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

Supplementary Table 3: TRIPOD-LLM 확장 체크리스트(설명 및 해설 요약본)

섹션	항목	체크리스트 항목(설명 포함)	연구 설계	LLM 과업
제목	1	해당 연구가 LLM 의 개발, 미세조정 및/또는 성능 평가임을 명시하고, 수행 작업, 대상 집단, 예측할 결과를 구체적으로 기술해야 한다. 정보성 있는 제목은 LLM 기반 연구의 식별 및 체계적 검토에 도움이 되므로, 대상 집단과 예측 결과에 대한 주요 정보를 반드시 제공해야 한다.	전체	전체
초록	2	TRIPOD-LLM for Abstracts 의 모든 항목을 포함하여 초록을 작성해야 한다. 체크리스트의 모든 내용을 반영해야 한다.	전체	전체
서론: 배경	3a	의료 맥락/활용 사례(예: 행정, 진단, 치료, 임상 워크플로우)와 LLM 개발 또는 평가의 근거를 설명하고, 기존 접근법 및 모델에 대한 참고문헌을 반드시 포함해야 한다. LLM 이 사용될 의료 환경이나 활용 사례를 구체적으로 서술해야 하며, 기존 접근법 또는 LLM 이 있을 경우 신규 LLM 개발의 타당한 근거를 명확히 기술해야 한다. 기존 모델을 평가하는 연구라면 평가의 근거와 모든 평가 모델의 참고문헌을 제시해야 하며, 신규 LLM 개발 및 방법론 연구에서는 잠재적 의료 맥락/활용 사례의 예시를 제공하는 것이 바람직하다.	전체	전체
	3b	대상 집단과 케어 경로 내에서 LLM 의 의도된 사용, 그리고 현재 표준 실무 내 주요 사용자를 명확히 설명해야 한다(예: 의료 전문가, 환자, 대중, 행정가 등). 개발/평가된 LLM 의 대상 집단(특정 연령, 국가, 질환 등)을 명확히 밝히고, LLM 의 의도된 목적 및 임상적 의사 결정 지원(예: 검사 의뢰, 입원 결정, 환자 메시지, 청구 등), 그리고 케어 경로 내에서의 적용 위치를 기술한다. 또한, LLM 의 최종 사용자(의료진, 환자, 대중, 기타 이해관계자)를 명확히 하고, LLM 이 대체하거나 상호작용하는 기존 표준 실무를 반드시 설명해야 한다.	E, H	전체
목적	4	연구의 목적을 명확하게 기술해야 하며, LLM 의 최초 개발, 미세조정, 또는 검증 중 어떤 단계에 대한 것인지(혹은 여러 단계를 포함하는지) 여부를 포함해야 한다. 즉, 연구가 LLM 을 새로 개발하는지, 기존 LLM 을 조정 또는 통합하는지, 또는 단순히 성능 평가를 목적으로 하는지, 혹은 이 중 복수 단계를 포함하는지를 구체적으로 밝힌다.	전체	전체
방법: 데이터	5a	학습, 튜닝, 평가 데이터 세트의 출처를 각각 명확히 기술하고, 해당 데이터를 사용한 근거(예: 웹 코퍼스, 임상시험 데이터, EHR 등)를 반드시 제시해야 한다. 즉, 데이터가 특정 웹, 무작위 임상시험, 등록 자료, EHR 등인지 명확히 밝히고, 기존 데이터 사용 시 그 적합성과 대표성을 논의하며, 합성 데이터 사용 시 생성 및 활용 방법을 상세히 기술(코드 공개 포함, 14f 참조)해야 한다.	전체	전체

