Towards User-Centric Chatbot for Q&A in Industry: Evaluation of Employees' Experience

Marko Vještica, Elena Akik, Vladimir Dimitrieski
University of Novi Sad, Faculty of Technical Sciences
Department of Computing and Control Engineering
Trg Dositeja Obradovića 6, 21102, Novi Sad, Serbia
{marko.vjestica, elena, dimitrieski}@uns.ac.rs

Lukas Hinterleitner, Jovan Erić, Frank-Christian Weidenfelder
KEBA Industrial Automation GmbH
Reindlstraße 51, 4040, Linz, Austria
{hilu, eri, weid}@keba.com

Jonas Wieczorek
KEBA Industrial Automation Germany GmbH
Gewerbestraße 5-9, 35633, Lahnau, Germany
Jonas.Wieczorek@keba.com

Abstract. *Chatbots are increasingly being adopted in industrial settings to assist employees in their daily tasks, typically by leveraging the Retrieval-Augmented Generation (RAG) architecture. While technical benchmarking of such chatbots is common, assessing their performance, evaluations of user preferences and chatbots' practical usability are often neglected. To address this gap, we conducted an in-company experiment by deploying an RAG-based chatbot for use by the company's employees. In this paper, we present an analysis of their experience with the chatbot and highlight usability challenges. Based on the experiment findings, we also propose design recommendations for developing user-centric chatbots tailored to industrial environments.*

Keywords. User Experience, Retrieval-Augmented Generation, Chatbot, Large Language Model, Multi-Agent System

1 Introduction

A profound digital transformation is being driven across the global industrial sector through the adoption of advanced Artificial Intelligence (AI) and Machine Learning (ML) technologies (Sarker et al., 2023), enhancing productivity, decision-making, and data management. Among diverse ML technologies, Large Language Models (LLMs) are acknowledged as transformative ML models, automating tasks such as text generation, language translation, and content summarization, in the form of conversational agents (Chen et al., 2024; L. Wu et al., 2024).

While generative and comprehension capabilities are demonstrated by LLMs, they remain limited by their inability to access proprietary or domain-specific knowledge in enterprise settings. To address this deficiency, external retrieval components are employed by the Retrieval-Augmented Generation (RAG) architecture, grounding the LLM's output in domain-specific information (Chen et al., 2024). The RAG architecture that includes an LLM is typically implemented as a chatbot through which natural language communication with end-users is facilitated.

Various RAG-based chatbots serve industrial use cases, yet user experience evaluations, including interaction ease, result satisfaction, and workplace acceptance, are typically neglected. Such evaluations are essential, as practical utility in corporate environments is determined by both technical outcomes and user adoption capacity (Pulapaka et al., 2024). Existing research has focused predominantly on technical performance metrics rather than practical utility and user acceptance. Accordingly, the expectations that users hold toward RAG-based chatbots, such as communication style and response delivery, remain underexplored.

In this paper, a comprehensive user evaluation of an RAG-based chatbot for Question and Answer (Q&A) was conducted with employees from a medium-sized manufacturing company. Participants across various roles interacted with the chatbot to retrieve the company's internal knowledge from various documents, followed by the structured feedback collection. Various user experience dimensions related to the chatbot usage were assessed, including perceived accuracy, response latency, answer structure preferences, ease of use, context selection workflows, language support, and overall satisfaction.

The contribution of this research centers on deploying RAG-based chatbots in authentic industrial environments and systematically measuring multi-

dimensional user experience factors. Thus, empirical connections were established between findings and actionable design recommendations for AI-assisted information retrieval in corporate settings.

Following the Introduction, this paper is organized as follows. In Section 2, related work regarding the application and evaluation of RAG-based systems is discussed. The basic RAG architecture and implementation of a chatbot for Q&A are outlined in Section 3. The evaluation of user experience with the RAG-based chatbot is presented in Section 4. The discussion regarding research findings is presented in Section 5, while conclusions and future work are outlined in Section 6.

2 Related Work

In this section, existing research on the application of RAG-based systems in industry and their evaluation is reviewed. Such systems have been investigated across various industrial domains in order to enhance information retrieval, knowledge management, and decision-support systems.

A novel system was introduced to support internal knowledge sharing within manufacturing setups by leveraging LLMs to interpret and retrieve information from factory documentation and expert insights (Kernan Freire et al., 2024). User studies confirmed accelerated document retrieval, with benchmarking showing state-of-the-art models excelling in factual accuracy while open-source alternatives offered customization and privacy benefits.

A conversational monitoring framework for shop floor decision-making integrated LLMs within RAG to retrieve real-time data from Computer Numerical Control (CNC) machines, Industrial Internet of Things (IIoT) sensors, and manufacturing databases (Jeon et al., 2025). Framework evaluations encompassed experiments with various LLM-based agents and human usability assessments, demonstrating that the framework reliably provided context-aware responses, improved operational efficiency, and achieved high user satisfaction.

