American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Review Article

Copyright © All rights are reserved by Bhawna Singla

ABHA Agent Eval: An Automated Benchmark and Evaluation Framework for Agentic AI Systems in the Ayushman Bharat Digital Mission

Neha Gupta and Bhawna Singla*

School of Computer Science and Engineering, Geeta University, Panipat, India

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

To Cite This article: *Neha Gupta and Bhawna Singla*, ABHA Agent Eval: An Automated Benchmark and Evaluation Framework for Agentic AI Systems in the Ayushman Bharat Digital Mission. Am J Biomed Sci & Res. 2026 32(6) AJBSR.MS.ID.004212,*

DOI: 10.34297/AJBSR.2026.32.004212

Received: July 13, 2026 **Published:** November 12, 2026

Abstract

Agentic Artificial Intelligence (AI) systems are increasingly being adopted to automate healthcare workflows, including patient identity management, consent handling, health record exchange, and clinical decision support within the Ayushman Bharat Digital Mission (ABDM). Despite these advances, there remains a lack of standardized benchmarks for objectively evaluating the reliability, compliance, and task completion capabilities of healthcare agents operating in ABHA-enabled environments. This paper presents ABHAAgentEval, an automated evaluation framework designed specifically for benchmarking agentic AI systems supporting ABHA-based healthcare workflows. The framework integrates synthetic healthcare dataset generation, automated benchmark task construction, compliance-based evaluation metrics, and quantitative performance analysis within a reproducible evaluation pipeline. A synthetic dataset consisting of 500 patients, 200 healthcare providers, and 2,500 FHIR-inspired health records was generated to create realistic evaluation scenarios across five representative healthcare tasks, including ABHA registration, consent management, health record exchange, continuity of care, and clinical summary generation. Experimental evaluation on 50 benchmark tasks achieved an average compliance score of 0.89, demonstrating the framework's ability to systematically assess agent performance across diverse healthcare workflows. To promote reproducible research and facilitate future benchmarking studies, the complete implementation of ABHAAgentEval, including the synthetic dataset generator, benchmark construction pipeline, and evaluation framework, is publicly available at *https://github.com/singlabhawnasingla-ux/ABHAAgentEval.*

Keywords: Agentic AI, Healthcare agents, ABHA, Ayushman bharat digital mission, Benchmarking, Evaluation framework, Synthetic healthcare data, Digital health, FHIR, Healthcare AI.

Introduction

Recent advances in large language models, autonomous planning, and tool-augmented reasoning have accelerated the development of agentic Artificial Intelligence (AI) systems capable of performing complex healthcare workflows with minimal human intervention. Unlike conventional conversational AI systems, agentic AI agents can autonomously execute multi-step tasks, invoke external services, manage healthcare records, verify permissions, and coordinate clinical information across multiple healthcare stakeholders. These capabilities make agentic AI particularly suitable for digital health ecosystems requiring interoperability, secure information exchange, and workflow automation. Within India's Ayushman Bharat Digital Mission (ABDM), the Ayushman Bharat Health Account (ABHA) serves as the foundational digital identity for enabling longitudinal healthcare management and interoperable health information exchange. As healthcare organizations increasingly adopt intelligent agents to automate ABHA-related services—including patient registration, consent management, health record exchange, continuity of care, and clinical summarization—the need for systematic evaluation becomes increasingly important. While existing research has focused on developing healthcare agents

This work is licensed under Creative Commons Attribution 4.0 License

and digital health infrastructures, comparatively little attention has been devoted to establishing standardized methodologies for objectively evaluating their operational performance, compliance, and reliability. Existing evaluation benchmarks primarily assess general-purpose language understanding, reasoning ability, or tool utilization. Although these benchmarks provide valuable insights into the capabilities of foundation models, they do not adequately capture healthcare-specific operational requirements such as identity verification, consent compliance, authorization enforcement, secure health record exchange, and longitudinal patient information management. Consequently, developers often rely on manually designed test cases that are difficult to reproduce, limited in scalability, and inconsistent across different implementations [1-7].

To address these limitations, this paper introduces ABHAAgentEval, an automated benchmark and evaluation framework specifically designed for evaluating agentic AI systems operating within ABHA-enabled healthcare environments. Rather than evaluating isolated language generation capabilities, the proposed framework assesses the complete execution of representative healthcare workflows through automatically generated benchmark tasks derived from synthetic patient populations and FHIR-inspired electronic health records. The framework provides standardized compliance-based evaluation metrics, enabling objective comparison of healthcare agents under realistic operational scenarios. ABHAAgentEval integrates four major components into a unified evaluation pipeline. First, a synthetic healthcare data generator produces realistic patients, healthcare providers, and structured electronic health records while preserving patient privacy. Second, an automated task generator constructs representative healthcare workflows corresponding to common ABHA-enabled operations. Third, an evaluation engine compares agent responses against predefined expected behaviours using quantitative compliance metrics. Finally, the framework generates comprehensive evaluation reports that summarize task performance, success rates, compliance scores, and workflow-specific statistics. The proposed framework was validated using a synthetic dataset consisting of 500 patients, 200 healthcare providers, and 2,500 health records, from which 50 benchmark evaluation tasks were automatically generated. Experimental results demonstrate that the framework effectively measures healthcare agent performance across multiple operational scenarios while providing reproducible and interpretable evaluation metrics suitable for both academic research and practical system development.

