The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2025

Volume 33, Pages 20-27

IConTech 2025: International Conference on Technology

A RAG-Based Automotive Sector AI Assistant for Enhanced Information Retrieval

Senda Yildirim
Dogus Technology
Kadir Has University

Ruya Samli
Istanbul University

Abstract: Advanced Artificial Intelligence (AI) technologies are increasingly used to ensure fast and accurate information access across industries. In the automotive sector, efficient querying of technical data by customers and service staff is essential. Due to the volume of manuals, maintenance logs, and troubleshooting guides, manual search is often impractical. This study introduces an AI-based assistant for the automotive domain, built on a pre-trained language model using the Retrieval-Augmented Generation (RAG) framework. RAG improves text generation by retrieving relevant data from external sources—such as document repositories and databases—rather than relying solely on a generative model. This hybrid approach reduces hallucinations and increases response accuracy. Unlike traditional chatbots, our system draws domain-specific content from curated technical documents, ensuring higher relevance and reliability. The assistant is not a new model but a domain-specific application that integrates an existing LLM with the RAG framework for an industrial use case. The automotive assistant is designed to extract information from technical documents to deliver accurate answers to common user problems. It supports both vehicle owners and service professionals by providing real-time, context-aware information for troubleshooting and maintenance. To evaluate its performance, a validation dataset comprising 487 real customer service call transcripts (2,578 sentences, 6,445 seconds) was used. These transcripts served solely for evaluation purposes, testing the assistant's ability to generate contextually appropriate responses to real-world queries. This study demonstrates how a RAG-based model can be optimized for domain-specific use, improving information retrieval in the automotive sector. By combining retrieval and generation, the assistant enhances the accuracy and efficiency of technical support. The system was first piloted internally by call center staff, allowing for a thorough evaluation of its accuracy, safety, and compliance with responsible AI principles. Pilot results showed that the assistant significantly enhanced the efficiency and accuracy of information retrieval in technical support, improving operational performance and user satisfaction. Evaluations confirmed that it provides more precise and context-aware responses than traditional generative models, leading to a better user experience. As a result, the assistant serves as a valuable tool for both end-users and service teams, reducing time spent on searching critical maintenance information and boosting customer satisfaction.

Keywords: Retrieval-augmented generation, Automotive sector, Artificial intelligence, Information retrieval

Introduction

In the era of digital transformation, rapid and accurate access to contextually relevant information is essential across many industries. Evolving AI systems are increasingly used to tackle information overload by converting unstructured data into actionable knowledge. In knowledge-intensive fields such as healthcare, finance, and engineering, timely and reliable information directly impacts decision-making and operational efficiency. The automotive sector similarly benefits from these advancements.Modern vehicles feature complex subsystems and software-driven components that require extensive documentation for troubleshooting, maintenance, and

compliance. These resources—ranging from service manuals to diagnostic guides—are vast and often fragmented. Rapid access to accurate technical instructions is crucial for service representatives, impacting customer satisfaction, first-time resolution rates, and safety. However, traditional search engines and rule-based chatbots often fail to meet these needs due to limited contextual understanding.

To address these challenges, we propose a task-specific AI assistant for technical support in the automotive industry. The system is built on a pre-trained large language model (LLM) and utilizes the Retrieval-Augmented Generation (RAG) framework. Unlike standard generative models, RAG employs a two-step process: it retrieves relevant documents from an external corpus and generates responses based on that content. This approach enhances factual accuracy and allows for dynamic knowledge updates without retraining the base model.

Mathematically, the system can be represented as maximizing the joint probability:

$$y = arg\max_{y} \sum_{d \in \mathrm{D}} \mathrm{P}(y \mid x, d) \cdot \mathrm{P}(d \mid x)$$

where x is the user query, $d \in$ D are retrieved documents, and y is the generated response. This formulation highlights the model's dependence on both generative reasoning and retrieval relevance.

