IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

# Beyond Chatbots: Building More Reliable and Context-Aware AI Conversational Agents

**Sanap Shital Sunil[1], Prof. K. M. Borude[2], Prof. L. S. Bhadake[3], Dr. H. B. Jadhav[4]**
Student, Department of Computer Engineering[1]
Guide, Department of Computer Engineering[2]
Co-Guide, Department of Computer Engineering[3]
HOD, Department of Computer Engineering[4]
Adsul's Technical Campus, Chas, Ahilyanagar, India

**Abstract:** *Conversational artificial intelligence has changed the way people interact with digital systems. Users can now ask questions in everyday language, receive explanations, obtain recommendations, and complete tasks without learning complicated software interfaces. However, a conversation that sounds natural is not always accurate or dependable. An AI assistant may misunderstand a question, lose track of earlier information, produce an unsupported answer, or use an external tool in an inappropriate way. These issues have become increasingly important as conversational systems move beyond simple question answering towards more autonomous tasks.This paper examines recent developments in AI-based conversational agents through a critical review of selected research published between 2024 and 2026. The 2022 scoping review by Kusal et al. provides the foundation for understanding earlier conversational technologies, including natural-language processing, deep learning, context management, and emotion-aware dialogue. More recent studies extend this discussion to large language models, autonomous agents, retrieval-augmented generation, hallucination mitigation, and agent evaluation.The review suggests that improving the language model alone is not sufficient to address the practical challenges of conversational AI. Reliability also depends on how information is retrieved, how previous conversation details are managed, how responses are checked, and how risky actions are controlled. To address these concerns, this paper proposes a Trustworthy Context-Grounded Conversational Agent Framework. The framework combines context management, evidence-based response generation, response verification, privacy controls, restricted tool access, and human intervention when necessary. The proposed framework has not yet been experimentally evaluated; therefore, its effectiveness must be established through future implementation and comparative testing*

**Keywords**: Conversational AI, chatbots, large language models, context management, retrieval-augmented generation, trustworthy AI, agent evaluation, human oversight

## I. INTRODUCTION

Imagine asking a digital assistant the same question in two different ways. A traditional chatbot may understand the first sentence but fail to recognize that the second sentence expresses the same intention. A modern AI assistant may handle both questions, remember the earlier discussion, and provide a detailed response. This difference illustrates how conversational artificial intelligence has developed over time.

Conversational agents are software systems that communicate with people through text or speech. They are increasingly used in education, healthcare, online shopping, customer service, and information access. Their purpose is not simply to return predefined answers. Ideally, they should understand what a person wants, consider the context, and provide a useful response.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

Earlier systems commonly depended on fixed rules, keywords, or predefined dialogue patterns. Machine-learning and deep-learning approaches improved their ability to recognize user intentions and process natural language. Transformer-based language models introduced another major development by allowing systems to generate more flexible and detailed responses.

The review published by Kusal et al. in 2022 provides an important starting point for understanding these technologies [1]. It discusses conversational-agent architectures, natural-language understanding, response generation, deep-learning methods, emotional information, and the challenges associated with context and training data.

Since then, conversational AI has developed rapidly. Research in 2024 explored large language model agents and their applications [2], [3]. Studies published in 2025 placed greater emphasis on hallucinations, reliability, and evaluation [4]–[6]. Research in 2026 further examined autonomous agents, tool use, and the assessment of complex agent behaviour [7], [8].

These developments raise an important question: if AI systems can already generate fluent and convincing responses, what is needed to make them dependable in real situations?

The problem is that fluent language can hide uncertainty. An answer may appear reasonable even when it is based on incomplete information. An assistant may also retain irrelevant details, misunderstand the user's intention, or make an incorrect decision while using an external service. These limitations become especially important when a conversational system is used for tasks involving sensitive information or consequential decisions.

This paper argues that conversational agents should be designed around reliability as well as language generation. It reviews recent research, identifies connected design challenges, and proposes a framework that combines context, evidence, verification, privacy, and controlled actions.

### 1.1 Research problem

Existing conversational-agent technologies offer increasingly flexible interaction, but the quality of a conversation does not necessarily indicate whether its information is correct or its actions are safe.

The main research problem addressed in this paper is how to organize the different components of a conversational agent so that the system can use context appropriately, produce evidence-supported answers, recognize uncertainty, and avoid unauthorized actions.

