Benchmarking Large Language Models in Retrieval-Augmented Generation

Jiawei Chen[1,3], **Hongyu Lin**[1,*], **Xianpei Han**[1,2,*], **Le Sun**[1,2]

[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
{jiawei2020,hongyu,xianpei,sunle}@iscas.ac.cn

Abstract

검색 증강 생성(Retrieval-Augmented Generation, RAG)은 대형 언어 모델(LLMs)의 환각 현상을 완화하는 데 유망한 접근 방식이다. 그러나 기존 연구는 검색 증강 생성이 다양한 대형 언어 모델에 미치는 영향을 철저히 평가하지 않아, RAG의 잠재적 병목 현상을 식별하는 데 어려움이 있다. 이 논문에서는 검색 증강 생성이 대형 언어 모델에 미치는 영향을 체계적으로 조사한다. 우리는 RAG에 필요한 4가지 기본 능력, 즉 노이즈 강인성, 부정 거부, 정보 통합, 반사실 강인성에 대한 다양한 대형 언어 모델의 성능을 분석한다. 이를 위해, 우리는 영어와 중국어 모두에서 RAG 평가를 위한 새로운 말뭉치인 검색 증강 생성 벤치마크(RGB)를 구축한다. RGB는 사례를 해결하는 데 필요한 앞서 언급한 기본 능력에 따라 벤치마크 내의 인스턴스를 4개의 별도 테스트베드로 나눈다. 그런 다음 RGB에서 6개의 대표적인 LLM을 평가하여 RAG를 적용할 때 현재 LLM의 도전 과제를 진단한다. 평가 결과, LLM이 어느 정도의 노이즈 강인성을 보이는 반면, 부정 거부, 정보 통합, 잘못된 정보 처리 측면에서는 여전히 상당한 어려움을 겪고 있음을 드러낸다. 이러한 평가 결과는 RAG를 LLM에 효과적으로 적용하기 위해서는 아직 상당한 여정이 남아 있음을 나타낸다.

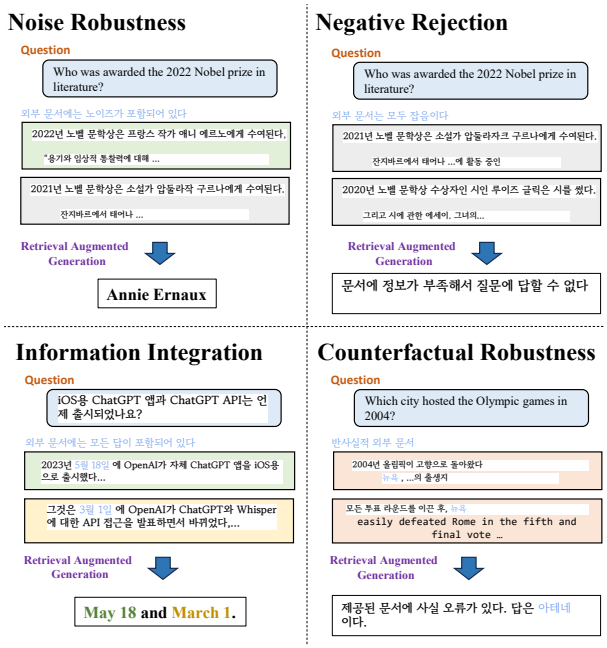

그림 1: LLM의 검색 증강 생성에 필요한 4가지 능력을 설명한 그림.

Introduction

최근 ChatGPT (OpenAI 2022)와 ChatGLM (THUDM 2023a) 같은 대형 언어 모델(LLMs)에서 인상적인 발전이 있었다. 이러한 모델들은 놀라운 일반 능력을 보여주었지만 (Bang et al. 2023; Guo et al. 2023), 여전히 사실 왜곡 (Cao et al. 2020; Raunak, Menezes, and Junczys-Dowmunt 2021; Ji et al. 2023), 지식 노후화 (He, Zhang, and Roth 2022), 그리고 특정 도메인 전문성 부족 (Li et al. 2023c; Shen et al. 2023) 같은 문제에 심각하게 시달리고 있다.

정보 검색을 통한 외부 지식 통합, 즉 검색 증강 생성(RAG)은 위의 문제를 해결하는 유망한 방법으로 여겨지고 있다. (Guu et al. 2020; Lewis et al. 2020; Borgeaud et al. 2022; Izacard et al. 2022). 외부 지식을 활용하면 LLM들이 더 정확하고 신뢰할 수 있는 응답을 생성할 수 있다. 가장 일반적인 방법은 New Bing과 같은 검색 엔진을 검색기로 사용하는 것이다. 인터넷에 방대한 양의 정보가 존재하기 때문에, 검색 엔진을 사용하면 더 실시간 정보를 제공받을 수 있다.

하지만 Retrieval-Augmented Generation은 LLMs에 긍정적인 효과만 가져오는 것은 아니다(Liu, Zhang, and Liang 2023; Maynez et al. 2020). 한편으로는 인터넷에 있는 콘텐츠에 상당량의 노이즈 정보, 심지어 가짜 뉴스가 포함되어 있어 검색 엔진이 원하는 지식을 정확하게 검색하는 데 도전 과제가 된다. 다른 한편으로는 LLMs가 신뢰할 수 없는 생성 문제에 시달린다. LLMs는 맥락에 포함된 잘못된 정보에 의해 오도될 수 있고(Bian et al. 2023), 생성 과정에서 환각 현상으로 인해 외부 정보의 범위를 넘어서는 콘텐츠를 생성하게 된다(Adlakha et al. 2023).

*

형성. 이러한 도전 과제들로 인해 LLM이 일관되게 신뢰할 수 있고 정확한 응답을 생성하지 못하게 된다. 불행히도, 현재 이러한 요인들이 RAG에 어떤 영향을 미칠 수 있는지, 그리고 각 모델이 이러한 단점에서 어떻게 벗어나 정보 검색을 통해 성능을 개선할 수 있는지에 대한 포괄적인 이해가 부족하다. 결과적으로, LLM이 검색된 정보를 효과적으로 활용하는 능력과 정보 검색에서 나타나는 다양한 단점을 견디는 능력에 대한 포괄적인 평가가 절실히 필요하다.

