Journal of Information Systems Engineering and Management
2026, 11(3s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

An Empirical Analysis of Reliability Challenges in Multi-Step Agentic AI Systems

Sai Viswa Teja Arumilli

Independent Researcher, USA

ARTICLE INFO

Received: 01 Apr 2026

Revised: 28 May 2026

Accepted: 10 June 2026

ABSTRACT

The application of Agentic Artificial Intelligence (AI) systems attracts a lot of interest to undertake complex tasks through planning, reasoning, memory usage, tool interaction, and validation processes. It is common knowledge that while the applications are widely used in various fields like software engineering, healthcare, and business automation, reliability issues are still one of the primary challenges. The aim of this study is to highlight possible reliability issues in the multi-step agentic AI system, employing an experimental framework. A quantitative research design is used in which the AgentBench benchmark dataset with 5,000 task instances and 12 variables related to the workflow is used. The data pre-processing, reliability modelling and performance evaluation are performed using Python libraries and agent simulation frameworks. Task success rate, failure rate, time to execute task, task recovery rate, consistency of output and error propagation rate are used as measures of reliability. Results indicated that the reliability reduces as difficulty of the task escalates and reasoning and tool related errors are the leading causes of failure. In addition, it is found that the propagation of error over extended workflows could drastically escalate resulting in domino effects. It extends an established reliability evaluation framework, and offers ideas for developing more robust, transparent, and dependable autonomous systems that incorporate agentic AI in complex settings.

Keywords: Agentic Artificial Intelligence, Multi-Step AI Systems, Reliability Analysis, Autonomous Agents, Error Propagation

Introduction

A. Background

Progression of Agentic AI has led to the development of autonomous agents that are capable of executing complex tasks through planning and applications of tools. Agentic systems differ from traditional AI models in that they function over multiple steps, each one connected to the other and are used to reach a target goal [1]. They have become popular in the software industry, the healthcare field, customer service, and business analytics.

B. Problem Statement

Multi-step agentic AI systems are capable of performing complex tasks, but they can still fail at various stages of execution. Other errors may propagate through the workflow, such as during planning, reasoning, memory retrieval and interaction with tools, thus introducing inaccuracies to the output and compromising the dependability of the system [2]. Therefore, failure modes, reliability issues, and error propagation in agentic AI systems need to be explored.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

C. Research Aim and Objectives

Aim: The study aims to examine and explore the reliability issues involved in multi-step agentic AI systems with modelling and experiments.

➔ To find and categorize typical failure modes of multi-step agentic AI systems.

➔ To compare how reliability and performance varies with the task complexity.

➔ To build reliability measures and mitigation techniques for improving system robustness.

D. Novel Contribution

- The present research depicts a framework for analyzing reliability challenges in multi-step agentic AI systems.
- The paper presents a systematic method of the failure categories, measurement of reliability indicators, and an analysis of error propagation through successive interactions between the agents.
- The research focuses on the reliability and dependability of agentic workflows, in contrast to previous work, which has mostly examined task accuracy.
- The proposed ideas involve developing a quantitative evaluation framework and visual analytics to foster practical insights for the development of more resilient, transparent and trustworthy autonomous AI systems.

