.19	73.07	55.09
GPT-3.5-turbo-instruct										
Size: Undisclosed										
MeLLo	57.43	28.8	40.87	28.13	35.27	25.3	88.12	52.84	74.57	53.53
PokeMQA (Ours)	67.27	56.37	56.0	49.63	45.87	39.77	76.98	68.09	78.16	67.88

기존의 Editing methods들은 멀티홉 추론에 edited facts를 적용하는 데에 매우 취약한 모습을 보임
 + 특히 파라미터를 수정하는 방법들은 멀티홉 QA 성능이 엄청 안 나온다

PokeMQA: Programmable knowledge editing for Multi-hop Question Answering

**Hengrui Gu¹, Kaixiong Zhou², Xiaotian Han³, Ninghao Liu⁴,
Ruobing Wang¹, Xin Wang¹⁺**

¹School of Artificial Intelligence, Jilin University

²Department of Electrical and Computer Engineering, North Carolina State University

³Department of Computer Science and Engineering, Texas A&M University

⁴School of Computing, University of Georgia

Limitation of Existing KE methods

1. Parameter를 수정하는 KE methods의 한계
 - 모델의 internal weights를 수정하는 방법론들은, MQA에서 요구하는 연쇄적인 지식 업데이트에 있어서 편집된 지식이 reasoning에 유연하게 적용되지 못함
2. External Memory 기반 편집의 한계
 - 메모리 기반 방법은 지식 편집과 question decomposition을 하나의 프롬프트 내에서 동시에 수행되도록 설계되어 있음 (MeLLO)

이는 두 가지 문제를 발생시킴

- 1) Edited facts와 original facts의 의미를 파악하고 충돌 여부를 판단하는 작업이 QA와 통합된 few-shot prompt 내에서 충분히 수행되기 어려움
- 2) Q- Decomposition 과 충돌 검출이라는 기능이 하나의 프롬프트에 혼합되면 서로 간섭을 일으킬 수 있음
>> LLM이 edited facts를 사용해서 멀티홉 추론하는데 방해가 됨

PokeMQA

Question Decomposition과 knowledge editing을 명확히 분리하여 각각에 집중할 수 있도록 설계

1) Programmable Scope Detector

Multi-hop question의 각 atomic question이 편집된 사실의 범위(scope)에 해당하는지 여부를 결정

$$g(t, q) : \mathcal{T} \times \mathcal{Q} \rightarrow [0, 1],$$

Scope Detector는 Pre-detector, Conflict Disambiguator의 두 모델로 구성

- Pre-detector: t와 q의 임베딩을 독립적으로 계산한 후, 두 임베딩의 유사도를 계산
- Conflict Disambiguator: t와 q를 concat 한 후, sequence classification을 수행
>> 가장 유사한 edit statements를 고르기 위함

PokeMQA

1) Programmable Scope Detector

Scope Detector의 훈련은 edit statement와 이에 대응하는 sub-question으로 구성된 데이터셋을 사용

$$\mathcal{L} = -\log g(t_i, q_i) - \mathbb{E}_{q_n \sim P_n(q)} [\log(1 - g(t_i, q_n))], \quad (1)$$

훈련된 모델(DistilBERT를 사용)은 accuracy 대신 아래의 두 지표를 통해 모델을 평가

Success Rate (SR): 각 q 에 대해 올바른 edit statement t 가 가장 높은 $g(t, q)$ 를 가지는지 평가

$$SR = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \left[\bigwedge_{(t,q) \in \mathcal{D}_{val}} (g(t_i, q_i) \geq g(t, q_i)) \right], \quad (2)$$

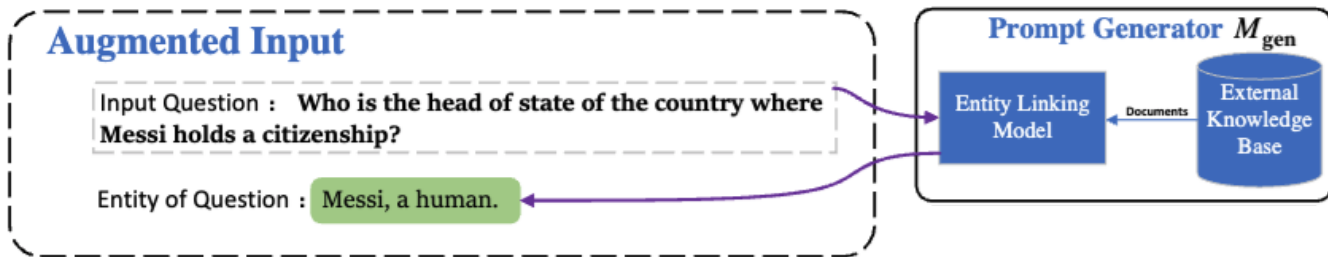
Block Rate (BR): 각 q 에 대해 올바르지 않은 t 의 $g(t, q)$ 가 0.5 미만으로 억제되는지 평가

$$BR = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \left[\bigwedge_{(t,q) \in \mathcal{D}_{val}^-} (g(t, q_i) < 0.5) \right], \quad (3)$$