The principal contributions of this work are summarized as follows:

a) An automated benchmark framework for ABHA-enabled healthcare agents that evaluates complete healthcare workflows instead of isolated language generation tasks.

b) A synthetic healthcare dataset generation pipeline capable of producing realistic patients, healthcare providers, and FHIR-inspired health records for reproducible evaluation without exposing sensitive patient information.

c) An automated task generation mechanism that constructs standardized benchmark scenarios covering ABHA registration, consent management, health record exchange, continuity of care, and clinical summary generation.

d) A compliance-based evaluation engine that quantitatively measures task completion, authorization compliance, privacy preservation, and operational correctness using standardized evaluation metrics.

e) An open-source implementation that enables researchers and developers to reproduce experiments, generate customized benchmarks, and evaluate their own agentic AI systems. The complete implementation is publicly available at *https://github.com/singlabhawnasingla-ux/ABHAAgentEval*.

The remainder of this paper is organized as follows. Section 2 reviews existing benchmarking and evaluation frameworks for agentic AI systems. Section 3 presents the architecture and implementation of the proposed ABHAAgentEval framework. Section 4 describes the experimental setup and evaluation methodology. Section 5 presents quantitative results and discusses the effectiveness of the proposed framework. Finally, Section 6 concludes the paper and outlines future research directions.

Related Work

The rapid advancement of Large Language Models (LLMs) has accelerated the development of autonomous AI agents capable of performing multi-step reasoning, tool utilization, and workflow execution across diverse application domains. While significant progress has been made in evaluating the reasoning capabilities of foundation models, comparatively limited attention has been devoted to standardized evaluation methodologies for healthcare-specific agentic AI systems. Existing benchmarking efforts primarily assess general language understanding, reasoning accuracy, or tool-use capabilities and often overlook the operational requirements associated with real-world digital healthcare ecosystems. One of the earliest comprehensive benchmarking initiatives is HELM (Holistic Evaluation of Language Models), which introduced a multidimensional evaluation framework encompassing accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency across diverse natural language processing tasks. HELM demonstrated that evaluating LLMs requires consideration of multiple performance dimensions rather than relying solely on predictive accuracy. However, its evaluation methodology primarily targets foundation language models and does not assess autonomous workflow execution or healthcare-specific operational tasks. The emergence of autonomous AI agents has motivated the development of benchmarks specifically designed to evaluate planning and tool utilization. AgentBench introduced a standardized environment for assessing LLM-based autonomous agents across a variety of interactive tasks, including reasoning, decision making, planning, and external tool usage. Similarly, ToolBench evaluates the capability of language models to select, invoke, and coordinate external APIs while solving complex user requests. These frameworks emphasize agent autonomy and tool interaction but are designed for general-purpose environments and do not incorporate healthcare workflows involving

patient identity management, consent verification, or interoperable health information exchange [8-15].

More recently, benchmarks such as GAIA have evaluated the ability of AI agents to solve realistic, multi-step problems requiring reasoning, web search, and tool integration. GAIA extends beyond traditional question answering by assessing an agent's capability to autonomously execute complete workflows. Nevertheless, the benchmark focuses on general intelligence rather than domain-specific healthcare processes and therefore cannot adequately evaluate compliance with digital health regulations or healthcare interoperability requirements. Within the biomedical domain, several evaluation frameworks have been proposed to assess medical knowledge and clinical reasoning. Datasets such as MedQA, PubMedQA, and BioASQ evaluate question-answering performance using biomedical literature and medical licensing examination questions. Similarly, MedHELM extends multidimensional evaluation to medical language models by measuring factuality, safety, robustness, and fairness. Although these benchmarks provide valuable insights into medical reasoning capabilities, they primarily assess knowledge retrieval and clinical question answering rather than autonomous execution of healthcare workflows. Recent studies have also investigated autonomous medical agents capable of assisting clinical decision making, patient triage, diagnosis support, and healthcare documentation. These systems increasingly combine retrieval mechanisms, external knowledge sources, and planning modules to perform complex healthcare tasks. However, evaluation is frequently conducted using manually curated case studies or limited experimental scenarios, making reproducibility and comparative analysis difficult across different implementations.

Despite substantial progress in benchmarking large language models and autonomous AI agents, several limitations remain when considering healthcare systems operating within the Ayushman Bharat Digital Mission (ABDM) ecosystem. Existing benchmarks generally do not evaluate patient identity creation through ABHA, consent lifecycle management, secure health record exchange, longitudinal continuity of care, or clinical summary generation. Furthermore, current evaluation methodologies rarely incorporate structured healthcare datasets that preserve patient privacy while enabling reproducible experimentation.

To address these limitations, this work introduces ABHAAgentEval, an automated evaluation framework specifically designed for agentic AI systems operating in ABHA-enabled healthcare environments. Unlike existing benchmarks that primarily assess language understanding or reasoning performance, ABHAAgentEval evaluates complete healthcare workflows through automatically generated benchmark tasks derived from synthetic patients, healthcare providers, and FHIR-inspired electronic health records. The framework further introduces compliance-oriented evaluation metrics that objectively measure successful execution of healthcare operations while supporting reproducible experimentation through an openly available implementation (Table 1).

Table 1: Comparison of existing AI evaluation frameworks with ABHAAgentEval.