Related Work

Agentic AI Systems

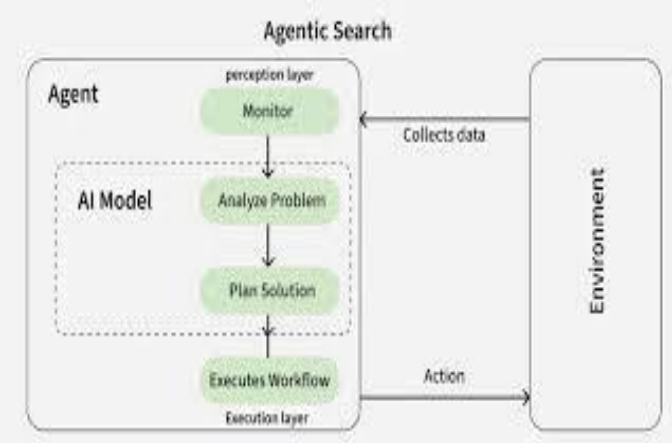

Fig. 1: Agentic AI Systems

Agentic Artificial Intelligence systems can autonomously plan, reason and act on their own to perform a desired action with little to no human intervention. Agentic systems can not only break down complex tasks into smaller subtasks but also make decisions, use external tools and adjust their actions according to evolving conditions regarding the task and the environment [3]. The proliferation of agentic architectures has driven the application of recent Large Language Models (LLMs) developments in software development, business process automation, customer support, and research assistance [4]. The recent advances in the Large Language Models (LLMs) have propelled the use of agentic architectures for various applications, including software development, business process automation, customer support, and research assistance. Agentic systems are highlighted as performing complicated workflows that involve perception, memory, planning and action mechanisms [5]. Multi-agent environments like ***AutoGen, CrewAI, LangChain, and LangGraph*** provide a framework for collaboration between different types of agents, each with their own specific skills, to work together and achieve a common goal [6]. Agentic AI can monitor, analyze, plan and execute the task in a proper way. Agentic AI can be capable of enhancing productivity and optimizing the operational workflow, as it reduces human effort and speeds up decision-making process.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multi-Step Agent Architectures

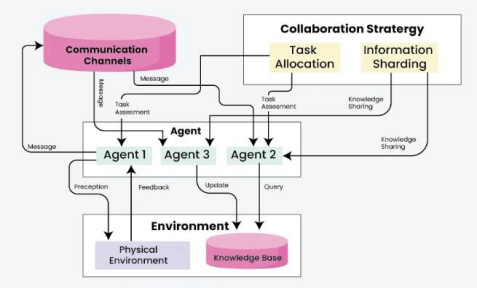

Fig. 2: Multi-Step Agent Architectures

Multi-step agents are capable of executing a series of steps by a collection of agents to get the job done instead of just generating a single answer [7]. In general, in these architectures, there are several process steps such as planning, reasoning, using tools, retrieving information from memory, performing and checking the work with multiple agents like agent1, agent 2 and agent 3. All of the stages are interconnected to ensure that the ultimate goal is realized and that the result of previous stages feeds into the next. Multi-step architectures have been shown to show benefit in problem-solving ability, by dividing complex problems into smaller pieces that can be solved by the agents [8]. Through a structured approach, a greater degree of flexibility is ensured and more complex workflows can be implemented in quite different domains. As part of this, collaborative agent frameworks have been recently examined in which several different specialized agents are employed to carry out overlapping functionality in a workflow [9]. An example plan generation agent is the agent that regulates the generation of a plan, an example reasoning agent is the agent that runs reasoning, and an example validation agent is the agent that validates output [10]. Complexity in workflows can lead to a loss of predictability and can complicate measuring workflow performance. Thus, the study of the impact of multi-step architectures on reliability is an important topic of research for developing agentic AI systems.

Reliability and Robustness in AI Systems

One of the key components of trustworthy Artificial Intelligence systems is reliability and robustness. Reliability is the capacity of an AI system to produce always accurate and predictable results in different conditions, robustness is the capability of an AI system to keep performing in the face of uncertainty, noisy inputs or unexpected situations [11]. The qualities are the focus of recent AI studies as autonomous systems become common not only in areas like health care, finance, transportation, and cybersecurity, but also in other critical fields [12]. Several methods for verifying the reliability of AI have been proposed such as First is by measuring accuracy, the second is by consistency testing, the third is by fault tolerance analysis, and the fourth is by resilience analysis [13]. In the case of agentic AI, reliability is even more critical as the actions taken during each stage can affect subsequent decisions and actions. An incorrect prediction or calculation can be transferred up the line and take a toll on the results. Self-correction strategies and human-in-the-loop strategies have also been studied to increase the robustness and decrease the failure risks [14]. Most reliability studies, however, are targeted towards one model performance rather than the interactions that happen in a multi-step agentic system. There is still a need for an extensive framework that could measure the individual agent performance as well as the overall behavior of interconnected agents in the environment.

