Automated Evaluation of RAG Systems for Customer Support in Automotive Applications

Luis Alexander Wagner[1], Dr. Gayane Sedrakyan[1] and Prof. Dr. Jos van Hillegersberg[1]

[1]University of Twente, Drienerlolaan 5, 7522 NB Enschede, Netherlands

Abstract

Large language models (LLMs) are increasingly used in industry applications, where the output needs to meet special requirements. This paper proposes a method to automatically evaluate outputs from LLMs in specialised applications using relevant domain-specific criteria. The proposed method is applied to a case study, evaluating the output of a Retrieval-Augmented Generation (RAG) system in automotive aftermarket customer support at a filtration specialist. According to stakeholders LLM answers can be judged using the criteria for corporate language adherence, correctness, relevance and user satisfaction. For each criterion, we instructed a LLM to act as a judge, rating the response in consideration of the question, the retrieved evidence and the company context in the range 0 to 10. A threshold per criterion determines whether the output or a redirect message to a human is shown to the user. We then tested the automated evaluation using ten sample questions from actual customer interaction. For the test case, we also compared the answers of the models to the actual customer support responses.

Keywords

Retrieval-Augmented Generation (RAG), RAG evaluation metrics, semantic faithfulness, evaluation framework, domain-specific QA systems, enterprise NLP applications

1. Introduction

In the automotive aftermarket, incorrect recommendations from LLMs can lead to severe consequences. Applied to the customer support at an automotive filtration company, recommending the wrong air filter could lead to engine damage and warranty claims.

According to [1] and [2], traditional deterministic metrics evaluating the LLM response like *BLEU* and *ROUGE* fail to address domain-specific correctness in technical Q&A systems. The limits stem from their judgement by lexical overlap, weighting overlapping boilerplate content equally to crucial facts like a product name. While *BERTScore* can judge semantic alignment, it cannot score compliance with brand guidelines or validate technical accuracy against internal documentation [3].

To address this topic, multiple authors explore mission-critical LLM-based evaluation frameworks in various industries, such as medical applications and tourism [4, 5]. However, the evaluation of special requirements in the customer support of the automotive aftermarket is not covered in academia. To bridge this gap, we apply the *Design Science Research Methodology* (DSRM) to propose a LLM-based evaluation framework for the automotive area.

The DSRM is focusing on iterative increment refinement while ensuring alignment between academia and industry. Both subjects are crucial as the academic contribution of this paper is linked to an industry case study. The first step of the DSRM includes the problem definition and its context. Also, the section aims to point out the motivation for the development of a new artifact [6].

2. Problem Identification & Motivation

The title of the paper refers to *automated evaluation* of a RAG system. According to [7], the term refers to the scoring of work automatically through computer programmes. Applied to the paper, the term

IR-RAG @ SIGIR'25: Workshop on Information Retrieval's Role in RAG Systems,July 2025,Padua, Italy

luis.wagner@utwente.nl (L. A. Wagner); g.sedrakyan@utwente.nl (Dr. G. Sedrakyan); j.vanhillegersberg@utwente.nl (Prof. Dr. J. v. Hillegersberg)

refers to scoring the output (answer) of a RAG system compare to its human input (question) using multiple criteria. We break down the problem in three sub-problems to facilitate the solution design.

2.1. Criteria Definition

For an evaluation, the first step is to design the criteria on which the answer should be scored. The criteria should holistically measure interests of the different parties involved. For the case study, *marketing*, *brand management*, *product management*, *sales*, and *customer support* are interviewed.

2.2. Threshold Design

The second step is to collaboratively define a scoring range and an acceptance level for each criterion. In case the answer scores below the threshold, the question is flagged to be processed by a human and the answer is not displayed to the questioner. Thresholds were calibrated iteratively with stakeholders in a co-creation study to balance risk tolerance and automation efficiency.

2.3. Automated Evaluation

Having criteria and thresholds, the automated evaluation is integrated in the operational RAG system. The step also includes the automatic evaluation of ten questions from customer support.
The next phase of the DSRM aims at creation of the conceptual and artifactual solution to address the outlined problems. In this paper, we will introduce the final output of the phase.

3. Design and Development

Collaborative workshops with the stakeholders defined four evaluation criteria:

- **Corporate Language Adherence:** Alignment with brand guidelines, including tone, terminology, and style. Lower threshold: 7.
- **Correctness:** Factual accuracy of the answer against retrieved technical documentation. Lower threshold: 7.
- **Relevance:** Appropriateness of the response to the user's query and context. Lower threshold: 7.
- **User Satisfaction:** Perceived clarity, helpfulness, and politeness of the answer. Precisely formulated negative answers should also reach high user satiafaction. Lower threshold: 7.