LLMs were integrated within RAG for real-time industrial troubleshooting by combining technical manuals, incident reports, and sensor data (Narimani & Klarmann, 2024). This context-aware approach enhanced production efficiency and minimized downtime. Evaluation employed quantitative metrics to assess the relevance of retrieved documents, complemented by expert reviews. Results indicated effective information retrieval, high user satisfaction, and troubleshooting efficiency, although specific quantitative outcomes were not detailed.

Expanding on the manufacturing domain, a framework was proposed for extracting process-level knowledge in additive manufacturing (Liu et al., 2025). LLMs were integrated within the RAG architecture to automatically customize taxonomies for specific manufacturing processes and facilitate the extraction of technical entities. The framework was evaluated by benchmarking in-context learning and fine-tuning against a baseline model, demonstrating superior extraction accuracy and reduced model hallucination.

In digital twin environments, an automated tool was developed for generating Asset Administration Shell (Xia et al., 2024). LLMs converted raw datasheet text into structured formats, while RAG enhanced semantic understanding. Evaluations by graduate annotators included usability assessment, LLM comparisons, and an ablation study, with effectiveness confirmed by bypass rates, helpfulness scores, and quality ratings.

Additionally, a framework was proposed for energy infrastructure management, integrating digital twins and data-driven approaches (Ieva et al., 2024). RAG was employed to enhance a conversational assistant, enabling the retrieval of domain-specific information and enrichment of sensor data and knowledge graphs. Prototype implementations were evaluated in a real-world case study, and improved management, predictive capabilities, timeliness, and accuracy were observed in high-voltage network operations.

In building energy optimization, an LLM-based multi-agent system was described as incorporating RAG to extract information from unstructured energy audit reports (Xiao & Xu, 2024). Case studies demonstrated robust performance at low operational costs through performance testing and benchmarking. Results confirmed the system's ability to handle diverse inputs reliably.

A framework proposed in the construction industry witnessed advancements where RAG applications enhanced document retrieval (C. Wu et al., 2025). A software system parsed documents, such as inspection reports, standards, and working plans. Quantitative benchmarking was employed to evaluate the system, demonstrating improvements in retrieval accuracy and response efficiency.

In the same domain, LLMs were integrated within RAG to enhance information retrieval and process automation (Taiwo et al., 2025). The evaluation combined a Delphi study including domain experts and a benchmarking case study. The findings indicated improvements in output quality, relevance, and reproducibility.

In the telecommunications sector, a system was devised to navigate complex technical standards by constructing a knowledgebase from industry-relevant technical documentation (Yilma et al., 2024). The evaluation was conducted through technical benchmarking, wherein system-generated responses were assessed for accuracy, technical depth, and verifiability. Comparative analyses demonstrated that the proposed system outperformed generic LLMs, producing more precise and verifiable outputs.

In the oil and gas sector, a comprehensive leakage detection approach was presented where LLM-based agents and RAG were used to construct a knowledge vector library from professional documents (Wei et al.,

2024). Improvements in Q&A accuracy and complex reasoning were observed compared to traditional methods. The evaluation was performed through benchmarking against experimental data and further validated in practical deployment scenarios.

In an industrial setting, a conversational RAG-based chatbot for query resolution in Continuous Integration and Continuous Delivery (CI/CD) workflows was developed (Chaudhary et al., 2024). The chatbot was benchmarked on CI/CD questions from an industrial context, showing mostly correct responses with some partial or incorrect answers. It also demonstrated high accuracy in technical queries.

In the research on Electron-Ion Collider (EIC), a RAG-based summarization system was developed to extract information from EIC documentation (Suresh et al., 2024). Relevant content was indexed within a vector database, and citation-enriched summaries were produced by a fine-tuned LLM. System performance was evaluated by using the Retrieval-Augmented Generation Assessments (RAGAs) framework (Es et al., 2024) to improve researcher accessibility and literature navigation.

Evaluation techniques across the reviewed studies on RAG-based applications in industrial domains predominantly relied on quantitative benchmarking to assess performance improvements. Human-based evaluations, such as expert reviews, usability studies, and Delphi evaluations, were used in a smaller subset of investigations to measure user satisfaction with the chatbot's response.

Although benchmarking methods have been widely applied, comprehensive user-centric evaluations have been relatively underrepresented. This gap highlights the need for further research incorporating human usability assessments in order to provide a more complete understanding of the real-world applicability of RAG-enhanced systems. Especially in cases of Q&A applications in industrial domains, a user-centric evaluation of using and applying chatbots for retrieving relevant information is needed. Such an evaluation would contribute to the understanding of what users expect from such systems to use them in practice. Therefore, in the following sections, we present one such RAG-based system for searching documents in a company, and the evaluation results based on the examination of user experience utilizing the presented system and AI assistants in general.

3 Basic RAG Architecture for Documentation Search

To evaluate the user experience with chatbots used for company's internal documentation search and to gain insights into user expectations regarding such software solutions, we implemented an RAG-based chatbot named Document Search Bot (DSB). In this section, we outline a basic RAG architecture on which DSB is built, as well as its implementation. Leveraging content retrieved from document storage, DSB assists users in finding relevant information by enabling natural language interactions.

