SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Information Systems

Investigating the Adoption of Conversational Search by Customer Service Agents

Yaren Mändle

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Master's Thesis in Information Systems

Investigating the Adoption of Conversational Search by Customer Service Agents

Untersuchung der Einführung der konversationellen Suche durch Kundendienstmitarbeiter

Author:	Yaren Mändle
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	M.Sc. Nektarios Machner
Submission Date:	10.02.2025

I confirm that this master's thesis in information systems is my own work and I have documented all sources and material used.

Munich, 10.02.2025 Yaren Mändle

AI Assistant Usage Disclosure

Introduction

Performing work or conducting research at the Chair of Software Engineering for Business Information Systems (sebis) at TUM often entails dynamic and multi-faceted tasks. At sebis, we promote the responsible use of *AI Assistants* in the effective and efficient completion of such work. However, in the spirit of ethical and transparent research, we require all student researchers working with sebis to disclose their usage of such assistants.

For examples of correct and incorrect AI Assistant usage, please refer to the original, unabridged version of this form, located at this link.

Use of *AI Assistants* for Research Purposes

I have used AI Assistant(s) for the purposes of my research as part of this thesis.

Yes ◉ No ○

Explanation:

- Transcribing the voice recordings from the interviews.
- Translating interview transcripts from German to English.
- Translating interview and survey questions from German to English.
- Translating Abstract from English to German.
- Generating and fixing R code.
- Correcting grammar mistakes, refining language.

I confirm in signing below, that I have reported all usage of AI Assistants for my research, and that the report is truthful and complete.

Munich, 10.02.2025 Yaren Mändle

Acknowledgments

I would like to take this opportunity to thank everyone who contributed to the completion of this thesis. Firstly, I want to thank Nektarios Machner and Christoph Drösemeier for their guidance and support throughout the thesis. Additionally, I would like to express my gratitude to Prof. Florian Matthes for giving me the opportunity to conduct my thesis under the Chair of Software Engineering for Business Information Systems.

Next, I want to thank Martin Fürst for providing valuable feedback on my interview questions, as well as all my other team members at the insurance firm for their support. I would also like to thank all the customer service agents who took time out of their day to participate in my interviews and surveys.

Finally, I extend my heartfelt thanks to my dear family, who were always there for me throughout this journey, and my loving husband, who always motivated me to aim for greater achievements.

Abstract

The importance of artificial intelligence (AI)-based conversational agents (CAs) for customer service is increasing rapidly. Companies willing to improve their efficiency are integrating AI-based features into their existing systems. While there are many studies focus on the adoption of customer-facing CAs, there is a lack of research on employee-facing CAs within the specific organizational context. This thesis addresses this research gap by conducting interviews with 12 customer service agents and a survey with 17 agents to identify the key factors in adopting a newly integrated large language model (LLM)-based CA to an existing knowledge management software. We identify scenarios where the CA is preferred over traditional keyword search, factors influencing agents' choice of tool, the strengths and limitations of both search tools, and evaluation metrics that play a key role in user adoption. Our study advances research on the adoption of conversational agents integrated into existing knowledge management software within an organization and offers valuable insights for those adopting CAs in their specific contexts.

Keywords: Artificial Intelligence (AI), Conversational Agent (CA), Customer Service, LLM integration, Knowledge Management, User Adoption

Kurzfassung

Die Bedeutung von auf künstlicher Intelligenz (KI) basierenden Conversational Agents (CAs) für den Kundenservice nimmt rapide zu. Unternehmen, die ihre Effizienz verbessern wollen, integrieren KI-basierte Funktionen in ihre bestehenden Systeme. Während es viele Studien gibt, die sich auf die Einführung von kundenorientierten CAs konzentrieren, gibt es einen Mangel an Forschung über mitarbeiterorientierte CAs im spezifischen Unternehmenskontext. Diese Arbeit schließt diese Forschungslücke, indem sie Interviews mit 12 Kundendienstmitarbeitern und eine Umfrage mit 17 Mitarbeitern durchführt, um die Schlüsselfaktoren bei der Einführung einer neu integrierten, auf einem large language model (LLM) basierenden CA in eine bestehende Wissensmanagement-Software zu ermitteln. Wir identifizieren Szenarien, in denen die CA gegenüber der traditionellen Stichwortsuche bevorzugt wird, Faktoren, die die Wahl des Tools durch die Agenten beeinflussen, die Stärken und Grenzen beider Suchtools und Bewertungsmetriken, die eine Schlüsselrolle bei der Benutzerakzeptanz spielen. Unsere Studie bringt die Forschung über die Akzeptanz von Conversational Agents, die in bestehende Wissensmanagement-Software innerhalb einer Organisation integriert sind, voran und bietet wertvolle Erkenntnisse für diejenigen, die CAs in ihrem spezifischen Kontext einsetzen.

Schlüsselwörter: Künstliche Intelligenz (KI), Conversational Agent (CA), Kundenservice, LLM-Integration, Wissensmanagement, Benutzerakzeptanz

Contents

1. Introduction

In this chapter, first, the motivation behind the thesis will be explained. Next, the research questions and the underlying research methodology will be presented. Lastly, an outline of the overall structure of the thesis will be provided.

1.1. Motivation

The importance of knowledge is increasing day by day for the companies. According to the researchers, "Successful companies are those that create new knowledge, disseminate it widely throughout the organization, and quickly embody it into new technologies and products. This process further fuels innovation and develops lasting competitive advantage" [1]. Knowledge management is the process of using an organization's knowledge to create value. It focuses on effectively sharing knowledge within the organization, with customers and stakeholders [2]. In order to share knowledge in the best way, many companies are using knowledge management tools. Knowledge management tools are the tools that "support the performance of applications, activities or actions such as knowledge generation, knowledge codification, and knowledge transfer" [3].

Artificial Intelligence can be integrated into different areas of knowledge management, such as using AI-related knowledge discovery and data/text mining approaches to determine relationships and trends in the knowledge repositories, helping codify knowledge in knowledge management systems, or using intelligent agents to enhance the search and retrieval methods of knowledge [4]. While digitalization in companies increases with the aim of improving efficiency, complexity also rises. Employees can be supported by providing systems that integrate both human capabilities and smart technologies, leading to value co-creation [5]. Wiethof et al. explain that "as conversational agents still regularly fail to answer complex issues, the concept of Hybrid Intelligence suggests combining artificial with human intelligence in a Hybrid Intelligence System to overcome the weaknesses of CAs and service employees and promote their strengths leading to enhanced performance results and collaborative learning through mutual augmentation" [6].

However, it is also important to know about the willingness of employees to use such systems and the factors influencing their decisions [5]. In their systematic literature review where they did a state-of-the-art analysis of adopting AI-based conversational agents in organizations, Lewandowski et al. [7] concluded that the introduction of Conversational Agents goes beyond a typical software implementation and there are several key factors for success such as ensuring compliance and system transparency, training employees to use the system, and fostering an understanding that CAs will be ongoing self-learning systems.

Lewandowski et al. argue that the success of adoption depends on employees' cooperation and acceptance of their role as continuous knowledge integrators. Training is crucial for adoption, and a holistic approach to collaboration and engagement is necessary for long-term success [7].

A research gap exists in understanding the effects of AI integration into existing knowledge management software for customer service agents within a specific organizational context. Thus, in the form of a case study, customer service agents' adoption of an AI-based conversational search tool into an existing knowledge management software will be investigated in the context of an European insurance firm. While numerous studies in the literature examine the adoption and acceptance of conversational AI and chatbots by customers in the context of customer service, there is a noticeable lack of research on the adoption of these tools by customer service employees within organizations. Thus, the main goal of this study is to examine: How do customer service agents perceive the adoption of an AI-based conversational search as a tool for addressing customer inquiries, particularly in comparison to traditional keyword search methods, and what are their evaluations regarding its effectiveness and usability in improving their workflow?

This research aims to address the gap in understanding the adoption of AI integration within an insurance firm's knowledge management software. Specifically, it will explore how customer service agents respond to this integration and assess its effectiveness. Insights will be offered into how AI adoption affects the process of finding the correct information. By examining when and how AI-based tools are most beneficial, the research will identify scenarios within the context of customer service. The outcomes are expected to offer practical strategies for increasing the adoption of new technology in customer service and guide processes after the integration of AI tools in knowledge management systems, serving as a reference for other organizations adopting similar technologies.

1.2. Research Questions

The following research questions will be addressed throughout the thesis:

RQ1: What factors influence customer service agents' choice between the conversational search and the traditional keyword search?

Different types of questions that customer agents deal with will be identified and categorized by analyzing the customer agents' logs. Using these different question categories, different scenarios will be created. Anonymous and voluntary interviews will be conducted with the customer service agents and during the interviews, for each scenario, they will be given an example customer request and they will be asked which search tool they would prefer and why.

In the context of this case study, the conversational search was integrated alongside the traditional keyword search and it did not intend to replace the traditional keyword search but rather to serve as an additional tool. The qualitative analysis will identify patterns where

one tool is preferred over the other.

Apart from the seven scenarios that were identified, the agents will also be asked an open-ended question about the factors influencing their choice between the two search tools.

Furthermore, at the beginning of the interview, demographic variables will be collected from the interviewees to ensure that the collected data are free of biases such as age or gender bias.

RQ2: How can an LLM-based conversational agent be evaluated?

The existing literature will be analyzed to find different criteria and metrics for evaluating an LLM-based conversational agent. Multiple studies about evaluating conversational agents and various frameworks will be combined, and the measurement instruments that fit our research context will be identified. After gathering the metrics and their corresponding measurement instruments, an online, anonymous, and voluntary survey will be conducted with customer service agents, who will evaluate each metric using Likert scale statements on a scale of 1 to 5. Moreover, correlation analysis will be conducted on the collected data to investigate the relationships between the metrics and other factors further.

In addition to that, customer agents' logs, which include the questions they asked the LLM-based conversational search tool and their ratings of the tool's generated answers, will be analyzed. This approach will allow us to assess whether the survey responses align with the rating agents previously provided for the tool's performance.

RQ3: "What are the benefits and challenges of adopting LLMs in existing knowledge management systems after integration?

After analyzing the results of interviews, surveys, and correlation analysis, key insights will be identified regarding the advantages and limitations of leveraging LLMs in knowledge management systems, as well as factors influencing user adoption.

1.3. Outline

The following is how the rest of the thesis is structured: in Section 2, an overview of knowledge management is given, as well as the usage of artificial intelligence in the context of knowledge management. Moreover, the usage of AI-based conversational agents was discussed, and metrics to evaluate the conversational agents were mentioned. Section 3 gives detailed information about the case study and describes the procedures and processes of both qualitative interviews and quantitative surveys in depth. Section 4 presents and discusses the results of the interviews conducted with customer service agents, as well as the results of the survey where they evaluated the conversational search (CS) based on different metrics. Additionally, the outcomes of different correlation analyses are also included in this section. Lastly, in section 6, areas for future research are outlined, and the research's significant contributions and limitations are summarized.

2. Background

2.1. Knowledge Management Overview

In this section, a detailed overview of knowledge and knowledge management will be given.

2.1.1. Definition of Knowledge

A common definition of knowledge is "justified true belief" [8]. There are three conditions for knowing: the truth condition, the belief condition, and the justification condition [9]. Ayer summarizes these conditions as follows: "The necessary and sufficient conditions for knowing that something is the case are first that what one is said to know be true, secondly that one be sure of it, and thirdly that one should have the right to be sure" [10, 9]. Moreover, according to Bollinger et al. [11], knowledge is an understanding, awareness, or familiarity gained over time by study, investigation, observation, or experience. It is a person's interpretation of information based on their particular experiences, skills, and abilities. The importance of knowledge within the organizational context increases day by day, which makes knowledge a very important factor in the success of an organization [12]. Bollinger et al, argue that knowledge can be found in databases, shared experiences, and best practices, or from other internal and external sources. Organizational knowledge grows throughout time, allowing organizations to achieve deeper degrees of understanding and observation that lead to business astuteness and acumen, both of which are traits of wisdom [11]. In a similar way, Davenport et al. also define it as "a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms" [13].

2.1.2. Types of Knowledge

Knowledge can be classified into two types: explicit knowledge and tacit knowledge.

Explicit Knowledge

Explicit knowledge refers to knowledge that is clearly defined without ambiguity, easily documented, and straightforward to communicate. It typically appears in forms such as manuals, databases, and policy and procedure documents [14, 15]. Furthermore, it is easier to transfer within the organization or between individuals without losing the meaning since it is highly codifiable [16].

Tacit Knowledge

On the other hand, tacit knowledge is hard to express, informal, uncodified, and thus not readily transferable. The reason for that is tacit knowledge is based on individuals' or organizations' experiences so it resides in individuals' minds [16]. Although it is challenging, it is important for organizations to be able to transfer tacit knowledge to support the knowledge management process [16]. A way to transfer this knowledge is "personal communications through discussion and demonstrations" [17].

2.1.3. Knowledge Management Definitions

The ability of companies and individuals within them to share knowledge with one another, particularly organizational knowledge, has been highlighted as a contributing element to organizational competitiveness [14]. As a result, the concept of Knowledge Management is currently of special interest to organizations concerned about their employees' learning and competitiveness, because effective knowledge management can assist organizations in increasing member collaboration and encouraging knowledge sharing among them [18]. Organizations must thus identify knowledge as a valuable resource and build a system for tapping into the collective intelligence and talents of employees in order to generate a larger organizational knowledge base [11]. To implement and fully utilize knowledge management, businesses must have a clear understanding of how knowledge is produced, communicated, and applied inside the organization [12, 19, 20].

Organizations have been effectively motivated to adopt knowledge management practices [12]. In literature, there are different definitions for knowledge management. Lytras et al. define it as "the cumulative ability to utilize the value incorporated in the various stakeholders of an organization. Knowledge management is the integration of knowledge assets in reusable formats, that sets a win-win relation for all the parts of the knowledge Web" [21]. O'Dell et al. suggest that knowledge management means making sure that knowledge workers are effectively connected to the right people or resources at the right time to support well-informed decision-making [22]. In this thesis, we adopt the definition of knowledge management provided by O'Dell et al as it aligns with our focus on evaluating how LLM-based conversational search can enhance knowledge accessibility and usability within knowledge management systems. Furthermore, it can also be defined as a practice that includes the processes of creating, gathering, organizing, diffusing, using, exploiting, transferring, and storing an organization's vital knowledge [23]. Moreover, two different definitions are provided by Liebowitz et al. [4]: "knowledge management is the systematic, explicit, and deliberate building, renewal, and application of knowledge to maximize an enterprise's knowledge-related effectiveness and returns from its knowledge assets." and " knowledge management is the formalization of and access to experience, knowledge, and expertise that create new capabilities, enable superior performance, encourage innovation, and enhance customer value.".

It is important that the organization are aware of the importance of the knowledge management activities. In order to measure the outcomes and benefits of knowledge management,

Chong et al. [23] conducted a comprehensive literature review examining the knowledge management performance outcomes. After analyzing previous research, 38 performance outcomes from knowledge management efforts were found. Some of the identified performance outcomes are as follows: "Better decision making, better customer handling through better client interaction and sharing knowledge with clients, faster response to key business issues, immediate results in solving organizational-wide problems, development and constant improvement of competitive long-range service and technology strategies". [23].

2.1.4. The Knowledge Management Lifecycle

Many theories, models, and frameworks have been produced over time to help guide knowledge management research and practice [24]. In order to prevent confusion about which framework to use for research and practice, Shongwe [24] examined 20 important knowledge management lifecycle frameworks and suggested a unified framework, which can be seen in Table 2.1.

The KM process:	Knowledge transfer	Knowledge storage	Knowledge application	Knowledge creation	Acquisition

Table 2.1.: Processes of the Proposed Unified Framework [24].

Knowledge Transfer

Knowledge transfer is the process of transferring a specific ability from a source to a user. Its effectiveness is directly proportional to the degree to which the user possesses the ability to transmit from the source [25]. The transfer process depends not only on the user's cognitive qualities, which drive his or her interpretation but also on how knowledge is communicated from the source to the user [25]. Knowledge can be transferred at various levels: "transfer of knowledge between individuals, from individuals to explicit sources, from individuals to groups, between groups, across groups, and from the group to the organization" [26].

Knowledge Storage

Next comes knowledge storage. Transferred knowledge should be maintained in a repository for easy retrieval by other members of the organization, without requiring contact with the original source [27]. The term organizational memory is used as a synonym for knowledge storage by many researchers. Organizational memory encompasses knowledge in written documentation, electronic databases, expert systems, documented procedures and processes, and tacit knowledge held by individuals and networks [26, 28]

Knowledge Application

Previous research indicates that applying knowledge is crucial for developing new products, and fostering creativity and performance [29]. Knowledge application refers to how organiza-

tions employ internal knowledge to improve operations, develop new products, and create new knowledge assets [30]. Song et al. describe knowledge application as "an organization's timely response to technological change by utilizing the knowledge and technology generated into new products and processes" [31]. Knowledge application involves decision-making protection, action, and problem-solving which according to Allameh et al., ultimately lead to knowledge creation [32].

Knowledge Creation

A learning firm is one that continuously integrates, communicates, and creates knowledge. The continuous development of organizational knowledge is a dynamic capability that results in continual organizational learning and the expansion of knowledge assets [33]. According to van Aalst [34], knowledge creation is not only about creating a new idea, it also "requires discourse (talk, writing, and other actions) to determine the limits of knowledge in the community, set goals, investigate problems, promote the impact of new ideas, and evaluate whether the state of knowledge in the community is advancing". Nonaka suggests that there are four modes of knowledge creation: socialization, combination, externalization, and internalization [35]. Socialization is the process of creating tacit knowledge through social interactions and shared experiences. Combination is the process of creating explicit knowledge from explicit knowledge through re-configuring of existing information. Externalization is the process of converting tacit knowledge to explicit knowledge and lastly, internalization is the conversion of explicit knowledge to tacit knowledge.

Knowledge Acquisition

According to Hu, knowledge acquisition is a fundamental process in knowledge management and it contains the tasks: "elicitation, collection, analysis, modeling, and validation of knowledge" [36]. Huber defines the knowledge acquisition process as the process by which knowledge is obtained [37]. There are several ways to obtain the knowledge such as "customer surveys, research and development activities, performance reviews, and analyses of competitor's products" [37].

Some other frequent processes, which were a part of the examined frameworks by Shongwe [24] but not included in the unified framework were: organizing, identifying, learning, and analyzing.

2.2. Artificial Intelligence in the Context of Knowledge Management

Because of its transformational potential, artificial intelligence has received a lot of attention and is being broadly used in a variety of industries [38]. Table 2.2 shows various definitions of artificial intelligence.

AI Definition	Source
AI is the technology to "make something that has been comparatively expensive abundant and cheap. The task that AI makes abundant and inexpensive is prediction — in other words, the ability to take information you have and generate information you didn't previously have."	[39]
AI is the "behavior of a machine which, if a human behaves in the same way, is considered intelligent."	[40]
"Artificial intelligence systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions."	[41]
"AI refers to the capability of a computer system to show human-like intelligent behavior characterized by certain core competencies, including perception, understanding, action, and learning. In line with this, our understanding of an AI application refers to the integration of AI technology into a computer application field with human–computer interaction and data interaction."	[42]
"AI refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech, and face recognition systems), or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones or Internet of Things applications)."	[43]

Table 2.2.: Definitions of AI.