### 1.2 Objectives

The objectives of this research are:

- To understand the development of conversational agents from traditional dialogue systems to modern LLM-based assistants.
- To critically examine selected research published between 2024 and 2026.
- To identify challenges involving factual reliability, context retention, privacy, and agent evaluation.
- To examine the limitations of a conventional conversational-agent architecture.
- To propose an integrated framework for more reliable and context-aware conversational interaction.
- To identify suitable methods for evaluating the proposed framework in future work.

## II. BACKGROUND AND EVOLUTION OF CONVERSATIONAL AGENTS

### 2.1 From predefined responses to generative systems

The earliest conversational systems generally followed predefined rules. When a user entered a recognized phrase, the system selected a response associated with that pattern. This approach was useful for simple tasks, such as answering frequently asked questions, but it struggled with unfamiliar wording and unexpected follow-up questions.

Machine-learning techniques made it possible to identify patterns in user messages. Natural-language understanding could be used to classify intentions, identify entities, and extract relevant information. Deep-learning methods further improved the representation of language and the handling of complex linguistic patterns.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

Modern language models have expanded these capabilities. Instead of selecting only from a fixed collection of answers, a generative model can produce a response based on the user's message and the surrounding context.

However, each approach has limitations. Rules can be predictable but inflexible. Learned models may misclassify an intention. Generative models can produce detailed answers that contain unsupported information. The development of conversational AI should therefore be understood as an improvement in capability, not as the complete removal of earlier problems.

### 2.2 Large language models and autonomous agents

Large language models have made conversational systems more useful for open-ended questions, summarization, explanation, and multi-turn dialogue. Some systems now connect language models to external tools, databases, search services, and task-planning components.

Wang et al. [2] reviewed LLM-based autonomous agents, examining their construction, applications, evaluation, and future challenges. Their work is relevant because an agent may need to perform several steps rather than produce a single response.

For example, an assistant asked to summarize recent research may need to search for articles, identify relevant sources, compare findings, and prepare a summary. If the system retrieves an incorrect article or misunderstands a source, the final response may also be unreliable.

Consequently, an autonomous conversational agent needs more than language-generation capability. It also needs suitable controls over the information it uses and the actions it performs.

### 2.3 Retrieval-augmented generation

Retrieval-augmented generation, commonly known as RAG, connects a language model to an external collection of information. The system retrieves documents related to a user's question and uses them as supporting material when generating an answer.

This approach is useful when information must come from a particular document collection or needs to be updated independently of the language model. Gupta et al. [9] reviewed the development of RAG methods and their role in knowledge-intensive language tasks.

Nevertheless, retrieval is not a guarantee of correctness. The system may select an irrelevant document, retrieve outdated information, or produce a response that goes beyond what its sources establish. A reliable system should therefore examine both the quality of retrieved evidence and the relationship between that evidence and the final answer.

### 2.4 Context and personalization

Human conversation depends heavily on context. A person may say, “What about the second option?” without repeating the full question. A conversational agent needs access to the relevant earlier information to understand the request.

Kusal et al. [1] identified context understanding and the representation of dialogue state as important issues in conversational-agent design. Modern systems can use conversation history and memory mechanisms to maintain continuity, but more memory is not necessarily better.

Retaining every detail may introduce privacy concerns and irrelevant information. A better design should identify which information is useful, determine whether it may be retained, and avoid reusing sensitive details without an appropriate basis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

## III. CRITICAL LITERATURE REVIEW: 2024–2026

### 3.1 Developments in 2024

Research in 2024 examined conversational agents from several perspectives. Casheekar et al. [3] reviewed conversational AI systems, including ChatGPT, their applications, open challenges, and future research directions. Their work helps explain the growing role of generative systems in everyday interaction.

Wang et al. [2] examined LLM-based autonomous agents. Their review broadens the discussion from generating dialogue to designing systems that can plan, use tools, and perform tasks. This shift creates additional questions about how an agent should select actions and determine whether a task has been completed correctly.

Ding et al. [10] examined evaluation frameworks for AI conversational agents used in health interventions. Their work illustrates why an evaluation method must reflect the purpose of the system. A chatbot used for general information and a chatbot supporting a health intervention cannot be evaluated solely by the same measures of conversational quality.

Taken together, these studies show that conversational AI is not a single technology with a single performance measure. Its usefulness depends on the task, the information available, and the consequences of an incorrect answer.

**Critical observation:** Research in 2024 provides a broader understanding of agent capabilities and applications. However, increased flexibility also creates new responsibilities. A system that can search and act needs mechanisms for checking its evidence and controlling its actions. These requirements should be considered part of the architecture rather than optional improvements added after deployment.

