

TRIPOD-LLM 지침: 대형 언어 모델 활용 보고를 위한 표적화된 가이드라인

Supplementary Table 1: NYUTron 논문에 대한 완성된 TRIPOD-LLM 체크리스트

아래는 NYUTron 논문 [1]에 대해 작성 완료한 체크리스트이다. 이 연구에 해당하는 설계 범주는 신규 LLM 개발(de novo LLM development), 평가(evaluation), 의료 환경에서의 평가(evaluation in healthcare settings)이고, 적용되는 과업 범주는 분류(classification)와 결과 예측(outcome forecasting)이다.

섹션	항목	체크리스트 항목	연구 설계	LLM 과업	족수
제목	1	해당 연구가 LLM의 개발, 미세조정, 또는 성능 평가 연구임을 명시하고, 수행한 작업, 대상 집단, 예측 결과를 구체적으로 제시	전체	전체	1
초록	2	TRIPOD-LLM for Abstracts 참고	전체	전체	1
서론	3a	의료 맥락/활용 사례(예: 행정, 진단, 치료, 임상 워크플로우)와 LLM 개발/평가의 근거 및 기존 모델·접근법 인용	전체	전체	1
	3b	대상 집단,케어 경로 내 LLM의 의도된 사용, 표준 임상 실무에서의 주요 사용자를 포함해 설명	E, H	전체	2
목적	4	연구 목적 및 LLM의 개발, 미세조정, 검증(또는 복수 단계) 포함 여부 명확히 제시	전체	전체	2
방법	5a	학습, 튜닝, 평가 데이터 세트 출처를 각각 기술하고, 사용 이유를 제시(예: 웹 코퍼스, 임상연구, EHR 등)	전체	전체	7
	5b	주요 데이터 항목과 분포 등 데이터 세트 특성 및 설명자(출처, 언어, 국가 등) 정량적·정성적 기술	전체	전체	7
	5c	개발(학습/미세조정/보상 모델링), 평가 데이터 세트 각각에서 사용한 텍스트의 가장 오래된 날짜와 최신 날짜 명시	전체	전체	7
	5d	데이터 전처리 및 품질 점검 방법과, 각 집단(텍스트 코퍼스, 기관, 사회인구학적 집단) 간의 일치 여부 기술	전체	전체	9
	5e	결측/불균형 데이터 처리 방법, 제외 사유 등을 명확히 기술	전체	전체	9
분석 방법	6a	LLM 이름, 버전, 최종 학습 날짜 보고	전체	전체	9
	6b	LLM 개발 과정(아키텍처, 학습, 미세조정, 얼라인먼트 전략[예: 강화학습 등], 얼라인먼트	M, D	전체	9

		목표) 상세 보고			
LLM 출력	6c	LLM 기반 텍스트 생성과 프롬프트 엔지니어링, 추론 세팅(시드, 온도 등) 상세 기술	M, D, E	전체	9-10
	6d	LLM의 초기 및 후처리 출력(확률, 분류, 비정형 텍스트 등) 명시	전체	전체	9-10
	6e	분류 및 확률 산출, 임계값 결정 과정 등 상세 설명	전체	C, OF	10
	7a	생성 결과 품질(일관성, 관련성, 정확성 등) 평가 지표 명시	전체	QA, IR, DG, SS, MT	10-11
	7b	배포 시점에서의 결과 지표 적합성 및 인간 평가와의 상관관계 보고	E, H	전체	10-11
주석	7c	결과 정의, LLM 예측 산출 방식(공식, 코드, API 등), 폐쇄형 LLM의 추론 날짜, 평가 지표 명확화	E, H	전체	10-11
	7d	주관적 해석 필요 시 평가자 자격, 지침, 인구통계, 평가자 간 합의 포함	전체	전체	10-11
	7e	성능 비교의 기준(LLM, 인간, 벤치마크 등) 명시	전체	전체	10-11
	8a	주석 작성 시 텍스트 라벨링 방식, 구체적 가이드라인 및 예시 기술	전체	전체	해당 없음(계산적 표현형만 사용)
	8b	주석자 수, 다중 주석 비율, 평가자 합의율 등 기술	전체	전체	해당 없음
프롬프트	8c	주석자의 배경, 경험, 라벨링 모델 특성 등 정보 제공	전체	전체	해당 없음
	9a	프롬프트 연구 시 설계, 선별, 선정 과정 상세 기술	전체	전체	해당 없음
	9b	프롬프트 개발에 사용된 데이터 명시	전체	전체	해당 없음
요약	10	요약 전 데이터 전처리 방법 기술	전체	SS	해당 없음
지시 조정/ 얼라인먼트	11	사용 시 평가에 사용된 지침, 데이터, 인터페이스, 평가 집단 특성 명시	M, D	전체	해당 없음(타겟 과업에 대한 미세조정만 적용)
연산 자원	12	사용한 연산 자원과 대체 지표(기계 수, 시간, 비용, 추론시간, FLOPs 등) 보고	M, D, E	전체	6
윤리 승인	13	연구 승인기관(윤리위원회 등) 및 동의(또는 면제) 여부 명시	전체	전체	11
오픈 사이언스	14a	연구비 출처 및 역할 명시	전체	전체	12
	14b	모든 저자의 이해상충 및 재정 공개 명시	전체	전체	12