However, Mikalef and Gupta [44] argue that AI technologies are unlikely to provide significant competitive advantages on their own and thus, they are not sufficient to develop an AI capability, which "is the ability of a firm to select, orchestrate, and leverage its AI-specific resources". They further suggest that to differentiate from competitors, firms need a unique combination of physical, human, and organizational resources to develop AI capabilities that add value. In their findings, they identified that there are eight types of complementary resources, that contribute to the development of an overall AI capability. First, there are tangible resources that consist of data, technology, and basic resources. Next, there are human resources that consist of technical skills and business skills. Lastly, there are intangible resources that include inter-department coordination, organizational change capacity, and risk proclivity [44]. In the same vein, Olan et al.'s findings also suggest that simply installing AI technologies is insufficient to improve organizational effectiveness [45]. They argue that a more sustainable strategy is to combine AI with knowledge sharing, such as incorporating lessons learned from completed projects, to improve performance and efficiency in a rapidly changing digital society.

Recent literature emphasizes the evolving role of AI in augmenting knowledge management systems. AI technologies play a crucial role in enhancing organizational knowledge, leading to improved performance and competitive advantage, and helping employees manage complicated knowledge that they might otherwise struggle to integrate into company operations [45]. As mentioned before, the key to knowledge management is making sure that knowledge

workers are effectively connected to the right people or resources at the right time to support well-informed decision-making [22]. The integration of AI holds promises in further enhancing this process [46]. However, researchers also agree that the future of AI in knowledge work must prioritize collaborative techniques where humans and AI collaborate closely, rather than complete automation [47]. By using AI-enabled technologies such as AI-based search features, AI has the potential to transform traditional knowledge management techniques. Table 2.3, adapted by Jarrahi et al. [46], shows an overview of the knowledge management processes and the possibilities to use AI systems in a particular process.

The KM process	**Possibilities created using AI systems**
Knowledge creation	Predictive analytics via self-learning analytical capacities Recognizing previously unknown patterns
Knowledge storing and retrieving	Harvesting, classifying, organizing, storing, and retrieving explicit knowledge
Knowledge sharing	Connecting people working on the same issues Facilitating collaborative intelligence and shared organizational memory
Knowledge application	Enhancing situated knowledge application by searching and preparing knowledge sources Offering more natural and intuitive system interfaces

Table 2.3.: Potential AI Usage in Various KM Processes [46].

2.3. Usage of AI-based Conversational Agents

Conversational Agents fall within the category of knowledge application and in recent years, it has developed rapidly due to improvements in AI and machine learning [48]. According to Von Wolff et al., conversational agents can be defined as "an application system that provides a natural language user interface for the human-computer-integration. It usually uses artificial intelligence and integrates multiple (enterprise) data sources (like databases or applications) to automate tasks or assist users in their (work) activities". CAs are used in a variety of industries such as healthcare, education, entertainment, and customer service [49]. For example, in the healthcare sector, Bin Sawad et al. [50], in their systematic review on CAs, reported that in more than half of the reviewed studies, users gave positive feedback on the helpfulness, satisfaction, and ease of use of CAs within the healthcare context. Moreover, in the context of education, Wambsganss et al. [51] conducted a study to determine how CAs can improve student engagement and response quality in course evaluations compared to traditional web surveys and they found out that using conversational agents resulted in higher response quality and enjoyment levels. Furthermore, in the context of customer service, the results of Wang et al. showed that using CAs in both regular as well as innovative ways have a positive influence on a company's internal and external agility, and this increased agility leads to improved customer service performance [52].

The ability of CAs to save time and reduce costs makes them highly attractive to many companies. Xu et al. [53] define AI in the context of customer service as "a technology-enabled system for evaluating real-time service scenarios using data collected from digital and/or physical sources in order to provide personalized recommendations, alternatives, and solutions to customers' inquiries or problems, even very complex ones". Some studies predict that CAs have the ability to lower global business costs by 30%, which presently amount to $1.3 trillion per year due to 265 billion customer support questions [54, 55]. This reduction can be achieved by shortening response times, freeing up agents for other tasks, and handling up to 80% of routine questions [54, 55]. However, Poser et al. argue that even though CAs are efficient at processing simple, routine customer inquiries, they are unable to handle complex issues, which can lead to service failures and customer dissatisfaction [56]. In a similar vein, Xu et al.'s results indicated that for low-complexity tasks, consumers perceived AI as more effective at solving problems than human customer service and were more inclined to use it. However, for high-complexity tasks, they regarded human customer service as superior and were more likely to prefer it over AI [53]. Moreover, it was stated by research that more than 80% of consumers still prefer human interaction over CAs. Respondents indicated that human agents are significantly better at addressing various queries, particularly when it comes to understanding complex situations, as they believe humans offer better comprehension [57].

To prevent this trade-off between service efficiency and quality, some companies adopted a combined approach where humans handle personal customer interactions while AI assists with problem-solving [58]. Therefore, rather than replacing humans with AI-based agents, a collaborative approach where humans and AI agents work together will be necessary to deliver improved services [59]. This collaborative approach is known as the Hybrid Intelligence System (HIS) [60]. HIS is defined by Dellermann et al. [60] as "the ability to achieve complex goals by combining human and artificial intelligence, thereby reaching superior results to those each of them could have accomplished separately, and continuously improve by learning from each other". According to Dellermann et al., human intelligence has strengths such as flexibility, empathy, creativity, and common sense. On the other hand, machine intelligence has strengths such as pattern recognition, probabilistics, consistency, speed, and efficiency. By leveraging these strengths together, collaborative efforts can lead to very high performance in particular tasks [61]. Wiethof et al. [6] report that the "initial research has shown that CAs have a positive impact on employees' performance across various digital workplaces allowing intuitive dyadic, dialog-based interaction to receive output from information systems and to provide input and commands as well as feedback to improve the AI."

Several studies investigate whether a hybrid approach, combining AI with human expertise in customer service, can assist newly hired agents with limited experience, reduce their time-to-performance, and decrease the turnover rate, which can be up to 70% [58, 62, 63]. In their paper, Reinhard et al. [58] explore how generative artificial intelligence (GenAI) collaborates with newly hired frontline service employees (FSEs) during their onboarding process in customer service roles. The authors propose a hybrid intelligence system where generative AI acts as a "co-agent" to support novice employees by assisting them in solving

customer queries in real time. Results showed that although interaction and time investment increased, service employees reported a lower perceived workload when working with a GenAI-based collaborator. The hybrid intelligence approach allowed service employees to focus on complex, human-centric problems while the AI assisted with routine or data-heavy tasks.

The study by Gao et al. [64] examined whether AI chatbot suggestions help people answer knowledge-based questions more quickly and easily. Researchers conducted a large experiment with different system versions, combining humans with chatbots of varying quality. It was found that AI chatbot suggestions can make answering questions faster and easier, even if the chatbot quality is low, reducing response time by up to 35% and keystrokes by up to 60%. However, low-quality chatbot suggestions sometimes led to worse answers, as users didn't always ignore poor suggestions. This resulted in responses that were lower in quality than those a human would produce without any chatbot assistance. Thus, while using hybrid systems, it is important to use a high-performing chatbot in order not to experience a quality-efficiency trade-off.

2.4. Conversational Search Evaluation Criteria

Evaluating the performance of CAs is crucial for ensuring their effectiveness and reliability across various applications. Chatbot evaluation metrics provide a systematic way to assess key aspects such as accuracy, relevance, and user satisfaction. These metrics, which can be both quantitative and qualitative, help developers understand how well a chatbot achieves its intended goals.

A comprehensive chatbot evaluation framework was developed by Peras [65] by systematically unifying various metrics. This framework is organized into five key perspectives, each containing multiple categories:

- **User Experience Perspective:** usability, performance, affect, and satisfaction
- **Information Retrieval Perspective:** accuracy, accessibility, and efficiency
- **Linguistic Perspective:** quality, quantity, relevance, manner, and grammatical accuracy
- **Technology Perspective:** humanity
- **Business Perspective:** business value

These perspectives provide a holistic approach to evaluating chatbot performance. Moreover, each category consists of several attributes such as task completion, support for a minimal set of commands, correctness and categorization of the responses, relevancy of responses to the context of the conversation as well as efficiency and qualitative cost.

Kuligowska et al. [66] also provide several measurement metrics to evaluate the performance, usability as well as overall quality of a CA. Their main focus is on commercially used CAs in the B2C sector and the proposed metrics are as follows: "visual look, the form

of implementation on the website, speech synthesis unit, built-in knowledge base (with general and specialized information), presentation of knowledge and additional functionalities, conversational abilities and context sensitiveness, personality traits, personalization options, emergency responses in unexpected situations, the possibility of rating chatbot and the website by the user".

In order to evaluate the quality of chatbots and intelligent conversational agents, Radziwill et al. [67] extract quality attributes from existing papers in the literature. They assign these quality attributes under three main measurements: efficiency, effectiveness, and satisfaction.

Efficiency includes the subcategory performance, with quality attributes such as graceful degradation, robustness to manipulation, and robustness to unexpected input [67].

Effectiveness is divided into two subcategories: functionality and humanity. Accurate speech synthesis, interpreting commands accurately, linguistic accuracy of outputs, and general ease of use are some quality attributes assigned to the functionality category. The humanity category includes attributes like disclosing the chatbot identity, convincing, satisfying, & natural interaction, being able to maintain themed discussion [67].

Lastly, satisfaction has three subcategories: affect, ethics & behavior, and accessibility. Quality attributes for affect include providing greetings, conveying personality, and providing emotional information through tone, inflection, and expressivity. For ethics & behavior, attributes include respect, inclusion, and preservation of dignity, ethics and cultural knowledge of users, and protecting and respecting privacy. Under accessibility, attributes include responding to social cues or lack thereof and detecting meaning or intent [67].

A study conducted by Ashfaq et al. [57] investigated drivers of users' satisfaction and continuance intention toward chatbot-based customer service. The analytical framework they proposed combines the expectation-confirmation model, information system success model, technology acceptance model (TAM), and the need for interaction with a service employee. They identified seven main constructs, as shown in Figure 2.1. They found out that information quality and service quality, perceived enjoyment, and perceived usefulness positively influence consumers' satisfaction. Additionally, perceived enjoyment, perceived usefulness, and perceived ease of use were found to support continuance intention.

Furthermore, two automated evaluation frameworks could be found in the literature to evaluate retrieval-augmented generation systems [68, 69]. Both frameworks RAGAS and ARES don't rely on human annotations and thus focus on evaluation metrics that are fully self-contained and reference-free. The focused metrics are: Faithfulness, Answer Relevance, and Context Relevance [68, 69].

Answer Faithfulness refers to the concept that the answer should be based on the provided context, which helps to avoid fabrications. An answer is considered faithful to the context "if claims that are made in the answer can be inferred from the context" [68]. Answer Relevance refers to the concept that "the generated answer should directly address the provided actual question [68]. Lastly, Context Relevance refers to the importance of retrieving focused and concise context to answer a question, minimizing irrelevant or redundant information [68]. Longer, unfocused passages can hinder the effectiveness of LLMs, especially when important details are buried in the middle. This metric aims to ensure that the context provided contains

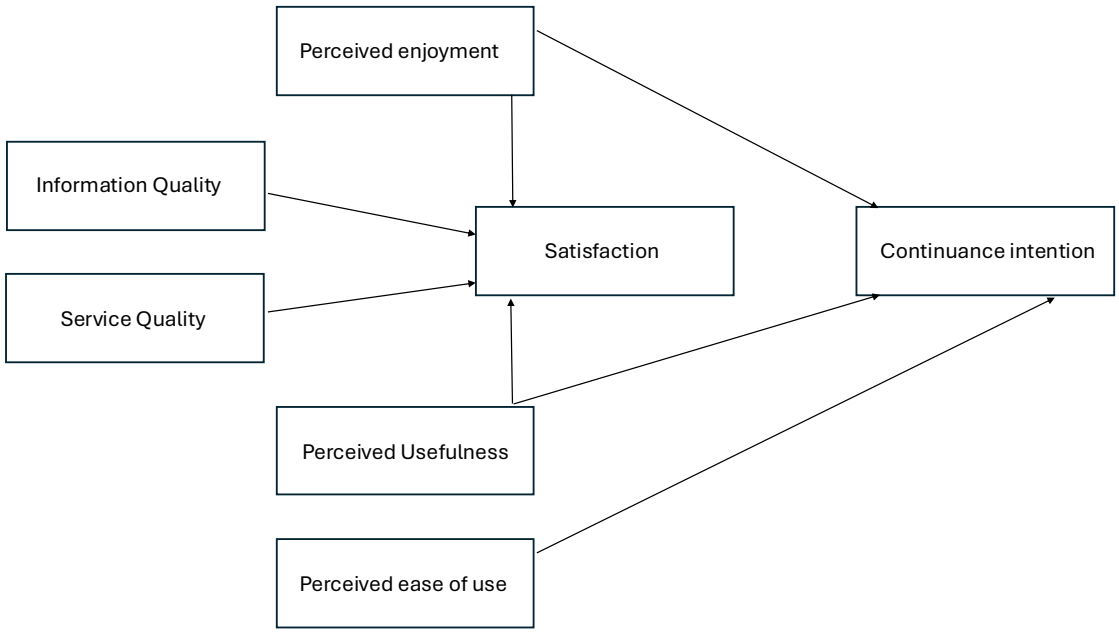

Figure 2.1.: Conceptual Framework Adapted by [57].

only the necessary information to address the question, avoiding unnecessary content that could reduce the model's performance [68].

In addition to the chatbot evaluation criteria discussed, understanding the context of user acceptance also adds valuable insights, as both aspects are closely intertwined. Many of the factors that influence chatbot evaluation, such as usability, perceived usefulness, and ease of use, are also foundational elements in frameworks on user acceptance. A study conducted by De Cicco et al. [70] identified seven constructs to use in their study: perceived ease of use, perceived usefulness, compatibility, trust, perceived enjoyment, attitude, and intention to use. The study used the Technology Acceptance Model as a base that was developed by Davis et al. [71], who stated that users' intention to use technology is driven by their perceived usefulness of the technology and perceived ease of use.

In a similar vein, also building on the TAM, Brachten et al. [5] use 11 scales for their studies: attitude towards using, perceived usefulness, perceived ease-of-use, trust, perceived behavioral control, facilitating conditions, efficacy, subjective norm, superior influence, peer influence and intention to use enterprise bots.

In 2003, Venkatesh et al, [72] evaluated the TAM together with 7 other models and proposed a unified model, Unified Theory of Acceptance and Use of Technology (UTAUT), that integrates elements across the eight models. The UTAUT model identifies four main constructs: performance expectancy, effort expectancy, social influence, and facilitating conditions. Performance expectancy refers to an individual's belief that using a system would lead to better job performance. The second construct, effort expectancy, is defined as the degree of ease with which the system is used. Next, social influence is described as an

individual's perception that other important individuals believe he or she should use the new system. Lastly, facilitating conditions are defined as the extent to which users believe there are resources and support available to use the technology effectively [72].

Despite the usefulness of these models, some researchers argue that traditional frameworks like TAM and UTAUT have various shortcomings that may make them unsuitable for use in determining the acceptability of AI technologies [73]. These models focus on predicting the acceptance of non-intelligent technologies and place an undue emphasis on the user's effort (e.g., ease-of-use) necessary for operating these technologies [73]. Additionally, traditional theories typically assess technology performance based on repetitive and utilitarian tasks. However, AI service robots are designed for social interaction, meaning their performance is evaluated through both cognitive (task efficiency) and emotional (user connection and engagement) lenses [73].

In 2019 Gursoy et al. [74] developed a theoretical model of artificially intelligent device use acceptance (AIDUA) that has the goal of explaining customers' willingness to accept AI device use in service encounters. This framework focuses specifically on AI devices and thus differentiates from TAM and UTAUT frameworks. The model consists of three acceptance generation stages and six antecedents. In the Primary Appraisal Stage, customers evaluate the relevance and importance of AI devices in service settings. Three factors are identified as critical constructs that play a role in influencing consumers' primary evaluation of AI devices: social influence, hedonic motivation, and anthropomorphism. During the Secondary Appraisal Stage customers perform a deliberate evaluation of the AI devices' performance and effort expectancy. Thus, the identified constructs are performance expectancy and effort expectancy. In the final Outcome Stage, the customer's willingness to accept the use of AI devices and/or objection to the use of AI devices is determined.

2.5. Related Work

This section reviews the relevant literature across two key areas: comparing AI-based conversational search with the traditional keyword search, and evaluating the conversational search using different metrics.

- Sakirin et al. conducted a study to identify the user preferences for ChatGPT-powered conversational interfaces versus traditional keyword search methods. They collected data from college students using a survey questionnaire. The collected data was analyzed using descriptive statistics and inferential analysis. They found that the majority of participants preferred ChatGPT-powered conversational interfaces over traditional keyword search methods, due to factors such as convenience and efficiency. However, this study differs from ours since it focused on a different type of conversational AI i.e. ChatGPT and it wasn't in an organizational context.

- Preininger et al. [75] researched information accessed through conversational agents versus traditional keyword search methods in a widely used pharmacologic knowledge base. Conversational agent use was associated with higher frequencies of access to some

specific topics than searching using traditional keyword search methods and vice versa. However, the limitation of the study is that they did not determine the reasons for the selection of each search method or the usability and user satisfaction associated with each method.

- Liu et al. [76] designed a study to gather users' interaction behaviors and explicit feedback signals while engaging with both traditional and agent-mediated conversational legal case retrieval systems. They investigate the differences in search interaction behavior and in user search outcomes between traditional and conversational legal case retrieval. They found that better retrieval performance was achieved by the users with the conversational case retrieval system. However, this study is different than ours as it is in the legal context.

- Wazzan et al. [77] compared the traditional keyword search and the LLM-based search for image geolocation tasks, where users determined the location where an image was captured. They focused on how the usage of each tool impacts participants' performance in geolocation tasks and how participants adapt their query formulation strategies. The results showed that participants using traditional keyword search were more effective in accurately predicting the location of the image than LLM-based search. This study is distinct from ours as ours compares the traditional keyword search and the LLM-based search in an organizational context and evaluates the LLM-based search using different evaluation metrics.

- Spatharioti et al. [78] conducted two online experiments and assigned participants to complete a consumer product research task using either a traditional keyword search tool or an LLM-based search tool. The experiments aimed to investigate four key research questions: (1) Efficiency – comparing task completion time and query complexity between LLM and traditional keyword search conditions, (2) Accuracy – examining differences in decision accuracy, (3) Perceptions – assessing user experience and perceived reliability, and (4) Confidence & Errors – exploring how participants handle LLM mistakes with or without confidence cues. Their results showed that participants who used the LLM-based tool were faster in completing their tasks by using fewer but more complex queries. However, they also discovered that the participants overly relied on incorrect information generated by LLM. This study takes a different approach from ours by conducting experiments and not surveys or interviews.

- Ling et al. [79] and Lewandoski et al. [7] both conducted systematic literature reviews to analyze factors influencing the adoption of conversational agents. While Ling et al., examined user-related, agent-related, and attitude-based factors, proposing a model that explains how consumers interact with and accept CAs in business and marketing settings, Lewandovski et al. identified organizational, technical, and environmental factors affecting adoption and explored the strategic and governance aspects of managing CAs in work environments. A key difference between their studies and ours is that while both papers are systematic literature reviews analyzing existing research

on the adoption of conversational agents, our study is empirical, based on surveys and interviews. Moreover, the aspect of comparing traditional keyword search with LLM-based search lack in their research context.

3. Methodology

3.1. Case Study

This study is conducted as a case study within an European insurance firm. In 2008, the company launched its in-house developed knowledge management software. This software comprises 168,000 pages across 5,600 documents and has approximately 40,000 active users, with 12,500 daily logins and 9.6 million annual logins. The software serves as the central help system of the company and supports employees across all divisions in case processing. The application includes business process-related documents for work support as well as system-related information. It is mainly used by customer service agents to answer customer inquiries.