### 3.2 Developments in 2025

Research published in 2025 gave substantial attention to hallucinations and the evaluation of language-model-based agents.

Huang et al. [4] surveyed hallucinations in large language models, examining their principles, categories, challenges, and mitigation approaches. Zhang et al. [5] also reviewed hallucination research, including detection, explanation, and mitigation.

These studies address a central limitation of generative AI: a model may produce an answer that sounds convincing without sufficient factual support. This problem is particularly important when users cannot easily verify the response themselves.

Mohammadi et al. [6] examined evaluation and benchmarking of LLM agents. Their work is relevant because agent evaluation must consider more than the quality of the final sentence. It may also need to examine whether the system used appropriate tools, followed instructions, completed the task, and behaved safely when something went wrong.

**Critical observation:** The 2025 literature makes reliability a central research concern. However, detecting hallucinations, retrieving evidence, and evaluating agent behaviour are distinct tasks. A retrieval component cannot independently guarantee factual correctness, and a benchmark cannot prevent a failure during a live interaction. A useful architecture must connect these functions so that evidence and safety checks can influence the response before it reaches the user.

### 3.3 Developments in 2026

Research in 2026 has continued to examine agents that operate with greater autonomy and interact with external systems.

Bajoria et al. [7] reviewed the transition from language models to agentic AI, including autonomous operation, tool use, collaboration, and governance. This development is important because a conversational assistant may now be expected to carry out actions rather than merely provide suggestions.

Yehudai et al. [8] surveyed evaluation methods for LLM-based agents. Their work highlights the need to evaluate agents in ways that reflect their actual behaviour, including their ability to handle complex tasks and operate reliably.

Research on RAG also continues to evolve, as illustrated by the review from Wu et al. [11]. This body of work provides useful background for systems that combine generated responses with retrieved information.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

**Critical observation:** The move towards agentic systems increases the importance of access control and failure management. An incorrect informational answer may mislead a user, whereas an incorrect external action may change a record, submit a request, or trigger another operation. The system architecture should therefore distinguish between providing information and performing an action.

### 3.4 Comparison of research directions

| Period | Main research emphasis | Critical insight |
|---|---|---|
| 2024 | Conversational AI applications and autonomous-agent design | Greater capability requires attention to how agents use information and tools |
| 2025 | Hallucinations, reliability, and benchmarking | Fluent answers need independent evaluation and evidence checks |
| 2026 | Agentic AI, tool use, and realistic evaluation | Systems need safeguards for actions, failures, and oversight |

This comparison represents a thematic synthesis of selected publications. It is not a complete statistical analysis of all papers published during these years.

## IV. RESEARCH GAP AND PROPOSED CONTRIBUTION

The literature reviewed in this paper identifies several connected opportunities for further research.

**First, evidence retrieval and response verification need to work together.** Retrieval can provide useful information, but the final answer may still contain claims that are not supported by the retrieved material.

**Second, context management must balance continuity and privacy.** A system should use relevant conversation history without treating every past message as information that must be retained indefinitely.

**Third, agent evaluation should include the process as well as the outcome.** A correct final answer does not necessarily mean that the system used appropriate sources or followed safe procedures.

**Fourth, uncertainty requires an explicit response strategy.** When the available evidence is inadequate or a request is ambiguous, the system should be able to ask a question, acknowledge uncertainty, or involve a human.

These are research opportunities derived from the selected literature. They should not be interpreted as proof that existing conversational systems have never implemented these features.

### 4.1 Original contribution

This paper proposes the **Trustworthy Context-Grounded Conversational Agent Framework (TCG-CAF)**.

The proposed framework connects six responsibilities:

Understanding the user's request and relevant context.

Retrieving information from approved sources.

Generating a response based on the available evidence.

Checking the response for unsupported claims and policy violations.

Restricting external actions through permissions and risk controls.

Monitoring performance and escalating uncertain or high-risk cases.

The contribution is a conceptual integration of these responsibilities into one reference design. It is not a newly trained language model, and its effectiveness has not yet been established experimentally.

## V. RESEARCH METHODOLOGY

### 5.1 Research approach

This study uses a focused literature synthesis and conceptual framework design. The uploaded paper by Kusal et al. [1] provides the foundational understanding of conversational-agent technologies. Selected publications from 2024, 2025,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, October 2026 Impact Factor: 8.2

and 2026 are then examined to identify developments in LLM agents, hallucination mitigation, retrieval, and evaluation.

