질문 유형에 따른 검색 증강 생성의 질의응답 성능 분석

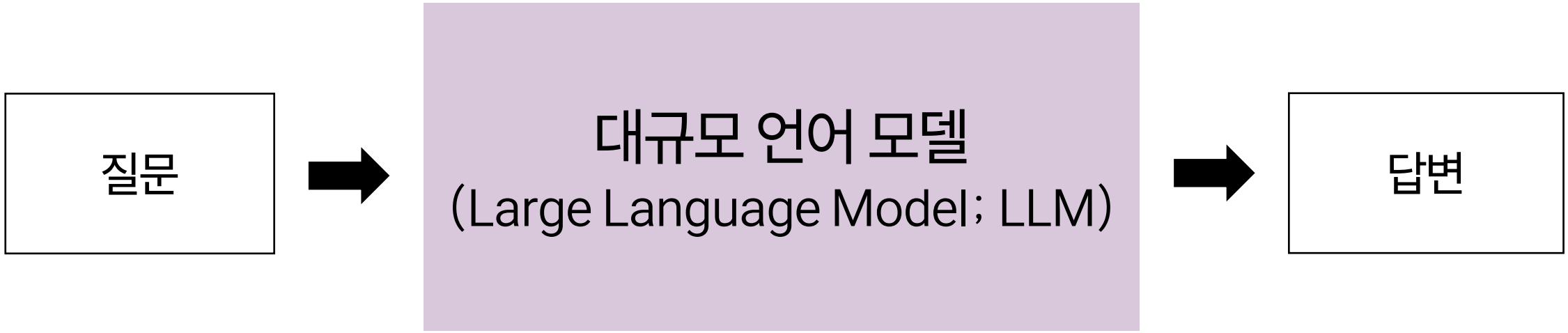
Question Types Matter: An Analysis of Question-Answering Performance in Retrieval-Augmented Generation Across Diverse Question Types

이동건*, 박아정*, 이혜리, 남현서, 맹윤호

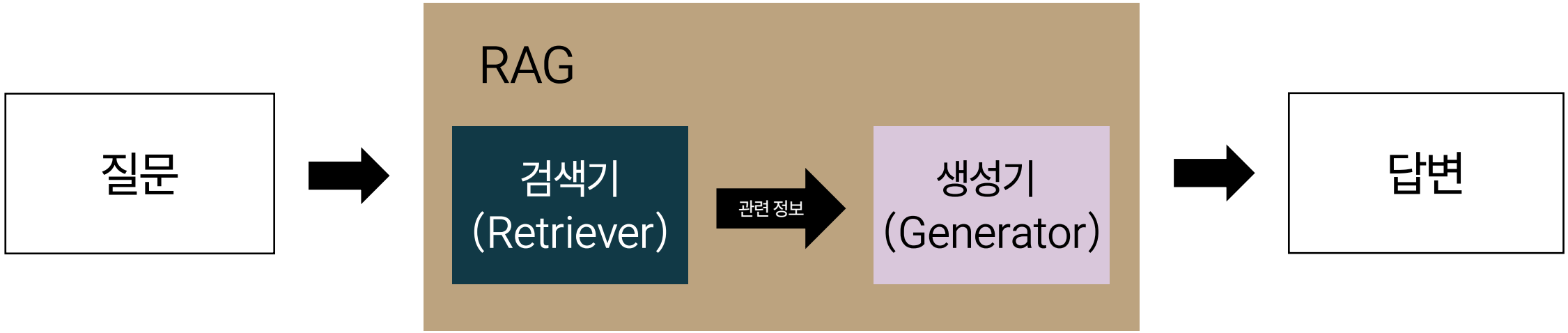
donggeonlee@postech.ac.kr

POSTECH TmaxCoreAI kt 이화여자대학교

LLM 기반 질의응답 시스템



검색 증강 생성 (Retrieval-Augmented Generation; RAG)



[1] Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," In *NeurIPS*, 2020.

질문 유형 분류

사실적 질문 (Factoid Question)

간단하고 명확한 사실적 정보를 요구하는 질문

- 예시:
  - "대한민국의 수도는 어디인가요?"
  - "한글날은 언제인가요?"
  - "물의 화학식은 무엇인가요?"

