KFREB: 생성형 한국어 대규모 언어 모델의 검색 기반 생성 평가 데이터셋

이정섭[1○], 손준영[1○], 이태민[2○], 박찬준[3], 강명훈[1], 박정배[2*], 임희석[1,2*]
고려대학교 컴퓨터학과[1], Human-inspired AI 연구소[2], Upstage[3]
{omanma1928, s0ny, taeminlee}@korea.ac.kr,
{bcj1210}@naver.com, {chaos8527, insmile, limhseok}@korea.ac.kr

KFREB: Korean Fictional Retrieval-based Evaluation Benchmark for Generative Large Language Models

Jungseob Lee[1○], Junyoung Son[1○], Taemin Lee[2○], Chanjun Park[3], Myunghoon Kang[1], Jeongbae Park[2*], Heuiseok Lim[1,2*]
[1]Department of Computer Science and Engineering, Korea University, [2]Human-inspired AI Research, [3]Upstage

요 약

본 논문에서는 대규모 언어모델의 검색 기반 답변 생성능력을 평가하는 새로운 한국어 벤치마크, KFREB(Korean Fictional Retrieval Evaluation Benchmark)를 제안한다. KFREB는 모델이 사전학습 되지 않은 허구의 정보를 바탕으로 검색 기반 답변 생성 능력을 평가함으로써, 기존의 대규모 언어모델이 사전학습에서 보았던 사실을 반영하여 생성하는 답변이 실제 검색 기반 답변 시스템에서의 능력을 제대로 평가할 수 없다는 문제를 해결하고자 한다. 제안된 KFREB는 검색기반 대규모 언어모델의 실제 서비스 케이스를 고려하여 장문 문서, 두 개의 정답을 포함한 골드 문서, 한 개의 골드 문서와 유사 방해 문서 키워드 유무, 그리고 문서 간 상호 참조를 요구하는 상호참조 멀티홉 리즈닝 경우 등에 대한 평가 케이스를 제공하며, 이를 통해 대규모 언어모델의 적절한 선택과 실제 서비스 활용에 대한 인사이트를 제공할 수 있을 것이다.

주제어: 대규모 언어모델, 환각지식, 검색 기반 답변 시스템, KFREB

1. 서론

최근 ChatGPT와 같이 사람처럼 언어를 이해하고 생성하는 능력을 보유한 다양한 대규모 언어모델(Large Language Model)의 등장으로, 다양한 서비스에서 이를 활용하는 추세가 점점 더 확산되고 있다 [1, 2, 3, 4, 5].

이러한 대규모 언어모델의 등장에도 불구하고, 대규모 언어모델은 공통적인 범위의 지식으로 답변을 생성하므로, 특정 기업에 대한 챗봇 등으로 활용하는 데 제한이 있다 [6, 7]. 또한, 대규모 언어모델은 지식을 묻는 것에도 환각지식(Hallucination) 문제를 겪는다 [6]. 이는 모델이 실제로는 존재하지 않는 정보나 사실을 생성하는 현상을 말한다.

이를 해결하기 위해 제한적인 문서를 검색하여 프롬프트에 포함하고, 문서를 통해 답변을 생성하는 검색 기반의 대규모 언어모델이 실제 산업에서 많이 활용되고 있다[1] [1, 8].

그러나, 이러한 검색 시스템과 결합된 대규모 언어모델을 선택하는 관점에서 벤치마크가 존재하지 않는다. 더욱이, 이러한 벤치마크가 개발되었다고 하더라도, 대규모 언어모델이 학습된 내제된 정보를 통해 답변을 생성하여 정답을 추론한 것인지, 프롬프트에 주어진 문서 정보에서 적절한 정보를 잘 찾아와서 답변하는지에 대한 제대로된 평가를 할 수 없다는 문제가 있다. 이를 극복하기 위해, 허구가 아닌 최신 기반의 실제 문서를 사용하여 벤치마크를 설정한다 하더라도, 추후 개발된 대규모 언어모델이 벤치마크 내의 데이터를 학습하여 내재된 정보를 통해 결과를 추론할 수 있으므로 모델의 검색 기반 생성 성능을 정확하게 평가하기 어려울 수 있다.

