질문 유형에 따른 검색 증강 생성의 질의응답 성능 분석

이동건[○,1,*] 박아정[2,*] 이혜리[3] 남현서[4] 맹윤호[5,†]
[1]포항공과대학교 [2]TmaxCoreAI [3]독립연구자 [4]KT [5]이화여자대학교
donggeonlee@postech.ac.kr, {ajeongi59, keira.hyeri.lee, namhs2030}@gmail.com, yunhomaeng@ewha.ac.kr

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

DongGeon Lee[○,1,*] Ahjeong Park[2,*] Hyeri Lee[3] Hyeonseo Nam[4] Yunho Maeng[5,†]
[1]POSTECH [2]TmaxCoreAI [3]Independent Researcher [4]KT [5]Ewha Womans University

요 약

초거대 언어 모델(Large Language Model; LLM)의 발전으로 질의응답 시스템의 성능이 크게 향상되었으나 최신 정보나 전문 지식이 요구되는 질문에 대한 한계가 여전히 존재한다. 이러한 한계를 극복하기 위해 검색 증강 생성(Retrieval-Augmented Generation; RAG) 기술이 제안되었지만 기존 연구들은 주로 사실적인 질문에만 초점을 맞추고 비사실적인 질문에 대한 성능 분석은 부족한 실정이다. 이에 본 논문은 사실적 질문, 비사실적 질문, 그리고 비질문을 포함한 8가지 질문 유형에 대해 RAG와 LLM 기반 질의응답 시스템의 성능을 비교 분석한다. 7종류의 단일 및 다중 추론 질의응답 데이터셋으로 구성한 126,000건의 대규모 평가 데이터셋을 통해 실험한 결과, 두 시스템 모두 사실적 질문과 일부 비사실적 질문에서 높은 성능을 보이지만, 대부분의 비사실적 질문에서는 낮은 성능을 보임을 확인한다. 본 연구는 다양한 질문 형태에 따른 RAG의 성능을 최초로 분석한 연구로 향후 질문 유형에 따른 차별화된 접근의 필요성을 제시한다.

주제어: Evaluation of Question-Answering, Question Type, Retrieval-Augmented Generation, Large Language Model

1 서론

최근 초거대 언어 모델(Large Language Model; LLM) [1]의 발전과 함께, 질의응답 시스템의 자연어 이해와 생성 능력은 크게 향상되었다 [2]. 그러나 LLM은 환각 현상과 학습 데이터의 한계로 최신 정보나 전문적인 지식을 다루는 질문에는 어려움을 겪는다 [3, 4]. 이러한 한계를 극복하기 위해, 검색기로 외부 지식 베이스에서 이용자의 질문과 관련된 정보를 검색하여 LLM의 생성 과정을 보완하는 검색 증강 생성(Retrieval-Augmented Generation; RAG) [5, 6, 7] 기술이 주목받고 있다.

한편, [8]은 사실적인 질문(factoid question)에 비해 고유한 답변 구조를 갖는 비사실적인 질문(non-factoid question)의 중요성을 지적하고, 비사실적인 질문을 지침(instruction), 이유(reason), 증거 기반(evidence-based), 비교(comparison), 경험(experience), 토론(debate) 등 6종류로 분류하는 범주를 제시하였다. 질의응답 시스템을 사용하는 이용자들은 사실적인 질문, 비사실적인 질문 등 다양한 형태의 질문을 사용하지만, RAG 기반 질의응답 시스템을 연구한 기존 방법들은 이용자의 질문 형태에 상관없이 주로 사실적인 질문에 초점을 맞춘 전략을 택해 RAG 파이프라인을 구성하였다 [5, 7, 9]. RAG는 검색기와 LLM 기반의 생성기가 연속적으로 구성된 복잡한 파이프라인이므로, 질문의 의도를 고려하지 않는 기존 RAG 기반 질의응답 시스템은 관련성이 낮은 답변을 내뱉을 가능성이 크고 이용자들의 불만족을 초래할 수 있으나 [10], 아직까지 질문 형태에 따른 RAG의 질의응답 성능을 확인하는 연구는 이루어지지 않았다.

