에이전트 기반 AI 모델 개방성 평가 시스템

한현준, 이강원
세종대학교 컴퓨터공학과

super2002010119@daum.net, kangwon.lee@sejong.ac.kr

Agent-based AI model openness evaluation system

Hyeon-Jun Han, Kang-Won Lee
Dept. of Computer Science, Sejong University

요 약

AI 모델 개방의 긍정적인 효과와 요구에 따라 AI 모델의 개방성 평가도 중요시되고 있다. 기존 AI 모델 개방성 평가는 수작업으로 진행되어 많은 시간과 노력을 요구한다. 본 논문에서는 모델 개방성 평가를 자동화하는 에이전트 기반 AI 모델 개방성 평가 시스템 아키텍처를 제안한다.

1. 서론 및 기존 연구

최근 AI 의 발전으로 오픈 소스 AI 모델들이 대거 등장하고 있다. 하지만 AI 모델마다 개방된 항목과 정도가 상이하고, 정부(예: EU AI Act)와 기술 커뮤니티(예: OSI)는 높은 개방성(Openness)을 요구함에 따라 AI 모델의 개방성 평가는 중요시되고 있다.

기존 AI 모델 개방성 평가는 연구자가 직접 논문, 오픈소스 커뮤니티, 홈페이지를 참고하여 수작업으로 이루어져 시간과 노력이 많이 들고, 평가의 일관성이 떨어진다는 단점이 존재한다.

White et al.[1]는 17 가지 모델 개방성 평가항목과, Open Model, Open Tooling Model, Open Science Mode 로 3 가지 등급으로 나누었고, 평가항목을 3 가지 등급에 매핑하여 모델의 개방성을 등급으로 나눈 MOF (Model Openness Framework)를 제시하였다. 또한, MOT (Model Openness Tool)와 리더보드를 개발하여 모델의 개방성을 비교하였다. MOT 는 AI 모델의 개방성을 평가하는 툴로 개방성 평가항목을 대표적으로 코드, 데이터, 문서로 분야를 나누었고, 각 분야를 세분화하여 총 16 가지의 평가항목으로 구성하였다. 각 평가항목별로 라이선스 정보와 구성 요소의 경로를 사용자가 수동으로 툴에 입력을 한다. 툴은 각 구성 요소를 직접 분석하지 않고, 라이선스의 정보를 이용하여 각 평가항목별 공개, 비공개 평가한 후, MOF 에서 제시한 3 단계 등급 중 하나로 지정하였다.

그 외의 Liu et al.[2], Eiras et al.[3] 등 다수의 연구자들이 모델 개방성 평가 프레임워크를 제안하고, AI 모델을 평가하였다.

기존 연구는 특정 영역 (코드, 가중치)의 개방성을 위주로 평가한 경향이 보이고, 개방성 평가 프레임워크가 OSI (Open Source Initiative)에서 제시한 공개모델의 재사용과 연구 가능여부와 같은 오픈소스 AI 정의(OSAID)에 적합하지 않다[4].

2. 평가 프레임워크

<표 1> 각 항목별로 제안한 프레임워크가 기존 연구의 세부 사항을 어느정도 포함하는가에 관한 표

	모델 기본 개방성	재현성과 이용가능성	학습 방법론 개방성	데이터 개방성
Ours	○	○	○	○
OSD[5]	△	△	×	×
OSAID[4]	○	△	×	○
Liu et al.[2]	△	×	△	○
Eiras et al.[3]	△	×	×	○
White et al.[6]	○	△	×	○

○: 2/3 이상 포함, △: 2/3 미만 1/3 이상 포함, ×: 불 포함

에이전트 기반 자동화 시스템에서 활용할 AI 모델 개방성 평가 프레임워크는 이전 연구에서 제안한 프레임워크로 OSI 의 오픈소스 AI 정의와 모델 재현가능성에 중점을 두고, 기존 연구들의 평가항목들도 포괄적으로 포함(표 1)하여 16 개의 평가항목으로 구성하였다. 평가항목은 모델 기본 개방성(가중치, 코드, 라이선스, 논문, 아키텍처, 토크나이저), 재현성과 이용가능성(하드웨어 스펙, 소프트웨어 스펙, API), 학습 방법론 개방성(사전학습, 파인 튜닝, 강화학습/DPO), 데이터 개방성(사전학습, 파인 튜닝, 강화학습/DPO, 데이터 필터링)으로 구성되어 있다. 각 평가항목당 완전히 재현 가능할 정도로 공개되어 있으면 개방, 일부만 공개 되어있을 때 준개방, 그 외는 비개방으로 개방의 정도를 분리한다[7].

3. 시스템 개요

3.1 입력 프로세스

사용자로부터 받은 모델명이나 URL 을 URI 스키마나 키워드 검사 등을 활용하여 모델명을 추출하고, 공통으로 사용할 “모델 식별자”로 정규화 한다. 예를 들어 https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E 혹은 Llama 4 Scout 17B 16E 에서 Llama-4-Scout-17B-16E 를 얻는 과정이다. 모델 식별자와 입력 원본을 결합한 후, 이를 JSON 형식으로 직렬화 하여 FetchRequest 를 생성하고, Dispatcher Agent 로 전달한다.