The approach is appropriate for developing a research framework because the selected literature covers different technologies and application domains. However, this paper does not claim to be a completed systematic review with exhaustive database searches and independent screening.

### 5.2 Literature selection

The selected sources were considered according to their relevance to the research questions. Priority was given to scholarly articles, reviews, and conference papers that address at least one of the following areas:

Conversational-agent architecture and language technologies.

LLM-based agents and external tool use.

Retrieval-augmented generation.

Hallucination detection and mitigation.

Evaluation, safety, privacy, and responsible AI.

The 2022 review was included as a foundational source, while the recent publications were used to examine developments in the requested three-year period.

### 5.3 Thematic analysis

The selected literature was organized into five themes: architecture, retrieval and grounding, context management, reliability and evaluation, and safety and privacy.

The analysis compared the issues discussed across these themes and identified requirements that could be addressed through a common architecture. Those requirements informed the design of the proposed TCG-CAF framework.

Because the literature selection and screening have not been documented as an exhaustive systematic process, the findings should be interpreted as a focused critical synthesis rather than a comprehensive census of all research.

### 5.4 Future experimental validation

The proposed framework should be implemented and compared with a baseline conversational agent before any claims about performance improvement are made.

A future evaluation should include questions with verified answers, questions for which the knowledge base contains no answer, ambiguous requests, multi-turn conversations, conflicting sources, tool failures, and privacy-sensitive scenarios.

The comparison should use the same test set and, where practical, the same language model and knowledge sources. This would help determine whether the proposed safeguards provide measurable benefits beyond those of a simpler system.

## VI. EXISTING SYSTEM ARCHITECTURE

A conventional conversational agent can be represented as a sequence of processing stages. The user submits a message, the system processes the input, identifies the likely intention, maintains the dialogue state, and generates a response.

The following figure is a simplified conceptual representation based on the workflow described by Kusal et al. [1].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

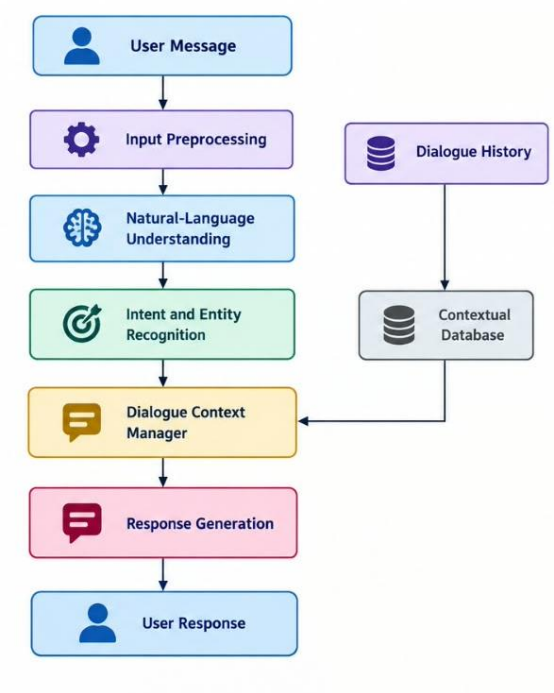



**Fig.1: Existing conversational-agent architecture**

**Working of the existing system**

The input-processing component prepares the message for analysis. Natural-language understanding then identifies the likely meaning of the request. Intent and entity recognition determine what the user wants and which important details appear in the message.

The dialogue manager uses relevant conversational context to support the next response. A response-generation component then constructs the reply and returns it to the user. Some systems store dialogue history to support follow-up questions.

This design provides a useful foundation for ordinary conversations. However, the simplified model does not explicitly include independent evidence verification, permission-controlled external actions, or a risk-based escalation mechanism.

These missing components are limitations of the simplified architecture shown here, not proof that every conventional system lacks them.