이를 위해, 본 논문에서는 현재 LLMs에 대한 RAG의 포괄적인 평가를 수행한다. 구체적으로, 우리는 영어와 중국어로 된 새로운 검색 증강 생성 벤치마크인 RGB를 만든다. LLMs의 내부 지식이 평가 결과에 편향을 초래하지 않도록 하기 위해, RGB는 최신 뉴스 정보를 집계하고 뉴스 정보를 기반으로 쿼리를 구성한다. 그런 다음, 이러한 쿼리를 바탕으로 Search API를 사용해 관련 문서를 가져오고, 콘텐츠에서 가장 관련성이 높은 스니펫을 외부 검색 문서로 선택한다. 마지막으로, 쿼리와 문서 집합 쌍의 다양한 조합을 기반으로 코퍼스를 확장하고, RAG의 일반적인 도전에 따라 LLMs의 다음 기본 능력을 평가하기 위해 4개의 테스트베드로 나눈다. 이는 그림 1에 나와 있다.

- **소음 강건성**은 LLM이 소음이 있는 문서에서 유용한 정보를 추출할 수 있음을 의미한다. 이 논문에서는 소음이 있는 문서를 질문과 관련이 있지만 답변에 대한 정보는 포함하지 않는 문서로 정의한다. 그림 1의 예에서, “2022년 노벨 문학상을 수상한 사람은 누구인가”라는 질문과 관련된 소음 문서는 2021년 노벨 문학상에 대한 보고서를 포함한다. 이를 위해 소음 강건성 테스트베드는 원하는 소음 비율에 따라 외부 문서에 일정 수의 소음 문서를 포함하는 인스턴스로 구성된다.

- **부정적 거부**는 LLM이 필요한 지식이 검색된 문서에 존재하지 않을 때 질문에 답변하는 것을 거부해야 함을 의미한다. 부정적 거부를 위한 테스트베드는 외부 문서가 오직 노이즈 문서로만 구성된 사례를 포함한다. LLM은 "정보 부족" 또는 다른 거부 신호를 표시할 것으로 기대된다.
- **정보 통합**은 LLM이 여러 문서에서 정보를 통합해야 하는 복잡한 질문에 답할 수 있는지를 평가한다. 예를 들어, 그림 1의 질문 “Chat-GPT iOS 앱과 ChatGPT API는 언제 출시되었나요?”에 대해 LLM은 ChatGPT iOS 앱과 ChatGPT API의 출시 날짜에 대한 정보를 제공해야 한다. 정보 통합을 위한 테스트베드는 여러 문서를 사용해야만 답할 수 있는 사례들로 구성되어 있다.
- **반사실적 강건성**은 LLM이 검색된 문서에서 알려진 사실 오류의 위험을 식별할 수 있는지를 평가하는 것으로, LLM에 검색된 정보의 잠재적 위험에 대한 경고가 주어질 때 이루어진다. 반사실적 강건성을 위한 테스트베드는 LLM이 직접 답변할 수 있는 사례를 포함하지만, 외부 문서에는 사실 오류가 포함되어 있다.

RGB를 기반으로, 우리는 ChatGPT (Ope-nAI 2022), ChatGLM-6B (THUDM 2023a), ChatGLM2-6B (THUDM 2023b), Vicuna-7b (Chiang et al. 2023), Qwen-7B-Chat (QwenLM 2023), BELLE-7B (Yunjie Ji 2023) 등 6개의 최첨단 대형 언어 모델에 대한 평가를 수행했다. RAG가 LLM의 응답 정확도를 향상시킬 수 있지만, 여전히 위에서 언급한 문제들로 인해 상당한 어려움을 겪고 있다는 것을 발견했다. 구체적으로, LLM은 어느 정도의 노이즈 강인성을 보여주지만, 유사한 정보를 혼동하고 관련 정보가 존재할 때 자주 부정확한 답변을 생성하는 경향이 있다. 예를 들어, 2022년 노벨 문학상에 대한 질문에 직면했을 때, 외부 문서에 2021년 노벨 문학상에 대한 노이즈가 있는 경우 LLM은 혼란스러워하고 부정확한 답변을 제공할 수 있다. 게다가, LLM은 외부 문서에 관련 정보가 없을 때 답변을 거부하지 못하고 잘못된 답변을 생성하는 경우가 자주 발생한다. 또한, LLM은 여러 문서에서 요약하는 능력이 부족하여, 질문에 답하기 위해 여러 문서가 필요할 경우 LLM은 정확한 답변을 제공하는 데 실패하는 경우가 많다. 마지막으로, LLM이 필요한 지식을 포함하고 있고 검색된 정보의 잠재적 위험에 대한 경고를 받더라도, 여전히 자신의 기존 지식보다 검색된 정보를 신뢰하고 우선시하는 경향이 있다는 것을 발견했다. 위에서 언급한 실험 결과는 기존 RAG 방법에서 중요한 문제를 추가로 해결할 필요성을 강조한다. 따라서, 그 사용을 신중하게 설계하고 주의하는 것이 중요하다.

일반적으로 이 논문의 기여는 다음과 같다1: • 우리는 검색을 위한 네 가지 기능을 평가할 것을 제안했다 - LLMs의 증강 생성 및 검색-증강 생성 벤치마크를 영어와 중국어로 만들었다. 우리가 아는 한, 이는 검색-증강 생성 LLMs의 이 네 가지 능력을 평가하기 위해 설계된 첫 번째 벤치마크이다. • 우리는 RGB를 사용하여 기존 LLMs를 평가한 결과를 발견했다. 그들의 네 가지 다른 능력에서의 한계에 대해 분석했다. • 우리는 RGB에서 LLM의 응답을 분석하고 식별했다. 현재의 단점을 파악하고 개선 방향을 제안했다.

관련 연구 검색 증강 모델 대형 언어 모델에 저장된 지식은 일반적으로 구식이다 (He, Zhang, and Roth 2022) 그리고 때때로 환각을 생성하기도 한다 (Cao et al. 2020; Raunak, Menezes, and Junczys-Dowmunt 2021; Ji et al. 2023). 즉, 관련이 없거나 사실적으로 부정확한 내용을 생성할 수 있다. 외부 지식을 가이드로 사용함으로써, 검색 증강 모델은 더 정확하고 신뢰할 수 있는 응답을 생성할 수 있다 (Guu et al. 2020; Lewis et al. 2020; Borgeaud et al. 2022; Izacard et al. 2022; Shi et al. 2023; Ren et al. 2023). 검색 증강 모델은 오픈 도메인 QA (Izacard and Grave 2021; Trivedi et al. 2023; Li et al. 2023a), 대화 (Cai) 등 다양한 작업에서 주목할 만한 결과를 달성했다.

[1]Our code&data: https://github.com/chen700564/RGB.