The application currently uses a full-text-based search. However, in 2021, the idea of implementing and integrating an AI-based conversational search was formulated for the first time. This conversational search aimed to provide an intelligent and optimized way to search for relevant information and work instructions. Its implementation was completed in 2023. The motivation behind this idea was that customer service agents were spending too much time and effort searching inefficiently for information and work instructions using the full-text search, especially when handling customer inquiries simultaneously. Newly hired customer service agents required even more effort to find relevant information. This led to inefficiencies in customer service, dissatisfaction and frustration among both customer service agents and customers during periods of high workload, and significant challenges for newly hired customer service agents.

There are three personas: newly hired agents, agents with a couple of years of experience, and expert agents. It has been observed that since expert agents have many years of experience, they know the majority of answers by heart and don't often use the search features. Thus, the target group for this application is newly hired agents and those with a couple of years of experience.

With the new AI-based conversational search functionality, it is expected that customer satisfaction will increase as they will receive answers to their questions more quickly, employee satisfaction will rise as they will use less effort while searching, and the company will save on operational costs since the new solution reduces processing time.

The newly integrated conversational search is not a substitute for the full-text-based search but serves as an additional tool. Thus, users can freely choose which search tool they want to use. Figure 3.1 shows the process goal of the new conversational agent. There are multiple authentication groups within the application. An agent can only access documents within their authentication group. When the agent types a question, only the documents corresponding to the agent's authentication group are considered. With the help of a retriever,

documents are searched and the data volume is reduced. Then, the ranker evaluates and organizes the information selected by the retriever in relevance to the query. Next, a GPT-4-based large language model generates an AI-based response based on the initial query and the evaluated information. The agent now has the possibility to evaluate the answer by giving a rating from 1-5. In addition to generating an LLM-based answer, the conversational search also provides three documents where the answer can be found. Agents can give either a thumbs up or a thumbs down to these document suggestions. The generated response is derived from these documents.

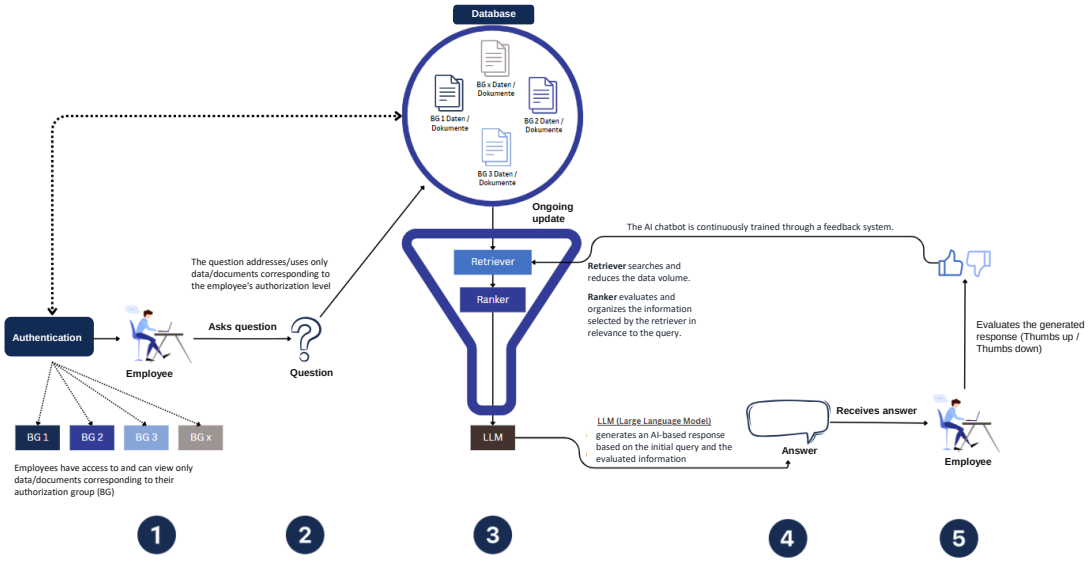

Figure 3.1.: Process Goal.

3.1.1. IT-Architecture of Conversational Search

Figure 3.2 presents the architecture of the Enterprise Knowledge Assistant (EKA) system, outlining the components and workflows that allow customer service agents to interact with a knowledge base enhanced by AI models and expert feedback loops. A centralized authentication service, KeyCloack, is used to manage users and customer service agents log in via this system to ensure secure access. When the customer service agent submits a question through the EKA Application, the application uses a set of AI models in order to process the query. The EKA interacts with the Customer Service Business Application to retrieve data from the source database. Retriever model searches and retrieves relevant information from the knowledge base (vector database). The architecture also includes an intermediate data storage layer, where data from source systems is pre-processed and embedded into the vector database for efficient retrieval. The ranker model ranks the retrieved results based on

relevance and an LLM refines or generates responses based on the retrieved information. To improve the EKA application, a feedback system is designed that involves first-line feedback directly from agents as well as second-line feedback provided by experts. The first-line feedback from agents is stored in a feedback database, where it is reviewed by experts. Based on the experts' evaluation of this feedback, a fine-tuning dataset is created, which is used to continuously refine the models (Retriever, Ranker, and LLM). This iterative fine-tuning process helps to achieve a high level of accuracy and reliability in the responses generated by the EKA system.

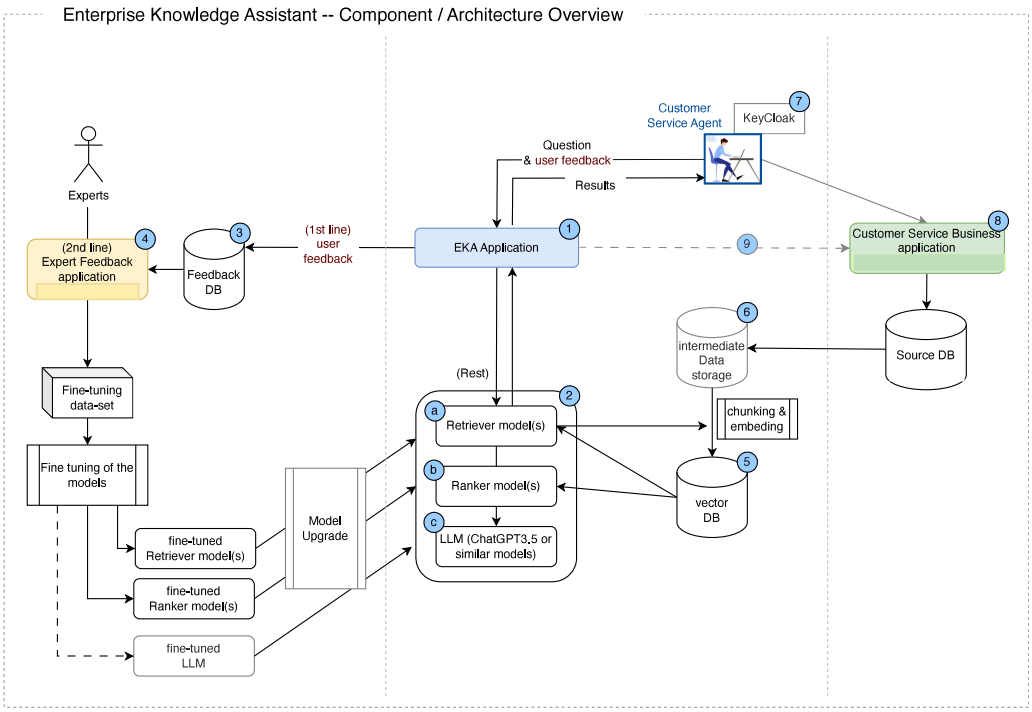

Figure 3.2.: Enterprise Knowledge Assistant Architecture.

3.1.2. IT-Architecture of Full-Text-Based Search

The current full-text search has no technical similarities with the AI-based search. It is purely a text-based search without any AI components. Prioritization of search results is done through weighting, considering matches in the title, heading, text, and document type. Documents are stored in JSON format with a unique ID, mandatory fields, and metadata. Queries are processed in Python without using an LLM. In order to authenticate the users, there is an authentication process, which is based on users' email addresses and global IDs to determine which documents a user is allowed to see. The search not only checks for the presence of a word but also evaluates its frequency, where overly frequent occurrences reduce its weight. A tokenized query is matched with documents and then passed to an algorithm. Additionally,

there is an auto-complete function that suggests possible next words while typing. While full-text search primarily focuses on content, metadata searches allow filtering.

3.2. Interview Design and Implementation

In order to understand when customer service agents prefer conversational search over keyword search, and vice versa, as well as to evaluate conversational search, semi-structured interviews were conducted with customer service agents. Three team leaders were contacted from three different locations, and agents were invited to the interviews by their team leads. Participation in the interviews was voluntary and anonymous. 13 agents participated in the interviews, and the mean time for the interviews was 16.13 minutes. Interviews were conducted online via a web conferencing application. With the consent of the participating agents, all interviews were voice-recorded and the voice recordings were transcribed using a transcription AI. Afterward, the transcriptions were refined and translated from German to English using a GPT-based AI.

The interviews consisted of four main parts. The first part was the introduction and an explanation of the purpose of the interview. In the second part, agents were asked about their individual characteristics, such as how many years they had been working in the company, and whether they had any chatbot experience before the integration of conversational search in the company, and if yes, how often. As stated in Ashfaq et al's study, control variables in the study were used to ensure that the results from the empirical analysis were not influenced by variance in demographic variables, as previous academic literature suggests these variables could impact empirical outcomes in technological environments [57]. Table 3.1 shows an overview of the individual characteristics of the respondents. Individual characteristics data were collected from the agents who participated in the interviews, and not from the agents who only took part in the surveys. There were more female (8; 61.54%) respondents than male (5; 38.46%), the mean age was 29.62 (SD = 12.05; minimum = 19; maximum 52), the mean of years that respondents are working in the company was 14.19 (SD = 12.34; minimum = 1.5; maximum 37). Moreover, the majority (10; 76.92%) of the respondents stated that they had a chatbot experience before the integration of the CS into the company, compared to the minority (3; 23.08%). The respondents who stated that they had chatbot experience were asked about their frequency of chatbot usage. The majority (5; 38.46%) stated that they use it 1-2 times per year, while the second largest group (3; 23.08%) said they use it 1-2 times per week. Lastly, respondents were asked how long they have been using the CS in the company. Five respondents (38.46%) indicated they had used it for less than 1 month, and another five (38.46%) said they had used it for 2-3 months.

Characteristics	Distribution	Frequency	%	Mean	SD
Gender	Female	8	61.54%		
	Male	5	38.46%		
Age	20-25	8	57.14%	29.62	12.05
	26-35	2	14.29%		
	36-45	2	14.29%		
	46-55	2	14.29%		
Work Experience in the Company	1–5 years	5	38.46%	14.19	12.34
	6-10 years	4	30.77%		
	11-20 years	0	0.00%		
	21-30 years	2	15.38%		
	31+ years	2	15.38%		
Chatbot/CS experience?	No	3	23.08%		
	Yes	10	76.92%		
Frequency	Daily	1	7.69%		
	1-2 times per week	3	23.08%		
	1-2 times per month	1	7.69%		
	1-2 times per year	5	38.46%		
CS Usage Period	Less than 1 month	5	38.46%	1.66	1.1
	1-2 months	2	15.38%		
	2-3 months	5	38.46%		
	More than 3 months	1	7.69%		

Table 3.1.: Profile of the Respondents.

In the third part, agents were asked about seven different scenarios. The goal was to understand which search tool they preferred in which scenario. Scenarios were categorized as simple, complex, open-ended, close-ended, short, long, and procedural queries. Simple queries were the queries that required manual lookup in a single document of the knowledge management software, whereas for complex queries agents needed to look at multiple documents to find the answer. Furthermore, open-ended queries were defined as questions that required detailed and extensive answers while for a close-ended question, the required answer was a yes or no. The next metric was the length of the question. A short query was defined as shorter than 8-10 words and a long query was defined as more than 10 words. Lastly, procedural queries were defined as queries that required a step-by-step process. Table 3.2 shows the explanation and the example customer question for each scenario. Different scenarios were created based on the customer agent's logs, along with input from the product owner of the knowledge management tool. The example questions are taken from real queries asked by customer agents, which were saved in the database. After explaining each scenario and reading the example question, agents were asked which search tool they would use for that specific scenario and to explain their reasoning.

Simple Query
Requires manual lookup in a single document of the knowledge base by the agent. Example Question: Are e-scooters insurable under private insurance?
Complex Query
Requires more intensive research, such as multiple documents or entries in the knowledge base. Example Question: How do I insure a minor policyholder?
Open-ended Query
Requires more detailed and extensive answers. Example Question: What should I consider as a buyer or seller during a change of ownership?
Close-ended Query
Yes/no questions. Example Question: Is Addison's disease a master illness?
Short Query
Less than 8-10 words. Example Question: How long is the immediate coverage valid?
Long Query
More than 10 words. Example Question: Are damages caused by my pet, such as bite injuries or property damage, covered under liability insurance?
Procedural Query
Requires guidance or a description of how to perform a specific task step-by-step. Example Question: How do I withdraw a balance?

Table 3.2.: Scenarios with Different Query Types.

The fourth part of the interview aimed to evaluate the strengths and weaknesses of the CS compared to the current keyword search, as well as gather insights on potential improvements and additional features that could enhance the user experience. The questions that were asked for this part are as follows:

- What factors influence your choice between CS and the current keyword search?
- In your opinion, what are the strengths and limitations of CS?
- In your opinion, what are the strengths and limitations of the current keyword search?
- Are there any additional features or improvements you would like to see in CS?
- Is there anything else you would like to add to the topic?

3.3. Survey Design and Implementation

After completing the interviews, customer service agents were invited to participate in an online survey. A survey link was sent to each agent at the end of their interview. Of the 13 invited agents, 12 completed the survey and thus, in the case of 12 agents, qualitative and

quantitative datasets involved the same participants. Afterward, the survey link was sent via email to the agents who did not wish to participate in the interviews, and they were asked if they would like to participate in the survey. 5 more agents took part in the survey, which made the total number of agents that participated 17.

At the beginning of the survey, participants were informed about the purpose of the survey and assured of their anonymity. Participation was voluntary, and agents could choose to skip questions they did not want to answer. The survey aimed to assess customer service agents' perceptions of the conversational search tool across multiple evaluation metrics.

The initial survey consisted of 12 sections, with 20 Likert-scale questions. The questions ranged from Strongly Disagree" to "Strongly Agree". Additionally, at the end of the survey, there was one open-ended feedback prompt asking whether the agents wanted to add something. The study adapted various evaluation metrics from the existing literature. The measurement instruments were adjusted to fit the research context. Table 3.3 presents the evaluated metrics and their corresponding survey questions. Additionally, the survey for the agents who didn't take part in the interviews also included the seven scenarios, and they were asked to choose between the CS and the keyword search.

The metric perceived ease of use was assessed using a 2-item scale, while perceived usefulness was measured with a single-item scale. Both metrics were adapted from Davis [71]. For performance evaluation, a 3-item scale was employed, adapted from Peras [65] and Lu et al. [80]. Three additional metrics—answer faithfulness, answer relevance, and context relevance—were adapted from Ares [69] and Ragas [68]. The three metrics were measured using 2-item, 3-item, and 1-item scales, respectively. Satisfaction was evaluated using a 1-item scale adapted from Oliver [81], while quality was assessed using a 2-item scale adapted from Peras [65] and Oghuma et al. [82]. Lastly, agents' openness to new technologies was measured using a 1-item scale adapted from McKnight et al. [83].

First, the responses to the survey questions were analyzed using descriptive statistics. Mean scores and standard deviations were calculated for each question. Besides that, multiple correlation analysis was conducted in order to determine possible correlations between the variables. The items AF1 and RN2 were reverse coded for the analysis, because of their negative wording. The correlations were analyzed by using the programming language R, which is used for statistical computing and data visualization. Pearson's correlation coefficient was computed to measure the strength and direction of the linear relationship between two continuous variables. The correlation significance test was used to assess the statistical significance of the observed correlations. The null hypothesis for this test is that there is no linear relationship between the variables.

Perceived Ease of Use
Definition: The degree to which a person believes that using a particular system would be free of effort [71].
PEOU1 CS helps me find the information I am looking for without needing additional support.
PEUO2 CS's user interface is easy to understand and requires minimal effort to use.

Performance
Definition: Refers to completion of a task in terms of completeness, promptness and appropriateness [65].
PER1 CS remains stable and functional when faced with unusual requests.
PER2 CS's answers are consistent and logically connected to my questions.
PER3 CS efficiently delivers fast and relevant answers without unnecessary steps or delays.

Answer Faithfulness
Definition: The degree to which the responses generated by the language model are properly grounded in the retrieved context [69].
AF1 I noticed cases where the response didn't match the context retrieved.
AF2 In most cases, CS makes statements that are supported by the information retrieved.

Answer Relevance
Definition: Indicates how well the response corresponds to the question asked [69].
AR1 CS's answers are directly relevant to the questions I asked.
AR2 CS provides complete answers without leaving out important information.
AR3 CS's answers are free of unnecessary details and focus only on what is being asked.

Context Relevance
Definition: The extent to which the context retrieved by the system is focused and contains minimal irrelevant information [68].
CR1 I feel that CS only uses information that is relevant to my request.

Satisfaction
Definition: The summary psychological state resulting when the emotion surrounding disconfirmed expectations is coupled with the consumer's prior feelings about the consumption experience [81].
SAT1 CS met or exceeded my expectations in terms of functionality and performance.

Perceived Usefulness
Definition: The degree to which a person believes that using a particular system would enhance his or her job performance [71].
PU1 Using CS allows me to complete tasks faster and improves my work performance.

Quality
Definition: User perception of the superiority of a service [82].
QUA1 CS consistently provides accurate, correct, and reliable information.
QUA2 CS's answers are structured to make it easy to understand and follow the information provided.

Business Value
Definition: The difference between the effectiveness and the costs of the chatbot [65]
BV1 CS saves time and resources compared to today's keyword search.
BV2 CS's performance justifies its use as a business tool.

Openness to New Technologies
Definition: The general tendency to be willing to depend on technology across a broad spectrum of situations and technologies [83].
OPE1 I am always open to trying new technologies as I believe they will expand my skills and support my professional development.

Replaceability and Necessity of CS
Definition: The evaluation of whether the current keyword search can be effectively replaced by CS and whether CS is an essential addition.
RN1 Today's keyword search can be replaced by CS.
RN2 CS is a useful extension but not absolutely necessary.

Table 3.3.: Items Used to Measure Evaluation Metrics.

3.4. Analysis of Customer Agent Logs

The questions that were asked to the CS by the customer agents were saved in a data labeling tool. A large JSON file containing the logs was exported from this tool. The logs included information such as queries, time and date of interactions, the CS's answers, the CS's three document suggestions, and the agents' ratings of both the answers and document suggestions. As mentioned in the third chapter, agents could rate the answers generated by the CS from 1-5, 5 meaning they are satisfied with the answer. Additionally, they could give thumbs up or down for the document suggestions. Thumbs up identified the document as "good" and thumb down as "bad". If there were at least one document evaluated as "good", the success rate of CS was considered 100%, as the agents could find the information they were looking for in that document.

Moreover, in order to understand if there were any trends over time, an analysis was done with R using the libraries ggplot2, dplyr, lubridate, and readxl. Various visualizations were created to examine how different factors evolved from February to November 2024. The analysis focused on:

- **Average Ratings per Month:** To observe fluctuations in document evaluations.
- **Percentage of Good Documents Over Time:** To determine whether the number of high-quality documents changed.
- **Number of Documents Rated Per Month:** To track user activity and engagement trends.

Furthermore, 400 of the queries found in the logs were divided into seven categories corresponding to the seven scenarios (simple, complex, open-ended, close-ended, short, long, and procedural queries). To correctly categorize the simple and complex queries, all questions were manually searched in the knowledge management tool using the keyword search function. A query was labeled as simple if the answer could be found within one document; otherwise, it was marked as complex.