	5b	관련 데이터 항목을 구체적으로 명시하고, 데이터 세트의 정량적·정성적 특성(예: 분포, 출처, 언어, 국가 등)과 기타 관련 설명자를 제공해야 한다. 예시로 데이터 분포, 사용된 언어, 국가, 의료기관 또는 기타 사회인구학적 특성을 포함할 수 있다.	전체	전체
	5c	개발(학습, 미세조정, 보상 모델링 등) 및 평가 데이터 세트에 사용된 텍스트 중 가장 오래된 날짜와 최신 날짜를 명확하게 보고해야 한다. 데이터의 시간적 범위(예: 수집/추출 시점), 또는 데이터베이스 내 포함된 텍스트의 연도 구간 등을 구체적으로 명시한다.	전체	전체
	5d	데이터 전처리 및 품질 검사 방법을 구체적으로 기술하고, 텍스트 코퍼스, 기관, 사회인구학적 집단 등 그룹 간에 동일한 처리가 적용되었는지도 명확히 해야 한다. 예시로 정제, 익명화, 언어 교정, 분류, 라벨링, 임계값 조정 등의 방법을 구체적으로 기술한다.	전체	전체
	5e	결측값 또는 불균형 데이터 처리 방법, 그리고 데이터 제외(누락) 사유를 명확히 기술해야 한다. 예시로 결측 데이터 보간, 삭제, 오버샘플링, 언더샘플링, 기준 미달 데이터의 제외 등을 구체적으로 명시한다.	전체	전체
방법: 분석 방법	6a	사용한 LLM 의 이름, 버전, 최종 학습 날짜를 반드시 보고해야 한다. 예시로 'NYUTron v0.9, 사전학습 2021 년 12 월, 미세조정 2022 년 2 월'과 같이 명확히 기술한다.	전체	전체
	6b	LLM 개발 과정(아키텍처, 학습, 미세조정, 얼라인먼트 전략[예: 강화학습, 직접 선호 최적화 등]와 그 목표[예: 유용성, 정직성, 무해성 등])의 세부사항을 구체적으로 보고해야 한다. 가능한 한 상세하게, 공개된 코드/공식 문서/참고문헌 등으로 구체적으로 명시한다.	M, D	전체
	6c	LLM 을 활용한 텍스트 생성 과정, 프롬프트 엔지니어링(출력의 일관성 포함), 추론 설정(예: 시드 값, 온도, 최대 토큰 길이, 페널티 등)을 상세히 보고해야 한다. 프롬프트 세부 설계, 반복 실험, 하이퍼파라미터, 임의성 통제 방법(시드 고정 등)을 명시한다.	M, D, E	전체
	6d	LLM 의 초기 및 후처리 출력값(예: 확률, 분류, 비정형 텍스트 등)을 명확히 보고해야 한다. 출력값의 수치화, 텍스트 후처리, 후속 변환(예: 점수로 변환, 임상지표 매핑 등) 과정을 포함한다.	전체	전체
	6e	분류 관련 세부내용 및 확률 산출 방법, 임계값 식별 과정이 있을 경우 이를 구체적으로 설명해야 한다. 예시로 확률 예측, ROC 분석, cut-off 산출 방식, 분류 기준 정의 등을 포함한다.	전체	C, OF
방법: LLM 출력	7a	생성 결과의 품질(일관성, 관련성, 정확성 등)을 평가하는 지표를 반드시 명시해야 한다. 예시로 BLEU, ROUGE, METEOR 등 자동화된 텍스트 평가 지표 및 사용자(평가자) 설문, 전문가 평가 점수 등을 포함할 수 있다.	전체	QA, IR, DG, SS, MT

	7b	배포 시점에서의 하위 과업 결과 지표의 적합성 및 해당 용도에 대해 인간 평가와의 상관관계를 반드시 보고해야 한다. 실제 적용 환경(임상, 행정 등)에서의 정량적/정성적 성능 및 인간 평가자와의 일치도, 상관계수 등을 명확히 기술한다.	E, H	전체
	7c	결과 정의, LLM 예측 산출 방식(예: 공식, 코드, 객체, API), 폐쇄형 LLM 의 추론 날짜, 평가 지표를 반드시 명확히 해야 한다. 결과가 수치, 분류, 텍스트 등 어떤 방식으로 산출되었는지, 관련 코드, 공식, API 사양, 평가 날짜(버전) 등을 명확하게 명시한다.	E, H	전체
	7d	결과 평가에 주관적 해석이 필요한 경우, 평가자의 자격, 제공된 지침, 평가자 인구통계 정보, 평가자 간 합의율 등을 반드시 포함해야 한다. 평가 기준의 명확성, 평가자의 전문가 여부, 훈련 유무, 평가 설명서 제공 여부, 평가자 간 신뢰도 등 구체적 방법을 기술한다.	전체	전체
	7e	성능 비교가 있을 경우, 기준(비교 LLM, 인간 평가자, 기타 벤치마크/표준)을 명확히 명시해야 한다. 대조군, 비교군, 기존 임상관행, 이전 모델 등과의 상대적 성능을 명확히 보고한다.	전체	전체
방법: 주석	8a	주석을 작성한 경우, 텍스트 라벨링 방식, 구체적 가이드라인 및 예시를 포함하여 상세히 기술해야 한다. 예시로 라벨링 지침서, 주석 기준, 실 예제, 합의 프로세스 등을 포함한다.	전체	전체
	8b	주석자 수, 각 데이터 세트별 다중 주석 비율, 주석자 간 합의(예: 일치율, Kappa 등)도 반드시 기술한다. 데이터별 라벨러 수, 다중 주석 비율, 주석자간 신뢰도 지표 등을 포함한다.	전체	전체
	8c	주석자(라벨러)의 배경 및 경험, 라벨링에 관여한 모델 특성 등 정보도 제공해야 한다. 라벨러의 경력, 직종, 전문성, 훈련 여부 등 구체적 프로필을 명시한다.	전체	전체
방법: 프롬프트	9a	프롬프트 관련 연구인 경우, 프롬프트 설계, 선별, 선정 과정을 상세히 기술해야 한다. 설계 프로세스, 반복 실험, 샘플 프롬프트, 선별 기준 등을 포함한다.	전체	전체
	9b	프롬프트 개발에 사용된 데이터도 명확히 밝혀야 한다. 데이터의 출처, 유형, 샘플 수, 전처리 과정 등 상세히 명시한다.	전체	전체
방법: 요약	10	요약 과업에 대해 데이터 전처리 방법을 반드시 기술해야 한다. 예시로, 원문 추출 방식, 길이 제한, 문장 분할, 중요도 판단 기준, 불필요한 정보 제거 방법 등을 상세히 명시한다.	전체	SS
방법: 지시 조정/ 얼라인먼트	11	해당 전략(지시 조정, 얼라인먼트 등)을 사용한 경우, 평가에 사용된 지침, 데이터, 인터페이스, 평가 집단 특성을 반드시 명시해야 한다. 예시로, 평가용 지시문, 인터페이스 스크린샷, 평가자 집단의 전문성/분포 등을 포함한다.	M, D	전체

