과학기술정보 전문분야 한국어 벤치마크

이민지[a], 육정훈[a], 이병호[a], 임경태[b], 이경하[c], 박천음[a*]
국립한밭대학교 컴퓨터공학과[a]
서울과학기술대학교 인공지능학과[b]
한국과학기술정보연구원[c]
{20231418, 20191780, 20201779}@edu.hanbat.ac.kr, ktlim@seoultech.ac.kr, kyongha@kisti.re.kr
parkce@hanbat.ac.kr

Korean Benchmark for Science and Technology Information in Specialized Fields

Minji Lee[a], Jeonghun Yuk[a], Byeongho Lee[a], Kyungtae Lim[b], Kyongha Lee[c], Cheoneum Park[a*]
Dept. of Computer Engineering, Hanbat National University
Dept. of Applied Artificial Intelligence, Seoul National University of Science and Technology
Korea Institute of Science and Technology Information

요 약

본 논문에서는 거대 언어 모델의 과학기술정보 분야의 전문 지식을 평가하기 위해 한국어 벤치마크를 구축한다. 범용 분야 벤치마크와 차별화된 다양한 과학기술 도메인 특화 데이터 구축을 위하여 과학기술정보 전문 분야 다섯 가지 카테고리를 정의하고 각 문제별 난이도와 평가 기준을 수립한다. Synthetic 데이터 자동 생성 파이프라인을 2단계로 구성하여 벤치마크 전반의 데이터 품질 및 전문성을 향상시키며, 구축 시간을 단축하는 방법을 제안한다. 공개된 거대언어모델을 활용하여 성능 평가를 진행하여 구축한 벤치마크가 과학기술정보 전문 분야 평가 데이터의 유효성을 검증하였다. 특히, 본 벤치마크는 과학기술정보 분야의 전문 지식을 평가할 수 있는 한국어 최초의 벤치마크로서 도메인 특화 성능 평가 연구의 중요한 토대를 마련할 수 있을 것이다.

주제어: 전문분야 벤치마크, 과학기술 도메인, 거대언어모델 벤치마크, 한국어 평가 데이터

1. 서 론

거대 언어 모델(LLM)의 등장과 이를 활용한 연구들은 다양한 자연어처리 태스크에서 성능 향상을 이끌어내며, 기존 문제들을 효과적으로 해결하는 데 기여한다 [1, 2, 3, 4]. 거대 언어 모델은 다국어 범용 및 웹 데이터를 기반으로 학습되기 때문에 언어 간 일반화, 상호 작용, 코드 스위칭이 가능하며, 범용적인 도메인에 대한 응답을 유연하게 처리할 수 있다. 이런 거대 언어 모델의 언어 이해와 생성, 추론 능력을 평가하기 위해, 문제 유형, 도메인, 언어와 같은 다양한 기준으로 구성된 벤치마크(Benchmark) 테스트가 활발히 연구되고 있다 [5, 6, 7, 8, 9]. 벤치마크 테스트는 언어 모델의 실제 성능을 객관적으로 비교하고, 개선 방향을 도출하는 데 중요한 역할을 한다. 주로 일반 분야에 대한 벤치마크가 연구되었으며, MT-Bench [8], MMLU [9], C-EVAL [10], Xeizhi [11], KMMLU [12] 등이 있다. 전문 분야에서는 고유한 문제 해결 방식과 지식을 필요로 하며, 이를 검증할 벤치마크의 중요성이 점점 더 강조되고 있다.

