RAG-QA Arena: Evaluating Domain Robustness for Long-Form Retrieval-Augmented Question Answering

Announcement Date	@2025년 3월 26일
Conference Name	EMNLP 2024
Keywords	Benchmark LLM RAG

1. Introduction

전통적인 Reading Comprehension QA 모델은 고정된 문맥 내에서 정답을 추출하는 **extractive QA 방식**에 의존해 왔습니다. 하지만 실제 환경에서는 질문에 해당하는 문맥이 주어지지 않는 경우가 많으며, 시스템은 대규모의 오픈 코퍼스(예: Wikipedia)에서 스스로 정보를 찾아야 합니다.

이러한 상황에서 **Retrieval-Augmented Generation (RAG)** 기반 QA는 정보 검색과 생성 능력을 결합하여 유망한 접근 방식으로 부상하고 있습니다. RAG-QA는 정보 검색기로부터 관련 문서를 검색하고, 그 문서를 기반으로 LLM이 답변을 생성하는 방식입니다.

문제는 기존 벤치마크 데이터셋들—**Natural Questions (NQ)**, **TriviaQA**, **ROBUSTQA** 등—이 대부분 **단일 출처(Wikipedia)** 또는 짧은 추출식 답변만을 포함한다는 것입니다. 이러한 형식은 LLM 기반의 장문 응답 생성 능력을 정확히 평가할 수 없고, **도메인 일반화 성능(out-of-domain, OOD)**을 제대로 측정하기 어렵습니다.

이 논문은 이러한 문제를 해결하기 위해 두 가지 기여를 하였습니다.

1. **LFRQA** (Long-form RobustQA) 데이터셋:
    - 26K개의 질문에 대해 여러 문서에서 정보를 추출하고 통합한 **coherent long-form 답변** 수작업 생성.
    - 총 7개 도메인 포함: 금융, 과학, 기술, 라이프스타일, 레크리에이션, 글쓰기, 바이오메디컬

- 답변은 단일 문서가 아닌 **다수 문서의 정보를 통합**하고, 정보 충돌을 해결해 **자연스럽고 일관된 서술형 응답**으로 구성됨.

2. **RAG-QA Arena**:
   - 인간 및 LLM 을 사용해 LLM의 응답을 **LFRQA의 정답과 직접 비교**.
   - 평가 기준: **유용성, 진실성, 완전성**.
   - 사람 평가자와 LLM 평가자의 판단 결과가 높은 상관관계를 보여 신뢰성 확보.

Question	why should a company go public?
Document 1	The purpose is to go public but also to generate more wealth. The real money comes when market values you at a price more than your cash flow. If a company brings in $1000 of cash flow, then that is what the employees and owners have to distribute among themselves. But if they are likely to increase to $2000 next and $4000 next year and they go public then the stock will do well. In this case, the promoters and employees with options/RSUs will benefit as well…….
Document 2	You go public to raise money, to invest in the business and/or pay off the existing shareholders. It's really as simple as that. The advantage of being public is that your shares can easily be bought and sold, and so you can issue and sell new ones and your existing shareholders can sell out if they want to…….
Document 3	The reason to go public is to get money. Not to be snarky, but your question is like asking, "Why should a company try to sell its products……?" The answer, of course, is because they want the money…….
Long-form Answer	A company goes public to raise money because the shares can be easily bought and sold, it can issue and sell new ones. [2, 3] Also, it is a means to generate more wealth among employees who own company's options/RSUs. [1]

Figure 1: LFRQA annotation example. There are three documents (some text removed for brevity) relevant to the query. We instruct annotators to combine ROBUSTQA's answers into a coherent long-form answer with added text if necessary. Citations [1], [2] and [3] indicate the supporting documents of each sentence.

"회사가 왜 상장해야 하나요?" 라는 질문에 대해 Doc 1/2/3 에서 ROBUSTQA에서 추출한 short extractive answers 를 노란색으로 표기하고, annotator들은 하이라이트 된 내용을 통합해서 자연스럽고 일관된 <Long-form Answer>을 만듬.

