



SupportGuardAI: A Domain-specific Conversational AI Framework for Hallucination-Free Enterprise Customer Support

Kshitija churi¹, Lalita Sav²

¹(Department Master of Computer Application, VIVA Institute of Technology, India)

²(Department Master of Computer Application, VIVA Institute of Technology, India)

Abstract : *Conversational AI systems are being used more and more in contemporary businesses to automate customer care tasks and boost response times. However, hallucinated responses—information that seems convincing but is factually inaccurate or unsupported—are frequently produced by general-purpose conversational AI models, posing significant hazards in business settings. Current customer support systems either use massive language models without adequate domain grounding or rely on strict rule-based techniques. In order to reduce hallucinations in business customer service, this article introduces SupportGuardAI, a domain-specific conversational AI platform. To guarantee accurate and dependable interactions, the suggested system incorporates enterprise knowledge bases, retrieval-based grounding, answer verification techniques, and consistency constraints. To provide reliable enterprise-level customer service, the framework places a strong emphasis on regulated response generation, descriptive analytics, and structured knowledge access.*

Keywords: Conversational AI, Enterprise Customer Support, Hallucination Reduction, Domain-specific AI, Knowledge-grounded Systems

1. Introduction

Conversational AI's capacity to effectively handle high client query volumes has led to its rapid acceptance in enterprise customer care. Businesses use chatbots and virtual assistants to help with billing questions, technical problems, service requests, and policy explanations. Large language models have advanced, but using them in corporate settings is still difficult. Large general-purpose datasets are used to train these models, and the results they provide might not be in line with company rules or existing knowledge. These hallucinogenic reactions may result in false information, noncompliance, and disgruntled customers.

The majority of workplace chatbots now in use either use generative AI without sufficient domain grounding or adhere to strict rule-based procedures that restrict flexibility. SupportGuardAI, a domain-specific conversational AI system that guarantees hallucination-free responses through regulated knowledge access and verification procedures, is proposed in this research to address this difficulty.

2. Literature Review

Research on conversational AI and enterprise customer support spans multiple domains, including natural language processing, dialogue systems, knowledge-based systems, and business analytics. Early conversational agents relied on rule-based and decision-tree approaches, which provided deterministic and reliable responses but lacked flexibility and scalability in handling diverse customer queries.

With the advancement of machine learning and deep learning techniques, data-driven conversational AI

systems emerged. These systems, particularly those based on large language models (LLMs), demonstrated strong natural language understanding and generation capabilities. However, several studies have highlighted that generative models often produce hallucinated responses— outputs that are fluent but factually incorrect or unsupported by source knowledge—posing challenges in enterprise and high-risk domains.

2.1 Conversational AI in Enterprise Customer Support

Existing enterprise chatbot solutions primarily focus on automating frequently asked questions and basic support workflows. Retrieval-based chatbots improve reliability by matching user queries with predefined answers stored in knowledge bases. While these systems reduce incorrect responses, they struggle with complex, multi-turn, or contextual queries.

Recent research indicates that hybrid conversational systems combining retrieval mechanisms with generative models provide improved flexibility. However, without strict domain constraints, these systems may still generate responses that deviate from enterprise policies or operational guidelines.[2][6][7]

Existing AI-based support systems lack sufficient controls to prevent hallucinated responses.

- To study existing hallucination mitigation approaches Techniques such as retrieval augmentation and fine- tuning were reviewed to understand their limitations.
- **To identify gaps in domain-specific conversational frameworks**
 - Most systems do not enforce strict enterprise knowledge boundaries.
- **To propose a hallucination-free conversational AI framework**
 - SupportGuardAI integrates knowledge grounding, verification, and response constraints.
- **To improve reliability and trust in enterprise customer support**
 - The framework aims to deliver consistent, accurate, and compliant responses [3][4][1]

3. Methodology

This research adopts a qualitative and conceptual methodology. A comprehensive review of literature related to conversational AI, hallucination mitigation, and enterprise knowledge systems was conducted.

The study analyzes existing conversational frameworks, identifies limitations, and proposes an integrated system architecture. No real customer data is used, ensuring ethical compliance. The proposed framework is evaluated conceptually through comparative analysis with existing generative AI-based customer support systems.[5]

4. Proposed System: SupportGuard

SupportGuardAI is a domain-specific conversational AI framework designed for enterprise customer support with minimal hallucination risk.

The proposed framework has four principal layers:

4.1 Hallucination in Large Language Models

Hallucination has been identified as a critical limitation of large language models, especially in applications requiring factual accuracy. Studies by Ji et al. and others categorize hallucinations into intrinsic hallucinations (incorrect generation contradicting source data) and extrinsic hallucinations (generation of unsupported or fabricated information).

