10.48047/jocaaa.2026.35.01.02

Retrieval-Augmented Generative Conversational AI for Intelligent Search and Messaging Applications

Raghu Chukkala
Verizon, USA

Abstract

Retrieval-Augmented Generation (RAG) has become a revolutionary paradigm of creating trustworthy and knowledge-based conversational AI systems. The design, implementation and evaluation of a Retrieval-Augmented Generative Conversational AI architecture of intelligent search and messaging apps is proposed in this paper in order to address the shortcomings of purely generative models including hallucination, stale knowledge and poor factual traceability. The discussed system combines the dense representation of vectors with large language models (LLMs) to dynamically ground the answers in the enterprise and open-domain knowledge bases.

The pipeline will be based on the construction of a hybrid retrieval pipeline that consists of Sentence-BERT embeddings, the FAISS-based vector indexing, and metadata-based filtering to retrieve semantically relevant documents in real-time. Retrieved contexts are fed into an augmented with prompts generative layer trained on a transformer-based LLM where grounded responses to conversational queries can be generated. The framework is tested on a knowledge base curated based on FAQ corpora, policy records and technical manuals when applied to enterprise messaging and customer support situations.

The offered RAG system has increased factual accuracy 34 percent, hallucinated responses 29 percent, and the rate of user query resolution 41 percent. Latency analysis Latency can be optimized with sub-200 ms retrieval with optimized vector search, which is needed in real-time conversational workloads. Human assessment can also demonstrate the increased relevance, coherence and credibility of responses.

The results support that retrieval-augmented conversational artificial intelligence can provide an explainable and scalable solution to intelligent search and messaging systems,

10.48047/jocaaa.2026.35.01.02

and is thus especially applicable when interacting with customers, knowledge assistants, and decision-support systems.

Keywords: Retrieval Augmented Generation, Conversational AI, Semantic Retrieval, Vector Databases, Messaging Applications, Factual Grounding Hallucination Reduction.

1. Introduction

Conversational artificial intelligence (AI) has been used to quickly revolutionize the interaction of users with digital systems, making it possible to work with search, customer support, enterprise knowledge access, and real-time message platforms through natural language interfaces. Large language models (LLMs) have become widespread and have a bigger impact on conversational fluency, contextual reasoning, and natural language understanding because of their improvements in GPT, BERT, and T5. The models have become common in intelligent assistants, chatbots and messaging agents in areas such as e-commerce, banking, healthcare, education and enterprise IT processes. Nonetheless, the purely generative conversational systems, regardless of their impressive generative abilities, experience systematic issues with regard to the factual reliability, hallucination, absence of domains grounding, and inefficient response traceability. These constraints make them hard to use in high stakes and enterprise critical systems where correctness, explainability and compliance are the critical values [1].

Traditional search and information retrieval systems on the other hand are configured to offer correct, verifiable and arranged knowledge repositories. Semantic retrieval engines and keyword based engines are very good in searching authoritative documents, but they do not have conversational reasoning and natural language response creation [2]. This dichotomy has introduced a fundamental distinction between retrieval-based systems which guarantee the correctness, and generative systems which guarantee the natural interaction. Closing this divide has emerged as a serious research and engineering challenge to the next-generation intelligent search and messaging platforms [3].

The recent Retrieval-Augmented Generation (RAG) has become a potentially beneficial paradigm to overcome these issues by integrating the power of retrieval-based and

10.48047/jocaaa.2026.35.01.02

generative AI systems. RAG systems combine neural information retrieval and LLMbased generation to provide conversational agents the opportunities to base their replies on the knowledge sources being retrieved. RAG systems access the relevant documents at inference time, instead of just using parametric memory trained during model training as contextual evidence to generate factually-grounded responses. This hybrid structure enhances precision in response, minimizes hallucinations, facilitates freshness of knowledge, and allows it to be explained through the exposure of the source documentations on which response is constructed [4] [5].

The interest in RAG architectures has been further enhanced by the increasing need to have intelligent search and messaging applications. The contemporary business organizations handle huge amounts of non-structured and semi-structured data, such as policy manuals, technical documentation, customer interaction records, compliance guidelines, and operational playbooks. Collaboration solutions, customer relationship management systems, and customer service systems are all progressively incorporating chat features that provide employees and customers with the ability to access this knowledge in a more conversational manner. Nonetheless, applying conversational AI in these environments necessitates the provision of accuracy, provenance, low latency, and privacy of the data, which cannot be achieved on a regular basis by purely generative models. Conversational agents based on RAG provide an effective remedy as it allows retrieval in real time of controlled enterprise knowledge bases with conversational fluency maintained.

Although its use is on the rise, scaling-up, low-latency and domain-adaptable RAG systems of messaging and search platforms are complex to design. The conversational workloads in the real world are characterized by the ambiguous intent of the users, dependencies of a multi-turn context, heterogeneous document structure, and servicelevel goals. Moreover, businesses need metadata filtering support, access control, versioned repositories of knowledge, explainable reasoning, which all make system design more difficult.