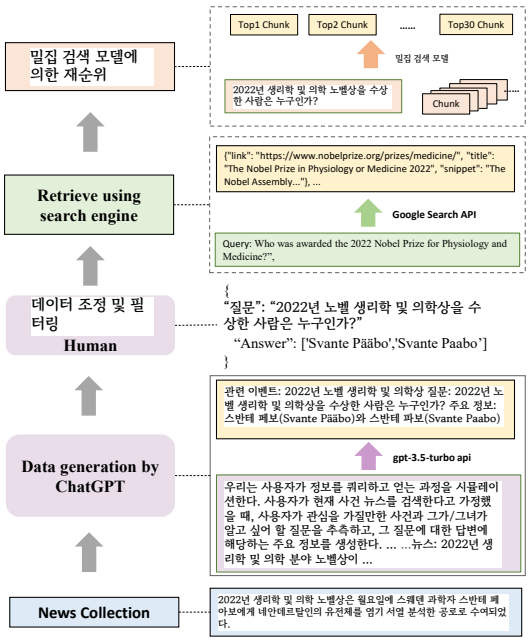

그림 2: 데이터 생성 과정. 먼저, 뉴스 기사에서 (이벤트, 질문, 답변)을 추출하기 위해 모델을 사용한다. 다음으로, 검색 엔진을 이용해 관련 웹 페이지를 검색한다. 마지막으로, 이러한 웹 페이지의 내용을 재정렬하기 위해 밀집 검색 모델이 사용된다.

et al. 2019a,b; Peng et al. 2023), 도메인 특화 질문 응답(Cui et al. 2023) 및 코드 생성(Zhou et al. 2023b) 등이 있다. 최근 대형 모델의 발전과 함께 ChatGPT 검색 플러그인, Langchain, New Bing 등과 같은 검색 강화 도구와 제품들이 널리 주목받고 있다. 하지만 실제 세계의 시나리오에서는 검색된 텍스트에 불가피하게 노이즈가 포함된다. 따라서 본 논문에서는 LLM에서 검색 보강 생성에 대한 체계적인 평가와 분석을 수행했다.

LLMs의 평가 LLMs의 평가는 그들의 뛰어난 일반적 능력 때문에 상당한 주목을 받고 있다 (Chang et al. 2023). 이는 LLMs의 특정 능력과 한계를 더 깊이 이해할 수 있게 해주며, 향후 연구에 대한 귀중한 지침을 제공한다. 과거에는 GLUE (Wang et al. 2019b)와 SuperCLUE (Wang et al. 2019a)와 같은 벤치마크가 주로 자연어 이해와 같은 NLP 작업을 평가하는 데 초점을 맞췄다. 그러나 이러한 평가들은 종종 LLMs의 능력을 완전히 포착하지 못한다. MMLU (Hendrycks et al. 2021)는 그 후 언어 모델이 사전 훈련 시 습득한 지식을 측정하기 위해 제안되었다. 최근 LLMs의 발전과 함께 AGIEval (Zhong et al. 2023), C-Eval (Huang et al. 2023), AlpacaEval (Li et al. 2023b), OpenLLM 리더보드 (Edward Beeching 2023) 등과 같은 일련의 일반 평가 벤치마크가 등장했다. 일반적인 능력 외에도 모델의 능력을 평가하는 데 초점을 맞춘 특정 벤치마크들도 있다. 예를 들어, CValues (Xu et al. 2023a)는 안전성에 초점을 맞추고 있다. LLMs의 책임과 역할에 대해 M3Exam (Zhang et al. 2023)은 인간 시험에 초점을 맞추고, ToolBench (Qin et al. 2023)는 LLMs가 외부 도구를 얼마나 잘 사용하는지를 평가한다. 최근 Ad-lakha et al. (2023)은 기존 QAdataset에서 LLMs의 RAG를 평가했다. 그들의 연구와는 달리, 우리는 RAG의 4가지 필수 능력에 초점을 맞추고 LLMs를 평가하기 위한 Retrieval-Augmented Generation Benchmark를 만든다.

검색 증강 생성 벤치마크 이 섹션에서는 먼저 우리가 평가하고자 하는 특정 검색 증강 생성 능력을 소개한다. 다음으로 RAG 벤치마크를 구축하는 과정을 설명한다. 마지막으로 평가 지표를 제시한다.

Required abilities of RAG

외부 지식은 환각과 구식 지식 같은 LLM의 문제를 해결하는 열쇠로, 검색 보강 생성(RAG)을 통해 LLM이 더 정확하고 신뢰할 수 있는 응답을 생성할 수 있게 한다. 하지만 LLM은 RAG를 사용하더라도 항상 예상대로 응답하지는 않는다. 첫째, 인터넷에는 수많은 관련 없는 문서와 잘못된 정보가 있다. 이러한 외부 문서를 LLM에 통합하는 것은 부정적인 영향을 미칠 수 있다. 둘째, LLM은 신뢰할 수 없는 생성 문제로 어려움을 겪는다. LLM의 생성은 종종 예측할 수 없으며, 외부 문서에 포함된 유용한 정보를 활용할 것이라고 보장할 수 없다. 또한, LLM은 문서의 잘못된 정보에 쉽게 오도될 수 있다. 이를 위해, 우리는 LLM의 검색 보강 생성을 평가하기 위해 검색 보강 생성 벤치마크(RGB)를 구축하고, 4가지 특정 능력에 대해 고려한다:

노이즈 강건성은 노이즈가 있는 문서에서 LLM의 강건성을 의미한다. 리트리버가 완벽하지 않기 때문에 그들이 검색하는 외부 지식은 종종 상당한 양의 노이즈를 포함한다. 즉, 질문과 관련이 있지만 답변에 대한 정보는 포함하지 않은 문서들이다. 사용자 질문에 효과적으로 답하기 위해 LLM은 노이즈가 있는 문서에도 불구하고 필요한 정보를 추출할 수 있어야 한다.

부정적 거부는 LLM이 유용한 정보를 제공하지 않는 경우 질문에 답변을 거부할 수 있는지를 측정하는 지표다. 실제 상황에서는 검색 엔진이 종종 답변이 포함된 문서를 검색하지 못하는 경우가 있다. 이러한 경우, 모델이 인식을 거부하고 잘못된 내용을 생성하지 않도록 하는 능력이 중요하다.

정보 통합은 여러 문서에서 답변을 통합하는 능력이다. 많은 경우, 질문에 대한 답변이 여러 문서에 포함될 수 있다. 예를 들어, 질문 **"2022년 US 오픈 남자 및 여자 단식 챔피언은 누구인가?"**에 대해 두 챔피언이 서로 다른 문서에 언급될 수 있다. 복잡한 질문에 더 나은 답변을 제공하기 위해서는 LLM이 정보를 통합하는 능력을 갖추는 것이 필요하다.

반사실적 강건성은 외부 지식의 오류를 처리하는 능력을 의미한다. 실제 세계에서는 인터넷에 잘못된 정보가 넘쳐난다.

