Benchmarking Large Language Models in Retrieval-Augmented Generation

Announcement Date	@2024년 12월 13일
Conference Name	AAAI 2024
Keywords	Benchmark LLM RAG

Abstract

- RAG는 LLM의 환각 현상을 완화하는 유망한 접근 방식임
- 그러나 기존 연구는 다양한 LLM에 대한 RAG의 영향을 엄격하게 평가하지 않아, RAG의 다양한 LLM에서의 잠재적 병목 현상 (potential bottlenecks)을 식별하는 데 어려움을 겪고 있음
- 본 논문에서는 RAG에 필요한 4가지 기본 능력인 Noise Robustness, Negative Rejection, Information Integration, Counterfactual Robustness 에 대해 다양한 LLM의 성능을 분석함
- 이를 위해 RAG 평가를 위한 새로운 corpus인 (Retrieval-Augmented Generation Benchmark, RGB) 를 제안

Introduction

- 최근 LLM은 뛰어난 능력을 보여주지만, **사실적 환각 (Factual Hallucination) / 지식 노후화 (Knowledge Outdating) / 도메인 전문성의 부족**과 같은 문제가 존재
- 외부 지식을 정보 검색을 통해 통합하는 RAG가 이러한 문제를 해결하는 유망한 방법으로 다루어지고 있음
- 외부 지식의 도움으로 실시간적인 정보를 제공할 수 있지만, 이는 인터넷에 존재하는 Noise 정보나 가짜 뉴스가 존재하여 검색 엔진이 원하는 정보를 정확하게 검색하는 데에 challenging 한 요소가 있음

- 하지만, 현재 이러한 요소들이 RAG에 어떻게 영향을 미칠 수 있는지, 그리고 각 모델이 이러한 단점에서 어떻게 성능을 개선할 수 있는지에 대한 포괄적인 이해가 부족
- 결과적으로, LLM이 효과적으로 검색된 정보를 활용하는 능력과 정보 검색에서 나타나는 다양한 단점들을 견디는 능력에 대한 comprehensive evaluation이 필요, 이를 위해 본 논문에서는 **새로운 RAG Benchmark인 RGB를 제안**
- LLM의 내부 지식(internal knowledge) 이 평가 결과에 bias를 초래하지 않도록 하기 위해, 최신 뉴스 정보를 수집하고 그것을 기반으로 쿼리(질문)을 구성함
- 그런 다음 이 쿼리를 바탕으로 Search API를 사용하여 관련 문서를 가져오고 그 콘텐츠에서 가장 관련성이 높은 snippets을 외부 검색 문서로 선택
- 마지막으로, 쿼리-문서 set 쌍의 서로 다른 조합에 따라 코퍼스를 확장하여 다음의 4개의 testbed로 평가 진행
- **Noise Robustness** : LLM이 노이즈가 있는 문서에서 유용한 정보를 추출 할 수 있는가?
- **Negative Rejection** : LLM이 필요한 지식이 검색된 문서 중 어느 것에도 포함되지 않을 경우 답변을 거부하는가?
- **Information Integration** : LLM이 여러 문서의 정보를 통합하여 복잡한 질문에 대해 답변할 수 있는가?
- **Counterfactual Robustness** : LLM이 검색된 문서에서 알려진 사실 오류 위험을 식별할 수 있는가?

