

코드 스위칭 코퍼스 기반 다국어 LLM의 지식 전이 연구

김성현[○], 이강희, 정민수, 이정우

신한AI

Mrbananahuman.kim@gmail.com[○], dlrkdgm198@gmail.com, ms.jeong@shinhan.com, jungwoo.l2.rs@gmail.com

Knowledge Transfer in Multilingual LLMs Based on Code-Switching Corpora

Seonghyun Kim[○], Kanghee Lee, Minsu Jeong, Jungwoo Lee
Shinhan AI

요약

최근 등장한 Large Language Models (LLM)은 자연어 처리 분야에서 눈에 띄는 성과를 보여주었지만, 주로 영어 중심의 연구로 진행되어 그 한계를 가지고 있다. 본 연구는 사전 학습된 LLM의 언어별 지식 전이 가능성을 한국어를 중심으로 탐구하였다. 이를 위해 한국어와 영어로 구성된 코드 스위칭 코퍼스를 구축하였으며, 기본 모델인 LLAMA-2와 코드 스위칭 코퍼스를 추가 학습한 모델 간의 성능 비교를 수행하였다. 결과적으로, 제안하는 방법론으로 학습한 모델은 두 언어 간의 의미론적 정보가 효과적으로 전이되었으며, 두 언어 간의 지식 정보 연계가 가능했다. 이 연구는 다양한 언어와 문화를 반영하는 다국어 LLM 연구와, 소수 언어를 포함한 AI 기술의 확산 및 민주화에 기여할 수 있을 것으로 기대된다.

주제어: Large language model, 언어모델, 한국어 코퍼스, 한국어 언어 모델

1. 서론

최근 수년 간 인공지능 및 자연어 처리 분야에서 가장 두드러진 변화 중 하나는 Large Language Models (LLM)의 등장이다. LLM은 방대한 양의 텍스트 데이터를 학습함으로써 언어의 복잡한 패턴과 구조를 파악하게 되었으며, 이를 통해 사람처럼 자연스러운 텍스트를 생성하거나 이해하는 능력을 획득하였다[1,2].

이들은 단순한 텍스트 생성 뿐만 아니라 문장 간의 관계 파악, 정보 검색, 텍스트 분류 등 다양한 자연어 처리 작업에서 뛰어난 성능을 보이게 되었다. GPT[1,2], BERT[3], T5[4]와 같은 대표적인 LLM들은 방대한 양의 데이터와 고성능의 컴퓨팅 리소스를 통해 훈련되었으며, 이러한 높은 학습 비용에도 불구하고 그 성능의 효과 때문에 많은 관심과 자원이 투입되고 있다.

LLM의 등장은 기업, 교육, 의료 등 다양한 산업 분야에 활용 가능성을 제시하며, 우리의 일상에도 큰 영향을 미치게 되었다. 이러한 LLM의 활용 가능성과 그로 인한 사회적, 경제적 변화는 앞으로도 지속적인 연구와 논의의 주제가 될 것이다. 그러나 이러한 발전의 뒷면에는 중요한 문제점이 숨어 있다.

대부분의 LLM은 주요 언어, 특히 영어를 기반으로 연구 및 개발되고 있다. 이는 대규모의 데이터와 연구 자원이 주로 영어와 같은 주요 언어에 집중되어 있기 때문이다. 이렇게 되면 소수 언어권이나 개발이 덜된 언어들은 LLM의 혜택을 누리기가 어려워진다. 다양한 언어와 문화를 가진 세계에서 일부 언어만을 중심으로 기술 발전이 이루어진다면, 그것은 근본적으로 언어적 다양성과 국가 간의 디지털 격차를 심화시킬 위험이 있다.

LLM의 언어적 한계는 그저 기술적인 문제가 아니다. 언어는 특정 문화와 지역의 역사, 가치, 생각 방식을 반영하기 때문에 다양한 언어와 문화에 대한 폭넓은 연구와 이해 없이 LLM을 발전시키는 것은 그 자체로 큰 한계와 위험성을 내포하게 된다. 따라서, LLM의 발전과 활용에 있어서 다양한 언어와 문화를 포괄하는 방향으로의 연구 확장이 요구된다.