4. Results

4.1. Interview Results

4.1.1. Scenarios

As explained in section 3.2, customer service agents were asked to state their preference between keyword search and conversational search and explain their reasons for their choice in seven different scenarios (simple, complex, open-ended, close-ended, short, long, and procedural queries). Below are the results of the agents' choices.

Simple Query

For the first scenario, which contained a simple query, 7 out of 13 agents preferred using CS, while 6 agents preferred keyword search. Agents that preferred keyword search stated their reasons for their choice to be speed and familiarity with the tool. They value these two, especially during live customer calls where time is a critical factor. Some agents added that typing a few keywords to the keyword search takes less time and is more intuitive than formulating a full question in the CS. Additionally, agents mentioned that they prefer the keyword search for simple and frequently referenced queries, as they are sure of where to find the correct information quickly and directly. Some example reasons for choosing keyword search for simple questions are as follows:

"For a simple question that I know where I can find it, I always rely on traditional sources. So, I would always first refer to the source I originally used before the conversational search. Because I am confident that this answer is 99% correct."

"I would prefer keyword search because I have actually looked up these e-scooters several times before, and I know which search term will bring me to the documents I need. Especially when I'm on a call and need a quick answer, I would definitely rely on keyword search because I know I can get there quickly and provide help promptly."

"I would actually look it up myself using keyword search. Because during a phone, I would be faster, using keywords in keyword search. With conversational search, you have to formulate the questions directly, and by the time you find something correct, it will take too much time. So, I would rather use keywords in keyword search myself."

On the other hand, some agents preferred using conversational search. According to the

agents, one advantage of using CS is that they can easily differentiate between different types of policies (e.g., liability vs. household insurance) and identify the relevant coverage areas:

"I would use conversational search because it allows me to quickly check the relevant area. For example, it helps me determine if the query relates to liability insurance, or possibly even household insurance in some cases.

Additionally, when agents are uncertain about the exact keywords to use, conversational search is considered a more effective tool since it doesn't rely on exact terms and instead allows them to phrase their query in a natural way:

"Because with regular keyword search, it's more like a collection of loose pages, where without knowing the exact keyword, you often end up making the wrong choice."

"I think I would ask that to CS because I wouldn't know where to start clicking in the keyword search."

The conversational search was highlighted as a beneficial tool for newer policies, whereas keyword search was used for older policies or when a specific keyword guaranteed fast results. Agents stated that CS doesn't contain information regarding older policies and since e-scooters are a topic found in newer policies, they would prefer CS in the case of the e-scooter scenario. Below are exemplifying statements:

"It depends on which tariff it is. If it's the new one, then I would use conversational search. For the older tariffs, I would probably rely on the keyword search."

"For this, I would actually use conversational search instead of any other system, because e-scooters are quite new. For example, with conversational search, we can ask questions about the newer tariffs, but not about the older ones. That's why I would use conversational search."

Lastly, in contrast to some other agents, one agent stated that even if they know where to find the information, typing in the conversational search is often faster than navigating through documents manually. Thus, each agent prefers the tool that they deem faster.

Complex Query

For the scenario of a complex question, of 13 agents, all preferred using CS. A complex question required a look-up to multiple documents. Agents prefer conversational search for its ability to quickly provide relevant information from multiple sources. It allows them to view related topics and documents at once, helping them assess if the information matches their query. This ability is particularly useful for cases where more time is required in order to narrow down the results of the keyword search. Conversational search helps agents view multiple sources and links and find comprehensive answers without the need for searching

manually in databases. According to the agents, the ability to gain a broad overview of relevant documents and relevant links while searching with the CS significantly speeds up the process. Agents also find it helpful when the tool presents answers directly, which reduces the need for additional searches or looking through various sources. In addition to that, CS's preformulated search prompts are considered helpful by the agents, as they can communicate complex information easily, especially for newer policies. Some example responses for choosing CS for complex questions are as follows:

"I would use conversational search because it provides various sources and links from different places where I can find information on the topic. That's why I don't find it bad at all; I can extract the right information. In this case, if I need to consider many factors for the minor policyholder and it's not always found on a single page in the keyword search, conversational search allows me to find everything right away without having to click through the keyword search further."

"I would also use conversational search because it shows me almost everything related to minors, whereas, in other tools, I would have to search for and open each document individually. If it works well and I enter the query correctly, it will display all the documents related to minors as policyholders."

Additionally, one agent stated that they prefer starting with the CS since it gives helpful formulation suggestions that make it easier to explain the answer to the customers. They value that conversational search includes the latest documents and added that while they rely on conversational search for initial guidance, they also double-check the information through keyword search, as they are familiar with the documents and trust their accuracy.

Close-ended Query

For the close-ended question, of 13 agents 10 prefer CS and 3 prefer keyword search. Agents prefer conversational search because it allows them to quickly find answers with minimal effort. It provides immediate responses and relevant links and agents agree that the CS saves time compared to keyword search in this scenario. Agents also find CS fast and efficient when they require a direct answer such as "yes" or "no". Additionally, it can offer more precise and specific answers that help agents avoid the complexity of navigating extensive condition manuals or searching manually in other databases.

"I would use conversational search for this. For such questions, I handle the conversational search well. You can quickly get a yes or no answer."

Many agents agreed that if they search for "master diseases" in the keyword search, it would also show other documents related to health insurance and it would take a longer time to go through all the documents:

"For the question about master diseases, I can enter "master diseases" in the keyword search, and it will show up. However, it will also show other documents related to health insurance. I think with

conversational search, I would be faster."

Some agents stated that they are not familiar with master diseases and they don't often deal with them. Thus, they don't know where to find it using the keyword search and prefer CS for this case:

"Master diseases, are completely new territory for me, so I would actually ask through conversational search and enter the disease to see what comes up."

The agents that chose the keyword search were already familiar with master diseases and therefore they knew where to find the answer using the traditional keyword search methods:

"I would probably use the old reference material simply because I already know exactly where to search. If I didn't know where to search, I would use conversational search again, hoping that it would provide the information when I just entered the disease."

Open-ended Query

For the open-ended question, of 13 agents 6 preferred CS, and 7 preferred the keyword search. The reason for choosing the keyword search was that it allows access to the sources and documents that agents are already familiar with. Since they are experienced in that matter, they are sure that they can find these documents quickly and that these documents will provide them with the right answer. This reason is reflected in the following statements:

"Because I just know where to find it and I know exactly which link I will find it under when I query it through the keyword search. Simply because I know I type in 'ownership change' and it's the third link."

"I actually use the keyword search because it's a question I use very frequently, and when I have a query that I simply want to confirm for myself, I go to my original source. And I always know which page I need to flip to, where the solution suggestions for the specific situation are."

"Okay, it used to be my area of expertise, so I would first use the old documents in keyword search because they are well-updated. I would place them next to each other and check if anything has changed or if there are any updates I don't know about, but I would definitely start with the old documents."

Another reason to choose the keyword search was that they feel more comfortable using their old documents and current databases for queries such as ownership transfers that require a detailed answer. In their old documents, they are able to find comprehensive legal information. They mention that even though the CS can give them summarized answers, they prefer the keyword search as the information there is detailed, current, and verified. They trust their familiarity with the old tool instead of the new tool and they would like to be sure they are giving the customers the correct information. Some relevant answers are:

"Because the topic is very extensive, I would still need to check in keyword search, as there are many documents regarding the legal regulations. We have multiple letters for both buyers and sellers, and there are a lot of things to consider. The AI does summarize these well in the response options, but in keyword search, I can find everything, especially when I need more than just a short answer."

"I would use my old documents. I think I would still rely on my traditional sources since I've never really searched for ownership transfer using keyword or conversational search. If I were in that situation, I would probably stick to my old documents and sources to be on the safe side. Even though conversational search might provide the answer directly, I would still choose the more familiar path for now."

One agent stated that they would use the keyword search for this scenario since they are familiar with it. However, they added that they were also curious about what results the CS would give and that they would also try it there. If they see that the results are accurate, they will use the keyword search in the future for similar cases. On the other hand, another agent said that they already tried using the CS for such detailed questions and it didn't work and required more time than using the keyword search.

Of the agents who preferred the CS stated that they were not familiar with this topic and thus, they wouldn't know where to start using the keyword search. Another agent mentioned that they would use the CS as it is a detailed question and that the CS gives accurate answers if they prompt the question in a detailed way. A third agent noted that they would use the CS however they wouldn't check the document link suggestions but they would check the generated answer. They would quickly evaluate whether the generated answer is correct and adjust it if necessary. Lastly, one agent mentioned that finding the answer using the keyword search is sometimes more complex and not as quick as using the CS.

Short Query

For the scenario of a short question of 13 agents, 10 preferred CS, and 3 preferred the keyword search. There were several reasons for the choice of the CS. Many agents value that they can find the answer directly and quickly using the CS instead of looking through multiple links and documents. One agent mentioned that they see it as a big advantage that they avoid multiple steps that they had to go through using the keyword search. Moreover, for broad or general topics like the question in our scenario, conversational search is seen as effective, as it doesn't require searching within specific categories or segments. Agents also find the CS useful for current and broad topics, and they trust that the system will provide them with the right answer in that case. In addition to that, for the questions that don't require detailed answers, agents agree that the search time is reduced by using the CS, and the process is simplified as it gives concrete answers directly. Example answers for the choice of using CS are as follows:

"I think I would prefer conversational search because it's easy to formulate queries, even as a question. It's quickly typed, and the answer will likely come just as quickly."

"I would use conversational search. I think it would be faster if it directly outputs "18 months" rather than having to search through everything again. That would simplify the search, I think, because the answers are simpler and don't go into much depth. I believe that works better with conversational search."

Additionally, one agent mentioned that they would prefer conversational search if it is already open on their computer. However, if a session timeout required a new log-in, then they would switch to the keyword search:

"It is a current topic, and I know it must be stored in conversational search, as it's a quick query and a general question. However, if I have to log back into conversational search because it closes after a certain period, I find it a bit inconvenient. That's why, as I said, if it's not already open, I would use keyword search, because the information is also there, and I can find my answer relatively quickly."

Agents' reasons for preferring the keyword search is that the agents are familiar with the process and they know exactly where to find the answer. Because of that, they believe that they are faster using the keyword search. Since this scenario contains a short and direct question, they think the keyword search is more efficient. They don't have to type detailed questions or wait for a response.

"I would use keyword search because if I enter "immediate protection" it directly gives me 18 months, and I wouldn't need to rephrase the question."

"I would prefer the old method because of speed. I would just enter "immediate protection" briefly and have everything, so I wouldn't need an answer. That's why I would choose the shorter option for me, as I know exactly where to find it."

Long Query

For the scenario of a long question, of 13 agents, 7 preferred the CS and 6 the keyword search. Their reason for choosing the CS was that agents believed that if they could formulate the query well enough, the CS would give them a precise answer. Agents agree that it is easier to find the information they are looking for, especially for newer policies where there are predefined answers or formulations. Thus, agents aim to save time using the CS. However, some agents added that in some cases they return to the keyword search for more complex matters to make sure the answer given by the CS is correct. Some example answers are as follows:

"I would use conversational search because it helps you get to the target and the relevant documents faster, especially for longer and detailed questions if you ask it correctly."

"I would like to use conversational search because it usually provides good documents, as it's a current topic, and since there are already predefined formulations, I would verify them for accuracy. I

might go back to keyword search if the answer seems unclear, but so far I have had good experiences with these predefined answers. That's why conversational search is very interesting for me in this case."

On the other hand, agents who preferred the keyword search also had several reasons for their choice. One of the main reasons was their lack of trust in the CS. Agents stated that for more detailed and specific questions, they trust the keyword search more. One agent mentioned that sometimes even though they think that the CS would be a better option, they don't opt to use it. Moreover, they added that for more complex matters, even if the CS could give them perfect answers, they would still prefer the keyword search to be sure they are getting the correct information. Another agent also mentioned that if the query is a complex query with multiple aspects to consider, they would not trust the CS with it just as they don't trust ChatGPT. They believe that the CS is not yet ready to handle checking multiple aspects of a complex query.

Moreover, agents mentioned that to find answers easily to complex questions, they feel more confident using the keyword search, as it has a more straightforward approach. They believe the keyword search is more reliable when the query involves multiple aspects, as they can validate those aspects from official documents such as terms and conditions. This makes them feel more in control of the process. They find this approach better as the insurance themes require precision and clarity. In addition to that, since they are familiar with the process and can easily locate the relevant documents, they find using the keyword search faster and more efficient to find the detailed information.

"I would rather check through keyword search and not conversational search. Simply because it's too specific, I would say. The question is so extensive, down to the smallest detail. I would rather check in the keyword search and read a bit more about the insurance coverage there. I prefer CS for not-so-detailed questions. Not so deep, more for general, smaller questions. Therefore, when it comes to the details, I don't think it will give me exactly what I need. I'm just faster with keyword search because I am more familiar with it."

Next, agents were asked whether they would search for the long question by writing the full sentence in the conversational search or just using keywords. The majority stated that they would shorten the question, as it is faster. However, some agents mentioned that they prefer to use full sentences, as they were advised that the more words included in the sentence, the easier it is to find the information:

"I would search with a full sentence. I would phrase it out because we were told that the more words in the sentence, the easier it is to find the information. I have preferred using the keyword search for the keywords because the formulation there was often not ideal. However, in conversational search, I find the long-form phrasing more effective. It's more stressful for me to write because it involves more text, but I feel the hit rate is higher."

Procedural Query

In a procedural question scenario, 7 out of 13 agents preferred the conversational search, 4 preferred the keyword search, and 2 were undecided. The agents that preferred the CS mentioned that the CS only focuses on what was asked and shows relevant results, while the keyword search is too broad and shows results about multiple topics that are not necessarily relevant. With the keyword search, agents have to click through each link to find out whether they are relevant to the question or not.

For this scenario, many agents stated that they never had such a question before and they didn't know where to find the answer using the keyword search. If the agents don't know where to start, they tend to use the CS as they can see whether the information matches the query more easily and faster. If they see that the result aligns with the query, they will proceed to the relevant document.

On the other hand, some agents were skeptical about whether CS would provide the correct answers, as they had never used it for such a procedural question before. These agents preferred the keyword search instead.

"I think I would rather check in the database using the keyword search for such a work instruction. I'm not sure if conversational search can provide that. Therefore, it would probably take more time because I'm not sure if the system can even pull up such a work instruction. So, I would prefer to look through the keyword search. You really need a step-by-step guide for that, and I don't think CS can handle that yet. But it's possible I'm mistaken, I haven't tried it yet."

One agent, who was also uncertain about whether CS would provide the correct answer in this case, wanted to test it during the interview. They typed the question into CS to see if a work instruction would appear. They were not satisfied with the suggestion and stated that, if they had to rate it, they would give it a thumbs down.

4.1.2. Factors Affecting the Choice of the Tool

Familiarity and their Confidence in Knowledge

After the seven scenarios, agents were asked which factors influenced their choice between keyword search and CS. The factor that was mentioned the most was their familiarity with the tool and their trust and confidence in their existing knowledge. The agents tend to use the keyword search if they know where to find the information. Because of their familiarity with the old tool, they feel more comfortable using it. One agent mentioned that sometimes, even though they know they would probably be faster using the CS, they still tend to use the keyword search because of the comfort they feel. On the other hand, agents use the CS if they don't have any knowledge about the topic or if they are not sure where to find the information. One agent stated that if they don't know where to start searching, they always start by using the CS. Some example answers are as follows:

"If I'm already quite familiar with the topic and confident that I can find it using keyword search,

and that I can quickly locate it because I already think I know the answer and just want to confirm it, I would handle it myself. For topics where I'm very unsure or have no idea, I would use CS because it provides a better overview of the referenced documents."

"Because I've done it so many times and have searched for it before if I know exactly where it is, keyword search is definitely more helpful. However, given the vast amount of information we have to go through, keyword search is usually not very helpful when I have no idea what I'm looking for."

This suggests that newly hired customer service agents may prefer using the CS more, as they have no familiarity with the existing keyword search and don't know where specific information lies within the tool.

Query Complexity

The complexity of the query is another factor that influences the agents' choice. Some agents stated that they prefer the keyword search for simple questions, while they prefer the CS for more complex or detailed questions. CS can give multiple suggestions relevant to the topic and can provide an overview of the various aspects that a question should cover in its answer. According to the agents, if they are unsure of what to search for when they are dealing with unfamiliar topics, then the CS is a helpful tool. One agent explains this as follows:

"The complexity of the question also plays a role. The more complex the question, the more likely I am to use conversational search because it can provide multiple answers or suggestions. Sometimes, I'm not entirely sure what exactly I need to know or what areas I need to consider. When it shows me everything related to the keyword, I can approach the issue with a broader perspective. That's why I would turn to conversational search for more complex questions. For simpler queries, I would probably stick to the methods I've used so far."

Speed and Time Pressure

Speed and time pressure were mentioned as other key factors. Agents prefer the keyword search for live customer calls where they are under time pressure. They believe that they are faster using the keyword search. On the other hand, they find the CS is a valuable tool for situations where there is time to review and consider detailed suggestions, such as when writing customer confirmations or conducting post-processing tasks:

"When it comes to phone inquiries, I always stick to keyword search because it takes time to find the right information using conversational search initially. For tasks like handling mail, however, I find conversational search excellent. In such cases, there's time to read through the information since there isn't someone on the line waiting for an answer. That's when I would prefer to use conversational search."

"In tasks like document processing, it's a 50-50 choice depending on the situation. If I have time

and no immediate pressure, I might explore conversational search to browse through the available documents and find what fits best. However, if time is tight during document processing, I would rely on keyword search, as I know the old system well and can quickly find what I need. Time is always the deciding factor, depending on how quickly I need the information."

One agent explained that it is very easy to find the answer immediately using the keyword search and that they only have to type one keyword. However, they mentioned that if they search for something more specific, the duration of finding it would be higher. On the other hand, if they use the CS, it takes more time for them to formulate and type the question, but the CS would provide the correct result faster as it provides multiple suggestions that are relevant to the query. This aligns with the finding of Wazzan et al. [77] who also reported that "participants using the LLM-based search issued longer, more natural language queries, but had shorter search sessions".

Moreover, one agent stated that while they were on the phone with a customer, they always thought of which tool would bring them faster to the correct answer. If they have a rough idea about how the answer should look and where to find the information, they would prefer the keyword search. However, if they already know the answer and only want validation and help with the formulation, then they would use the CS.

Trust

Reliability and trust in the tools also play a role. Agents accept that CS is a valuable tool, however, some hesitate to fully trust it since it is a new tool for them. *"Because we are still testing conversational search, I don't trust it 100 percent yet."* Keyword search, by contrast, is seen as a more reliable option, particularly for familiar tasks.

Nature and Scope of the Task

The nature and scope of the task further determine the choice of tool. The agents prefer the keyword search for well-known and frequent matters. For broader topics that require knowledge about various aspects and a detailed exploration, they choose the CS. Furthermore, agents state that the CS only contains newer policies, and the old policies are not included. Thus, for questions about the older policies agents' only option is to use the keyword search. Example answers are as follows:

"First, whether it's something general or specific. For general matters, I would use conversational search, and for specific things, I would search on my own. As I mentioned, it often happens that documents are displayed that aren't helpful."

"Newer policies are only available in conversational search for now. It would be even better if older policies were included too, as it would make the tool more comprehensive. Currently, I have to use other methods for older policies. So, speed and the scope of available information, such as policy coverage, are important factors when choosing a search method."

Search Experience and Usability

Agents had different opinions on which tool they found easier to use. One agent stated that they are more confident and comfortable with the keyword search. They find it easier to use since they don't have to navigate through links and they have it set as their homepage for easy access.