이러한 문제점을 극복하기 위해 본 논문에서는 검색 기반 대규모 언어모델에 대한 평가 벤치마크인 KFREB(Korean Fictional Retrieval Evaluation Benchmark)를 제안한다. KFREB는 대규모 언어모델이 사전학습에서 보았던 사실을 반영하여 생성하는 답변이 검색 기반 답변 시스템에서의 실제 능력을 제대로 평가할 수 없음을 지적한다. 이를 극복하기 위해, KFREB는 실제 사실이 아닌 허구 사실을 기반으로 하여 제작되었고, 이는 모델이 사전학습되지 않은 내용으로 모델의 순수 검색 기반 답변 생성 능력을 평가하는 것을 목표로 한다. 또한, 실제 서비스에서 사용되는 다음의 다섯 가지 케이스로 데이터의 카테고리를 구분하여 모델의 검색기반 생성 성능을 종합적으로 판단하도록 데이터를 구성하였다.

첫째, 한 개의 장문 (800글자 이상)의 골드 문서가 주어진 경우의 multiple choice이다. 둘째, 각각이 정답을 포함한 두 개의 골드 문서가 포함된 경우의 multiple choice이다. 셋째, 한 개의 골드 문서와 검색 시스템에서 검색됐을 것 같은 유사 방해 문서가 주어진 경우의 multiple choice로, 질문의 키워드가 포함된다. 넷째, 한 개의 골드 문서와 검색 시스템에서 검색됐을 것 같은 유사 방해 문서가 주어진 경우의 multiple choice로, 질문의 키워드가 포함되지 않는다. 마지막으로, 문서를 상호참조하는 경우의 멀티홉 리즈닝을 요구하는 두 개의 골드 문서의 multiple choice이다.

본 논문은 다양한 실제 서비스 케이스를 고려하여 적절한 목

*교신저자(Corresponding author)
[1]https://tools.zmo.ai/webChatGPT

적에 따른 서비스 선택을 할 수 있도록 인사이트를 제공하기 위한 벤치마크인 KFREB를 설계하고 벤치마크를 공개한다[2].

2. 관련연구

관련 연구로, 대화식 평가를 위한 다중 에이전트 프레임워크인 ChatEval을 제안한 연구가 있다 [9]. 이 프레임워크는 여러 대규모 언어 모델이 협력하여 다양한 모델이 생성한 응답의 품질을 자동으로 평가하고 토론한다. 실험 결과, ChatEval은 사람의 평가와 높은 일치도를 보여주며, 다양한 역할 프롬프트가 다중 에이전트 토론 과정에서 필수적임을 밝혔다. 이 연구는 단순한 텍스트 점수화를 넘어서 신뢰할 수 있는 평가를 제공하는 인간 모방 평가 과정을 제시하였다.

또한, 대규모 언어 모델의 조정을 개선하기 위해 PandaLM이라는 새로운 평가 모델을 제시한 연구도 있다 [10]. PandaLM은 여러 대규모 언어모델 중 우수한 모델을 판별하는 데 학습되며, 답변의 객관적 정확성뿐만 아니라 상대적 요약성, 명확성, 지시사항 준수, 포괄성, 공식성과 같은 주관적 요소에 초점을 맞춘다. 이 모델의 신뢰성을 보장하기 위해 다양한 인간이 주석을 단 테스트 데이터셋을 수집하였다. 결과적으로, PandaLM은 Alpaca의 기본 하이퍼파라미터로 학습된 모델들에 비해 향상된 성능을 보여주며, API 기반 평가에 의존하지 않아 데이터 유출 가능성을 줄인다.

또한, 대형 언어 모델의 추론 능력을 측정하는 새로운 평가 프레임워크를 제안한 연구도 있다 [11]. 이 연구는 행동과 변화에 대한 추론, 즉 인간 지능의 핵심적인 부분을 평가하는 복잡한 테스트 케이스를 제공한다. GPT-3, Instruct-GPT3, BLOOM 등의 모델을 이용한 결과, 이러한 복잡한 추론 작업에서는 미흡한 성능을 보였다. 이 연구는 대규모 언어모델의 실제 한계를 측정하기 위해 더욱 정교한 추론 문제를 살펴보는 것이 필요함을 보여준다.

위의 연구들은 대규모 언어모델을 활용하는 것에 있어서 다양한 평가 방법 혹은 데이터를 제시하였으나, 요즘 대규모 언어모델에서 선두에 있는 검색 기반 언어모델에 활용되는 점을 평가하지 않는다. 이러한 문제를 해결하기 위해 본 논문에서는 검색 시스템에 대한 평가 벤치마크를 제안한다. 본 논문에서 채택하는 벤치마크인 KFREB(Korean Fictional Retrieval Evaluation Benchmark)는 대규모 언어모델이 사전학습에서 보았던 사실을 반영하여 생성하는 답변이 검색 기반 답변 시스템에서의 실제 능력을 제대로 평가할 수 없음을 지적하고 있다. KFREB는 실제 사실이 아닌 허구 사실을 기반으로 하여, 모델이 사전학습되지 않은 내용으로 모델의 순수 검색 기반 답변 생성 능력을 평가하는 것을 목표로 한다.