## VII. PROPOSED SYSTEM ARCHITECTURE

The proposed TCG-CAF framework extends the existing design by introducing explicit controls for information grounding, response verification, and external actions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

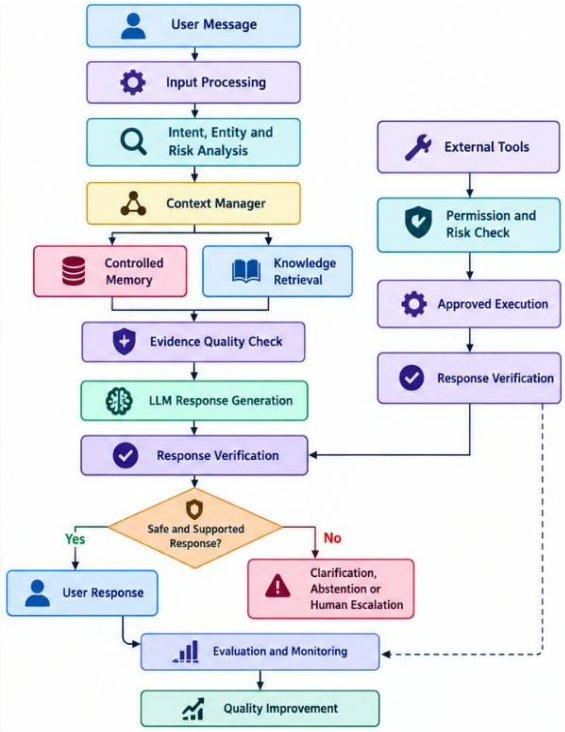



**Fig. 2: Proposed TCG-CAF architecture**

### 7.1 Main components

**Context manager:** Maintains the information required for the current conversation and selects relevant earlier details.

**Controlled memory:** Stores only information permitted by the system's retention and privacy rules. Memory should not be treated as an unrestricted record of every user interaction.

**Knowledge retrieval:** Searches approved documents or databases for information relevant to the request.

**Evidence quality check:** Examines whether the retrieved information is sufficiently relevant and suitable for answering the question.

**LLM response generation:** Produces the response using the request, relevant context, and retrieved evidence.

**Response verification:** Checks whether important claims are supported by the available sources and whether the response follows applicable policies. Verification can also fail, so it must be evaluated rather than assumed to be infallible.

**Permission and risk check:** Determines whether an external action is permitted and whether confirmation or human review is required.

**Clarification and escalation:** Provides a safe alternative when the system lacks sufficient evidence, encounters an ambiguous request, or detects a high-risk situation.

**Evaluation and monitoring:** Tracks performance, failure cases, and safety-related outcomes so that the system can be improved through controlled updates.

### 7.2 Expected benefits

The proposed architecture is intended to make conversational-agent behaviour easier to assess and govern. Evidence retrieval may help ground answers in approved information. Response verification may identify unsupported claims. Controlled memory may reduce unnecessary retention, while permission checks may prevent some unintended external actions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

These are expected benefits rather than measured findings. The framework may also introduce additional processing time, operational costs, and new points of failure. Its overall value must be established through implementation and testing.

## VIII. PROPOSED EVALUATION FRAMEWORK

A meaningful evaluation should examine several dimensions rather than relying on a single measure of conversational quality.

| Dimension | Possible metric | Purpose |
|---|---|---|
| Answer accuracy | Expert-verified accuracy | Measures factual correctness |
| Evidence grounding | Supported-claim ratio | Examines whether claims are supported by sources |
| Retrieval quality | Precision@k or Recall@k | Measures the relevance of retrieved information |
| Context handling | Multi-turn task success | Evaluates use of relevant conversation history |
| Safety | Unsafe-response rate | Measures policy-violating responses |
| Tool use | Authorized-task success rate | Assesses permitted action execution |
| Privacy | Unauthorized disclosure rate | Examines inappropriate information exposure |
| Efficiency | Latency and cost per task | Measures operational overhead |

The supported-claim ratio can be defined as:

$$G = \frac{\text{Number of supported checkable claims}}{\text{Total number of checkable claims}}$$

Where:

G = Supported-claim ratio.

Supported checkable claims = Claims supported by reliable evidence.

Total checkable claims = All claims that can be verified.

Interpretation: The value of G ranges from 0 to 1. A value closer to 1 indicates that more claims are supported by evidence.

This measure should be used alongside factual accuracy because a response can cite a source while still misrepresenting it.

The proposed framework should be compared with a baseline system using identical test cases wherever possible. Evaluation should also include cases in which the correct behaviour is to abstain, request clarification, or escalate rather than answer immediately.

No performance scores are reported here because the proposed framework has not been implemented or tested.

## IX. DISCUSSION

The literature reviewed in this paper shows that conversational AI is developing from systems designed mainly to produce responses into systems expected to maintain context, retrieve information, and complete tasks using external tools.

This transition changes how reliability should be understood. In a simple chatbot, the primary question may be whether the response is appropriate. In a tool-enabled agent, the evaluation must also consider the information used, the actions taken, and the consequences of a failure.