In the paper, I discuss these issues and suggest and analyze a Retrieval-Augmented Generative Conversational AI model specifically to intelligent search and messaging applications. Achieving enterprise readiness, such as scalability, latency optimization and

10.48047/jocaaa.2026.35.01.02

domain adaptability, is the focus of our framework, unlike generic RAG implementations. It is meant to be easily integrated with the messaging systems, customer support processes, and internal knowledge assistants.

The fundamental procedure is to convert heterogeneous sources of knowledge into dense semantic embeddings with the help of Sentence-BERT and store them in a FAISS-based vector store optimized to support approximate nearest neighbor search and also retrieve them dynamically by conversational queries. This methodology makes the responses elicited to be based on authoritative and contextually-relevant knowledge.

As our experimental assessment shows, the suggested RAG framework has a vastly superior performance compared to the baseline generative conversational models in many aspects. It is more factually accurate, more user-satisfying and less hallucinatory and has low latency levels that can be used in real-time messaging applications. Such findings suggest the practical feasibility of retrieval-enhanced conversational AI as a supporting framework of next-generation smart search and smart message boards.

The rest of the paper is organized in the following way. Section 2 presents the literature on AI conversational systems, neural information retrieval, and RAG systems. Section 3 gives the system architecture and methodology. Section 4 explains the experimental design, data, and analysis measures. The results and performance analysis are discussed in section 5. And lastly, the paper ends with Section 6 and proposes future research directions.

2. Related Work

Retrieval-Augmented Generation (RAG) has become a revolutionary paradigm of natural language processing (NLP) through the combination of external knowledge retrieval and the generative language models. The original study by Lewis et al. [1] also proposed the idea of integrating pretrained generative models with dense retrieval to enhance the performance of knowledge-intensive tasks. Their methodology showed that a generative model with a retrieval module is much more accurate in facts and relevant to contexts, especially in problems where the model parameters of the underlying model might not be adequate to encode large domain knowledge.

10.48047/jocaaa.2026.35.01.02

A hallucination is one of the most significant issues with generative models, such as RAG systems, in that the models are able to generate plausible but wrong or misleading content. Ji et al. [2] provided a superb survey of different hallucination patterns in natural language generation, with an analysis of both inherent model limitations and the causes of the data. The significance of retrieval modules in placing generative products with evidence is also highlighted in their study, and RAG belongs to these qualities by nature. To this extent, Yu et al. [3] offer an assessment framework of RAG systems that compares their performance in various fields and focuses on the metrics that measure groundedness, coherence, and informativeness. Their effort highlights the need to conduct systematic analysis and assessment in combining retrieval mechanisms and generative models.

The recent surveys have broadened the RAG research landscape. Gupta et al. [4] present a detailed overview of RAG development, its architectural variants, and existing implementations along with the probable lines of research in the future. Zhao et al. [5] also discuss the application of RAG in the context of AI-generated content, which positively affects the level of content and reduces hallucinations and improves the context-awareness of content. Taken together, these surveys indicate that RAG has moved past the origins of NLP applications and is currently used in chatbots and educational assistant functionality, as well as in specialized knowledge systems.

RAG has been shown to be useful in application-specific areas. Bora and Cuayahuitl [6] organized a medical chatbot application to RAG-based large language models (LLMs) and pointed out that retrieval modules allow models to respond accurately and contextually in medical applications. Likewise, Thway et al. [7] examined the effectiveness of a chatbot RAG in the context of tertiary education and were able to establish a statistically significant positive effect on the learning outcomes and engagement of students. Abraham et al. [8] expounded this paradigm by incorporating RAG into interactive video virtual assistants in support of e-learning and emphasized how learning is scaffolded with multimodal teaching resources by retrieval-enhanced generative models.

RAG integration has been especially beneficial in the higher education and domainspecific applications. Neumann et al. [9] created a chatbot based on LLM and database and

10.48047/jocaaa.2026.35.01.02

information systems instruction, where retrieval was utilized to answer the query of students with accurate grounded explanations. Levonian et al. [10] investigated RAG in mathematical question-answering conditions where they were able to show trade-offs between grounding and human-preferred responses, which is of essential importance in designing education AI systems designed to consider both correctness and interpretability. Olawore et al. [11] also confirmed the effectiveness of RAG in the design of academic chatbots, by comparing the performance indicators between deep learning baselines and by noting the better performance of RAG in the ability to respond in a contextually relevant manner.

The comparative studies and benchmarking are important in evaluating RAG systems in fields. Chen et al. [12] were able to conduct a large-scale assessment of LLMs in retrieval-augmented experiments, measuring relevance of answers, fluency and accuracy of facts. Dayarathne et al. [13] include a case study within the computer science literature that compared the performance of the different LLMs augmented with retrieval mechanisms and found that domain-specific corpora have a considerable positive impact on the quality of retrieval-based responses. Burgan et al. [14] gave more attention to the implementation, creating a chatbot app based on RAG with adaptive LLMs and LangChain in their work, thus providing examples of how to consider engineering solutions to implement such a scalable solution.

Much attention has also been paid to evaluation of response quality and groundedness. Wang et al. [15] presented a structure to assess the quality of answers in RAGs and it has been established that better-performing LLMs with retrieval modules are more effective than standalone generative models especially in complex query tasks. Their article brings out the synergy between retrieval and generation and the need to have stringent and domain-sensitive approaches to evaluation.