범용 분야 벤치마크는 검증된 데이터인 공인 시험으로부터 문제를 선별해서 평가 데이터에 포함한다. 이때 영어 벤치마크는 주로 STEM I과 II를 사용하고, 중국어 벤치마크는 글쓰기, 과학, 수학, 공학, 법률 등의 도메인 문제를 포함한다. 특정 도메인에 대한 전문 지식을 알아야 풀 수 있는 문제는 거대 언어 모델의 범용적 지식 추론 능력뿐만 아니라 도메인 특화 추론 능력을 평가할 수 있다. 이러한 벤치마크는 도메인 특화 데이터를 포함하지만, 데이터의 출처는 STEM이나 공인 시험에 국한된다. 범용 분야 벤치마크는 특정 도메인에 대한 전문 지식을 충분히 다루지 못하며, 거대 언어 모델이 도메인 특화 문제 해결 능력을 정확하게 평가하는 데 한계를 드러낸다. 이에 따라, 특정 도메인에 대한 해결 능력을 평가하기 위하여 MultiMedQA [13], ChemLLMBench [14], MATH [15] 등과 같은 과학 기술 분야 데이터 연구가 주로 진행되었다. 본 논문에서는 도메인의 범위를 더욱 확장하여 일반적인 과학기술정보 분야 벤치마크를 구축하며, 도메인 특화 언어 추론 능력을 평가한다.

본 논문에서는 과학기술정보 전문분야 한국어 평가셋을 구축하며, 일반적인 과학기술 범위에서 전문성을 겸비하기 위하여 과학기술정보 전문분야 다섯가지 카테고리를 정의하고 카테고리 별 문제의 난이도를 선별한다. 과학기술정보 전문분야 벤치마크는 언어 모델의 다양한 추론 능력을 평가하기 위해, 다섯가지 카테고리에 포함된 논문에서 문제를 구축한다. 문제는 단답형, 사지선다 형식의 객관식, 다중선택 문제, 참/거짓과 같이 네가지 유형으로 나누고, 각 유형별로 난이도를 부여하여 생성한다. 본 논문에서 구축한 과학기술정보 전문분야 벤치마크에 대한 유효성을 판단하기 위해 공개된 거대 언어 모델들을

*교신 저자

사용하여 성능 비교를 수행한다. 실험 결과, GPT-4o 모델을 기준으로, 객관식에서 0.6421, 다중선택 문제에서 0.4048, 참/거짓 문제에서 0.61의 점수를 보였으며, 구축한 벤치마크가 과학기술정보 전문분야 평가 데이터로 유의미함을 알 수 있다.

본 논문의 컨트리뷰션은 다음과 같다:

- 과학기술정보 전문분야 한국어 벤치마크 최초 구축
- 벤치마크의 전문성을 강화하기 위하여 다섯개의 카테고리와 각 카테고리 별 문제 난이도 정의
- 공개된 거대 언어 모델을 활용하여 성능 평가 및 벤치마크 유효성 검증

2. 관련 연구

영어와 중국어 관련된 벤치마크 테스트 데이터가 많이 공개되었으며, 한국어 또한 범용적인 평가 벤치마크와와 문화적 특성을 분석하는 벤치마크가 연구되었다.

영어 벤치마크 대표적으로 MT-Bench [8]와 MMLU [9]가 있으며, 각각 8개 태스크에 대한 multi-turn 문제와 57가지 태스크에 대한 객관식 문제로 구성된다. 두 벤치마크 데이터는 범용적으로 평가되는 데이터를 기반으로 한다는 점이 특징이다.

중국어 벤치마크 C-EVAL [10]과 Xeizhi [11]가 있으며, 먼저 C-EVAL은 거대언어모델의 중국어 관련 문제 이해 능력과 추론 능력을 포괄적으로 평가하기 위한 위한 데이터셋이다. C-EVAL은 나라의 고유한 특성을 반영하기 위해 중국 문화, 역사, 법률 도메인을 포함한 52개의 태스크로 구성되며, 중국의 모의고사나 지역 시험 등에서 객관식 문제를 수집하여 정제와 검증을 수행한다. Xeizhi는 중국어 기반 의학, 문학, 경제 등 516개의 다양한 태스크와 50개의 보기 중 3개의 정답을 포함하는 객관식 문제를 구성하는 특징이 있다. 이와 같은 객관식 문제를 평가하기 위해 ranking score를 계산하는 Mean Reciprocal Rank (MRR)을 평가 지표로 사용한다.