"I think mainly because I feel very confident with keyword search and know where to click. The instructions are easier for me. With conversational search, you always have to navigate through the links, and it might turn out that it's not the right one. You have to search through it first. I'm just not as experienced with it. That's the main reason. Also, with conversational search, you have to activate it and log in. I always have keyword search open as my homepage, although you can do the same with conversational search."

Another agent, however, finds conversational search easier to use as they can directly see where the answer is located using the CS, and they can easily assess whether the answer is accurate. In comparison, keyword search is seen as less precise since it returns too many irrelevant links and results.

"I find keyword search more complicated—it spits out things I don't need, whereas CS is more precise, I'd say. I can also see, for example, with private protection, it lists which folder it's in, so I can immediately see if it's relevant or not. In keyword search, there are a lot of links. Exactly, it's a bit overwhelming for me, I was a bit overwhelmed and didn't know, okay, which one should I click on to get my answer."

4.1.3. Strengths and Limitations of the CS

Strengths of the CS

According to the agents, one of the main strengths of the CS is its ease of use and efficiency. Agents stated that they handle a variety of work instructions from different areas, and using the CS has significantly eased the process of finding the information they need. They added that the system had already improved a lot, and they were impressed by the well-formulated summaries and answers, increasing the efficiency. This also makes their daily work much easier as they can use the response directly, they don't have to spend extra time thinking about how to phrase it themselves, and they *"do not need to take raw information and turn it into something polished for the customer"*. Thus, they think the CS is significantly more customer-friendly.

"I am genuinely pleased because it provides tremendous relief in managing tasks. We handle a wide variety of areas, countless work instructions, and clauses, and using conversational search has made finding information significantly more comfortable compared to our previous methods. The system has already improved to the point where it delivers impressively well-formulated summaries and answers, which greatly enhance efficiency and ease of use."

Moreover, agents state that it is relatively easy to find the correct document using the CS, and it avoids the confusion of a keyword search that provides a large number of results. They think that the range of the keyword search is too broad and that they easily get lost between too many documents. They agree that CS is practical and more precise than keyword search since *"it doesn't give an overwhelming amount of links all at once"*. Furthermore, using the CS saves agents time because it allows them to find the particular area of the document where the solution is located. This minimizes the overall difficulty of finding the correct information.

In addition to that, agents agree that one of the important strengths of the CS is its ability to give comprehensive results that address various aspects of a query at the same time. According to the agents, this ability is particularly helpful for complex customer questions. It presents the agents with multiple answers that may be relevant to their questions. That way, agents don't have to search separately for different areas. When the agents type a query to the CS, they get all the results where the term can be found. Agents consider this as a major strength. Moreover, agents find it a nice feature that they can give feedback on whether the response was helpful or not. They hope that this will allow the system to continuously improve and with time, the system will work more efficiently by providing accurate answers.

Limitations of the CS

Despite its benefits, CS also has numerous limitations according to the agents. One limitation is that the older policies are not included in the CS. Agents noted that older policies make up 50% of their workload and it would be helpful for them if the CA not only covered the new policies but also the old ones. Agents want to be able to use the CS for the old policies and thus, this limitation makes CS less useful for agents who frequently handle questions about older policies and contracts.

"The limitation is that the old tariffs are not included; it is initially focused solely on the new ones. There are many questions about the old system, older tariffs, and specific issues that are simply not covered yet. I am not sure what will be added gradually, but for now, when it comes to current tariffs, I feel confident and can use conversational search. However, for questions about older contracts, I have to rely on the traditional system because that's where the old conditions are stored."

Another restriction, according to the agents, is that technological issues sometimes make it difficult to use CS. One agent mentioned an event in which changes were expected to take one day but took a week. The agents could not use the CS that week, they noted that afterward, even though the issue was resolved, the system remained slow or unresponsive.

"I find that the technology still has some shortcomings. For instance, there was an update that was supposed to take one day, but it ended up taking about a week. During that time, we couldn't use conversational search at all, which was quite inconvenient. After the issue was resolved, I still couldn't access it properly for a while, even though everything seemed to be functioning. I would enter a question, but it just kept loading without providing an answer. Aside from that, I haven't experienced any major problems yet. Sure, the links aren't always accurate, and the answers aren't always right,

but since the AI is still learning, I believe it will improve significantly over time."

Moreover, agents must critically assess the accuracy of responses, as they are not always correct. The agents can't rely on the accuracy of the responses 100% as the CS is still in the development phase. However, agents want to make sure they give the correct information to the customers on the phone and for that reason, especially in challenging situations, they double-check and validate the CS's responses by using the keyword search.

Another disadvantage of the CS is that if users don't provide enough information, it offers multiple document links, which the agents sometimes find overwhelming. CS's performance depends on how detailed the user queries are. If queries are ambiguous or incomplete, the system may return irrelevant results. This increases the time to find the correct answer as the agents have to sort through the provided information. Some agents believe that having to write a full sentence or carefully select keywords to ensure the outcome is accurate, is inconvenient. Furthermore, the tool's response times can be slow, especially when users formulate detailed queries or when they don't specify the category. According to the agents, for situations where a quick response is required, this delay can be a significant disadvantage.

"If you don't write exactly, it will return too many links, and you have to check all of them. While this is less than a keyword search, it's still a lot. What I've noticed is that it's also a bit slow. When a customer or representative calls me, I need a quick answer. But before I can formulate a sentence in conversational search, I have to wait for it to load. It still takes a bit too long for me."

"A weakness, at least currently, is that I have to think carefully about what I enter. I can't just randomly type in keywords that I think might lead to the right result. Instead, I need to carefully consider what I'm searching for, or else it takes a while to navigate through it. So, the keywords themselves need to be chosen carefully to ensure the correct result."

Additionally, some agents expressed that they would like to refine their queries or ask follow-up questions to the CS. This is currently not possible, and agents have to open a new chat and start with their queries from the beginning. Agents believe that this makes the search process less efficient for them.

Furthermore, agents mentioned that currently, they receive the answer from the CS as a large text without any highlighted parts, and they see this representation as a disadvantage. They expressed that it would be a nice feature if the most important and relevant parts of the text are presented in bold or highlighted so that it gets easier to extract the information.

4.1.4. Strengths and Limitations of the Keyword Search

Strengths of the Keyword Search

According to the agents, the keyword search has several strengths, such as its speed, broadness, and the fact that agents are familiar with it. If the agents know the right term, using the keyword search can get them the information quickly and they don't have to write long

queries as in the CS. According to the agents, the keyword search uses a very comprehensive database where they can find all the documents they are looking for. Agents find the system reliable, and they are confident that if they have the right keywords, even information that is hidden or difficult to retrieve could be retrieved.

"Some things are difficult to find, but one of keyword search's strengths is its ability to locate even hidden information effectively."

"Keyword search, in general, is really good because it feels like everything is documented somewhere. You just have to know where to look, but it's all there."

Another advantage is familiarity with the tool, as agents are used to it and know how to use it. One agent describes this as:

"I think it's mainly the familiarity here that is the strength, the fact that one simply doesn't know it any other way and has already gotten used to it. It's not really a strength, but I wouldn't know what would be better than CS, except that I know this tool better."

Furthermore, agents find some features of the keyword search very useful. They can save the documents they use often with the saved favorites feature and access them with ease when they need it. Agents also felt that the system provides good visibility of updates and document statuses. They can also provide feedback if they notice a document is incorrect, and it gets corrected relatively fast.

Limitations of the Keyword Search

Agents expressed that one of the major limitations of the keyword search is the fact that if they do not know the exact term they should use to search for something, they may not find it. Using synonyms or different versions of what they are searching for may lead them to fail in finding what they are looking for. Therefore, the keyword search is not flexible regarding the keywords the agents type, and it requires them to know about the terms associated with each document.

"The limitation is that sometimes the keywords in the system are not entirely accurate. For example, an agent might search for "change of possession" ("Besitzwechsel") but not find relevant results because the system uses the term "change of ownership" ("Eigentumswechsel"), even though they mean the same thing. These small discrepancies, or "form errors," can make finding the right information challenging."

Another disadvantage of the keyword search, as the agents said, is that *"it returns too many results"*. This makes it difficult to find the correct answer. Agents reported that too many links appear for one query, and it is not easy to determine the most relevant documents. As a result, they have to go through every link to find information suitable for their needs. Therefore,

they find the keyword search is less effective and less friendly than the CS. Clicking through the documents is especially hard when agents are under time pressure. They appreciate that the CS returns a small number of links and does not require an extensive review of the documents.

"Sometimes it feels like there is simply too much information. When a keyword is entered, too many results are displayed, and one has to click through each link to find the correct answer. This is a major disadvantage. Personally, I am not very impressed with keyword search because I find it too cluttered and overwhelming."

"I find keyword search overwhelming and unorganized. There's just too much at once, and it's not immediately clear where to go. Sometimes I feel a bit lost when using the current system. I've never quite gotten the hang of it. I often end up somewhere else, not where I need to be. From my perspective, it's sometimes just too chaotic and excessive."

Agents mention that the keyword search also lacks the ability to precisely interpret the context and provide related results. A good example would be when the agent wants to query about a specific category, "legal protection insurance" ("Rechtsschutzversicherung"), but it nonetheless shows findings concerning "pet health insurance" ("Tierkrankenversicherung") as both results include a common term like "waiting period" ("Wartezeit"). The agents can only find the right answers if they are already in the correct category, or else the number of unnecessary steps increases, and it slows down the process.

"The weaknesses, without a doubt, include the need for more targeted searching. I can't simply enter a keyword anywhere; instead, I need to already be in the correct area to use the keyword search effectively. This is definitely a drawback because, as mentioned, there are times when you're not entirely sure where you should be searching or where to input the keyword to get the desired result."

Another limitation for the agents is that the keyword search system is hard to navigate. Agents believe that the filter options they can apply to the results are inefficient. Sometimes, they are unable to find the documents that have been relocated or updated, and there are cases where they receive error messages about missing documents. These complications may raise the time and labor required to identify suitable information. In addition, the agents point out that when they only utilize keywords they are familiar with, they sometimes miss the changes or additions made to the knowledge base since the system does not highlight the new information. Because of the lack of awareness about the latest developments, agents' ability to provide the most current information to customers may be affected.

"Some agents prefer keyword searches simply because they've worked with them longer and know which terms lead to specific information. This familiarity can be limiting, as it assumes agents already have a good idea of what they're looking for. It also risks overlooking new information. For example, if a new article is added under a different keyword, it might not be noticed if agents habitually search

under the same familiar terms to review information."

4.1.5. Additional Functionalities

Next, agents were asked whether there were improvements and additional functionalities that they wished for the CS. Some agents wanted to have more enhanced input methods, such as the ability to ask questions via voice instead of typing. They feel this would save them time and allow them to multitask when talking with a customer over the phone. For example, instead of writing down the question of the customers, they would directly ask the question to the CS with their voice. Also, agents wanted the CS not to require writing full sentences and to work better with typing only keywords, which would be more practical and would save time.

"I would like to try controlling it with voice. So, not only typing everything on the keyboard but actually being able to link into CS while I'm on a call and ask the question verbally. This way, I wouldn't always have to type and have a pen in my hand because I'm taking notes from the representative or customer, but I would be freer and could perhaps formulate the question further. Typing takes longer than speaking."

"With CS, you get the best results when you ask the question as thoroughly as possible. You should formulate full sentences and avoid using abbreviations. I've noticed that otherwise, the results are different and not as satisfying. Perhaps a long-term improvement could be made so that you don't have to phrase the questions so elaborately. That could be helpful in the long run." Furthermore,

agents wished that the answers provided by CS were 100% accurate and directly relevant to their query. That way, they can use these answers directly, and they do not have to check the answers for their correctness. According to the agents, that would enhance trust and efficiency.

"I don't know how realistic this is, but it would be great to get a 100% correct answer. Because as it is now, you can't rely completely on it and always have to double-check. It would be great if there was an answer you could trust 100%."

"I would wish that the response time could be a bit faster, because right now, as I mentioned, we receive document links and are supposed to ignore the answer because it is usually incorrect. It would be much more helpful if we received a direct answer instead of just document links, through which we then have to search for the answer again."

Additionally, agents wanted a more direct presentation of the results. They state that the system presents too many links, which makes it harder to detect the correct link. Hence, they would like the number of links provided as the response to a query to be limited and make sure that only the most relevant ones are displayed.

"It's a bit overwhelming when 3 or 4 answer options appear, and it takes a moment to figure out which link to click on."

Agents also wished that they were able to open different chats, which they could divide into categories, such as liability insurance or health insurance. They believe that this might help to improve clarity and retrieve past queries more easily.

"It could be divided into categories. I could imagine opening a chat where I only ask liability insurance questions, and then later, if I have the same question again, I could find it more quickly in that chat."

Moreover, it is mentioned that the user interface of the CS system feels outdated and not visually pleasing, like older systems. Agents suggest improving the system to make the interface more modern. They also wanted an interface where the most important points are printed in bold or highlighted. Another agent mentioned their desire that near the responses, percentages are shown, stating how accurate the answer is. They further suggest that, for instance, the answers with high accuracy could be displayed in green while the low-quality answers are highlighted in another color. The agent thinks this would make the system more user-friendly and efficient.

Agents further wish the coverage to be extended so that they can ask about not only new tariffs but also older ones, which are currently out of the scope of CS. Also, in cases where agents need to transfer a customer from one department to another department, they would like to find the corresponding internal phone numbers. Agents believe including all this information would help them to better assist customers.

"It would be great if the system could cover all sectors and tariff generations. It would be helpful to be able to ask about any tariff, not just specific topics like what is covered under a homeowner's insurance policy, but also for phone support. For example, if I need to transfer a customer to the health department but can't find the number, it would be great if the system could provide those contact numbers. That way, the system would give the correct number to transfer the customer, simplifying things for phone support and preventing misdirected transfers. I think that would be really useful."

Lastly, agents wanted dynamic interaction with the system. The fact that CS can refer to previous questions and follow up dynamically will make it more conversational and friendly for the user.

"I also suggest that it should be dynamic and allow referencing previous questions. The CS should be able to integrate previous questions and follow-up questions. This works, for example, with ChatGPT. When you ask something and say it's not the answer, you can add more context, and the answer will be re-evaluated and provided again."

4.1.6. Additional Comments

At the end of the interviews, agents were asked if they wanted to add anything. One agent mentioned that they find CS weak and not as perfect as they would like. They mentioned that sometimes it is very difficult to judge an answer critically after reading it once, and they are unsure whether the answer is correct, as the answers are very general and not as targeted. They said that in terms of how much they trust the newly integrated CS, it is 50%, while for ChatGPT, their percentage is 70%. One of the agents expressed they are glad now to have CS, which they find very practical. They express that they are confused using the keyword search because they always need to find the right word to get a result. Another point raised was that CS is not yet fully integrated into their daily routine since they haven't been using it for long. However, they believe that once it gets habitual like the old systems, it will be a greater relief. Additionally, one agent suggested that CS should not turn off after a certain period. They said it would be nice if it could be used continuously so that they could open it and have it work right away.

4.2. Evaluation of Conversational Search Systems

As outlined in Section 2.4, various metrics and criteria are used to evaluate conversational search (CS) systems. Are CS's answers consistent and logically connected to customer agents' questions? Does CS provide complete answers without leaving out important details? Does CS make statements that are supported by the information retrieved? In order to find answers to such questions, a voluntary and anonymous survey was conducted with customer service agents within the organization and 17 agents participated. The survey assessed 11 key metrics for evaluating conversational search systems:

- **Perceived ease of use**
- **Performance**
- **Answer faithfulness**
- **Answer relevance**
- **Context relevance**
- **Satisfaction**
- **Perceived usefulness**
- **Quality**
- **Business value**
- **Openness to new technologies**
- **Replaceability and necessity of CS**

Two tables are provided in Appendix A.3.: Table A.1 presents descriptive statistics for each metric as a whole. Table A.2 details which survey questions correspond to each metric along with descriptive statistics for each statement.

Perceived Ease of Use

The metric perceived ease of use was evaluated to find out how easy it is to use the conversational search for customer agents.

First, customer agents were asked whether CS helps them find the information they need without requiring additional support. Among the 17 customer agents, one agent (5.88%) remained neutral, 12 (70.59%) agreed, and 4 (23.53%) strongly agreed with the statement. None of the agents strongly disagreed or disagreed, which means the CS enables them to find information independently without relying on additional assistance. However, even though the result indicates they do not need assistance, the interview findings showed that to be more efficient in finding the right information, agents need training in formulating their queries.

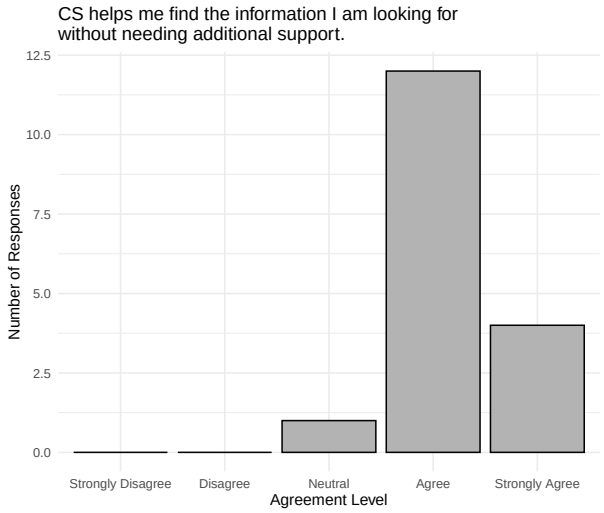

Figure 4.1.: CS helps me find the information I am looking for without needing additional support (n=17).

Another statement to measure the perceived ease of use was whether CS's user interface is easy to understand and requires minimal effort to use. None of the customer agents strongly disagreed or stayed neutral. 1 agent disagreed (5.88%), 9 (52.94%) agents agreed, and 7 (41.18%) agents strongly agreed with the statement. The fact that all the agents agreed with the statement except one indicates that the agents perceive the system as user-friendly and intuitive. The disagreement of one agent could be due to individual preferences or lack of familiarity.

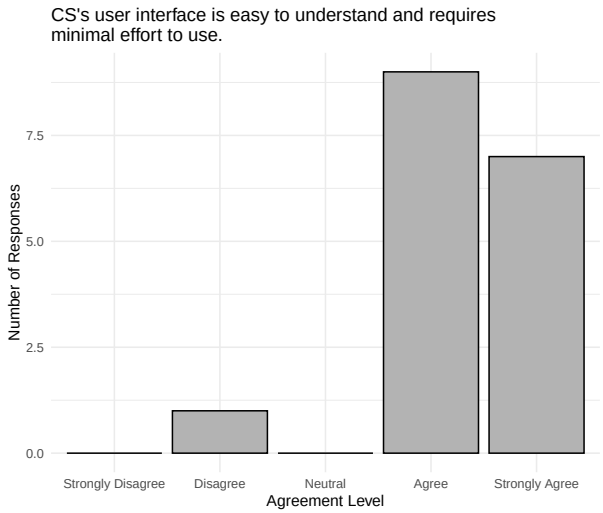

Figure 4.2.: CS's user interface is easy to understand and requires minimal effort to use (n=17).

Overall, the metric Perceived Ease of Use has a mean of 4.24, indicating that customer service agents generally found the system easy to use. The standard deviation of 0.39 suggests a low variability, meaning that most agents rated the system similarly.

Performance

The next metric evaluated was performance, with the objective of assessing the stability and coherence of the conversational search under various conditions. There were 3 statements to evaluate performance.

The first was "CS remains stable and functional when faced with unusual requests." One of the customer agents (5.88%) strongly disagreed, 3 (17.65%) disagreed, 4 (23.53%) stayed neutral, 6 (35.29%) agreed, and 3 (17.65%) strongly agreed with the statement. While slightly over half of the agents found the system stable and functional even when faced with unusual requests, the perception of the remaining agents suggests that the system may struggle with unusual requests. To prevent any inefficiencies, such specific areas should be identified, and the ability of the system to handle edge cases should be improved.