이에 본 논문에서는 [8]에서 제안한 6가지 비사실적인 질문의 유형에 더불어, 사실적인 질문 및 비질문(not a question) 유형을 더한 8가지 질문 유형에 대해 RAG 기반 질의응답 시스템의 성능을 LLM 기반 질의응답 시스템과 함께 비교한다. 실험에 사용한 평가 데이터셋은 4종류의 단일 추론(single-hop) 질의응답 데이터셋 [11, 12, 13, 14]와 3종류의 다중 추론(multi-hop) 질의응답 데이터셋 [15, 16, 17]로 총 126,000건을 구성하였으며, 비교 실험 결과, LLM 및 RAG 모두 사실적인 질문에는 성능이 뛰어나지만, 대부분의 비사실적 질문에는 낮은 성능을 보이는 것을 확인하고, 그 이유를 질문 유형 별로 분석한다.

본 연구의 주요 기여는 다음과 같다.

- 기존 연구 [8]에서 제시된 비사실적 질문의 6가지 유형과 사실적 질문과 비질문 유형을 포함한 총 8가지 질문 유형에 대한 LLM 및 RAG 기반 질의응답 시스템의 성능을 체계적으로 분석한다.
- 7종류의 단일 추론 및 다중 추론 질의응답 데이터셋을 포함하여 일반화한 대규모 평가 데이터셋(126,000건)을 구성하여 질문유형을 분류하고, 유형 별 LLM 및 RAG 기반 질의응답 시스템의 성능을 다각도로 평가한다.

* 공동 저자. (Both authors contributed equally.)
† 교신 저자. (Corresponding author.)

- LLM 및 RAG 기반 시스템이 사실적 질문과 일부 비사실적 질문의 유형에서 높은 성능을 보이는 반면, 대부분의 비사실적 질문에서 낮은 성능을 보임을 확인한다. 이러한 결과는 질의응답에 대한 생성형 언어모델의 한계점을 명확히 하고, 다양한 질문 유형에 적응할 수 있는 질의응답 시스템 설계의 필요성을 강조하며, 향후 개선 방향을 제시한다.

2 관련 연구

검색 증강 생성 기반 질의응답 시스템 LLM (초거대 언어모델)은 환각 현상과 도메인 지식 부족으로 인한 정확도 저하 등의 한계를 가지고 있다 [3, 4]. 이를 보완하기 위해, 검색로 검색한 검색한 외부 지식 데이터를 생성기가 활용하는 연속적 프레임워크인 RAG (검색 증강 생성)가 제안되었으며, 생성형 언어 모델의 정확도와 신뢰성을 높이는 데 기여해왔다 [5, 6, 7].

현재, RAG를 구성하는 각 모듈에 대한 성능 개선 연구가 활발히 진행 중이며 [18], 질문에 내재된 특징에 맞춰 검색 및 생성 전략을 선택하는 방식도 주목받고 있다. [5]는 질문의 복잡도를 평가하여 단일 혹은 다중 단계의 질의응답 전략을 선택하고, [6]은 모호한 질문에 대해 긴 답변을 생성하는 전략을 사용한다. [7]은 질문의 복잡도와 모호성을 모두 고려하여 답변을 생성하는 전략을 제안했지만, 아직까지 RAG는 이용자 질문의 표면적인 형태와 직접적인 의도를 충분히 고려하지 않고 통일된 전략을 적용하는 한계를 보인다.

질문 유형 분류 [10]은 질의응답 시스템에서 질문 유형이 답변에 직접적인 영향을 끼친다는 것과, 질문을 6가지 유형으로 분류했을 때, 각 유형에 따른 답변 방식이 다르다는 것을 밝혔다. [8]은 질의응답 시스템이 복잡한 범주의 데이터셋에 대해 일반화가 어렵다는 문제를 지적하며, 비사실적인 질문(non-factoid question)의 6가지 분류와 각 유형별 예상 답변을 제안하였다. 더불어, 각 유형에서 발견되는 패턴을 표 1과 같이 정리하였다.

