Evaluation of User Experience with RAG-based Chatbots for Searching Documentation: Industrial Case Study

Marko Vještica
University of Novi Sad, Faculty of Technical Sciences
Novi Sad, Serbia *marko.vjestica@uns.ac.rs*

Elena Akik
University of Novi Sad, Faculty of Technical Sciences
Novi Sad, Serbia *elena@uns.ac.rs*

Vladimir Dimitrieski
University of Novi Sad, Faculty of Technical Sciences
Novi Sad, Serbia *dimitrieski@uns.ac.rs*

Lukas Hinterleitner
KEBA Industrial Automation
Linz, Austria *hilu@keba.com*

Jovan Erić
KEBA Industrial Automation
Linz, Austria *eri@keba.com*

Frank-Christian Weidenfelder
KEBA Industrial Automation
Linz, Austria *weid@keba.com*

Milan Pisarić
KEBA Industrial Automation
Linz, Austria *pisa@keba.com*

Abstract

Artificial Intelligence (AI) has accelerated digital transformation across industries, with Large Language Models (LLMs) powering content generation, summarization, and dialogue systems, yet struggling with domain-specific knowledge. In industrial settings, Retrieval-Augmented Generation (RAG) architectures, often implemented as chatbots, address this issue by grounding responses in internal company knowledge. Despite increased industrial deployment, user experience evaluations of RAG-based chatbots remain limited, particularly regarding their effectiveness in supporting domain-specific workplace tasks. In this paper, we present a user evaluation of an RAG-based chatbot conducted in a medium-sized company. Employee feedback on chatbot usability and acceptance is analyzed to guide digitalization efforts in future AI-assisted enterprises.

Keywords: Evaluation, Chatbot, Retrieval-Augmented Generation, Vector Database, Large Language Model

1. Introduction

Large Language Models (LLMs) are acknowledged as Machine Learning (ML) models with transformative impact across applications, including automated text generation, translation, conversational agents, and content summarization, marking a significant advancement in Artificial Intelligence (AI) [5]. However, despite excelling at text

processing, LLMs often struggle with proprietary or domain-specific tasks due to pretraining on general-purpose corpora and lack of access to internal knowledge. Retrieval-Augmented Generation (RAG) architectures address this by using an external retrieval component to select relevant context from knowledge bases and integrate it into the generative process, thereby providing domain-specific information [1].

In document searching, RAG architectures are primarily represented through chatbots communicating with end-users via natural language. Various RAG-based chatbots serve industrial use cases, and different studies evaluated their performance metrics, but to the best of our knowledge, no systematic evaluation of user experience in industrial contexts has been conducted. As practical utility is determined by technical outcomes and end-user adoption, in this paper, we present a user evaluation of an RAG-based chatbot, deployed with employees of a medium-sized manufacturing company. The findings are expected to provide insights into the future development of RAG-based chatbots for Question and Answer (Q&A) and support the digital transformation of companies.

Besides the Introduction and Conclusions, this paper is structured as follows. In Section 2, literature related to RAG-based systems and their evaluation is reviewed. The experiment setup and user experience evaluation results are discussed in Section 3.

2. Background and Related Work

In this section, research on the industrial applications and evaluations of LLMs and RAG architectures is reviewed. In manufacturing, an LLM-based system improved internal knowledge sharing by interpreting technical documents and expert input [4]. A user study reported efficient retrieval, while benchmarks found that open-source models excel in data protection and customization, whereas proprietary models yield higher accuracy.

By combining LLMs with RAG, a conversational monitoring system was implemented to enable real-time data fetching and natural language interaction for shop floor decision-making [3]. Evaluations across multiple LLMs and usability tests showed high user satisfaction, operational efficiency, and context-aware responses. In the energy optimization domain, a multi-agent RAG system extracted information from audit reports [7], with case studies and benchmarking highlighting its low operational costs and reliability with diverse inputs.