PokeMQA

2) Knowledge Prompt Generator

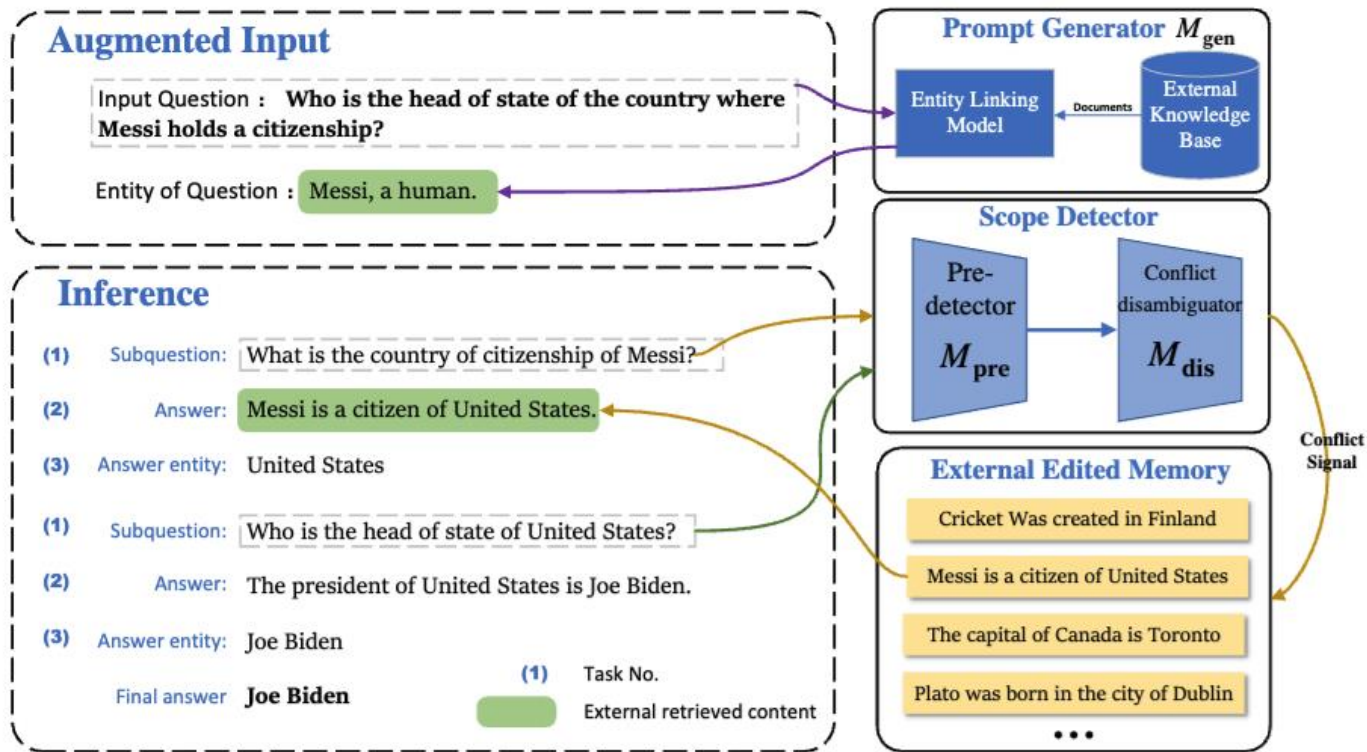
멀티홉 질문 Q에서 핵심 entity를 자동으로 인식한 후, 해당 entity와 관련된 외부 지식을 활용해 보조 문맥 정보를 생성



보조 문맥 정보를 통해, LLM은 질문의 핵심 entity를 명확히 파악할 수 있어 첫 번째 sub-question을 올바르게 도출하기 쉬워짐

>> LLM의 pre-trained knowledge에 멀티홉 질문 Q가 다루는 지식이 포함되어 있지 않을 경우 성능 향상 효과가 있음

PokeMQA



Main Result

Method	MQUAKE-CF-3K						MQUAKE-T			
	1 edited		100 edited		All edited		1 edited		All edited	
	Acc.	Hop-Acc	Acc.	Hop-Acc	Acc.	Hop-Acc	Acc.	Hop-Acc	Acc.	Hop-Acc
LLaMa-2										
Size: 7B										
FT _{COT}	22.3	-	2.13	-	OOM	-	47.32	-	3.75	-
FT	28.2	7.3	2.37	0.03	OOM	OOM	56.48	33.89	1.02	0.37
ROME _{COT}	11.17	-	2.87	-	2.77	-	28.96	-	14.4	-
ROME	13.13	5.37	3.5	0.03	3.63	0.1	24.89	17.99	1.71	0.32
MEMIT _{COT}	11.83	-	9.23	-	5.57	-	36.88	-	31.58	-
MEMIT	14.97	6.43	9.4	2.47	2.3	0.37	30.89	23.98	25.21	20.13
MeLLo	33.57	9.9	20.0	10.07	17.33	9.9	97.7	0.21	62.58	3.96
PokeMQA (Ours)	44.13	30.6	37.33	27.83	32.83	23.87	75.43	60.44	74.36	60.22
Vicuna										
Size: 7B										
MeLLo	22.7	7.03	12.83	6.77	10.9	6.7	42.24	1.12	19.86	1.28
PokeMQA (Ours)	45.83	34.8	38.77	31.23	31.63	25.3	74.57	55.19	73.07	55.09
GPT-3.5-turbo-instruct										
Size: Undisclosed										
MeLLo	57.43	28.8	40.87	28.13	35.27	25.3	88.12	52.84	74.57	53.53
PokeMQA (Ours)	67.27	56.37	56.0	49.63	45.87	39.77	76.98	68.09	78.16	67.88

Table 1: Evaluation results on MQUAKE-CF-3K and MQUAKE-T. The best result is indicated in **Bold**. The term ‘k edited’ means the size of edit batch is k. ‘COT’ means that the current method uses chain-of-thought prompt, otherwise the question decomposition prompt; The metrics are multi-hop accuracy (Acc) and Hop-wise answering accuracy (Hop-Acc) presented in Section 4.1. ‘-’ means the current metric is not applicable to this setting.

