

Supplementary Table 1. The TRIPOD-LLM statement: a targeted guideline for reporting large language models use					
Section	Item	Checklist item	Research design	LLM task	Page / section
Title	1	Identify the study as evaluating the performance of an LLM, specifying the task, target population, and outcome	All	All	Title
Abstract	2	See TRIPOD-LLM for Abstracts	All	All	Abstract
Introduction – Background	3a	Explain the healthcare context/use case and rationale, including references to existing models	All	All	Introduction
	3b	Describe the target population and intended use of the LLM in the care pathway and intended users	E, H	All	Introduction
Objectives	4	Specify study objectives and whether this is development, fine-tuning, or evaluation	All	All	Introduction
Methods – Data	5a	Describe data sources for training/tuning/evaluation and rationale	All	All	Methods (Study Design, Data Collection)
	5b	Describe data points and dataset characteristics (source, size, content)	All	All	Methods (Data Collection)
	5c	State oldest and newest items of text used	All	All	Not applicable (closed-source model; evaluation only)
	5d	Describe data preprocessing and quality checking	All	All	Methods (Data Collection)
	5e	Describe handling of missing/imbalanced data	All	All	Not applicable (case-based dataset with predefined labels)
Methods – Analytical Methods	6a	Report LLM name, version, and last training date	All	All	Methods (Browser-based GPT-5; API GPT-5)
	6b	Report LLM architecture, training, fine-tuning, alignment strategy	M, D	All	Not available (closed-source model)
	6c	Report text generation process, prompting, inference settings	M, D, E	All	Methods (Prompting & Hyperparameters)
	6d	Specify initial and post-processed LLM outputs	All	All	Methods (Response Evaluation)
	6e	Provide classification rationale and probability thresholds if applicable	All	C, OF	Not applicable (categorical outcomes; no probabilities)
LLM Output	7a	Include quality metrics (accuracy, consistency, relevance) vs gold standard	All	QA, IR	Results
	7b	Report relevance of metrics to downstream clinical task	E, H	All	Results
	7c	Define outcomes, inference date, evaluation metrics	E, H	All	Methods & Results
	7d	Describe assessor qualifications and inter-assessor agreement	All	All	Methods (Response Evaluation)
	7e	Specify comparison to humans, benchmarks, or other models	All	All	Results
Annotation	8a	Describe annotation process and guidelines	All	All	Not applicable (reference labels predefined)
	8b	Report number of annotators and agreement	All	All	Methods (Response Evaluation)
	8c	Describe annotator background	All	All	Methods (Study Design)
Prompting	9a	Describe prompt design, curation, selection	All	All	Methods (Prompting Process)
	9b	Report data used to develop prompts	All	All	Methods (Prompting Process)
Summarization	10	Describe preprocessing before summarization	All	SS	Not applicable
Instruction Tuning / Alignment	11	Describe instruction tuning or alignment strategies	M, D	All	Not applicable (no fine-tuning performed)
Compute	12	Report compute, cost, inference time or proxies	M, D, E	All	Methods (Compute description)
Ethical Approval	13	Name ethics committee and consent/waiver	All	All	Methods (Study Design)
Open Science	14a	Report funding source and role	All	All	Declarations
	14b	Declare conflicts of interest	All	All	Declarations
	14c	Provide protocol access or state none prepared	H	All	Methods (Study Design)
	14d	Provide registration information or state not registered	H	All	Methods (Study Design)
	14e	Provide data availability	All	All	Methods (Study Design)
	14f	Provide code availability	All	All	Not applicable (closed-source LLM; no custom code)
Public Involvement	15	Describe patient/public involvement or state none	H	All	Methods (Study Design)
Results – Participants	16a	Describe flow of cases through study	E, H	All	Results (Flowchart)
	16b	Report case characteristics and sample size	E, H	All	Results
	16c	Compare important variables between splits	E, H	All	Not applicable (single evaluation cohort)
	16d	Specify number of cases per analysis	E, H	All	Results
Performance	17	Report LLM performance per pre-specified metrics	All	All	Results
LLM Updating	18	Report results of LLM updating if applicable	All	All	Not applicable
Discussion – Interpretation	19a	Interpret results, including fairness and prior studies	All	All	Discussion
Limitations	19b	Discuss limitations, bias, uncertainty, generalizability	All	All	Limitations
Usability in Context	19c	Describe challenges with data representation and bias	E, H	All	Discussion
	19d	Define intended use, autonomy, human oversight	E, H	All	Discussion
	19e	Describe handling poor/unavailable inputs	E, H	All	Discussion
	19f	Specify user expertise required	E, H	All	Discussion
	19g	Discuss future research directions	All	All	Discussion & Conclusion
For studies using existing LLMs, users should include reference(s) to reportable information if provided by the original developers or state that this information is not available. LLM, large language model; M, LLM methods; D, <i>de novo</i> LLM development; E, LLM evaluation; H, LLM evaluation in healthcare settings; C, classification; OF, outcome forecasting; QA, long-form question-answering; IR, information retrieval; DG, document generation; SS, summarization and simplification; MT, machine translation; HER, electronic health record.					