2. 다국어 LLM

최근에는 다양한 국가에서 그들만의 언어에 특화된 언어 모델을 개발하는 움직임이 활발하다. 최근 세계에서 초거대 AI 언어모델(LLM)을 보유한 나라는 미국·중국·한국·영국·이스라엘 등이다. 그럼에도 현재 많은 언어 모델, 특히 대규모 모델들은 영어를 중심으로 개발되고 있다. 영어 데이터의 풍부함은 영어 언어 모델의 발전을 주도하고 있다. 이러한 영어 언어 모델은 다른 언어들에 대한 모델 구축의 기본 토대로 활용되곤 한다. 대부분의 국가별 언어 모델은 기존의 영어 모델을 기반으로 한 재학습 과정을 거치게 된다. 이 재학습은 전이 학습(Transfer Learning)의 원칙에 근거하며, 기존의 영어 모델이 가진 대규모의 지식과 학습 능력을 활용하여 특정 언어나 문화에 맞게 미세 조정하는 과정이다. 특히, Meta에서 공개한 LLAMA[15]를 기반으로 만들어진 중국어[5], 일본어[6] 모델과 FlanT5 모델[7] 기반의 스페인어

특화 모델[8] 등이 있다. 한국어 역시 Llama-2-ko¹나 koalpaca²등, 공개된 영어 LLM 모델을 이용해 한국어 특화 모델을 만드려는 시도들이 이어지고 있다.

언어 변환에 관한 최근의 연구들은 다음과 같은 두 가지 주요 방법론을 중심으로 이루어져 왔다.

- (1) 사전 학습된 모델의 구조적 특징을 변화시키지 않으며, 특정 언어의 데이터를 통해 재학습 [5-6,8]
- (2) 사전 학습 모델의 임베딩 층을 재 정의하거나, 모델의 구조를 수정하여 특정 언어의 데이터를 재학습 [9-12]

임베딩 레이어를 직접 재학습하는 접근 방법은 기존 지식이 손실될 수 있는 문제점이 있었다. 또한, 이러한 연구는 BERT와 같은 임베딩 모델을 중심으로 진행되었으나 [13], LLM과 같은 대형 모델에서의 실험은 대체로 성공적이지 않았다[9]. 더불어 새로운 언어 변환 작업은 추가적인 언어 코퍼스를 필요로 하며, 그 양은 기존 소스 언어의 20~50% 정도로 매우 큰 부담이 되었다[11]. 또한, 다양한 언어 간의 관계성과 상호작용을 이해하는 데는 여전히 제한적이다

3. 코드 스위칭(Code-Switching) 코퍼스

코드 스위칭(Code-Switching)은 두 가지 이상의 언어나 방언을 사용하는 이들이 대화 중에 한 언어에서 다른 언어로 전환하는 현상을 의미한다[14]. 이 현상은 다양한 사회적, 문화적, 언어적 문맥에서 발생할 수 있으며, 대화자의 의도나 상황에 따라 다양한 방식으로 나타난다. 이러한 코드 스위칭은 언어학, 사회언어학, 심리언어학 분야에서 오랫동안 연구되어왔다. 특히, 이중 언어 사용자의 언어 전환 현상을 중심으로 다양한 연구가 진행되었다. 본 연구에서는 LLM의 언어 변환 과정에서 언어 간의 관계성과 사전 학습된 지식의 전이를 유도하기 위해 ‘코드 스위칭 코퍼스’를 통한 모델의 재학습 방법론을 제안하고자 한다.

코드 스위칭 코퍼스란, 두 가지 이상의 언어가 혼합된 데이터를 의미한다. 본질적으로, 이는 특정 언어의 용어나 구문을 다른 언어의 문맥 내에서 사용하여 두 언어 사이의 연결고리를 유도한다. 예를 들어, "트랜스포머 모델이 개발되었다"와 "transformer 모델이 develop되었다."와 같은 혼합된 표현은 한국어와 영어 간의 지식 전달 및 연계를 도울 수 있다.

이러한 코드 스위칭 코퍼스의 활용은 LLM에게 두 언어 사이의 의미적 관계와 구조적 특성을 자연스럽게 학습하게 한다. 결과적으로, LLM이 다양한 언어에 걸쳐, 보다 일관된 지식의 표현과 추론 능력을 갖추게 되는지 실험하고자 한다.

