American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Mini Review

Offline and Online Evaluation of Agentic AI Frameworks for Healthcare: Challenges, Current Practices, and Future Directions

Neha Gupta, Bhawna Singla*

School of Computer Science and Engineering, Geeta University, India

***Corresponding author:** Bhawna Singla, School of Computer Science and Engineering, Geeta University, Panipat, India.

To Cite This article: *Neha Gupta, Bhawna Singla*, Offline and Online Evaluation of Agentic AI Frameworks for Healthcare: Challenges, Current Practices, and Future Directions. Am J Biomed Sci & Res. 2026 31(5) AJBSR.MS.ID.004080,* **DOI:** 10.34297/AJBSR.2026.31.004080

Received: July 12, 2026 **Published:** July 17, 2026

Abstract

Agentic Artificial Intelligence (AI) systems are transforming healthcare by autonomously planning, reasoning, retrieving information, invoking external tools, and executing multi-step clinical workflows. Unlike conventional large language models that primarily generate text, agentic AI systems interact continuously with electronic health records, clinical knowledge repositories, and healthcare information exchanges. Consequently, evaluating these systems requires methodologies that extend beyond traditional natural language processing metrics. Current evaluation approaches generally fall into two complementary categories: offline evaluation and online evaluation. Offline evaluation measures an agent's capability using predefined benchmark datasets, synthetic scenarios, and reproducible task collections under controlled conditions, whereas online evaluation assesses real-time performance during live deployment by monitoring reliability, safety, latency, user satisfaction, and clinical outcomes. This mini-review summarizes recent advances in both evaluation paradigms, discusses their advantages and limitations, and highlights the growing need for standardized healthcare-specific evaluation frameworks. The review argues that comprehensive assessment of agentic healthcare systems requires combining reproducible offline benchmarking with continuous online monitoring to ensure trustworthy, safe, and clinically reliable deployment.

Introduction

The rapid evolution of Large Language Models (LLMs) has shifted artificial intelligence research from passive question-answering systems toward autonomous intelligent agents capable of planning, tool utilization, memory management, and sequential decision making. These agentic systems execute complex workflows by decomposing user objectives into multiple reasoning steps while interacting with external databases, APIs, software tools, and domain-specific knowledge sources. Healthcare has emerged as one of the most promising application domains for agentic AI because modern clinical workflows naturally involve multiple interconnected tasks. A healthcare agent may retrieve longitudinal medical histories, verify patient identity, obtain consent, summarize previous diagnoses, coordinate referrals, exchange clinical records, and generate treatment summaries before responding to clinicians or patients. Such workflows are considerably more complex than conventional question-answering tasks and therefore require more comprehensive evaluation methodologies.

Traditional language model evaluation metrics, including BLEU, ROUGE, METEOR, and semantic similarity scores, primarily measure textual quality and are insufficient for assessing autonomous healthcare agents. Agentic systems must additionally satisfy requirements related to workflow completion, reasoning consistency, regulatory compliance, patient privacy, tool utilization, and clinical safety. Consequently, the research community has increasingly focused on two complementary evaluation paradigms. Offline evaluation provides reproducible benchmarking under controlled

experimental settings, whereas online evaluation measures agent behavior in real-world operational environments. Together, these paradigms provide a comprehensive understanding of system performance and reliability.

Offline Evaluation of Agentic AI Systems

Offline evaluation refers to assessing an agent using predefined datasets, benchmark tasks, and controlled experimental protocols without interacting with real users. Since every task possesses known inputs and expected outputs, results are fully reproducible and easily comparable across different agent architectures.

Recent benchmarks such as AgentBench, ToolBench, GAIA, WebArena, SWE-bench, and MedHELM evaluate various aspects of autonomous reasoning including planning ability, tool usage, software engineering, web navigation, and medical question answering. These benchmarks have significantly improved standardized evaluation for general-purpose agents. Healthcare introduces additional complexity because clinical workflows involve structured electronic health records, longitudinal patient histories, interoperability standards such as HL7 FHIR, consent management, and privacy regulations. Consequently, offline healthcare benchmarks increasingly employ synthetic patient datasets, simulated clinical scenarios, and predefined workflow objectives that allow repeatable experimentation without exposing sensitive patient information.