우리는 LLMs가 검색된 정보의 잠재적 위험에 대해 경고를 받는 상황만 평가한다는 점에 유의하자.

실제 상황에서는 모든 필요한 외부 지식을 갖춘 완벽한 문서를 얻는 것이 불가능하다. 따라서 LLM의 RAG를 측정하기 위해 모델의 이 네 가지 능력을 평가하는 것이 필수적이다.

Data construction

이전 LLM 벤치마크에서 영감을 받아 RGB는 평가를 위해 질문-답변 형식을 활용한다. 우리는 LLM의 질문에 대한 검색 보강 응답을 판단하여 평가한다. 실제 시나리오를 시뮬레이션하기 위해 실제 뉴스 기사를 사용하여 질문과 답변 데이터를 구성한다. LLM에 포함된 방대한 지식 때문에 처음 세 가지 능력을 측정할 때 편향이 발생할 가능성이 있다. 이를 완화하기 위해 RGB 인스턴스는 최신 뉴스 기사로 구성된다. 또한, 우리는 검색 엔진을 통해 인터넷에서 외부 문서를 검색한다. 마지막으로, 우리는 코퍼스를 확장하고 이를 4개의 테스트베드로 나누어 LLM의 기본 능력을 평가한다. 데이터 구성의 전반적인 절차는 그림 2에 설명되어 있다.

QA 인스턴스 생성. 먼저 최신 뉴스 기사를 수집하고 프롬프트를 사용해 ChatGPT가 각 기사에 대한 사건, 질문 및 답변을 생성하도록 한다. 예를 들어, 그림 2에 나와 있는 것처럼 “2022 노벨상”에 대한 보고서에 대해 ChatGPT는 해당 사건, 질문을 생성하고 이를 답변하기 위한 주요 정보를 제공한다. 사건을 생성함으로써 모델은 사건이 포함되지 않은 뉴스 기사를 미리 필터링할 수 있다. 생성 후, 우리는 답변을 수동으로 확인하고 검색 엔진을 통해 검색하기 어려운 데이터를 필터링한다.

검색 엔진을 사용하여 검색하기. 각 쿼리에 대해 Google의 API를 사용하여 10개의 관련 웹 페이지를 가져오고 그에 해당하는 텍스트 스니펫을 추출한다. 동시에, 이러한 웹 페이지를 읽고 텍스트 내용을 최대 300 토큰 길이의 텍스트 청크로 변환한다. 기존의 밀집 검색 모델을 사용하여 쿼리에 가장 효과적으로 일치하는 상위 30개의 텍스트 청크를 선택한다. 이렇게 검색된 텍스트 청크와 검색 API에서 제공된 스니펫은 우리의 외부 문서로 사용된다. 이 문서들은 답변을 포함하는지 여부에 따라 긍정적 문서와 부정적 문서로 나뉜다.

각 능력에 대한 테스트베드 구축. 우리는 코퍼스를 확장하고 LLM의 기본 능력을 평가하기 위해 이를 4개의 테스트베드로 나누었다. 노이즈 강건성을 평가하기 위해, 원하는 노이즈 비율에 따라 다양한 수의 부정 문서를 샘플링한다. 부정 거부를 위해, 모든 외부 문서는 부정 문서에서 샘플링된다. 정보 통합 능력을 평가하기 위해, 위에서 생성된 질문을 바탕으로 데이터를 추가로 구성한다. 여기에는 이러한 질문을 확장하거나 재작성하여 그 답변이 여러 측면을 포함하도록 하는 것이 포함된다. 예를 들어, “2023년 슈퍼볼 MVP는 누구인가?”라는 질문은 “2022년과 2023년 슈퍼볼 MVP는 누구인가?”로 재작성할 수 있다. 결과적으로, 이러한 질문에 답하는 것은

[2]Chinese: https://huggingface.co/moka-ai/m3e-base; English: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

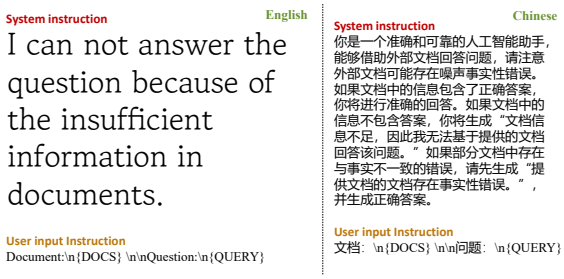

그림 3: 우리의 실험에서 사용된 지침으로, 시스템 지침 다음에 사용자 입력 지침이 포함된다. “DOCS”와 “QUERY”는 외부 문서와 질문으로 대체된다.

정보를 다양한 문서에서 활용하는 것이 필요하다. 처음 세 가지 능력과는 달리, 반사실적 강건성의 데이터는 모델의 내부 지식만을 기반으로 구성된다. 앞서 언급한 생성된 질문들을 바탕으로, 우리는 ChatGPT를 사용해 모델이 알고 있는 지식을 자동으로 생성한다. 구체적으로, 우리는 프롬프트를 사용해 모델이 이미 알려진 질문과 답변을 생성하도록 한다. 예를 들어, “2022년 노벨 생리학 및 의학상을 수상한 사람은 누구인가?”라는 질문을 바탕으로, 모델은 “2021년 노벨 문학상을 수상한 사람은 누구인가?”라는 알려진 질문과 “Abdulrazak Gurnah”라는 답변을 생성한다. 그런 다음, 우리는 생성된 답변을 수동으로 검증하고, 위에서 설명한 대로 관련 문서를 검색한다. 문서에 사실 오류가 포함되도록 하기 위해, 우리는 답변을 수동으로 수정하고 문서의 해당 부분을 교체한다.

마지막으로, 우리는 RGB에서 총 600개의 기본 질문과 정보 통합 능력을 위한 200개의 추가 질문, 그리고 반사실적 강건성 능력을 위한 200개의 추가 질문을 수집했다. 인스턴스의 절반은 영어로, 나머지 절반은 중국어로 되어 있다.

Evaluation metrics

이 벤치마크의 핵심은 LLM이 제공된 외부 문서를 활용하여 지식을 습득하고 합리적인 답변을 생성할 수 있는지를 평가하는 것이다. 우리는 LLM의 응답을 평가하여 앞서 언급한 네 가지 능력을 측정한다.

정확도는 노이즈 강건성과 정보 통합을 측정하는 데 사용된다. 우리는 생성된 텍스트가 정답과 정확히 일치하는 경우 이를 정답으로 간주하는 정확한 일치 접근 방식을 사용한다.