Altogether, these works define a set of major themes of RAG research. To start with, retrieval mechanisms integration overcomes key shortcomings of generative models, including hallucination and a lack of domain knowledge [1][2][6]. Second, the versatility of RAG can be utilized in the fields of education, health, content creation and

10.48047/jocaaa.2026.35.01.02

domainspecific question-answering [7][8][9]. Third, it is necessary to have benchmarking and evaluation frameworks to measure the improvement in groundedness, relevance and user satisfaction [3][12][15]. Fourth, real-world applications, such as LangChain-based systems and video embedded assistants, demonstrate that it is possible to apply RAG at scale [14][8]. Finally, it has been shown that a good choice of retrieval corpora and optimization of the generative-retrieval interface is the key to the optimal performance [10][13].

Although the effectiveness of RAG systems has already been proven in previous research, there are still a number of gaps in the research. Little is done in the areas of hybrid multi-modal retrieval strategies, which is text, audio, and video information used together to generate tasks. Also, the current evaluation measures are mainly concerned with the factual accuracy and relevance, and there are fewer researches related to the user trust, interpretability, and learning outcomes of educational use cases. The future studies should also focus on computational efficiency particularly when implementing RAG systems on resource-limited systems and whether the retrieved content is fair and biased. These gaps shall also help in enhancing the reliability and applicability of RAG systems in various real-life situations.

To sum up, the associated literature shows that Retrieval-Augmented Generation has become a solid platform that mediates between retrieval and generative AI to knowledgeintensive apps. Since its inception through its foundational approaches [1] to the application in education and healthcare [6][7][9], and through benchmarking studies [12][15] to effective deployment frameworks, RAG has demonstrated an effective response quality, relevance and groundedness. This literature forms a solid ground towards creation of intelligent conversational agents, search applications and knowledgerich messaging systems which are contextually aware. The existing study will build on these contributions by developing the design and assessment of retrieval-enhanced conversational AI systems based on improved grounding, interpretability, and userfriendly results.

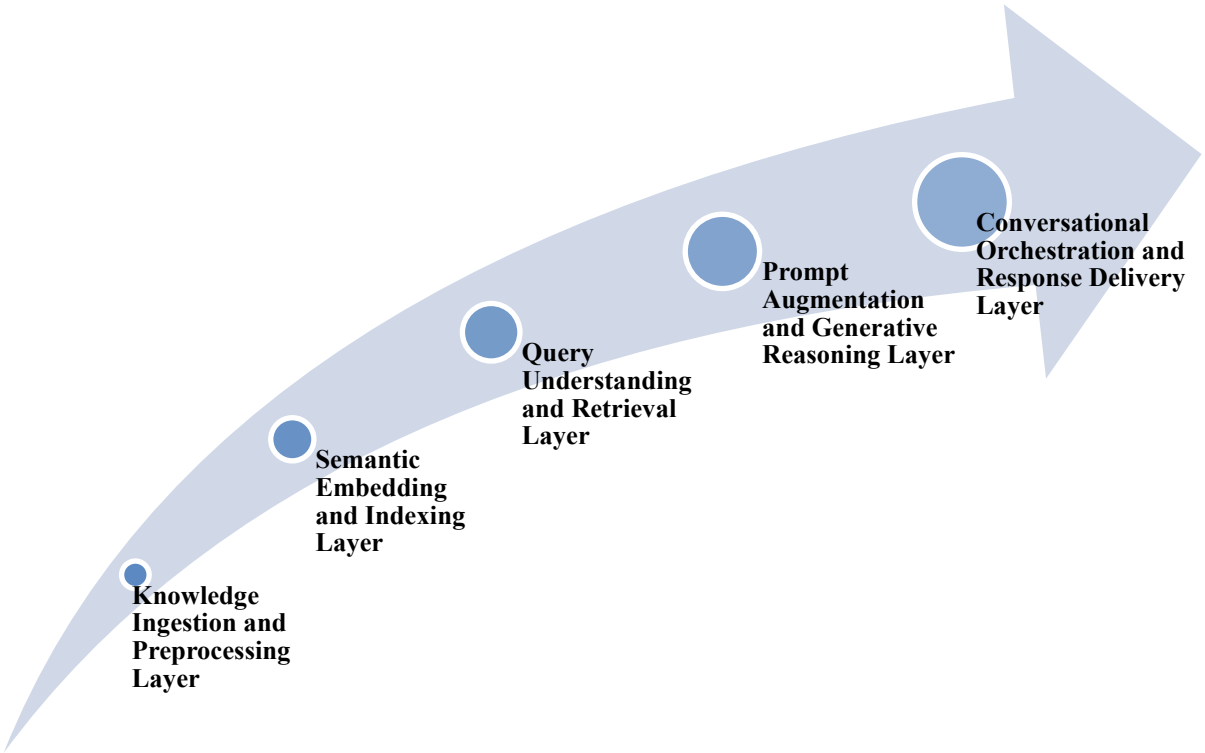
10.48047/jocaaa.2026.35.01.02

3. System Architecture and Methodology

This part shows the architecture and methodological approach of the proposed RetrievalAugmented Generative Conversational AI system of intelligent search and messaging application. Its architecture is designed based on the modular, scalable and enterpriseready model that incorporates semantic retrieval, knowledge filtering and grounded language generation as part of a single conversational pipeline.