Researchers have proposed several mitigation techniques, including prompt engineering, model fine-tuning, and post-generation verification. While these methods reduce hallucination rates, they do not fully eliminate the issue, particularly when models lack direct access to domain-specific knowledge.[3][4][9]

4.2 Knowledge-Grounded and Retrieval-Augmented Systems

By conditioning model outputs on validated external knowledge sources, knowledge-grounded conversational AI systems seek to increase answer accuracy. Frameworks for Retrieval-Augmented Generation (RAG) obtain pertinent materials and use them as context when creating responses. RAG-based methods greatly increase factual accuracy in knowledge-intensive activities, according to numerous studies.

Nevertheless, the majority of RAG implementations currently in use are generic and do not impose enterprise-specific limitations such as controlled response bounds, policy validation, or compliance checks. Their direct application in corporate customer service settings is limited by this constraint.[1][5]

4.3 Enterprise Constraints and Trustworthiness

Enterprise customer support systems operate under strict requirements including data privacy, regulatory compliance, and consistent policy interpretation. Research in trustworthy AI emphasizes the need for transparency, accountability, and controllability in automated systems.

Despite this, many conversational AI solutions prioritize response fluency over correctness. Limited attention has been given to verification layers and fallback mechanisms that prevent speculative responses when knowledge is insufficient.[8][10]

4.4 Research Gap Identification

From the reviewed literature, it is evident that:

- Rule-based systems offer reliability but lack flexibility.
- Generative AI systems offer flexibility but suffer from hallucinations.
- Retrieval-augmented systems improve grounding but lack enterprise-level constraints.

There exists a clear research gap for a domain-specific, enterprise-oriented conversational AI framework that integrates knowledge grounding, hallucination mitigation, and response verification within a unified architecture. This gap forms the foundation for the proposed Support GuardAI framework.[2][3][4][1]

5. Problem Definition

Conversational AI is being used more and more by enterprise customer support systems, however these systems lack established ways to guarantee domain conformance and response accuracy. Instead of using validated enterprise knowledge, general-purpose AI models produce replies based on probabilistic patterns.

This leads to inaccurate justifications for policies, manufactured troubleshooting procedures. Reactions that are inconsistent or noncompliant.

The lack of verification and grounding methods makes it difficult for organizations to trust AI-driven customer assistance.

The Problem's Mathematical Basis

Let $Q = \{q_1, q_2, \dots, q_n\}$ represent client queries. Enterprise knowledge sources are denoted by $K = \{k_1, k_2, \dots, k_m\}$. H stands for the likelihood of hallucinations. A stands for reaction accuracy.

In conventional generative systems:

$$A = f(Q)$$

However, response limitations and confirmed enterprise knowledge determine accuracy. Response accuracy is modeled by the suggested framework as follows: $A = f(Q, K, V)$

where V represents verification and grounding mechanisms. Reducing hallucinations requires minimizing H through controlled knowledge access and validation. [3][9][10]

6. Objectives

- **To analyze challenges in enterprise conversational AI systems** Existing AI-based support systems lack sufficient controls to prevent hallucinated responses.
- To study existing hallucination mitigation approaches Techniques such as retrieval augmentation and fine-tuning were reviewed to understand their limitations.
- **To identify gaps in domain-specific conversational frameworks**
Most systems do not enforce strict enterprise knowledge boundaries.
- **To propose a hallucination-free conversational AI framework**
SupportGuardAI integrates knowledge grounding, verification, and response constraints.
- **To improve reliability and trust in enterprise customer support**
The framework aims to deliver consistent, accurate, and compliant responses [3][4][1]

7. Methodology

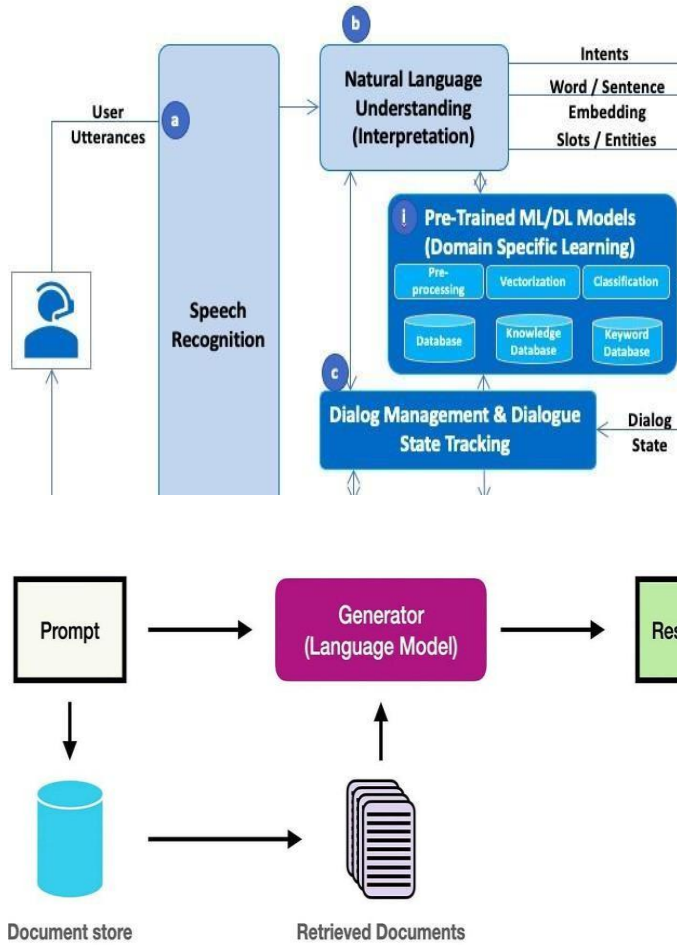
This research adopts a qualitative and conceptual methodology. A comprehensive review of literature related to conversational AI, hallucination mitigation, and enterprise knowledge systems was conducted.