- 토론(debate)은 가설적 질문을 물어보는 유형이며, 각 질문에 대해 상반된 의견이 제시되는 답변 구조를 가진다.
- 증거 기반(evidence-based)은 특정 개념이나 사건의 정의를 요구하며, 사실에 기반한 설명이 답변으로 제공된다.
- 지침(instruction)을 물어보는 질문은 절차나 방법을 이해하려는 질문으로, 단계별 지침이 제공되는 답변을 갖는다.
- 이유(reason)는 현상의 원인을 찾는 질문으로, 여러 근거와 설명이 포함된다.
- 경험(experience)은 조언이나 추천을 구하는 질문으로, 개인적 경험에 대한 설명이 답변으로 주어진다.
- 비교(comparison)는 차이점과 유사점을 파악하는 질문으로, 대상 간의 차이와 유사점이 나열된다.

이 같은 기존 연구들은 질문의 유형 별로 적합한 접근 방식을 적용하는 것이 답변의 품질을 개선할 수 있고 [10, 19, 20], 질문의 형태에 맞는 올바른 분류와 그에 따른 차별화된 처리 전략이 RAG와 같은 시스템에도 필수적임을 시사한다.

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 [8]

3 연구 방법

3.1 데이터셋 구성

본 연구에서는 7개의 공개된 질의응답 데이터셋을 활용하여 총 126,000건의 질문-답변 쌍으로 구성된 평가 데이터셋을 구축하였다. 구축한 평가 데이터셋에는 사실적인 질문과 비사실적인 질문이 종합적으로 구성되어 있으며, 질의응답의 복잡성이 유형 별 성능에 미치는 영향을 함께 확인하고자, 단일 추론을 요구하는 질의응답 데이터셋 뿐만 아니라 다중 추론을 요구하는 질의응답 데이터셋을 함께 포함하였다.

3.1.1 단일 추론(single-hop) 질의응답 데이터셋

단일 추론 질의응답 데이터셋으로는 MS MARCO [11], Natural Questions (NQ) [12], TriviaQA [13], SQuAD [14]를 사용하였다. 이들 데이터셋은 주로 단일 문서나 문단에서 답변을 찾을 수 있는 질문들로 구성되어 있다.

3.1.2 다중 추론(multi-hop) 질의응답 데이터셋

다중 추론 질의응답 데이터셋으로는 HotpotQA [15], 2WikiMultiHopQA (2WMHQA) [16], MuSiQue [17]을 활용하였다.

표 2. 질의응답 데이터셋의 질문 유형 별 분포

질문 유형		단일 추론 질의응답 데이터셋 [11, 12, 13, 14]				다중 추론 질의응답 데이터셋 [15, 16, 17]			전체 데이터셋 (126,000건)
		MS 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%)

표 3. 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

이들 데이터셋은 여러 문서나 문단에서 정보를 종합하여 답변을 도출해야 하는 복잡한 질문들을 포함하고 있다.

3.2 질문 유형 분류

모든 질문-답변 쌍은 비질문(not a question), 사실적인 질문, 토론, 증거 기반, 지침, 이유, 경험, 비교 등 8가지 질문 유형으로 분류되었다. 질문 유형 다중 분류에는 deepset/roberta-base-squad2[1] 모델을 미세조정함으로써 0.901의 F1-score를 달성한 Lurunchik/nf-cats [8] 모델을 활용하였다.

3.3 데이터셋 분포 분석

구축한 데이터셋의 질문 유형별 분포를 분석한 결과는 표 2에 나타내었다. 사실적 질문이 전체 데이터셋의 82.90%로 가장 높은 비중을 차지하였다. 그 다음으로는 증거 기반 유형(9.21%), 토론 유형(3.94%) 순으로 나타났다. 반면, 비교 유형의 질문과(0.55%), 경험을 물어보는 질문(0.58%), 지침을 묻는 질문(0.73%)은 상대적으로 낮은 비중을 보였다.