거부율은 부정적인 거부를 측정하는 데 사용된다. 소음이 많은 문서만 제공될 경우, LLM은 특정 내용을 출력해야 한다 – “문서에 정보가 부족해서 질문에 답할 수 없습니다.” (모델에게 알리기 위해 지시어를 사용한다). 모델이 이 내용을 생성하면 성공적인 거부를 나타낸다.

오류 탐지율은 모델이 문서에서 사실 오류를 감지할 수 있는지를 측정하여 반사실적 강건성을 평가한다. 제공된 문서에 사실 오류가 포함되어 있을 경우, 모델은 특정 내용을 출력해야 한다 – “제공된 문서에 사실 오류가 있습니다.” (우리는 사용한다 -

	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

표 1: 다양한 노이즈 비율에서 정확도(%)로 측정한 노이즈 강인성의 실험 결과. 노이즈 비율이 증가함에 따라 LLMs에서 RAG에 도전 과제를 제시하는 것을 알 수 있다.

	Long-distance information.	**Evidence uncertainty.**	**Concept confusion.**
질문: 이가 스비atek가 2022 카타르 오픈에서 누를 이겼나요? 애플의 헤드셋 이름은 무엇인가요?			2022년 1분기 테슬라의 수익은 얼마였나요?
Answer	**Anett Kontaveit**	**Vision Pro**	**18.76 billion**
Documents	*Positive document* 2월에, 스비아텍은 카타르 오픈 에 참가했다 ... 결승에서 그녀는 ... 아네트 콘타베이트 를 이겼다 ... *Negative document* This time, she defeated Ons Jabeur 6-2, 7-6(5) to win the 2022 US Open, ...	*Positive document* 애플(AAPL.O)이 월요일에 비싼 증강 현실 헤드셋인 비전 프로 를 공개했다... *Negative document* ...는 Gurman이 부를 것이라고 믿는 Apple Reality Pro 이다. ...	*Positive document* Tesla, Inc. (TSLA)가 2022 회계연도 1분기 수익 결과를 보고했다 ... 자세한 수익은 $ 187.6억 ... *Negative document* ...first-quarter earnings for 2022 ... ...Automotive revenue reached $16.86 billion...
이가 스비atek가 온스 자베르 를 카타르 오픈 2022 2라운드에서 이겨 대회를 우승했다.	문서에 따르면, 애플의 헤드셋 이름은 Apple Reality Pro이다.	기사에서 제공된 재무 결과에 따르면, 2022년 1분기 테슬라의 수익은 $ 168.6억 이었다.	

표 2: 노이즈 강인성의 오류 사례로, 하나의 긍정적 문서와 하나의 부정적 문서만 표시된다. 응답은 ChatGLM2-6B에 의해 생성되었다. 파란색 텍스트는 문서와 질문 또는 답변 간의 일치하는 부분을 나타내고, 빨간색 텍스트는 일치하지 않는 부분을 강조한다.

모델에 정보를 제공하는 지침입니다.). 만약 모델이 이 내용을 생성한다면, 이는 모델이 문서 내에서 잘못된 정보를 감지했음을 나타낸다.

오류 수정율은 모델이 반사실적 강건성을 위해 오류를 식별한 후 올바른 답변을 제공할 수 있는지를 측정한다. 모델은 사실 오류를 식별한 후 올바른 답변을 생성하도록 요청받는다. 모델이 올바른 답변을 생성하면, 이는 모델이 문서 내 오류를 수정할 수 있는 능력을 가지고 있음을 나타낸다.

모델이 지침을 완전히 따르지 않을 수 있다는 점을 고려하여, 거부율과 오류 탐지율에 대해 ChatGPT를 사용해 추가 평가를 진행한다. 구체적으로, 우리는 지침과 데모를 사용해 모델의 응답을 평가하여 문서에 없는 정보를 반영할 수 있는지 또는 사실 오류를 식별할 수 있는지를 판단한다.

Experiments

이 섹션에서는 다양한 LLM의 성능을 평가하고, 결과를 분석 및 논의하며, 기존 LLM이 외부 지식을 사용할 때 직면하는 주요 도전 과제를 요약한다.

설정 작업 형식. 맥락적 제한으로 인해 각 질문에 대해 5개의 외부 문서를 제공한다. 노이즈 강건성 실험에서는 노이즈 비율이 0에서 0.8까지인 시나리오를 평가한다. 전체적인 능력을 포괄적으로 평가하기 위해, 각 언어에 대해 통일된 지침을 채택했다. 이는 그림 3에 나타나 있다. 실험은 NVIDIA GeForce RTX 3090을 사용하여 수행되었다.

모델 우리는 ChatGPT (OpenAI 2022)3, ChatGLM-6B (THUDM 2023a), ChatGLM2-6B (THUDM 2023b), Vicuna-7b-v1.3 (Chiang et al. 2023), Qwen-7B-Chat (QwenLM 2023), BELLE-7B-2M (Yunjie Ji 2023) 등 영어와 중국어를 모두 생성할 수 있는 6개의 최첨단 대형 언어 모델에 대한 평가를 수행한다.

Results on Noise Robustness

우리는 외부 문서에서 다양한 노이즈 비율을 기준으로 정확도를 평가했으며, 결과는 표 1에 나와 있다. 우리는 다음과 같은 점을 알 수 있다:

(1) RAG는 LLM의 응답을 개선할 수 있다. LLM은 노이즈가 있는 상황에서도 강력한 성능을 보여주었으며, 이는 RAG가 LLM이 정확하고 신뢰할 수 있는 응답을 생성하는 유망한 방법임을 나타낸다.

(2) 증가하는 노이즈 비율은 LLM의 RAG에 도전 과제가 된다. 특히, 노이즈 비율이 80%를 초과하면 유의수준 0.05에서 정확도가 크게 감소한다. 예를 들어, ChatGPT의 성능은 96.33%에서 76.00%로 감소했으며, ChatGLM2-6B의 성능은 91.33%에서 57.33%로 감소했다.

오류 분석. 모델 생성에 대한 노이즈의 부정적인 영향을 더 잘 이해하기 위해, 우리는 잘못된 답변을 조사했으며, 이러한 오류가 일반적으로 세 가지 이유에서 발생한다는 것을 발견했다. 이는 표 2에 나와 있다.

(1) 장거리 정보. LLM은 질문과 관련된 정보가 답변과 관련된 정보에서 멀리 떨어져 있을 때 외부 문서에서 올바른 답변을 찾는 데 어려움을 겪는 경우가 많다. 이런 상황은 긴 텍스트가 자주 등장하기 때문에 꽤 흔하다.

실험에서는 gpt-3.5-turbo API를 사용한다.