Ablation Study

Knowledge Prompt Generator, Conflict Disambiguator의 효과를 분석

M_{dis} M_{gen}	GPT-3.5-turbo-instruct				LLaMa-2-7B				Vicuna-7B				
	MQUAKE-CF-3K		MQUAKE-T		MQUAKE-CF-3K		MQUAKE-T		MQUAKE-CF-3K		MQUAKE-T		
	1 edited	All edited	1 edited	All edited	1 edited	All edited	1 edited	All edited	1 edited	All edited	1 edited	All edited	
-	-	49.0	29.93	67.99	55.67	29.33	19.47	59.31	52.19	27.37	16.43	54.23	48.39
✓	-	49.0	34.27	68.09	67.77	29.33	22.87	59.31	59.1	27.37	19.37	54.23	54.12
-	✓	56.07	33.83	68.04	56.32	30.6	20.3	60.44	53.21	34.8	22.23	55.19	49.68
✓	✓	56.37	39.77	68.09	67.88	30.6	23.87	60.44	60.22	34.8	25.3	55.19	55.09

Table 2: Ablation results of PokeMQA and its variants in terms of Hop-Acc. We also provide the results in terms of Acc. in Appendix A.

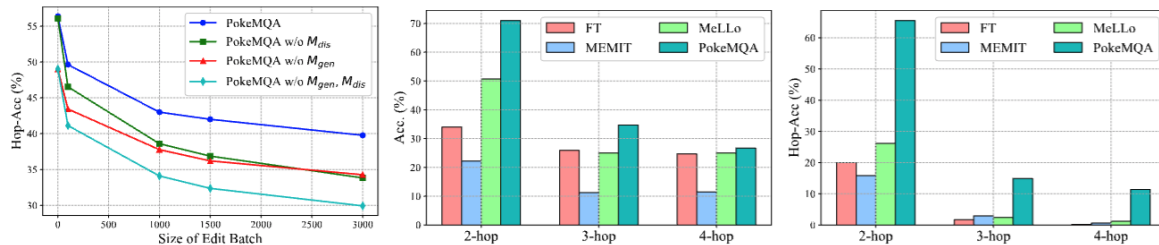
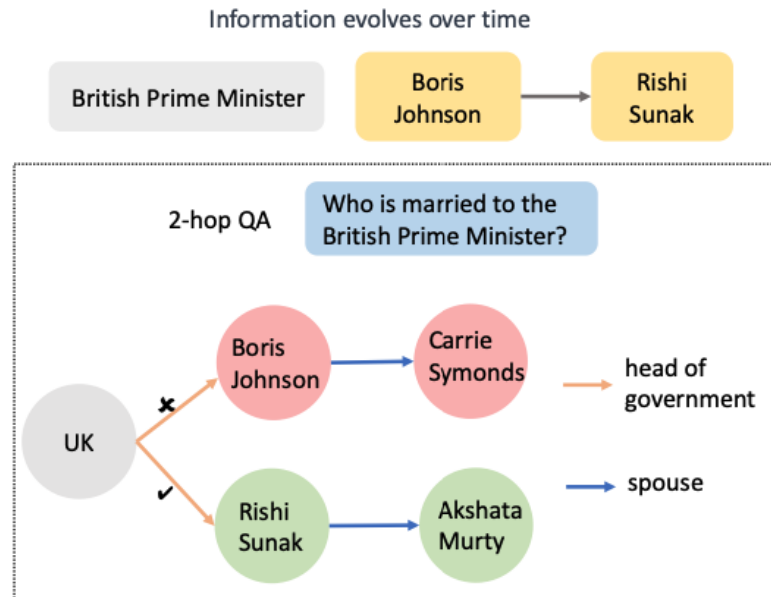


Figure 3: **Left:** Hop-Acc across multiple variants of PokeMQA with varying size of edit batch, utilizing GPT-3.5-turbo-instruct as the base language model on MQUAKE-CF-3K. **Middle, Right:** On MQUAKE-CF-3K, Acc. and Hop-Acc results for 2,3,4-hop questions, utilizing different knowledge editing methods. The experiments is conducted on LLaMa-2-7B with the size of edit batch is 1. Extra results is provided in Appendix A.

Introduction

Multi-hop QA에서 기존 지식 편집 방법들의 한계점

- 1) single-hop에 대한 edit 여부만 검증하는 경향 (파라미터를 수정하는 방법론들)
- 2) QA에 필요한 related facts를 검색해 오는 과정에 의존 + 추론이 복잡할수록 더 어려움
- 3) 지식이 시간에 따라 변하는 동적 환경에서 일관성 있게 반영하기 어려움