Asset Administration Shell (AAS) generation in digital twins was automated using LLMs for datasheet processing and RAG for enriched semantics [6]. Graduate annotators confirmed usability through bypass rates and quality scores. Similarly, in energy infrastructure management, an RAG-powered assistant supported data retrieval within a digital twin [2], improving operational oversight, forecasting, and responsiveness in high-voltage networks.

While industrial RAG evaluations have prioritized quantitative benchmarks, human-centered approaches remained underexplored. Thus, in the following section, an RAG-based documentation search chatbot and user experience evaluation are presented.

3. Evaluation of User Experience with Documentation Search Chatbot

In this section, we discuss the experiment and evaluation results of user experience with a chatbot for documentation search. First, we present an overview of the RAG-based chatbot provided to participants, followed by the experiment setup and results derived from the questionnaire filled in by the experiment participants.

3.1. RAG-based Chatbot for Documentation Search

To search internal documents in a company and evaluate user experience with chatbots, we applied the basic RAG architecture, as presented in Fig. 1, and developed the chatbot named Document Search Bot (DSB). The architecture includes two flows of activities: (i) the document insertion flow (dashed lines); and (ii) the querying flow (solid lines).

The insertion of documents is initialized by a user (step I), who forwards them to the documentation parser, implemented primarily with the Unstructured library. The parser extracts text from documents, creating text chunks (step II) that are forwarded to the

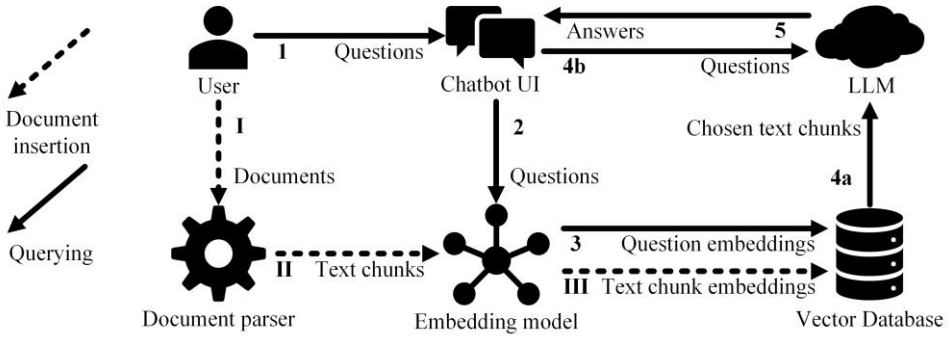

Fig. 1. The basic RAG architecture for searching documentation

Multilingual-E5-large-instruct embedding model. It transforms text chunks into vector embeddings (step III) that are stored in the Milvus vector database.

The querying of stored documents is initiated by a user, who writes a question in a natural language (step 1) through the chatbot's user interface. The question is sent to the embedding model (step 2) to be transformed into a vector embedding, which is then forwarded to the vector database (step 3). By using the cosine similarity metric, the Vector Database Management System (VDBMS) finds the most similar text chunks compared to the question asked. These text chunks (step 4a) and the question (step 4b) are sent to the Claude 3.5 Sonnet LLM, which forms an answer based on the provided context and passes it to the chatbot's user interface (step 5).

3.2. Experiment Setup and Execution

The experiment for evaluating user experience with chatbots was conducted in KEBA Industrial Automation[1], an international medium-sized manufacturing company that offers various hardware and software products. Several use cases were identified for applying DSB to reduce search time: (i) technical documents of products; (ii) standards and norms; (iii) general-purpose documents, such as frequently asked questions; (iv) meeting notes; and (v) Application Programming Interfaces (APIs) needed for software development.

Vector embeddings generated from the documents were stored inside DSB vector database collections, including: (i) 8706 pages of 50 English and 5 non-English technical documents divided into 33 vector database collections to distinguish different products; (ii) 5195 pages of 51 English and 14 non-English standards and norms, divided into 2 collections based on their type and usage; (iii) 103 pages of 4 English and 4 non-English general-purpose documents stored in a single collection; (iv) 34 meetings notes written in English and stored in a single collection; and (v) 10 English specifications of APIs divided into 2 collections. Before asking a question, a user chooses a topic for which a question is to be asked, based on the provided vector database collections. A composite collection containing all technical documents was also available, allowing participants to select it once and submit any questions regarding products. However, this required them to include full context in their questions (e.g., product name and version).