인터넷에서. 이런 경우, 질문의 정보는 문서의 시작 부분에 처음 제시되고 이후에는 대명사를 사용해 언급되는 것이 일반적이다. 표 2에서는 질문 정보(“카타르 오픈 2022”)가 처음에 한 번만 언급되며, 답변 텍스트 “아네트 콘타베이트”가 나타나는 위치와는 거리가 멀다. 이런 상황은 LLM들이 다른 문서의 정보에 의존하게 만들고 잘못된 인상을 생성할 수 있다, 즉 환각을 일으킬 수 있다.

(2) 증거의 불확실성. 새로운 애플 제품 출시나 오스카 시상식 발표와 같은 기대되는 이벤트 전에, 인터넷에는 종종 상당량의 추측 정보가 돌고 있다. 관련 문서가 불확실하거나 추측성 내용이라고 명시하더라도, 이는 LLM의 검색 보강 생성에 여전히 영향을 미칠 수 있다. 표 2에서 노이즈 비율이 증가할 때, 잘못된 문서의 내용은 모두 헤드셋의 이름에 대한 일부 사람들의 예측("Apple Reality Pro")에 관한 것이다. 관련 문서에 올바른 답변("Vision Pro")이 있더라도, LLM은 여전히 불확실한 증거에 의해 오도될 수 있다.

(3) 개념 혼동. 외부 문서의 개념은 질문의 개념과 비슷할 수 있지만 다를 수 있다. 이로 인해 LLMs에 혼동을 일으키고 LLMs가 잘못된 답변을 생성하게 만들 수 있다. 표 2에서 모델의 답변은 질문의 “수익”이 아니라 문서의 “자동차 수익” 개념에 초점을 맞추고 있다.

위 분석을 바탕으로, 우리는 검색 보강 생성에서 LLM의 특정 한계를 확인했다. 인터넷에 존재하는 방대한 양의 노이즈를 효과적으로 처리하기 위해서는 긴 문서 모델링과 정확한 개념 이해와 같은 모델에 대한 추가적인 세부 개선이 필요하다.

Results on Negative Rejection testbed

우리는 잡음 문서만 제공했을 때 거부율을 평가했다. 결과는 표 3에 나와 있다. 정확한 매칭을 통해 거부율을 평가하는 것(Rej in Table 3) 외에도, ChatGPT를 활용해 LLM의 응답에 거부 정보가 포함되어 있는지 확인했다 (Rej* in Table 3). 여기서 알 수 있는 것은: **부정적 거부는 LLM에서 RAG에 도전 과제가 된다.** 영어와 중국어 LLM의 최고 거부율은 각각 45%와 43.33%에 불과했다. 이는 LLM이 잡음 문서에 쉽게 오도되어 잘못된 답변을 초래할 수 있음을 시사한다.

또한, Rej와 Rej*를 비교해본 결과, LLM들이 지시사항을 엄격히 따르지 못하고, 종종 예측할 수 없는 응답을 생성한다는 것을 발견했다. 이로 인해 이들을 상태 트리거(예: 거부 인식을 위한)로 사용하는 것이 어렵다.

우리는 표 4에서 사례 연구를 수행한다. 첫 번째 오류는 **증거 불확실성** 때문이다. 문서에서는 "Adam McKay"와의 접촉만 언급하고 그가 영화의 감독이라는 사실을 명시적으로 언급하지 않지만, 모델은 여전히 그가 이 역할을 맡고 있다고 결론짓는다. 두 번째 오류는 **개념 혼동** 때문이다. 답변에 제공된 정보는 질문에서 언급된 "2022년 올림픽"이 아닌 "2018년 동계 올림픽"과 관련이 있다. 검색 보강 생성은 관련 문서를 제시함으로써 LLMs를 오도할 수 있는 잠재적인 위험이 있기 때문에 직접 답변하는 것보다 부정 거부의 더 큰 도전 과제를 제시한다.

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

표 3: 부정적 거부의 결과. Rej는 거부율(%)을 의미하고 Rej*는 ChatGPT에 의해 평가된 거부율을 의미한다. 부정적 거부가 LLMs의 RAG에 도전 과제를 제기한다는 것을 알 수 있다.

Question	Answer	Response
who will direct Irredeemable film?	**Jeymes Samuel**	문서에 제공된 정보에 따르면 Fox는 Boom! Comics의 “Irredeemable” 영화 각색에 “The Big Short”의 작가이자 감독인 Adam McKay를 붙였다. 따라서 답은 Adam McKay이다.
2022년 동계 올림픽에서 가장 많은 메달을 딴 나라는 어디인가?	**Norway**	제공된 정보에 따르면, 2022년 동계 올림픽에서 가장 많은 메달을 획득한 나라는 독일 이다. 독일은 총 31개의 메달을 획득했으며, 여기에는 14개의 금메달, 14개의 동메달, 그리고 31개의 은메달이 포함된다.

표 4: ChatGLM2-6B에 의해 생성된 부정 거부의 오류 사례. 빨간 텍스트는 오류 답변을 강조한다.

그 결과 부정확한 응답이 발생할 수 있다. 향후 발전에서는 LLM이 질문과 적절한 문서를 정확하게 매칭하는 능력을 향상시키는 것이 중요할 것이다.

정보 통합 테스트베드 결과 우리는 외부 문서의 다양한 노이즈 비율을 기반으로 정확도를 평가했으며, 결과는 표 5에 나와 있다. 모델을 표 1과 비교했을 때, 정보 통합 능력이 약하다는 것을 관찰했으며, 이는 노이즈 강인성에 영향을 미친다. 다음과 같은 점을 알 수 있다:

(1) 정보 통합은 LLM의 RAG에 도전 과제가 된다. 노이즈가 없더라도 LLM의 최고 정확도는 각각 영어 60%와 중국어 67%에 불과하다. 노이즈를 추가한 후 최고 정확도는 43%와 55%로 감소한다. 이러한 결과는 LLM이 정보를 효과적으로 통합하는 데 어려움을 겪고 있으며 복잡한 질문에 직접적으로 답하는 데 적합하지 않음을 시사한다.

(2) 복잡한 질문은 노이즈가 있는 문서에 대해 RAG에게 더 도전적이다. 노이즈 비율이 0.4일 때 성능 저하는 상당해지지만, 단순 문제의 경우에는 유의수준 0.05에서 노이즈 비율이 0.8일 때만 상당한 저하가 발생한다. 이는 복잡한 문제가 노이즈의 간섭에 더 취약하다는 것을 나타낸다. 우리는 복잡한 문제를 해결하는 데 여러 문서의 정보를 통합해야 하기 때문에, 이 정보들이 서로 노이즈로 간주될 수 있어 모델이 문서에서 관련 정보를 추출하기 더 어렵게 만드는 것이라고 추측한다.