Dataset Name	Answers grounded in corpus	Long-form answers	Multiple documents	Coherent answers	Multiple domains	Human annotated	# Test queries
LFRQA	✓	✓	✓	✓	✓	✓	16.1K
ROBUSTQA (Han et al., 2023)	✓	✗	✓	✗	✓	✓	16.1K
NQ (Kwiatkowski et al., 2019)	✓	✗	✗	✗	✗	✓	3.6K
MULTIHOP-RAG (Tang and Yang, 2024)	✓	✗	✓	✓	✓	✓	2.5K
ASQA (Stelmakh et al., 2022)	✓	✓	✓	✓	✗	✓	1.0K
LONGFACT (Wei et al., 2024)	✗	✓	✗	✓	✓	✗	2.3K
ELI5 (Fan et al., 2019)	✗	✓	✓	✓	✗	✓	25.0K

Table 1: Comparison of datasets. LFRQA distinguishes from previous work by uniquely encompassing seven features: 1) RAG-QA dataset with answers annotated based on underlying corpus; 2) Long-form answers of paragraph length; 3) Multiple documents that provide different facts/views; 4) Coherent answers that handle conflicting information; 5) Multiple-domain corpus to benchmark domain robustness; 6) Human annotated high-quality answers; 7) Large-scale evaluation set.

1. 정답이 실제 문서(코퍼스)에 기반했는가? 2. 단락 길이의 장문 answer 인가? 3. multiple document 에서 정보를 추출했는가? 4. 정답이 논리적이고 일관적인가? 5. 다양한 주제/도메인을 포함하는가? 6. 사람이 직접 응답을 작성했는가?

---

2. RAG-QA Task Formulation

1) Retrieval 단계 – 정보 검색

- RAG-QA 시스템은 먼저 질문에 대해 관련이 있을 법한 문서들을 **검색**합니다.
- 이 문서들은 너무 길기 때문에, 작은 **문단 단위(passage)**로 잘게 나눕니다. (예: 100단어씩)
- 그런 다음, 시스템은 그 중에서 질문과 **가장 관련성이 높은 문단 몇 개(K개)**를 선택합니다.

2) Generation 단계 – 정답 생성

- 검색된 문단들을 바탕으로 **LLM이 정답을 생성**합니다.
- 여기서 중요한 건, 정답이 단순한 한 줄짜리 문장이 아니라, **자연스럽고 설명이 포함된 long-form 서술형**이라는 점입니다.
- 예전에는 모델이 정답 문장과 **얼마나 비슷한 단어를 썼는지(단어 중복률)**만 보고 평가했지만,
    - 이런 방식은 LLM이 문장을 다르게 말하거나, 더 많이 설명하면 **오히려 낮은 점수**를 받게 되어 **공정하지 않습니다.**

왜 이게 중요한가?

- 실제 세상에서 QA 시스템은 항상 새로운 질문을 받습니다.

  → 예: 금융, 의료, 과학, 기술 등 여러 도메인.

- 기존 시스템은 **한정된 도메인(예: Wikipedia)**에만 훈련되어 있어서, **다른 도메인에서는 성능이 떨어집니다.**
- 그래서 LFRQA는 다양한 도메인의 문서와 정답을 포함시켜서,

  → **LLM이 새로운 상황에서도 잘 작동하는지(도메인 일반화 성능)** 평가할 수 있도록 설계된 것입니다.

---

3. Data Creation (LFRQA)

3.1 Annotated Data

Queries	RobustQA answers	LFRQA answers
Will the word schadenfreude be understood in an English text?	Yes it would not be understood by the majority of English-speaking adults No mostly anecdotal evidence here suggests not	The term "schadenfreude" has seen an increased use in the English language and is understood to mean deriving pleasure from others' misfortunes. However, it would not be widely recognized by the majority of English-speaking adults, as its current usage is still negligible. Furthermore, there is a consensus that the term isn't generally understood, and anecdotes reflect that only a minor portion of certain professional groups, such as IT personnel, might be familiar with it.
How much can I trust Tor?	can not say Tor can solve all your problems many ways to compromise your identity give you considerably more protection than browsing directly Tor is probably what I would recommend don't let it be the only type of thread in your safety net use Tor everywhere except on an Internet connection	You cannot say Tor can solve all your problems, as there can be many ways to compromise your identity, so while Tor will be useful as one thread in a safety net, don't let it be the only type of thread in your safety net. Tor is probably to be recommended as it gives considerably more protection than browsing directly, but be aware that your ISP is in a position to see that your IP address is using Tor, even though it can't tell what you're using Tor for. You should use Tor everywhere, except on an Internet connection which can be strongly associated with you.