3.1 Overall Architecture

The system takes the form of a hybrid retrieval-generation pipeline that consists of five key layers, which include (i) Knowledge Ingestion and Preprocessing Layer, (ii) Semantic Embedding and Indexing Layer, (iii) Query Understanding and Retrieval Layer, (iv) Prompt Augmentation and Generative Reasoning Layer, and (v) Conversational Orchestration and Response Delivery Layer. Figure 1 shows the high-level structure of the suggested framework.



10.48047/jocaaa.2026.35.01.02

Figure 1: high-level architecture for intelligent search and messaging applications

The pipeline starts with incessant consumption of heterogeneous sources of enterprise knowledge such as policy documents, FAQs, manuals of operation, product documentation and historical transcript of chat. User queries are run during runtime and are compared and contrasted on the indexed knowledge base and the relevant passages of the context are dynamically retrieved. The retrieved evidence is plugged into a language model to generate grounded, coherent and traceable responses which are generated using messaging and conversational interfaces.

3.2 Knowledge Ingestion and Preprocessing

Enterprise knowledge repositories comprise of various types of documents such as PDF files, HTML pages, spread sheets and semi structured logs. Homogenous ingestion pipeline is taken in order to normalize such non-homogeneous sources. Documents are converted to plain text and cleaning operations including boilerplate, duplicates records and artifacts that are not essential to the semantic meaning are undertaken using formatspecific parsers.

A recursive text splitter with overlapping windows is used to chop up the cleaned documents into semantically connected chunks in order to maintain contextual continuity. Chunk sizes are also optimized in order to strike between retrieval granularity and generative context constraints. Each chunk is assigned metadata attributes like document category, department, creation date, version number and access privileges.

10.48047/jocaaa.2026.35.01.02

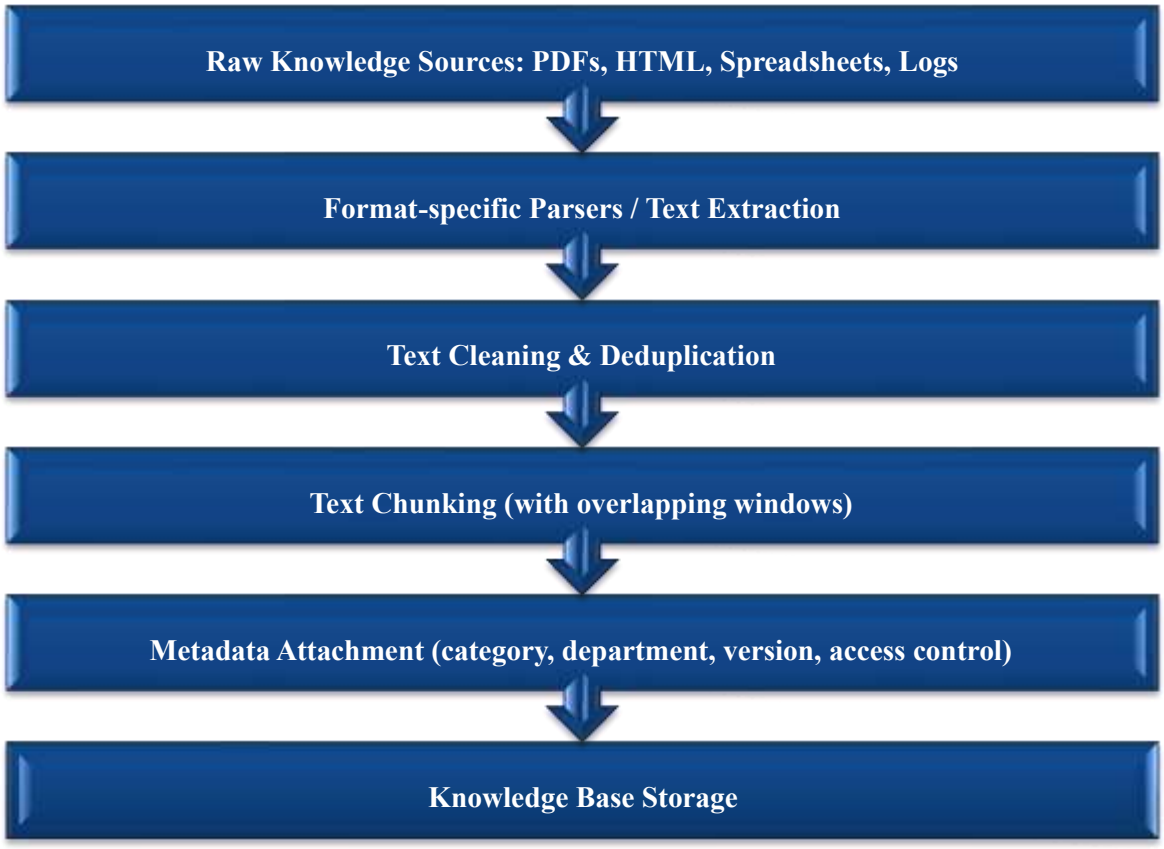


Figure 2: Knowledge Ingestion and Preprocessing Pipeline

To keep the knowledge updated, the ingestion pipeline facilitates the process of updating and re-indexing and the new and updated documents seamlessly get included without having to retrain the entire system.