Error Propagation in Agentic Workflows

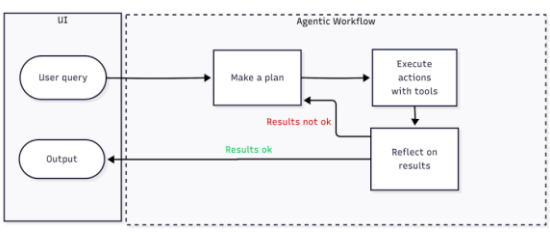

Fig. 3: Error Propagation in Agentic Workflows

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In error propagation, a small error tends to propagate over the course of tasks, resulting in an unnecessarily low average overall performance. This is a major issue across the success of overall task success and reliability of those systems [15]. Errors made in task decomposition, reasoning, information retrieval, and tool selection often affect further decisions, leading to sequences of errors outputs. The information generated at each stage is required for the next stage and small errors in one stage can make significant errors in the following stages [16]. Cascading failures have been recognized as a frequent phenomenon in autonomous workflows in the case of a complex task-based agent system [17]. An incorrect planning decision might lead to other reasoning steps towards an inappropriate solution, and the incorrect use of tools might produce misleading information that can impact the validation stages. As information flows get longer and the task even more intricate, propagation of error is increasingly difficult [18]. To lessen cascading failures, there are some proposals of mitigation methods including intermediate verification, feedback loops, redundancy mechanisms, and dynamic error correction [19]. Even with all these advances, little empirical evidence exists on the frequency, severity and effects of error propagation in various agentic architectures. This constraint provides for importance of more investigations into failure modes in multi-step AI systems.

Reliability Evaluation Frameworks for Agentic AI

Evaluation frameworks for reliability offer systematic methods to measure and evaluate AI systems' performance and reliability. The standard practice in the evaluation is mostly based on the standard measures of accuracy, precision, recall, and the quality of the answer [20]. The advent of multi-step agentic AI systems has brought in greater demands for accounting for reliability at the workflow level. Evaluation models, which recently have been put forward, consider planning effectiveness, consistency of planning decisions, degree of task completion, accuracy of tool use, and error resistance [21]. Benchmarking research programs like ***AgentBench, GAIA, ToolBench,*** and other evaluations tools have provided insightful methodology for evaluating autonomous agent performance for various tasks [22]. Such frameworks can be used to evaluate the capabilities of agents in known environments and assess what needs to be improved. A number of works highlight the need to incorporate reliability measures into evaluation procedures [23]. It is critical, then, that the creation and growth of specialized reliability assessment methods continue to be a critical factor in the successful and safe implementation of agentic AI technology.

Literature Gap

Several existing works have investigated agentic AI architectures, autonomous decision-making, and reliability in task performance, but few works have specifically tackled reliability challenges in multi-step agentic workflows. Most studies focus on output accuracy, but do not consider how failures develop, propagate or impact later stages of task execution. In addition, current evaluation tools offer limited insight of errors that create reliability problems.

Methodology

A. Research Design

The methodology of this study is a quantitative experimental research design to examine the reliability issues of multi-step agentic AI systems. The design allows a systematic evaluation of the agent's performance with respect to varying task complexities and execution conditions. A simulation environment is built and developed in Python that simulates the planning, reasoning, memory retrieval, tool use and validation steps in autonomous work flows. In the study, the reliability is measured based on the task success rate, the number of errors, execution time, the recovery rate and the level of consistency of the output. The study provides insights into prevalent failure modes and views on the overall reliability of agentic AI processes by analyzing these metrics in various task scenarios.

B. Data Collection

The AgentBench benchmark datasets are publicly available datasets for autonomous AI agent evaluation that are used in the study. There are around ***5,000 rows*** showing various aspects of agent behavior and task execution and ***12 columns. The key variables are Task_ID, Task_Type, Complexity_Level, Planning_Steps, Reasoning_Steps, Tool_Calls, Memory_Usage,***

Journal of Information Systems Engineering and Management
2026, 11(3s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

Execution_Time, Success_Status, Error_Type, Recovery_Attempt, and Final_Output_Score. These variables give a full picture of workflow execution, the occurrence of failures, and the outcome on the performance.

C. Data Preprocessing

Pre-processing of the data takes place in order to enhance the data quality and thus ensure the reliability of the data analysis. First, loading the data into Python using the Pandas library and reviews it for any missing values, duplicate records, and missing or inconsistent values. Observations and duplications are separated out to avoid redundant information, categorical variables are converted to numbers and standardized [24]. To boost the model's performance, numerical features like execution time, memory usage, planning steps, reasoning steps and tool calls are normalized with StandardScaler. Statistically based methods are used to detect outliers, while related classes of errors are categorized into failure classes. The cleaned data set is then ready to be used for modelling and evaluation [25].