Company employees and external collaborators used DSB to search internal documents over a period of at least two weeks, submitting more than 600 questions. A human-based evaluation of chatbot accuracy was conducted on a sample of 32 questions. When directed to the composite collection, the chatbot achieved an accuracy of 81.25%, but when the same questions were posed to product-specific collections, the accuracy increased to 90.63%, which is expected due to the narrower search scope. Participants included 2 service and 8 software engineers, 1 tester, 2 managers, 3 researchers, and 2 students. After using DSB, participants were asked to complete a feedback questionnaire.

The questionnaire items were grouped into 6 categories: (i) Background Knowledge – participants' familiarity with LLMs and RAG; (ii) Expectations and Preferences – the perceived practical value of chatbots, the response-speed importance, native language

[1] https://www.keba.com/en/industrial-automation/industrial-automation

support, and expected accuracy; (iii) Usage Experience – DSB usage patterns, question formulation ease, and modification frequency; (iv) Performance Perception – perceived DSB accuracy, answer clarity, and response-speed satisfaction; (v) Value Proposition – perceived DSB benefits, like time savings and search efficiency; and (vi) Overall Assessment – ratings of DSB interface intuitiveness, recommendation likelihood, and overall satisfaction. Responses on a five-level Likert scale are analyzed as follows.

3.3. Questionnaire Analysis

A comprehensive evaluation of the questionnaire was conducted to identify key factors influencing user satisfaction. Spearman's rank correlations with the False Discovery Rate correction were calculated to identify significant associations, Kruskal–Wallis tests were performed to compare role-based differences among respondent categories, and internal consistency reliability was assessed via Cronbach's alpha coefficients.

Thirteen statistically significant associations were identified following a multiple-testing correction. The strongest observed relationship was between participants' perception of DSB's time-saving benefits in finding information and their overall satisfaction with DSB ($\rho \approx 0.74$, $p \approx 0.0004$). The perceived answer accuracy provided by DSB was also strongly correlated with overall user satisfaction ($\rho \approx 0.62$, $p \approx 0.0059$). Within usage experience variables, a strong correlation was found between the frequency of modifying questions to obtain accurate answers from DSB and the perceived importance of precise wording in questions addressed to DSB ($\rho \approx 0.59$, $p < 0.05$), indicating that more engaged users often refined their queries to improve results. These findings suggest that satisfaction is predominantly driven by perceptions of efficiency and accuracy, and that iterative user behavior reflects a pursuit of these attributes.

Regression analysis using Ordinary Least Squares confirmed these insights, explaining approximately 73% of the variance in overall satisfaction ($R^2 = 0.729$; adjusted $R^2 = 0.671$; $F = 12.54$, $p \approx 0.0003$). Perceived time-saving benefits ($\beta = 0.5145$, $p = 0.008$) and perceived answer accuracy DSB provided ($\beta = 0.5459$, $p = 0.008$) emerged as significant independent predictors of overall user satisfaction, while the satisfaction with DSB's replying speed did not ($\beta = 0.1230$, $p = 0.483$). These results imply that speed enhancements only contribute to satisfaction when they reinforce efficiency and answer reliability. No statistically significant differences were found across roles, suggesting that satisfaction drivers are consistent regardless of job category.

Although internal consistency across questionnaire categories was moderate (Cronbach's α between 0.49 and 0.66), perceived efficiency and accuracy were confirmed as primary satisfaction drivers for RAG-based chatbots. Based on participants' feedback, they expected response accuracy to exceed 90%. Enhancing reliability may involve not only correct answers but also features such as displaying confidence scores or answer reasoning. System design should prioritize observable time savings and factual accuracy over speed optimization alone. Meeting these goals depends on robust infrastructure (e.g., high-performance local GPUs or cloud services). Statistically significant questionnaire items and percentages of participant responses are presented in Table 1.