Typical offline evaluation metrics include:

a) Task completion rate

b) Workflow accuracy

c) Tool invocation correctness

d) Compliance score

e) Clinical reasoning accuracy

f) Information retrieval accuracy

g) Hallucination rate

h) Execution latency

i) Cost per completed task

Offline evaluation offers several important advantages. Experiments are reproducible, relatively inexpensive, and completely safe because no real patients are involved. Researchers can evaluate thousands of clinical scenarios under identical conditions while systematically comparing different prompting strategies, reasoning architectures, retrieval systems, and orchestration frameworks.

Nevertheless, offline benchmarks remain approximations of real clinical practice. Synthetic scenarios cannot fully represent the uncertainty, ambiguity, interruptions, and incomplete information encountered during routine healthcare delivery.

Online Evaluation of Agentic AI Systems

Online evaluation measures system performance after deployment by continuously monitoring interactions between users and the agent during real operational use. Unlike offline benchmarking, online evaluation captures dynamic environmental conditions including changing user behavior, evolving clinical workflows, network delays, tool failures, and unexpected inputs. Healthcare agents operating within hospital information systems or digital health platforms must satisfy stringent requirements concerning reliability, patient safety, regulatory compliance, and operational efficiency. Consequently, online evaluation extends beyond technical accuracy to include system robustness and human-centered outcomes.

Common online evaluation metrics include:

a) Response latency

b) System availability

c) Workflow success rate

d) User satisfaction

e) Clinical acceptance

f) Escalation frequency

g) Error recovery capability

h) Safety incidents

i) Privacy violations

j) API reliability

Continuous monitoring enables developers to identify performance degradation caused by software updates, changes in external APIs, evolving clinical guidelines, or distribution shifts in incoming patient data. Modern observability platforms increasingly support real-time tracing of agent reasoning, tool execution, memory utilization, and workflow completion, thereby facilitating rapid debugging and quality assurance. Despite its practical advantages, online evaluation presents significant challenges. Controlled experimentation is difficult because each user interaction differs. Ethical considerations restrict experimentation involving clinical decision support, and patient privacy regulations impose additional constraints on data collection and performance analysis. Moreover, evaluating failures in production environments may expose patients to unnecessary risks if adequate safeguards are not implemented.

Comparing Offline and Online Evaluation

Offline and online evaluation address complementary aspects of agent assessment rather than competing methodologies. Offline evaluation emphasizes reproducibility, controlled experimentation, and benchmarking across standardized datasets, whereas online evaluation measures operational effectiveness under real-world conditions. Offline evaluation is particularly valuable during sys-

tem development because it enables rapid iteration, quantitative comparison, and safe experimentation using synthetic healthcare data. Conversely, online evaluation becomes essential once systems are deployed within healthcare environments, where reliability, robustness, clinician trust, and patient safety determine practical usefulness. A comprehensive evaluation strategy therefore combines both approaches. Offline benchmarking establishes baseline capability before deployment, while online monitoring continuously validates performance after deployment. Together, these evaluation paradigms support continual improvement throughout the entire lifecycle of an agentic AI system.

Future Directions

As healthcare increasingly adopts autonomous AI agents, evaluation methodologies must evolve beyond isolated benchmark datasets toward continuous lifecycle assessment. Future research should focus on standardized healthcare benchmarks that integrate longitudinal patient histories, multi-agent collaboration, interoperability testing, and regulatory compliance within realistic clinical workflows.

Emerging evaluation frameworks should incorporate explainability, fairness, uncertainty estimation, safety verification, and human-in-the-loop assessment alongside traditional performance metrics. Continuous monitoring systems capable of detecting workflow failures, hallucinations, privacy violations, and reasoning inconsistencies will become increasingly important for maintaining trustworthy deployment. Ultimately, combining reproducible offline evaluation with comprehensive online observability provides the most reliable pathway toward safe, transparent, and clinically acceptable agentic AI systems capable of supporting next-generation digital healthcare ecosystems.