

Supplementary Information

5.1 ProbSum Evaluation Rubric

<Accuracy>

DESCRIPTION: Does the output qualify as an official medical diagnosis according to the provided definition?

NOTE: According to the National Library of Medicine Unified Medical Language System has defined diagnosis as the following:

- o (1) MESH definition: The determination of the nature of a disease or condition, or the distinguishing of one disease or condition from another. Assessment may be made through physical examination, laboratory tests, or the likes. Computerized programs may be used to enhance the decision-making process.
- o (2) NCI: The investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation. A diagnosis is made from the study of the signs and symptoms of a disease.
- o UMLS Semantic Types: Disease or Syndrome

GRADES:

0: No

1: Yes

<\\Accuracy>

<Plausibility>

DESCRIPTION: Is the diagnosis plausible given the input?

NOTE: A diagnosis is plausible if it is not contradicted by any information in the input and would be included as a potential diagnosis in the differential diagnosis process. When answering questions based upon plausibility, \Strongly Disagree" indicates that a diagnosis is so implausible that it has the potential to cause harm, create bias, or negatively impact the patient's care. If the diagnosis is likely just incorrect but not potentially harmful then just mark \Disagree".

GRADES:

- 0: The output is not an official medical diagnosis
- 1: Strongly Disagree
- 2: Disagree
- 3: Neutral
- 4: Agree
- 5: Strongly Agree

<\\Plausibility>

<Specificity>

DESCRIPTION: How specific is the diagnosis given the input?

NOTE: A diagnosis is specific if the level of detail provided in the diagnosis. A diagnosis can be too broad, where the diagnosis ignores information from the input that would imply a diagnosis that is more granular or abstract a more granular version of the problem, or very narrow where the diagnosis is as granular as possible.

GRADES:

0: The diagnosis is not plausible

1: Very General

2: General

3: Neutral

4: Specific

5: Very Specific

<\\Specificity>

<Omission_Diag>

DESCRIPTION: What type of diagnoses are omitted from the output?

NOTE: A diagnosis is considered omitted if is not included in the list of outputted diagnoses, but would be considered by a clinician in the clinical setting based upon the input data to the LLM.

NOTE: A diagnosis is direct if it

- (1) is the primary diagnosis/problem listed for hospitalization and available in the input to the LLM
- (2) is a problem/diagnosis related to the primary signs/symptoms in the input to the LLM

NOTE: A diagnosis is indirect if it

- (1) is a complication/subsequent event or organ failure related to the primary diagnosis/problem
- (2) is another listed diagnosis/problem from the overall progress note that is not part of the primary diagnosis/problem

(3) is a diagnosis/problem that is not previously mentioned but closely related (i.e., same organ system) to the primary diagnoses/problems

GRADES:

- 1: All Direct
- 2: Majority Direct
- 3: Majority Indirect
- 4: All Indirect
- 5: No Omissions

<\\Omission_Diag>

<Uncertainty>

DESCRIPTION: Were the omissions due to incorrect utilization or missing information?

NOTE: In the case of this project an omission due to incorrect utilization uncertainty results when the model has been provided with the necessary information, but has not utilized it. The human evaluator can deduce the diagnosis but the model was not able to (i.e., inherent limitation of the model and not the input data). IF the Gold Standard contains a diagnosis that is also apparent from the input data THEN this is incorrect utilization. In the case of this project an omission due to missing information uncertainty results when the input to the model does not contain the data needed to make a diagnosis. The human evaluator would also not be able to deduce a diagnosis without more information (i.e., inherent limitation of the data input itself). IF the Gold Standard contains

a diagnosis that is NOT apparent from the input data THEN this is missing information.

NOTE: The uncertainty type can be determined by comparing the omissions, gold standards, and input to determine if the model has been given the opportunity to make the correct diagnosis

GRADES:

- 1: All Incorrect Utilization
- 2: Majority Incorrect Utilization
- 3: Majority Missing Information
- 4: All Missing Information
- 5: No Omissions

<\\Uncertainty>

<Bad_Comprehension>

DESCRIPTION: The sentence DOES contain evidence of incorrect reading comprehension

NOTE: Reading comprehension refers to the ability of the model to understand the input it has been given.