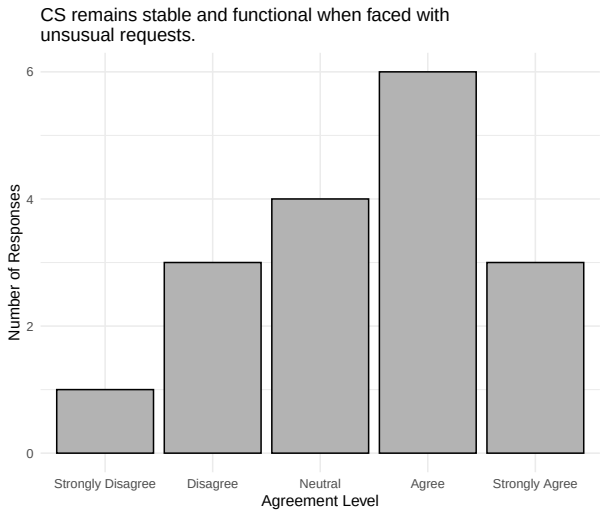

Figure 4.3.: CS remains stable and functional when faced with unusual requests (n=17).

Next, they were asked whether CS's answers were consistent and logically connected to their questions. None of the agents strongly disagreed. Two of the customer agents (11.76%) disagreed, 3 (17.65%) stayed neutral, 10 (58.82%) agreed, and 2 (11.76%) strongly agreed with the statement. The fact that 70.58% of agents agreed or strongly agreed suggests that the system is generally effective in providing consistent and logically connected answers, which is an important factor for maintaining trust and efficiency in customer service operations. The small percentage of disagrees may point out specific instances such as ambiguous queries or detailed scenarios. The neutral responses indicate that some agents may not have had enough exposure to the system or may have encountered mixed results regarding the answers' consistency.

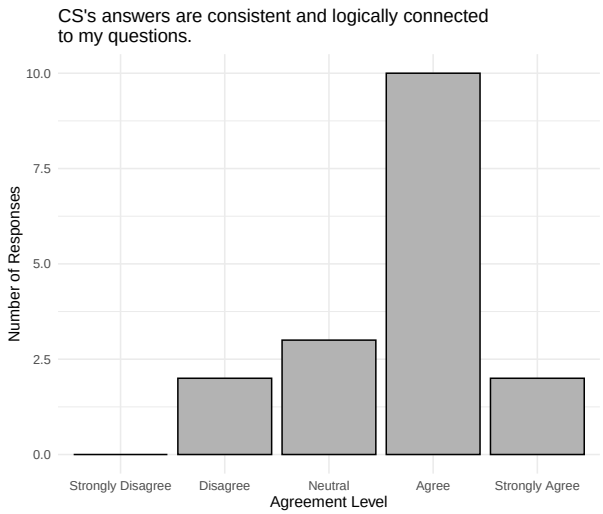

Figure 4.4.: CS's answers are consistent and logically connected to my questions (n=17).

The last statement was whether CS efficiently delivers fast and relevant answers without unnecessary steps or delays. None of the agents strongly disagreed. 4 (23.53%) of the customer agents disagreed, 3 (17.65%) stayed neutral, 4 (23.53%) agreed, and 6 (35.29%) strongly agreed with the statement. While the majority of the agents find the system efficient in providing quick and relevant answers, disagreeing agents suggest that there were also some instances where the system may have failed to deliver fast and relevant answers efficiently. This could be due to technical limitations, which is an area for the system to improve.

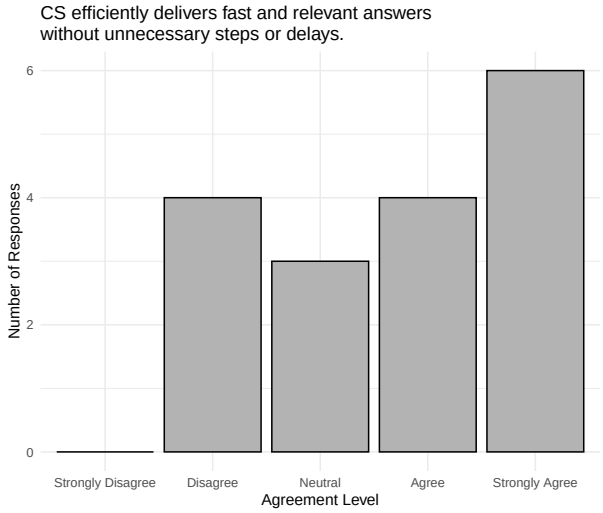

Figure 4.5.: CS efficiently delivers fast and relevant answers without unnecessary steps or delays (n=17).

Overall, the performance metric has a mean score of 3.55, indicating a slightly above average perception of performance among customer service agents. However, the high standard deviation of 1.13 suggests that different agents have significantly different opinions about the system's performance.

Answer Faithfulness

Furthermore, CS's answer faithfulness was evaluated with the objective of determining whether the answers provided by the conversational search were accurate and based on the information retrieved.

The first statement was, "I noticed cases where the response didn't match the context retrieved." None of the agents strongly agreed. One of the customer agents strongly disagreed (5.88%), 7 of the customer agents (41.18%) disagreed, 1 (5.88%) stayed neutral, and 8 (47.06%) agreed with the statement. While nearly half of the agents did not notice significant issues with the system's contextual accuracy, an equal proportion observed cases where the response did not match the retrieved context. Some agents may have encountered more problematic cases than others, possibly depending on the types of queries or scenarios they handled.

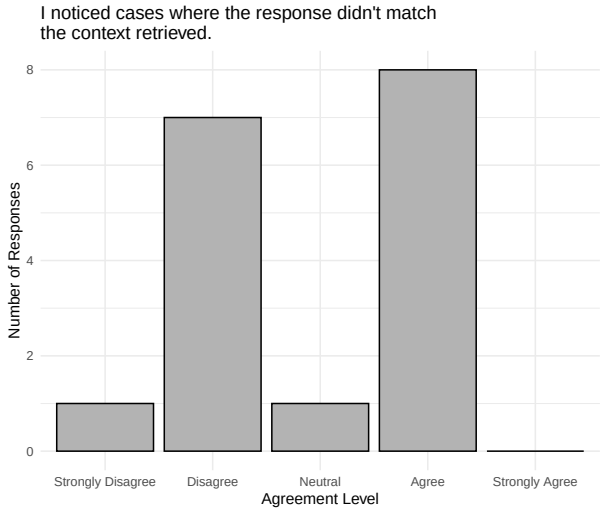

Figure 4.6.: I noticed cases where the response didn't match the context retrieved (n=17).

The next statement for the metric of answer faithfulness was, "In most cases, CS makes statements that are supported by the information retrieved." None of the agents strongly disagreed or disagreed. 3 of the customer agents (17.65%) stayed neutral, 10 (58.82%) agreed, and 4 (23.53%) strongly agreed with the statement. The fact that 82.35% of agents agreed or strongly agreed suggests that the system is highly proficient in making statements which are supported by the retrieved information.

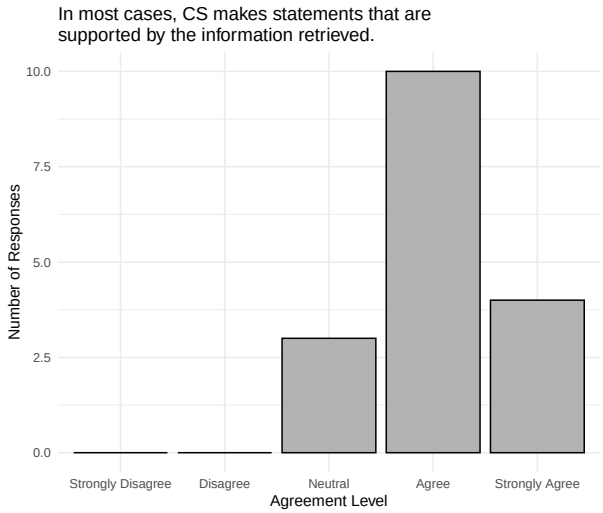

Figure 4.7.: In most cases, CS makes statements that are supported by the information retrieved (n=17).

The answer faithfulness metric has a mean score of 3.56, reflecting a slightly above-average level of agreement among customer service agents regarding the faithfulness of the answers

provided by the CS. The standard deviation of 0.70 indicates a moderate level of variability, suggesting that while some agents may find the system's answers faithful, others might have different opinions.

Answer Relevance

Moreover, answer relevance was one of the evaluated metrics, with the objective of ensuring the conversational search provides relevant, complete answers without unnecessary information. Three statements were asked of customer agents for this metric.

The first one was, "CS's answers are directly relevant to the questions I asked." None of the agents strongly disagreed, and one (5.88%) agent disagreed. 2 (11.76%) of the agents stayed neutral, 13 (76.47%) agreed, and 1 (5.88%) strongly agreed with the statement. While there is a small portion of neutral and negative feedback, the overall results suggest that the system performs well in understanding and addressing user queries and providing directly relevant answers to the questions asked.

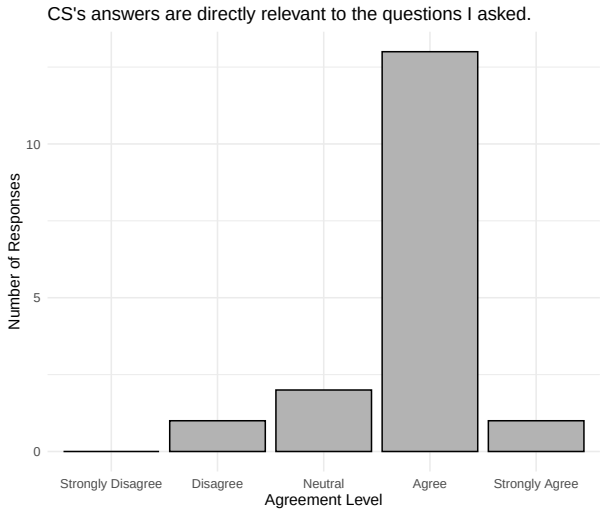

Figure 4.8.: CS's answers are directly relevant to the questions I asked (n=17).

For the second statement, customer agents were asked whether "CS provides complete answers without leaving out important information." None of the agents strongly disagreed, and 2 (11.76%) disagreed with the statement. 4 (23.53%) of the agents stayed neutral, 10 (58.82%) agreed, and 1 (5.88%) strongly agreed. The CS system is largely perceived as providing complete answers by the majority of agents. However, 11.76% disagreed, pointing to specific instances where the system failed to include important information. This highlights potential areas for improvement, such as refining the LLM that generates responses to ensure all relevant information is covered.

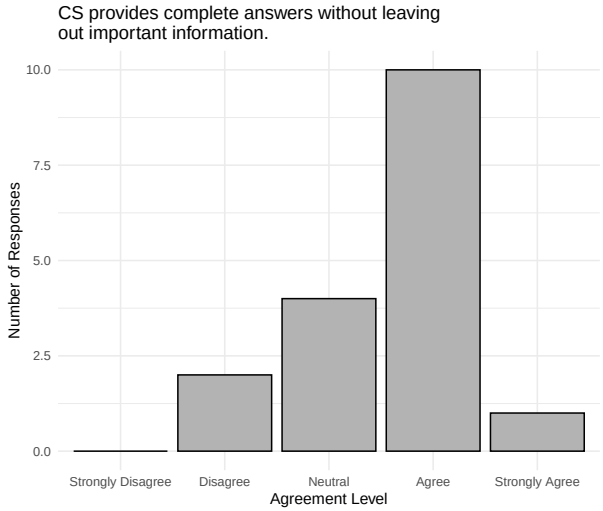

Figure 4.9.: CS provides complete answers without leaving out important information (n=17).

The last statement for this metric was, "CS's answers are free of unnecessary details and focus only on what is being asked." None of the agents strongly disagreed, but 3 (17.65%) disagreed, 5 (29.41%) of the agents stayed neutral, 6 (35.29%) agreed, and 3 (17.65%) strongly agreed. The results suggest that the system generally performs well in generating relevant and focused answers. However, a significant portion remained neutral, indicating that they either did not have strong opinions or found that the system's performance varied. This might suggest occasional inconsistencies in the system's responses. Moreover, some agents stated that they disagree with the statement, meaning there may be a need to further fine-tune and improve the LLM's ability to prioritize key information and avoid irrelevant content.

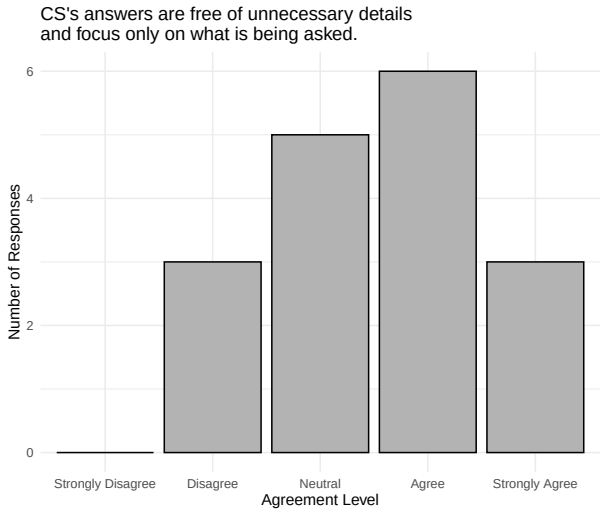

Figure 4.10.: CS's answers are free of unnecessary details and focus only on what is being asked (n=17).

The answer relevance metric has a mean score of 3.65, suggesting that customer service agents generally agree that the conversational search system provides relevant answers. However, the moderate standard deviation of 0.71 indicates some degree of variability, with some agents perceiving the answers' relevance differently than others.

Context Relevance

Another evaluated metric is context relevance, with the objective of assessing whether the conversational search retrieves and uses context appropriately. Customer agents gave their responses to the statement, "I feel that CS only uses information that is relevant to my request." None of the agents strongly disagreed. 3 (17.65%) disagreed, 7 (41.18%) of the agents stayed neutral, 6 (35.29%) agreed, and 1 (5.88%) strongly agreed. A significant portion of agents remained neutral, which may indicate that they have had both positive and negative experiences, making it difficult to decide and thus, they are unsure whether the system consistently maintains context relevance.

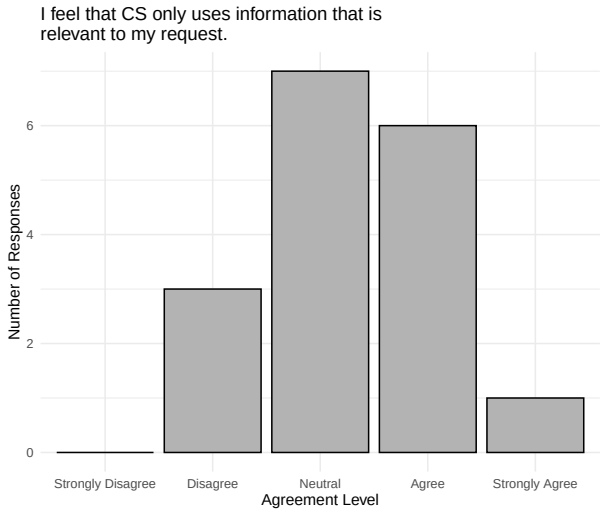

Figure 4.11.: I feel that CS only uses information that is relevant to my request (n=17).

The context relevance metric has a mean score of 3.29, indicating an average perception of the context relevance. The standard deviation of 0.88 shows moderate agreement among agents, which means most of the agents have similar views on this metric.

Satisfaction

Next, users' satisfaction was evaluated, with the objective of measuring how well the conversational search meets user expectations. The statement was, "CS met or exceeded my expectations in terms of functionality and performance." None of the agents strongly disagreed, and one (5.88%) agent disagreed. 6 (35.29%) of the agents stayed neutral, 9 (52.94%) agreed, and 1 (5.88%) strongly agreed. The results indicate that overall satisfaction is above

average, as over half of the agents believe CS meets their expectations. The high neutrality rate suggests that agents' experiences may be inconsistent, and the system sometimes meets their expectations but not always.

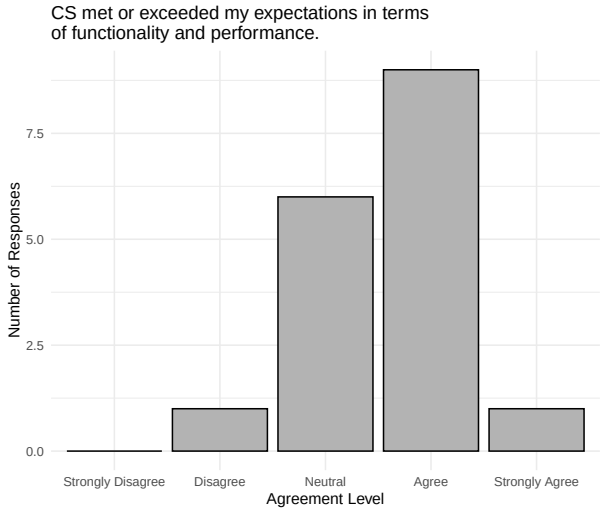

Figure 4.12.: CS met or exceeded my expectations in terms of functionality and performance (n=17).

The satisfaction metric has a mean score of 3.59, implying a generally positive perception regarding satisfaction. The moderate standard deviation of 0.71 indicates that while some agents share similar views on satisfaction, there are also agents who have different opinions.

Perceived Usefulness

For the metric of perceived usefulness, the objective was to evaluate how efficiently the conversational search helps complete tasks. Agents evaluated the statement, "Using CS allows me to complete tasks faster and improves my work performance." None of the agents strongly disagreed. 5 (29.41%) disagreed, 2 (11.76%) stayed neutral, 7 (41.18%) agreed, and 3 (17.65%) strongly agreed. More than half of the agents feel the CS system improves their task efficiency and work performance. Thus, the CS is a valuable tool for many agents. However, agents who disagree may be facing scenarios where the CS system doesn't provide significant time-saving or performance benefits. The interview results also showed that for certain types of queries, such as open-ended or long queries, CS was not identified as beneficial.

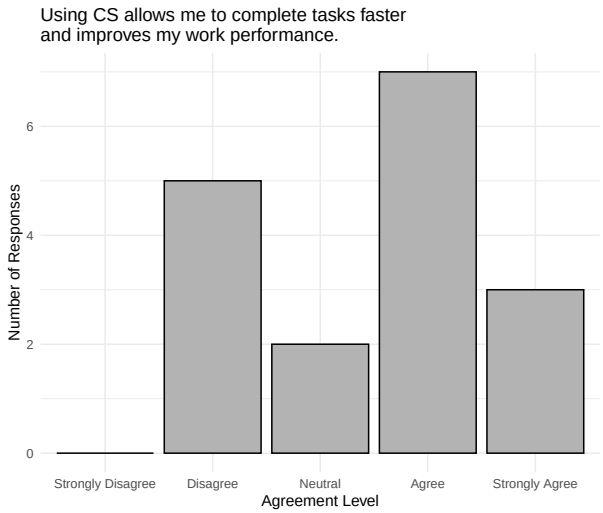

Figure 4.13.: Using CS allows me to complete tasks faster and improves my work performance (n=17).

The mean score of the perceived usefulness metric is 3.47, which is a moderate mean. The standard deviation of 1.57 is very high, reflecting significant variability in responses and, thus, differences in how useful agents perceive the CS.

Quality

Next, quality was evaluated, with the objective of ensuring the accuracy and reliability of the conversational search. Customer agents were asked to state their opinions on two statements.