공공 참여	14c	연구 프로토콜 위치 공개 또는 미작성했음을 명시	H	전체	해당 없음
	14d	등록 기관명과 등록 번호를 포함한 연구 등록 정보 제공 또는 등록되지 않은 연구임을 명시	H	전체	해당 없음
	14e	연구 데이터의 이용 가능성 안내	전체	전체	12
	14f	연구 결과 재현을 위한 코드 이용 가능성 안내	전체	전체	12
	15	연구 과정(설계, 수행, 해석, 결과 확산 등)에서의 환자 및 대중 참여 여부 기술 또는 미참여 사실 명시	H	전체	없음
결과	16a	환자/EHR 데이터 사용 시, 데이터 흐름 및 참여자/문서/질문별 결과 유무, 추적 기간 등 기술	E, H	전체	22
성능	16b	환자/EHR 데이터 사용 시, 전체 및 각 데이터 출처/설정/개발·평가 분할별 특성, 주요 날짜, 표본 크기 등 보고	E, H	전체	15,21,22,23
	16c	임상 결과 포함 LLM 평가 시, 주요 임상 변수 분포 비교(가능 시)	E, H	전체	4,17-20
	16d	환자/EHR 데이터 사용 시, 분석별 참여자 및 결과 발생 건수 명시	E, H	전체	2-4,7
	17	사전 지정 평가 지표(7a), 또는 인간 평가(7d)에 따른 LLM 성능 보고	전체	전체	4,13-14,16-20
LLM 업데이트	18	해당 시, LLM 업데이트 결과 및 후속 성능 보고	전체	전체	해당 없음
논의	19a	주요 결과의 해석(공정성 포함), 목표 및 선행연구 맥락에서 논의	전체	전체	5
한계	19b	연구 한계, 편향, 통계적 불확실성, 일반화 가능성 등 논의	전체	전체	6
LLM 활용성	19c	특정 과업 및 도메인에서의 데이터 활용 관련 알려진 도전(표현, 결측, 통합, 편향 등) 설명	E, H	전체	6
	19d	평가 대상 구현의 의도된 사용 목적, 입력, 최종 사용자, 자율성/인간 감독 수준 정의	E, H	전체	6
	19e	해당 시, 품질 저하 또는 결측 입력 데이터 평가 및 처리 방법, 임상 현장 내 활용성 기술	E, H	전체	6
	19f	해당 시, 사용자의 입력 데이터 처리·LLM 사용 개입 여부, 필요한 전문성 수준 명시	E, H	전체	6
	19g	후속 연구의 방향성 및 적용성, 일반화 가능성 중심으로 논의	전체	전체	6

용어:

LLM = 대형 언어 모델;

M = LLM 방법;

D = 신규 LLM 개발;

E = LLM 평가;

H = 의료환경에서의 LLM 평가;

C = 분류;

OF = 결과 예측;

QA = 장문 질의응답;

IR = 정보검색;

DG = 문서 생성;

SS = 요약 및 단순화;

MT = 기계번역;

EHR = 전자의무기록

참고:

기존 LLM 활용 연구의 경우, 원 개발자가 제공한 보고 가능한 정보에 대한 참고문헌을 포함하거나 해당 정보가 없음을 명시해야 한다.

참고문헌:

1. Jiang, L. Y. 외. Health system-scale language models are all-purpose prediction engines. Nature 619, 357–362 (2023).