D. Proposed Model Architecture

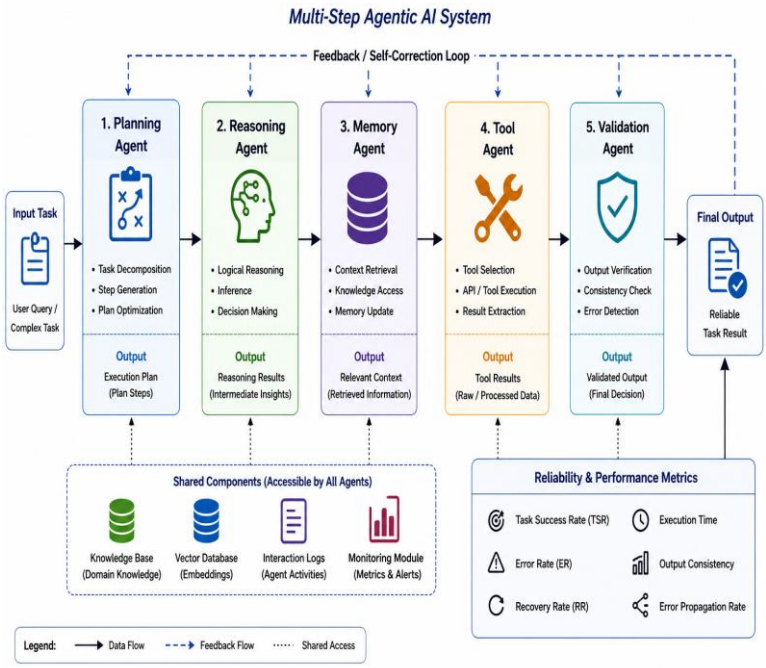

Fig. 4: Proposed Model Architecture

E. Model Implementation

An agent architecture comprising Planning agent, Reasoning agent, Memory agent, Tool agent and Validation agent is proposed and implemented using Python programming language. The tasks are carried out one by one to simulate the agentic process that takes place in real life. Reliability parameters are computed at the end of each execution cycle to help evaluate performance. The model provides a tool to identify failure points, error propagation and reliability degradation at each one of the stages of the workflow.

$$TSR = \frac{Ns}{Nt} \times 100 \quad ------ (1)$$

Where Ns represents successfully completed tasks and Nt represents total assigned tasks.

$$ER = \frac{Nf}{Nt} \times 100 ------ (2)$$

Where Nf represents failed tasks.

Journal of Information Systems Engineering and Management
2026, 11(3s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

$$RR = \frac{Nr}{Nf} \times 100 - - - - - - (3)$$

Where Nr represents recovered failures.

F. Pseudocode

```
BEGIN
Load AgentBench Dataset
Perform Data Cleaning
Remove Duplicate Records
Handle Missing Values
Encode Categorical Variables
Normalize Numerical Features
FOR each Task
        Send Task to Planning Agent
        Generate Execution Plan
        Pass Plan to Reasoning Agent
        Perform Logical Reasoning
        Retrieve Context from Memory Agent
        Execute Required Tools
        Validate Output
        IF Output Correct
            Mark Task Successful
        ELSE
            Record Failure Type
        ENDIF
        Calculate Reliability Metrics
        Store Results
ENDFOR
Generate Statistical Analysis
Visualize Reliability Results
END
```

Fig. 5: Pseudocode

The algorithm mainly starts with the AgentBench data set, and then proceeds to data pre-processing steps like data cleaning, encoding, and data normalization. Each of these tasks is subsequently delegated to the Planning Agent which presents an execution strategy.