3.4 실험 설계

본 연구에서는 질문 유형이 분류된 데이터셋을 활용하여 LLM 기반 질의응답 시스템과 RAG 기반 질의응답 시스템의 성능을 질문 유형 별로 비교 평가한다. 두 시스템 모두 gpt-4o [1]을 기반으로 하며, 대용량 규모의 고성능 모델인 `gpt-4o-2024-08-06` 모델과 자원 효율적이지만 성능과 정확성이 상대적으로 낮은 `gpt-4o-mini-2024-07-18` 모델을 사용한다. RAG 시스템의 경우에는 BM25 알고리즘 [21]을 기반으로 하는 검색기를 사용한다. 이렇게 설계한 두 종류의 질의응답 시스템을 바탕으로 각 질문 유형별 성능을 평가하여, 질문 유형에 따른 성능 차이를 분석한다.

4 실험 결과

LLM을 단독으로 사용한 질의응답 시스템과 LLM에 검색기를 더한 RAG 기반 질의응답 시스템의 질문 유형 별 성능 비교는 표 3에 나타내었다. 실험 결과, 전반적으로 RAG 기반의 질의응답 시스템이 LLM을 단독으로 사용하는 시스템보다 우수한 성능을 보였다. 특히, 최신 모델인 `gpt-4o-2024-08-06`을 사용했을 때 성능 차이가 더욱 두드러졌다.

4.1 질문 유형별 성능 분석

4.1.1 비질문(not a question)

비질문 유형에서의 질의응답 성능은 다른 질문 유형에서의 성능보다 월등히 높은 결과를 보였는데, 이는 비질문 유형으로 판별된 데이터가 주로 클로즈 테스트(cloze test) [22] 형식인 것과 관련있는 것으로 분석된다. 클로즈 테스트 형식은 문장 내 특정 단어를 빈칸으로 만들고, 그 빈칸을 채우도록 요구함

[1] `https://huggingface.co/deepset/roberta-base-squad2`

으로써 지식을 묻는 방법이다. 이러한 형식은 디코더 기반의 LLM이 인과적 언어 모델링 [23] 방식의 사전 학습 과정에서 접하는 훈련 데이터의 형식과 유사하고, 다른 유형의 질문들보다 질문과 답변이 명확하기 때문에 상대적으로 높은 성능을 보인 것으로 판단된다. 한편, RAG 시스템은 LLM 단독 시스템보다 훨씬 우수한 성능을 보였으며, 이는 생성기 모델의 성능에 관계없이 RAG 시스템이 비질문 유형에서 일관된 높은 성능 향상을 가져온다는 것을 시사한다.

4.1.2 사실적 질문(factoid question)

답변 형태가 단답형이고 정답의 개수가 1개인 사실적 질문 [24]는 비사실적 질문에 비해 명확한 답변을 얻을 수 있으므로 다른 질문 유형에 비해 상대적으로 높은 성능을 보였다. 그러나 사실적 질문에 대한 성능은 비질문 유형에 비해 상대적으로 낮았다. 이는 사실적 질문이 더 구체적이고 정확한 정보를 요구하기 때문에 모델이 정확한 답변을 생성하는 데 더 큰 어려움을 겪었을 것으로 해석된다. 또한, 사실적 질문에 대한 답변은 종종 최신 정보나 특정 도메인의 전문 지식을 필요로 하는데, 이는 LLM의 학습 데이터에 포함되지 않았을 가능성이 있다. 이러한 경우 RAG 시스템이 외부 지식 베이스를 활용하여 부족한 정보를 보완할 수 있지만, 여전히 완벽한 답변을 생성하는 데는 한계가 있을 수 있다.

4.1.3 비사실적 질문(non-factoid question)

비사실적 질문 유형은 답변이 명확하지 않거나 주관적인 해석이 필요한 경우가 많아, 사실적 질문에 비해 성능이 전반적으로 낮게 나타났다. 표 3에서 볼 수 있듯이, LLM과 RAG 시스템 모두 비사실적 질문 유형에서 상대적으로 저조한 성능을 보였다. 특히, 토론을 요청하는 질문이나, 이유, 경험을 물어보는 질문 유형에서 성능이 매우 낮았다.