본 논문에서는 코드 스위칭 코퍼스 구축을 위해 GPT-4 API와 한국어 위키피디아 문서를 활용했다. 한국어 위키피디아 원본 문서를 GPT-4 API의 호출 프롬프트로 제공

하고, one-shot 예시를 사용하여 해당 문서를 코드 스위칭 코퍼스로 변환하는 방식을 적용하였다. 만들어진 코퍼스는 github을 통해 다운로드받을 수 있다³. 그림 1은 해당 변환 과정에서 얻어진 코드 스위칭 코퍼스의 일부를 나타낸다.

상형 문자는 'pictogram'에서 developed되었으며, 그림 문자와의 major difference는 문자가 단어와 combine되는지에 있습니다. Pictogram은 'words'와 언어의 관계를 consider하지 않고, 그림에 representation된 concept이나 sentence를 전달합니다. 하지만, 상형 문자는 언어와 문자가 correspond하는 'ideographic character'입니다. 상형 문자에서 각 character는 specific한 meaning을 가지며, 대부분 'words'나 morphemes를 represent합니다. 이 유형의 상형 문자 examples로는, 한자, hieroglyphics, Indus 스크립트 등이 있습니다.

그림 1. 코드 스위칭 코퍼스 예시

이를 통해 총 8,233건의 문서(4,198,779 음절, 664,579 어절)대한 코드 스위칭 코퍼스를 구축할 수 있었다. 이 코퍼스는 LLAMA-2 [15] tokenizer 기준으로 약 3M (3,268,659) 개 토큰을 포함하며, 이는 전체 LLAMA-2의 학습 데이터 2조 개 토큰 중에서 한국어가 차지하는 비율인 0.06%에 해당하는 약 12억 토큰의 0.0025% 다.

4. 코드 스위칭 코퍼스 활용 LLM 학습

코드 스위칭 코퍼스가 지식 전이에 미치는 효과를 검증하기 위해 LLAMA-2 백본 모델을 사용하였다. 3에서 언급한 바와 같이, LLAMA-2는 전체 학습 데이터 중 한국어 데이터가 0.06%만을 차지하는데, 이런 비율은 언어 모델의 관점에서 한국어를 소수언어로 분류할 수 있다. 본 연구에서는 임베딩 층이나 모델 사전의 수정을 수행하지 않았으며, 모델에 전이학습만 진행하였다. 학습 과정에서의 파라미터 설정은 다음과 같다: 시퀀스 길이 2048, 배치 크기 32, 에폭 수 5, 학습률은 4e-5로 설정하였다. 이렇게 만들어진 모델은 아래와 같다.

- (1) LLAMA-2-7B (baseline): Meta에서 공개한 LLM [15]
- (2) LLAMA-2 + Kor: 코드 스위칭 코퍼스의 원본 한국어만을 활용해 학습한 모델
- (3) LLAMA-2 + Pangyo⁴: 코드 스위칭 코퍼스로 학습한 모델
- (4) LLAMA-2-Kor-7B: 대량의 한국어 데이터를 통해 학습하여 공개된 모델 [16]
- (5) Polyglot-5.8B: 한국어 특화 언어모델 [17]

5. 언어 생성 능력 평가

먼저, 새롭게 구축한 코퍼스를 통해 학습한 언어 모델

¹ <https://huggingface.co/beomi/llama-2-ko-7b>

² <https://huggingface.co/beomi/KoAlpaca-30B-LoRA>

³ <https://github.com/MrBananaHuman/PangyoCopora>

⁴ 전문 용어를 대상으로 나타나는 대표적인 한국어 코드 스위칭 현상을 지칭하는 은어인 ‘판교어’를 의미한다.

의 생성 성능을 정성적으로 관찰했다. 본 연구의 주요 목적은 영어 지식이 한국어로의 지식 전이가 성공적으로 이루어졌는지를 확인하는 것이다. 이를 확인하기 위해, 영어로 주어진 질문에 대한 응답과 한국어로 주어진 질문에 대한 응답이 유사한 지식 정보를 포함하는지 확인했다.