GRADES:

- 1: Strongly Disagree
- 2: Disagree
- 3: Neutral
- 4: Agree

5: Strongly Agree

<\\Bad_Comprehension>

<Bad_Recall>

DESCRIPTION: The sentence DOES contain evidence of incorrect recall of knowledge. (Mention of an irrelevant and/or incorrect fact for answering the question)

GRADES:

1: Strongly Disagree

2: Disagree

3: Neutral

4: Agree

5: Strongly Agree

<\\Bad_Recall>

<Bad_Rationale>

DESCRIPTION: The sentence DOES contain evidence of incorrect reasoning steps. (Incorrect rationale for a diagnostic choice)

GRADES:

1: Strongly Disagree

2: Disagree

3: Neutral

4: Agree

5: Strongly Agree

<\\Bad_Rationale>

<Omission_Reas>

DESCRIPTION: Did the reasoning output provide an explanation for every generated diagnosis?

GRADES:

1: Yes

0: No

<\\Omission_Reas>

5.2 TRIPOD-LLM Checklist

TRIPOD+LLM Checklist

Section / Topic	Item Number	Checklist Item	Research Design	LLM Task	Reported on Page
Abstract					
Title	2a	Identify the study as developing, fine-tuning, and/or evaluating the performance of an LLM, specifying the task, the target population, and the outcome to be predicted.	All	All	1
Abstract	2b	Provide a brief explanation of the healthcare context, use case and rationale for developing or evaluating the performance of an LLM.	E,H	All	Not Required
Objectives	2c	Specify the study objectives, including whether the study describes LLMs development, tuning, and/or evaluation	All	All	1
Methods	2d	Describe the key elements of the study setting.	All	All	1
	2e	Detail all data used in the study, specify data splits and any selective use of data.	M,D,E	All	1
	2f	Specify the name and version of LLM used.	All	All	1
	2g	Briefly summarize the LLM-building steps, including any fine-tuning, reward modeling, reinforcement learning with human feedback (RLHF), etc.	M,D	All	1
	2h	Describe the specific tasks performed by the LLMs (e.g., medical QA, summarization, extraction), highlighting key inputs and outputs used in the final LLM.	All	All	1
	2i	Specify the evaluation datasets/populations used, including the endpoint evaluated, and detail whether this information was held out during training/tuning where relevant, and what measure(s) were used to evaluate LLM performance.	All	All	1
Results	2j	Give an overall report and interpretation of the main results.	All	All	1
Discussion	2k	Explicitly state any broader implications or concerns that have arisen in light of these results.	All	All	1
Other	2l	Give the registration number and name of the registry or repository (if relevant).	H	All	Not Required
Introduction					

Section / Topic	Item Number	Checklist Item	Research Design	LLM Task	Reported on Page
Background	3a	Explain the healthcare context / use case (e.g., administrative, diagnostic, therapeutic, clinical workflow) and rationale for developing or evaluating the LLM, including references to existing approaches and models.	All	All	2
	3b	Describe the target population and the intended use of the LLM in the context of the care pathway, including its intended users in current gold standard practices (e.g., healthcare professionals, patients, public, or administrators).	E,H	All	Not Required
Objectives	4	Specify the study objectives, including whether the study describes the initial development, fine-tuning, or validation of an LLM (or multiple stages).	All	All	2
Methods					
Data	5a	Describe the sources of data separately for the training, tuning, and/or evaluation datasets and the rationale for using these data (e.g., web corpora, clinical research/trial data, EHR data).	All	All	15
	5b	Describe the relevant data points and provide a quantitative and qualitative description of their distribution and other relevant descriptors of the dataset (e.g., source, languages, countries of origin)	All	All	4
	5c	Specifically state the date of the oldest and newest item of text used in the development process (training, fine-tuning, reward modeling) and in the evaluation datasets.	M,D,E,H	All	15
	5d	Describe any data pre-processing and quality checking, including whether this was similar across text corpora, institutions, and relevant sociodemographic groups.	All	All	15
	5e	Describe how missing and imbalanced data were handled and provide reasons for omitting any data.	M,D,E	All	N/A

Section / Topic	Item Number	Checklist Item	Research Design	LLM Task	Reported on Page
Analytical Methods	6a	Report the LLM name, version, and last date of training or use during inference.	All	All	17
	6b	Specify the type of LLM architecture, and LLM building steps, including any hyperparameter tuning (e.g., temperature, length limits, penalties), prompt engineering, and any inference settings (e.g., seed, temperature, max token length) as relevant.	M,D,E	All	17
	6c	Report details of LLM development process from text input to outcome generation, such as training, fine-tuning procedures, and alignment strategy (e.g., reinforcement learning, direct preference optimization, etc.) and alignment goals (e.g., helpfulness, honesty, harmlessness, etc.).	M,D	All	17
	6d	Specify the initial and post-processed output of the LLM (e.g., probabilities, classification, unstructured text).	All	All	17
	6e	Provide details and rationale for any classification and how the probabilities were determined and thresholds identified.	All	C,OF	Not Required
	6f	Include metrics that capture the quality of generative outputs, such as consistency, relevance, and accuracy, compared to gold standards.	All	QA,IR,DG,SS,MT	20
	6g	Report the outcome metrics' relevance to downstream task at deployment time and correlation of metric to human evaluation of the text for the intended use.	E,H	All	Not Required
LLM Output	7a	Clearly define the outcome, how the LLM predictions were calculated (e.g., formula, code, object, API), and evaluation metrics.	E,H	All	Not Required
	7b	If outcome assessment requires subjective interpretation, describe the qualifications of the assessors, any instructions provided, relevant information on demographics of the assessors, and inter-assessor agreement.	All	All	N/A
	7c	Specify how performance was compared to other LLMs, humans, and other benchmarks or standards.	All	All	20