[2] https://github.com/js-lee-AI/KFREB

3. Korean Fictional Retrieval Evaluation Benchmark (KFREB)

본 논문에서 제안하는 KFREB는 대규모 언어모델의 검색 기반 답변 생성 능력을 평가하기 위한 벤치마크이다. 이를 위해 실제 사실이 아닌 허구 사실을 기반으로 데이터셋을 구성한다. 이는 모델이 사전학습에서 보았던 사실을 반영하여 생성하는 답변이 검색 기반 답변 시스템에서의 실제 능력을 제대로 평가할 수 없음을 보완하기 위한 것이다.

대규모 언어모델의 여러 케이스의 검색 생성 성능을 확인하기 위해, 장문 문서, 두 개의 정답을 포함한 골드 문서, 한 개의 골드 문서와 유사 방해 문서 (키워드 포함), 한 개의 골드 문서와 유사 방해 문서 (키워드 미포함), 상호참조 멀티홉 리즈닝의 다섯 가지 케이스로 KFREB를 구성하였다.

구체적으로, KFREB 데이터셋은 나무위키 데이터셋[3]을 바탕으로 1,500개의 문서를 무작위로 선정하였다. 이후, GPT-4를 이용하여 데이터 생성 과정을 네 단계로 진행하였다. 첫 번째 단계에서는 GPT-4를 이용하여 각 문서를 허구의 문서로 교체하였다. 두 번째 단계에서는 GPT-4를 이용하여 각각의 케이스에 맞도록 텍스트를 변환하였다. 세 번째 단계에서는 변환된 텍스트를 바탕으로 GPT-4를 이용하여 질문과 정답을 포함한 multiple choice 후보군을 생성하였다. 마지막 단계에서는 GPT-4를 통해 정답을 추론하여 이전에 생성한 GPT-4의 정답과 일치하는지 검수하였고, 일치하지 않을 경우 해당 데이터를 제외하였다.

이 과정을 통해 총 96개의 데이터를 필터링하였으며, 필터링된 데이터 중 인간 검토 및 수정을 통해 각 케이스별 100개의 데이터만 선정하고 케이스에 맞도록 추가 수정하였다.

3.1 장문 문서

장문 문서 케이스는 한 개의 장문 (800글자 이상)의 골드 문서가 주어진 경우를 대상으로 한다. 이 케이스는 언어 모델이 장문의 문서에서 필요한 정보를 효과적으로 추출하고 이해하는 능력을 평가하는 것을 목표로 한다. 장문문서 케이스의 생성된 데이터 예시는 표 1에 나타나있다.

3.2 정답을 포함한 두 골드 문서

정답을 포함한 두 골드 문서 케이스는 각각이 정답을 포함한 두 개의 골드 문서가 주어진 경우를 대상으로 한다. 이 케이스는 검색 모델이 답변을 생성하는 것에 적절한 여러 정보를 검색한 상황에서 해당 골드 문서 두 개 중 한 개 이상의 문서를 적절히 참조하여 통해 언어 모델이 정답을 반영할 수 있는지를 확인하기 위함이다.

[3] https://huggingface.co/datasets/heegyu/namuwiki

표 1. KFREB의 장문 문서, 골드 및 유사 방해 (키워드 포함) 케이스 데이터 예시

카테고리	문서	질문	정답
장문 문서	6300년(렬즈시대 노턴 왕제 49년)부터 6428년(노턴 왕제 76년)까지 약 9번의 베조르 제국 베조르의 페르키아 학살로 인해 시작된 전쟁을 말한다. '대베수항전', '베조르 저항전쟁'등 다양한 명칭으로 ... 베조르 제국과의 전쟁에서 그나마 자리, 정치, 문화적으로 기존의 세력이 보존된 곳을 찾는다면, 서부 대륙 일대와 아프리카, 동남아, 미나리아, 카로에덴 정도를 언급할 수 있겠다. ... 하지만, 이들 역시 베트레아는 서부 대륙이나 아프리카는 사실 너무 멀어서 베조르가 전력을 다한 곳이라고 보기는 어려웠다. 카로에덴 역시 베조르의 침략을 어렵게 막았지만, 이후 베조르 제국의 후손을 주장한 디모리아 제국의 후예인 무그나 제국에게 정복당하고 ...	Q: 베조르 제국의 침략으로 인해 피해를 입은 국가 중, 베조르의 침략을 어렵게 막았지만 이후에 무그나 제국에게 정복당한 국가는? 1) 페르키아 2) 인라쿠르 3) 카로에덴 4) 미나리아	3)
골드 및 유사 방해 (키워드)	**골드 문서 1** # 흑철쌀 푸레이크 조선시대를 살아가던 만인재는 건강에 관심이 많아 ... 그는 특히 아침에 먹는 식사에 집중하여, 특별한 레시피를 만들었다. 이것이 바로 "흑철쌀푸레이크"이다. ... 이 다크올은 아침을 출발시키는 힘찬 울음소리와 흑철쌀푸레이크를 상징하며, ... **유사 방해 문서** # 흑철쌀 도시락 고구려 시대의 농부였던 백초백은 흑철쌀의 효능에 대해 깊이 연구하였다. ...	Q: 만인재가 아침 식사로 선호했던 식품은? 1) 흑철쌀 도시락 2) 흑철쌀푸레이크 3) 채소와 해산물, 고기 4) 다크올	2)