토론(debate) 토론 유형의 경우 LLM의 성능이 매우 낮게 나타났으며, RAG 시스템을 사용하더라도 성능 향상은 미미했다. 이러한 질문 유형에서는 RAG 시스템이 제공하는 외부 정보도 명확한 정답을 제시하기보다는 다양한 관점을 강화하는 데 그칠 수 있어, 성능 향상에 제한이 있는 것으로 판단된다. 또한, 토론 질문의 특성상 답변이 명확하지 않으며, 여러 해석이 가능하기 때문에 정답에 일치하는 답변을 생성하는 것이 어렵다는 한계를 가진다.

증거 기반(evidence-based) 증거 기반 질문 유형은 다른 비사실적 질문 유형들에 비해 비교적 높은 성능을 보였으며, 특히 F1 점수에서 두드러진 성과를 나타냈다. 이는 증거 기반 질문이 명확한 사실이나 정의를 요구하는 경우가 많아, LLM이 학습한 지식 또는 RAG 시스템이 검색한 외부 정보로부터 상대적으로 명확한 답변을 도출할 수 있었기 때문으로 해석된다.

지침(instruction) 지침 유형의 실험 결과, LLM 단독 시스템과 RAG 시스템 모두에서 성능이 비교적 낮게 나타났다. 이는 지침을 묻는 질문이 답변의 구조적 명확성을 요구하기 때문에, 모델이 학습한 지식만으로는 이러한 요구를 충족시키기 어려웠을 가능성이 있다. 특히, LLM 단독 시스템의 경우 EM 점수가 매우 낮게 나타났으며, 이는 모델이 명확한 절차나 방법을 단계별로 설명하는 데 어려움을 겪었음을 시사한다.

이유(reason) 이유를 묻는 질문 유형에서는 전반적으로 LLM과 RAG 모두 낮은 성능을 보였다. 해당 유형은 단순한 정보 제공을 넘어서 논리적 사고 능력이 요구되므로, 단일한 사실적 답변을 찾기보다는 모델이 다양한 이유와 그에 대한 증거를 종합적으로 고려해 답변을 생성해야 하기 때문에 어려움이 발생한 것으로 보인다. RAG 시스템을 사용했을 때도 성능 향상은 일부 있었지만, 여전히 완벽한 이유를 제공하는 데는 한계가 있었다.

경험(experience) 경험을 묻는 질문은 다른 질문 유형에 비해 성능이 낮게 나타났는데, 이는 경험 질문이 개인의 주관적인 경험에 기반한 답변을 요구하기 때문에, 모델이 이러한 질문에 대한 적절한 답변을 생성하는 데 어려움을 겪은 것으로 추정된다. 또한, RAG 시스템이 LLM 단독 시스템보다 성능이 다소 향상되었으나, 여전히 전체적인 성능이 낮았던 것은 외부 지식 베이스에서 제공하는 정보가 주관적인 경험을 충분히 대체할 수 없기 때문일 수 있다. 이는 경험 기반 질문에 대해서는 현재의 LLM 및 RAG 시스템이 가지는 한계를 보여주는 결과로 해석할 수 있다.

비교(comparison) 비교 질문은 두 가지 이상의 대상을 비교하는 질문으로, 비사실적 질문 유형 중 가장 높은 답변 일치율(Exact Match; EM)을 보였다. 이는 비교 질문이 명확한 기준을 가지고 두 대상을 비교하는 구조를 가지기 때문에, LLM이 학습한 지식이나 RAG 시스템이 검색한 외부 정보로부터 비교적 명확한 답변을 도출할 수 있었기 때문으로 해석된다. 특히, `gpt-4o-2024-08-06` 모델을 사용한 경우 RAG 시스템의 답변 일치율은 10.37로, 다른 비사실적 질문 유형에 비해 월등히 높은 성능을 보였다. 이는 RAG 시스템이 비교 질문에 대해 특히 강점을 보인다는 것을 시사한다.