The study analyzes existing conversational frameworks, identifies limitations, and proposes an integrated system architecture. No real customer data is used, ensuring ethical compliance. The proposed framework is evaluated conceptually through comparative analysis with existing generative AI-based customer support systems.[5]

8. Proposed System: SupportGuard

SupportGuardAI is a domain-specific conversational AI framework designed for enterprise customer support with minimal hallucination risk.

The proposed framework has four principal layers:



- 1. Input & Understanding Layer**
 - a. Intent detection (multi-intent)
 - b. NER + entity linking
 - c. User context, role & session tracking
- 2. Reasoning & Orchestration Layer**
 - a. Task planning & decomposition
 - b. Tool/data source selection
 - c. Fallback & confidence thresholds
- 3. Knowledge Retrieval Layer**
 - a. Secure enterprise data access
 - b. Hybrid RAG (vector + keyword)
 - c. Freshness & permission checks
- 4. Response Generation & Validation Layer**
 - a. Grounded generation with citations
 - b. Hallucination & policy checks
 - c. Confidence scoring

5. Feedback, Learning & Governance Layer

- a. Human-in-the-loop feedback
- b. Continuous optimization
- c. Monitoring, audit & security

8.1 Query Understanding Module

Processes user input through intent classification and entity recognition to accurately identify the customer's request. [5]

8.2 Knowledge Retrieval Engine

Retrieves relevant information from enterprise-approved knowledge bases such as FAQs, policy documents, and support manuals using retrieval-augmented techniques. [1]

8.3 Response Generation and Verification Layer

Generates responses strictly grounded in retrieved knowledge. Verification rules ensure factual consistency, policy compliance, and domain relevance. [3][4][9]

8.4 Consistency and Constraint Engine

Prevents responses outside enterprise scope by enforcing predefined constraints and fallback mechanisms when information is unavailable.[8][10]

8.5 Analytics and Feedback Module

Tracks query patterns, response accuracy, and escalation rates to support continuous system improvement.[6][7]

9. Results and Discussion

Conceptual analysis shows that by basing output on validated enterprise knowledge, SupportGuardAI dramatically lowers hallucinated responses. The suggested approach offers better accuracy, consistency, and customer trust when compared to generic conversational AI systems.

The system is protected from speculative or unsupported replies by structured verification and constraint enforcement. The framework is appropriate for enterprise deployment because it strikes a compromise between automation and dependability.[1][3][6][7]

10. Conclusion and Future Work

In order to eradicate hallucinations in business customer service, this article introduced SupportGuardAI, a domain-specific conversational AI platform. The framework overcomes the main drawbacks of current conversational systems by combining knowledge grounding, answer verification, and constraint mechanisms.

Future work might involve creating a working prototype, assessing performance using actual enterprise datasets, and integrating cutting-edge machine learning methods for uncertainty assessment and adaptive knowledge retrieval.

References

- [1] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge- Intensive NLP

Tasks. *NeurIPS*.

- [2] Roller, S., et al. (2021). Recipes for Building an Open-Domain Chatbot. *ACL*.
- [3] Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*.
- [4] Zhang, Y., & Chen, Q. (2022). Hallucination Mitigation in Neural Text Generation. *IEEE Access*.
- [5] Jurafsky, D., & Martin, J. (2023). *Speech and Language Processing*. Pearson.
- [6] Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics*. Harvard Business School Press.
- [7] Sharda, R., Delen, D., & Turban, E. (2018). *Business Intelligence, Analytics, and Data Science*. Pearson.
- [8] ISO/IEC 27001 (2022). Information Security Management Systems.
- [9] OpenAI. (2024). Reducing Hallucinations in Language Models. Technical Report.
- [10] Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots. *FAccT Conference*.
- [11] Open AI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.

NIST, "Artificial Intelligence

- [12] e Risk Management Framework (AI RMF 1.0)," National Institute of standards and Technology, 2023.