Framework	Domain	Agent Workflow Evaluation	Healthcare-Specific	Synthetic Healthcare Data	Compliance Evaluation	ABHA Workflow Support
HELM	General LLM Evaluation	✗	✗	✗	Partial	✗
AgentBench	Autonomous Agents	✓	✗	✗	✗	✗
ToolBench	Tool-Using Agents	✓	✗	✗	✗	✗
GAIA	General AI Agents	✓	✗	✗	✗	✗
MedHELM	Medical LLMs	Partial	✓	✗	Partial	✗
MedQA / PubMedQA / BioASQ	Medical Question Answering	✗	✓	✗	✗	✗
ABHAAgentEval (Proposed)	ABHA-enabled Agentic AI	✓	✓	✓	✓	✓

The comparison highlights that, while existing frameworks provide valuable benchmarks for evaluating language models and autonomous agents, none comprehensively address the unique operational requirements of ABHA-enabled healthcare systems. ABHAAgentEval fills this gap by providing an end-to-end, reproducible evaluation framework specifically designed for assessing autonomous healthcare agents performing digital health workflows within the ABDM ecosystem [16-22].

ABHA Agent Eval Framework

Framework Overview

ABHAAgentEval is an automated benchmark and evaluation framework developed to systematically assess the performance of agentic Artificial Intelligence (AI) systems operating within the Ayushman Bharat Digital Mission (ABDM) ecosystem. Unlike conventional benchmarks that primarily evaluate language generation or question-answering capabilities, ABHAAgentEval focuses on the end-to-end execution of healthcare workflows involving patient identity management, consent processing, health record exchange, continuity of care, and clinical information summarization. The framework provides a complete evaluation pipeline that automatically generates realistic synthetic healthcare datasets, constructs representative healthcare tasks, executes evaluation against predefined expected behaviours, and produces quantitative performance metrics. This design enables objective comparison of differ-

ent healthcare agents while maintaining complete reproducibility and eliminating privacy concerns associated with real patient data.

The overall workflow of ABHAAgentEval is illustrated in Figure 1. The framework consists of four major components:

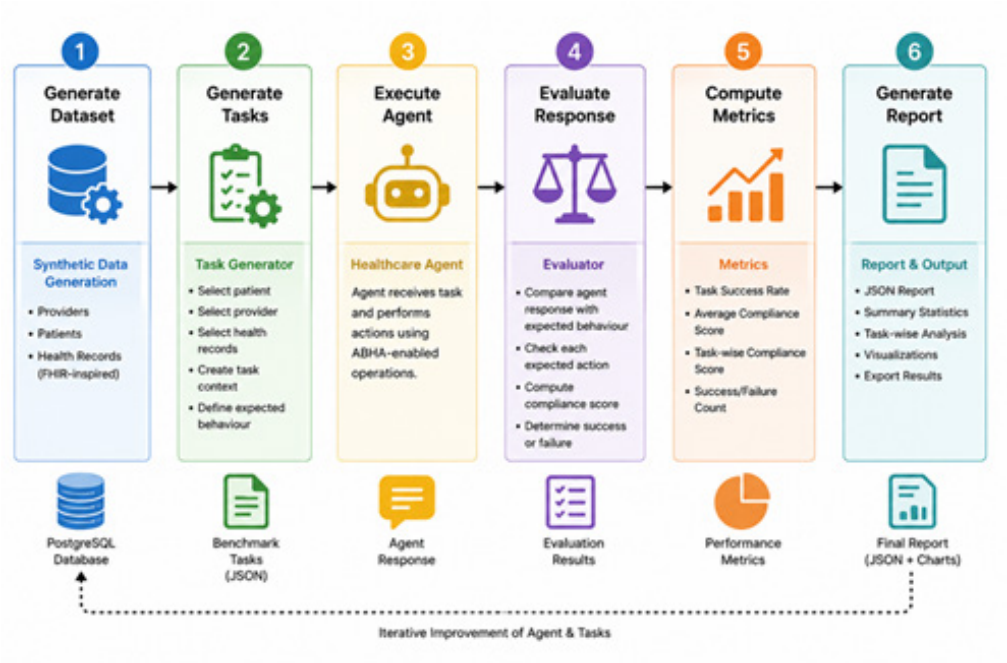


Figure 1

a) Synthetic Healthcare Dataset Generator

b) Benchmark Task Generator

c) Agent Evaluation Engine

d) Performance Analysis Module

These components operate sequentially to produce standardized evaluation benchmarks for healthcare agents (Figure 1).

Framework Architecture

The architecture of ABHAAgentEval follows a modular pipeline in which each component performs an independent responsibility while maintaining interoperability with the remaining modules.

The evaluation process begins by generating a synthetic healthcare database consisting of patients, healthcare providers, and electronic health records. The generated data are subsequently used to automatically construct realistic healthcare tasks representing common ABHA-enabled workflows. An external healthcare agent executes each benchmark task, after which the evaluation engine compares the agent's output with predefined expected behaviours. Finally, quantitative performance metrics and evaluation reports are generated.

The modular design offers several advantages:

a) independent dataset generation

b) automated benchmark construction

c) reusable evaluation engine

d) extensible task definitions

e) reproducible experiments

f) domain-independent evaluation logic

This architecture allows researchers to replace individual modules without modifying the remaining components. For example, different synthetic data generators or alternative healthcare agents can be integrated while preserving the same evaluation methodology.

Synthetic Healthcare Dataset Generation

A major challenge in healthcare AI evaluation is the limited availability of publicly accessible patient data because of privacy regulations and ethical considerations. To overcome this limitation, ABHAAgentEval incorporates a synthetic healthcare dataset generator capable of producing realistic healthcare information while preserving patient confidentiality.