한국어 벤치마크 국내에는 KMMLU [12], KLUE [7], CLIcK [16] 등이 있다. KMMLU는 MMLU 벤치마크 데이터셋을 한국 문화와 언어적 특성을 반영하여 변형한 벤치마크이다. STEM, 인문학, 기술, 공학 등 다양한 도메인으로 구성되며, 45개 주제를 포함한 총 35,030개의 다중 선택형 문항을 포함한다. KLUE는 한국어 이해 능력을 평가하기 위해 여덟 개의 자연어처리 태스크로 구성된 벤치마크이다. 태스크는 토픽 분류, 의미 유사도, 자연어 추론, 개체명 인식, 관계 추출, 의존 구문 분석, 기계 독해, 대화 상태 추적로 구성된다. CLIcK은 한국의 문화와 언어적 이해 능력을 평가하기 위한 벤치마크로 11가지 카테고리로 분류된 1,995개의 객관식 문항을 포함한다. 특히, 한국의 문화/언어적 특성을 반영하기 위해 CSAT와 같은 한국의 여러 공식 시험 및 한국어 통합 프로그램 교과서에서 데이터를 수집하였다.

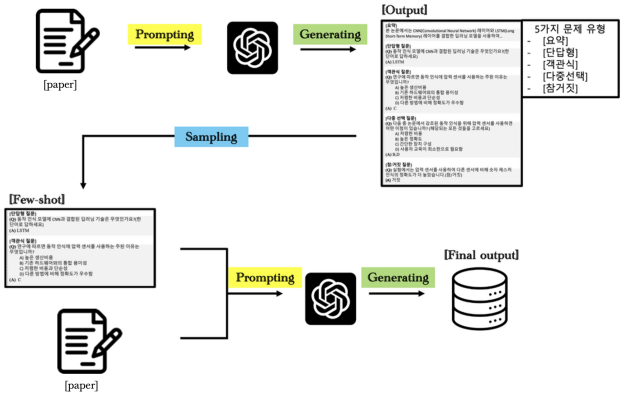


그림 1. Sythetic 데이터 자동 생성 파이프라인

표 1. 논문 주제 카테고리 및 문항 속성 예제

논문 주제 카테고리	문항 속성 대표 예제
경제/경영	추세의 주요 요인 분석,경제 모델링과 예측
화학공학/생화학	생화학 반응의 메커니즘, 데이터 분석 기법, 샘플 준비 방법
공학	시스템 메커니즘, 기술 시스템 성능 평가, 기술 및 연구의 한계점
생물학/지구과학	모델 및 생물의 특징, 샘플 분석 방법
물리학/수학(자연과학)	산술, 이론 및 방법론, 검증 방법

3. 과학기술정보 특화 벤치마크

본 논문에서는 거대언어모델의 전문 분야 언어 이해, 생성, 추론 능력을 평가할 수 있는 과학기술정보 특화 벤치마크를 구축한다. [17, 18, 19]에서는 GPT-4o를 사용하여 synthetic 데이터를 자동으로 생성하고, 이를 기반으로 학습하는 방법이 유의미함을 보인다. [18]에 따라 GPT-4o로부터 생성한 synthetic 데이터가 유효한 품질임을 가정하여, 본 논문에서는 벤치마크 문제 자동 생성 파이프라인(sampled few-shot-based generation pipeline)을 구축한다.

3.1 벤치마크 디자인

문제 유형 설계 제안한 벤치마크의 문제 유형은 네가지 종류로, 단답형, 객관식, 다중선택, 참/거짓 문제로 구성한다. 이는 문제의 다양성, 일반성, 검증 용이성을 함께 고려한다. 본 논문에서는 거대언어모델이 이미 학습한 지식으로 과학기술정보 도메인에 대하여 올바른 평가가 가능한지 확인하고자 하며, 이에 따라 문제 해결에 단서가 되는 문서나 문단 정보는 제공하지 않는다. 네가지 유형의 문제에서 단답형은 질문을 이해하고 명확한 정답을 생성해야 하며, 다양한 질의응답 태스크에서 일반적으로 많이 사용된다 [20]. 객관식은 MMLU, Xiezhi, MT-bench 등에 주로 구성된 문제 유형으로, 보기를 기반으로 언어모델의 판단력을 평가한다. 다중 선택은 가능한 모든 답을 찾아야하기 때문에 더욱 분별력있는 문제 유형이다. 마지막으로 참/거짓

