도메인 특화 LLM 지식 내재화 평가 프레임워크

양동헌, 임찬욱
한국과학기술정보연구원
{yangdonghun3, chanuklim}@kisti.re.kr

Evaluation Framework for Knowledge Internalization in Domain-specific LLMs

Donghun Yang, Lim Chanuk
Korea Institute of Science and Technology Information (KISTI)

요 약

최근 자연어 처리를 비롯한 다양한 분야에서 대형언어모델(LLM)이 폭넓게 활용되고 있으며, 특히 특정 분야에 특화된 도메인 특화 LLM 연구가 활발히 진행되고 있다. 도메인 특화 LLM은 일반적으로 오픈 LLM에 도메인 특화 데이터를 추가 학습하는 방식으로 개발되며, 이 과정은 크게 추가 사전학습(Continual Pre-training, CPT)과 지도 미세조정(Supervised Fine-tuning, SFT)으로 구분된다. SFT는 주로 모델에 특정 태스크 수행 능력과 응답 양식을 부여하는 데 초점을 두며, CPT는 방대한 도메인 코퍼스로부터 지식을 내재화하는 것을 목적으로 한다. 본 연구는 이러한 관점에서, CPT된 도메인 특화 LLM이 실제로 도메인 지식을 내재화했는지를 검증하기 위한 연속 문장 생성 기반 평가 프레임워크를 제안한다. 이를 위해 한국어와 영어 논문 초록 각각 1,000건을 수집해 문장 단위로 분할하여 평가 데이터셋을 구축하였다. 구축한 데이터셋의 초록 앞부분을 입력으로 모델이 마지막 1-3문장을 예측하도록 하고, 이를 실제 초록과 비교하여 지식 내재화 성능을 평가하였다. 본 평가 프레임워크를 Gemma3-4b-pt 모델을 과학기술정보 특화 코퍼스로 CPT한 KONI-4b-base 모델에 적용하고, 그 결과를 Gemma3-4b-pt 및 Gemma3-4b-it와 비교하여 제안 프레임워크의 우수성을 검증하였다. 실험 결과, 한국어 데이터셋에서는 KONI-4B-base 모델이 가장 우수한 지식 내재화 성능을 기록하였으며, 영어 데이터셋에서는 Gemma3-4b-pt가 상대적으로 더 뛰어난 성능을 보였다. 이는 Gemma3-4b-pt가 사전학습 단계에서 이미 대규모 영어 논문 데이터를 충분히 학습하였기 때문에 KONI-4b-base에서 수행된 CPT의 추가 지식 내재화 효과가 제한적이었던 것으로 해석된다. 한편, Gemma3-4b-it 모델은 전반적으로 낮은 성능을 보였는데, 이는 SFT 과정에서의 태스크 지향적 조정이 연속 문장 생성에 제약으로 작용했음을 시사한다. 본 연구는 CPT의 지식 내재화 여부를 정량적으로 평가할 수 있는 간단하면서도 효과적인 프레임워크 제안하였으며, 이는 향후 도메인 특화 LLM 개발에서 학습 데이터 구성과 평가 방법론에 중요한 시사점을 제공할 것으로 기대된다.

Ⅰ. 서 론

최근 대규모 데이터와 파라미터를 기반으로 한 대형언어모델(LLM)은 번역, 요약, 질의응답과 같은 범용 작업뿐만 아니라 다양한 응용 분야에서도 탁월한 성능을 보이며, 자연어 처리를 비롯한 여러 영역에서 기존 패러다임을 빠르게 대체하고 있다 [1]. 특히 특정 도메인에 최적화된 LLM 개발에 대한 수요가 급격히 증가하고 있으며, 의료, 법률, 과학기술 등과 같이 전문성이 요구되는 다양한 분야에서 해당 영역의 전문 지식을 효과적으로 내재화한 도메인 특화 LLM 연구가 활발히 진행되고 있다.

도메인 특화 LLM은 일반적으로 공개된 범용 LLM을 기반으로 도메인 특화 데이터를 추가 학습하는 방식으로 개발되며, 크게 추가 사전학습(Continual Pre-training, CPT)과 지도 미세조정(Supervised Fine-tuning, SFT)으로 구분된다 [2]. SFT는 특정 태스크에 맞추어 모델 출력을 조정하고 사용자 지향적 응답 양식을 학습시키는 데 효과적이지만, 모델 내부의 지식 구조를 근본적으로 변화시키는 데에는 한계가 있다. 반면 CPT는 대규모 도메인 코퍼스를 활용하여 새로운 지식을 학습하고 내재화하는 데 초점을 두며, 특정 도메인의 언어적 특성과 지식 구조를 모델 내부에 반영하기 위한 핵심 과정이다.