G. Workflow Diagram

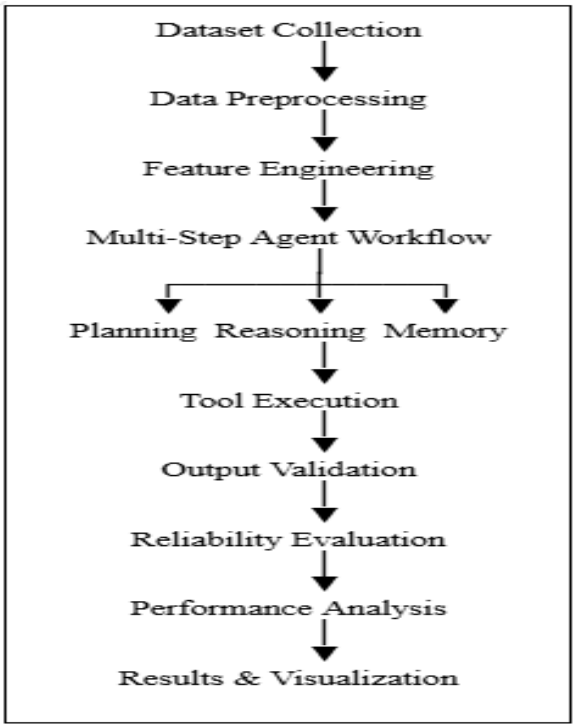

Fig. 6: Workflow Diagram

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The workflow is started with data collection, data quality, data consistency, and then on to data preprocessing. The cleaned data is further processed for feature engineering and preparation before being fed into the multi-step agentic framework.

H. Ethical Considerations

Ethical standards comprise transparency, fairness and accountability in the evaluation process, which are upheld in the research [26]. The AgentBench data set does not include any sensitive information and is made available in a public manner, reducing privacy concerns. Every experimental procedure is performed using an environment that is controlled and free from the possibility of misuse of autonomous agents.

Results And Discussion

A. Results

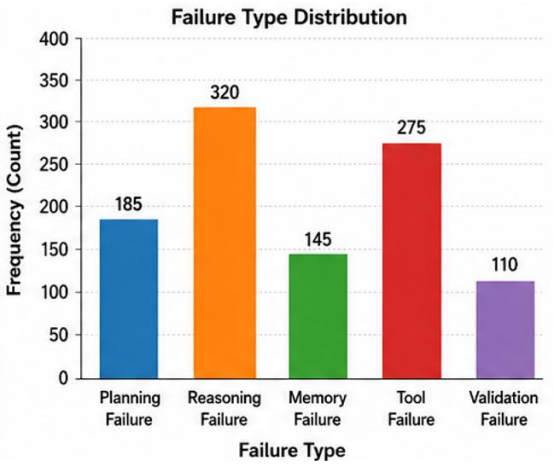

Fig. 7: Failure Type Distribution

The Failure Type Distribution chart displays the statistics of reliability issues with the agents along various phases of the agentic workflow. The next highest number of occurrences are reported under Reasoning Failure (320) followed by Tool Failure (275). Failure occurs in planning with 185 cases, failure when in memory with 145 and failure in validation with 110.

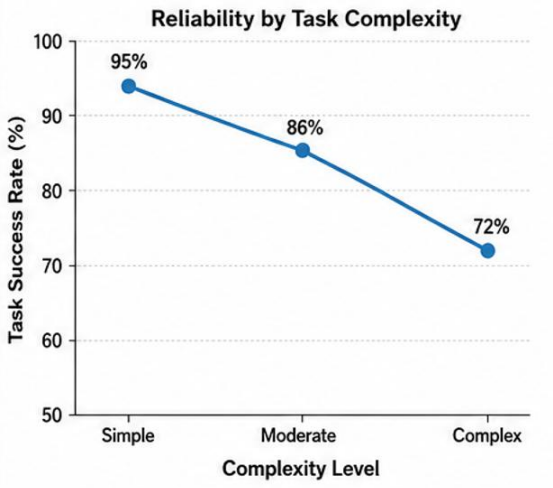

Fig. 8: Reliability vs Task Complexity

Journal of Information Systems Engineering and Management
2026, 11(3s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

The chart shows how the degree of task complexity correlates with the success of the task. Simple tasks 95% of reliability and moderate tasks 86% of reliability. The 72% peak for complex tasks shows that the more complex the application, the lesser the performance.