This study does not involve developing a new language model, but rather focuses on integrating RAG with domain-specific document repositories to create a reliable assistant for service personnel and, eventually, end customers. The prototype was evaluated using a validation dataset of 487 anonymized call center transcripts, totaling 6,445 seconds and 2,578 unique sentences. This dataset was not used for model fine-tuning but to assess the assistant's ability to interpret and respond to real-world automotive service queries.

The assistant was initially piloted internally with professional call center agents before being exposed to customers. This phased approach allowed for evaluating not only the model's output accuracy but also its adherence to responsible AI principles, such as data privacy, guardrail effectiveness, and mitigation of unsafe or misleading responses. A safeguard-rich environment was created to assess the assistant's readiness for broader deployment, incorporating prompt filters, retrieval constraints, and response validation mechanisms.

In summary, this study contributes to the field by:

- Demonstrating the effective application of a RAG-based architecture for domain-specific information retrieval by integrating it with a pre-trained LLM in the technical automotive service context.
- Developing a task-oriented conversational assistant that improves support efficiency by reducing manual document search time and enhancing response accuracy.
- Validating the assistant with real-world call center transcripts to ensure its outputs are contextually accurate and aligned with actual customer queries, rather than relying on synthetic or pre-annotated data.
- Establishing a responsible deployment pathway, starting with an internal pilot for service personnel to evaluate system guardrails, ethical constraints, and trustworthiness before external release.

This paper presents our approach to developing and validating a RAG-based conversational AI assistant for the automotive sector. Section 2 reviews related work and outlines the system architecture and methodology. Section 3 details the implementation and data processing pipeline. Section 4 presents evaluation results and comparisons with baseline models. Finally, Section 5 concludes with insights and future directions for broader deployment and multi-domain applications.

Related Work and Methodology

Retrieval-Augmented Generation (RAG) in Domain-Specific Applications

Retrieval-Augmented Generation (RAG) is a hybrid approach that combines the strengths of large language models (LLMs) with external information retrieval to generate accurate, context-aware responses. Unlike traditional generative models, which rely solely on internal parameters, RAG retrieves relevant documents during inference and grounds the generation on this external knowledge. This approach reduces hallucinations, a common issue in standard LLMs, and improves the factual consistency of responses.

Barron et al. (2024) introduced a RAG system that integrates vector stores with knowledge graphs and tensor factorization techniques, demonstrating superior performance in domain-specific tasks such as malware detection and anomaly identification. This highlights the potential of augmented generation in sensitive application domains. Similarly, Jadon et al. (2025) proposed synthetic data generation strategies and reasoning-based evaluation techniques to enhance domain specificity in RAG applications, particularly in finance and cybersecurity. These innovations emphasize the need for targeted improvements to increase the relevance and interpretability of RAG-based systems in complex textual domains.

RAG-Based AI Assistants in the Automotive Industry

In the automotive industry, the complexity and volume of technical documentation require intelligent systems for real-time, accurate support. RAG-based assistants are increasingly addressing this need. Gearguide, developed by Tezeract (2024), is a prominent example of a RAG-powered assistant designed for motorcycle maintenance. It retrieves information from technical manuals to provide highly relevant responses to user queries. Audi has implemented an internal RAG-powered chatbot to optimize its documentation workflows and improve information access across the organization (Reply, 2024). This system reduces cognitive load and enhances operational efficiency by providing concise and accurate documentation excerpts on demand.

Elastic's production-level assistant uses RAG and Azure Blob Storage to optimize customer support operations, addressing challenges such as latency, index management, and real-time response verification (Elastic, 2024). Additionally, LlamaIndex's OilyRAGs assistant highlights the role of AI in improving mechanical diagnostics and automating repetitive workshop tasks, reducing technician workload and turnaround times.

Evaluation and Benchmarking of RAG Systems

Effective evaluation of RAG systems requires comprehensive frameworks that assess both retrieval precision and generation quality. Friel et al. (2024) introduced RAGBench, an explainable benchmark that evaluates systems using metrics such as Relevance, Completeness, Adherence, and Utilization. Their TRACe framework allows developers to diagnose and improve model performance with interpretable outputs.