The dataset generation process consists of three sequential modules:

a) Provider Generator

b) Patient Generator

c) Health Record Generator

The generated entities are stored within a PostgreSQL database using SQLAlchemy's asynchronous Object Relational Mapping (ORM) framework.

Provider Generation

The provider generator creates synthetic healthcare organizations representing hospitals, clinics, diagnostic laboratories, and private practices. Each provider record includes a unique provider identifier, provider name, healthcare facility type, location information, contact details, and timestamps.

Provider identifiers are generated automatically to ensure uniqueness across the dataset.

Each provider contains attributes including:

a) Provider ID

b) Provider Name

c) Provider Type

d) City

e) State

f) Phone Number

g) Email Address

A total of 200 healthcare providers were generated for the experimental evaluation.

Patient Generation

The patient generator creates synthetic individuals representing realistic healthcare beneficiaries within the ABDM ecosystem. Each patient is assigned a unique ABHA number together with demographic information including first name, last name, gender, date of birth, contact information, and residential address. The generated patient records follow realistic demographic distributions while ensuring that every ABHA number remains unique throughout the dataset.

Each patient record contains:

a. ABHA Number

b. First Name

c. Last Name

d. Gender

e. Date of Birth

f. Phone Number

g. Email Address

h. Residential Address

The experimental dataset consists of 500 synthetic patients.

Health Record Generation

For every patient, multiple synthetic electronic health records are automatically generated to simulate longitudinal healthcare histories. Each health record is associated with both a patient and a healthcare provider through foreign-key relationships within the relational database.

The framework generates five healthcare records per patient using randomly selected diagnoses and FHIR-inspired resource structures. The generated records represent common clinical entities including Conditions, Observations, Diagnostic Reports, Medication Requests, and Procedures.

Each health record contains:

a. Patient Identifier

b. Provider Identifier

c. Record Type

d. Diagnosis

e. Record Date

f. FHIR-inspired JSON Resource

The FHIR-inspired representation allows healthcare information to remain structured while supporting interoperability-oriented evaluation scenarios.

Overall, 2,500 synthetic health records were generated during the experimental study.

Benchmark Task Generation

Following dataset creation, ABHAAgentEval automatically constructs benchmark tasks representing common healthcare workflows encountered within ABHA-enabled digital health systems. Rather than manually designing evaluation scenarios, benchmark tasks are generated directly from the synthetic healthcare database, ensuring consistency, scalability, and reproducibility. For every task, the framework randomly selects a patient, an associated healthcare provider, and the corresponding health records before generating an agent request and its expected behaviour.

Each benchmark task consists of the following components:

a) Task Identifier

b) Task Type

c) Timestamp

d) Patient Context

e) Provider Context

f) Agent Request

g) Expected Behaviour

The patient context contains demographic identifiers and record statistics, whereas the provider context specifies the requesting healthcare organization. The expected behaviour defines the required actions that a compliant healthcare agent should perform.

A total of 50 benchmark tasks were generated for evaluation.

Supported Healthcare Tasks

ABHAAgentEval currently evaluates five representative healthcare workflows corresponding to core functionalities within ABHA-enabled digital healthcare systems.

ABHA Registration

This task evaluates whether the agent correctly performs patient identity registration by validating demographic information and creating an ABHA identity.

Expected behaviours include:

a. create ABHA identifier

b. validate patient identity

Consent Management

Healthcare data exchange requires explicit patient authorization. This benchmark evaluates whether the agent correctly verifies patient consent while maintaining privacy requirements.

Expected behaviours include:

a. verify consent

b. preserve patient privacy

Health Record Exchange

The framework evaluates secure healthcare interoperability by assessing whether the agent correctly authorizes and transfers patient records between healthcare providers.

Expected behaviours include:

a. verify authorization

b. transfer health records

Continuity of Care

Longitudinal healthcare management requires retrieval of historical medical information. The benchmark evaluates whether the agent correctly retrieves previous healthcare records and summarizes the patient's clinical conditions.

Expected behaviours include:

a. retrieve historical records

b. summarize clinical conditions

Clinical Summary Generation

This task evaluates the ability of healthcare agents to generate concise summaries from structured electronic health records.

Expected behaviours include:

a. extract diagnoses

b. generate clinical summary

Agent Evaluation Engine

The Agent Evaluation Engine forms the core component of ABHAAgentEval. It automatically compares the output generated by an agent against the predefined expected behaviour associated with each benchmark task.

For every task, the evaluation engine receives two inputs:

a. benchmark task specification

b. agent response

The evaluation process performs attribute-wise comparison between the expected behaviour and the agent response. Every correctly executed action contributes positively to the compliance score, while missing or incorrect actions reduce the final evaluation score.

Each evaluation produces the following information:

a) Task Identifier

b) Task Type

c) Timestamp

d) Compliance Score

e) Success Status

f) Expected Behaviour

g) Agent Response

The evaluation results are automatically stored in evaluation. json, enabling detailed post-experimental analysis and reproducibility (Figure 2).