문제는 Trivia QA [21], BoolQ [22], TruthfulQA [23] 등에서 사용된 유형으로, 본 논문에서는 주어진 질문에 대하여 모델로부터 이진 결과를 추론한 것으로 사실적 정보 추출 능력을 평가한다. 이와 같은 문제 유형은 일반분야 벤치마크와 유사한 기준으로 비교할 수 있기 때문에 전문지식 벤치마크에 대하여 언어모델의 문제 풀이 능력을 확인하고 벤치마크의 난이도를 추정할 수 있다.

전문성 향상을 위한 문제 설계 본 논문에서는 과학기술정보 도메인에 대한 전문성을 향상시키기 위해 [표 1]와 같이 다섯가지 논문 주제 카테고리를 정의하고, 각 카테고리 별로 문제의 종류와 평가 목표에 맞는 문항 속성을 설계한다. 이때 논문 주제 선정 기준은 다음과 같다. 1) KCI 학술인용정보를 기준으로 피인용지수(5년)가 가장 높은 학회지를 선별하고, 2) 피인용지수를 기준으로 상위 10위에 해당하는 학회지에서 다섯개의 카테고리를 추출한다. 과학기술정보 도메인 특화 벤치마크는 거대언어모델이 학습된 지식을 활용하여 전문분야 문제를 추론하고, 해결할 수 있는지 평가하는 것을 목표로 한다. 이와 같은 이유로 벤치마크 문제는 closed-book QA (question answering)와 같이 구성하며, 각 논문 주제마다 세부적인 특징을 나눠 문제를 설계한다. 사용된 데이터는 논문 문헌이기 때문에 대부분의 문항 속성은 연구 방법, 연구 결과 비교, 실험 설계, 성능 평가, 결과 검증 등으로 구성하며, 각 논문 주제 별로 대표하는 속성 예시는 [표 3]와 같다. 이런 문제 설계 예제는 GPT-4o를 활용하여 synthetic data를 만들때 함께 입력하여 주어진 논문 주제에 적합한 예시를 선별하여 문제의 정확도와 난이도를 향상시킬 수 있도록 한다.

3.2 벤치마크 구축 과정

데이터 자동 구축 자동 생성 파이프라인은 [그림 1]과 같이 2단계로 구성되며 각 단계의 과정은 다음과 같다. 1) 주어진 과학기술 논문 데이터로부터 GPT-4o를 사용하여 사전에 정의한 네가지 문제 유형에 따라 1차 synthetic 데이터를 생성한다. 2) 생성된 1차 데이터를 임의로 샘플링하여 문제 생성 prompt에 함께 입력하고, 마찬가지로 GPT-4o를 사용하여 최종 synthetic 데이터를 생성한다. 이와 같은 과정으로 생성된 synthetic 데이터는 사람이 최종 검수하여 평가 문제의 품질을 향상시킨다.

데이터 검수 데이터 검수 작업은 2명의 검수자가 다음 규칙에 따라 진행한다. 추가로 검수 과정에서 검수자 간 평가 일치도를 측정하기 위해 카파(Kappa) 계수를 활용한다. 그 결과, 0.685로 나타났으며, 이는 높은 수준의 일치도를 의미한다. 검수 과정에서 확인된 주요 이슈는 단답형 문항의 경우, 1:N의 관계를 형성하여 답안이 다수가 되는 문제가 발생하였다. 이에 대한 방안으로 관형어(구)를 추가하여 답안이 하나로 특정될 수 있도록 한다. 검수 작업을 거친 2단계 파이프라인으로 구축된 데이터는 1차 데이터에 비해 더 세부적 지식을 요하는 문항으로 구성되는 것을 확인할 수 있다.