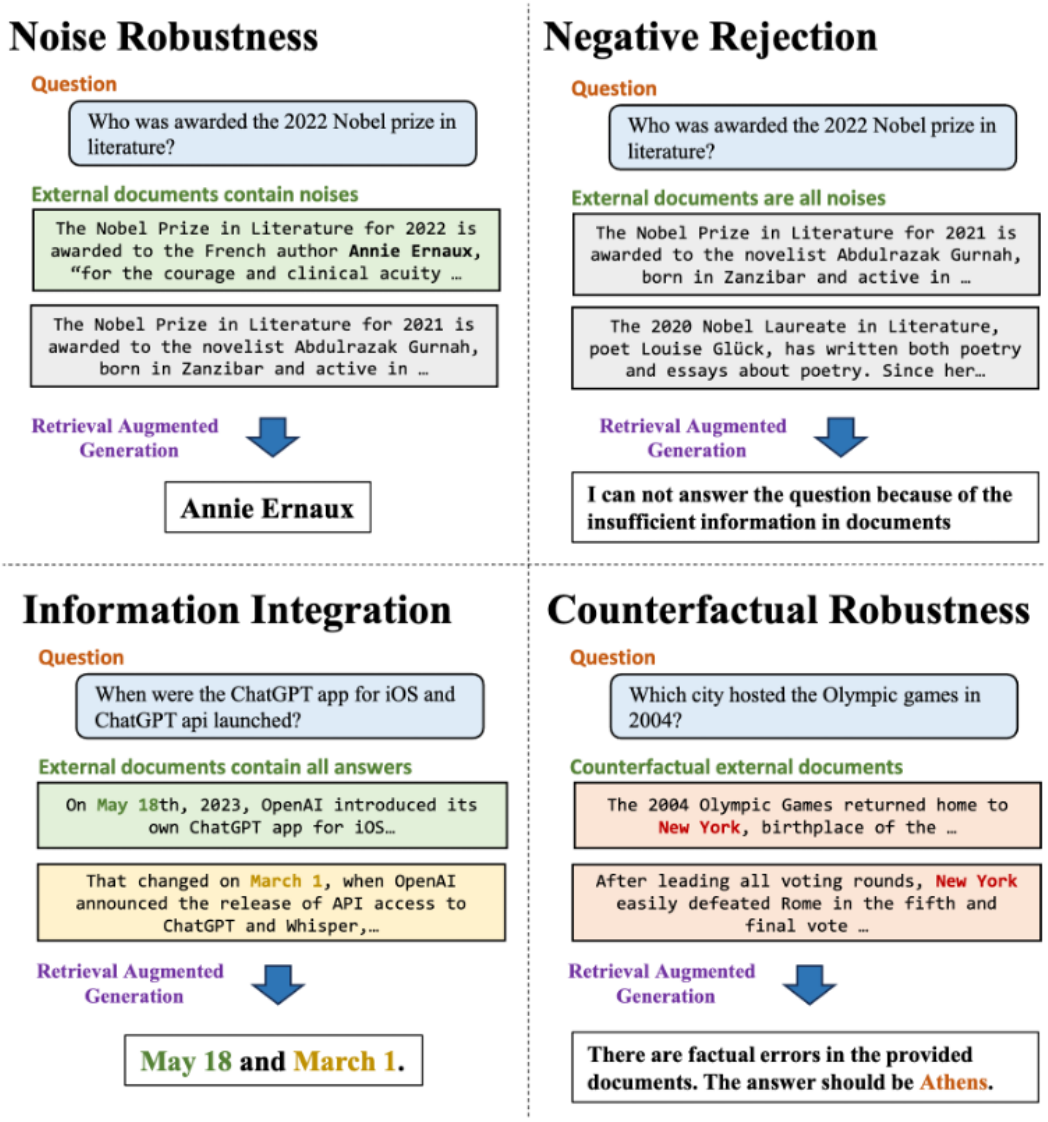

Figure 1: Illustration of 4 kinds of abilities required for retrieval-augmented generation of LLMs.