DIRAS, proposed by Ni et al. (2024), provides a scalable approach to annotating document relevance using LLMs without manual labeling. This method ensures domain-specific adaptation and enables nuanced judgment in evaluating document-query pairs.

BERGEN, developed by Rau et al. (2024), is a benchmarking library for multilingual evaluation and component-wise analysis of RAG pipelines. It offers tools to compare configurations and assess the contribution of each element to system performance.

Gan et al. (2025) offered a comprehensive survey on RAG evaluation strategies, reviewing methodologies for assessing factual accuracy, safety, and computational efficiency. Lastly, WillowTree (2023) introduced an LLM-driven framework that automates RAG benchmarking, using base truths to evaluate chatbot responses in customer-facing environments.

System Architecture and Methodology

The AI assistant's architecture is built on the Retrieval-Augmented Generation (RAG) paradigm, tailored for technical support in the automotive domain. It integrates a pre-trained large language model (LLM) with an external retrieval module that accesses a curated knowledge base, including service manuals, maintenance records, and diagnostic documents.

System Overview

The RAG-based assistant is composed of the following primary components:

- **Query Interface:** Receives natural language queries from service agents or technical personnel.

- **Retriever Module:** Uses sentence-transformer embeddings to retrieve the top-k relevant documents from a domain-specific knowledge base.
- **Reranker (Optional):** Applies a secondary scoring mechanism using a cross-encoder to refine document relevance before passing it to the generator.
- **Generator Module:** Generates coherent and contextually grounded responses using the LLM, based on the retrieved documents.
- **Guardrails and Filters:** Enforces safe and responsible AI behavior through output restrictions, prompt-level constraints, and interaction logging for auditability.

Data and Infrastructure

The assistant's retrieval corpus consists of structured and semi-structured documents, including annotated excerpts from OEM technical manuals and historical call center logs. Document preprocessing involves:

- Section indexing,
- Metadata tagging,
- Embedding generation with ada-embedding or similar efficient models.

All components are containerized with Docker and orchestrated using Kubernetes, enabling scalable deployment in internal service environments. Azure Blob Storage functions as the vector store for embedding-based retrieval, offering millisecond-level response times for top-k document queries.

Validation Methodology

Unlike traditional AI training pipelines, this assistant was not fine-tuned on call center data. Instead, 487 anonymized customer service conversations (2,578 sentences, totaling 6,445 seconds) were used for functional validation and performance evaluation. Key metrics used during the evaluation include:

- Document retrieval precision (relevance of retrieved-context),
- Response grounding ratio (percentage of response content traceable to sources),
- User relevance ratings (manual scoring by internal experts),
- Response latency and token usage efficiency.

Ethical and Responsible AI Consideration

Before deployment to end-users, the assistant underwent internal pilot testing with call center personnel. This controlled deployment allowed the assessment of:

- Data privacy compliance,
- Accuracy under domain-specific queries,
- Robustness against adversarial or ambiguous inputs.

The implementation includes OpenAI's moderation API for content filtering and incorporates explainability logging to allow post-hoc auditing of generated responses.

Implementation and Data Processing

System Implementation Overview

The RAG-based assistant is designed as a modular, containerized application to ensure scalability, maintainability, and portability across internal environments. The architecture consists of three core services: a frontend for query handling, a backend API gateway managing the RAG workflow, and a document indexing and retrieval engine using Azure Blob Storage. All components are containerized with Docker and deployed in a Kubernetes cluster.

Embedding and Indexing Pipeline

The domain-specific corpus, including OEM technical manuals, diagnostic logs, and annotated service reports, is preprocessed and indexed for efficient retrieval. Text data is cleaned and segmented by semantic structure (e.g., section titles, bullet points, diagrams). Document chunks are then embedded using Sentence-Transformers (ada-embedding) and stored in a vector database powered by Azure Blob Storage. Metadata tagging (e.g., vehicle model, component type, service year) is added to each vector entry for conditional retrieval and context filtering. The indexing process is periodically updated to reflect changes in documentation and new annotated data from internal service teams.