3.1 RAG Architecture

The basic RAG architecture includes two main flows of activities: (i) the document insertion flow; and (ii) the querying flow. The document insertion flow is discussed first, as it is a prerequisite for querying the documents.

As text within documents represents unstructured data, one way to semantically search for relevant information inside documents is to use a Vector Database Management System (VDBMS). Such a system manages a vector database that stores vector embeddings, representing data of various types in a multi-dimensional vector space.

To insert a document in a vector database, its text needs to be parsed and transformed into vector embeddings. Therefore, the following steps of the document insertion flow are needed, as presented in Fig. 1. First, a user initiates the insertion of documents by sending them to the document parser. The document parser is a component that extracts text from documents and splits it into multiple text chunks. These text chunks are sent to the embedding model, which represents a generative AI model that transforms data into vector embeddings. After the transformation, text chunk embeddings are sent and stored inside the vector database. Meta-data regarding vector embeddings, such as the name of a document and the page number from where they originate, as well as their original content, can be stored inside the vector database as well. The VDBMS can be used afterward to search vector embeddings for relevant information.

Whenever a user aims to search for relevant information in documents, a question can be written in

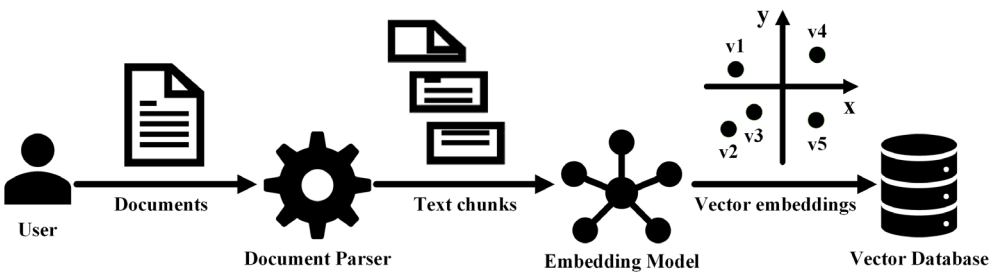

Figure 1. The document insertion flow within a basic RAG architecture

natural language, and the answer is received in the same manner. To achieve such a Q&A feature, the following steps are included in the querying flow, as presented in Fig. 2.

A user initiates querying by writing a question using the chatbot's User Interface (UI). The question is then forwarded to the embedding model, transforming it into a vector embedding, which is sent to the vector database. The distances between the question embedding and the documents' text chunk embeddings are measured by the VDBMS using a chosen similarity metric. Therefore, text chunk embeddings nearest to the question embedding are searched in a multi-dimensional vector space, and their textual representations are sent to the LLM, alongside the question being asked. Based on the provided question, text chunks, and instructions in the form of a prompt, the LLM generates the answer in natural language and sends it to the chatbot's UI, displaying the answer to a user.

3.2 Document Search Bot

DSB represents a Python implementation of the RAG architecture discussed. The document parser is implemented by using the Unstructured library (*Unstructured.io*, 2025). It extracts the text content from paragraphs, tables, and images, but eliminates irrelevant or repeating content, such as that exists in header and footer sections or the table of contents. As documents and questions may be written in various languages, the embedding model needs to support transforming text written in different languages to vector embeddings. Therefore, the Multilingual-E5-large-instruct model (L. Wang et al., 2024) is used for such a purpose. The Milvus VDBMS (J. Wang et al., 2021) is used to store vector embeddings generated from the embedding model, calculate similarity measures using the cosine similarity metric when a question is asked by a user, and retrieve relevant text chunks.

User questions are written on the DSB UI implemented using the Streamlit framework (Khorasani et al., 2022). A user may choose a collection of documents to be searched for an answer before writing a question. A question is added to the prompt sent to the LLM, instructing it to form the answer in the language in which the question is being asked, based on the text chunks provided, or to clearly state whether the answer cannot be found in the given context. The Claude 3.5 Sonnet LLM (*Claude 3.5 Sonnet*, 2025) is used to generate an answer to a user's question, which is displayed on the DSB UI. The LLM is utilized through the Amazon Bedrock service (*Amazon Bedrock*, 2025) to provide an answer rapidly. After receiving an answer, a user may rate the answer and leave a comment using the DSB UI. All questions, received text chunks, generated answers, and user feedback are stored for analysis purposes.

The presented chatbot is given to the participants of an experiment conducted to evaluate the applicability and usability of chatbots in a company. The experiment and evaluation results are presented in the following section.

4 Evaluation of User Experience with Chatbots in Company

In this section, we present an evaluation of the user experience with a chatbot designed for searching documentation in a medium-sized manufacturing company. We outline the preparation of the experiment conducted and the feedback provided by the experiment participants, thus evaluating the applicability and usability of chatbots.

4.1 Experiment Preparation

We prepared an experiment to evaluate user experience with chatbots in the industry. The experiment participants used DSB, outlined in Section 3.2, which was implemented to store and search the company's internal documents. After using DSB, participants filled in a questionnaire regarding the application and usage of chatbots in their daily work. In this section, we present an overview of the experiment conducted and the use cases covered by the experiment.