Figure 2: LFRQA v.s. ROBUSTQA. Citations are removed in LFRQA's answers, and a few answer spans are removed for clarity. Green and orange texts represent positive and negative opinions, respectively.

"영어 문맥에서 schadenfreude 라는 단어가 이해될까?" 라는 질문에 RobustQA는 짧고 단편적인 정답을 찾아서 나열했지만, LFRQA에서는 사람이 여러 정답을 하나의 자연스러운 문단형 응답으로 통합함. (긍,부정 의견도 하나의 서술로 통합)

기존 방식 기반:

- LoTTE, FiQA와 같은 IR 기반 데이터셋에서 질문과 관련 문서를 수집.
- ROBUSTQA는 이 문서에서 3개의 짧은 정답 span(최대 16단어)을 추출하는 방식.

LFRQA 방식:

- 문서 내의 여러 단편적 정답을 **통합하고 확장하여 자연스러운 서술형 정답**을 만듦.
- 정답 문장은 어떤 문서의 정보를 기반으로 했는지 [1], [2] 같은 citation으로 표시(평가 시 제거).

품질 관리:

- 전문 데이터 작업자들(annotators) 이 사람이 직접 문서를 읽고, 여러 단편적인 정답을 **하나의 일관된 long-form 문장**으로 통합해서 작성

- 별도의 언어학 전문가 팀(data linguists)이 작성된 데이터 중 **무작위로 10% 정도를 샘플링해서 검토**
- 10% 샘플 중 90% 이상이 기준에 부합하지 못하면 해당 작업 전체 반려, 작성자에게 다시 수정 요구
- 주요 실패 유형: 불완전성, 중복, 문법 오류, 출처 오표기 등.

3.2 Adapted Data

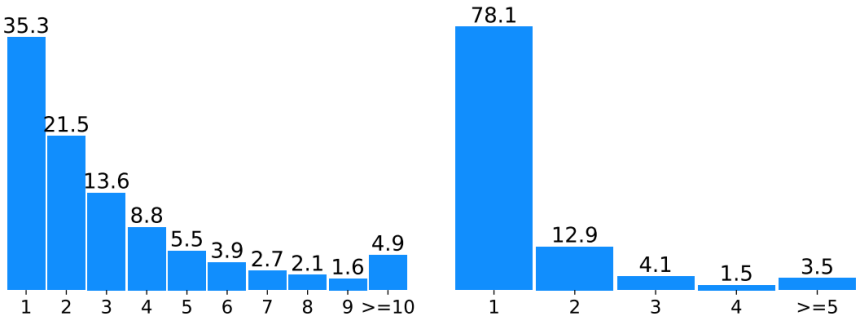
- LFRQA는 바이오메디컬 도메인을 포함하기 위해 기존 BioASQ 데이터셋에서 제공하는 두 가지 정답 중 **서술형(long-form)인 ideal answer만 사용**
- 이로써 별도의 재작업 없이도 **LFRQA의 long-form 기준에 맞는 고품질 정답**을 활용할 수 있었음

3.3 데이터 분석

			Test Set			RobustQA		LFRQA	
Domain	**Source**	**Label**	$\|Q\|$	$\|D\|$	$\|P\|$	A/Q	W/A	A/Q	W/A
Biomedical	BioASQ	[BI]	1,956	15,559,026	37,406,880	2.6	2.4	1.0	30.0
Finance	FiQA	[FI]	3,612	57,638	105,777	3.0	9.4	1.0	69.1
Lifestyle	LoTTE	[LI]	2,208	119,461	241,780	5.7	8.7	1.0	99.5
Recreation	LoTTE	[RE]	2,094	166,975	315,203	3.2	7.2	1.0	60.3
Technology	LoTTE	[TE]	2,111	638,509	1,252,402	6.0	8.7	1.0	99.7
Science	LoTTE	[SC]	1,423	1,694,164	3,063,916	5.3	7.8	1.0	92.0
Writing	LoTTE	[WR]	2,695	199,994	347,322	6.2	6.6	1.0	88.0

