거대언어모델에 대한 원자력 안전조치 용어 적용 가능성 평가

윤성호O
O한국원자력통제기술원
e-mail: shyoon@kinac.re.krO

A Training Feasibility Evaluation of Nuclear Safeguards Terms for the Large Language Model (LLM)

Sung-Ho YoonO
OKorea Institute of Nuclear Nonproliferation and Control

● 요 약 ●

본 논문에서는 원자력 안전조치 용어를 미세조정(fine tuning) 알고리즘을 활용해 추가 학습한 공개 거대언어모델(Large Language Model, LLM)이 안전조치 관련 질문에 대해 답변한 결과를 정성적으로 평가하였다. 평가 결과, 학습 데이터 범위 내 질문에 대해 학습 모델은 기반 모델 답변에 추가 학습 데이터를 활용한 낮은 수준의 추론을 수행한 답변을 출력하였다. 평가 결과를 통해 추가 학습 개선 방향을 도출하였으며 저비용 전문 분야 언어 모델 구축에 활용할 수 있을 것으로 보인다.

키워드: 안전조치(Safeguards), 거대언어모델(Large Language Model)

I. Introduction

2022년 11월 OpenAI에서 'ChatGPT'를 공개한 이후 챗봇(Chatbot)을 비롯한 사용자 인터페이스(UI) 서비스들이 사용자의 질문에 대해 사전 정의된 답변을 출력하는 방식에서 거대언어모델(LLM)을 통해 추론하고 생성된 답변을 하는 방식으로 바뀌고 있다.

ChatGPT 공개 이후 메타(Meta)에서 오픈 소스 기반 LLM인 'LLaMA'를 공개하며 대중들은 거대언어모델을 무료로 사용하고 개발할 수 있게 되었으며 이로부터 파생된 거대언어모델들도 공개되고 있다.

이에 따라 정보 유출에 민감하고 거대언어모델 및 학습데이터 저작권에 대한 법률적 문제를 피하길 원하는 사용자들이 로컬(local) 환경에 추가학습을 통해 특화된 LLM을 구축하려고 노력하고 있다.

II. Method

1. Development Environment

거대언어모델에 대한 미세조정 학습과 답변 생성을 위한 시스템 환경은 표 1과 같다.

Table 1. Computer specification

Item	Model
CPU	Intel i7-13700
GPU	Nvidia RTX 3060 12GB
RAM	DDR5 32 GB
OS	Windows 11 pro

모델 구동을 위한 소프트웨어로는 Python 3.12.0과 Visual Studio Code를 사용하였다.

2. LLM Model

기반 LLM 모델은 성능이 검증되었으며 한국어에 특화된 Polyglot-Ko 5.8B 모델[1]에 한국어 추론 능력 향상을 위해 네이버 지식iN 베스트[2]에 등재된 질문과 답변 21,155개를 학습(full-finetune) 시킨 KoAlpaca-Polyglot-Ko 5.8B 모델[3]로 선정했다.

LLM 학습기법은 저사양 PC에서 구동할 수 있도록 4bit 양자화 학습기법인 QLOLA[4]를 활용하여 전체 매개변수를 재학습하지 않고 추가학습 데이터를 Adapter 방식으로 추론성능을 향상시키는 미세조정기법(Fine tuning)을 사용하였다.

3. Dataset

학습 Dataset은 IAEA에서 발간한 안전조치 용어집[5] 중 15개 용어에 대한 설명을 한국어로 번역하여 학습시켰다. 학습한 용어는 안전조치 협정, 추가의정서, 소량의정서, 사찰, 최초사찰, 정기사찰, 단기 통보 사찰, 사찰의 빈도, 사찰과 방문의 사전 통보 등이다.

III. Result

원자력 안전조치 용어 학습에 따른 거대언어모델 답변 결과는 표 2와 같이 나타났다.

Table 2. Comparison of Q/A Result in LLMs

질문 1	원자력 안전조치란?
기반 모델	원자력 안전조치는 방사성 물질의 폭발이나 연쇄적인 사고로 인한 위험을 줄이기 위한 예방적 조치입니다. (생략)
학습 모델	안전조치는 핵물질의 사용을 안전하게 관리하기 위한 시설 및 기술적 조치의 조합으로 규정된다. (생략)
질문 2	IAEA의 역할은?
기반 모델	국제원자력기구(IAEA)는 국제적인 원자력 발전 설비 및 기술 개발에 대한 협력을 촉진하는 기구입니다. (생략)
학습 모델	IAEA는 International Atomic Energy Agency의 준말로, 안전조치에 관한 것은 국가 간에 협상할 필요에 따라 IAEA에 의해 시행된다. (생략)

질문 1에 대해 기반 모델은 원자력 안전조치(Safe-guards)를 원자력 안전(Safety)을 위한 조치 사항으로 해석한 답변을 내놓았으며, 학습 모델은 안전조치를 핵물질의 평화적 이용을 위한 조치라는 답변을 하여 학습한 데이터에 대한 직접적인 답변은 가능한 것으로 확인하였다.

질문 2는 학습 모델의 추론 성능을 확인하기 위한 질문을 입력하였다. 기반 모델은 IAEA와 관련하여 백과사전 수준의 설명에 네이버 지식iN 데이터가 혼합된 답변으로 평가할 수 있었다면, 학습 모델은 기존 답변에 IAEA가 안전조치를 수행하는 기관이라는 추가 답변을 내놓아 학습한 안전조치 데이터로부터 IAEA의 역할을 추론할 수 있음을 확인하였다.

IV. Conclusions

본 연구에서는 저사양 PC 환경에서 공개 한국어 거대언어모델에 원자력 안전조치 용어를 학습시키고 안전조치 관련 질의에 대한 답변 성능을 정성적으로 평가하였다. 평가 결과 학습 모델이 낮은 수준의 추론이 가능함을 확인하였으며 기반 모델의 해당 분야에 대한 추론 성능이 뛰어나다면 세부 분야에 대해서는 저비용 추가 학습을 통해 추론 성능을 높일 수 있을 것으로 예상되었다.

향후 용어 데이터 및 학습 연산 작업량을 늘린다면 내부 저경력 직원을 대상으로 안전조치 교육용 챗봇을 운영하거나 지식정보 데이터베이스용 인터페이스로 활용할 수 있을 것이다.

REFERENCES

[1] EleutherAI, https://github.com/EleutherAI/polyglot.

[2] https://kin.naver.com/best/listaha.naver

[3] J. Lee, https://huggingface.co/beomi/KoAlpaca-Polyglot-5.8B.

[4] T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, "QLoRa: Efficient Finetuning of Quantized LLMs," arXiv: 2305.14314, May 2023.

[5] IAEA, IAEA Safeguards Glossary, 2022.