The proposed TCG-CAF framework addresses this broader problem by separating retrieval, generation, verification, and action control. Each component has a defined responsibility, making the overall workflow easier to describe and evaluate.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

The framework also recognizes that a dependable assistant should not always attempt to answer. When the evidence is incomplete, asking a clarifying question or acknowledging uncertainty may be more useful than producing a confident but unsupported statement.

Nevertheless, the framework has limitations. Retrieved documents may be inaccurate, verification mechanisms may miss errors, and risk classifiers may incorrectly assess a request. Additional processing stages can also increase latency and cost.

The contribution of this paper is therefore a structured proposal for future investigation, not evidence that the proposed architecture already outperforms existing systems.

## X. CONCLUSION

Conversational AI has progressed considerably from predefined dialogue patterns to systems that can generate detailed responses, use external information, and perform multi-step tasks. The research reviewed from 2024 to 2026 demonstrates the growing importance of agent architecture, hallucination mitigation, evaluation, and responsible deployment.

At the same time, improved language generation does not remove the need to verify information, manage context carefully, and control external actions. A conversational agent must be assessed not only by how naturally it communicates but also by how it handles uncertainty and behaves when it cannot complete a task reliably.

This paper proposes the Trustworthy Context-Grounded Conversational Agent Framework as a way to organize these responsibilities within one architecture. The framework combines context management, approved knowledge retrieval, evidence-aware generation, response verification, privacy controls, and risk-based escalation.

Its contribution is a conceptual design that can guide implementation and future experiments. Whether it provides better factual reliability, privacy, or safety remains to be demonstrated through comparative evaluation. The next step is therefore to build and test the proposed system, using measurable criteria rather than relying on architectural expectations alone.

## ACKNOWLEDGEMENT

The authors acknowledge the researchers whose published work on artificial intelligence, conversational agents, large language models, context management, information retrieval, and trustworthy AI informed this study. Their contributions provided the foundation for examining the challenges of reliability, contextual understanding, privacy, and safe response generation in AI conversational agents. Any datasets or external resources used in future experimental evaluations will be acknowledged in accordance with their respective licences and citation requirements.

## REFERENCES

[1] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra, and A. Abraham, "AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions," *IEEE Access*, vol. 10, pp. 92337–92356, 2022. DOI: 10.1109/ACCESS.2022.3201144.

[2] L. Wang et al., "A Survey on Large Language Model Based Autonomous Agents," *Frontiers of Computer Science*, vol. 18, 2024. DOI: 10.1007/s11704-024-40231-1.

[3] A. Casheekar, A. Lahiri, K. Rath, K. S. Prabhakar, and K. Srinivasan, "A Contemporary Review on Chatbots, AI-Powered Virtual Conversational Agents, ChatGPT: Applications, Open Challenges and Future Research Directions," *Computer Science Review*, vol. 52, Art. no. 100632, 2024. DOI: 10.1016/j.cosrev.2024.100632.

[4] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, 2025. DOI: 10.1145/3703155.

[5] Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *Computational Linguistics*, vol. 51, no. 4, 2025. DOI: 10.1162/coli.a.16.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, October 2026
Impact Factor: 8.2

[6] M. Mohammadi, Y. Li, J. Lo, and W. Yip, "Evaluation and Benchmarking of LLM Agents: A Survey," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2*, 2025. DOI: 10.1145/3711896.3736570.

[7] S. Bajoria et al., "From Language Models to Agentic AI: A Survey of Autonomous, Action-Enabled, and Collaborative LLM Agents," *Cognitive Computation*, 2026. DOI: 10.1007/s12559-026-10619-1.

[8] A. Yehudai et al., "A Survey on Evaluation of LLM-based Agents," in *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. DOI: 10.18653/v1/2026.findings-acl.1330.

[9] S. Gupta, R. Ranjan, and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," arXiv:2410.12837, 2024. DOI: 10.48550/arXiv.2410.12837.

[10] H. Ding, J. Simmich, A. Vaezipour, N. Andrews, and T. Russell, "Evaluation Framework for Conversational Agents with Artificial Intelligence in Health Interventions: A Systematic Scoping Review," *Journal of the American Medical Informatics Association*, vol. 31, no. 3, pp. 746–761, 2024. DOI: 10.1093/jamia/ocad222.

[11] S. Wu et al., "Retrieval-Augmented Generation for Natural Language Processing: A Survey," *Artificial Intelligence Review*, 2026. DOI: 10.1007/s10462-026-11605-7.