[Prompt]

주어진 {document}와 {few-shot}의 내용을 바탕으로 평가셋을 만들기 위해,
우리는 다음과 같은 네가지 유형의 질문과 답변을 만들어야 합니다.
:<단답형 질문, 객관식 질문, 다중선택 질문, 참/거짓 질문>

1.<Attribute> 중 4가지를 선택하여, 각 주제에 해당되는 질문을 생성하세요.
질문의 주제는 중복을 허용하지 않습니다.
2. <Attribute> 에 대한 기준 및 설명은 <criteria>입니다.
내용을 참고해서 문제 생성시 반영하세요.
3. <few-shot>의 내용을 반영하되, 동일한 문제를 만들지 않도록 합니다.

그림 2. 프롬프트 예시

- 해당 문항과 답안이 올바르게 매칭되도록 한다.
- 단답형(short-answer)의 경우, 문항과 답안은 1:1로 매칭되도록 한다.
- 모든 문항의 중심 소재는 반드시 원문 데이터에 존재해야 한다.

Synthetic 데이터 프롬프트 [그림 2]는 synthetic 데이터 프롬프트이다. 각 문항 속성 정의 및 기준을 criteria로 정의하여 1차 데이터로부터 샘플링된 데이터인 few-shot과 함께 프롬프팅한다. 1차 데이터 생성을 위한 프롬프트는 few-shot을 제거하여 시행한다.

표 2. 데이터 통계

논문 주제 카테고리	데이터 수
경제/경영	28
화학공학/생화학	80
공학	100
생물학/지구과학	100
물리학/수학(자연과학)	184

카테고리별 데이터 구축 통계 [표 2]는 생성된 데이터의 논문 주제 카테고리 별 데이터 통계를 나타낸다. 각 카테고리에는 4가지 유형(단답형, 객관식, 다중선택, 참/거짓)은 균등하게 분포한다. 본 벤치마크의 도메인 특성을 반영하고자, 과학 및 기술 도메인의 데이터의 비중을 상대적으로 높인다.

4. 실험 및 결과

이번 장에서는 제안한 벤치마크 데이터의 유효성을 파악하기 위해 공개된 한국어-영어 이중언어 모델들의 성능평가를 진행하고자 한다. 이러한 실험은 벤치마크의 변별력을 증명할 수 있

표 3. 각 모델과 문제 유형별 실험결과

문제유형(threshold)	Bllossom	EXAONE	GPT4
다중 선택(all)	0.2857	0.3334	0.4048
다중 선택(1)	0.5595	0.4524	0.4524
다중 선택(2)	0.4881	0.4048	0.4524
다중 선택(3)	0.369	0.3452	0.4048
객관식	0.6211	0.6105	0.6421
참/거짓	0.63	0.6	0.61

표 4. 각 모델의 단답형 문제 유형 실험결과

성능 평가 방법	Bllossom	EXAONE	GPT4
Exact Matching	0.0532	0.1064	0.1596
BERT Score	0.5495	0.5795	0.6076

는 동시에 학습 데이터와 대상이 다른 여러 모델의 비교평가를 토대로 과학기술 분야 이해에 적합한 언어모델을 조사하고자 한다.

4.1 실험조건 및 평가지표

본 연구에서는 가장 최근에 발표된 3개의 서로 다른 한국어 공개 모델과, 상업 모델인 GPT-4o를 활용해 해당 벤치마크의 성능을 측정했다. 제안하는 벤치마크는 다중 선택 문제나나 객관식 문제, 그리고 참/거짓의 경우 전체 문제 대비 정답 개수를 토대로 정확도를 산정하는 방식으로 성능평가를 진행했다. 다만 단답형 질의응답의 경우 가능한 변수가 많기 때문에, 정답과 정확히 일치할 경우만 인정하는 Exact Matching과 BERT 모델 임베딩을 이용한 의미론적 유사성을 나타내는 BERT Score 토대로 점수를 산정했다. 평가 대상 모델은 다음과 같다:

- Bllossom [4]: LLama3-8B에 대량의 한국어-영어 데이터로 Continual Training을 진행한 모델
- EXAONE [24]: Scratch 부터 한국어-영어 이중 언어만 집중 학습한 모델
- GPT-4o: OpenAI의 상업용 모델