4.2 단일 및 다중 추론 질문 성능 비교

그림 1은 RAG 기반 질의응답 시스템에서 단일 추론과 다중 추론 질문에 대한 F1-score 성능 비교 결과를 보여준다. 모든 유형의 질문에서 다중 추론 질문이 단일 추론 질문보다 낮은 성능을 기록했다. 이는 다중 추론 질문이 더 많은 정보 검색과 복잡한 추론 과정을 요구하기 때문이다.

주목할 점은 단일 추론 질문에서는 질문 유형 간 성능 편차가

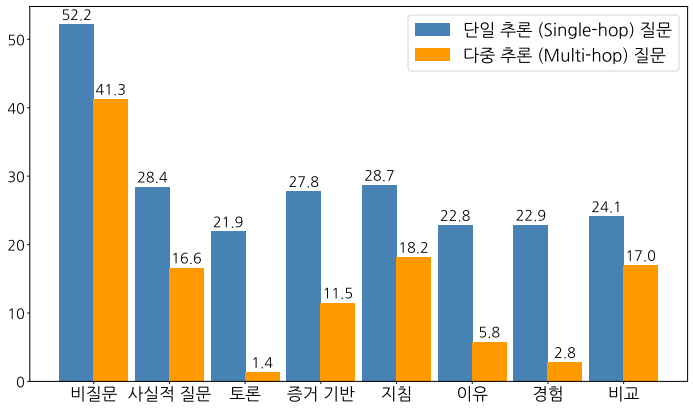

그림 1. 단일 추론 및 다중 추론 질문에 대한 RAG 시스템의 F1-score (%) 성능 비교 (`gpt-4o-mini-2024-07-18` 모델)

크지 않은 반면, 다중 추론 질문에서는 성능 편차가 매우 크다는 것이다. 이는 다중 추론 과정에서 검색기를 통해 더 많은 정보를 제공받아, 결과적으로 잘못된 정보가 포함될 가능성이 높아지기 때문으로 보인다. 특히, 복잡한 추론 과정이 필요한 토론, 이유, 경험과 같은 세 가지 질문 유형에서 급격한 성능 저하가 관찰된다. 이는 앞으로 복잡한 추론 과정과 방대한 정보 검색이 필요한 RAG 시스템을 연구 및 개발할 때, 각 질문 유형에 맞는 차별화된 접근이 필요함을 시사한다.

5 결론 및 논의

본 연구에서는 RAG 기반 질의응답 시스템과 LLM 단독 질의응답 시스템을 다양한 질문 유형에 대해 비교 분석하였다. 특히, [8]에서 제안된 비사실적 질문 유형과 더불어 사실적 질문 및 비질문을 포함한 8가지 질문 유형에 대해 두 시스템의 성능을 체계적으로 평가하였다. 실험 결과, RAG 시스템은 전반적으로 LLM 단독 시스템보다 우수한 성능을 보였으며, 특히 증거 기반, 비교, 지침 질문과 같은 유형에서 성능 향상이 두드러졌다.

비질문 유형에서는 두 시스템 모두 다른 질문 유형에 비해 월등히 높은 성능을 보였다. 이는 비질문 유형이 주로 LLM이 학습 과정에서 접한 클로즈 테스트(cloze test) 형식으로 구성되어 있기 때문으로 추정되며, 특정 유형의 입력에 대해서는 LLM의 한계를 극복할 수 있음을 보여준다. 특히 비질문 유형에서의 이러한 성능 향상은, 향후 질의응답 시스템 설계 시 입력 유형에 따라 차별화된 접근이 필요하며, 비질문과 같은 유형의 입력에 대해서는 RAG 시스템이 매우 효과적인 해결책이 될 수 있음을 시사한다.

반면, 토론, 이유, 경험과 같은 비사실적 질문 유형에서는 두 시스템 모두 상대적으로 낮은 성능을 보였으며, 이는 이러한 질문들이 명확한 정답이 없거나 복잡한 논리적 연결을 요구하기 때문으로 해석된다. 또한, 단일 추론 질문이 다중 추론 질문보다 전반적으로 더 높은 성능을 기록하였으며, 이는 다중 추론 질문이 여러 단계를 거쳐야 하는 복잡한 정보 처리 과정을 요구하기 때문에 성능 저하가 발생했음을 보여준다.