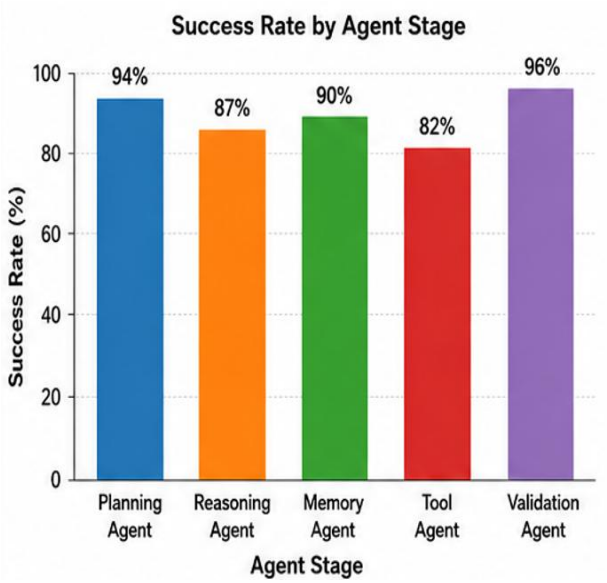

Fig. 9: Task Success Rate by Agent Stage

The most successful Agent is the Validation Agent with a success rate of 96%, then the Planning Agent with a success rate of 94%. Memory Agents have a success rate of 90% while Reasoning Agents a score of 87%. The results of the Tool Agent show the minimum performance of 82%, where the external tool interactions represent one of the important reasons for failures and reliability loss in multi-step agentic environments.

Fig. 10: Execution Time by Task Complexity

The figure shows that simple tasks take around 8 seconds, Moderate tasks take about 17 seconds. At 31 seconds, the longest time to be taken is to perform complex tasks. The trend indicates that the number of reasoning, planning, memorization and tool interaction steps contributes towards the high computational needs. These results support the direct relationship between workflow complexity and operational efficiency resource usage.

Journal of Information Systems Engineering and Management
2026, 11(3s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

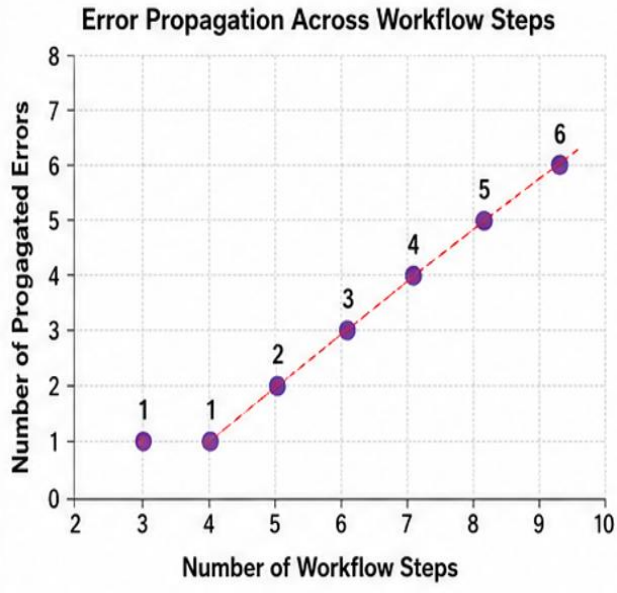

Fig. 11: Error Propagation Analysis

The Error Propagation Analysis chart looks at the relationship between how much the workload increases and how many propagated errors there can be. Tasks with three or four steps have only one propagated error, while workflows with nine steps have nine propagated errors. The number of failures has a clear positive correlation with the complexity of the workflow. The results illustrate that earlier errors can affect later judgements, ultimately resulting in the loss of trust and efficacy in the entire agentic processes.

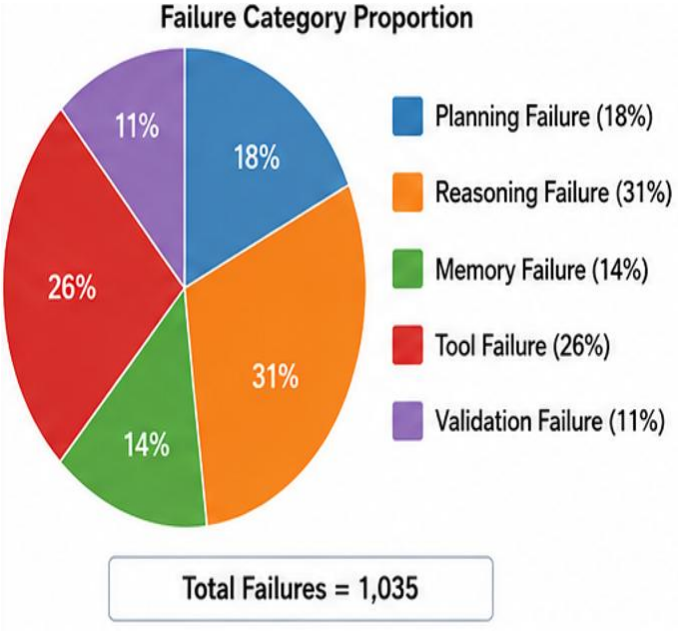

Fig. 12: Failure Category Proportion

The Failure Category Proportion chart shows the proportion of failures categorized by the type of failure. The biggest percentage is Reasoning Failure (31%) followed by Tool Failure (26%). 18% is due to Planning Failure, 14% is due to Memory Failure, and 11% is due to Validation Failure. The results of these findings indicate that every process of cognitive decision making and process of tool execution is the major source of reliability problem and special mitigation strategy is needed to improve the performance of the agents.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