3.2 정보 수집 프로세스

Dispatcher Agent 는 FetchRequest 를 받아 분석한 후 ArxivFetcher, HuggingFaceFetcher, GitHubFetcher, CoWebSiteCrawler 의 4 개의 모듈에 16 개의 평가항목을 적절하게 할당하여 병렬로 FetchRequest 와 함께 전파한다. 각 Fetcher 는 arXiv API, Hub API, GitHub API, 웹 크롤러(web crawler)를 활용하여 비동기적으로 평가항목의 정보(논문 본문, 모델카드 텍스트, README, 허깅페이스 Dataset, 기업 블로그 등)를 가져와 중앙 큐로 전달한다. 모델의 신뢰성 있는 정보를 얻기 위하여 모델 논문(기술보고서), 허깅페이스, 깃허브, 기업 혹은 단체 웹사이트로 Fetcher 범위를 제한한다.

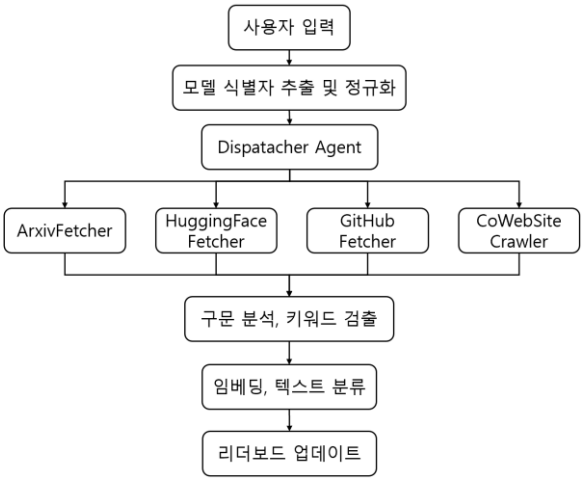
각 Fetcher 로부터 전달받은 정보에서 경량 NLP 모델(Lightweight NLP Model)을 활용하여 16 개 평가항목에 해당하는 단어 혹은 문장을 추출한다. 의존 구문 분석(Dependency Parsing)과 키워드 검출(Keyword Spotting)을 통해 논문 속 아키텍처, 학습 방법론, 데이터 공개 범위 등 문장을 해석하여 정보를 추출한 후 평가항목과 매핑을 한다.

3.3 평가 프로세스

추출한 정보들의 임베딩을 생성한 뒤 개방성 평가 프레임워크를 토대로 사전 학습된 텍스트 분류 모델을 활용하여 각 평가항목의 사후확률 (posterior probability)이 0.95 이상일 경우 개방, 0.95 미만 0.1 이상일 경우 준개방, 나머지는 비개방으로 판단한다. 코드와 가중치의 경우 Docker 컨테이너에 실제로 빌드, 로딩을 시도하여 재현가능 여부를 검증한다. 또한 개방: 1 점, 준개방: 0.5 점, 비개방: 0 점으로 지정하고 합산하여 모델의 개방성 점수를 도출한다.

3.4 리더보드 생성

평가 에이전트로부터 전달된 모델의 정보를 리더보드 데이터베이스에 저장한다. 리더보드 상에 각 모델의 개방성 점수 외에도 16 개의 평가항목의 개방의 정도, 회사명 혹은 단체명, 출시일 등을 표기하여 모델 간 비교가 용이하도록 순위를 도출한다.



(그림 1) 다이어그램.

4. 결론

본 연구에서는 AI 모델의 개방성 평가에 있어서 비효율성과 일관성부족을 해결하기 위하여, AI 에이전트 기반 자동화 평가 시스템 아키텍처를 제안하였다. 4 단계로 구성된 시스템은 각 단계를 거치며 16 개의 평가 항목의 개방성을 평가한 후 모델의 최종 개방성 점수를 도출하고 리더보드를 생성하는 과정을 자동화 하였다. 새로운 모델이 공개될 때마다 신속한 개방성 평가가 가능해진다는 장점이 있다. 하지만 개방성 평가 프레임워크의 평가항목이 대형 언어 모델에 국한되어 있다는 한계점이 존재한다. 이에 따라 이미지, 오디오 생성 모델과 같은 다양한 AI 모델의 개방성 평가 프레임워크를 확립하고, 자동화 시스템을 실제로 구현하는 과정이 필요하다.

참고문헌

[1] M. White et al., “The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence,” arXiv preprint arXiv:2403.13784, 2024

[2] Z. Liu et al., “LLM360: Towards fully transparent open-source LLMs,” arXiv preprint arXiv:2312.06550, 2023

[3] F. Eiras et al., “Risks and opportunities of open-source generative AI,” arXiv preprint arXiv:2405.08597, 2024

[4] OSI, “The Open Source AI Definition– 1.0” [Internet], https://opensource.org/ai/open-source-ai-definition

[5] OSI, “The Opensource Definition” [Internet], https://opensource.org/osd

[6] M. White et al., “The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence,” arXiv preprint arXiv:2403.13784, 2024

[7] G. W. Jeon, H. J. Han, and K. W. Lee, “Evaluating the Openness of Impactful AI Models with a Focus on LLMs”, Sejong Tech Report, 2025