For each section, we set up an independent *Mistral 7B* LLM instances per criterion. The LLM is instructed with a role description to act as a judge, the criterion definition, and the goal to rate the answer based on the assigned criterion. The scoring happens as a last step in the RAG system, after the answer is already generated and ready to be displayed to the user. To perform the scoring, the LLM receives the answer, the question and the retrieved data. For each instance, we extract the score integer and check against the threshold. Finally, we check if at least one of the answers does not meet the minimum required score.
The next phase of the DSRM shows that the artifact can solve the problem in a relevant environment.

4. Demonstration

The automated evaluation of the RAG-based Q&A system demonstrated high adherence to corporate language, with consistent scores above 8 out of 10. Only in one case, the corporate language adherence was rated below 7, resulting in the answer not being displayed to the user. Correctness achieved a median score of 10/10 with variability and outliers reaching 7/10 as the lower limit. Response speed remained under 10 seconds for all queries, making the system suitable for email-based user support but not yet optimised for real-time chatbot applications. User satisfaction scores varied most, reflecting diverse

user expectations. As anticipated, traditional metrics such as BLEU and ROUGE did not accurately reflect output quality, underscoring the need for more context-aware evaluation methods. These results highlight the evaluation system's potential in reflecting stakeholder-relevant metrics in a demonstration. The final step in the DSRM process assesses how well the artifact meets the defined objectives.

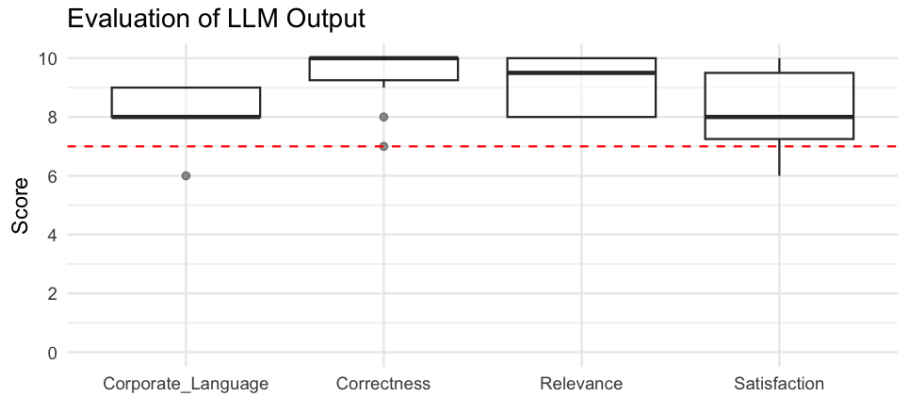

Figure 1: Scores across four criteria for ten test queries.

5. Evaluation & Recommendation

The case study demonstrated the framework's effectiveness in balancing automation with risk mitigation. The evaluation step rated the answers with median scores of 8/10 for corporate language adherence (range: 6-10) and 10/10 for technical correctness across 10 real-world queries. Notably, 80% of responses met all four criteria thresholds simultaneously, while one query was redirected to human agents per the safety thresholds. This happened likely as the answer matched the slang of the questioner. Compared to human responses, the system showed equivalent correctness (100% match on product data lookup) but lower user satisfaction scores on queries requiring multiple answer sections. This supports our view that domain-specific evaluation requires moving beyond traditional metrics. This approach establishes a template for mission-critical LLM evaluation across regulated industries where traditional NLP metrics prove inadequate. The modular criterion design enables customization for other automotive applications like technical documentation validation or service recommendation systems. Key Limitations of this case study includes the small validation sample (10 queries) limits statistical significance. Moreover, the same model was used for creating and judging the answer. This could lead to a self-reinforcement bias. In this case different models could take over the different roles as proposed in [8]. Finally, the theresholds were designed statically, requiring further adaptations based on criticality.

Acknowledgments

The authors would like to thank the industry partners at the filtration specialist for their contribution to the criteria definition and case study data.

Declaration on Generative AI

During the preparation of this work, the authors used Mistral 7B and GPT-4 in order to: Generate responses for the RAG system and perform automated scoring of those responses as part of the research artifact. After using these services, the authors reviewed and edited the content and take full responsibility for the publication's content.

References

[1] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, Bleu: a method for automatic evaluation of machine translation (2002). doi:10.3115/1073083.1073135.

[2] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[3] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. arXiv:1904.09675.

[4] N. T. Ngo, C. V. Nguyen, F. Dernoncourt, T. H. Nguyen, Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering, 2024.

[5] B. S. Ahmed, L. Otto, Baader, Quality assurance for llm-rag systems: Empirical insights from tourism application, 2025.

[6] K. Peffers, T. Tuunanen, M. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, Journal of Management Information Systems 24 (2007) 45–77.

[7] ShuaiZhang, Journal of china computer assisted language learning, 2021. doi:.org/10.1515/jccall-2021-2007.

[8] C.-M. Chan, W. Chen, Y. Su, J. Yu, Z. Liu, Chateval: Towards better llm-based evalua- tors through multi-agent debate, 2023.