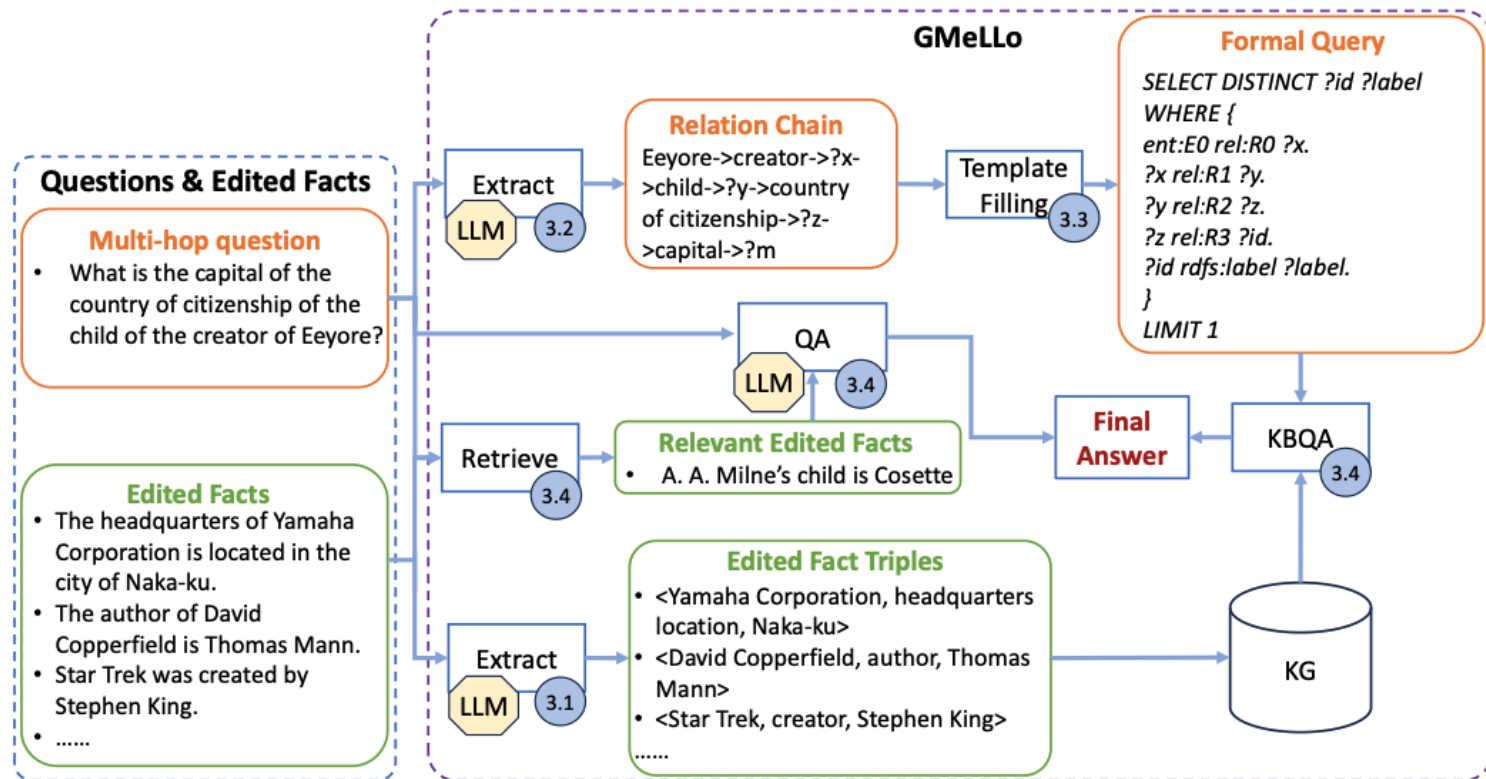


GMeLLO

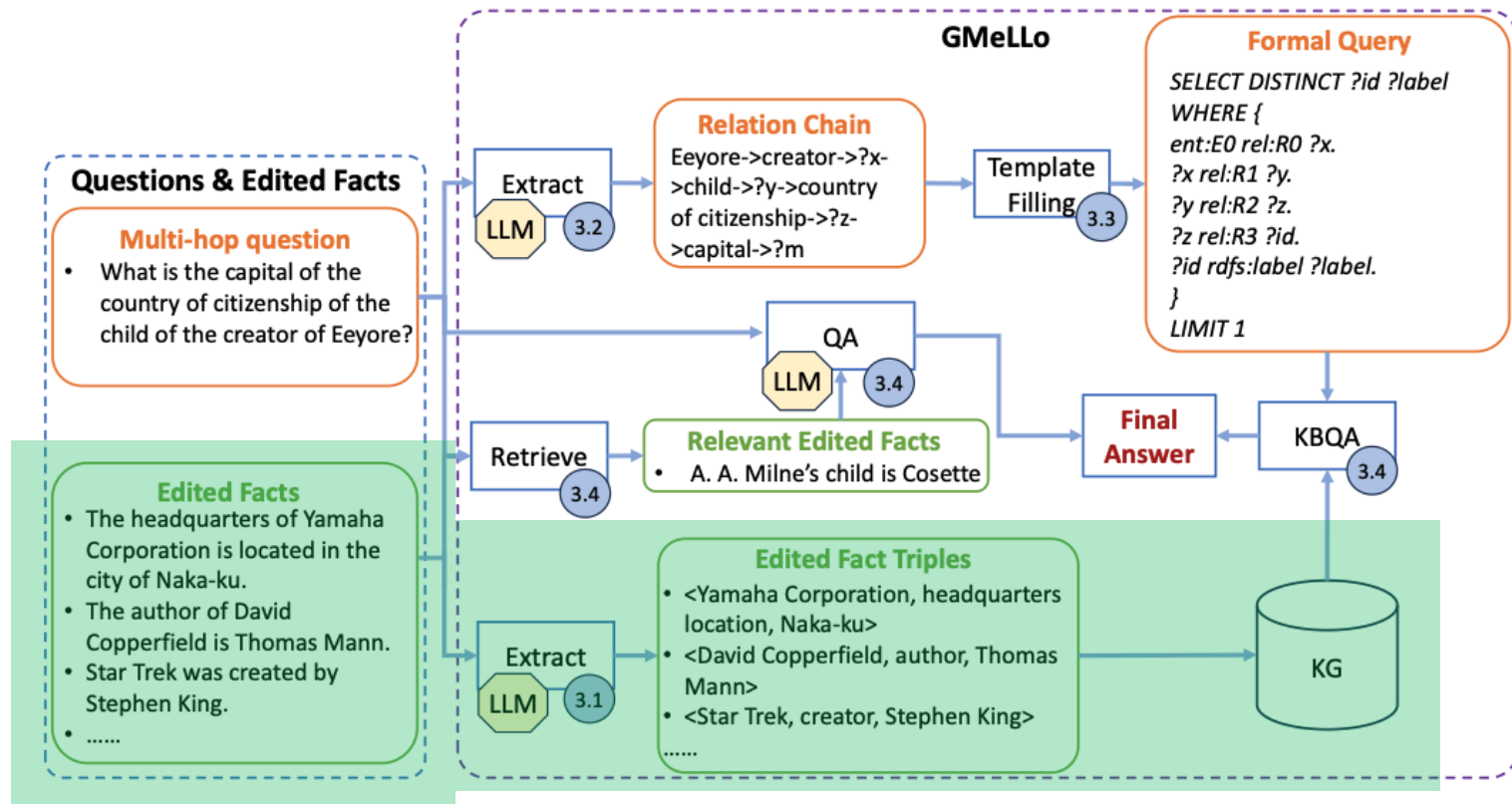
GMeLLO(Graph Memory-based Editing for LLMs)