이러한 CPT의 지식 내재화 능력을 평가하려는 연구가 다양하게 수행되고 있으나, 대부분 분류, 질의응답, 요약 등 특정 태스크 성능을 기준으로 한 간접 평가 수준에 그쳐, 모델이 실제로 지식을 학습했는지, 아니면 단순히 태스크 수행 능력만 향상된 것인지를 명확히 구분하기 어렵다. 또한 기존 평가 방법은 태스크 설계와 데이터셋 구성에 크게 의존하기 때문에, 모델 내부에 축적된 지식의 양과 질을 직접적으로 평가하기 어렵다는 한계가 있다.

따라서 본 연구는 CPT된 도메인 특화 LLM이 실제로 도메인 지식을 내재화했는지를 직접적으로 검증하기 위한 연속 문장 생성 기반 지식 내재화 평가 프레임워크를 제안한다. 제안 프레임워크는 논문 초록을 문장 단위로 분할하고 앞부분을 입력으로 제공하여 모델이 마지막 문장을 생성하도록 한 뒤, 이를 실제 초록과 비교하는 방식으로 지식 내재화 정도를 평가한다. 이를 통해 기존의 태스크 기반 간접 평가를 보완하고, CPT가 도메인 특화 LLM의 지식 내재화에 미치는 효과를 보다 직접적이고 명확하게 검증할 수 있는 방법론적 기반을 마련하고자 한다.

Ⅱ. 관련 연구

기존 LLM 평가 연구는 주로 분류, 질의응답, 요약 등 특정 태스크 성능 평가에 한정되어 왔다. 이러한 접근은 모델의 태스크 수행 능력을 평가 할 수는 있으나, 실제로 특정 지식을 내재화했는지를 직접적으로 검증하기에는 한계가 있다. 최근 언어모델의 지식 보유 여부를 평가하기 위해 지식 추출형 프롬프트나 상식 질의응답 벤치마크를 활용한 연구도 보고되었으나, 이는 프롬프트 설계나 특정 포맷 의존성에 크게 좌우되는 문제가 있다 [3]. 따라서 기존의 간접 평가를 넘어, 보다 직접적으로 모델의 지식 내재화를 검증할 수 있는 새로운 평가 방식의 필요성이 제기되고 있다.

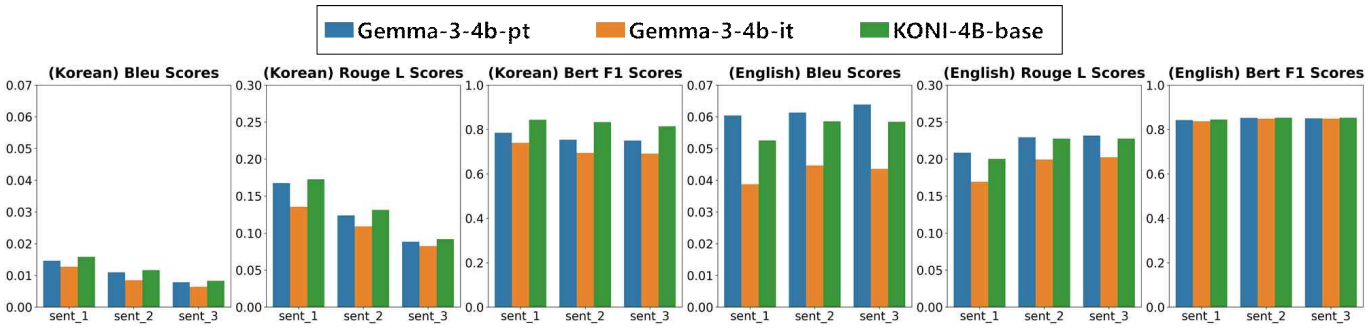

그림 1. 제안 평가 프레임워크 기반 지식 내재화 성능 비교

Ⅲ. 연구 방법

본 연구는 기존 태스크 중심 평가의 한계를 극복하고자, CPT된 도메인 특화 LLM의 지식 내재화 여부를 직접 검증하기 위한 연속 문장 생성 기반 평가 프레임워크를 제안한다.

3.1 평가 데이터셋 구축

지식 내재화 평가 프레임워크 구축을 위해 한국어와 영어 논문 초록 각각 1,000건을 수집하고 문장 단위로 분할하여 평가 데이터셋을 구성하였다. 초록 앞부분은 모델의 입력으로 제공하고, 마지막 1 - 3문장은 정답으로 활용하였다. 논문 초록은 도메인 지식이 밀집된 텍스트로서, 이를 통해 모델이 해당 지식을 실제로 내재화했는지를 효과적으로 검증할 수 있다.