이를 구체화하기 위해 다른 조건 없이 질문만 주어지고, 답변을 유도하는 제로-샷 생성 테스트를 수행하였다. 모델에 입력으로 주어진 프롬프트와 결과는 아래 표 1과 같다. 디코딩 전략으로는 빔서치를 선택하였으며, 빔의 수를 10개로 제한하여 테스트를 진행하였다.

표 1. 모델의 자연어 생성 예시

Question: What caused the Great Depression?	
Answer:	
LLAMA-2	The Great Depression was caused by a combination of factors, including the stock market crash of 1929, the failure of banks and businesses, and the protectionist policies of the Smoot-Hawley Tariff Act.
LLAMA-2 + Pangyo	여러 요인이 있었습니다. 첫 번째로, 1929년 10월 29일 발생한 'Black Tuesday'이며, 이는 뉴욕 증권거래소의 주식 가격이 dramatic하게 떨어졌을 때 시작된 것입니다. Black Tuesday 이후, 미국은 economic recession으로 기울어 가던 상태였고, 1903년 12월 5일에는 Smoot-Hawley Tariff가 시행되어 미국과 다른 나라들 간의 trade 관계가 더욱 악화되었습니다.
질문: 경제 대공황의 원인은 무엇인가?	
답변:	
LLAMA-2	실업률이 급격히 증가함으로써 경제가 침체 상태에 빠질 수 있다는 것을 의미한다.
LLAMA-2 + Pangyo	여러 요인들이 있었습니다. 첫째, 1929년 10월 29일, 뉴욕 증권거래소에서 큰 폭락이 발생했고, 이로 인해 세계 금융 시장은 크게 붕괴되었습니다. 'Black Tuesday'이라는 이름으로 알려진 이 날은 미국의 경제 history에서 가장 어려운 시기 중 하나로, 실업률이 25%, 빈곤선 아래로 떨어진 사람들의 수가 30%에 달했습니다.

LLAMA-2의 출력 결과를 관찰해보면, 영어 질의에는 정확한 배경 지식을 기반으로 한 답변이 반환되는 것을 확인할 수 있었다. 그러나 동일한 내용의 한국어 질의에 대해서는 영어 질의와 비교하여 상대적으로 지식 정보의 풍부함이 부족한 응답이 반환되었다. 그러나 코드 스위칭 코퍼스를 추가적으로 학습한 LLAMA-2 + Pangyo 모델에서는 영어와 한국어 질의에 대한 답변 간의 유사성이 높게 나타났다. 특히, LLAMA-2 + Pangyo 모델이 반환한 영어 질의의 응답은 LLAMA-2 모델의 영어 질의 응답의 지식 정보를 큰 부분 포함하고 있었다. 이런 결과는 영어 모델을 한국어 모델로 전환하는 과정에서 상대적으로 적은 양의 데이터로도 효과적인 전이 학습이 가능함을 시사한다.

6. In-context learning 평가

다음으로 언어 모델의 정량적 성능을 평가하기 위해

퓨-샷 테스트를 진행하였다. 퓨-샷 테스트는 제로-샷 테스트와 달리, 실험하고자 하는 과제에 대한 예시를 모델에 제시한 후, 정답을 맞추도록 유도하는 것이다. 테스트 수행을 위해, 지식 학습 능력을 평가할 수 있는 huggingface의 lm evaluation harness 오픈소스[18]를 활용하였다. 특히, 학습 데이터에서 무작위로 선택된 5개의 퓨-샷 예시를 입력하고, 평가 데이터(dev)로 평가를 실시하였다. 'BoolQ'는 주어진 문장에 대한 지식 평가를 수행하는 태스크로, 언어 모델이 보유하고 있는 지식의 깊이와 정확성을 평가하는 데 있어 중요한 지표로 활용되었다. 앞선 실험과 마찬가지로, 한국어 BoolQ 데이터[19]와 영어 BoolQ[20]를 각각 분석하였으며, 정확도(accuracy)를 비교했다.