3.3 키워드 포함 골드 문서 및 유사 방해 문서

키워드가 포함된 한 개의 골드 문서와 유사 방해 문서 케이스는 한 개의 골드 문서와 검색 시스템에서 검색됐을 것 같은 유사 방해 문서가 주어진 경우를 대상으로 한다. 이 케이스는 검색 모델이 정답 문서와 그에 해당하는 후보 문서를 검색했을 때, 그리고 두 문서 모두 키워드가 포함되어 있을 때를 가정으로 한다. 가령, 위키피디아[4]나 나무위키[5] 등의 문서가 여기에 해당한다. 여기에서, 후보 문서는 검색 모델의 특성 상 유사한 문서가 검색되기 때문에 실제 정답 유추에 방해가 되는 문서로 처리한다. 이러한 유사 방해 문서는 실제 검색 기반 대규모 언어모델 서비스에서의 검색 성능을 정확히 평가하는 것과 더불어 언어모델에게 혼란을 주어 정확한 추론을 할 수 있는지를 평가할 수 있다. 키워드는 생성된 질문의 키워드가 문서에 포함되어 있는 경우로, 모델에게 정확한 근거 힌트를 주는 역할을 한다.

3.4 키워드 미포함 골드 문서 및 유사 방해 문서

키워드가 미포함된 한 개의 골드 문서와 유사 방해 문서 케이스는 키워드가 포함된 한 개의 골드 문서와 유사 방해 문서 케이스와 마찬가지로, 한 개의 골드 문서와 검색 시스템에서 검색됐을 것 같은 유사 방해 문서가 주어진 경우를 대상으로 한다. 가령, 논문이나 사내의 질의응답 문서가 여기에 해당한다. 하지만, 차이점은 문서 내에 모델이 명확한 힌트를 얻을 수 있는 키워드가 존재하지 않아 모델에게 직접적인 힌트가 주어지지 않는 경우로 한다. 즉, 이 케이스는 언어 모델이 키워드를 포함하지 않은 골드 문서와 방해 문서에서 필요한 정보 만을 효과적으로 추출하고 이해하는 능력을 평가하는 것을 목표로 한다. 키워드 미포함 골드 문서 및 유사 방해 문서 케이스의 생성된 데이터 예시는 표 1에 나타나있다.

두 케이스는 키워드가 포함되거나 키워드 포함되지 않는 문서의 특성에 따라 대규모 언어모델의 성능을 평가하여 실제 서비스에서 생성 능력을 평가하여 실제 서비스에서 문서 가공 혹은 문서 특성에 따른 대규모 언어모델 선택 등을 제공하기 위함이다.

3.5 상호참조 멀티홉 리즈닝

상호참조 멀티홉 리즈닝 케이스는 문서를 상호참조하는 경우의 멀티홉 리즈닝을 요구하는 두 개의 골드 문서가 주어진 경우를 대상으로 한다. 이 케이스는 언어 모델이 여러 문서간의 관계를 이해하고, 그 관계를 바탕으로 필요한 정보를 효과적으로 추출하는 능력을 평가하는 것을 목표로 한다. 상호참조 멀티홉 리즈닝 케이스의 생성된 데이터 예시는 표 2에 나타나있다.