비사실적 질문 (Non-Factoid Question)

더 복잡하고 상세한 설명이나 추론을 요구하는 질문

- 예시:
  - "AI가 사회에 미치는 영향은 무엇인가요?"
  - "아이폰16을 추천하시겠습니까?"
  - "발라드는 트로트에 비해 어떻게 다른가요?"

[2] Bolotova et al. "A Non-Factoid Question-Answering Taxonomy," In *SIGIR*, 2022.
[3] Reddy et al. "A Survey on Types of Question Answering System," *IOSR Journal of Computer Engineering*, Vol. 19, pp. 19–23, 2017.
[4] Mohasseb et al. "Factoid vs. Non-factoid Question Identification: An Ensemble Learning Approach," *International Conference on Web Information Systems and Technologies*, 2022.

질문 유형 분류

사실적 질문 (Factoid Question)

간단하고 명확한 사실적 정보를 요구하는 질문

비사실적 질문 (Non-Factoid Question)

더 복잡하고 상세한 설명이나 추론을 요구하는 질문

비질문 (Not A Question)

질문 형태가 아닌 문장

- 예시:
  - “이 곳은 대한민국의 수도이다.”
  - “대한민국의 수도는 ____이다.”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.
[3] Reddy et al. “A Survey on Types of Question Answering System,” *IOSR Journal of Computer Engineering*, Vol. 19, pp. 19–23, 2017.
[4] Mohasseb et al. “Factoid vs. Non-factoid Question Identification: An Ensemble Learning Approach,” *International Conference on Web Information Systems and Technologies*, 2022.

비사실적 질문의 6가지 유형 [2] – (1/6)

- **토론 (debate)**
  - 가설적 질문을 물어보는 유형
  - 상반된 의견이 제시되는 답변 구조

- **예시 패턴**
  - “...가 ...라고 생각하시나요?”
  - “...가 성공할 수 있나요?”
  - ...가 정말 ...인가요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적 질문의 6가지 유형 [2] – (2/6)

- **증거 기반 (evidence-based)**
  - 특정 개념이나 사건의 정의를 요구
  - 사실에 기반한 설명이 답변으로 요구

- **예시 패턴**
  - “…는 어떻게 작동하나요?”
  - “…의 의미는 무엇인가요?”
  - “…를 어떻게 설명하나요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적 질문의 6가지 유형 [2] – (3/6)

- **지침 (instruction)**
  - 절차나 방법을 이해 하려는 질문
  - 단계별 지침이 제공되는 답변 구조

- **예시 패턴**
  - “...를 어떻게 하나요?”
  - “...의 과정은 무엇인가요?”
  - “...하는 가장 좋은 방법은 무엇인가요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적 질문의 6가지 유형 [2] – (4/6)

- 이유 (reason)
  - 현상의 원인을 찾는 질문
  - 여러 근거와 설명이 포함하는 답변 구조

- 예시 패턴
  - “...의 이유는 무엇인가요?”
  - “무엇이 ...를 유발하나요?”
  - “어떻게 ...가 일어났나요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적 질문의 6가지 유형 [2] – (5/6)

- 경험 (experience)
  - 조언이나 추천을 구하는 질문
  - 개인적 경험에 대한 설명이 답변

- 예시 패턴
  - “…를 추천하시겠습니까?”
  - “…에 대해 어떻게 생각하시나요?”
  - “제가 … 해야 하나요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적 질문의 6가지 유형 [2] – (6/6)

- 비교 (comparison)
  - 차이점과 유사점을 파악하는 질문
  - 답변에는 대상 간의 차이와 유사점을 나열

- 예시 패턴
  - “X는 Y와 어떻게 … 하나요?”
  - “X가 Y보다 …한 점은 무엇인가요?”
  - “X는 Y에 비해 어떻게 … 하나요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

RAG 관련 기존 연구

- Adaptive-RAG [5]
  - **질문의 복잡성**에 따라 검색 전략을 유동적으로(adaptive) 사용
- RQ-RAG [6]
  - **질문의 복잡성**이나 **질문의 모호성을 판별**하고 질문을 재작성(rewriting)하여 사용

[5] Jeong et al, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," In *NAACL,* 2024.
[6] Chan et al, "RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation," In *COLM,* 2024.