The first statement was, "CS consistently provides accurate, correct, and reliable information." None of the agents strongly disagreed. 3 (17.65%) of the agents disagreed, and 3 (17.65%) stayed neutral. 9 (52.94%) agreed, and 2 (11.76%) strongly agreed. Most of the agents agree that the system provides accurate and reliable information, which is a strong indicator that most agents are satisfied with the quality of the information. However, even though it is a relatively low percentage, agents who disagreed and the neutral responses indicate that the system may sometimes provide inaccurate or incorrect information.

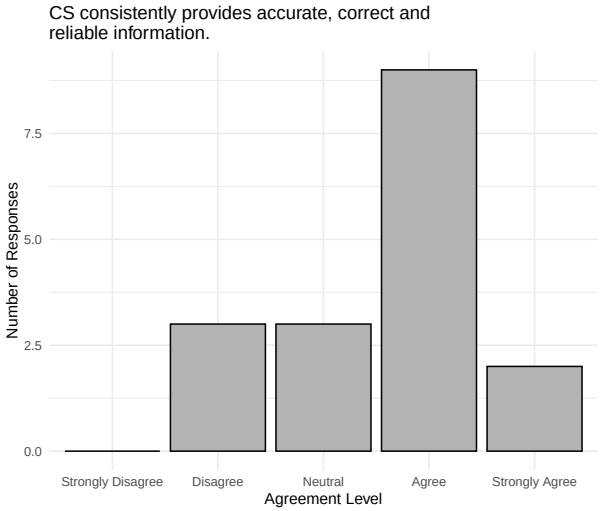

Figure 4.14.: CS consistently provides accurate, correct, and reliable information (n=17).

The second statement was, "CS's answers are structured to make it easy to understand and follow the information provided." None of the agents strongly disagreed. 2 (11.76%) disagreed, 4 (23.53%) stayed neutral, 8 (47.06%) agreed, and 3 (17.65%) strongly agreed. The majority of agents find the CS system's answers well-structured and easy to follow, suggesting general satisfaction with the structure. However, the agents who disagreed or were neutral indicate that there is still room for improvement to ensure all responses are clear and easy to follow. This could involve simplifying answer formatting by breaking down information into smaller parts or using techniques like colors or bolding key points, as suggested by agents during the interviews.

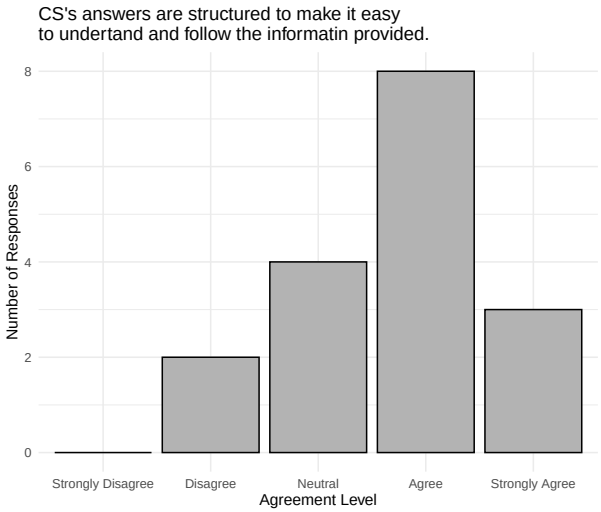

Figure 4.15.: CS's answers are structured to make it easy to understand and follow the information provided (n=17).

The mean score for this metric is 3.69, which implies an overall positive evaluation of the quality of the system. The standard deviation is 0.8, meaning the variability is moderate and opinions on this metric are not entirely consistent.

Business value

Another metric evaluated was business value. The goal was to assess whether the conversational search adds business value by saving time and resources.

"CS saves time and resources compared to today's keyword search" was the first statement that agents were asked to state their opinions on. None of the agents strongly disagreed. 3 (17.65%) disagreed, 4 (23.53%) stayed neutral, 8 (47.06%) agreed, and 2 (11.76%) strongly agreed. Over half of the agents believe the CS system saves time and resources compared to the traditional keyword search. This suggests that for most agents, the CS system is an efficient tool for several reasons, such as faster task completion. However, there are also a few agents who don't perceive the CS as a time-saving and efficient tool, possibly for reasons such as irrelevant or inaccurate results. The agents that stayed neutral suggest that their experience was mixed, they could be uncertain about the overall efficiency improvements, and they might have seen some benefits in time or resource savings but not consistently.

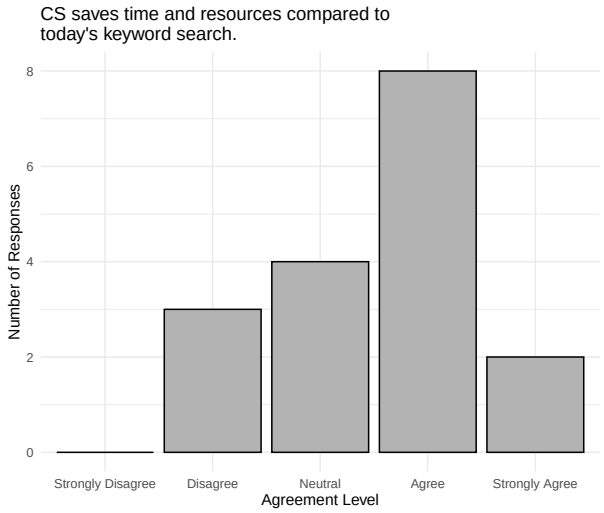

Figure 4.16.: CS saves time and resources compared to today's keyword search (n=17).

The second statement was, "CS's performance justifies its use as a business tool." None of the agents strongly disagreed. 1 (5.88%) disagreed, 4 (23.53%) stayed neutral, 8 (47.06%) agreed, and 4 (23.53%) strongly agreed. The majority of the agents agree that the CS system is generally seen as a valuable tool for business purposes. Only one agent disagreed with the statement, which is a very low percentage, showing that only a small number of agents did not find the system's performance satisfactory enough to justify its use in a business context. Moreover, the agents who stayed neutral might see potential but are unsure if the system

currently offers enough benefits to fully justify its use. However, if the system improves over time, these agents might change their minds.

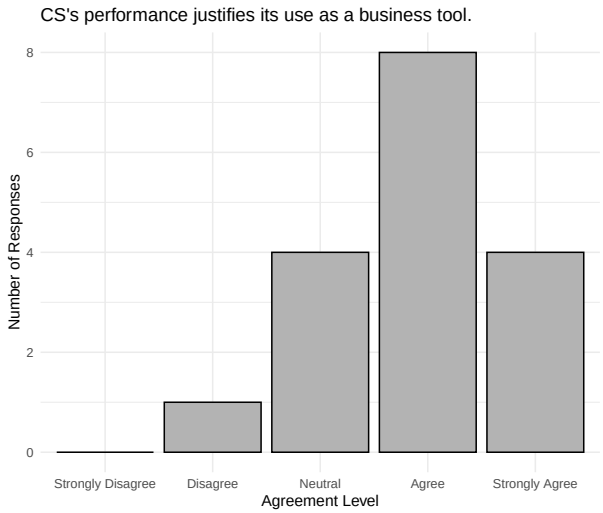

Figure 4.17.: CS's performance justifies its use as a business tool (n=17).

The business value metric has a mean score of 3.71, meaning that, generally, customer service agents perceive the system as beneficial for business purposes. The standard deviation was 0.89, with a moderate variability. This indicated that the agents have diverse opinions regarding the system's value to the organization.

Openness to new technologies

Additionally, customer agents' openness to new technologies was evaluated with the statement, "I am always open to trying new technologies as I believe they will expand my skills and support my professional development." None of the agents strongly disagreed or stayed neutral. 1 (5.88%) agent disagreed, while 8 (47.06%) agreed and 8 (47.06%) strongly agreed. Although the vast majority of the agents agreed that they are open to new technologies, interviews showed that habits and familiarity with the existing keyword search are strong factors that prevent the agents from fully adopting the CS. Agents feel comfortable using the keyword search as they know how to use it in an efficient way. This makes it difficult to adopt something new, even with the potential benefits of CS. This resistance could be because of the time required to learn about the new system or because of the uncertainty of its efficiency. To ease this transition, it is crucial to provide training and hands-on practice that will make agents understand CS's long-term benefits, such as time savings and increased productivity. This will help agents gain confidence in the new system and encourage adoption over time.

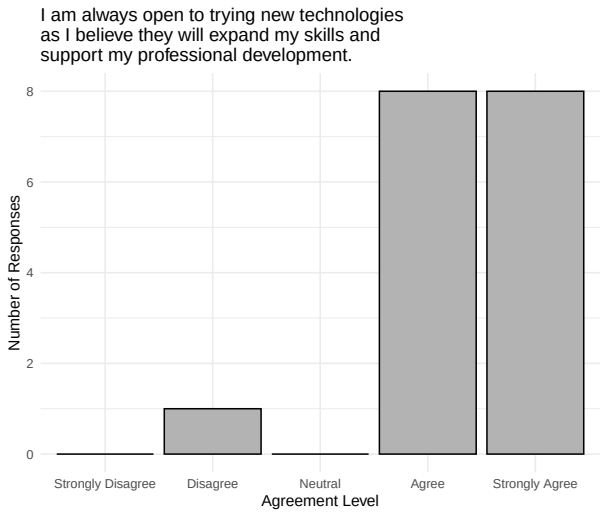

Figure 4.18.: I am always open to trying new technologies as I believe they will expand my skills and support my professional development (n=17).

The openness to new technologies metric has a mean score of 4.35, meaning that, generally, agents stated that they are open to new technologies. The standard deviation of 0.53 indicates a low variability, which means most of the agents share similar views.

Replaceability and necessity of CS

Lastly, the replaceability and necessity of CS were evaluated. The first statement was, "Today's keyword search can be replaced by CS." None of the agents strongly disagreed. 5 (29.41%) disagreed, 5 (29.41%) stayed neutral, 5 (29.41%) agreed, and 2 (11.76%) strongly agreed. The majority of the agents are not fully convinced that CS can replace the keyword search, possibly due to reasons such as comfort and familiarity they have with keyword search.

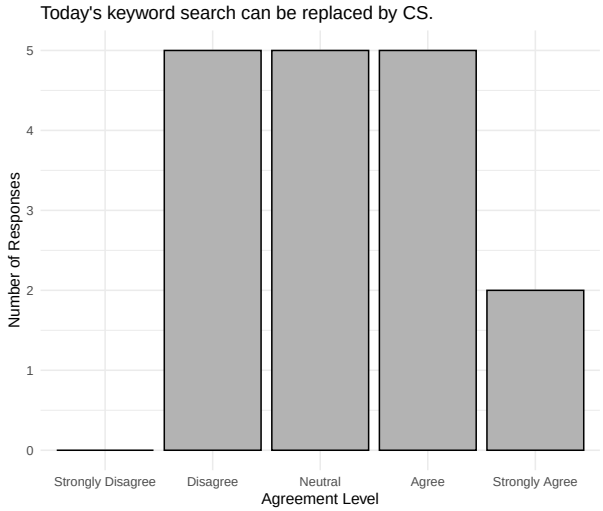

Figure 4.19.: Today's keyword search can be replaced by CS (n=17).

The second statement was, "CS is a useful extension but not absolutely necessary." Two agents (11.76%) strongly disagreed, 5 agents (29.41%) disagreed, 4 agents (23.53%) remained neutral, 6 agents (35.29%) agreed, and none of the agents strongly agreed. Over half of agents view CS as a useful but non-essential tool. This means they see it as a valuable tool, but it is mostly seen as a useful extension rather than a necessary tool, suggesting that agents do not yet view it as essential for their day-to-day work.

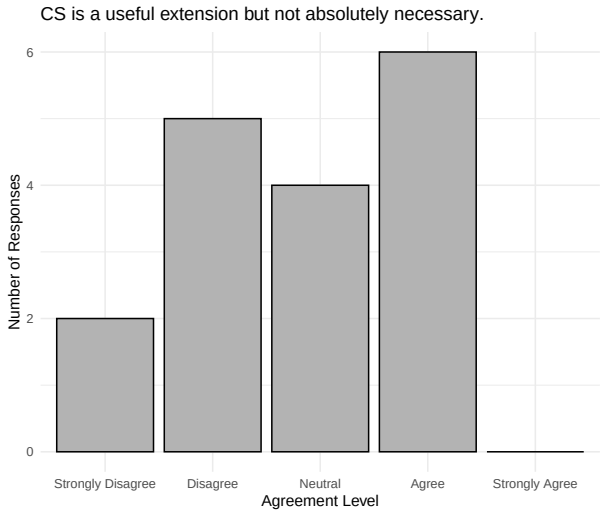

Figure 4.20.: CS is a useful extension but not absolutely necessary (n=17).

The Replaceability and Necessity of CS metric has a mean score of 3.27, suggesting a moderate agreement on the system's necessity and replaceability. The standard deviation is 0.95, which reflects a moderate variability with different views on the matter.

4.2.1. Correlation Analysis

Correlation Analysis Between Individual Characteristics

To assess the linear relationships between continuous variables (age, frequency of using a chatbot, openness to new technologies, and duration of CS use), a Pearson correlation test was conducted. Pearson's correlation coefficient (r) and associated p-values were computed to determine both the strength and statistical significance of these associations. Moreover, to examine the relationship between the binary variable gender and the continuous variables frequency and openness, Point-Biserial correlations were computed. The results showed that there are no significant correlations between the individual characteristics. Table 4.1 shows an overview of the results.

Variable 1	Variable 2	Correlation Coefficient
Age	Frequency	-0.17
	Openness	0.19
	Duration	-0.19
Frequency	Openness	0.45
	Duration	-0.02
Gender	Openness	-0.07
	Frequency	-0.19
Openness	Duration	-0.41

Table 4.1.: Correlation Results Between Individual Characteristics (n=12).
Note. *p < 0.05

Correlation Between Individual Characteristics and Choices

A possible correlation between the individual characteristics (such as age, openness, frequency, and duration) and the choice of using the CS or keyword search was examined. Pearson's correlation coefficient was computed to measure the strength and direction of the linear relationship between two continuous variables. To assess the statistical significance of the observed correlations, a correlation significance test (t-test for correlation) was performed.

It is found that there is a moderate positive correlation between choosing CS and openness to new technologies, suggesting that those who choose CS more frequently tend to be more open to new technologies. The result is statistically significant. In a similar vein, it is found that choosing the keyword search shows a moderate negative correlation with openness to new technologies, indicating that people who choose keyword search more often tend to be less open to new technologies. The correlation is statistically significant. No other significant correlation is found between choices and individual characteristics. Table 4.2 shows the overview of the results.

Next, we examined the correlation between each choice case and individual characteristics. The results indicated that there is a significant moderate positive correlation between choosing CS for an open-ended question that required a more detailed and thorough answer and

	Age	Openness	Duration	Frequency	Gender
Keyword Search	-0.095	-0.651*	0.185	-0.176	0.191
Conversational Search	0.201	0.636*	-0.237	0.230	-0.171
Undecided	-0.345	0.088	0.160	-0.169	-0.076

Table 4.2.: Correlation Results and Significance Levels Between Individual Characteristics and Choices (n=12).
Note. *p < 0.05

openness to new technologies. This moderate positive correlation suggests that generally the agents that chose CS to find the answer to an open-ended question, are more open to new technologies. Moreover, it was found that there is a significant moderate negative correlation between choosing CS for an open-ended question and the duration of CS usage. This implies that agents who chose to use CS for answering an open-ended question have generally been using the system for a shorter period of time. Between other choices and individual characteristics, there is no significant correlation. There is no correlation between choosing CS for a complex question and other variables as the standard deviation for the case of a complex question is 0. An Overview of the results can be found in Table 4.3.

	Simple Question	Complex Question	Close-end Question	Open-end Question	Short Question	Long Question	Procedural Question
Age	0.098		-0.150	0.157	-0.028	0.292	-0.310
Openness	0.331		-0.088	0.588*	0	0.133	0.523
Duration	-0.023		0.382	-0.653*	-0.165	0.014	0.127
Frequency	-0.064		0.338	0	-0.073	-0.192	0.275

Table 4.3.: Correlation Results and Significance Levels Between Choice Cases and Individual Characteristics (n=12).
Note. *p < 0.05

Correlation Between the Evaluation Metrics

Next, Pearson's correlation analysis was conducted between the different evaluation metrics, and the results showed that many metrics have significant positive correlations with each other. This aligns with other findings from the literature [70, 84].

Two correlations had a very strong positive correlation ($0.8 \leq r \leq 1.0$) with a very strong level of significance ($p < 0.001$), meaning the probability that these correlations occurred by chance is less than 0.1%. Four correlations had a strong positive correlation ($0.7 \leq r \leq 0.79$) with a very strong level of significance ($p < 0.001$), while two correlations had a very strong correlation ($0.8 \leq r \leq 1$) with a very strong level of significance. Table 4.4 shows all correlation values and their significance.

- **Perceived Usefulness and Business Value:** A very strong positive correlation exists between perceived usefulness and business value with a very strong significance. This implies that as users perceive the system as more useful, they also believe it contributes more value to the business.
- **Openness to New Technologies and Replaceability/Necessity of CS:** A very strong positive correlation with a very strong significance was found between openness to new technologies and replaceability/necessity of CS. This suggests that individuals who are more open to new technologies also tend to view CS as absolutely necessary.
- **Perceived Usefulness and Replaceability/Necessity of CS:** A strong positive correlation with a very strong significance exists between the perceived usefulness and necessity of CS. This implies that as users perceive the system as more useful, they also assess it as a necessary tool.
- **Answer Relevance and Perceived Usefulness:** A strong positive correlation with a very strong significance was found between answer relevance and perceived usefulness. This suggests that users who find the answers more relevant tend to view the system as more useful.
- **Ease of Use and Satisfaction:** A strong positive correlation with a very strong significance was found between ease of use and satisfaction. This suggests that the easier the system is to use, the higher the perceived satisfaction.
- **Ease of Use and Quality:** A strong positive correlation with a very strong significance was found between ease of use and quality. This indicates that users who find the CS easy to use, asses the system has higher quality.

	1	2	3	4	5	6	7	8	9	10	11
1. Ease of Use		0.582*	0.309	0.662**	0.441	0.772***	0.544*	0.748***	0.6*	0.099	0.373
2. Performance			0.452	0.632**	0.464	0.637**	0.53*	0.636**	0.631**	0.089	0.601*
3. Answer Faithfulness				0.244	0.396	0.221	0.488*	0.12	0.414	0.343	0.574*
4. Answer Relevance					0.539*	0.702**	0.782***	0.634**	0.692**	0.219	0.674**
5. Context Relevance						0.523*	0.435	0.263	0.21	0.303	0.439
6. Satisfaction							0.491*	0.7**	0.5*	-0.059	0.44
7. Perceived Usefulness								0.492*	0.824***	0.437	0.793***
8. Quality									0.586*	-0.045	0.188
9. Business Value										0.58*	0.836***
10. Openness to New Tech.											0.496*
11. Replaceability and Necessity of CS											

Table 4.4.: Correlation Results and Significance Levels Between the Evaluation Metrics (n=17). *Note.* *p < 0.05, **p < 0.01, ***p < 0.001

Correlation Between the Evaluation Metrics and Choices

Pearson's correlation analysis was conducted to find the correlations between different evaluation metrics and the choice of a search tool. To perform this, we aggregated the

responses across all seven choice cases and quantified the number of times a participant chose the CS option or the keyword search option. This allowed us to examine the overall relationship between the metrics and the likelihood of selecting a CS versus keyword search, based on the aggregated data.

Table 4.5 shows the results of the analysis. It was found that there is a very strong negative correlation with a very strong significance between ease of use and choosing keyword search. This implies that as the ease of use ratings for the CS decreases, the frequency of choosing the keyword search increases. Thus, agents who do not find the CS easy to use are more likely to choose keyword search instead. Moreover, there is a significant strong positive correlation between ease of use and choosing the CS. This indicates that the agents who find the CS easy to use tend to choose CS.