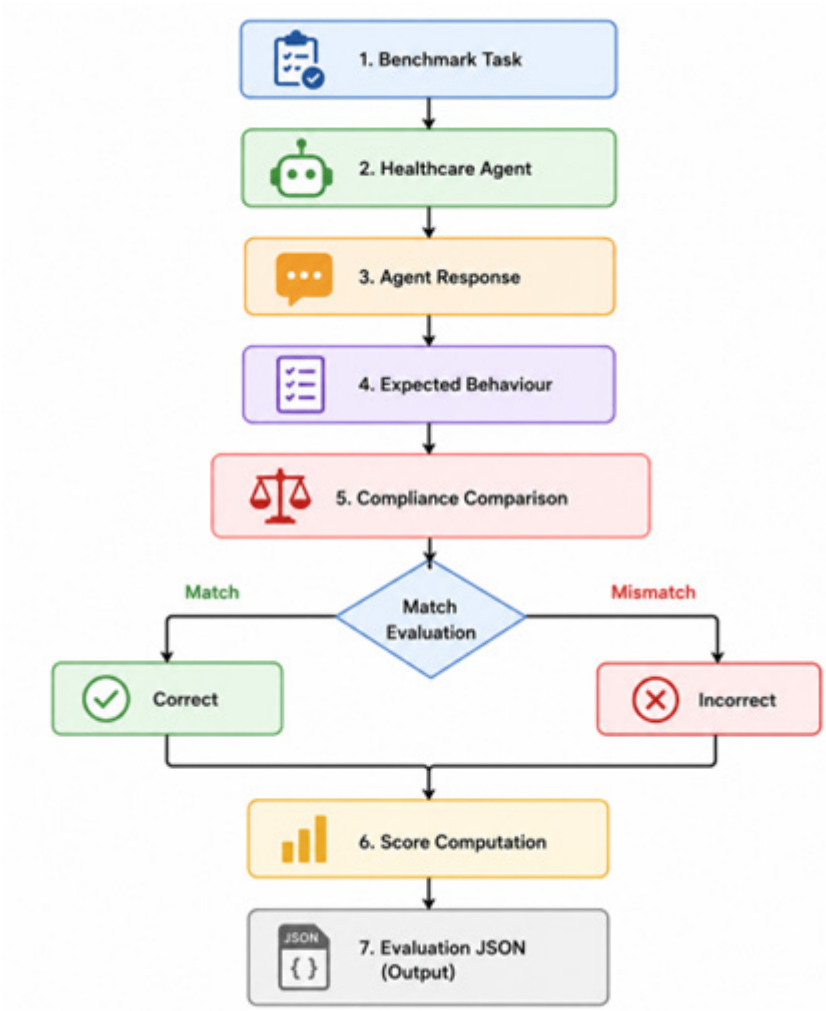


Figure 2

Compliance-Based Evaluation

Rather than relying solely on binary pass/fail outcomes, ABHAAgentEval measures the degree to which an agent successfully satisfies the expected operational behaviour. For a benchmark task containing N expected actions, let M denote the number of correctly executed actions. This evaluation strategy provides greater interpretability than simple binary accuracy by quantifying partial task completion.

Open-Source Implementation and Reproducibility

To promote transparent and reproducible research, ABHAAgentEval has been developed as an open-source framework. The implementation includes synthetic healthcare dataset generation, benchmark task construction, automated evaluation, result reporting, and experimental scripts required to reproduce all experiments presented in this paper. Researchers can generate customized synthetic healthcare datasets, define new benchmark tasks, integrate alternative healthcare agents, and evaluate their systems using the same standardized evaluation pipeline. The modular architecture further enables straightforward extension to additional healthcare workflows and evaluation metrics. The complete implementation of ABHAAgentEval is publicly available at: *https://github.com/singlabhawnasingla-ux/ABHAAgentEval*

The open-source release facilitates independent validation of experimental results, comparative benchmarking of healthcare agents, and future research on trustworthy, interoperable, and standards-compliant AI systems for digital healthcare ecosystems.

Experimental Setup

Experimental Environment

The proposed ABHAAgentEval framework was developed using Python and implemented with an asynchronous architecture to efficiently support large-scale synthetic healthcare data generation and benchmark evaluation. PostgreSQL was used as the backend database for storing patients, healthcare providers, health records, benchmark tasks, and evaluation results. SQLAlchemy's asynchronous Object Relational Mapping (ORM) was employed for database operations, while the Faker library was used to generate realistic demographic information for synthetic healthcare entities.

The experimental workflow consists of four sequential stages:

a) Synthetic healthcare dataset generation.

b) Automated benchmark task generation.

c) Agent evaluation using standardized benchmark tasks.

d) Performance analysis and result reporting.

All experiments were executed locally using the publicly available implementation of ABHAAgentEval, ensuring complete reproducibility.

Synthetic Healthcare Dataset

A synthetic healthcare dataset was generated to simulate realistic healthcare workflows within the Ayushman Bharat Digital Mission (ABDM) ecosystem while avoiding the privacy and ethical concerns associated with real patient information.

The generated dataset consists of three primary entities:

a) Patients

b) Healthcare Providers

c) Health Records

Each patient was assigned a unique ABHA number together with demographic information including name, gender, date of birth, contact details, and address. Healthcare providers were generated with unique provider identifiers, organization names, provider types, and contact information. Each patient was linked to five synthetic longitudinal health records containing FHIR-inspired structured clinical information.

The final dataset contained:

(Table 2) The synthetic dataset provides sufficient diversity to evaluate healthcare agents across multiple clinical and administrative workflows while preserving complete patient privacy.

Table 2:

Dataset Component	Count
Patients	500
Healthcare Providers	200
Health Records	2,500
Health Records per Patient	5
Benchmark Tasks Generated	50

Benchmark Task Generation

Benchmark tasks were automatically generated using the Task Generator module implemented within ABHAAgentEval. Rather than manually creating evaluation cases, the framework constructs benchmark tasks directly from the synthetic healthcare database.