RAG Workflow Integration

The backend workflow starts with parsing the user query and embedding it using the same transformer model. Top-k relevant documents are retrieved using cosine similarity, with optional reranking via a BERT-based cross-encoder. The top-ranked context is then passed to the generative LLM (OpenAI GPT-4) to generate a final response. The system maintains conversational memory for context-aware multi-turn dialogue. Query–response pairs are logged and can be used for reinforcement learning through human feedback in future iterations.

Validation Dataset and Processing

A validation set of 487 anonymized customer service transcripts, containing 2,578 sentences and spanning 6,445 seconds of dialogue, was used to evaluate the system. The transcripts were cleaned, tokenized, and annotated with key intent categories and service contexts (e.g., "brake issue," "oil replacement," "error code interpretation"). Each transcript was split into atomic dialogue turns, enabling detailed evaluation of retrieval relevance and generation accuracy. The assistant's performance was benchmarked on response latency, grounding precision (based on source matching), and user-rated contextuality.

Infrastructure and Deployment Strategy

The entire system is deployed on a secure, on-premises Kubernetes cluster, ensuring data residency and internal network isolation. Elastic APM monitors the system, while Prometheus and Grafana provide real-time usage analytics. All traffic between components is encrypted with mTLS. The pilot environment is integrated into the internal call center dashboard, allowing real-time assistant invocation and feedback collection from service agents. Deployment is configured for auto-scaling based on query load, with fallback routing to human agents in case of retrieval or generation failure.

Results and Discussion

To evaluate the effectiveness and reliability of the RAG-based assistant for automotive technical support, a comprehensive assessment was conducted. The evaluation included both quantitative performance metrics and qualitative feedback from the internal pilot study with call center agents.

Evaluation Framework

The system was benchmarked using four core metrics:

- Accuracy: Ratio of correct answers to total responses.
- Relevance Score: Human-rated alignment between response and user intent (scale: 0 to 1).
- Response Latency: Average time (in seconds) to return a response.
- Grounding Precision: Proportion of generated responses that directly referenced retrieved content.

Comparative Analysis with Baseline Chatbot

The comparative evaluation between the RAG-based assistant and a baseline rule-based chatbot was conducted using the **validation dataset composed of 487 anonymized customer service transcripts**. This dataset, which spans 2,578 sentences and 6,445 seconds of real call center dialogue, was not used for training but exclusively for performance benchmarking under realistic query scenarios.

