

## Supplementary Material 2: AD-GPT Checklist

| Section             | Item | Checklist Item                                                                                                                                                                                                                                                           | Research Design | LLM Task | Page |
|---------------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|----------|------|
| <b>Title</b>        |      |                                                                                                                                                                                                                                                                          |                 |          |      |
| Title               | 1    | 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    |
| <b>Abstract</b>     |      |                                                                                                                                                                                                                                                                          |                 |          |      |
| Abstract            | 2    | See TRIPOD-LLM for Abstracts                                                                                                                                                                                                                                             | All             | All      | 1–2  |
| <b>Introduction</b> |      |                                                                                                                                                                                                                                                                          |                 |          |      |
| 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      | 2–4  |
|                     | 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      | 5    |
| 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      | 5    |
| <b>Methods</b>      |      |                                                                                                                                                                                                                                                                          |                 |          |      |
| 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, or unknown).                                                             | All             | All      | 5–7  |
|                     | 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      | 5–7  |
|                     | 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.                                                                                               | All             | All      | 5–7  |
|                     | 5d   | Describe any data pre-processing and quality checking, including whether this was similar across text corpora, institutions, and relevant socio-demographic groups.                                                                                                      | All             | All      | 7–8  |
|                     | 5e   | Describe how missing and imbalanced data were handled and provide reasons for omitting any data.                                                                                                                                                                         | All             | All      | N/A  |
| Analytical Methods  | 6a   | Report the LLM name, version, and last date of training.                                                                                                                                                                                                                 | All             | All      | N/A  |
|                     | 6b   | Report details of LLM development process, such as LLM architecture, 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      | 8–10 |

| Section                        | Item | Checklist Item                                                                                                                                                                                                           | Research Design | LLM Task           | Page   |
|--------------------------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|--------------------|--------|
|                                | 6c   | Report details of how text was generated using the LLM, including any prompt engineering (including consistency of outputs), and inference settings (e.g., seed, temperature, max token length, penalties), as relevant. | M, D, E         | All                | Suppl. |
|                                | 6d   | Specify the initial and post-processed output of the LLM (e.g., probabilities, classification, unstructured text).                                                                                                       | All             | All                | N/A    |
|                                | 6e   | Provide details and rationale for any classification and, if applicable, how the probabilities were determined and thresholds identified.                                                                                | All             | C, OF              | N/A    |
| LLM Output                     | 7a   | 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 | 10-14  |
|                                | 7b   | Report the outcome metrics' relevance to downstream task at deployment time and, where applicable, correlation of metric to human evaluation of the text for the intended use.                                           | E, H            | All                | 10-14  |
|                                | 7c   | Clearly define the outcome, how the LLM predictions were calculated (e.g., formula, code, object, API), the date of inference for closed-source LLMs, and evaluation metrics.                                            | E, H            | All                | 10-14  |
|                                | 7d   | 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                | 10-14  |
|                                | 7e   | Specify how performance was compared to other LLMs, humans, and other benchmarks or standards.                                                                                                                           | All             | All                | 14–21  |
| Annotation                     | 8a   | If annotation was done, report how text was labeled, including providing specific annotation guidelines with examples.                                                                                                   | All             | All                | 13-14  |
|                                | 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, and the inter-annotator agreement.             | All             | All                | 13-14  |
|                                | 8c   | If annotation was done, provide information on the background and experience of the annotators or characteristics of any models involved in labelling.                                                                   | All             | All                | 13-14  |
| Prompting                      | 9a   | If research involved prompting LLMs, provide details on the processes used during prompt design, curation, and selection.                                                                                                | All             | All                | 7–8    |
|                                | 9b   | If research involved prompting LLMs, report what data were used to develop the prompts.                                                                                                                                  | All             | All                | 7–8    |
| Summarization                  | 10   | Describe any preprocessing of the data before summarization.                                                                                                                                                             | All             | SS                 | Suppl  |
| Instruction tuning / Alignment | 11   | If instruction tuning/alignment strategies were used, what were the instructions, data, and interface used for evaluation, and what were the characteristics of the populations doing evaluation?                        | M, D            | All                | 7–8    |
| 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)),                                         | M, D, E         | All                | Suppl. |

| Section            | Item | Checklist Item                                                                                                                                                                                                           | Research Design | LLM Task | Page  |
|--------------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|----------|-------|
|                    |      | required to carry out methods.                                                                                                                                                                                           |                 |          |       |
| Ethical 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      | N/A   |
| Open Science       | 14a  | Give the source of funding and the role of the funders for the present study.                                                                                                                                            | All             | All      | 25    |
|                    | 14b  | Declare any conflicts of interest and financial disclosures for all authors.                                                                                                                                             | All             | All      | 25    |
|                    | 14c  | Indicate where the study protocol can be accessed or state that a protocol was not prepared.                                                                                                                             | H               | All      | N/A   |
|                    | 14d  | Provide registration information for the study, including register name and registration number, or state that the study was not registered.                                                                             | H               | All      | N/A   |
|                    | 14e  | Provide details of the availability of the study data.                                                                                                                                                                   | All             | All      | 25    |
|                    | 14f  | Provide details of the availability of the code to reproduce the study results.                                                                                                                                          | All             | All      | N/A   |
| 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      | N/A   |
| <b>Results</b>     |      |                                                                                                                                                                                                                          |                 |          |       |
| 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 as applicable. | E, H            | All      | N/A   |
|                    | 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 characteristics, and sample size.              | E, H            | All      | N/A   |
|                    | 16c  | For LLM evaluation that include clinical outcomes, show a comparison of the distribution of important clinical variables that may be associated with the outcome between development and evaluation data, if available.  | E, H            | All      | N/A   |
|                    | 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      | N/A   |
| Performance        | 17   | Report LLM performance according to pre-specified metrics (see item 7a) and/or human evaluation (see item 7d).                                                                                                           | All             | All      | 15-19 |
| LLM Updating       | 18   | If applicable, report the results from any LLM updating, including the updated LLM and subsequent performance.                                                                                                           | All             | All      | N/A   |
| <b>Discussion</b>  |      |                                                                                                                                                                                                                          |                 |          |       |
| 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      | 20-21 |

| Section                         | Item | Checklist Item                                                                                                                                                                                               | Research Design | LLM Task | Page  |
|---------------------------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|----------|-------|
| Limitations                     | 19b  | Discuss any limitations of the study and their effects on any biases, statistical uncertainty, and generalizability.                                                                                         | All             | All      | 21-23 |
| 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      | 21-22 |
|                                 | 19d  | Define the intended use for the implementation under evaluation, including the intended input, end-user, level of autonomy/human oversight.                                                                  | E, H            | All      | 21-23 |
|                                 | 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      | 21-23 |
|                                 | 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      | 21-23 |
|                                 | 19g  | Discuss any next steps for future research, with a specific view to applicability and generalizability of the LLM.                                                                                           | All             | All      | 23-24 |

Abbreviations: LLM = large language model; M = LLM methods; D = de novo 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; EHR = electronic health record.

Note: 'N/A' indicates items not applicable to this study. 'Suppl.' indicates information provided in the Supplementary Material 1: AD-GPT Implementation Details.