표2. 한국어, 영어 BoolQ 벤치마크 성능 비교

	LLAMA-2	LLAMA-2 + Kor	LLAMA-2 + Pangyo	LLAMA-2-Ko	Polyglot
BoolQ (Ko)	0.59	0.60	0.64	0.69	0.59
BoolQ (En)	0.78	0.80	0.81	0.63	0.59

표 2의 결과에 따라, 본 연구에서 baseline으로 사용된 백본 언어 모델 LLAMA-2는 한국어 BoolQ에서 미흡한 성능을 보였다. 그러나 코드 스위칭 코퍼스를 활용하여 추가 학습된 LLAMA-2 + Pangyo 모델은 기존의 LLAMA-2보다 향상된 성능을 보였다. LLAMA-2-Ko 모델은 한국어 BoolQ에서 LLAMA-2 + Pangyo 대비 높은 정확도를 보였으나, 영어 BoolQ에서 기존의 LLAMA-2에 비해 정확도가 크게 저하된 결과를 나타냈다. 반면, LLAMA-2 + Pangyo 모델은 영어 BoolQ에서도 LLAMA-2보다 우수한 성능을 보이며, 이는 제안된 방법론이 한국어와 영어 간의 지식 정보 연계에 효과적으로 작용하였음을 시사한다.

7. 의미 임베딩 평가

LLM이 사전에 학습한 영어 기반의 지식과 한국어와의 연계성을 검증하기 위해 번역 말뭉치를 활용하였다. 이를 위해 LLM의 한국어와 영어의 임베딩 벡터 간 유사성을 분석하여, 한-영 간의 의미론적 연계가 얼마나 잘 이루어졌는지를 판단할 수 있다. 본 연구에서는 일반적인 용어를 포함하고 있는 뉴스 도메인의 병렬 말뭉치와, 기술과학 분야의 전문 말뭉치를 활용하여 이러한 비교를 수행하였다. 아래 표3은 사용된 번역 병렬 말뭉치의 예시를 표현한다.

표 3. 병렬 말뭉치의 예시

병렬 말뭉치	한국어	영어
기술과학⁵	상기 발음듣기는 데이터	The pronunciation
	베이스관리부에 저장된	listening is linked
	영어사전 및 한글사전과	with the English
	연동되어, 플레이어가 선	dictionary and the

뉴스 ⁶	택한 단어를 원어민 발음 및 음성으로 재생시키는 역할을 한다.	Korean dictionary stored in the database management unit, and plays a role of reproducing the word selected by the player as the pronunciation and voice of a native speaker.
	창설 작업의 실무자였던 이범석의 “국민당 정부와 사전협의 없이 우선 만들어 놓고 내밀어 보자는 뉘스 뿐이었다”라는 회고가 독자적인 창설 의지를 말해준다.	Lee Bum-suk, a practitioner of the founding work, said, “It was only a jute to make it out without prior consultation with the nationalist government.”, which shows that independent will of the foundation.

위 실험을 위해 각 병렬 말뭉치로부터 1,000개의 병렬 병렬 쌍을 무작위로 추출하였다. 이후, 해당 쌍에 속하는 각 언어의 문장을 모델에 입력으로 제공하였다. 입력된 토큰 중에서 마지막 토큰의 임베딩 벡터 값을 추출하고, 한국어 문장과 영어 문장의 임베딩 벡터 간의 유사성을 코사인 유사도를 통해 평가하였다.