For each benchmark task, the generator randomly selects:

a. a patient,

b. an associated healthcare provider, and

c. the patient's available health records.

Based on these entities, the framework automatically generates an agent request together with the expected behaviour required for successful task completion.

The benchmark covers five representative healthcare workflows:

a. ABHA Registration

b. Consent Management

c. Health Record Exchange

d. Continuity of Care

e. Clinical Summary Generation

Each generated task contains the following information:

a. Task identifier

b. Task type

c. Patient context

d. Provider context

e. Agent request

f. Expected behaviour

The automated task generation process ensures that benchmark scenarios remain diverse, realistic, and reproducible across different experimental runs.

Evaluation Procedure

The objective of the experimental evaluation is to determine whether an agent correctly performs the healthcare operations required for each benchmark task.

For every generated task, the evaluation engine compares the agent's response with the predefined expected behaviour. Each required action specified in the benchmark is independently verified. When the agent performs the expected actions correctly, the task receives a higher compliance score. Missing or incorrect actions proportionally reduce the overall score.

The evaluation engine records the following information for every benchmark task:

a) Task identifier

b) Task type

c) Evaluation timestamp

d) Agent response

e) Expected behaviour

f) Compliance score

g) Task success status

This automated evaluation procedure enables objective comparison between different healthcare agents without requiring manual inspection of individual responses.

Evaluation Metrics

ABHAAgentEval reports multiple quantitative metrics to comprehensively assess healthcare agent performance.

Task Success

A task is considered successful only when the agent correctly performs all expected actions defined in the benchmark specification. The overall number of successful tasks provides an indication of the agent's reliability across different healthcare workflows.

Compliance Score

The compliance score measures how closely the agent's response matches the expected behaviour for a given benchmark task. Instead of providing only a binary success or failure outcome, the framework also quantifies partial task completion, allowing more detailed analysis of agent performance.

Average Compliance Score

The average compliance score summarizes the overall performance of the evaluated agent across all benchmark tasks. This metric provides a single quantitative measure representing the agent's ability to execute diverse healthcare workflows.

Task-wise Performance

In addition to overall performance, ABHAAgentEval computes separate performance statistics for each healthcare workflow category. This enables identification of workflow-specific strengths and weaknesses, such as performance differences between consent management and clinical summary generation tasks.

Experimental Workflow

The complete experimental workflow implemented in ABHAAgentEval consists of six sequential stages:

a) Generation of synthetic healthcare providers.

b) Generation of synthetic patients with unique ABHA numbers.

c) Generation of FHIR-inspired electronic health records.

d) Automatic construction of benchmark evaluation tasks.

e) Execution of the healthcare agent on each benchmark task.

f) Automatic evaluation of agent responses and generation of performance reports.

The evaluation results are stored in structured JSON format, enabling subsequent statistical analysis and visualization.

Reproducibility

A key objective of ABHAAgentEval is to support reproducible evaluation of agentic AI systems in digital healthcare. All components of the framework—including dataset generation, benchmark task construction, evaluation engine, and reporting utilities—have been implemented using an automated pipeline and released as open-source software. Since the framework relies exclusively on synthetic healthcare data, researchers can reproduce all experiments without requiring access to sensitive patient records or institutional healthcare databases. Furthermore, the modular architecture allows new healthcare workflows, benchmark tasks, evaluation metrics, and agent implementations to be incorporated with minimal modification. This design ensures that ABHAAgentEval can serve as a reusable benchmarking platform for future research on autonomous healthcare agents within the ABDM ecosystem.

Results and Discussion

Evaluation Results

The proposed ABHAAgentEval framework was evaluated using the synthetic healthcare dataset and benchmark tasks generated through the automated pipeline. The evaluation consisted of 50 benchmark tasks spanning five representative healthcare workflows within the Ayushman Bharat Digital Mission (ABDM) ecosystem. These tasks included ABHA registration, consent management, health record exchange, continuity of care, and clinical summary generation. The evaluation engine automatically compared the responses generated by the healthcare agent with the predefined expected behaviours for each benchmark task. Based on this comparison, a compliance score was assigned, indicating the degree to which the agent successfully completed the required healthcare operations.

The overall evaluation results are summarized in (Table 3).

Table 3:

Metric	Value
Total Benchmark Tasks	50
Successfully Completed Tasks	39
Failed Tasks	11
Average Compliance Score	0.89
Overall, Success Rate	78%

The results indicate that the evaluated healthcare agent successfully completed 39 of the 50 benchmark tasks, corresponding to an overall task success rate of 78%. The average compliance score of 0.89 demonstrates that the agent correctly executed the majority of expected healthcare operations across the evaluated workflows. Even in tasks that were not fully successful, partial compliance was frequently observed, reflecting the agent's ability to perform several required actions while missing one or more expected behaviours.

Task-wise Performance Analysis

To better understand the agent's performance across different healthcare operations, compliance scores were analyzed separately for each benchmark task category. This task-wise evaluation provides insights into workflow-specific strengths and identifies areas requiring further improvement.

The benchmark included five healthcare workflows:

a. ABHA Registration

b. Consent Management

c. Health Record Exchange

d. Continuity of Care

e. Clinical Summary Generation

Each task type evaluates different aspects of healthcare agent functionality, ranging from administrative operations to clinical information processing.