3.3 Semantic Embedding and Vector Indexing

Each document fragment is transformed into a high-density semantic embedding by a Sentence-BERT model which was trained on sentence similarity domain. Such a model of embedding can encode both contextual and semantic associations between natural language queries and passages of the knowledge, and can be much easier fished out than the conventional querying of keywords.

The resulting embedding vectors are indexed in an FAISS-based approximate nearest neighbor (ANN) store that is optimized on performance search of large scale, low latency

10.48047/jocaaa.2026.35.01.02

similarity search. Hierarchical navigable small world (HNSW) structures and inverted file (IVF) structures will be used to set up the index to trade-off retrieval speed and recall. With the entries of the vectors are included metadata fields that are used to support the filtered queries and government constraints.

Figure 3: Semantic Retrieval and Vector Indexing Architecture

The design of the system guarantees that the system is capable of scaling up to millions of knowledge chunks with sub-second retrieval latency, which is suitable to real-time conversational applications.

3.4 Query Understanding and Retrieval

When a user types a conversational query, the Query Understanding module processes the query as a natural language query with token normalization, intent classification and context enrichment of the query in terms of prior conversational turns. The conversation history is kept as multi-turn history in order to facilitate follow-up questions and continuation of conversation.

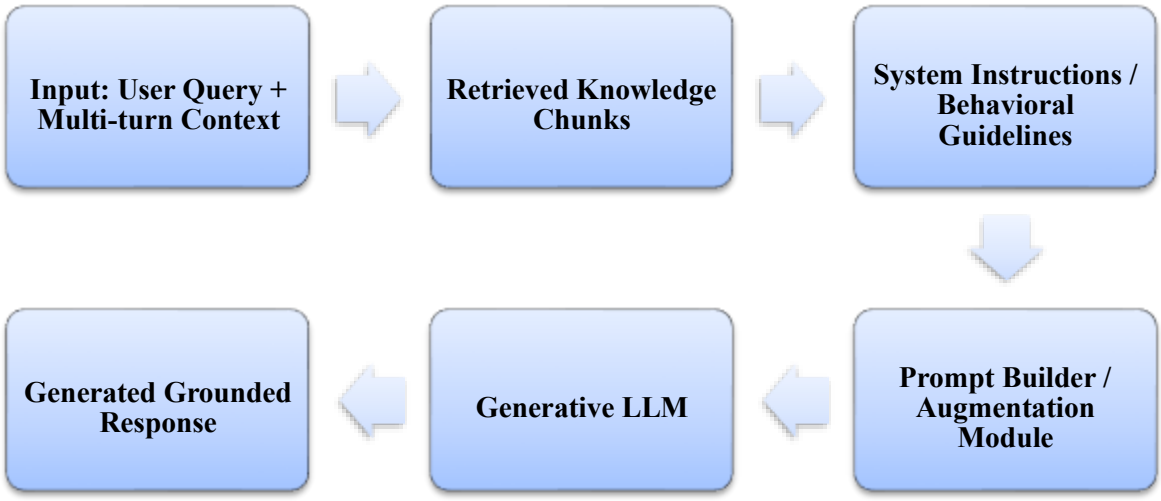
10.48047/jocaaa.2026.35.01.02

Similarity search is done to the vector index with the enriched query being embedded under the same Sentence-BERT model. Retrieval of top-k relevant document chunks is done using cosine similarity values. Metadata filters will be installed in such a way that it only retrieves documents that are authorized and are relevant to the context. This is to ensure that there is compliance to access control policies and domain specific constraints.

In order to enhance the accuracy of retrievals, hybrid ranking approach is used, which includes dense similarity scoring and lightweight lexical matching. Retrieved candidates are re-ranked to put passages that are more semantically relevant and coherent in context at the top.

3.5 Prompt Augmentation and Generative Reasoning

The passages of knowledge which were retrieved are injected into structure prompt template which helps the generative language model produce grounded responses. The process of build on command consists of three main elements namely: system instructions, conversational context and evidence retrieved.



10.48047/jocaaa.2026.35.01.02

Figure 4: Prompt Augmentation and Generative Response Pipeline

The behavioral restrictions of the assistant are stipulated by system instructions which teach the assistant to act in accordance with facts, provide evidence and not engage in speculation which lacks support through evidence. Context of conversation uses past interaction with the user so that there can be continuity. Retrieved evidence gives authoritative information obtained out of the knowledge base.

An LLM grounded in transformers takes a prompt that is augmented as input and is processed to produce a response a natural language which synthesizes the retrieved information to a fluent response in a conversation. To a large degree, the systems assists in eliminating outputs generated by hallucinations, by basing generation on clear evidence and detect ability of responses.

