Appendix 1. Candidate CHART Checklist Items.

Heading	Subheading	Candidate CHART Checklist Item
General		
	Disclosures	Lists any relevant author disclosures or conflicts of interest.
		Describes the process used to identify conflicts of interest.
		Describes how relevant conflicts of interest were handled, as well as their impact in the conduct and reporting of the study, if applicable.
		Reports sources of funding.
		Reports whether ethical approval was needed.
Title		Clearly states that the study is assessing one or more NLP-based chatbots.
Abstract		Applies a structured format (eg: background, methods, results, conclusion).
		States hypothesis and/or purpose.
Introduction		Explains the scientific background and rationale for the study.
		References relevant literature, including any previous applications of the chatbot(s) for this study topic, if applicable.
		Identifies a knowledge gap regarding the performance of the chatbot(s) in relation to the study topic.
Objectives & Aims		Primary and secondary objectives are explicitly stated.
		Objectives may appear in the introduction or methods section, depending on journal guidance.
		Objectives include: → Audience (who are the responses relevant for). → Topic → Intervention. → Comparator, when applicable. → Outcome(s) of interest.
		One of the following aims are explicitly identified: → Health prevention → Screening → Differential diagnosis → Diagnosis → Treatment

		→ General information
Methods		
	Query Strategy	Identifies the number of team members involved in querying the LLM(s)/chatbot(s).
		Identifies the LLM(s)/chatbot(s) being queried.
		Identifies whether the LLM(s)/chatbot(s) being queried are accessible or closed proprietary models.
		Identifies the version identifier(s) of LLM(s)/chatbot(s) being queried, as well as their date of release.
		Identifies the date of release or last update of the LLM(s)/chatbot(s) being queried.
		Briefly explains the rationale for querying the selected LLM(s)/chatbot(s).
		Describes the LLM characteristics, including: → Temperature → Token length → Fine-tuning availability → Penalties → Add-on availability → Date accessed → LLM version → Language → Layers
	Prompt Engineering	Declares whether prompts were developed prior to formal query initiation.
		Elaborates on the evolution of study prompt development.
		Describes whether test prompts were used to guide prompt development.
		Identifies barriers to chatbot output encountered during the development of study prompts (e.g.: legal disclaimers).
		Outlines follow-up prompts used, if applicable (e.g.: to handle legal disclaimers, to trigger the LLM(s)/chatbot(s) to make a decision/choice, etc).
		Reports whether prompts were reviewed for grammatical accuracy.
		Describes whether approaches were applied to mitigate biased responses from the LLM(s)/chatbot(s) (e.g.: not mentioning a specific society or organizations in the prompts).
		Describes the number of team members involved in prompt development.
		Describes the sources of prompts, including:

		→ Expert opinion → Clinical practice guidelines → Professional society website → Social media (patient questions) → Textbook → Website (non-evidence-based) → Website/forums (patient questions)
	Formal Query	For closed proprietary models, identifies the date(s) of queries for the LLM(s)/chatbot(s) including the day, month, and year.
		For closed proprietary models, reports the location(s) where the LLM(s)/chatbot(s) was/were queried including city and country.
		If the LLM(s)/chatbot(s) was/were queried in multiple locations, explicitly outlines any differences in responses and performances, if present, of the LLM(s)/chatbot(s) across locations.
		Reports whether standardized prompts were used including follow-up prompts, if applicable.
		If follow-up prompts are included, describes the specific circumstances in which they were used.
		Clearly outlines any post-hoc changes to prompts after formal query initiation and provides a rationale.
		Provides prompts in the manuscript body.
		Describes whether prompts were inputted into one or more chat(s) with no prior prompts (chat instance).
		Describes whether prompts were inputted into one or more chat(s) with prior prompts (chat session).

		Where a comparator group is included (e.g.: other LLM or clinician), reports whether queries conducted for intervention and comparator groups were inputted in separate chat windows.
	Performance	Provides responses in manuscript body or appendix, if LLM(s)/chatbot(s) allow(s).
	Evidence	If LLM/chatbot-provided evidence was critically appraised, describes explicitly and in detail what is meant by the performance of the bot regarding the quality of the evidence which it summarizes.
		If LLM/chatbot-provided evidence was critically appraised, describes the strategy for determining the performance of the bot, including criteria for determining a successful performance.
		If LLM/chatbot-provided evidence was critically appraised, describes the process of critical appraisal, including any aspects that were conducted by more than one evaluator.
		Reports whether evaluators were aware of the LLM/chatbot being assessed during chatbot performance evaluation for evidence appraisal.
	Advice	If clinical advice was evaluated, describes explicitly and in detail what is meant by the performance of the bot regarding the clinical advice which it provides.
		If clinical advice was evaluated, describes explicitly and in detail the strategy for determining the performance of the bot, including criteria for determining a successful performance. E.g.: → Clinical practice guidelines (evidence-based) → Clinical practice guidelines (non-evidence-based) → Systematic review (evidence-based)

		→ Systematic review (non-evidence-based) → Traditional textbook → Electronic compendium → Organization website → Investigators Without Reference to Source → Investigator Panel → Primary article (RCT, cohort study, etc.)
		If clinical advice was evaluated, describes the process of evaluation, including any aspects that were conducted by more than one evaluator.
		Reports whether evaluators were aware of the chatbot being assessed during chatbot performance evaluation for clinical advice.
	Data Analysis	Reports statistical analysis methods.
		Reports which statistical performance measures and evaluation criteria were adopted to determine successful performance. → Prognostic/diagnostic: F1 score, sensitivity (recall), specificity, AUC, PPV (precision) ?Combine with above (how to define performance) and this is how you calculate performance
		Reports how data included in the query were handled (transformation, rescaling, standardisation.
		Reports whether data were partitioned or altered.
		Reports a justification for the sample size determination as appropriate .
	Check Query	Once performance evaluation is complete, authors explicitly state whether check queries were needed and explain their rationale for doing so for the applicable LLM(s)/chatbot(s) based on version updates, or continuous learning of the LLM(s)/chatbot(s).
		Explicitly states that for relevant LLM(s)/chatbots that have been updated through new versions or technological improvement, whether check queries were conducted following data analysis to confirm the consistency of findings.
		Identifies the date(s) of check queries for the LLM/chatbot(s) including tthe day, month, and year.
		Reports the location(s) where the LLM(s)/chatbot(s) was/were queried including city and country.
		If the LLM(s)/chatbot(s) was/were queried in multiple locations, explicitly outlines differences in responses and performances of the chatbot(s) across locations.
Results		
	Flow Diagram	States whether a Flow Diagram is included.
		If included, identifies the LLM(s)/chatbot(s) evaluated, the use of accessible vs inaccessible/proprietary models, the number of test prompts used during prompt development, the number of prompts and follow-up prompts used during formal query initiation, the number of unanswered prompts, and the number of discrepancies following the check query.
		Reports findings of the study.

	Missing Data	Reports instances where the chatbot(s) did not answer prompts.
Discussion		
		Reports main findings of the study.
		Addresses relevant literature, including any previous assessments of the performance of the chatbot(s) for this study topic.
		When applicable for the interpretation of results, explains the basic functioning of the chatbot(s) including the process in which it synthesizes information.
		Reports strengths of the study.
		Reports limitations of the study.
		Reports relation to prior work.
		Reports implications for clinical practice and research
Additional		
		Reports ethical considerations of the clinical integration of the LLM(s)/chatbot(s) being evaluated.
		Reports regulatory considerations of the clinical integration of the LLM(s)/chatbot(s) being evaluated.

Case study would be single case