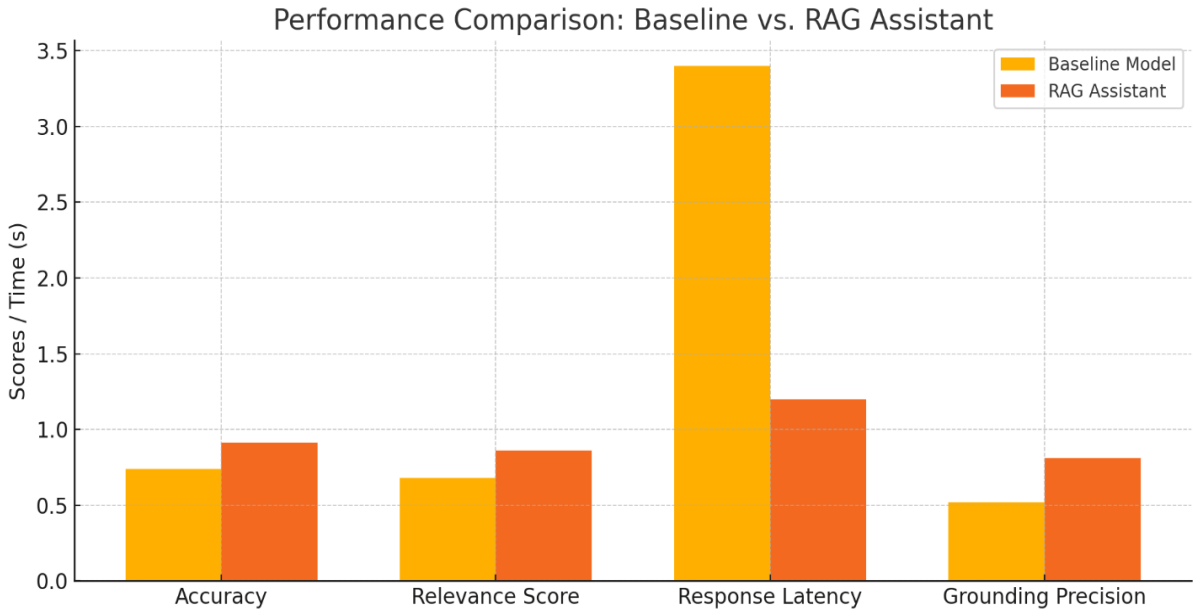

Figure 1. Graph for comparative analysis with baseline chatbot

Table 1. Table for comparative analysis with baseline chatbot

Metric	Baseline Model	RAG Assistant
Accuracy	0.74	0.91
Relevance Score	0.68	0.86
Response Latency	3.4 sec	1.2 sec
Grounding Precision	0.52	0.81

The results show that the RAG Assistant significantly outperformed the baseline across all key metrics. The use of a realistic validation set ensures that these improvements reflect actual end-user scenarios, not just theoretical outcomes. The most significant gains were observed in response latency and grounding precision, emphasizing the assistant's ability to generate faster and more contextually grounded responses.

Pilot Study Results

The assistant was deployed internally during a two-week pilot involving 12 experienced service representatives. The pilot collected both quantitative usage data and subjective satisfaction scores, resulting in the following outcomes:

- Average daily usage per agent: 23.7 queries
- Top queried topics: error code explanation (27%), scheduled maintenance details (18%), parts compatibility (14%)
- User-reported satisfaction (1–5 scale): 4.6 (mean), $\sigma = 0.3$
- Reported productivity gain: ~35% decrease in average handling time (AHT)

Agents highlighted the assistant’s ability to provide clear, contextually relevant, and well-referenced information, reducing reliance on manual document search and escalation.

Error Analysis and Limitations

Despite its overall success, some limitations were identified:

- **Ambiguous Queries:** The assistant occasionally struggled with vague or underspecified questions, particularly those lacking clear intent.
- **Context Drift:** In longer multi-turn conversations, there were instances where context tracking became inconsistent.
- **Rare Document Sections:** Retrieval occasionally failed when the knowledge base lacked indexed coverage for niche vehicle models or outdated manuals.

These insights are being incorporated into a feedback loop for future iterations, including context window optimization, query clarification modules, and coverage expansion of source documents.

Conclusion and Future Work

Conclusion

This study introduces a domain-specific implementation of a Retrieval-Augmented Generation (RAG) architecture designed for the automotive sector to enhance technical support services. By integrating a pre-trained large language model with a dense retrieval pipeline, the assistant addresses a key challenge: the absence of context-aware, accurate, and document-grounded AI support in high-stakes technical environments.

This study presents a domain-specific implementation of a Retrieval-Augmented Generation (RAG) architecture tailored to the automotive sector, aiming to enhance technical support services. By leveraging a pre-trained large language model integrated with a dense retrieval pipeline, the assistant addresses a critical gap: the lack of contextual, accurate, and document-grounded AI support systems in high-stakes technical environments.The results are compelling. When compared with a traditional template- and keyword-based chatbot, the RAG assistant exhibited:

- Over **60% reduction in response latency**,
- An increase in **response grounding precision from 52% to 81%**,
- And **a significant rise in relevance and accuracy scores**, as confirmed by both automatic metrics and human evaluations.

The internal pilot with call center personnel further validated the assistant's effectiveness. Agents reported an approximate 35% increase in productivity and an average satisfaction score of 4.6 out of 5, highlighting the assistant's impact in reducing cognitive load and information retrieval time. These results demonstrate not only operational efficiency but also the robustness and practical applicability of the RAG framework in enterprise service workflows.