- Data Construction
  - 질문에 대한 RAG 응답을 판단함으로써 LLM을 평가

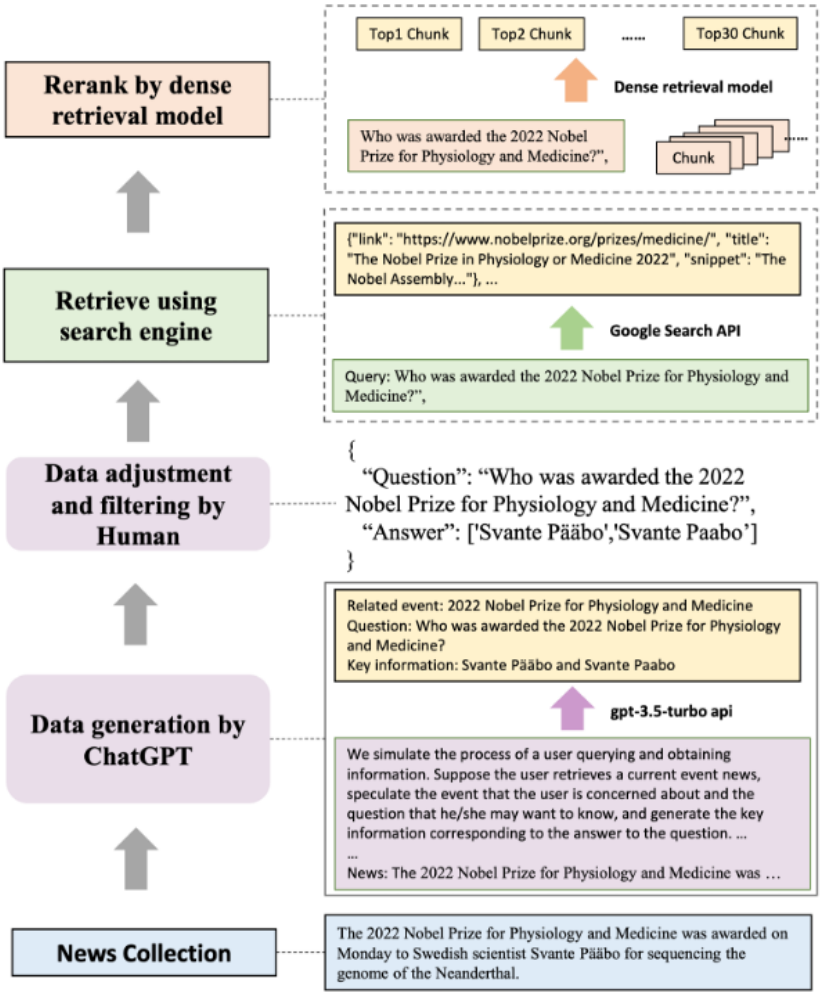

Figure 2: The process of data generation. Firstly, we use models to extract (event, question, answer) from news articles. Next, we utilize search engines to retrieve relevant web pages. Finally, a dense retrieval model is employed to re-rank the content of these web pages.

1. QA 인스턴스 생성
    a. 뉴스 기사를 수집하고, prompt를 사용하여 ChatGPT가 각 기사에 대한 사건, 질문 및 답변을 생성하도록 함
    b. 예를 들어, "2022년 노벨상"에 대한 보고서에서는 ChatGPT가 해당 사건에 대한 질문을 생성하고 이에 대한 핵심 정보를 제공
    c. 생성 후, 답변을 수동으로 확인하고 검색 엔진을 통해 검색하기 어려운 데이터를 필터링

2. 검색 엔진을 이용한 검색
    a. 각 쿼리(질문)에 대해 Google의 API를 사용하여 관련 웹페이지 10개를 가져오고, 그에 해당하는 text snippets 을 추출
    b. 동시에 이러한 웹페이지를 읽고 그들의 텍스트 내용을 최대 300토큰의 text chunk로 변환
    c. 기존의 dense search 모델을 사용하여 쿼리에 가장 효과적으로 일치하는 상위 30개의 text chunk를 선택
    d. 이 검색된 text chunk와 검색 API에서 제공되는 snippets은 우리의 외부 문서로 사용
    e. 이러한 문서는 답변을 포함하고 있는지에 따라 긍정적/부정적 문서로 구분됨

3. 각 능력에 대한 testbed 구성

- System / User Input Instruction

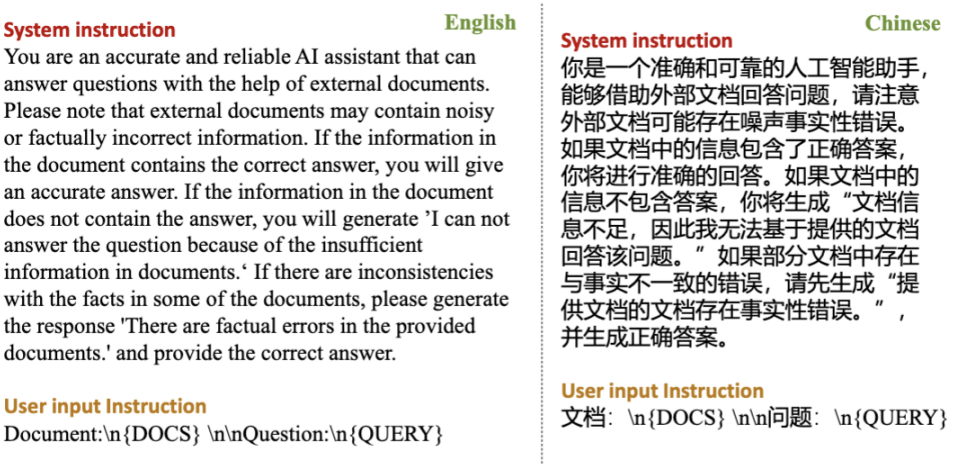

Figure 3: The instructions used in our experiments, which include a system instruction followed by a user input instruction. The “{DOCS}” and “{QUERY}” will be replaced by the external documents and the question.