The experiment was performed in an international medium-sized manufacturing company, KEBA Industrial Automation[1], that develops hardware and

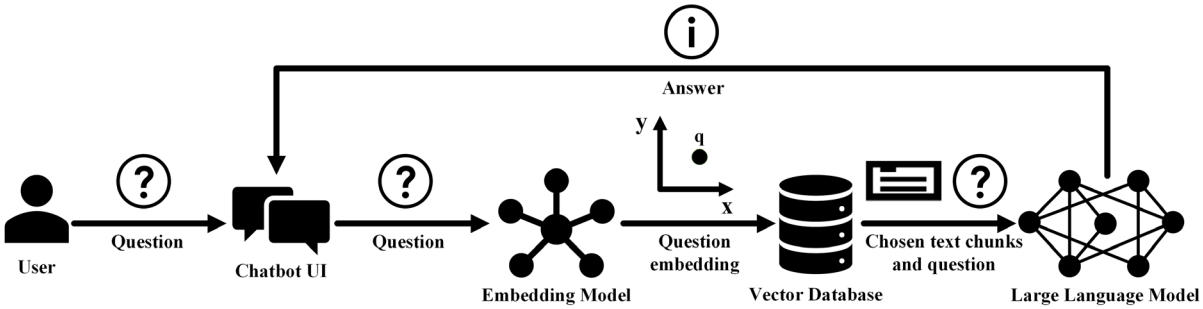

Figure 2. The querying flow within a basic RAG architecture

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

software for various industry domains. During the initial discussion with the company's employees, several use cases were identified on where to apply chatbots in the company. First, as the company has a wide range of products, its technical documentation is available to end-users. However, manually searching for relevant information can be time-consuming for end-users, and they usually have to contact the company's service engineers for help. Chatbots may be applied to help: (i) end-users find the information quickly; (ii) service engineers whenever they receive queries from the end-users; and (iii) software engineers when developing the products. Similarly, various standards and norms are applied in the company; thus, searching such extensive documents may be difficult. Additionally, various general-purpose documents in the company, such as house rules and Human Resource (HR) policies, need to be read by employees and usually demand time to locate relevant information. Sometimes, employees contact an HR assistant instead of searching for the documentation. Also, employees sometimes search for or summarize meeting notes, receiving insights into benefits, drawbacks, or conclusions derived from various meetings, which may be time-consuming. Finally, when developing applications, software engineers usually need to use various Application Programming Interfaces (APIs) already implemented by other engineers and consider what parameters to send and what response to expect. Searching for the required APIs may be tedious when working on immense software solutions. LLM-based chatbots may help employees search for relevant information rapidly in all these use cases.

However, as pre-trained LLMs do not contain knowledge regarding the company's internal documents, DSB is utilized for the presented use cases. The DSB's vector database is filled with vector embeddings generated from technical documentation, standards and norms, general-purpose documents, meeting notes, and APIs. The technical documentation included user and operation manuals, application notes, and error lists for various products and their variations, having 50 English and 5 non-English documents with 8706 pages in total. As variations of the same product can be similar to each other, and thus have their documentation, searching for relevant information may be challenging, due to difficulties in making a distinction between similar product variations and their properties. Thus, to make a better distinction between products and their variations, for each one of them, a vector database collection is created, 33 in total. Standards and norms are stored in 51 English and 14 non-English documents with 5195 pages in total. As there are two types of standards and norms, they are split into two vector database collections to narrow the search scope. The general-purpose documentation included 4 English and 4 non-English documents with 103 pages in total, such as house rules, HR policies, and frequently asked questions, stored in a single collection. Meeting notes covered records from 34 meetings written in English and stored in a single collection. DSB also included two collections for 10 APIs written in the JavaScript Object Notation (JSON) and HyperText Markup Language (HTML) formats.

Each collection of documents is presented in the DSB UI, so that a user may choose a specific product variation, a type of standard, general-purpose documents, meeting notes, or a set of APIs, for which a question is to be asked. When a collection is chosen, a link to the stored documents is presented, allowing users to check which exact documents are related to the collection. DSB also provides a single collection of all technical documents for users who prefer asking questions without choosing a specific product variation in advance, but at the expense of lower accuracy in receiving correct answers. Also, when using such a composite collection, a user must ask a question more precisely, thus including the specific product variation name within the question. The DSB UI provides a short description of how to use the chatbot, helping users search for relevant information.

DSB was used by experiment participants for at least two weeks, after which the questionnaire was provided to them. The experiment participants, questionnaire structure, and evaluation results are presented in the following section.

4.2 Evaluation Results

As a part of the experiment, DSB was available for use in the company, allowing employees and external collaborators to ask questions regarding the topics they were interested in. The company's employees who participated in the experiment included 2 service engineers, 8 software engineers, 1 tester, and 2 managers, and the company's external collaborators included 3 researchers and 2 students. Over 600 questions were asked by participants during the experiment.