[4]https://en.wikipedia.org/
[5]https://namu.wiki/

표 2. KFREB의 상호참조 멀티홉 리즈닝 케이스 데이터 예시

카테고리	문서	질문	정답
상호 참조	**문서1** 1976년 6월에 푸른하늘라디오라는 방송으로 시작되었다. … 아날로그 케이블은 전송을 원칙적으로 하지 않는다. … **문서2** 2010년 7월 12일에 푸른하늘라디오는 디지털 방송 채널로 발전하였다. … 더욱 다양한 방법으로 접근하게 되었다. 이렇게 노력한 결과로 푸른하늘라디오는 교육 방송의 새로운 모델로 자리 잡게 되었다.	Q: 푸른하늘라디오 방송이 교육 방송의 새로운 모델로 자리 잡게 된 이유는? 1) 일반적인 지상파나 DMB로 접근이 가능했기 때문 2) 푸른하늘라디오가 아날로그 케이블을 통해 전송되었기 때문 3) 푸른하늘라디오가 다양한 방법으로 접근 가능했던 것이 큰 역할을 함 4) 푸른하늘라디오의 신기술인 브라이트라이프가 실패했기 때문	3)

4. 결론 및 한계

본 논문에서는 대규모 언어모델의 검색 기반 답변 생성 능력을 평가하기 위한 벤치마크, KFREB(Korean Fictional Retrieval Evaluation Benchmark)를 제안하였다. KFREB는 실제 사실이 아닌 허구 사실을 기반으로 하여, 모델이 사전학습되지 않은 내용으로 모델의 순수 검색 기반 답변 생성 능력을 평가하는 것을 목표로 한다. 이 벤치마크는 다섯 가지 케이스를 종합적으로 판단하도록 구성되어 있다. 이를 통해 대규모 언어모델이 검색 시스템과 결합되었을 때, 모델의 내제된 정보를 활용하지 않고 주어진 문서 정보에서 적절한 정보를 잘 찾아와서 답변하는지에 대한 평가가 가능하다. 본 논문은 KFREB를 통해 적절한 서비스 목적에 따른 대규모 언어모델 선택에 대한 인사이트를 제공하고자 한다. 이 벤치마크를 통해 대규모 언어모델의 검색 기반 답변 생성 능력을 보다 정확하게 평가하고, 목적에 맞게 사용할 수 있으며, 이를 바탕으로 최신 대규모 언어모델의 개선 방향을 제시할 수 있을 것으로 기대된다.

하지만, KFREB에 대한 연구는 명확한 한계점을 가지고 있다. 첫 째로, KFREB는 허구 기반으로 작성되어 대규모 언어모델이 겪는 환각 지식을 강화할 수 있으며, 둘 째로 대규모 언어모델에 따른 여러 프롬프트에 따라 KFREB의 성능이 달라질 수 있다. 셋 째로, 용이한 평가를 위해 multiple choiece로 제작이 되어 모델의 정확한 생성 능력을 평가하지는 못한다. 마지막으로, 인간이 최종 데이터를 검수하고 수정하였음에도 불구하고, 해당 허구 데이터는 GPT-4의 내재적 지식을 통해 생성된 것이다. 이는 GPT-4와 GPT-3.5에게 유리하게 작용할 수 있다. 우리는 이러한 점들을 극복할 수 있도록 KFREB를 확장하고 다양한 프롬프트에 대해 한국어 대규모 언어모델의 성능을 비교 할 것이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2022-2018-0-01405). 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발).

참고문헌

[1] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.

[2] N. M. S. Surameery and M. Y. Shakor, "Use chat gpt to solve programming bugs," *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, Vol. 3, No. 01, pp. 17–22, 2023.

[3] R. W. McGee, "Using chatgpt to conduct literature searches: A case study," *Journal of Business Ethics*, Vol. 95, No. 2, pp. 165–178, 2023.

[4] M. Liebrenz, R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, "Generating scholarly content with chatgpt: ethical challenges for medical publishing," *The Lancet Digital Health*, Vol. 5, No. 3, pp. e105–e106, 2023.

[5] S. R. Ali, T. D. Dobbs, H. A. Hutchings, and I. S. Whitaker, "Using chatgpt to write patient clinic letters," *The Lancet Digital Health*, Vol. 5, No. 4, pp. e179–e181, 2023.

[6] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, Vol. 13, No. 4, p. 410, 2023.

[7] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding chatgpt: A taxonomy of existing research, current challenges, and possible future directions," *Journal of King Saud University-Computer and Information Sciences*, p. 101675, 2023.

[8] S. Liu, J. Wang, Y. Yang, C. Wang, L. Liu, H. Guo, and C. Xiao, "Chatgpt-powered conversational drug editing using retrieval and domain feedback," *arXiv preprint arXiv:2305.18090*, 2023.

[9] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better llm-based evaluators through multi-agent debate," *arXiv preprint arXiv:2308.07201*, 2023.

[10] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie *et al.*, "Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization," *arXiv preprint arXiv:2306.05087*, 2023.

[11] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati, "Large language models still can't plan (a benchmark for llms on planning and reasoning about change)," *arXiv preprint arXiv:2206.10498*, 2022.