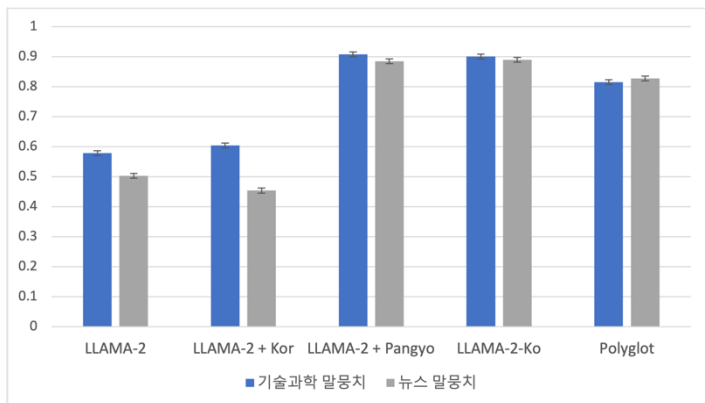


그림 2. 병렬 말뭉치의 문장 간 임베딩 유사도 비교

그림 2는 각 모델에 따른 평균 유사도를 시각화한 결과이다. 유사도는 0에서 1 사이의 값으로 표현되며, 두 문장이 매우 유사할 경우 1에 근접한 값을 나타낸다. 기본 모델로 설정한 LLAMA-2는 평균 유사도로 0.58 및 0.50의 값을 보였다. 이 결과는 LLAMA-2 모델 내의 한국어와 영어 간 의미론적 연계가 상대적으로 약함을 나타낸다. 반대로, LLAMA-2 + Pangyo 모델은 다른 모델들에 비해 눈에 띄게 높은 유사도를 보였다. 이는 제한된 코드 스위칭 코퍼스를 통한 학습이 두 언어 간의 의미론적 정보와 지식의 연계를 만들어내는 가교 역할을 할 수 있음을 증명한다.

8. 결론

본 연구에서는 대형 언어 모델의 언어별 지식 전이에 대한 가능성을 탐구하였다. 실험 결과, 한국어를 대상으로한 지식 전이는 성공적으로 이루어졌음을 확인하였다. 그러나, 이러한 방법론이 모든 언어에 일반화될 수 있는지에 대한 확신을 얻기 위해서는 추가적인 실험 및 연구가 필요하다. 특히, 다른 언어와 문화 배경을 가진 데이터셋에서의 성능을 검증해보는 것이 중요하다.

또한, 현재 AI와 관련된 기술 및 지식의 소유가 주로 영어를 중심으로 확산되고 있음을 고려할 때, 본 연구의 방법론은 AI의 민주화와 세계화에 중요한 기여를 할 것으로 기대된다. 특히 소수 언어와 소수 문화 지역에서의 AI 접근성이 제한적인 문제를 해결하는 데 있어, 본 연구의 접근 방식이 중요한 역할을 할 수 있을 것이다.

결국, AI 기술의 폭넓은 확산을 위해서는 다양한 언어와 문화를 아우르는 연구와 방법론의 개발이 필수적이다. 본 연구는 그 첫 걸음을 떼는 데 있어 중요한 단계라고 할 수 있다.

참고문헌

- [1] Lagler, Klemens, et al. "GPT2: Empirical slant delay model for radio space geodetic techniques." *Geophysical research letters* 40.6 (2013): 1069-1073.
- [2] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
- [3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [4] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.
- [5] Cui, Yiming, Ziqing Yang, and Xin Yao. "Efficient and effective text encoding for chinese llama and alpaca." *arXiv preprint arXiv:2304.08177* (2023).
- [6] Hirano, Masanori, Masahiro Suzuki, and Hiroki Sakaji. "llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models and its Methodology." *arXiv preprint arXiv:2305.12720* (2023).
- [7] Penedo, Guilherme, et al. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." *arXiv preprint arXiv:2306.01116* (2023).

- [8] <https://huggingface.co/clibrain/lince-zero>
- [9] de Vries, Wietse, and Malvina Nissim. "As good as new. How to successfully recycle English GPT-2 to make models for other languages." *arXiv preprint arXiv:2012.05628* (2020).
- [10] Zeng, Qingcheng, et al. "GreenPLM: Cross-Lingual Transfer of Monolingual Pre-Trained Language Models at Almost No Cost."
- [11] Ostendorff, Malte, and Georg Rehm. "Efficient language model training through cross-lingual and progressive transfer learning." *arXiv preprint arXiv:2301.09626* (2023).
- [12] Marchisio, Kelly, et al. "Mini-Model Adaptation: Efficiently Extending Pretrained Models to New Languages via Aligned Shallow Training." *arXiv preprint arXiv:2212.10503* (2022).
- [13] Lee, Chanhee, et al. "Exploring the data efficiency of cross-lingual post-training in pretrained language models." *Applied Sciences* 11.5 (2021): 1974.
- [14] Gardner-Chloros, Penelope. *Code-switching*. Cambridge university press, 2009.
- [15] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288* (2023).
- [16] <https://huggingface.co/beomi/llama-2-ko-7b>
- [17] Ko, Hyunwoong, et al. "A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models." *arXiv preprint arXiv:2306.02254* (2023).
- [18] <https://github.com/EleutherAI/lm-evaluation-harness>
- [19] https://huggingface.co/datasets/skt/kobest_v1
- [20] <https://huggingface.co/datasets/boolq>