B. Discussion

TABLE 1: Summary Statistics

Metric	Value
Total Tasks Evaluated	5,000
Overall Success Rate	88.0%
Overall Failure Rate	12.0%
Average Execution Time	18.7 sec
Recovery Rate	76.0%
Output Consistency Score	89.0%
Planning Agent Success Rate	94.0%
Reasoning Agent Success Rate	87.0%
Memory Agent Success Rate	90.0%
Tool Agent Success Rate	82.0%
Validation Agent Success Rate	96.0%

The results highlight the degradation in reliability when the complexity of the task goes higher in multi-step agentic AI systems. The most frequently appearing reliability problem is Reasoning Failure, which suggests that logical decision making is still a big hurdle for autonomous agents. Furthermore, other performance degradations are due to failures within the tool itself, underscoring the potential problems with the interaction of external systems. The analysis also found that workflow length had a strong link with error propagation for example, errors made at early phases had impact on subsequent decisions. Though 88% are successful at the overall level, there are varying success rates across different agents' stages, highlighting the need for validation mechanisms, recovery strategy and the monitoring of workflow.

C. Limitation

- The use of the dataset not captures all the operational uncertainties and conditions of agentic AI system deployment in the real world.

- A single multiagent architecture and dataset are used, so it is complex enough for domain applications.

Conclusion And Future Work

The present research aimed to explore the reliability problems of multi-step agentic AI systems using an experimental setup created in Python. The results showed that increasing the complexity of the task greatly compromises the reliability of the system and increases the execution time. Sources of performance degradation identified are reasoning and tool-related failures which are the more significant factors, and error propagation is viewed as a significant issue in the outcomes of the workflows. It found that errors that took place at the earlier phases of execution would impact later phases, causing cascading reliability problems. While the entire success rate is fairly consistent, each agent did not perform equal well, leading to the conclusion that there is a need for improved validation and recovery operations. The evaluative scheme proposed yielded positive results in measuring the reliability with several indicators, such as success, failure ratio, execution time, recovery and consistency of execution result. The findings can help to generate a better and more reliable framework of agent-based AI that can be a foundation for further development of such AI agents.

More advance mechanism of learning with self-correction, reinforcement learning, or adaptive feedback loops, etc. will be more effective in the case of agentic AI system. In addition, one should explore explainable AI (XAI) systems and methods that will create a more transparent and trustful relationship with the autonomous decision-making cycles as this will be the key to future efforts. Generalizing it to more data sets and other use cases such as healthcare, cyber security, finance, software development, etc. will further push the generalizability of results.

References

[1] Xing, L., 2025. Looking Forward: Challenges and Opportunities in Agentic AI Reliability. *arXiv preprint arXiv:2511.11921*.
[2] Hughes, L., Dwivedi, Y.K., Malik, T., Shawosh, M., Albashrawi, M.A., Jeon, I., Dutot, V., Appanderanda, M., Crick, T., De', R. and Fenwick, M., 2025. AI agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, *65*(4), pp.489-517.
[3] Raheem, T. and Hossain, G., 2025, May. Agentic ai systems: Opportunities, challenges, and trustworthiness. In *2025 IEEE International Conference on Electro Information Technology (eIT)* (pp. 618-624). IEEE.
[4] Flehmig, N., Lundteigen, M.A. and Yin, S., 2025. Perspectives on a Reliability Monitoring Framework for Agentic AI Systems. *arXiv preprint arXiv:2511.09178*.
[5] Mehta, S., 2025. Beyond Accuracy: A Multi-Dimensional Framework for Evaluating Enterprise Agentic AI Systems. *arXiv preprint arXiv:2511.14136*.
[6] Crespo-Márquez, A. and Fernandez, J.F., Agentic AI for Maintenance Management: A Process-Centric Review and a Staged Framework for Industrial Adoption. *Available at SSRN 5985157*.
[7] Bandi, A., Kongari, B., Naguru, R., Pasnoor, S. and Vilipala, S.V., 2025. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, *17*(9), p.404.
[8] Abou Ali, M., Dornaika, F. and Charafeddine, J., 2025. Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review*, *59*(1), p.11.
[9] Wang, H., Gong, J., Zhang, H., Xu, J. and Wang, Z., 2025. Ai agentic programming: A survey of techniques, challenges, and opportunities. *arXiv preprint arXiv:2508.11126*.
[10] Zhang, G., Niu, L., Fang, J., Wang, K., Bai, L. and Wang, X., 2025. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*.
[11] Hu, Y., 2024. Autonomous agent architecture for complex tasks via hierarchical planning and language model reasoning. *Transactions on Computational and Scientific Methods*, *4*(12).
[12] Brohi, S., Mastoi, Q.U.A., Jhanjhi, N.Z. and Pillai, T.R., 2025. A research landscape of agentic ai and large language models: Applications, challenges and future directions. *Algorithms*, *18*(8), p.499.