Section / Topic	Item Number	Checklist Item	Research Design	LLM Task	Reported on Page
Annotation	8a	If annotation was done, report how text was labeled, including providing specific annotation guidelines with examples.	All	All	N/A
	8b	If annotation was done, report how many annotators labeled the dataset(s), including the proportion of data in each dataset that were annotated by more than 1 annotator.	All	All	N/A
	8c	If annotation was done, provide information on the background and experience of the annotators, and the inter-annotator agreement.	All	All	N/A
Prompting	9a	If research involved prompting LLMs, provide details on the processes used during prompt design, curation, and selection.	All	All	17;19
	9b	If research involved prompting LLMs, report what data were used to develop the prompts.	All	All	17;19
Summarization	10	Describe any preprocessing of the data before summarization.	All	SS	N/A
Instruction Tuning / Alignment	11	If instruction tuning/alignment strategies were used, what were the instructions and interface used for evaluation, and what were the characteristics of the populations doing evaluation?	M,D	All	17
Compute	12	Report compute, or proxies thereof (e.g., time on what and how many machines, cost on what and how many machines, inference time, floating-point operations per second (FLOPs)), required to carry out methods.	M,D,E	All	9
Ethics Approval	13	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent.	All	All	15

Section / Topic	Item Number	Checklist Item	Research Design	LLM Task	Reported on Page
Open Science	14a	Give the source of funding and the role of the funders for the present study.	All	All	22
	14b	Declare any conflicts of interest and financial disclosures for all authors.	All	All	22
	14c	Indicate where the study protocol can be accessed or state that a protocol was not prepared.	H	All	Not Required
	14d	Provide registration information for the study, including register name and registration number, or state that the study was not registered.	H	All	Not Required
	14e	Provide details of the availability of the study data.	All	All	22
	14f	Provide details of the availability of the code to reproduce the study results.	All	All	11
Public Involvement	15	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	H	All	Not Required
Results					
Participants	16a	When using patient/EHR data, describe the flow of text/EHR/patient data through the study, including the number of documents/questions/participants with and without the outcome/label and follow-up time.	E,H	All	Not Required
	16b	When using patient/EHR data, report the characteristics overall and, for each data source or setting, and for development/evaluation splits, including the key dates, key predictors, and sample size.	E,H	All	Not Required
	16c	For LLM evaluation, show a comparison of the distribution of important predictors between development and evaluation data.	E,H	All	Not Required
	16d	When using patient/EHR data, specify the number of participants and outcome events in each analysis (e.g., for LLM development, hyperparameter tuning, LLM evaluation).	E,H	All	Not Required
Performance	17	Report LLM performance according to pre-specified metrics (see item 7a) and/or human evaluation (see item 7d).	All	All	7
LLM Updating	18	If applicable, report the results from any LLM updating, including the updated LLM and subsequent performance.	All	All	N/A
Discussion					

Section / Topic	Item Number	Checklist Item	Research Design	LLM Task	Reported on Page
Interpretation	19a	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies.	All	All	11
Limitations	19b	Discuss any limitations of the study and their effects on any biases, statistical uncertainty, and generalizability.	All	All	11
Usability of the LLM in context	19c	Describe any known challenges in using data for the specified task and domain context with reference to representation, missingness, harmonization, and bias.	E,H	All	Not Required
	19d	Define the intended use for the implementation under evaluation, including the intended input, end-user, level of autonomy/human oversight.	E,H	All	Not Required
	19e	If applicable, describe how poor quality or unavailable input data should be assessed and handled when implementing the LLM, i.e., what is the usability of the LLM in the context of current clinical care.	E,H	All	Not Required
	19f	If applicable, specify whether users will be required to interact in the handling of the input data or use of the LLM, and what level of expertise is required of users.	E,H	All	Not Required
	19g	Discuss any next steps for future research, with a specific view to applicability and generalizability of the LLM.	All	All	11