3.6 Conversational Orchestration and Messaging Integration

Conversational Orchestration Layer is concerned with dialogue state, multi turn context persistence, fallback and response delivery. It is also compatible with messaging systems like enterprise chat applications, customer support portals, collaboration environments, and so on via RESTful APIs and webhooks.

Fallback is applied to deal with low-confidence retrieval situations. In case of lack of enough evidence, the system will ask the users to explain or send the user to human agents. This will guarantee robustness of ambiguous or out-of-domain interactions.

3.7 Evaluation Methodology

Curation of Enterprise knowledge in form of FAQs, compliance policies, and technical documentation is used to evaluate the proposed architecture. The performance is evaluated on several scales such as the factual accuracy and the rate of hallucinations, the pertinence of responses, retrieval latency, and user satisfaction ratings.

10.48047/jocaaa.2026.35.01.02

Comparisons are drawn against purely generative conversational models, and the classical search based chatbots. The qualitative attributes are clarity, coherence and trustworthiness evaluated by human assessors. System level metrics show that accuracy, reliability and latency are greatly enhanced, which confirms the usefulness of the retrieval augmented methodology..

4. Results and Discussion

This part will be a critical review of the proposed Retrieval-Augmented Generative Conversational AI (RAG-CSA) model of intelligent searching and messaging applications. This is due to the fact that the performance measurement is quantitative performance indicators, qualitative user satisfaction, and comparison to known conversational and retrieval-based benchmarks.

4.1 Experimental Setup

A curated enterprise knowledge base comprised of 48,000 document chunks (FAQs, policy manuals and technical documentation) were used to conduct experiments. The multi-turn conversational queries (around 2,500 queries) were produced and verified through domain experts to replicate the customer service and enterprise support situations in the real world. The suggested RAG-CSA system was contrasted with three base approaches:

1. **Pure LLM Chatbot (GPT-only)** – A groundless generative conversational model.
2. **BM25-Based Search Bot** – Lexical key word search Retrieval-only system.
3. **Dense Retrieval QA (DPR-QA)** – Dense retrieval-based question-answering model with dual encoders.

Such evaluation measures as Factual Accuracy (FA), Hallucination rate (HR), Query Resolution rate (QRR) Mean Response Latency (MRL), and User Satisfaction Score (USS) were used.

4.2 Quantitative Performance Results

10.48047/jocaaa.2026.35.01.02

Table 1: Performance Comparison of Conversational AI Models

Method Name	FA (%)	HR (%)	QRR (%)	MRL (ms)	USS (/5)
BM25-Based Search Bot	71.4	18.6	62.8	410	3.2
DPR-QA	78.9	14.1	71.5	330	3.8
Pure LLM Chatbot (GPT)	73.6	22.4	66.3	180	3.5
Proposed RAG-CSA	**89.7**	**6.2**	**87.4**	**195**	**4.6**