Journal of Information Systems Engineering and Management
2026, 11(3s)

e-ISSN: 2468-4376

https://www.jisem-journal.com/

Research Article

[13] Shukla, M., 2025. Evaluation and Benchmarking of Generative and Agentic AI Systems: A Comprehensive Survey. *Available at SSRN 5927324*.
[14] Maheshkar, J.A., 2025. Bridging the Gap: A Systematic Framework for Agentic AI Root Cause Analysis in Hybrid Distributed Systems. *Acta Scientiae*, *26*(1), pp.228-245.
[15] Passi, S., 2025. Agentic AI has a Human Oversight Problem. *Available at SSRN 5529058*.
[16] Goyal, S., 2025. A Critical Review of Agentic AI: Core Technologies, Applications, Ethical Implications, and Future Research Directions. *Jurnal Masyarakat Informatika*, *16*(2), pp.268-283.
[17] Akbar, M.A., Khan, A.A., Hamza, M., Ghaffar, A. and Hajikhani, A., 2025. Agentic AI in Software Engineering: Practitioner Perspectives Across the Software Development Life Cycle. *Software Engineering: Practitioner Perspectives Across the Software Development Life Cycle (September 16, 2025)*.
[18] Khan, R., Joyce, D. and Habiba, M., 2025. AGENTSAFE: A Unified Framework for Ethical Assurance and Governance in Agentic AI. *arXiv preprint arXiv:2512.03180*.
[19] Banerjie, S., Zhu, Y., Freeman, I., Machado, J.V., Ahmed, A., Sarker, A. and Al-Garadi, M., 2025. Agentic AI in Healthcare: A Comprehensive Survey of Foundations, Taxonomy, and Applications. *Authorea Preprints*.
[20] Collaco, B.G., Haider, S.A., Prabha, S., Gomez-Cabello, C.A., Genovese, A., Wood, N.G., Bagaria, S.P., Gopala, N., Tao, C. and Forte, A.J., 2025. The role of agentic artificial intelligence in healthcare: a systematic review.
[21] Raza, S., Sapkota, R., Karkee, M. and Emmanouilidis, C., 2025. Responsible Agentic Reasoning and AI Agents: A Critical Survey. *Authorea Preprints*.
[22] Konda, R., 2025. Agentic AI for Software Development: Autonomous Agents in Requirements Engineering, Testing, and Deployment. *International Journal of Emerging Research in Engineering and Technology*, pp.139-148.
[23] Raza, S., Sapkota, R., Karkee, M. and Emmanouilidis, C., 2025. Responsible agentic reasoning and ai agents: A critical survey: Proposal for safe agentic AI via responsible reasoning ai agents (r2a2). *SuperIntelligence-Robotics-Safety & Alignment*, *2*(6).
[24] Cherukuri, R. and Yarram, V.K., 2024. From Intelligent Automation to Agentic AI: Engineering the Next Generation of Enterprise Systems. *International Journal of Emerging Research in Engineering and Technology*, *5*(4), pp.142-152.
[25] Nicoletti, B. and Appolloni, A., 2025. A digital twin framework for enhancing human–agentic AI–machine collaboration. *Journal of Intelligent Manufacturing*, pp.1-17.
[26] Tao, Z., Xu, W. and You, X., 2025. Toward Trustworthy Digital Twins in Agentic AI-based Wireless Network Optimization: Challenges, Solutions, and Opportunities. *arXiv preprint arXiv:2511.19961*.