- 1) 지식을 업데이트하기 위해 모델의 파라미터를 수정하거나, 외부의 메모리 파라미터를 수정하는 것 대신, LLM을 자연어 문장에서 획득한 정보를 KG에 반영하거나, 주어진 쿼리를 이해하는 역할로 사용
- 2) 다수의 facts가 지속적으로 업데이트되는 상황에서도 별도 모델의 훈련 없이 MQA가 가능하도록 하는 것이 목표
 - edit statement를 triple로 변환하여 KG에 업데이트
 - 복잡한 질문에서 LLM을 활용해 relation chain을 추출 -> 구조화 된 쿼리로 변환

GMeLLO



GMeLLO



Extracting Fact Triples

Edit facts(자연어 문장)가 주어지면, $\langle s, r, o \rangle$ 의 fact triple을 추출한 후, KG에 업데이트

- Few-shot prompt + instruction을 통해 triple을 추출하는 과정을 in-context learning
- Triple의 relation을 설정할 때, 사전에 정의해 둔 relation 리스트를 prompt에 포함시킴
- Subject, object는 문장에서 추출

>>> 최종적으로 자연어 문장을 입력으로 KG에 표준화된 $\langle s, r, o \rangle$ 를 추출한 후,

KG (Wikidata)에서 동일한 $\langle s, r, ? \rangle$ triple을 찾아서 대체하거나 추가

Prompt for Transforming the Edited Sentences to Triples

Sentence: The headquarters of University of Cambridge is located in the city of Washington, D.C.

Relation Chain: University of Cambridge->headquarters location->Washington, D.C.

.....

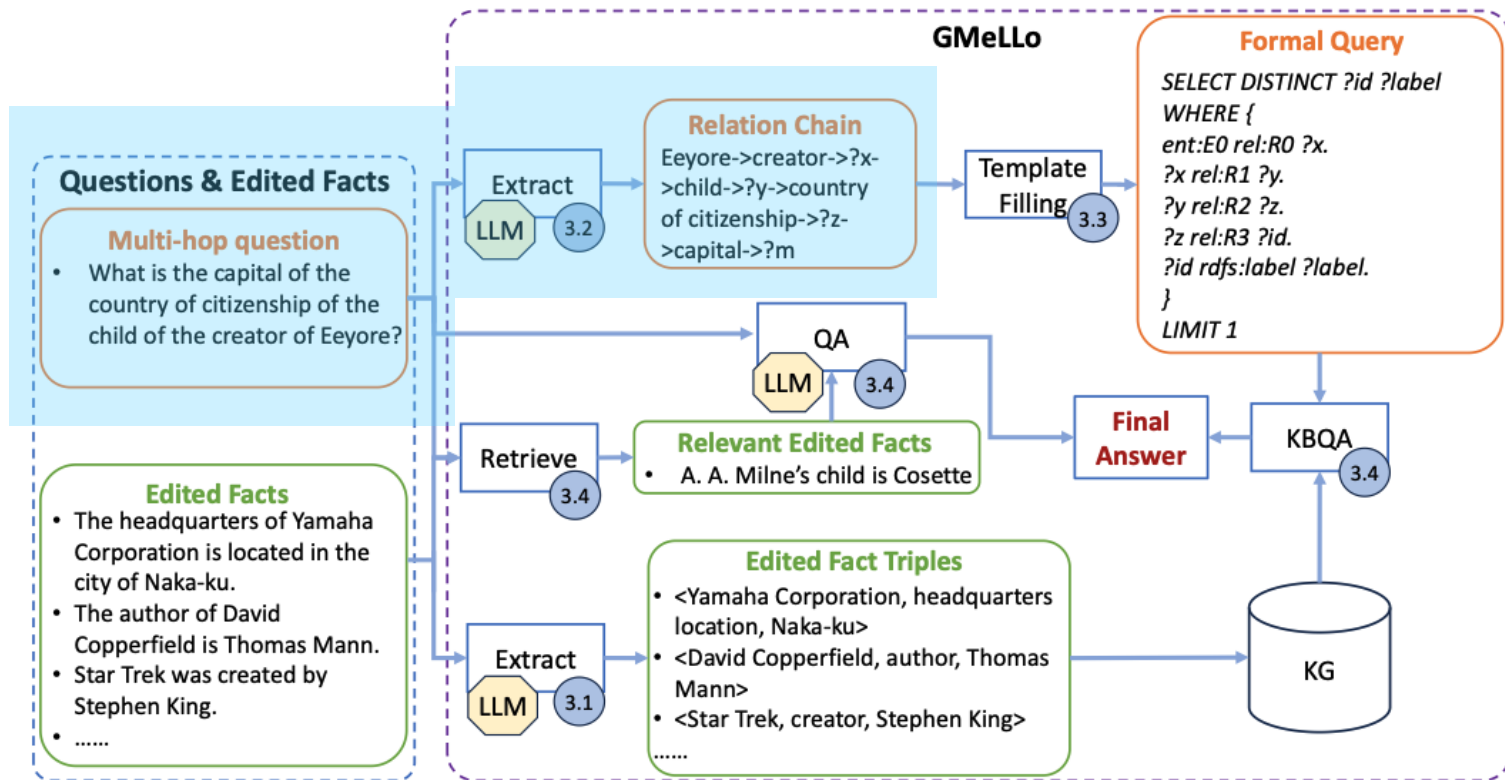
Given the above samples, please help me analyze the relation chain of the following sentence. All the relations should be selected from ['country of origin', 'sport', ...].