방법: 연산 자원	12	연구 수행에 필요한 연산 자원(컴퓨팅 리소스) 또는 그 대체 지표(예: 사용한 기계 수, 소요 시간, 비용, 추론 시간, FLOPs 등)를 보고해야 한다. 클러스터 규모, GPU/TPU 사양, 총 연산 시간, 전력 소비, 추론 시간 등을 상세히 기술한다.	M, D, E	전체
방법: 윤리 승인	13	연구를 승인한 기관윤리위원회(IRB) 또는 연구윤리위원회(IRB)의 명칭 및 참여자 동의(혹은 동의 면제) 여부를 반드시 명시해야 한다. 예시로, "MIT COUHES IRB, 면제 ID: E-5705, 전자 동의서 확보" 등 구체적으로 보고.	전체	전체
방법: 오픈 사이언스	14a	연구 자금 출처와 연구비 제공자의 역할을 명확히 밝혀야 한다. 자금 지원 기관명, 연구 설계·수행에 미친 영향 등 포함한다.	전체	전체
	14b	모든 저자의 이해상충(Conflicts of Interest) 및 재정적 이해관계를 반드시 공개해야 한다.	전체	전체
	14c	연구 프로토콜 공개 위치를 명시하거나, 미작성 시 미작성 사실을 밝혀야 한다.	H	전체
	14d	연구 등록 정보(등록 기관, 등록 번호 등)를 명시하거나, 미등록 시 미등록 사실을 밝혀야 한다.	H	전체
	14e	연구 데이터의 이용 가능성을 상세히 안내해야 한다. 공유 여부, 접근 방법, 라이선스, 데이터 세트 위치(예: DOI, 저장소 링크 등)까지 포함한다.	전체	전체
	14f	연구 결과 재현을 위한 코드의 이용 가능성도 상세히 안내해야 한다. 예시로, "모든 코드는 GitHub 저장소(링크)에서 공개되어 있다" 등과 같이 구체적으로 기술한다.	전체	전체
방법: 공공 참여	15	연구 설계, 수행, 보고, 해석, 결과 확산 등 과정에서 환자 및 대중이 참여했는지 여부를 명확히 기술해야 한다(미참여 시에는 미참여 사실 명시). 참여 방식, 참여 인원, 의견 반영 내역, 미참여 사유 등을 구체적으로 설명한다.	H	전체
결과: 참여자	16a	환자 또는 EHR 데이터를 사용한 경우, 데이터(텍스트/EHR/환자)의 흐름, 문서·질문·참여자 수(결과 보유/미보유 구분), 추적 기간 등을 반드시 기술해야 한다. 예시로, 연구에 포함된 전체 환자 수, 문서 수, 질문 수, 각각의 결과 보유/미보유 현황, 추적 기간 등을 표 또는 도식으로 명확히 제시한다.	E, H	전체
	16b	환자 또는 EHR 데이터를 사용한 경우, 전체 및 각 데이터 출처/설정/개발·평가 데이터 분할별 특성, 주요 날짜, 표본 크기 등 주요 특성치를 반드시 보고해야 한다. 데이터 세트의 크기, 연도, 출처 기관, 개발·검증용 데이터 분할 비율 등 구체적으로 기술한다.	E, H	전체
	16c	임상 결과를 포함하여 LLM 을 평가한 경우, 개발·평가 데이터 간의 주요 임상 변수 분포를 비교하여 보고해야 한다(가능한 경우). 예시로, 인구통계, 주요 임상 변수(예: 나이, 성별, 기저 질환 등)의 분포 비교 표/그래프를 제공한다.	E, H	전체