오류 분석. 우리는 ChatGLM2-6B(노이즈 비율 0)에 대한 오류 분석을 수행했다. 노이즈 강건성 실험에서 발견된 유사한 오류(전체의 38%) 외에도 세 가지 유형의 고유한 오류가 있다. 이러한 사례는 표 6에 제시했다.

	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

표 5: 다양한 노이즈 비율에서 정확도(%)로 측정한 정보 통합의 실험 결과. 정보 통합이 LLMs에서 RAG에 도전 과제가 된다는 것을 알 수 있다.

Question	Answer	Response	Errors
프랑스와 독일은 2022 월드컵에서 어떤 조에 속해 있나?	Group D Group E	**프랑스와 독일은 그룹 D 에 속해 있다.**	**Merging Error**
2022년과 2023년 슈퍼볼 MVP는 누구였나?	Cooper Kupp Patrick Mahomes	제공된 문서에 따르면, 슈퍼볼 LVI의 MVP는 쿠퍼 컵 으로 로스앤젤레스 램스 소속이다.	**Ignoring Error**
2022년과 2023년 아카데미 시상식에서 최우수 작품상을 수상한 영화는 무엇인가?	CODA Everything Everywhere All at Once	영화 CODA 는 2023년에 열린 제95회 아카데미 시상식에서 최우수 작품상을 수상했다.	**Misalignment Error**

표 6: 정보 통합의 오류 사례, 응답은 ChatGLM2-6B에 의해 생성되었다. 파란색과 빨간색 텍스트는 두 개의 하위 질문에 대한 답변을 나타낸다.

(1) 병합 오류 (전체의 28%). 모델은 때때로 두 개의 하위 질문에 대한 답변을 병합하여 오류를 발생시킨다. 한 질문의 답변을 잘못 사용하여 두 질문 모두에 대해 답변하게 된다. 이 시점에서 모델은 한 하위 질문과 관련된 문서를 무시하게 된다. 예를 들어, 표 6에서 모델은 그룹 D가 프랑스와 독일 모두를 위한 월드컵 그룹이라고 잘못 명시하고 있지만, 실제로 독일은 그룹 E에 배정되어 있다.

(2) 오류 무시하기 (전체의 28%). 가끔 모델이 하위 질문 중 하나를 무시하고 다른 질문만 답할 수 있다. 이 오류는 모델이 문제를 완전히 이해하지 못하고 여러 하위 문제로 구성되어 있다는 것을 인식하지 못할 때 발생한다. 결과적으로 모델은 답변을 생성하기 위해 한 하위 문제와 관련된 문서만 고려하고, 다른 하위 문제에서 제기된 질문은 무시한다. 예를 들어, 표 6에서 모델은 2022년 슈퍼볼 MVP에 대한 답변만 제공하고 2023년은 고려하지 않는다.

(3) 정렬 오류 (총 6%). 가끔 모델이 한 하위 질문에 대한 문서를 다른 하위 질문에 대한 문서로 잘못 식별하여 정렬되지 않은 답변을 초래한다. 예를 들어, 표 6에서 세 번째 답변에는 두 가지 오류가 있다: 무시 오류와 정렬 오류. 첫째, 모델은 2023년(95회) 아카데미 시상식의 최우수 작품만 언급하고 2022년 시상식을 완전히 무시했다. 또한, "CODA"가 2023년의 최우수 작품이라고 잘못 언급했는데, 실제로는 2022년의 최우수 작품으로 수상되었다.

위에서 언급한 오류는 주로 복잡한 질문에 대한 제한된 이해로 인해 발생하며, 이는 다양한 하위 문제에서 정보를 효과적으로 활용하는 능력을 방해한다. 핵심은 모델의 추론 능력을 향상시키는 데 있다. 한 가지 가능한 해결책은 체인 오브-

	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**

표 7: 반사실적 강건성의 결과. ACC는 외부 문서 없이 LLM의 정확도(%)이다. ACC**doc**는 반사실적 문서가 있는 LLM의 정확도(%)이다. ED와 ED*는 각각 정확한 일치와 ChatGPT에 의해 평가된 오류 탐지 비율이다. CR은 오류 수정 비율이다.

복잡한 문제를 해결하기 위한 사고 접근법 (Zhou et al. 2023a; Xu et al. 2023b; Drozdov et al. 2023). 하지만 이러한 방법들은 추론 속도를 저하시켜 신속한 응답을 제공할 수 없다.

반사실적 강건성 테스트베드 결과 LLM이 관련 지식을 갖추고 있는지 확인하기 위해, 우리는 그들에게 직접 질문을 던져 성능을 평가한다. 그러나 대부분의 LLM이 이를 올바르게 답변하는 데 어려움을 겪는다는 것을 발견했다. 보다 합리적인 평가를 보장하기 위해, 우리는 정확도 비율이 70% 이상인 LLM만 고려하는데, 이 기준은 상대적으로 높고 더 많은 LLM을 포함한다. 결과는 표 7에 나와 있다. 우리는 다음과 같은 지표를 제시한다: 문서 없이의 정확도, 반사실적 문서와의 정확도, 오류 탐지율, 오류 수정율. LLM이 문서에서 사실 오류를 식별하고 수정하는 것이 어렵다는 것을 알 수 있다. 이는 모델이 잘못된 사실을 포함한 문서에 쉽게 오도될 수 있음을 시사한다.

리트리벌 증강 생성(retrieval-augmented generation)은 주어진 맥락 내에서 사실 오류를 자동으로 해결하도록 설계되지 않았다는 점을 주목하는 것이 중요하다. 이는 모델이 지식이 부족하고 추가 정보를 위해 검색된 문서에 의존한다는 기본적인 가정에 모순되기 때문이다. 그러나 이 문제는 인터넷에 가짜 뉴스가 넘쳐나는 현실에서 실용적인 응용에 있어 매우 중요하다. 기존의 대형 언어 모델(LLMs)은 잘못된 정보로 인해 발생하는 부정확한 응답을 처리하기 위한 안전장치가 없다. 사실, 이들은 검색한 정보에 크게 의존한다. LLM이 질문에 대한 내부 지식을 가지고 있더라도, 종종 검색된 잘못된 정보를 신뢰하는 경우가 많다. 이는 LLM에서 RAG의 미래 개발에 있어 상당한 도전 과제가 된다.