Sentence: The chief executive officer of Boeing is Marc Benioff

Relation Chain:

Figure 3: The prompt used for transforming edited fact sentences to triples.

GMeLLO



Extracting Relation Chain from Question Using LLMs

Multi-hop QA에서는 단일 facts 만으로 답을 구할 수 없고, 여러 facts를 chain하여 추론해야 함

e.g.: “What is the capital of the country of citizenship of the child of creator of Eeyore?”

GMeLLo는 질문에 명시적으로 나타난 Entity (Eeyore)를 중심으로 문장의 의미를 해석하고, 어떤 relation으로 어떤 Entity와 연결되는지를 추출

Few-shot prompt와 instruction으로 relation chain을 추출하는 작업을 in-context learning

과정을 거치면 아래와 같은 relation chain을 추출

Question

What is the capital of the country of citizenship of the child of the creator of Eeyore?

Relation Chain

Eeyore->creator->?x->child->?y
->country of citizenship
->?z->capital->?m

Prompt for Transforming the Question Sentences to Relation Chains

Question: What is the birthplace of the author of "The Little Match Girl"?

Relation Chain: The Little Match Girl->author->?x->place of birth->?y

.....

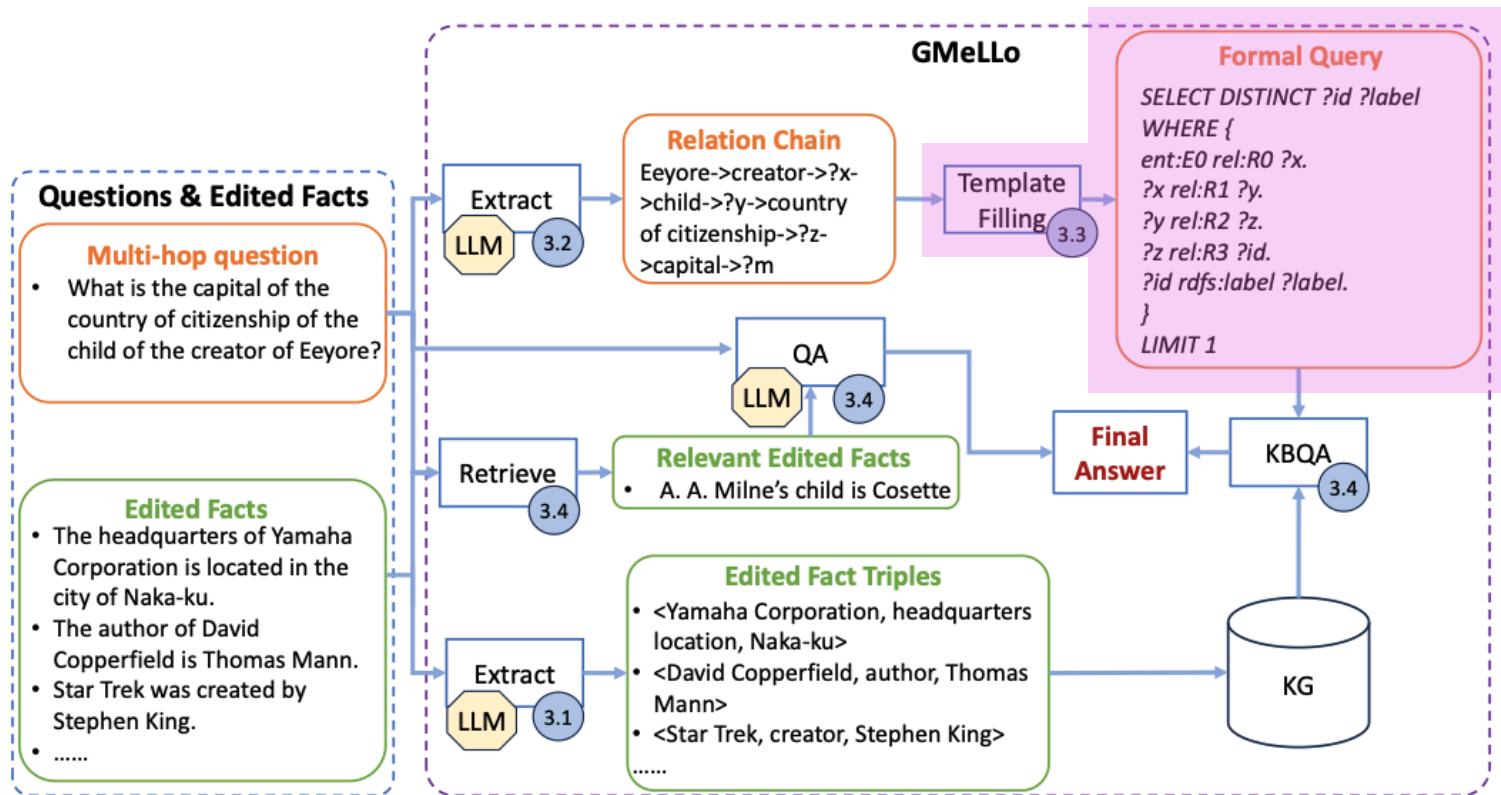
Given the above samples, please help me analyze the relation chain of the following sentence. All the relations should be selected from ['country of origin', 'sport', ...].