3.2 연속 문장 생성 기반 지식 내재화 평가

제안 프레임워크는 입력된 초록 앞부분을 기반으로 모델이 후속 문장을 생성하도록 하고, 이를 실제 초록과 비교하는 방식으로 지식 내재화 정도를 평가한다. 이를 통해 모델이 도메인 지식을 활용해 연속적이고 일관된 서술을 생성할 수 있는지를 검증할 수 있으며, 단순 태스크 성능 중심의 기존 방법과 달리 지식 내재화 여부를 직접적으로 평가할 수 있다.

Ⅳ. 실험

4.1 실험 환경

본 연구에서 제안한 지식 내재화 평가 프레임워크는 Python 3.11, PyTorch 2.6.0, transformers 4.51.3 및 vllm 0.9.0 환경에서 NVIDIA A100 (80GB) GPU를 활용하여 구현되었다. 본 연구의 우수성을 검증하기 위해 제안 프레임워크를 Gemma3-4b-pt 모델을 과학기술정보 특화 코퍼스로 CPT한 KONI-4b-base [4] 모델에 적용하고, 그 결과를 Gemma3-4b-pt [5] 및 Gemma3-4b-it [5]와 비교 분석하였다. 지식 내재화 평가 지표로는 생성된 문장과 실제 초록 간의 유사도를 정량적으로 측정할 수 있는 BLEU, ROUGE, BERTScore를 활용하였다.

4.2 실험 결과

그림 1은 제안한 지식 내재화 평가 프레임워크를 구축한 한국어와 영어 논문 초록 데이터셋을 기반으로 세 가지 모델(Gemma3-4b-pt, Gemma3-4b-it, KONI-4b-base)에 적용하여 성능을 비교한 결과를 보여준다. 한국어 데이터셋의 경우 KONI-4b-base가 모든 지표에서 가장 우수한 성능을 기록하였으며, 이는 CPT를 통해 과학기술정보 특화 지식을 효과적으로 내재화했음을 시사한다. 영어 데이터셋에서는 Gemma3-4b-pt가 BLEU, ROUGE-L, BERTScore 전반에서 가장 뛰어난 성능을 보였는데, 이는 사전학습 단계에서 이미 대규모 영어 논문 데이터를 충분히 학습하였기 때문에 KONI-4b-base에서 수행된 CPT의 추가 지식 내재화 효과가 제한적이었던 것으로 해석된다. 한편 Gemma3-4b-it은 모든 지표에서 가장 낮은 성능을 기록하였으며, 이는 SFT 과정에서의 태스크 지향적 조정이 연속 문장 생성 능력에 제약으로 작용했음을 보여준다.

V. 실험

본 연구는 도메인 특화 LLM의 지식 내재화 여부를 직접적으로 검증하기 위한 연속 문장 생성 기반 평가 프레임워크를 제안하였다. 제안된 프레임워크는 논문 초록 데이터셋을 활용해 모델의 연속 문장 생성 능력을 평가함으로써 기존 태스크 중심 평가의 한계를 보완하였으며, 이를 통해 CPT가 실제로 지식 내재화에 기여했는지를 정량적으로 확인할 수 있는 간단하면서도 효과적인 절차를 제공한다. 본 연구는 CPT된 도메인 특화 LLM의 지식 내재화 여부를 정량적으로 평가할 수 있는 새로운 관점을 제시하였으며, 이는 향후 도메인 특화 LLM 개발 과정에서 학습 데이터 구성과 평가 방법론 수립에 중요한 시사점을 제공할 것으로 기대된다.

ACKNOWLEDGMENT

This research was supported by the Korea Institute of Science and Technology Information (KISTI) in 2025 (No. (KISTI) 25L1M1C1), aimed at developing KONI (KISTI Open Neural Intelligence), a large language model specialized in science and technology.

참 고 문 헌

[1] A. Matarazzo and R. Torlone, "A survey on large language models with some insights on their capabilities and limitations," arXiv preprint arXiv:2501.04040, 2025.

[2] Ouyang, L., Wu, J., Jiang, X., et al. "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol. 35, pp. 27730-27744, 2022.

[3] G. Son, H. Lee, S. Kim, S. Kim, N. Muennighoff, T. Choi, C. Park, K. M. Yoo, S. Biderman, et al., "Kmmlu: Measuring massive multitask language understanding in Korean," arXiv preprint arXiv:2402.11548, 2024.

[4] KISTI, "KONI-4B-base-20250819," 2025, (https://huggingface.co/KISTI-KONI/KONI-4B-base-20250819).

[5] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Riviere, et al., "Gemma 3 technical report," arXiv preprint arXiv:2503.19786, 2025.