	16d	환자 또는 EHR 데이터를 사용한 경우, 각 분석(LLM 개발, 하이퍼파라미터 튜닝, 평가 등)별로 참여자 수 및 결과 발생 건수를 명확히 기술해야 한다. 데이터 세트별 분석 적용 인원, 결과 유무별 사례 수, 하위그룹별 분포 등 구체적 수치를 명시한다.	E, H	전체
결과: 성능	17	사전 지정한 평가 지표(7a) 및/또는 인간 평가(7d)에 따라 LLM 의 성능을 반드시 보고해야 한다. 예시로, 정확도, 민감도, 특이도, F1 점수, AUROC, 사용자 만족도, 전문가 평가 점수 등을 모두 포함한다.	전체	전체
결과: LLM 업데이트	18	해당되는 경우, LLM 의 업데이트 결과와 이후 성능을 반드시 보고해야 한다. 모델 업데이트(파인 튜닝, 추가 학습, 버전 변경 등) 후 성능 변화 및 재현성 결과를 명확히 기술한다.	전체	전체
논의: 해석	19a	주요 결과의 전반적 해석, 연구 목표 및 선행 연구 맥락에서의 논의, 공정성 이슈 등을 모두 포함해야 한다. LLM 의 성능 및 의의, 기존 연구와의 비교, 실제적 함의, 잠재적 영향력 및 한계점, 공정성·편향 분석 등을 폭넓게 해석한다.	전체	전체
논의: 한계	19b	연구의 한계와 이로 인한 편향, 통계적 불확실성, 일반화 가능성에 대한 영향을 반드시 논의해야 한다. 데이터·모델·방법론·평가 도구의 한계, 오차 원인, 일반화/재현성 문제, 통계적 불확실성(신뢰구간, 표준오차 등) 등을 구체적으로 명시한다.	전체	전체
논의: 맥락 내 LLM 활용성	19c	해당되는 경우, LLM 적용 시 저품질 또는 이용 불가 입력 데이터의 평가 및 처리 방법, 실제 임상 현장에서의 LLM 활용성을 상세히 기술해야 한다. 입력 데이터의 품질 관리, 결측/노이즈/오류 데이터 대처법, 실제 임상 적용 경험과 한계 등을 모두 포함한다.	E, H	전체
	19d	평가 대상 적용 사례의 의도된 사용 목적, 입력, 최종 사용자, 자율성/인간 감독 수준을 반드시 정의해야 한다. 실제 임상/현장 적용에서 모델의 사용 의도, 입력 정보, 최종 사용자(임상의, 환자 등), 모델의 자율성과 인간 개입 수준 등을 명확히 기술한다.	E, H	전체
	19e	해당되는 경우, LLM 적용 시 저품질 또는 이용 불가 입력 데이터의 평가 및 처리 방법, 실제 임상 현장에서의 활용성을 구체적으로 설명해야 한다. 예시로, 입력 데이터 오류 감지·수정, 임상 데이터에서의 노이즈 제어, 현실적 적용 한계 등 상세히 기술한다.	E, H	전체
	19f	해당되는 경우, 사용자가 입력 데이터 처리나 LLM 사용에 관여해야 하는지 여부, 그리고 필요한 전문성 수준을 반드시 명시해야 한다. 사용자가 직접 개입하는 과정, 요구되는 전문적 역량, 학습·교육 필요성 등도 포함한다.	E, H	전체
	19g	향후 연구의 다음 단계, LLM 의 적용 가능성 및 일반화 가능성 등을 중심으로 논의해야 한다. 후속 연구 필요성, 새로운 적용 분야, 실질적 확장성, 다른 인구집단/기관/국가 등에서의 일반화 전략 등을 구체적으로 제시한다.	전체	전체

LLM = 대형 언어 모델(large language model);

M = LLM 방법(methods);

D = 신규 LLM 개발(de novo LLM development);

E = LLM 평가(evaluation);

H = 의료 환경에서의 LLM 평가(LLM evaluation in healthcare settings);

C = 분류(classification);

OF = 결과 예측(outcome forecasting);

QA = 장문 질의응답(long-form question-answering);

IR = 정보 검색(information retrieval);

DG = 문서 생성(document generation);

SS = 요약 및 단순화(summarization and simplification);

MT = 기계 번역(machine translation);

EHR = 전자의무기록(electronic health record).