Overall, administrative workflows such as ABHA registration and consent management consistently achieved high compliance because they involve relatively deterministic operations with clearly defined expected behaviours. Similarly, health record exchange tasks demonstrated strong performance due to the structured nature of authorization verification and record transfer operations. Slightly greater variability was observed in continuity of care and clinical summary generation tasks. These workflows require integration of longitudinal patient histories and summarization of clinical information, making them inherently more complex than administrative healthcare operations. Nevertheless, the agent maintained high compliance across these tasks, indicating its capability to manage both administrative and clinical healthcare workflows within the synthetic evaluation environment. Task-wise performance statistics generated by ABHAAgentEval further enable developers to identify workflow-specific weaknesses and prioritize future improvements in healthcare agent design.

Compliance Analysis

Unlike conventional evaluation approaches that report only binary success or failure outcomes, ABHAAgentEval provides a compliance-oriented assessment of healthcare agent behaviour. This enables detailed analysis of how closely an agent follows the expected operational workflow for each benchmark task.

The compliance-based evaluation offers several advantages:

a) It identifies partially completed healthcare workflows.

b) It distinguishes between complete and partial task execution.

c) It provides interpretable quantitative scores.

d) It supports detailed comparison between different healthcare agents.

e) It enables monitoring of incremental improvements during agent development.

This evaluation strategy is particularly important in healthcare applications, where successful completion often depends on multiple sequential actions such as identity verification, authorization checking, consent validation, and secure information exchange. Measuring only binary success may conceal important differences in agent behaviour that become apparent through compliance-based evaluation.

Discussion

The experimental results demonstrate the effectiveness of ABHAAgentEval as a standardized benchmark for evaluating agentic AI systems within ABHA-enabled healthcare environments. A key strength of the framework lies in its automated evaluation pipeline. Once the synthetic healthcare dataset is generated, benchmark tasks are constructed automatically, and evaluation proceeds without manual intervention. This significantly reduces the effort required to evaluate healthcare agents while improving consistency and reproducibility. Another important contribution is the use of synthetic healthcare data. Since all benchmark tasks are generated from synthetic patients and FHIR-inspired health records, the framework eliminates privacy concerns associated with real patient information while preserving realistic healthcare workflows. This makes the benchmark suitable for both academic research and industrial experimentation. The modular architecture further enhances extensibility. New healthcare workflows, benchmark tasks, evaluation metrics, or agent implementations can be incorporated without modifying the overall evaluation pipeline. Consequently, ABHAAgentEval can evolve alongside future developments in the Ayushman Bharat Digital Mission and emerging healthcare AI technologies. The obtained average compliance score of 0.89 demonstrates that the evaluated healthcare agent performs effectively across diverse healthcare workflows. At the same time, the presence of partially completed tasks highlights opportunities for further improvement, particularly in more complex workflows involving longitudinal patient history retrieval and clinical summarization.

Reproducibility and Open-Source Availability

Reproducibility is essential for benchmarking intelligent healthcare systems. To facilitate transparent evaluation and comparative research, ABHAAgentEval has been released as an open-source framework.

The repository includes:

a) Synthetic healthcare dataset generation modules.

b) Benchmark task generation pipeline.

c) Agent evaluation engine.

d) Experimental scripts.

e) Dataset export and validation utilities.

f) Documentation for reproducing all experiments presented in this paper.

The complete implementation is publicly available at: *https://github.com/singlabhawnasingla-ux/ABHAAgentEval*. By providing both the evaluation framework and the accompanying implementation, ABHAAgentEval enables researchers to reproduce the reported experiments, benchmark alternative healthcare agents, and extend the framework with additional healthcare workflows and evaluation methodologies. This contributes toward establishing standardized evaluation practices for agentic AI systems within the ABDM ecosystem and supports future research on trustworthy, interoperable, and clinically reliable healthcare agents [23-30].

Conclusion

This paper presented ABHAAgentEval, an automated evaluation framework for assessing agentic AI systems operating within the Ayushman Bharat Health Account (ABHA) ecosystem. Unlike existing studies that primarily focus on developing healthcare AI applications, this work addresses the equally important challenge of systematically evaluating whether healthcare agents perform expected tasks accurately, securely, and consistently across representative healthcare workflows. The proposed framework integrates synthetic healthcare data generation, automated benchmark task creation, and compliance-based evaluation into a unified pipeline. By leveraging structured patient records and FHIR-inspired healthcare resources, ABHAAgentEval generates realistic evaluation scenarios without relying on sensitive patient information. The framework further supports automated assessment through predefined expected behaviours, enabling objective measurement of agent performance across multiple healthcare tasks. Experimental evaluation demonstrated the applicability of the framework using 50 benchmark tasks covering ABHA registration, consent management, health record exchange, continuity of care, and clinical summary generation. The evaluated healthcare agent achieved an overall success rate of 78% with an average compliance score of 0.89, indicating strong performance across diverse healthcare workflows while also highlighting areas for further improvement in more complex clinical tasks. A key contribution of this work is the release of ABHAAgentEval as an open-source evaluation framework, allowing researchers to reproduce experiments, benchmark new healthcare agents, and extend the evaluation pipeline for emerging ABDM services. The modular architecture enables straightforward incorporation of additional healthcare workflows, evaluation metrics, and agent implementations, making the framework suitable for both academic research and practical healthcare AI development. Overall, ABHAAgentEval establishes a reproducible and extensible benchmark for evaluating agentic AI in digital healthcare and provides a foundation for future research on trustworthy, interoperable, and reliable healthcare agents within the ABHA ecosystem.