After using DSB, participants were asked to fill in the online questionnaire, consisting of three main groups of questions: (i) the application of AI assistants in general; (ii) the application of chatbots for Q&A; and (iii) the application of DSB in the company. The questionnaire contained three types of questions whose answers can be: (i) chosen on a five-level Likert scale; (ii) selected from several options provided; or (iii) written as a free text. Questions with a five-level Likert scale are presented in tables in this section, showing the percentage distribution of responses, with level one indicating the lowest and level five the highest rating. Responses to the questions of the other two types are summarized in the text, without tabular representation.

4.2.1 Application of AI Assistants

The first group of questions was asked to examine how familiar participants are with AI assistants, as presented in Table 1. Participants are using AI assistants often, most frequently ChatGPT and MS

Table 1. Questions regarding the application of AI assistants

	Question	1	2	3	4	5
AI assistant application	How often are you using AI assistants?	5.6%	11.1%	33.3%	44.4%	5.6%
	How much practical use do you see in AI assistants in your company?	0.0%	0.0%	5.6%	44.4%	50.0%
	How familiar are you with Large Language Models (LLMs)?	0.0%	16.7%	27.7%	38.9%	16.7%
	How familiar are you with Retrieval-Augmented Generation (RAG) systems?	5.6%	50.0%	22.1%	5.6%	16.7%

Copilot, for gathering information, summarizing reports, writing text, checking grammar, simple coding tasks, building pieces of software, debugging code, and configuring software. Some participants also mentioned they are using AI assistants for transforming text into code, especially when dealing with programming languages they do not use often, and also for translating between different data structures.

Participants recognize much potential for the practical use of AI assistants in the company. They proposed using AI assistants in their company to assist them in writing or reviewing code on specific dialects made in the company, searching internal documents and information, automatically opening IT tickets, or searching the calendar for a free time slot.

While participants reported being mostly familiar with LLMs, their responses indicated limited familiarity with RAG systems. This discrepancy suggests that participants may not be thoroughly acquainted with the underlying technologies powering chatbots and AI assistants in general.

4.2.2 Application of Chatbots for Q&A

As presented in Table 2, when asked about the application of chatbots for Q&A in their company, participants stated that the speed of answering questions is important to them. Waiting too long for the answer may discourage employees from using a chatbot in the company. Despite all participants not speaking English natively, they mostly stated it is not so important for them to communicate with the chatbot in their native language, but there are also participants who would highly appreciate such a feature.

Participants mostly declared it as important to receive the reasoning behind the answer and the estimated confidence score, presumably raising their trust in the answer obtained from the chatbot. When asked about the minimum accuracy they expect from a chatbot to be usable for Q&A, none would expect very low (<50%) or low accuracy (50-70%), 27.7% would expect medium accuracy (70-85%), 55.6% high accuracy (85-95%), and 16.7% very high accuracy (>95%). The accuracy was defined in the questionnaire as the number of correct answers divided by the number of questions asked.

Regarding the answer structure they expect, 50% of participants stated they would like to get short and straightforward answers from a chatbot, while the other 50% of participants would like to get longer, more detailed answers with various notes. Asking any question without having to choose a collection of documents at the expense of having to ask questions more precisely is preferred by 77.8% of participants, while 22.2% of participants prefer choosing a collection before asking questions, but asking them without providing the whole context.

4.2.3 Application of DSB in Company

Regarding the application of DSB in the company, participants were first asked for which use cases they were using DSB during the evaluation process, and 13 used it for searching technical documentation, 4 for standards and norms, 8 for general-purpose documents, 5 for meeting notes, and 6 for APIs. Answers to the five-level Likert scale questions are summarized in Table 3 and grouped under four categories: (i) asking questions to DSB; (ii) receiving answers from DSB; (iii) performance and usability of DSB; and (iv) final remarks and conclusions to DSB.

Participants stated they used DSB occasionally in their daily work, as it was not necessary to frequently search for some information. They mostly declared that it was easy to formulate a question, but sometimes participants had to modify the question to get the correct or complete answer. Based on the analysis of questions being asked by participants, such a need for question alteration was mainly due to the question being ambiguously written or missing some relevant information. However, the question alteration was sometimes due to DSB's failure to find the answer, even if the question was asked precisely. Participants also agreed that the quality of the answer highly depends on the precision of the question's wording, which confirms our question analysis.

Table 2. Questions regarding the application of chatbots for Q&A

	Question	1	2	3	4	5
Chatbots for Q&A application	How important is the speed of answering a question given by a chatbot to you?	0.0%	11.1%	22.2%	55.6%	11.1%
	How important is it for you to communicate in your native language (mother tongue) with a chatbot?	5.5%	50.0%	16.7%	16.7%	11.1%
	How useful would it be for you to receive the reasoning behind the answer from a chatbot?	0.0%	0.0%	38.9%	38.9%	22.2%
	How important is it for you to receive an estimation of the chatbot's confidence in the provided answer?	0.0%	0.0%	16.7%	55.6%	27.7%