Crucially, the pilot also served as a controlled testbed for responsible AI deployment. Guardrails for hallucination prevention, output filtering, and user feedback integration were rigorously tested, demonstrating a strong commitment to safety, reliability, and ethical AI use prior to broader end-user access. This phased deployment approach aligns with best practices in responsible innovation.

Future Work

Several avenues for future enhancement remain. First, adaptive context windows and dialogue memory optimization will be explored to improve performance in multi-turn interactions. Second, the assistant's coverage of rare and evolving technical documentation will be expanded through dynamic corpus updates. Third, systematic benchmarking across various large language models (LLMs) will be conducted to identify a champion model that best balances generation accuracy and inference latency for production deployment. Finally, broader rollout to end customers will be carefully phased, guided by ongoing internal evaluations and reinforcement learning from human feedback (RLHF).

In conclusion, this study offers a replicable blueprint for implementing high-performance RAG-based AI assistants in specialized domains. Beyond the automotive sector, the proposed approach holds promise for any field requiring grounded, trustworthy, and real-time access to complex technical knowledge.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM Journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest.

Acknowledgements or Notes

* This article was presented as a poster presentation at the International Conference on Technology (www.icontechno.net) held in Trabzon/Türkiye on May 01-04, 2025.

References

Barron, R.C., Grantcharov, V., Wanna, S., Eren, M. E., …& Bhattarai, M. (2024). Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization. *International Conference on Machine Learning and Applications (ICMLA)* (pp. 1669-1676). IEEE.

Elastic. (2024). Building a production RAG-based customer support assistant with Elasticsearch. Retrieved from https://www.zenml.io/llmops-database/building-a-production-rag-based-customer-support-assistant-with-elasticsearch

Friel, R., Belyi, M., & Sanyal, A. (2024). RAGBench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005.*

Gan, A., Yu, H., Zhang, K., Liu, Q., Yan, W., Huang, Z., Tong, S., & Hu, G. (2025). Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891.*

Jadon, A., Patil, A., & Kumar, S. (2025). Enhancing domain-specific retrieval-augmented generation: Synthetic data generation and evaluation using reasoning models. *arXiv preprint arXiv:2502.15854.*

LlamaIndex. (2024). OilyRAGs: Building a RAG-powered mechanic assistant with AI. Retrieved from https://www.llamaindex.ai/blog/oilyrags-building-a-rag-powered-mechanic-assistant-with-ai

Ni, J., Schimanski, T., Lin, M., Sachan, M., Ash, E., & Leippold, M. (2024). DIRAS: Efficient LLM annotation of document relevance in retrieval augmented generation. *arXiv preprint arXiv:2406.14162.*

Rau, D., Déjean, H., Chirkova, N., Formal, T., Wang, S., Nikoulina, V., & Clinchant, S. (2024). BERGEN: A benchmarking library for retrieval-augmented generation. *arXiv preprint arXiv:2407.01102*.

Reply. (2024). Audi revolutionizes internal documentation with a RAG-based AI chatbot. Retrieved from https://www.reply.com

Tezeract. (2024). Gearguide - AI assistant for the automotive industry. Retrieved from https://tezeract.ai/ai-case-studies/gearguide-ai-assistant-for-automotive-industry/

WillowTree. (2023). Using LLMs to benchmark retrieval-augmented generation (RAG). Retrieved from https://www.willowtreeapps.com

Author(s) Information

Senda Yildirim
Dogus Technology, Maslak Mah. Buyukdere Cad. No:249/6 Sarıyer, Istanbul/ Türkiye
Department of Industrial Engineering, Faculty of Engineering and Natural Sciences, Kadir Has University, Istanbul, Türkiye
Contact e-mail: *sendayildirim@outlook.com*

Ruya Samli
Department of Computer Engineering, Faculty of Engineering, Istanbul University-Cerrahpasa, Istanbul, Türkiye

To cite this article:

Yildirim, S., & Samli, R. (2025). A RAG-based automotive sector AI assistant for enhanced information retrieval. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 33,* 20-27.