4.2 실험결과 및 분석

Overall [표 3]는 제안한 벤치마크 데이터셋을 토대로 각 모델의 성능평가 결과이다. 결과적으로는 GPT-4o의 성능이 가장 높았다. 다만 흥미로운 점은 규모가 훨씬 작은 한국어 특화 모델의 성능이 거대 모델인 GPT-4o와 크게 차이나지 않으며, 특히 참거짓 문제 유형에서 Bllossom 모델이 3% 정도 우수한 모습을 보였다. 이러한 결과는 기본적으로 일반적인 상식 수준의 문제의 경우 GPT-4o와 소형 한국어 특화 언어모델과의 편차가 크지 않다는 것을 의미한다. 다만 (1) 전문적인 지식을 요하는 문제이거나 (2) 문제 자체가 모호하여 추가 정보가 더 있어야 해결이 가능한 경우, 대규모 파라미터를 가진 GPT-4o 모델이 다른 모델에 비해 약간 우세한 모습을 보인다. 이를 통해 지식의 깊이와 추론 능력은 미묘하지만 GPT-4o가 더 우세하다고 판단할 수 있다.

다중 선택, 객관식 [표 3]의 다중 선택 문제 유형은 정답으로 인정하는 범위에 따라 네 가지의 성능 평가를 진행하였다. 다중 선택(n)은 적어도 n개만 맞춰도 정답으로 인정된다는 의미이며, 다중 선택(all)은 모든 정답을 맞추어야 정답으로 인정된다는 의미이다. 객관식 문제 유형은 네 가지 선택지 중 한 가지만이 정답인 문제 유형을 의미한다. 두 문제 유형에서의 전체적인 성능은 GPT-4o가 가장 좋은 모습을 보였다. 다만 흥미로운 점은, 다중 선택(1), (2)에서는 Bllossom이 더 좋은 성능을 보인다는 것이다. 이는 단순히 선택지 내에서 올바른 정답을 찾는 작업에 대해서는 한국어 언어모델이 GPT-4o에 비해 유리하다는 것을 의미한다. 이에 비해 다중 선택(all)의 경우에는 정답에 해당하는 모든 선택지를 찾아내야 하는 정밀한 작업이다. 이 때문에 대규모 파라미터에 의해 심도 깊고 정밀한 추론이 가능한 GPT-4o 모델이 유리함을 확인할 수 있다.

참/거짓 [표 3]의 진위형 문제의 결과는 세 모델이 비슷한 성능을 보였으며, 그 중 Bllossom 모델이 0.63으로 가장 높은 결과를 보였다. 진위형 문제는 타 문제 유형에 비해 정답의 후보가 적어 간단한 문제 유형에 속한다. 이 때문에 GPT-4o 모델의 정밀하고 심도 깊은 추론보다 한국어 언어모델의 언어의 이해를 바탕으로 한 추론이 더욱 유효했던 것으로 판단된다.

단답형 단답형 문제에 대한 실험 결과인 [표 4]에서, Exact Matching과 Bert Score 모두 GPT-4o가 가장 우세하였다. 이는 답변 형태의 범위가 매우 방대하여 보다 정밀한 추론 능력이 필요한 단답형 문제 유형에서, 대규모 파라미터 모델인 GPT-4o 모델이 정확한 답변 추론에 유리했기 때문이다. 다만 한국어 언어모델 또한 Exact Matching 결과에 비해 BERT Score가 매우 높게 측정되어, 의미적으로 정답에 가까운 추론을 하였음을 확인할 수 있다.

5. 결 론

본 논문에서는 과학기술정보 전문분야 한국어 벤치마크를 구축하였으며, 데이터 셋의 전문성능 향상시키기 위하여 과학기술정보에 대한 문제 설계를 수행하였다. Synthetic 데이터 자동생성 파이프라인을 2단계로 구성하여 보다 정확한 데이터를 자동으로 생성하였으며, 사람이 최종 검수를 수행하여 전문성을 주입하고 데이터 품질을 향상시켰다. 공개된 거대언어모델

(일반 분야 instruction 학습만 수행)을 활용하여 구축한 벤치마크를 평가했으며, 실험 결과 GPT-4o 기준으로 다중선택(all), 객관식, 참/거짓 문제에서 각각 0.4048, 0.6421, 0.61의 성능을 확인했다. 이와 같은 결과는 일반 분야 벤치마크에서 높은 성능을 보인 모델들이 전문 분야 벤치마크에서 낮은 성능을 보인 것으로, 본 논문에서 제안한 방법으로 구축한 전문 분야 벤치마크가 유효하다는 것을 알 수 있다.