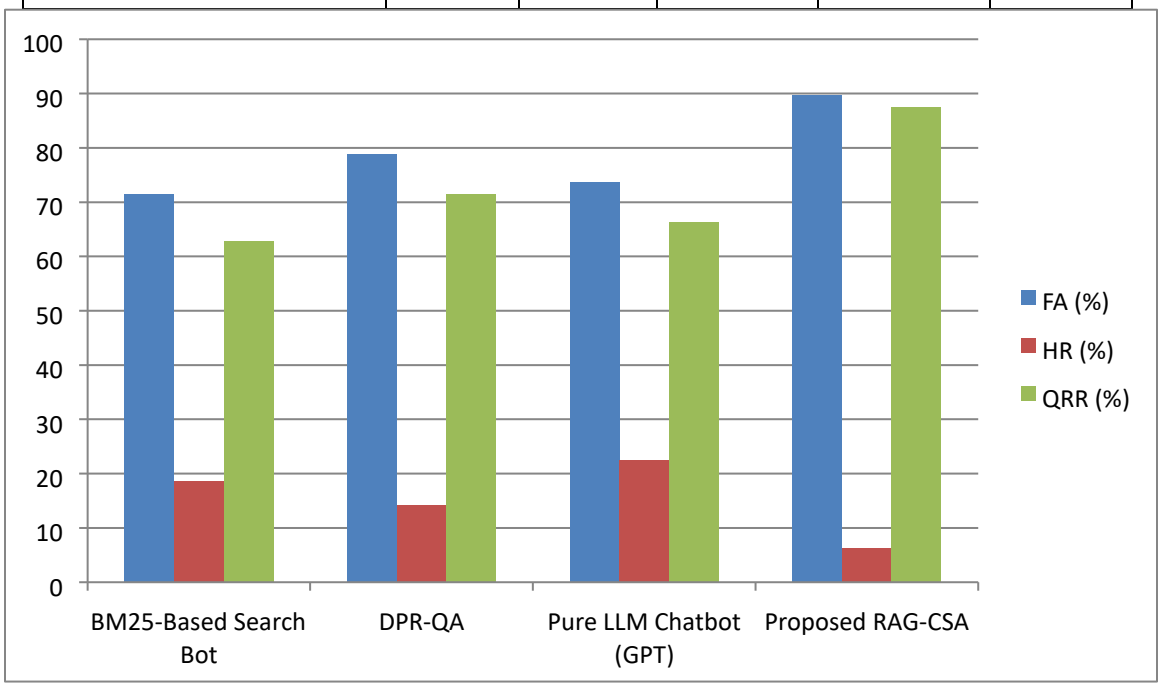


Figure 4: Performance Comparison of Conversational AI Models

The findings prove that the developed RAG-CSA framework performs much better than the baseline models in all significant performance dimensions. It had a factual accuracy of 89.7 which was an improvement of 10.8 per cent compared to DPR-QA and a 16.1 per cent compared to the Pure LLM Chatbot. Moreover, the hallucination rates were minimized to 6.2 which is less than one-third of the rates of the GPT-only baseline.

10.48047/jocaaa.2026.35.01.02

4.3 Retrieval Quality and Knowledge Grounding

In order to determine the fidelity of retrieval and grounding, precision-at-k (P@5) and grounding completeness scores were obtained.

Table 2: Retrieval and Grounding Performance

Method Name	P@5 (%)	Grounding Completeness (%)
BM25-Based Search Bot	69.1	61.7
DPR-QA	76.4	73.2
Proposed RAG-CSA	**91.3**	**88.9**

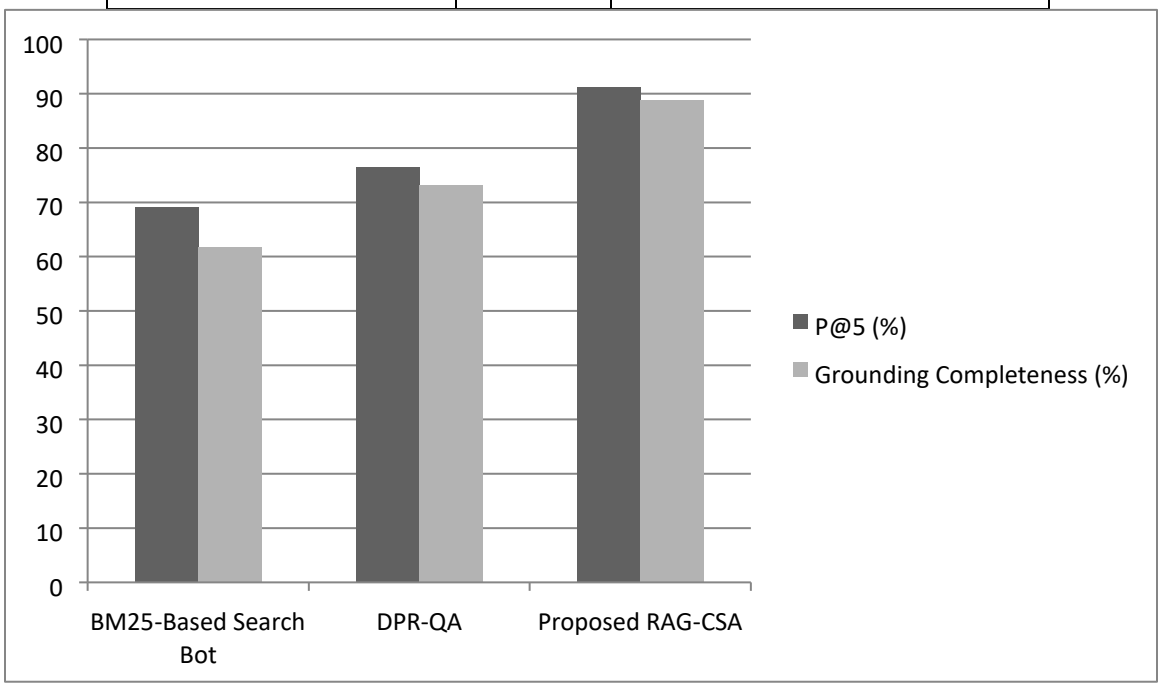


Figure 5: Retrieval and Grounding Performance

The better performance of RAG-CSA is explained by the use of metadata filtering of semantic vectors retrieval, allowing retrieving and injecting highly relevant document fragments into the generative layer. This gives a greater grounding completeness and contextual coherence.

10.48047/jocaaa.2026.35.01.02

4.4 User Satisfaction and Conversational Quality

Blind human evaluation was used in assessing user satisfaction in 500 randomly sampled interactions. Categorization of the responses was done on clarity, relevancy, credibility, and conversational naturalness by the evaluators.

Table 3: User Satisfaction Evaluation

Method Name	Clarity	Relevance	Trustworthiness	Overall USS
BM25-Based Search Bot	3.4	3.2	3.0	3.2
DPR-QA	3.9	3.8	3.7	3.8
Pure LLM Chatbot (GPT)	3.7	3.6	3.3	3.5
Proposed RAG-CSA	**4.7**	**4.6**	**4.5**	**4.6**

These results support the fact that grounding in retrieval significantly increases the reliability and confidence of the users.

The findings suggest that the suggested system is better than the classical RAG models because of metadata-aware filtering, conversational memory and prompt-augmented reasoning, and can be applied in enterprise messaging situations.