RAG 관련 기존 연구

- Adaptive-RAG [5]
  - **질문의 복잡성**에 따라 검색 전략을 유동적으로(adaptive) 사용
- RQ-RAG [6]
  - **질문의 복잡성**이나 **질문의 모호성을 판별**하고 질문을 재작성(rewriting)하여 사용

대부분의 관련 연구들은 주로 **사실적 질문**에 초점이 맞추어져 있음

[5] Jeong et al, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," In *NAACL*, 2024.
[6] Chan et al, "RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation," In *COLM*, 2024.

Research Question

1. LLM 및 RAG 기반 질의응답 시스템은 **질문의 여러 형태에 따라** 다른 성능을 보일 것인가?
2. 질의응답 시스템의 성능은 **비사실적 질문의 세부 유형에 따라** 어떻게 달라질 것인가?
3. **단일 추론(single-hop) 및 다중 추론(multi-hop) 질문**에 대한 질의응답 시스템의 성능은 질문의 형태에 따라 달라질 것인가?

실험 방법: 데이터셋 구성

- 7개의 질의응답 데이터셋을 통해 LLM 및 RAG의 성능 평가
  - 단일 추론(single-hop) 질의응답 데이터셋
    - MS MARCO
    - Natural Questions (NQ)
    - TriviaQA
    - SQuAD
  - 다중 추론(multi-hop) 질의응답 데이터셋
    - HotpotQA
    - 2WikiMultiHopQA (2WMHQA)
    - MuSiQue

- 각 데이터셋 별 18,000건 씩, 총 126,000건의 데이터셋 구성

실험 방법: 질문 유형 다중 분류

데이터셋의 모든 질문-답변 쌍은 아래와 같이 8가지 질문 유형으로 다중 분류

1. 비질문(not a question)
2. 사실적 질문(factoid question)
3. 토론
4. 증거 기반
5. 지침
6. 이유
7. 경험
8. 비교

3~8: 비사실적 질문 (non-factoid question)

실험 방법: 질문 유형 다중 분류

- 다중 분류에는 RoBERTa 기반의 Lurunchik/nf-cats [2, 7] 모델을 활용
  - 질문 유형 다중 분류에서 0.901의 F1-score을 달성함 [2]

- 데이터셋 내 질문 유형 분포

비질문	사실적 질문	비사실적 질문						합계
		토론	증거 기반	지침	이유	경험	비교	
1,481 (1.18%)	104,457 (82.90%)	4,959 (3.94%)	11,601 (9.21%)	922 (0.73%)	1,159 (0.92%)	727 (0.58%)	694 (0.55%)	126,000 (100%)

[2] Bolotova et al. "A Non-Factoid Question-Answering Taxonomy," In *SIGIR*, 2022.
[7] "Lurunchik/nf-cats," *Hugging Face Hub*, https://huggingface.co/Lurunchik/nf-cats, 2022.

실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06			
		LLM		RAG	
		EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06
사실적 질문		7.34	23.39	19.32	33.17
비사실적 질문	토론	0.56	3.0	1.05	4.66
	증거 기반	3.53	**23.34**	5.39	**31.44**
	지침	0.87	19.77	1.53	28.71
	이유	1.12	15.11	2.16	21.7
	경험	2.06	6.13	3.98	10.01
	비교	**4.18**	18.78	**10.37**	27.86

실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06			
		LLM		RAG	
		EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06
사실적 질문		7.34	23.39	19.32	33.17
비사실적 질문	토론	0.56	3.0	1.05	4.66
	증거 기반	3.53	**23.34**	5.39	**31.44**
	지침	0.87	19.77	1.53	28.71
	이유	1.12	15.11	2.16	21.7
	경험	2.06	6.13	3.98	10.01
	비교	**4.18**	18.78	**10.37**	27.86

실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06			
		LLM		RAG	
		EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06
사실적 질문		7.34	23.39	19.32	33.17
비사실적 질문	토론	0.56	3.0	1.05	4.66
	증거 기반	3.53	**23.34**	5.39	**31.44**
	지침	0.87	19.77	1.53	28.71
	이유	1.12	15.11	2.16	21.7
	경험	2.06	6.13	3.98	10.01
	비교	**4.18**	18.78	**10.37**	27.86