Table 3. Questions regarding the application of DSB in the company

	Question	1	2	3	4	5
DSB application: asking questions	How often do you use DSB in your company?	0.0%	33.3%	55.5%	5.6%	5.6%
	How easy is it for you to formulate your questions to DSB?	0.0%	5.6%	22.1%	66.7%	5.6%
	How often did you have to modify the question to get a good answer?	11.1%	16.7%	55.5%	16.7%	0.0%
	How much do you think the quality of an answer depends on the precision of the question's wording?	0.0%	0.0%	11.2%	44.4%	44.4%
DSB application: receiving answers	How satisfied are you with the accuracy of the answers provided by DSB?	0.0%	0.0%	11.1%	83.3%	5.6%
	To what extent do you think the answers are understandable and clearly formulated?	0.0%	0.0%	5.6%	72.2%	22.2%
DSB application: performance and usability	Are you satisfied with how fast DSB replies to you?	0.0%	0.0%	11.1%	61.1%	27.8%
	DSB saves your time in solving problems or getting information.	0.0%	0.0%	16.7%	61.1%	22.2%
	DSB reduces the time it takes you to manually search documents.	0.0%	0.0%	5.6%	61.1%	33.3%
	DSB helps you avoid having to ask other employees for certain information.	0.0%	0.0%	38.9%	50.0%	11.1%
	How intuitive and easily understandable is DSB's user interface to you?	0.0%	0.0%	16.7%	55.5%	27.8%
DSB application: conclusion	How satisfied are you with the current DSB solution overall?	0.0%	0.0%	11.1%	61.1%	27.8%
	In your company, how applicable is DSB in practice?	0.0%	0.0%	22.2%	50.0%	27.8%

Based on their experience, participants were mostly satisfied with the DSB's accuracy, with 44.4% of participants estimating it as medium (70-85%), while 55.6% as high (85-95%). Additionally, similar accuracy was estimated when answers received from DSB were manually checked by a human on a sample of 32 questions asked by participants. When these questions were directed to the composite collection, the chatbot achieved an accuracy of 81.25%. However, when the same questions were sent to product-specific collections, the accuracy rose to 90.63%, which is expected due to the narrower search scope.

Participants praised the received answers for being understandable and clearly formulated. When an answer cannot be found, 11.1% of participants expect to receive only a statement that the answer is not available in the context provided, while 22.2% expect additional information regarding the same topic, and 66.7% expect a reference to a document where an answer may be hidden.

Regarding the performance of DSB, participants were satisfied with how fast it can provide an answer. They stated that DSB saves their time on solving problems, manually searching documents, or asking other employees for certain information. The DSB's UI was intuitive to participants, and they usually did not need help with using it. When asked about drawbacks they encountered when using DSB, one participant stated that the accuracy of DSB is a bit lower when asking questions in a language that differs from the one on which documents are written. Also, two participants appear to receive lower accuracy when asking questions whose answers are distributed across tables, and in the documents provided, some tables are spread across many pages.

To conclude with the DSB application, participants were asked questions about Net Promoter Score (NPS) (Baehre et al., 2022), overall satisfaction with the solution, and their suggestions on how to improve it. With 10 promoters, 7 passives, and 1 detractor, the NPS value was 50, indicating a positive user experience. The reasons behind providing such a score, participants stated DSB to be useful, easy and fast, saving their time, and being accurate enough for daily operations. Also, two service engineers revealed that DSB can be used by new employees to delve into technical documentation, while another participant acknowledged that DSB helps when searching for documents not written in the user's known languages.

To improve their experience with DSB, participants suggested allowing them to insert their documents and thus have personal document collections. Also, they proposed including all documents in DSB that the company has. Additionally, they recommended integrating the chatbot within existing software used in the company to easily search for information related to that software. Participants were mostly satisfied with DSB overall, and they deemed it applicable in the company, both for internal usage by employees and external usage by customers of various company products.

4.3 Threats to Validity

Despite efforts to minimize them, several threats to the validity of evaluation results need to be discussed in this section, as the results may vary when conducting the experiment on different occasions.

The experiment was performed in a single company, including 18 participants in different roles who worked in various company sectors. However, having multiple companies participate in the experiment might provide more accurate evaluation results. Similarly, having a larger group of participants, either from a single company or multiple companies, might provide different results, but booking their time in a company for the experiment could be challenging.

The participants used DSB for at least two weeks during the experiment. However, using the chatbot during a longer period might change the user

experience with it, either positively by experiencing new ways to gather information, or negatively by receiving insights into possible drawbacks of the chatbot.

Regarding the questionnaire, we aimed to formulate precise and unambiguous questions in English, as this was the language known by all participants, but not their native language. Thus, some questions might appear less clear to participants.

DSB offered a reasonable solution for the experiment. However, if the accuracy of the provided chatbot was lower or higher, the user experience might be different, and thus the evaluation results as well. The DSB's accuracy might be influenced by its implementation and technology used, the number and structure of pages in documents, or the number of documents per collection. We aimed to provide participants with a basic RAG-based chatbot that would probably be developed in various companies as the first solution. However, we implemented the chatbot with contemporary technologies available at the time of conducting the experiment, thus offering a solution with decent accuracy.