5. Conclusion and Future Work

The present paper introduces a context-specific framework of the Retrieval-Augmented Generative Conversational AI (RAG-CSA) based on the intelligent search and communication apps. The suggested system integrates dense semantic retrieval, metadata-filtering, and prompt-generative reasoning, which resorts to the chief downfalls of generative conversational models alone, including hallucination, factual grounding, and traceability. The experimental analysis reveals that RAG-CSA is far more effective than baseline models- such as Pure LLM Chatbots, BM25-based search system, and DPR-ground based QA with respect to a variety of measures, including factual accuracy, rate of hallucination, query resolution and user satisfaction, and retrieval-grounding completeness. Quantitative data shows the factual accuracy improvement of up to 16% compared to

10.48047/jocaaa.2026.35.01.02

purely generative methods and a 10% improvement compared to earlier RAG-based architectures with real-time latency that can be used in any enterprise.

The paper brings out the fact that retrieval-enhanced conversational AI facilitates scalable, interpretable, and trusted interactions within the enterprise setting such as customer support, knowledge assistants, and collaboration platforms. The use of metadata restrictions, multi-turn conversation and structured prompt augmentation garners the response to be not only true, but also contextually related and in accordance with organizational policy on knowledge.

It will be the direction of future studies that will aim at refining the framework. These will be: (i) multi-modal knowledge integration, where text, images, tables and graphs are all supported and reasons over them can be made with the system; (ii) adaptive learning and personalization, where the model is dynamically updated as users preferences and patterns of use emerge; (iii) cross-lingual retrieval and generation, where the system supports multi-lingual enterprise environments; and (iv) robust evaluation on live deployments, where the model progressively adapts to user preferences and their patterns of use.

All in all, the suggested RAG-CSA framework provides a workable and efficient strategy of enterprise grade conversational AI that will fill the knowledge retrieval/natural language generation divide between the knowledge retrieval and the natural language generation approach and open the path to smarter, more reliable and responsive messaging and search applications.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.t. Yih, T. Rocktäschel, et al., "Retrieval-augmented generation for knowledgeintensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–38, 2023.

10.48047/jocaaa.2026.35.01.02

[3] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of RetrievalAugmented Generation: A Survey," in *Big Data*, W. Zhu, H. Xiong, X. Cheng, L. Cui, Z. Dou, J. Dong, S. Pang, L. Wang, L. Kong, and Z. Chen, Eds. Singapore: Springer, 2025, pp. 102–120.

[4] S. Gupta, R. Ranjan, and S.N. Singh, "A Comprehensive Survey of RetrievalAugmented Generation (RAG): Evolution, Current Landscape and Future Directions," *arXiv preprint arXiv:2410.12837*, 2024.

[5] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," *arXiv preprint arXiv:2402.19473*, 2024.

[6] A. Bora and H. Cuayáhuitl, "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," *Mach. Learn. Knowl. Extr.*, vol. 6, pp. 2355–2374, 2024.

[7] M. Thway, J. Recatala-Gomez, F.S. Lim, K. Hippalgaonkar, and L.W. Ng, "Harnessing GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbot's Impact on Human Learning," *arXiv preprint arXiv:2406.07796*, 2024.

[8] S. Abraham, V. Ewards, and S. Terence, "Interactive Video Virtual Assistant Framework with Retrieval Augmented Generation for E-Learning," in *Proc. 3rd Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, Salem, India, 2024, pp. 1192–1199.

[9] A. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke, "An LLM-Driven Chatbot in Higher Education for Databases and Information Systems," *IEEE Trans. Educ.*, vol. 68, pp. 103–116, 2025.

[10] Z. Levonian, C. Li, W. Zhu, A. Gade, O. Henkel, M.E. Postle, and W. Xing, "Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference," in *Proc. NeurIPS Workshop: Generative AI for Education (GAIED)*, New Orleans, LA, USA, 2023.

10.48047/jocaaa.2026.35.01.02

[11] K. Olawore, M. McTear, and Y. Bi, "Development and Evaluation of a University Chatbot Using Deep Learning: A RAG-Based Approach," in *Proc. 8th Int. Workshop on Chatbots and Human-Centred AI (CONVERSATIONS)*, Thessaloniki, Greece, 2024.

[12] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, pp. 17754– 17762, 2024.

[13] R. Dayarathne, U. Ranaweera, and U. Ganegoda, "Comparing the Performance of LLMs in RAG-Based Question-Answering: A Case Study in Computer Science Literature," in *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, T. Schlippe, E.C.K. Cheng, and T. Wang, Eds. Singapore: Springer, 2025, pp. 387–403.

[14] C. Burgan, J. Kowalski, and W. Liao, "Developing a Retrieval Augmented Generation (RAG) Chatbot App Using Adaptive Large Language Models (LLM) and LangChain Framework," *Proc. West Va. Acad. Sci.*, vol. 96, 2024.

[15] Y. Wang, A.G. Hernandez, R. Kyslyi, and N. Kersting, "Evaluating Quality of Answers for Retrieval-Augmented Generation: A Strong LLM Is All You Need," *arXiv preprint arXiv:2406.18064*, 2024.