Supplementary Information
(A)
Evaluation Dimensions, Metrics, and Examples for AI Agents in Healthcare

Indicator Dimension	Primary Indicator	Secondary Indicator
Basic Indicators	Objective correctness	Accuracy:It refers to the proportion of correctly classified samples to the total number of samples.
		Precision:Also known as positive predictive value, it is the proportion of samples correctly predicted as positive by the model to all samples predicted as positive.
		Recall:Also known as sensitivity, it is the proportion of samples correctly predicted as positive by the model to the actual positive samples.
		F1 - Score:It is the harmonic mean of precision and recall, used to comprehensively evaluate the performance of the model.
		ROC:The ROC curve is drawn with the false positive rate (False Positive Rate, FPR) as the horizontal coordinate and the true positive rate (True Positive Rate, TPR) as the vertical coordinate.
		AUC: The area under the ROC curve.
	Semantic correctness	BLEU/GLEU:Evaluate the overlap of n - grams (continuous sequences of n words or characters).
		METEOR:Considers synonyms and stems to calculate the mapping relationship between generated text and reference text to measure similarity.
		BERTScore:Input text into the BERT model to obtain word vector representations, then calculate cosine similarity and other indicators between words to measure text similarity.
		ROUGE:Calculate the overlap of units (such as words, n - grams, sentences, etc.) in the generated text with the reference text to measure the coverage of the generated text on the reference text.
Developmental Indicators	Task completion	Completion rate:The percentage of completed checklist items to the total number of items. An ideal system should complete all items on the list.
		Success rate:An indicator to evaluate whether the task is successfully completed to achieve the expected goal.
		Tool use: The use of tools is integral to completing specific tasks within AI agents. In this dimension, performance in selecting, activating, and utilizing relevant tools is sometimes evaluated as part of the overall task performance.
	Efficiency level	Response time: The time elapsed from when a user makes a request (such as asking a question, requesting a diagnosis, etc.) to the agent (such as a large - language - model agent in the medical field) to when the agent gives the first response.
		Interaction rounds: The number of times of information interaction between the user and the agent in a complete communication (such as medical consultation, diagnostic process, etc.).
	Content and presentation level	Content richness: Measure the diversity and depth of details, viewpoints, elements, etc. included in the information, which can be scored using a scale.
		Content detail: The degree of detail information included when describing a certain topic or object, which can be scored using a scale.
		Content usefulness: The actual value of the content for the audience in achieving specific goals, meeting specific needs, or obtaining valuable knowledge, which can be scored using a scale.
		Content safety: The degree to which the output content will not cause physical or psychological harm to users, and will not spread incorrect, harmful or misleading medical information, which can be scored using a scale.
		Ethical compliance: Evaluate the appropriateness of the medical agent in providing medical - related solutions in terms of morality and medical ethics, which can be scored using a scale.
		Readability: Quantified by the Flesch - Kincaid Grade Level, which combines the average sentence length, syllable count, and number of words per sentence. It can be obtained by importing a readability module.
		Reading ease: Assessed by sentence length and syllable count per word, with a score of 206.835 - (1.015 × ASL) - (84.6 × ASW).
		Content clarity: The degree to which information can be conveyed clearly, accurately, and unambiguously to the audience, which can be scored using a scale.

Humanistic care	Language coherence: The degree to which sentences and paragraphs in a text are connected and transition naturally, forming an organic whole, which can be scored using a scale. Accuracy under latent symptoms: The probability of correctly diagnosing diseases or providing correct medical advice when facing diseases with hidden and atypical symptoms. Human care: The degree of attention and respect for people's living conditions, dignity, rights, and values, which can be scored using a scale. Confidence: Confidence in this agent, which can be scored using a scale. Compliance: The likelihood of following up with treatment according to the diagnosis, which can be scored using a scale. Consultation: The likelihood of consulting this agent again, which can be scored using a scale. Satisfaction: A comprehensive and systematic measure of the user's experience and recognition of the agent's performance during the interaction process, which can be scored using a scale.

(B)

A structured research agenda of LLM-based AI agents in healthcare

Direction	Background and Significance	The current situation or predicament	Future Outlook
Integration with Embodied Robots	• The aging population is intensifying and there is a shortage of global healthcare resources • Medical operations are shifting from purely manual to a human-machine collaborative paradigm	• Breakthroughs have been made in specific medical fields • Large-scale application still faces challenges	Explore more personalized and humanized intelligent agents to interact with users
Hybrid Expert Model Combination	The requirements for the processing capabilities of complex medical data have increased	• Strong generalization ability • It is difficult to support precise decision support	Introduce a hybrid expert system
Expansion of Evaluation Metrics	The importance of the standardized application of AI agents in the medical field	Limited to traditional technical indicators	The evaluation framework has been expanded to multi-dimensional factors
Safety and Risk Management	The rapid evolution of autonomy of medical AI agents	• Insufficient clinical trust • The security risk control mechanism is lacking	Technical safety guidelines and regulatory framework
Moral and ethical review	Ethical issues have become a key constraint	The guidelines are constantly being explored and	Establish an ethical review and supervisory responsibility mechanism

		improved • It is difficult to balance the interests of multiple parties	
User trust and feedback adoption	There is still considerable room for improvement in its promotion and application	• The lack of an effective feedback loop • The users' doubts	Take into account user feedback and the demands of multiple parties
Career Development of Medical Staff	The development of agents has a significant impact on the careers of medical staff	Vocational skills assessment and Anxiety	Constantly explore new models of human-machine collaboration

(C)
Stakeholder Value Mapping of AI Agent Applications

Stakeholder	AI Agent Functionality	Value/Outcome
Developers	Hybrid expert models	Improved accuracy & interpretability in specialized tasks (e.g., oncology diagnostics).
Hospital Administrators	EHR-management agents (e.g., EHRAgent)	Reduced clinician burnout; 30% faster documentation (Shi et al. [64]).
Policymakers	Safety/ethics frameworks	Compliance with EU AI Act "high-risk" standards; accountable governance.
Clinicians	Multi-agent diagnostic debate	80% higher accuracy in complex cases (Ke et al. [30]).
Medical Educators	Simulated patient agents (e.g., MEDCO)	Enhanced clinical reasoning skills via realistic case training.

(D)
Literature Screening Flow Diagram