5 Discussion

Based on the evaluation results presented in this paper, users are aware of AI assistants and notice the potential to apply them in a company, especially in cases of information retrieval, text summarization, and code writing. However, they might not be familiar with the technologies behind AI and how AI assistants can be developed to help them in their daily work. Thus, various demonstrations of AI possibilities and prototypes developed in a company might help users recognize exact use cases in which AI assistants or chatbots, in particular, can be valuable.

As discussed in our previous study (Vještica et al., in press), in which statistical analysis was performed on participants' feedback, the key user satisfaction factor represents the time-saving benefit in information retrieval, compared to manual search. Achieving such a benefit is primarily influenced by a high answer accuracy and secondarily by a low response time.

Therefore, a user-centric chatbot for Q&A should provide answers relatively fast with high accuracy, offering an efficient and reliable solution. The evaluation participants mostly expect a chatbot to have an accuracy of at least 90%. Optionally, such a chatbot should provide users with reasoning and confidence scores, thus additionally increasing their trust in the answers received.

Also, users might find it beneficial that a user-centric chatbot is customizable, allowing them to: (i) upload their own documents; (ii) define how correct answers are structured; (iii) configure fallback responses when an answer is not found; and (iv) choose whether to select a document collection prior to posing a question or instead formulate more specific questions that include the exact topic of interest.

Due to the tendency to pose ambiguous or incomplete questions, i.e., without providing sufficient contextual information, users might not receive answers or receive inaccurate ones, often prompting them to reformulate their questions. Therefore, a chatbot needs to engage in interactive guidance, helping users to articulate precise questions with adequate context.

The presented study provided an in-situ deployment of an RAG-based chatbot within a company, contrasting with prior research that has focused on domain-specific prototypes evaluated primarily through technical benchmarks. Therefore, this work is positioned as complementary to existing research by emphasizing real-world usability and challenges of RAG-based chatbots. Overall, the main contribution of this study is the empirical investigation of chatbots' user acceptance, accuracy trade-offs, and deployment challenges. Based on this investigation, design recommendations are proposed for industrial RAG-based systems.

6 Conclusions and Future Work

In this paper, a user experience evaluation of an RAG-based chatbot was conducted in a medium-sized company. Accordingly, a preliminary case-specific observation of user requirements for a chatbot applied in industrial environments is provided. The chatbot for Q&A named DSB was developed for evaluation purposes and used by company staff who provided valuable feedback. Based on the feedback gathered from evaluation participants and the analysis of the DSB usage, the design recommendations for a user-centric Q&A chatbot were derived and discussed in this paper.

Accordingly, we plan to enhance DSB in a way that provides a better user experience and assess user satisfaction with the novel chatbot solution. The envisioned user-centric version of DSB might be developed as a multi-agent system offering: (i) higher accuracy through improved parsing of documents, especially when extracting text from tables and images; broader context coverage in question answering; and advanced search and reranking capabilities; (ii) enhanced management of vast amount of data and automatic selection of relevant collections based on topics inferred from a question; (iii) integration with various sources, including existing databases; (iv) interactive handling of ambiguous or inaccurate questions provided by users; (v) customizable answer formatting; and (vi) support for personal document collections. The proposed multi-agent system could prolong the time required for generating answers. Thus, a high-performance infrastructure, such as a cloud infrastructure, might be needed to provide answers fast and maintain user satisfaction.

To examine whether the proposed improvements yield measurable benefits, both user experience and the chatbot's accuracy should be evaluated and compared with the basic DSB solution described in this paper. Therefore, we plan to expand the user experience assessment by including a broader participant base, thereby gaining deeper insights into user requirements for an effective Q&A chatbot in industrial contexts. Furthermore, a systematic evaluation of the chatbot's accuracy is needed to complement the user experience evaluation, as the chatbot's accuracy has a high impact on user satisfaction. Frameworks such as RAGAs can be utilized for this purpose, enabling the evaluation of metrics including: (i) the faithfulness of generated answers with respect to the provided context; (ii) the relevance of the answer to the posed question; and (iii) the relevance of the retrieved context to the question's content and scope.

Acknowledgments

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

Amazon Bedrock. (2025). Retrieved March 25, 2025, from https://aws.amazon.com/bedrock/

Baehre, S., O'Dwyer, M., O'Malley, L., & Lee, N. (2022). The use of Net Promoter Score (NPS) to predict sales growth: Insights from an empirical investigation. *Journal of the Academy of Marketing Science*, *50*(1), 67–84. https://doi.org/10.1007/s11747-021-00790-2

Chaudhary, D., Vadlamani, S. L., Thomas, D., Nejati, S., & Sabetzadeh, M. (2024). Developing a Llama-Based Chatbot for CI/CD Question Answering: A Case Study at Ericsson. *Proceedings of the 2024 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 707–718. https://doi.org/10.1109/ICSME58944.2024.00075

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Claude 3.5 Sonnet. (2025). Retrieved March 25, 2025, from https://www.anthropic.com/claude/sonnet