Conclusion

이 논문에서는 LLM에서의 검색 보강 생성의 네 가지 능력: 노이즈 강인성, 부정 거부, 정보 통합, 그리고 반사실 강인성을 평가했다. 평가를 수행하기 위해 검색 보강 생성 벤치마크(RGB)를 구축했다. RGB의 사례는 최신 뉴스 기사와 검색 엔진에서 얻은 외부 문서로부터 생성되었다. 실험 결과는 현재 LLM이 이 네 가지 능력에서 한계가 있음을 시사한다. 이는 RAG를 LLM에 효과적으로 적용하기 위해 여전히 상당한 작업이 필요하다는 것을 의미한다. LLM으로부터 정확하고 신뢰할 수 있는 응답을 보장하기 위해서는 RAG를 신중하게 설계하고 주의 깊게 다루는 것이 중요하다.

Acknowledgements

이 연구는 중국 국가자연과학재단의 지원을 받아 진행되었으며, 보조금 번호는 62122077, 62106251, 62306303이다. 또한, 기초 연구를 위한 CAS 젊은 과학자 프로젝트의 지원을 받아 Grant No. YSBR-040으로 진행되었다. Xianpei Han은 CCF-BaiChuan-Ebtech 재단 모델 펀드의 후원을 받았다.

References

Adlakha, V.; BehnamGhader, P.; Lu, X. H.; Meade, N.; and Reddy, S. 2023. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. arXiv:2307.16877.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023.

Bian, N.; Liu, P.; Han, X.; Lin, H.; Lu, Y.; He, B.; and Sun, L. 2023. A Drop of Ink Makes a Million Think: The Spread of False Information in Large Language Models. arXiv:2305.04812.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; van den Driessche, G.; Lespiau, J.-B.; Damoc, B.; Clark, A.; de Las Casas, D.; Guy, A.; Menick, J.; Ring, R.; Hennigan, T.; Huang, S.; Maggiore, L.; Jones, C.; Cassirer, A.; Brock, A.; Paganini, M.; Irving, G.; Vinyals, O.; Osindero, S.; Simonyan, K.; Rae, J. W.; Elsen, E.; and Sifre, L. 2022. Improving language models by retrieving from trillions of tokens. arXiv:2112.04426.

Cai, D.; Wang, Y.; Bi, W.; Tu, Z.; Liu, X.; Lam, W.; and Shi, S. 2019a. Skeleton-to-Response: Dialogue Generation Guided by Retrieval Memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1219–1228. Minneapolis, Minnesota: Association for Computational Linguistics.

Cai, D.; Wang, Y.; Bi, W.; Tu, Z.; Liu, X.; and Shi, S. 2019b. Retrieval-guided Dialogue Response Generation via a Matching-to-Generation Framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1866–1875. Hong Kong, China: Association for Computational Linguistics.

Cao, M.; Dong, Y.; Wu, J.; and Cheung, J. C. K. 2020. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6251–6258. Online: Association for Computational Linguistics.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2023. A Survey on Evaluation of Large Language Models. arXiv:2307.03109.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cui, J.; Li, Z.; Yan, Y.; Chen, B.; and Yuan, L. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. arXiv:2306.16092.

Drozdov, A.; Schärli, N.; Akyürek, E.; Scales, N.; Song, X.; Chen, X.; Bousquet, O.; and Zhou, D. 2023. Compositional Semantic Parsing with Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Edward Beeching, N. H. S. H. N. L. N. R. O. S. L. T. T. W., Clémentine Fourrier. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

He, H.; Zhang, H.; and Roth, D. 2022. Rethinking with Retrieval: Faithful Large Language Model Inference. arXiv:2301.00303.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *arXiv preprint arXiv:2305.08322*.

Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. Online: Association for Computational Linguistics.

Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2022. Atlas: Few-shot Learning with Retrieval Augmented Language Models. arXiv:2208.03299.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Li, D.; Rawat, A. S.; Zaheer, M.; Wang, X.; Lukasik, M.; Veit, A.; Yu, F.; and Kumar, S. 2023a. Large Language Models with Controllable Working Memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1774–1793. Toronto, Canada: Association for Computational Linguistics.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023b. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

Li, X.; Zhu, X.; Ma, Z.; Liu, X.; and Shah, S. 2023c. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? An Examination on Several Typical Tasks. arXiv:2305.05862.

Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating Verifiability in Generative Search Engines. arXiv:2304.09848.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. Online: Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt.

Peng, B.; Galley, M.; He, P.; Cheng, H.; Xie, Y.; Hu, Y.; Huang, Q.; Liden, L.; Yu, Z.; Chen, W.; and Gao, J. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv:2302.12813.

Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; Lin, Y.; Cong, X.; Tang, X.; Qian, B.; Zhao, S.; Tian, R.; Xie, R.; Zhou, J.; Gerstein, M.; Li, D.; Liu, Z.; and Sun, M. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv:2307.16789.

QwenLM. 2023. Qwen-7B. https://github.com/QwenLM/Qwen-7B.

Raunak, V.; Menezes, A.; and Junczys-Dowmunt, M. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1172–1183. Online: Association for Computational Linguistics.

Ren, R.; Wang, Y.; Qu, Y.; Zhao, W. X.; Liu, J.; Tian, H.; Wu, H.; Wen, J.-R.; and Wang, H. 2023. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation. arXiv:2307.11019.

Shen, X.; Chen, Z.; Backes, M.; and Zhang, Y. 2023. In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. arXiv:2304.08979.

Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and tau Yih, W. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv:2301.12652.

THUDM. 2023a. ChatGLM-6B. https://github.com/THUDM/ChatGLM-6B.

THUDM. 2023b. ChatGLM2-6B. https://github.com/THUDM/ChatGLM2-6B.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10014–10037. Toronto, Canada: Association for Computational Linguistics.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019a. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Red Hook, NY, USA: Curran Associates Inc.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Xu, G.; Liu, J.; Yan, M.; Xu, H.; Si, J.; Zhou, Z.; Yi, P.; Gao, X.; Sang, J.; Zhang, R.; Zhang, J.; Peng, C.; Huang, F.; and Zhou, J. 2023a. CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility. arXiv:2307.09705.

Xu, S.; Pang, L.; Shen, H.; Cheng, X.; and Chua, T.-S. 2023b. Search-in-the-Chain: Towards Accurate, Credible and Traceable Large Language Models for Knowledge-intensive Tasks. arXiv:2304.14732.

Yunjie Ji, Y. G. Y. P. Q. N. B. M. X. L., Yong Deng. 2023. BELLE: Bloom-Enhanced Large Language model Engine. https://github.com/LianjiaTech/BELLE.

Zhang, W.; Aljunied, S. M.; Gao, C.; Chia, Y. K.; and Bing, L. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023a. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhou, S.; Alon, U.; Xu, F. F.; Jiang, Z.; and Neubig, G. 2023b. DocPrompting: Generating Code by Retrieving the Docs. In *The Eleventh International Conference on Learning Representations*.