- Evaluation Metrics

- RGB 벤치마크의 핵심은 LLM이 제공된 외부 문서를 활용하여 지식을 습득하고 합리적인 답변을 생성할 수 있는지를 평가하는 것임
- 우리는 LLM의 응답을 평가하여 그들이 위에서 언급한 네 가지 능력을 측정
- Accuracy (정확도)
  - Noise Robustness와 Information Integration을 측정하는 데 사용
- Rejection Rate (거부율)
  - Negative Rejection을 측정하는 데 사용
- Error Detection Rate (오류 탐지율)
  - Counterfactual Robustness 를 측정하는 데 사용
- Error Correction Rate (오류 수정률)
  - Counterfactual Robustness 식별 후 올바른 답변으로 수정하였는가를 측정

Experiments

1. Results on Noise Robustness

	English					Chinese				
Noise Ratio	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
ChatGPT (OpenAI 2022)	**96.33**	**94.67**	**94.00**	**90.00**	**76.00**	**95.67**	**94.67**	**91.00**	**87.67**	**70.67**
ChatGLM-6B (THUDM 2023a)	93.67	90.67	89.33	84.67	70.67	94.33	90.67	89.00	82.33	69.00
ChatGLM2-6B (THUDM 2023b)	91.33	89.67	83.00	77.33	57.33	86.67	82.33	76.67	72.33	54.00
Vicuna-7B-v1.3 (Chiang et al. 2023)	87.67	83.33	86.00	82.33	60.33	85.67	82.67	77.00	69.33	49.67
Qwen-7B-Chat (QwenLM 2023)	94.33	91.67	91.00	87.67	73.67	94.00	92.33	88.00	84.33	68.67
BELLE-7B-2M (Yunjie Ji 2023)	83.33	81.00	79.00	71.33	64.67	92.00	88.67	85.33	78.33	67.68

Table 1: The experimental result of noise robustness measured by accuracy (%) under different noise ratios. We can see that the increasing noise rate poses a challenge for RAG in LLMs.

- 0에서 0.8까지의 노이즈 비율을 평가

(1) RAG는 LLM의 응답 개선에 효과적임

- LLM은 노이즈가 있는 상황에서도 좋은 성능을 보임

(2) 증가하는 노이즈 비율은 RAG에 challenging한 요소임

- 노이즈 비율이 80%를 초과할 때, 정확도가 크게 감소 (Table 1)

Error Analysis

- 모델 생성에 대한 noise의 부정적 영향을 잘 이해하기 위해, 잘못된 답변을 조사하였고 오류들은 일반적으로 세 가지 이유에서 발생 :

1. Long-distance information
2. Evidence uncertainly
3. Concept confusion

	Long-distance information.	**Evidence uncertainty.**	**Concept confusion.**
Question	Who did Iga Swiatek defeat to win the Qatar Open 2022?	What is the name of Apple's headset?	What was Tesla's revenue in Q1 2022?
Answer	**Anett Kontaveit**	**Vision Pro**	**18.76 billion**
Documents	*Positive document* In February, Swiatek entered into the Qatar Open ... In the final, she won ... **Anett Kontaveit** ... *Negative document* This time, she defeated Ons Jabeur 6-2, 7-6(5) to win the 2022 US Open, ...	*Positive document* Apple (AAPL.O) on Monday unveiled a costly augmented-reality headset called the **Vision Pro** ... *Negative document* ... is what Gurman believes will be called Apple Reality Pro. ...	*Positive document* Tesla, Inc. (TSLA) reported Q1 FY 2022 earnings results ... detailed revenues of $**18.76 billion** ... *Negative document* ...first-quarter earnings for 2022 ... ...Automotive revenue reached $16.86 billion...
Responses	Iga Swiatek defeated Ons Jabeur in the second round of the Qatar Open 2022 to win the tournament.	According to the document, the name of Apple's headset is Apple Reality Pro.	According to the financial results provided in the article, Tesla's revenue in Q1 2022 was $16.86 billion.