향후 연구로는 과학기술정보에 포함되는 논문 데이터의 주제 카테고리를 확장하여 더욱 다양한 평가 데이터를 구축할 것이며, 한국어 뿐만 아니라 영어 데이터로도 확장하여 과학기술정보 전문분야 벤치마크의 유효성을 확인하고 도메인 특화 벤치마크의 필요성을 검증할 예정이다.

감사의 글

이 논문은 과학기술정보 특화 LLM, KONI(KISTI Open Natural Intelligence) 개발을 목표로 한 2024년도 한국과학기술정보연구원(KISTI)의 기본사업(과제번호: K24L4M2C5)으로 수행된 연구입니다.

참고문헌

[1] OpenAI, "Gpt-4 technical report," 2024.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[3] G. Team, "Gemma 2: Improving open language models at a practical size," 2024.

[4] Bllossom-Team, "Mlp-ktlim/llama-3-korean-bllossom-8b," 2024. [Online]. Available: https://huggingface.co/MLP-KTLim/llama-3-Korean-Bllossom-8B

[5] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," 2018.

[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," 2019.

[7] S. P. et al., "KLUE: Korean language understanding evaluation," *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=q-8h8-LZiUm

[8] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[9] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *International Conference on Learning Representations*, 2021.

[10] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, J. Lei, Y. Fu, M. Sun, and J. He, "C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models," *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23, 2024.

[11] Z. Gu, X. Zhu, H. Ye, L. Zhang, J. Wang, Y. Zhu, S. Jiang, Z. Xiong, Z. Li, W. Wu, Q. He, R. Xu, W. Huang, J. Liu, Z. Wang, S. Wang, W. Zheng, H. Feng, and Y. Xiao, "Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation," *AAAI*, pp. 18 099–18 107, 2024.

[12] G. Son, H. Lee, S. Kim, S. Kim, N. Muennighoff, T. Choi, C. Park, K. M. Yoo, and S. Biderman, "Kmmlu: Measuring massive multitask language understanding in korean," *CoRR*, Vol. abs/2402.11548, 2024.

[13] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, Vol. 620, No. 7972, pp. 172–180, 2023.

[14] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest, and X. Zhang, "What can large language models do in chemistry? a comprehensive benchmark on eight tasks," *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[15] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the MATH dataset," *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[16] E. Kim, J. Suk, P. Oh, H. Yoo, J. Thorne, and A. Oh, "Click: A benchmark dataset of cultural and linguistic

intelligence in korean," *Proceedings of the 2024 Joint International Conference on Computational Linguistics*, pp. 3335–3346, 2024.

[17] C. Park, J. Jiang, F. Wang, S. Paul, and J. Tang, "Llamaduo: Llmops pipeline for seamless migration from service llms to small-scale local llms," 2024. [Online]. Available: https://arxiv.org/abs/2408.13467

[18] C. Whitehouse, M. Choudhury, and A. F. Aji, "LLM-powered data augmentation for enhanced cross-lingual performance," *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[19] H. D. et al., "Auggpt: Leveraging chatgpt for text data augmentation," 2023. [Online]. Available: https://arxiv.org/abs/2302.13007

[20] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, I. Gurevych and Y. Miyao, Eds., pp. 784–789, Jul. 2018.

[21] M. Joshi, E. Choi, D. 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)*, R. Barzilay and M.-Y. Kan, Eds., pp. 1601–1611, Jul. 2017. [Online]. Available: https://aclanthology.org/P17-1147

[22] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *NAACL*, 2019.

[23] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., pp. 3214–3252, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.229

[24] LG-AI-Research, "Exaone 3.0 7.8b instruction tuned language model," 2024. [Online]. Available: https://arxiv.org/abs/2408.03541