Table 2: Data (test set) summary: LFRQA v.s. RobustQA. $|Q|$, $|D|$, $|P|$, A/Q, and W/A represent numbers of questions, documents, passages, answers per question, and words per answer, respectively. Each passage consist of 100 words at most. LFRQA has only one answer per query as we integrate multiple answers from RobustQA, which results in more words in (long-form) answers. Dev set statistics can be found in Appendix Table 7.

LFRQA와 ROBUSTQA의 test set 구성 차이를 비교한 표. |Q| : Question / |D|: Documents / |P| : Passage / A/Q: answers per question / W/A : words per answer. 해당 표는 LFRQA가 기존 ROBUSTQA와 동일한 질문과 문서를 사용하지만, **답변을 길고 일관성 있게 통합**했다는 점을 **수치로 보여주는 표임**



(a) # of documents used per answer.

(b) # of documents used per answer sentence.

Figure 3: Distribution of number (#) of documents used in LFRQA's answers. All numbers are %.

LFRQA에서 정답을 만들 때 몇 개의 문서를 참고했는가에 대한 그래프.

도메인	질문 수	문서 수	passage 수	정답 길이(단어)
Finance	3.6K	57K	105K	**69.1**
Tech	2.1K	638K	1.25M	**99.7**
Writing	2.6K	199K	347K	**88.0**

- 약 **65% 이상의 정답이 2개 이상 문서 사용**.
- **4.9%는 10개 이상 문서**의 정보를 통합.
- 문장 단위로도 **22% 이상이 복수 문서 통합**.
- Coherence, fluency, multi-perspective 통합 면에서 ROBUSTQA 대비 큰 향상.

4. RAG-QA Arena (Evaluation Framework)

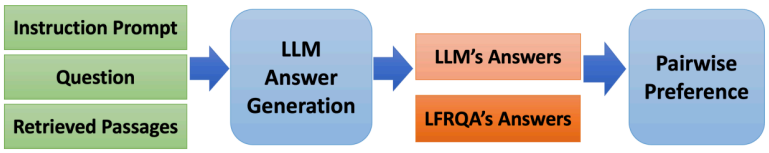


Figure 4: RAG-QA ARENA framework. Green blocks are LLM's inputs to generate answers. Orange blocks are LLM and LFRQA's answers presented to both human and LLM judges to determine pairwise preferences.

평가 framework 의 구조.

핵심 아이디어:

- 기존 평가 방식(EM, F1)은 short-answer extractive 기준이라 LLM long-form 응답 평가에 부적절.
- → **LFRQA의 정답과 직접 비교하는 방식** 채택

평가 지표:

1. **Truthfulness** (진실성): 정보의 정확성과 사실 기반 여부.
2. **Helpfulness** (유용성): 질문에 실제로 도움이 되는 정보.
3. **Completeness** (완전성): 가능한 많은 관련 정보 포함.

평가 방식:

- **Human Evaluation**: LFRQA vs LLM 응답, 3인 다수결 평가.
- **Model-based Evaluation (MBE)**:
  - GPT-4 시리즈를 평가자로 활용 (CoT, in-context learning 프롬프트 구성).
  - 사람 평가 결과와 높은 일치도(Pearson ≥ 0.52, Cohen's Kappa ≥ 0.43).

    → 앞으로 RAG-QA 벤치마크에서는 GPT 평가도 신뢰가 가능하다는 것을 입증

---

5. Experimental Setup

- **Retriever**: ColBERTv2 (100단어 단위 passage, 상위 5개 사용).
- **LLMs tested**:
  - GPT-4-Turbo, GPT-4o, GPT-4-0125
  - MIXTRAL-8×22B / 8×7B
  - LLAMA-3-70B / 8B
  - COMMAND R+ / R
  - QWEN1.5-110B / 32B
- **프롬프트 구성**:
  - CoT, in-context examples, HTML-style delimiter 등 사용.
  - GPT-4o는 CoT 포함 시 "답이 없다"는 응답이 과도하여 CoT 제거함.