Question: What is the continent where the CEO responsible for developing Windows 8.1 was born?

Relation Chain:

Figure 4: The prompt used for transforming question sentences to relation chains.

GMeLLO



Converting a Relation Chain into a Formal Query

자연어 문장 형태의 Question에서 추출한 Relation Chain은 질의응답을 위한 path로 사용됨

GMeLLo에서는 구조화된 SPARQL로 변환하기 위해, 사전에 정의한 템플릿에 추출된 relation chain을 끼워 넣음

Question

What is the capital of the country of citizenship of the child of the creator of Eeyore?

Relation Chain

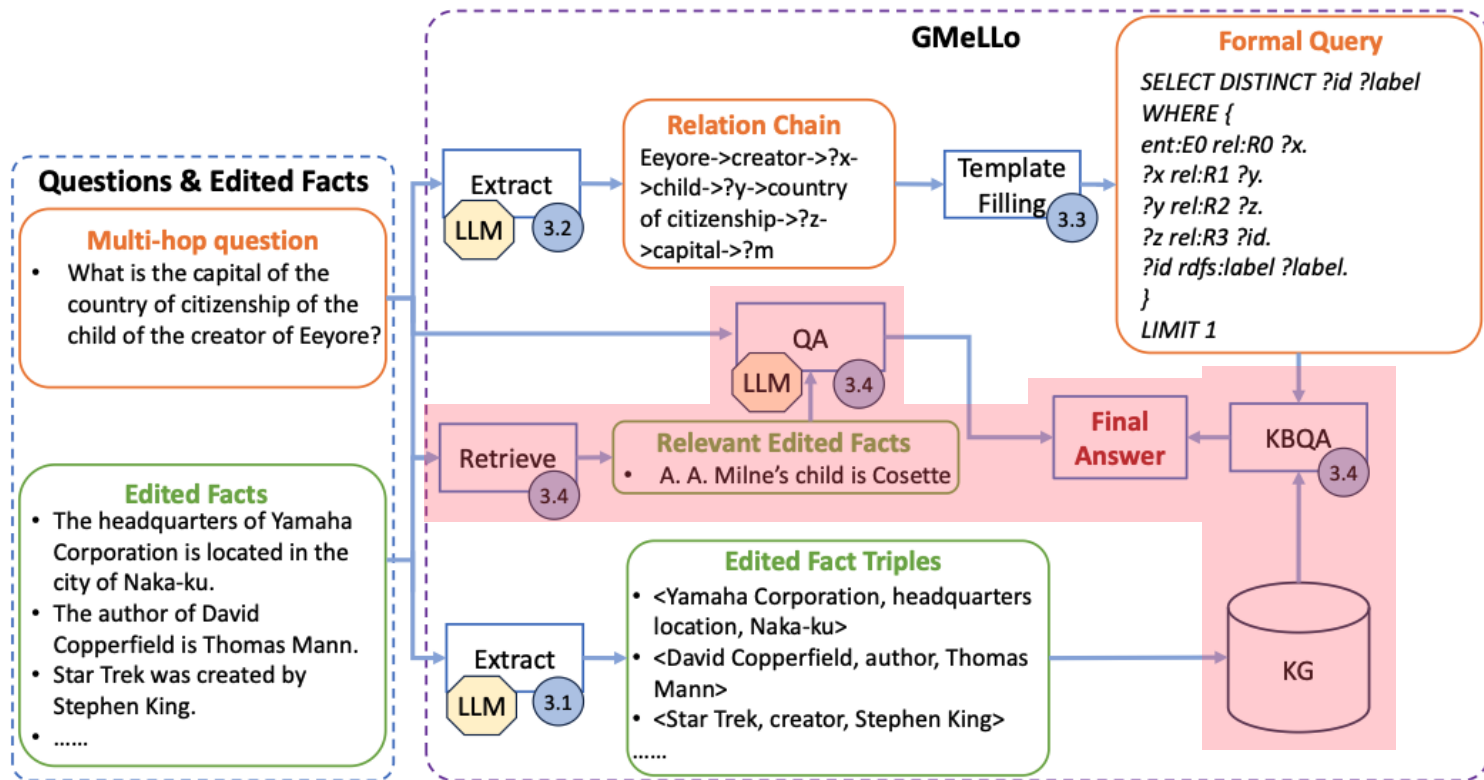
Eeyore->creator->?x->child->?y
->country of citizenship
->?z->capital->?m

```

PREFIX ent: <http://www.kg/entity/>
PREFIX rel: <http://www.kg/relation/>
SELECT DISTINCT ?id ?label WHERE {
  ent:E0 rel:R0 ?x.
  ?x rel:R1 ?y.
  ?y rel:R2 ?z.
  ?z rel:R3 ?id.
  ?id rdfs:label ?label.
}
LIMIT 1
  