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

Ieva, S., Loconte, D., Loseto, G., Ruta, M., Scioscia, F., Marche, D., & Notarnicola, M. (2024). A Retrieval-Augmented Generation Approach for Data-Driven Energy Infrastructure Digital Twins. *Smart Cities*, *7*(6), 3095–3120. https://doi.org/10.3390/smartcities7060121

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. (2025). ChatCNC: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. *Journal of Manufacturing Systems*, *79*, 504–514. https://doi.org/10.1016/j.jmsy.2025.01.018

Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., & Niforatos, E. (2024). Knowledge sharing in manufacturing using LLM-powered tools: User study and model benchmarking. *Frontiers in Artificial Intelligence*, *7*, 1293084. https://doi.org/10.3389/frai.2024.1293084

Khorasani, M., Abdou, M., & Hernández Fernández, J. (2022). *Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework*. Apress. https://doi.org/10.1007/978-1-4842-8111-6

Liu, X., Erkoyuncu, J. A., Fuh, J. Y. H., Lu, W. F., & Li, B. (2025). Knowledge extraction for additive manufacturing process via named entity recognition with LLMs. *Robotics and Computer-Integrated Manufacturing*, *93*, 102900. https://doi.org/10.1016/j.rcim.2024.102900

Narimani, A., & Klarmann, S. (2024). Integration of Large Language Models for Real-Time Troubleshooting in Industrial Environments based on Retrieval-Augmented Generation (RAG). *Proceedings of the 7th European Conference on Industrial Engineering and Operations Management*, 287–298. https://doi.org/10.46254/EU07.20240085

Pulapaka, S., Godavarthi, S., & Ding, Dr. S. (2024). GenAI-Powered Chatbots. In S. Pulapaka, S. Godavarthi, & S. Ding (Eds.), *Empowering the Public Sector with Generative AI: From Strategy and Design to Real-World Applications* (pp. 157–190). Apress. https://doi.org/10.1007/979-8-8688-0473-1_6

Sarker, S., Arefin, M. S., Kowsher, M., Bhuiyan, T., Dhar, P. K., & Kwon, O.-J. (2023). A Comprehensive Review on Big Data for Industries: Challenges and Opportunities. *IEEE Access*, *11*, 744–769. https://doi.org/10.1109/ACCESS.2022.3232526

Suresh, K., Kackar, N., Schleck, L., & Fanelli, C. (2024). Towards a RAG-based summarization for the Electron Ion Collider. *Journal of Instrumentation*, *19*(07), C07006. https://doi.org/10.1088/1748-0221/19/07/C07006

Taiwo, R., Bello, I. T., Abdulai, S. F., Yussif, A.-M., Salami, B. A., Saka, A., Ben Seghier, M. E. A., & Zayed, T. (2025). Generative artificial intelligence in construction: A Delphi approach, framework, and case study. *Alexandria Engineering Journal*, *116*, 672–698. https://doi.org/10.1016/j.aej.2024.12.079

Unstructured.io. (2025). Retrieved March 25, 2025, from https://unstructured.io/

Vještica, M., Akik, E., Dimitrieski, V., Hinterleitner, L., Erić, J., Weidenfelder, F.-C., & Pisarić, M. (in press). Evaluation of User Experience with RAG-based Chatbots for Searching Documentation: Industrial Case Study. *Proceedings of the 33rd International Conference on Information Systems Development (ISD 2025)*.

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., … Xie, C. (2021). Milvus: A Purpose-Built Vector Data Management System. *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627. https://doi.org/10.1145/3448016.3457550

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 Text Embeddings: A Technical Report* (No. arXiv:2402.05672). arXiv. https://doi.org/10.48550/arXiv.2402.05672

Wei, Q., Sun, H., Xu, Y., Pang, Z., & Gao, F. (2024). Exploring the Application of Large Language Models Based AI Agents in Leakage Detection of Natural Gas Valve Chambers. *Energies*, *17*(22), 5633. https://doi.org/10.3390/en17225633

Wu, C., Ding, W., Jin, Q., Jiang, J., Jiang, R., Xiao, Q., Liao, L., & Li, X. (2025). Retrieval augmented generation-driven information retrieval and question answering in construction management. *Advanced Engineering Informatics*, *65*, 103158. https://doi.org/10.1016/j.aei.2025.103158

Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., & Chen, E. (2024). A survey on large language models for recommendation. *World Wide Web*, *27*(5), 60. https://doi.org/10.1007/s11280-024-01291-2

Xia, Y., Xiao, Z., Jazdi, N., & Weyrich, M. (2024). 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. https://doi.org/10.1109/ACCESS.2024.3415470

Xiao, T., & Xu, P. (2024). Exploring automated energy optimization with unstructured building data: A multi-agent based framework leveraging large language models. *Energy and Buildings*, *322*, 114691. https://doi.org/10.1016/j.enbuild.2024.114691

Yilma, G. M., Ayala-Romero, J. A., Garcia-Saavedra, A., & Costa-Perez, X. (2024). TelecomRAG: Taming Telecom Standards with Retrieval Augmented Generation and LLMs. *ACM SIGCOMM Computer Communication Review*, *54*(3), 18–23. https://doi.org/10.1145/3711992.3711996