실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06			
		LLM		RAG	
		EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06
사실적 질문		7.34	23.39	19.32	33.17
비사실적 질문	토론	0.56	3.0	1.05	4.66
	증거 기반	3.53	**23.34**	5.39	**31.44**
	지침	0.87	19.77	1.53	28.71
	이유	1.12	15.11	2.16	21.7
	경험	2.06	6.13	3.98	10.01
	비교	**4.18**	18.78	**10.37**	27.86

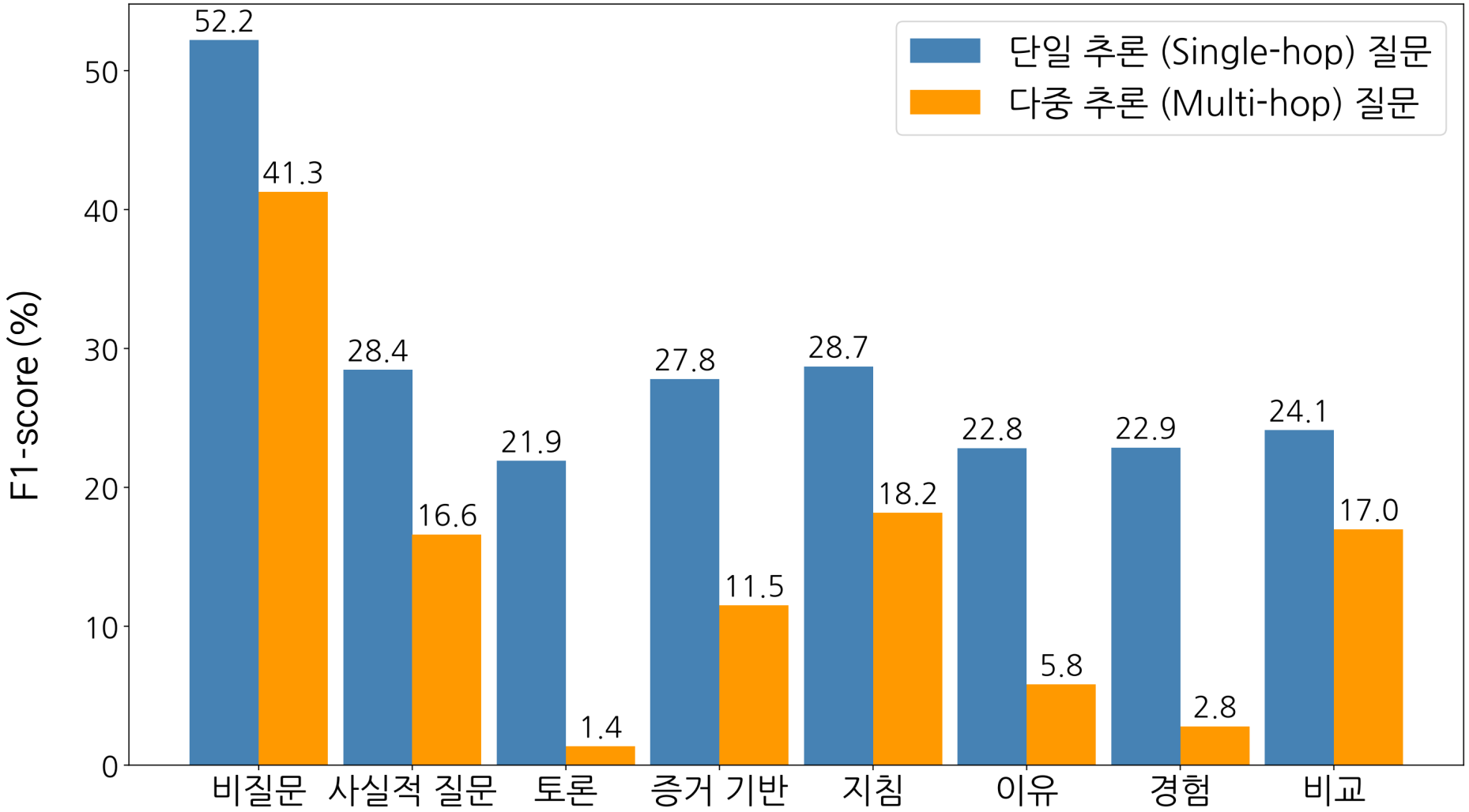
실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06			
		LLM		RAG	
		EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06
사실적 질문		7.34	23.39	19.32	33.17
비사실적 질문	토론	0.56	3.0	1.05	4.66
	증거 기반	3.53	**23.34**	5.39	**31.44**
	지침	0.87	19.77	1.53	28.71
	이유	1.12	15.11	2.16	21.7
	경험	2.06	6.13	3.98	10.01
	비교	**4.18**	18.78	**10.37**	27.86