```

Relation Chain에 해당하는
정답 Entity의 ID와 이름

GMeLLO



Integrating LLM-based QA and KBQA

여태까지 설명한대로,

자연어 문장 형태의 Multi-hop Question이 입력되면, 두 가지 방식으로 답변을 도출함

- 1) **Contriever(Izacard et al., 2022) 모델을 사용하여 edit statements의 목록에서 상위 k 개의 related facts를 검색 >> 검색한 facts를 Question과 함께 LLM에 입력하여 답변을 생성**
- 2) **자연어 문장에서 relation chain을 추출한 후, Edited KG를 활용하여 KBQA 식으로 답변을 검색**

Edit triple과 Relation chain이 올바르게 생성되면, KBQA 시스템은 올바른 답변을 생성함

만약 relation chain이 잘못 추출된 경우 >> 1) 의 방법으로 생성한 답변을 최종 답변으로 채택

Main Result

Base Model	Method	MQuAKE-CF				MQuAKE-T			
		k=1	k=100	k=1000	k=3000	k=1	k=100	k=500	k=1868
GPT-J-6B	MEMIT	12.3	9.8	8.1	1.8	4.8	1.0	0.2	0.0
	MEND	11.5	9.1	4.3	3.5	38.2	17.4	12.7	4.6
	MeLLo	20.3	12.5	10.4	9.8	85.9	45.7	33.8	30.7
	GMeLLo	76.3	53.4	49.5	49.0	86.9	82.1	81.5	81.5
Vicuna-7B	MeLLo	20.3	11.9	11.0	10.2	84.4	56.3	52.6	51.3
	PokeMQA	45.8	38.8	-	31.6	74.6	-	-	73.1
	GMeLLo	71.3	46.5	42.5	41.9	97.1	86.3	85.4	85.1

Table 1: Performance comparison of GMeLLo and other approaches on the MQuAKE-CF and MQuAKE-T datasets using GPT-J-6B or Vicuna-7B as the base language models. Adhering to the methodology outlined by [Zhong et al. \(2023\)](#), instances are grouped into batches of size k . For the MQuAKE-CF dataset, k varies from 1 to 3000, and for the MQuAKE-T dataset, it ranges from 1 to 1868. For example, in the MQuAKE-CF dataset, when $k = 100$, the 3000 instances are organized into 30 groups, and the average performance reported as the final result. The metric used is accuracy.






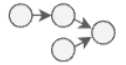
Ablation Study

Base Model	Method	MQuAKE-CF				MQuAKE-T			
		k=1	100	1000	3000	k=1	100	500	1868
GPT-J-6B	QA	71.0	24.2	14.3	12.2	32.3	18.0	15.7	15.5
	KBQA	43.3	43.3	43.3	43.3	80.2	80.2	80.2	80.2
	GMeLLo	76.3	53.4	49.5	49.0	86.9	82.1	81.5	81.5
Vicuna-7B	QA	72.6	27.0	16.5	13.5	96.9	63.0	59.2	58.2
	KBQA	35.9	35.9	35.9	35.9	73.6	73.6	73.6	73.6
	GMeLLo	71.3	46.5	42.5	41.9	97.1	86.3	85.4	85.1

Table 2: Ablation study of GMeLLo. QA involves directly using LLM for answering the multi-hop questions. KBQA involves using LLM to transform edited fact sentences into triples, update WikiData, convert question sentences into relation chains, and generate formal KG queries for question answering. GMeLLo combines these methods by using KBQA to correct answers from LLM-based QA.

Conclusion

- 1) Multi-hop QA에 edit facts를 적용하는 연구는 아직까지 거의 in-context learning 위주로 이루어지고 있음
- 2) 멀티홉 추론을 atomic sub-question으로 쪼개는 방법에 관해서는 의견이 나뉨 (좋다? 좋지 않다?)
- 3) 아직까지 KE for Multi-hop QA에서는 atomic facts가 chain 형태로 이어진 것만 다루고 있음 -> 앞으로 더 발전할 여지가 많다

Graph	Question	Decomposition
	Who succeeded the first President of Namibia? Hifikepunye Pohamba	<ol style="list-style-type: none"> 1. Who was the first President of Namibia? Sam Nujoma 2. Who succeeded Sam Nujoma? Hifikepunye Pohamba
	What currency is used where Billy Giles died? pound sterling	<ol style="list-style-type: none"> 1. At what location did Billy Giles die? Belfast 2. What part of the UK is Belfast located in? Northern Ireland 3. What is the unit of currency in Northern Ireland? pound sterling
	When was the first establishment that McDonaldization is named after, open in the country Hordean is located? 1974	<ol style="list-style-type: none"> 1. What is McDonaldization named after? McDonald's 2. Which state is Hordean located in? England 3. When did the first McDonald's open in England? 1974
	When did Napoleon occupy the city where the mother of the woman who brought Louis XVI style to the court died? 1805	<ol style="list-style-type: none"> 1. Who brought Louis XVI style to the court? Marie Antoinette 2. Who's mother of Marie Antoinette? Maria Theresa 3. In what city did Maria Theresa die? Vienna 4. When did Napoleon occupy Vienna? 1805
	How many Germans live in the colonial holding in Aruba's continent that was governed by Prazeres's country? 5 million	<ol style="list-style-type: none"> 1. What continent is Aruba in? South America 2. What country is Prazeres? Portugal 3. Colonial holding in South America governed by Portugal? Brazil 4. How many Germans live in Brazil? 5 million
	When did the people who captured Malakoff come to the region where Philipsburg is located? 1625	<ol style="list-style-type: none"> 1. What is Philipsburg capital of? Saint Martin 2. Saint Martin is located on what terrain feature? Caribbean 3. Who captured Malakoff? French 4. When did the French come to the Caribbean? 1625