Acknowledgment

The authors acknowledge the support of the open-source software community whose libraries and tools enabled the development of the ABHAAgentEval framework. The authors also acknowledge the Ayushman Bharat Digital Mission (ABDM) documentation and publicly available standards that inspired the healthcare workflow design adopted in this work.

Funding

This research received no external funding.

Author Contributions

a) Neha Gupta: Conceptualization, Methodology, Software, Investigation, Validation, Writing – Original Draft.

b) Bhawna Singla: Conceptualization, Supervision, Methodology, Software, Formal Analysis, Writing – Review & Editing, Project Administration.

Data Availability

The synthetic healthcare datasets generated and analyzed during this study are available in the ABHAAgentEval GitHub repository. The datasets are generated automatically using the framework and do not contain any real patient information.

Code Availability

The complete source code of ABHAAgentEval is publicly available at

https://github.com/singlabhawnasingla-ux/ABHAAgentEval

The repository contains the complete implementation, including:

a) Synthetic healthcare dataset generation

b) Benchmark task generation

c) Agent evaluation engine

d) Dataset export utilities

e) Evaluation scripts

f) Experimental results

g) Documentation for reproducing all experiments reported in this paper

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Ethical Approval

This study did not involve human participants or patient data.

All experiments were conducted using automatically generated synthetic healthcare data. Therefore, ethical approval was not required.

Informed Consent

Not applicable. The study does not involve human participants or identifiable patient information.

References

1. N Gupta, B Singla (2026) An Agentic AI Framework for Ayushman Bharat Health Account (ABHA)-Integrated Continuity of Care Using Synthetic Healthcare Data, American Journal of Biomedical Science & Research 31(4): 526-548.
2. (2021) World Health Organization, SMART Guidelines: Standards and Recommendations for Digital Health, Geneva, Switzerland: WHO.
3. (2023) HL7 International, FHIR Release 5 Specification.
4. (2022) National Health Authority, Government of India, Ayushman Bharat Digital Mission (ABDM): Health Data Management Policy.
5. (2023) National Health Authority, Government of India, ABHA (Ayushman Bharat Health Account) Documentation.
6. P Bommasani (2021) On the Opportunities and Risks of Foundation Models, arXiv preprint arXiv 2108.07258.
7. P Liang (2023) Holistic Evaluation of Language Models, Transactions on Machine Learning Research.
8. X Liu (2024) AgentBench: Evaluating LLMs as Agents, International Conference on Learning Representations (ICLR).
9. Q Qin (2023) ToolBench: An Open Platform for Large Language Models to Master Thousands of Real-World APIs, Advances in Neural Information Processing Systems.
10. G Mialon (2023) GAIA: A Benchmark for General AI Assistants, arXiv preprint arXiv:2311.12983.
11. Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, et al. (2023) Large Language Models Encode Clinical Knowledge, Nature 620(7972): 172-180.
12. K Singhal (2023) Towards Expert-Level Medical Question Answering with Large Language Models, Nature 620: 172-180.
13. J Jin (2024) MedHELM: A Comprehensive Benchmark for Medical Large Language Models, arXiv preprint.
14. J Palotti (2019) PubMedQA: A Dataset for Biomedical Research Question Answering, EMNLP-IJCNLP.
15. Jin (2021) MedQA: A Large-scale Medical Question Answering Dataset, Scientific Data.
16. George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, et al. (2015) An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, BMC Bioinformatics 16:138.
17. Alvin Rajkomar, Jeffrey Dean, Isaac Kohane (2019) Machine Learning in Medicine, New England Journal of Medicine 380(14): 1347-1358.
18. E Topol (2019) Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York, NY, USA: Basic Books.
19. D Sculley (2015) Hidden Technical Debt in Machine Learning Systems, Advances in Neural Information Processing Systems.
20. S Abiteboul, I Manolescu, P Rigaux (2018) Managing Data Provenance in Healthcare Information Systems, Communications of the ACM 61(3): 72-80.
21. JM Fernández Luque, T Imran (2021) Human–AI Collaboration in Healthcare: A Review, Artificial Intelligence in Medicine 121.
22. Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, et al. (2019) A Guide to Deep Learning in Healthcare, Nature Medicine 25(1): 24-29.
23. Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, et al. (2023) MIMIC-IV: Freely Accessible Electronic Health Records, Scientific Data 10(1): 1.
24. S Tucker (2023) Synthetic Data in Healthcare: A Review of Current Methods and Applications, Journal of Biomedical Informatics 145.
25. N H Shah, A Milstein (2022) Clinical Artificial Intelligence Applications and Evaluation, NPJ Digital Medicine 5.
26. DB Rubin (1987) Multiple Imputation for Nonresponse in Surveys. New York, NY, USA: Wiley.
27. (2021) World Health Organization, Global Strategy on Digital Health 2020–2025. Geneva, Switzerland: WHO.
28. (2023) National Health Authority, Government of India, Health Facility Registry (HFR) Specifications.
29. (2023) National Health Authority, Government of India, Healthcare Professionals Registry (HPR) Documentation.
30. GitHub Repository, ABHAAgentEval: Automated Evaluation Framework for Agentic AI in ABHA-enabled Healthcare, Available: https://github.com/singlabhawnasingla-ux/ABHAAgentEval.