---

6. Results and Analysis

	Overall		[BI]		[FI]		[LI]		[RE]		[TE]		[SC]		[WR]	
Compared Models	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T	W	W+T
GPT-4o* #1	**36.9**	**41.0**	**52.9**	**59.3**	38.4	42.3	**25.1**	**27.9**	**40.4**	**46.4**	35.6	38.8	42.8	47.6	**28.4**	**31.1**
GPT-4-TURBO #2	34.4	39.1	36.0	43.9	40.6	45.1	23.2	26.1	36.7	44.1	**36.6**	**39.6**	42.6	47.9	26.2	29.6
GPT-4-0125-PREVIEW #6	28.9	33.7	31.4	40.1	36.8	40.8	18.1	21.3	31.5	38.6	30.4	34.0	34.7	40.5	19.2	22.3
MIXTRAL-8X22B #3	34.5	38.8	37.0	46.0	**44.1**	**47.6**	21.3	24.4	34.4	41.0	33.9	36.8	**45.0**	**49.5**	25.9	28.1
MIXTRAL-8X7B #7	27.5	31.0	31.9	39.1	35.3	38.5	15.9	18.4	24.8	29.5	30.3	32.1	33.9	38.0	20.0	21.9
LLAMA-3-70B #8	21.7	25.2	30.3	37.2	24.6	27.7	12.9	15.1	22.3	27.3	22.4	24.4	25.6	30.0	15.5	18.2
LLAMA-3-8B #10	20.4	23.5	34.7	39.6	24.0	27.0	11.2	13.2	19.4	24.7	20.5	22.5	22.3	26.1	12.5	14.4
COMMAND R+ #9	21.1	25.8	26.0	33.5	25.8	30.3	13.5	16.4	22.6	30.0	22.4	25.4	24.9	31.2	13.6	16.0
COMMAND R #11	11.1	15.2	18.6	26.1	13.0	17.1	5.2	7.4	10.4	17.0	10.3	12.3	14.9	20.2	7.3	9.4
QWEN1.5-110B-CHAT #4	33.4	37.8	36.2	44.0	42.6	46.9	22.3	25.1	34.1	40.7	34.8	37.5	40.8	46.1	22.5	25.2
QWEN1.5-32B-CHAT #5	32.8	37.1	34.9	42.8	43.2	47.3	20.7	23.7	32.3	38.3	34.0	37.1	40.8	44.8	22.6	25.2

Table 3: Evaluation results on LFRQA test set. W and W+T indicate win and win+tie rate against LFRQA's answers. LLM's answers are generated based on the top 5 passages. **bold** and underline indicate the best and runner-up results. * means using the answer generation prompt w/o CoT. #n indicates the Elo ranking in Appendix Table 5.

여러 LLM들이 생성한 답변이 LFRQA 정답보다 얼마나 선호되었는지를 평가한 결과. (W : win / W+T : win+tie)

🔷 RAG-QA Arena Benchmark

모델	LFRQA와의 비교에서 승률 (W)
GPT-4O	**36.9%**
GPT-4-Turbo	34.4%
MIXTRAL-8×22B	34.5%
QWEN1.5-110B	33.4%
LLAMA-3-70B	21.7%
COMMAND R	11.1%

- 아직까지도 LFRQA의 응답이 대부분의 LLM보다 우수하다고 평가됨.
- **정답 참조로서의 LFRQA의 강력함**을 입증.

7. Conclusion

- **LFRQA**: 다양한 도메인을 포괄하는 고품질 long-form QA 데이터셋.
- **RAG-QA Arena**: 확장성 있는 평가 프레임워크로 사람/모델 기반 평가를 통합.
- 향후 RAG-QA 시스템의 도메인 적응성과 응답 품질 개선을 위한 **중요 벤치마크 플랫폼**으로 기대됨.