실험 결과: 단일 및 다중 추론 질문에 대한 RAG 성능 비교

(Generator: `gpt-4o-mini-2024-07-18`)



결론

- LLM 및 RAG 기반 질의응답 시스템은 **질문의 형태에 따라 성능 차이가 발생**한다.
  - 비질문 > 사실적 질문 > 비사실적 질문

- 질의응답 시스템의 성능은 **비사실적 질문의 세부 유형에 따라 달라진다**.
  - 토론, 이유, 경험 등의 질문 유형에서는 상대적으로 낮은 성능을 보임

- 질의응답 시스템의 성능은 단일 추론 질문보다 **다중 추론 질문에서 감소**한다.
  - 비사실적 질문 유형에서는 타 유형보다 감소폭이 큼

한계 및 향후 연구과제

- F1-score 등과 같은 **기존 질의응답 평가지표로 비사실적 질문을 평가하는 데에는 한계**가 있다.
  - 비사실적 질문의 질의응답 평가를 위한 새로운 평가지표 연구 필요

- **다중추론이 요구되는 비사실적 질문을 사용했을 때**의 질의응답 시스템의 성능이 낮다.
  - 질문 유형에 영향을 받지 않는 강건한(robust) 질의응답 시스템 연구 필요
  - 각 질문 유형에 맞는 차별화된 접근법을 반영한 질의응답 시스템 연구 필요

질문 유형에 따른 검색 증강 생성의 질의응답 성능 분석

Question Types Matter: An Analysis of Question-Answering Performance in Retrieval-Augmented Generation Across Diverse Question Types

이동건*, 박아정*, 이혜리, 남현서, 맹윤호

donggeonlee@postech.ac.kr

본 연구는 2024년 국가과학기술인력개발원(KIRD)의 지원을 받아 수행된 성과물임.

질문 유형에 따른 검색 증강 생성의 질의응답 성능 분석

Question Types Matter: An Analysis of Question-Answering Performance in Retrieval-Augmented Generation Across Diverse Question Types

Appendix

비사실적 질문의 유형 [2]

- **토론 (debate)**
  - 가설적 질문을 물어보는 유형
  - 상반된 의견이 제시되는 답변 구조
- **증거 기반 (evidence-based)**
  - 특정 개념이나 사건의 정의를 요구
  - 사실에 기반한 설명이 답변으로 제공
- **지침 (instruction)**
  - 절차나 방법을 이해 하려는 질문
  - 단계별 지침이 제공되는 답변 구조
- **이유 (reason)**
  - 현상의 원인을 찾는 질문
  - 여러 근거와 설명이 포함하는 답변 구조
- **경험 (experience)**
  - 조언이나 추천을 구하는 질문
  - 개인적 경험에 대한 설명이 답변
- **비교 (comparison)**
  - 차이점과 유사점을 파악하는 질문
  - 답변에는 대상 간의 차이와 유사점을 나열

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적 질문의 유형과 그에 따른 패턴 [2]

- 토론 (debate)
  - “...가 ...라고 생각하시나요?”
  - “...가 성공할 수 있나요?”
  - ...가 정말 ...인가요?”
- 증거 기반 (evidence-based)
  - “...의 의미는 무엇인가요?”
  - “...는 어떻게 작동하나요?”
  - “...를 어떻게 설명하나요?”
- 지침 (instruction)
  - “...를 어떻게 하나요?”
  - “...의 과정은 무엇인가요?”
  - “...하는 가장 좋은 방법은 무엇인가요?”
- 이유 (reason)
  - “...의 이유는 무엇인가요?”
  - “무엇이 ...를 유발하나요?”
  - “어떻게 ...가 일어났나요?”
- 경험 (experience)
  - “...를 추천하시겠습니까?”
  - “...에 대해 어떻게 생각하시나요?”
  - “제가 ... 해야 하나요?”
- 비교 (comparison)
  - “X는 Y와 어떻게 ... 하나요?”
  - “X가 Y보다 ...한 점은 무엇인가요?”
  - “X는 Y에 비해 어떻게 ... 하나요?”

