**Checklist for generative AI clinical studies**

|  |  |  |
| --- | --- | --- |
| **Before paper submission** | | |
| **Study design (Part 1)** | **Page number(s), if completed** | **Notes if not completed** |
| The clinical problem in which the model will be employed is clearly detailed in the paper. |  |  |
| The research question is clearly stated. |  |  |
| All cohort selection criteria and study design are detailed using precise, unambiguous language. |  |  |
| The characteristics of the cohorts are detailed in the text and are shown to be representative of real-world clinical settings. |  |  |
| Details on how labels were generated are described, including any annotation guidelines, level of experience of annotators, inter-annotator scores, etc. |  |  |
| Which step(s) have been taken to understand model biases, privacy and security concerns, and other potential harm? | ☐ Discussion  ☐ Identification  ☐ Mitigation |  |
| **Data and optimization (Part 2)** | **Page number(s) if completed** | **Notes if not completed** |
| The origin of the data is described and the original format is detailed in the paper. |  |  |
| All data preprocessing for model training or inference is described, including appropriate randomization and other transformations. |  |  |
| The independence between training, validation (including prompt evaluation), and test sets has been proven in the paper, and data is split at the patient level. |  |  |
| Details on the models that were evaluated and the code developed to select the best model are provided, including any prompt development or evaluation techniques. |  |  |
| Details on post-processing for model outputs should be detailed. |  |  |
| Is the output data type categorical, continuous, or unstructured? | ☐ Categorical  ☐ Continuous  ☐ Unstructured | |
| **Model performance and evaluation (Parts 3-4)** | **Page number(s) if completed** | **Notes if not completed** |
| The state-of-the-art solution used as a baseline for comparison has been identified and detailed. Both generative and non-generative approaches are considered. |  |  |
| The performance comparison between the baseline and the proposed model is presented with the appropriate statistical significance. |  |  |
| Identify which type(s) of evaluations were performed, and provide clear justifications for the primary metrics used for each evaluation. | ☐Overlap accuracy  ☐Semantic accuracy  ☐Clinical utility |  |
| If applicable, details on human evaluation are described, including any evaluation guidelines, level of experience of evaluators, inter-reviewer scores, etc. |  |  |
| **Model examination (Part 5)** | **Page number(s) if completed** | **Notes if not completed** |
| Relevant interpretability techniquesa, error analysis, and/or other approaches are applied to understand factors contributing to model behavior. |  |  |
| A discussion and/or assessment of the reliability and robustness of the model as the underlying data distribution shifts is included. |  |  |
| **Reproducibility (Part 6)** | **Page number(s) if completed** | **Notes** |
| **Data transparency: choose appropriate tier of transparency** | | |
| Tier 1: complete sharing of the code and data |  |  |
| Tier 2A: complete sharing of the code with synthetic data provided |  |  |
| Tier 2B: complete sharing of the code |  |  |
| Tier 3: no sharing of code or data |  |  |
| **Model transparency** | | |
| Model hyperparameters, along with infrastructure and compute requirements for running or developing the model are included, specifying hardware type, costs, and training time where applicable. |  |  |
| If applicable: Model cards detailing capabilities, intended use, training data, limitations, potential biases, and risks are provided. |  |  |

a Common examination approaches based on study type: for studies involving exclusively structured data, coefficients and sensitivity analysis are often appropriate; for studies involving unstructured data in the domains of image analysis or natural language processing, saliency maps (or equivalents) and sensitivity analyses are often appropriate.