Table 1. A summary of statistically significant questionnaire items

Questionnaire item	1	2	3	4	5
Perceived time savings DSB provided in finding information	0.0%	0.0%	16.7%	61.1%	22.2%
Perceived answer accuracy DSB provided	0.0%	0.0%	44.4%	55.6%	0.0%
Question modification frequency to obtain accurate answers from DSB	11.1%	16.7%	55.5%	16.7%	0.0%
Perceived importance of precise wording in questions addressed to DSB	0.0%	0.0%	11.2%	44.4%	44.4%
Satisfaction with DSB's replying speed	0.0%	0.0%	11.1%	61.1%	27.8%
Overall user satisfaction with DSB	0.0%	0.0%	11.1%	61.1%	27.8%

3.4. Threats to Validity

In this section, we address potential threats to validity, as results may vary across experimental iterations. The experiment involved 18 participants from a single company

who used DSB for at least two weeks. A broader evaluation across multiple companies, with larger group samples and longer duration, might yield more accurate findings. Since user experience might depend on both DSB implementation and answer accuracy, evaluation results might differ if another chatbot implementation was provided. We developed the basic RAG solution that provided satisfactory answer accuracy, offering a solution that is neither too advanced nor too inaccurate, balancing the user experience.

4. Conclusions

In this paper, we evaluated user experience with an RAG-based chatbot for Q&A in the industry and analyzed questionnaire responses. To enhance user satisfaction, companies need to offer efficient and reliable chatbot solutions, primarily prioritizing high answer accuracy and, secondarily, low response time. Therefore, time savings in information retrieval, compared to manual search, emerged as the key satisfaction factor. When participants' questions were analyzed, ambiguousness in some of them was observed, resulting in question modifications. This indicates that chatbots need to offer users assistance in formulating questions precisely. Thus, a chatbot integrating multiple LLMs and data storage filled in with documents labeled or grouped based on similar topics should operate on a capable local or cloud infrastructure and be deployed for daily use.

To gain deeper insights into chatbot user experience, future experiments will involve multiple companies and a larger group of participants. Substantial questionnaire revisions are also recommended, including clearer phrasing of items, addressing reverse-coded items, and regrouping or eliminating poorly aligned items to improve scale coherence.

Acknowledgements

This research has been supported by KEBA Industrial Automation; and by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-137/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025" (No. 01-50/295). We are thankful to the evaluation participants for their time and useful feedback.

References

1. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking Large Language Models in Retrieval-Augmented Generation. Proc. AAAI Conf. Artif. Intell. 38 (16), 17754–17762 (2024)
2. Ieva, S., Loconte, D., Loseto, G., Ruta, M., Scioscia, F., Marche, D., Notarnicola, M.: A Retrieval-Augmented Generation Approach for Data-Driven Energy Infrastructure Digital Twins. Smart Cities. 7 (6), 3095–3120 (2024)
3. Jeon, J., Sim, Y., Lee, H., Han, C., Yun, D., Kim, E., Nagendra, S.L., Jun, M.B.G., Kim, Y., Lee, S.W., Lee, J.: ChatCNC: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. J. Manuf. Syst. 79 504–514 (2025)
4. Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., Niforatos, E.: Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. Front. Artif. Intell. 7 1293084 (2024)
5. Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., Chen, E.: A survey on large language models for recommendation. World Wide Web. 27 (5), 60 (2024)
6. Xia, Y., Xiao, Z., Jazdi, N., Weyrich, M.: Generation of Asset Administration Shell With Large Language Model Agents: Toward Semantic Interoperability in Digital Twins in the Context of Industry 4.0. IEEE Access. 12 84863–84877 (2024)
7. Xiao, T., Xu, P.: Exploring automated energy optimization with unstructured building data: A multi-agent based framework leveraging large language models. Energy Build. 322 114691 (2024)