본 연구의 주요 기여는 비사실적 질문 유형에 대한 RAG 시스템의 성능을 처음으로 체계적으로 분석하였다는 점이다. 이를 통해 다양한 질문 유형에 적응할 수 있는 질의응답 시스템 설계의 필요성을 강조하며, 향후 연구 방향에 대한 중요한 시사점을 제공한다. 특히, 비사실적 질문에 대한 성능 향상을 위해서는 보다 정교한 검색 및 생성 전략이 필요하며, 다중 추론 질문에 대한 성능 개선을 위한 추가적인 연구가 요구된다.

감사의 글

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

참고문헌

[1] OpenAI, “GPT-4o System Card,” OpenAI, Tech. Rep., August 2024.

[2] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, “Large language models: A survey,” *arXiv preprint arXiv:2402.06196*, 2024.

[3] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi, “Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models,” *arXiv preprint arXiv:2309.01219*, 2023.

[4] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large Language Models Struggle to Learn Long-Tail Knowledge,” *International Conference on Machine Learning*, Vol. 202, pp. 15 696–15 707, 2023.

[5] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. Park, “Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity,” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024.

[6] G. Kim, S. Kim, B. Jeon, J. Park, and J. Kang, “Tree of Clarifications: Answering Ambiguous Questions with Retrieval-Augmented Large Language Models,” *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 996–1009, 2023.

[7] C. Chan, C. Xu, R. Yuan, H. Luo, W. Xue, Y. Guo, and J. Fu, "RQ-RAG: learning to refine queries for retrieval augmented generation," *arXiv preprint arXiv:2404.00610*, 2024.

[8] V. Bolotova, V. Blinov, F. Scholer, W. B. Croft, and M. Sanderson, "A Non-Factoid Question-Answering Taxonomy," *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1196–1207, 2022.

[9] S.-M. Lee, J. Lee, D. Seo, D. Jeon, I. Kang, and S.-H. Na, "In-Context Retrieval-Augmented Korean Language Model," *Annual Conference on Human and Language Technology*, pp. 443–447, 2023.

[10] M. Breja and S. K. Jain, "A survey on why-type question answering systems," *arXiv preprint arXiv:1911.04879*, 2019.

[11] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems*, Vol. 1773, 2016.

[12] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: a Benchmark for Question Answering Research," *Trans. Assoc. Comput. Linguistics*, Vol. 7, pp. 452–466, 2019.

[13] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pp. 1601–1611, 2017.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[15] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question answering," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

[16] X. Ho, A. D. Nguyen, S. Sugawara, and A. Aizawa, "Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

[17] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "MuSiQue: Multihop Questions via Single-hop Question Composition," *Trans. Assoc. Comput. Linguistics*, Vol. 10, pp. 539–554, 2022.

[18] W. Huang, M. Lapata, P. Vougiouklis, N. Papasarantopoulos, and J. Pan, "Retrieval augmented generation with rich answer encoding," *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1025, 2023.

[19] Y. Zhao, J. Zhang, X. Xia, and T. Le, "Evaluation of google question-answering quality," *Libr. Hi Tech*, Vol. 37, No. 2, pp. 312–328, 2019.

[20] B. B. Cambazoglu, V. Baranova, F. Scholer, M. Sanderson, L. Tavakoli, and B. Croft, "Quantifying human-perceived answer utility in non-factoid question answering," *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, p. 75–84, 2021.

[21] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends Inf. Retr.*, Vol. 3, No. 4, p. 333–389, 2009.

[22] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language Models as Knowledge Bases?" *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2463–2473, 2019.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., February 2019.

[24] A. C. O. Reddy and D. K. Madhavi, "A Survey on Types of Question Answering System," *IOSR Journal of Computer Engineering*, Vol. 19, pp. 19–23, 2017.