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적인 질문의 유형과 그에 따른 패턴 [2]

질문 유형	패턴
토론 (debate)	...가 ...라고 생각하시나요? ...가 성공할 수 있나요? ...가 정말 ...인가요?
증거 기반 (evidence-based)	...의 의미는 무엇인가요? ...는 어떻게 작동하나요? ...를 어떻게 설명하나요?
지침 (instruction)	...를 어떻게 하나요? ...의 과정은 무엇인가요? ...하는 가장 좋은 방법은 무엇인가요?
이유 (reason)	...의 이유는 무엇인가요? 무엇이 ...를 유발하나요? 어떻게 ...가 일어났나요?
경험 (experience)	...를 추천하시겠습니까? ...에 대해 어떻게 생각하시나요? 제가 ... 해야 하나요?
비교 (comparison)	X는 Y와 어떻게 ... 하나요? X가 Y보다 ...한 점은 무엇인가요? X는 Y에 비해 어떻게 ... 하나요?

표 1. 비사실적인 질문의 유형과 그에 따른 패턴

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

비사실적인 질문의 유형과 그에 따른 패턴 [2]

Table 1: The proposed taxonomy of NFQ categories and target answer structures

Category	Description	Expected Answer Structure	Patterns
INSTRUCTION	You want to understand the **procedure/method** of doing/achieving something.	Instructions/guidelines provided in a step-by-step manner.	How to ...? How can I do …? What is the process for …? What is the best way to …?
REASON	You want to find out **reasons** of/for something.	A list of reasons with evidence.	Why does …? What is the reason for …? What causes …? How come ... happened?
EVIDENCE-BASED	You want to learn about the **features/description/definition** of a concept/idea/object/event.	Wikipedia-like passage describing/defining an event/object or its properties based only on facts.	What is …? How does/do … work? What are the properties of …? What is the meaning of …? How do you describe …?
COMPARISON	You want to **compare/contrast** two or more things, understand their differences/similarities.	A list of key differences and/or similarities of something compared to another thing.	How is X … to/from Y? What are the … of X over Y? How does X … against Y?
EXPERIENCE	You want to **get advice** or **recommendations** on a particular topic.	Advantages, disadvantages, and main features of an entity (product, event, person, etc) summarised from personal experiences.	Would you recommend …? How do you like …? What do you think about …? Should I …?
DEBATE	You want to **debate on a hypothetical question** (is someone right or wrong, is some event perceived positively or negatively?).	Arguments on a debatable topic consisting of different opinions on something supported or weakened by pros and cons of the topic in the question.	Does … exist? Can … be successful? Do you think … are …? Is … really a …?

[2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.

데이터셋 내 질문 유형 분포

비질문	사실적 질문	비사실적 질문						합계
		토론	증거 기반	지침	이유	경험	비교	
1,481 (1.18%)	104,457 (82.90%)	4,959 (3.94%)	11,601 (9.21%)	922 (0.73%)	1,159 (0.92%)	727 (0.58%)	694 (0.55%)	126,000 (100%)

데이터셋 내 질문 유형 분포

질문 유형		단일 추론 질의응답 데이터셋	다중 추론 질의응답 데이터셋	전체 데이터셋
비질문		1.02%	1.38%	1,481 (1.18%)
사실적 질문		79.63%	87.26%	104,457 (82.90%)
비사실적 질문	토론	0.79%	8.13%	4,959 (3.94%)
	증거 기반	15.76%	0.47%	11,601 (9.21%)
	지침	1.19%	0.12%	922 (0.73%)
	이유	1.23%	0.51%	1,159 (0.92%)
	경험	0.20%	1.08%	727 (0.58%)
	비교	0.19%	1.04%	694 (0.55%)
합계		100% (72,000건)	100% (54,000건)	100% (126,000건)