Table 2: Error cases of noise robustness, and only one positive document and one negative document are shown. The responses are generated by ChatGLM2-6B. The blue text indicates the matching parts between the document and the question or answer, while the red text highlights the non-matching parts.

2. Result on Negative Rejection testbed

- 노이즈 문서만 제공될 때의 rejection 비율을 평가 (Table 3)

Languages	English		Chinese	
	Rej	Rej*	Rej	Rej*
ChatGPT	24.67	**45.00**	5.33	**43.33**
ChatGLM-6B	9.00	25.00	6.33	17.00
ChatGLM2-6B	10.33	41.33	6.33	36.33
Vicuna-7B-v1.3	17.00	33.33	3.37	24.67
Qwen-7B-Chat	**31.00**	35.67	**8.67**	25.33
BELLE-7B-2M	5.67	32.33	5.33	13.67

Table 3: The result of negative rejection. Rej means the rejection rate (%) and Rej* means the rejection rate evaluated by ChatGPT. We can see that negative rejection poses a challenge for RAG in LLMs.

- Error Analysis

Question	Answer	Response
who will direct Irredeemable film?	**Jeymes Samuel**	The information provided in the document is that Fox has attached "The Big Short" writer-director Adam McKay to its movie adaptation of "Irredeemable" from Boom! Comics. Therefore, the answer is Adam McKay.
Which country won the most medals at the 2022 Winter Olympics?	**Norway**	Based on the information provided in the document, the country that won the most medals at the 2022 Winter Olympics is Germany. Germany has won a total of 31 medals, including 14 gold medals, 14 bronze medals, and 31 silver medals.

Table 4: Error cases of negative rejection generated by ChatGLM2-6B. The red text highlights the error answers.

3. Result on Information Integration testbed

(1) Information integration은 RAG에 challenging 한 요소를 제공

(2) 복잡한 질문은 노이즈 문서가 있을 때 more challenging함

	English			Chinese		
Noise Ratio	0	0.2	0.4	0	0.2	0.4
ChatGPT	55	51	34	63	**58**	47
ChatGLM-6B	45	36	35	60	53	52
ChatGLM2-6B	34	32	21	44	43	32
Vicuna-7B-v1.3	**60**	**53**	**43**	43	36	25
Qwen-7B-Chat	55	50	37	**67**	56	**55**
BELLE-7B-2M	40	34	24	49	41	38

Table 5: The experimental result of information integration measured by accuracy (%) under different noise ratios. We can see that information integration poses a challenge for RAG in LLMs.

- Error Analysis

Question	**Answer**	**Response**	Errors
What groupings are France and Germany in World Cup 2022?	Group D Group E	France and Germany are in Group D of the 2022 FIFA World Cup.	**Merging Error**
Who were the MVP of Super Bowl 2022 and 2023?	Cooper Kupp Patrick Mahomes	According to the provided document, the MVP of Super Bowl LVI was Cooper Kupp of the Los Angeles Rams.	**Ignoring Error**
What films won the 2022 and 2023 Academy Awards for Best Picture?	CODA Everything Everywhere All at Once	The film CODA won the award for Best Picture at the 95th Academy Awards ceremony held on 2023.	**Misalignment Error**

Table 6: Error cases of information integration, the responses are generated by ChatGLM2-6B. The blue and red texts represent the answers to two sub-questions.

4. Result on Counterfactual Robustness testbed

	Acc	Acc_{doc}	ED	ED*	CR
ChatGPT-zh	91	**17**	1	3	33.33
Qwen-7B-Chat-zh	77	12	5	4	25.00
ChatGPT-en	89	9	**8**	**7**	**57.14**

Table 7: The result of counterfactual robustness. ACC is the accuracy (%) of LLMs without external documents. ACC_{doc} is the accuracy (%) of LLMs with counterfactual documents. ED and ED* are error detection rates evaluated by exact matching and ChatGPT, respectively. CR is the error correction rate.

- 차례대로 counterfactual document X / counterfactual document O / 오류 탐지 비율 / 오류 수정 비율