데이터셋 내 질문 유형 분포

질문 유형		단일 추론 질의응답 데이터셋				다중 추론 질의응답 데이터셋			전체 데이터셋 (126,000건)
		MARCO	NQ	TriviaQA	SQuAD	HotpotQA	2WMHQA	MuSiQue	
비질문		0.12%	0.18%	3.55%	0.23%	3.96%	0.01%	0.18%	1,481 (1.18%)
사실적 질문		50.01%	96.24%	90.02%	82.26%	87.29%	77.27%	97.23%	104,457 (82.90%)
비사실적 질문	토론	0.42%	0.01%	0.36%	2.36%	5.58%	18.32%	0.50%	4,959 (3.94%)
	증거 기반	43.18%	3.11%	5.46%	11.28%	0.42%	0.27%	0.73%	11,601 (9.21%)
	지침	3.78%	0.14%	0.09%	0.74%	0.08%	0.06%	0.22%	922 (0.73%)
	이유	2.14%	0.26%	0.20%	2.32%	0.05%	0.95%	0.53%	1,159 (0.92%)
	경험	0.08%	0.02%	0.23%	0.47%	0.21%	2.55%	0.48%	727 (0.58%)
	비교	0.28%	0.04%	0.09%	0.34%	2.41%	0.57%	0.13%	694 (0.55%)

실험 방법: LLM 및 RAG 기반 질의응답 시스템

- RAG
  - 검색기 (Retriever)
    - Wikipedia corpus로 검색용 데이터베이스 구축
    - BM-25 알고리즘으로 검색
  - 생성기 (Generator)
    - `gpt-4o-2024-08-06`
    - `gpt-4o-mini-2024-07-18`

- LLM
  - `gpt-4o-2024-08-06`
  - `gpt-4o-mini-2024-07-18`

실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06			
		LLM		RAG	
		EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06
사실적 질문		7.34	23.39	19.32	33.17
비사실적 질문	토론	0.56	3.0	1.05	4.66
	증거 기반	3.53	**23.34**	5.39	**31.44**
	지침	0.87	19.77	1.53	28.71
	이유	1.12	15.11	2.16	21.7
	경험	2.06	6.13	3.98	10.01
	비교	**4.18**	18.78	**10.37**	27.86

실험 결과: 질문 유형별 LLM과 RAG 시스템의 성능 비교

(EM: Exact Match, F1 : F1-score)

질문 유형		gpt-4o-2024-08-06				gpt-4o-mini-2024-07-18			
		LLM		RAG		LLM		RAG	
		EM	F1	EM	F1	EM	F1	EM	F1
비질문		35.92	52.23	53.21	67.06	8.64	31.33	26.2	46.68
사실적 질문		7.34	23.39	19.32	33.17	2.35	17.47	7.65	23.11
비사실적 질문	토론	0.56	3.0	1.05	4.66	0.16	2.65	0.34	3.7
	증거 기반	3.53	**23.34**	5.39	**31.44**	0.9	**20.57**	1.54	27.43
	지침	0.87	19.77	1.53	28.71	0.55	19.49	0.98	**27.96**
	이유	1.12	15.11	2.16	21.7	0.35	12.86	0.6	18.79
	경험	2.06	6.13	3.98	10.01	0.96	5.12	**1.93**	6.77
	비교	**4.18**	18.78	**10.37**	27.86	**1.15**	15.03	1.73	18.36

References

- [1] Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” In *NeurIPS*, 2020.
- [2] Bolotova et al. “A Non-Factoid Question-Answering Taxonomy,” In *SIGIR*, 2022.
- [3] Reddy et al. “A Survey on Types of Question Answering System,” *IOSR Journal of Computer Engineering*, Vol. 19, pp. 19–23, 2017.
- [4] Mohasseb et al. “Factoid vs. Non-factoid Question Identification: An Ensemble Learning Approach,” *International Conference on Web Information Systems and Technologies*, 2022.
- [5] Jeong et al, “Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity,” In *NAACL*, 2024.
- [6] Chan et al, “RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation,” In *COLM*, 2024.
- [7] “Lurunchik/nf-cats,” *Hugging Face Hub*, https://huggingface.co/Lurunchik/nf-cats, 2022.