Furthermore, a significant moderate negative correlation was found between answer faithfulness and choosing keyword search, suggesting that agents who perceive the CS's answers as less faithful are more likely to choose keyword search instead of relying on the CS. On the other hand, there is a significant moderate positive correlation between answer faithfulness and choosing CS. This suggests that agents who perceive the CS's answers as more faithful are more likely to choose CS over keyword search.

Additionally, it was seen that there is a significant moderate negative correlation between satisfaction and choosing the keyword search, implying that the agents who are not satisfied with the CS tend to use the keyword search. On the contrary, there is a significant moderate positive correlation between satisfaction and choosing the CS, suggesting that the agents who perceive the CS as satisfying, tend to use it more.

Moreover, a strongly significant negative correlation was found between perceived usefulness and the keyword search. This indicates that the less the agents perceive the CS as useful, the more they tend to choose the keyword search. There was also a strongly significant positive correlation between perceived usefulness and choosing the CS, meaning the agents who perceived the CS as useful chose the CS more often.

In addition to that, a significant moderate negative correlation was found between business value and choosing the keyword search, which indicates that agents who don't perceive the CS as having a high business value tend to choose the keyword search over the CS. However, agents who believe the CS has a high business value, prefer using the CS. This was implied by the strongly significant moderate positive correlation between business value and the CS.

Lastly, there is a significant moderate positive correlation between the necessity of CS and the CS. This suggests that as agents perceive conversational search as more necessary, they choose it over the keyword search.

Evaluation Metric	Keyword Search	Conversational Search
Ease of Use	-0.737***	0.666**
Performance	-0.274	0.274
Answer Faithfulness	-0.544*	0.551*
Answer Relevance	-0.461	0.479
Context Relevance	-0.257	0.245
Satisfaction	-0.488*	0.521*
Perceived Usefulness	-0.616**	0.681**
Quality	-0.426	0.391
Business Value	-0.588*	0.635**
Openness to New Technologies	-0.423	0.424
Replaceability and Necessity of CS	-0.439	0.539*

Table 4.5.: Correlation Results and Significance Levels Between Choice Frequency and Evaluation Metrics (n=17).
Note. *p < 0.05, **p < 0.01, ***p < 0.001

Additionally, to obtain a more detailed understanding of the relationship between the metrics and the choice of search method, we conducted a separate correlation analysis for each of the seven choice cases. This allowed us to examine how the metrics influenced the decision to choose either the CS or keyword search in each specific scenario. Table 4.6 shows an overview of the results.

- **Simple Question:** Results showed that there is a significant moderate positive correlation between choosing CS and answer faithfulness for a simple question. This implies agents who rate the CS's answers as faithful are more likely to choose the CS for a simple question.
- **Complex Question:** For complex questions, ease of use, perceived usefulness, quality, and business value were found to have a significant moderate positive correlation with the choice of CS. Thus, agents who gave higher ratings for ease of use, perceived usefulness, quality, and business value tend to prefer CS over the keyword search for complex questions.
- **Open-ended Question:** It was found that there is a significant moderate positive correlation between choosing CS and ease of use for an open-ended question. This implies that the agents who rate the CS as easy to use are more likely to choose the CS for an open-ended question. Moreover, there is a significant moderate positive correlation between choosing CS for this scenario and answer faithfulness. Higher ratings for answer faithfulness indicate that the agents tend to choose the CS for an open-ended question. Additionally, there is a significant moderate positive correlation between the quality and choosing the CS. Agents who believe the CS has a high quality,

prefer using the CS for an open-ended question. Lastly, there is a strongly significant moderate positive correlation between choosing CS and business value. This suggests that higher ratings for business value are associated with a greater likelihood of choosing the CS for an open-ended question.

- **Short Question:** Results showed that there is a significant moderate negative correlation between choosing CS and performance for short questions. As the performance rating decreases, the likelihood of choosing the CS increases. Furthermore, it was seen that there is a significant moderate negative correlation between choosing CS and answer faithfulness. When the answer faithfulness rating decreases, the likelihood of choosing the CS increases. This suggests that when participants are dissatisfied with the CS accuracy or relevance, they may still choose it over the keyword search. Lastly, it was found that there is a significant moderate negative correlation between choosing CS and quality. As quality ratings decrease, the likelihood of selecting the CS increases. This suggests that agents who feel that the CS's quality is not so high may still choose the CS over the keyword search for a short question.
- **Procedural Question:** It was found that there is a significant moderate positive correlation between choosing CS and ease of use. This suggests that if participants find the CS easier to use, they tend to prefer it over the keyword search for a procedural question.

	Simple Question	Complex Question	Close-end Question	Open-end Question	Short Question	Long Question	Procedural Question
Ease of Use	0.169	0.5868*	0.473	0.5455*	-0.3734	0.2996	0.5855*
Performance	0.1995	0.2863	0.0569	0.4203	-0.5988*	0.3722	0.0797
Answer Faithfulness	0.5139*	0.0913	0.2711	0.5657*	-0.497*	0.3108	0.2213
Answer Relevance	0.2884	0.2781	0.1843	0.2793	-0.2916	0.2884	0.2512
Context Relevance	0.4439	-0.0913	0.2305	0.0926	-0.3965	0.1537	0.2128
Satisfaction	0.1933	0.3108	0.3627	0.3913	-0.1978	0.1933	0.1377
Perceived Usefulness	0.2513	0.4922*	0.0418	0.4575	-0.3132	0.4704	0.257
Quality	-0.0801	0.5491*	0.234	0.5249*	-0.5305*	0.426	0.2657
Business Value	0.1337	0.5235*	0.0045	0.6799**	-0.3024	0.2048	0.3482
Openness to New Technologies	0.3873	-0.0704	-0.2092	0.3364	0.1295	0.0738	0.3875
Replaceability and Necessity of CS	0.263	0.2232	0.0496	0.4733	-0.257	0.121	0.1367

Table 4.6.: Correlation Results and Significance Levels Between Choice Cases and Evaluation Metrics (n=17).

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3. Findings from Customer Agent Log Analysis

In the database containing agent logs, a total of 508 queries could be found. The queries were entered into the CS tool between February 2024 and November 2024.

First, out of 508 queries, a total of 400 queries were analyzed and divided into seven categories: simple, complex, open-ended, close-ended, short, long, and procedural queries.

Each query belonged to multiple categories. Table 4.7 gives an overview of the categorization. 243 queries (60.75%) were categorized as simple, 157 (39.25%) as complex, 159 (39.75%) as open-ended, 241 (60.25%) as close-ended, 212 (53%) as short, 188 (47%) as long, and 7 (1.75%) as procedural. Furthermore, it was observed that out of the 400 queries, 387 (96.75%) were full sentences, while 13 (3.25%) were keyword search-like queries and were not complete sentences. This aligns with the interview results that the CS is preferred more for short and close-ended queries.

Scenarios	**Number of Queries**
Simple Query	243 (60.75%)
Complex Query	157 (39.25%)
Open-ended Query	159 (39.75%)
Close-ended Query	241 (60.25%)
Short Query	212 (53%)
Long Query	188 (47%)
Procedural Query	7 (1.75%)

Table 4.7.: Categorization of Queries.

Next, the ratings that agents gave to the answers generated by the CS were analyzed. Out of the 508 queries, there were 503 ratings for the answers. Table 4.8 gives an overview of the results. On a scale of 1-5, with 5 meaning the agent was fully satisfied with the answer, 251 questions (49.90%) had a rating of 1, 22 questions (4.37%) had a rating of 2, 39 questions (7.75%) had a rating of 3, 10 (1.99%) questions had a rating of 4, and lastly, 181 questions (35.98%) had a rating of 5. It was seen that most agents were either satisfied with the answer or not, with little middle ground. On the other hand, the evaluated survey metrics generally had a more balanced and even distribution across the scale.

Rating	**Number of Answers**
1	251 (49.90%)
2	22 (4.37%)
3	39 (7.75%)
4	10 (1.99%)
5	181 (35.98%)

Table 4.8.: User Ratings for Answers Generated by the LLM.

Moreover, agents also rated the 3 document suggestions provided by the CS search by giving a thumbs up or a thumbs down. Out of the 508 queries, there were document ratings for 495 queries. It was found that for 292 (58.99%) of the queries, all 3 document suggestions were evaluated with a thumbs down, and thus, agents were not satisfied with the document suggestions. On the other hand, for 203 (41.01%) questions, at least one document was evaluated with a thumbs up. If at least one document was evaluated with a thumbs up, it

indicated a success rate of 100% for that query, as the agents could find the answer to their questions in that document.

After analyzing the logs, we find that the ratings and thumbs-up/thumbs-down results do not fully align with the interviews or survey results. For five out of seven scenarios we asked in the interviews, the majority of the agents preferred conversational search over keyword search. Moreover, during the interviews, even though there were also agents who stated that the accuracy of the responses was not always 100% correct, the majority of the agents stated that the CS significantly eased the process of finding the information they needed and increased their efficiency. Also, during the survey, more than half of the agents stated that the CS system improves their task efficiency and work performance. While the mean scores from the survey results showed an overall moderately positive perception of CS, 58.99% of document suggestions receiving all thumbs down and 62.02% of documents having a rating of 1, 2, or 3 do not align with the moderately positive perception.

Furthermore, agents' ratings on a scale of 1-5, as well as the thumbs up or thumbs down that they gave to the documents were analyzed over time. The beginning of September 2024 was a breakpoint as, after that date, more agents started using the CS. The analysis in R also supports this. Possibly, due to most users being on vacation in August 2024, as shown in Figure 4.21, very few documents were rated in August (2 documents). Thus, the low rating percentages and good documents percentages were because of the low engagement in August. The massive increase in rated documents in September 2024 (234 documents) was likely due to users coming back after vacation and new users joining at the same time. However, in October and November, the number of rated documents declined, suggesting that some new users didn't stay active for long or their initial excitement faded.

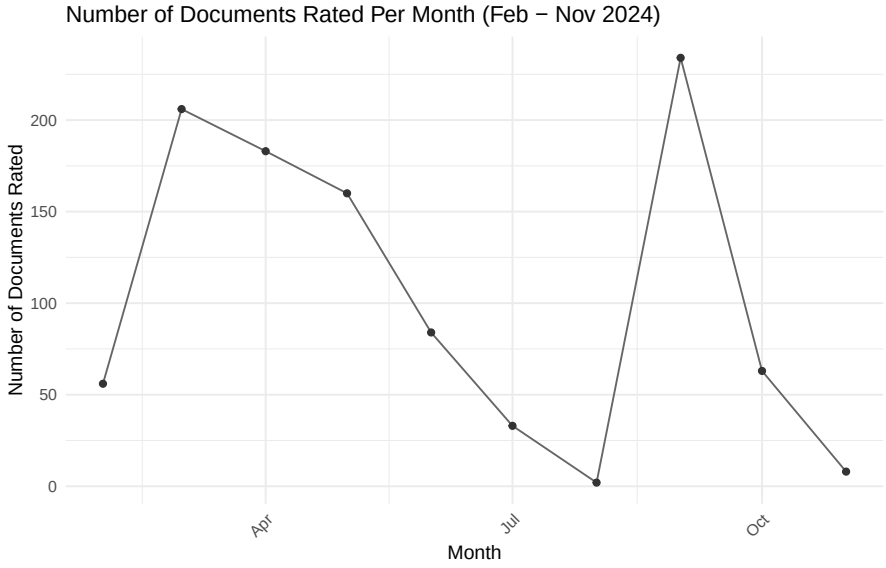

Figure 4.21.: Number of Documents Rated per Month (February-November 2024).

Figure 4.22 and Figure 4.23 show that before September 2024, the average rating on a scale

of 1-5 and the percentage of documents rated as good were lower, possibly indicating less document quality or a more experienced user base that was more critical in their evaluations. After the breaking point occurred, when new users started using the system, the average rating increased. This suggests that either the quality of documents improved or new users initially rated more positively than existing users. After September, the number of rated documents significantly dropped.

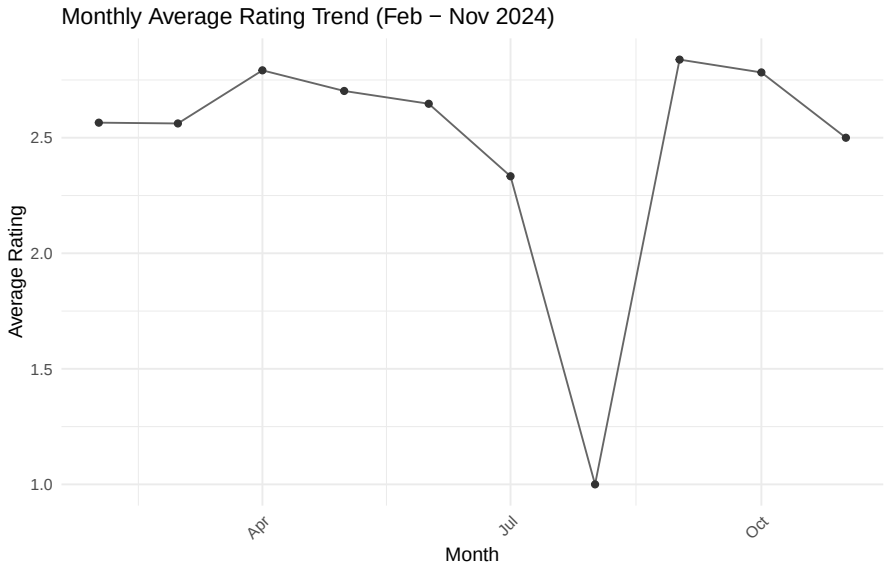

Figure 4.22.: Monthly Average Rating Trend (February-November 2024).

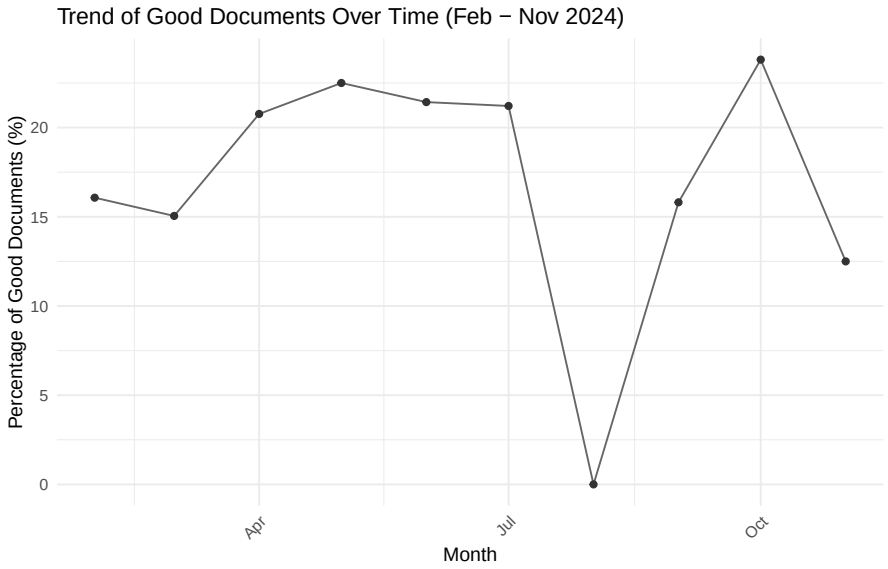

Figure 4.23.: Trend of Good Documents over Time (February-November 2024).

5. Discussion

5.1. Discussion of the Interview Results

Scenario Results Consequently, it was seen that the CS is favored by the agents for complex, close-ended and short questions with a very low standard deviation, suggesting that many agents shared the same opinion. As Pham et al. have explained, the time and effort that individuals had to invest to gather information decreased, and now they don't need to "sift through vast amounts of data to find relevant results. AI-driven algorithms now provide users with highly targeted and contextually relevant information, reducing the time and effort required for effective searches" [85]. Agents value that the CS is able to give direct and precise answers in the case of these scenarios, and they can get a quick response without sifting through the documents in the database. This supports findings from Pham et al., which showed that features such as speed and accuracy contribute to a more efficient and satisfying search experience.

For the other four scenarios (simple, open-ended, long, procedural questions), the standard deviation was high, and the agents had different opinions on which tool they preferred. Several factors played a role in their choice. It was found that familiarity with the tool plays a big role for agents when choosing between the CS and keyword search. Agents tend to rely on their well-established search habits even if the new tool could potentially be more efficient.

Importance of Habits Polites explains that many organizational tasks are repetitive. While not all frequently repeated organizational behavior is habitual, depending on the complexity and the nature of the task, there could potentially be a habitual behavior [86, 87]. Alsharo et al. state that habits are a way of "achieving a certain goal or accomplishing a certain task depending on the perception of the positive consequences of having performed the same actions in the past within the same environment" [87].

The agents who know that they can reach the desired results by using the keyword search, as they have used it many times before and are sure they would find the right information there, tend to choose the CS less. This choice is driven by the habits they have established and not by the efficiency that they perceive. During the interviews, an agent mentioned that the biggest strength of the traditional keyword search is the habit, as they don't know it any other way and have already settled into it. They mentioned that this is the only thing they can think of that makes the keyword search better than the CS. They added that the CS would be a more beneficial tool if one day it becomes a habit like the traditional keyword search. Moreover, in one of the interviews, an agent who preferred the keyword search over the CS in the majority of the scenarios couldn't think of any strengths when they were asked about

the strengths of the keyword search. This also can be interpreted as the strength for them was actually the habit. Overall, we can say that the agents who are more familiar with the keyword search are slower to adopt the CS.

On the other hand, if the agents aren't familiar with the query, are uncertain about the exact keywords to use, and don't know where to find the information using the traditional keyword search, they prefer using the CS as they deem it more efficient and think they will be faster in that case. This suggests that newly hired customer service agents are more likely to prefer using the CS, as they are not familiar with the keyword search. As agents mentioned, while using the keyword search if they do not know the exact term they should use to search for something, they may not find it. Therefore, the CS serves as a better starting point. It can be concluded agents agree that the speed of finding the correct information depends on their familiarity with the topic.

The role of trust Another key factor for the agents' choice was their trust in the CS tool. Especially for open-ended and long queries, agents lacked trust in the accuracy of the CS, as it is a new tool, and wanted to use the traditional keyword search methods where the information is verified and 100% correct. For detailed questions, agents feel more confident using the keyword search as they can validate the multiple aspects of a query using official documents. Several agents mentioned that even if the CS could give perfect answers very quickly, they would still not trust it with detailed questions, as opposed to one agent who stated that they are now hesitant but if they saw that the CS is working well with those kind of queries, then they would start using the CS. Currently, even if they use the CS, they cross-check the results using the keyword search. The trust factor aligns with findings from many studies that emphasize trust as a key factor in user acceptance of chatbots [5, 70, 73, 88, 89]. Features such as confidence scores for responses or clear explanations of how results are generated could help agents assess the reliability of the CS's outputs.

Strengths and Limitation of the CS Agents acknowledged that the CS has several strengths such as easing the process of finding the correct information, creating well-formulated summaries and answers, getting link recommendations practically, avoiding the keyword search that provides a large number of results, and getting comprehensive results that address various aspects of a query at the same time. However, the CS tool also has some limitations, such as not including the older policies or not giving 100% accurate results, that agents hope will be fixed.

Additionally, agents mention that in most cases, the CS's performance is related to how detailed they type their queries. Ambiguous or incomplete queries cause the system to return irrelevant results and make the process of finding the correct information slower. It was observed that agents have different strategies for formulating queries for the CS. While some agents use longer and well-phrased questions for better accuracy, others prefer shorter queries for speed and efficiency, and they believe that it is inconvenient to write full sentences. This suggests that the effectiveness of the CS is partly dependent on the agents' query formulating strategy, and educating agents on best practices for phrasing queries in CS may improve its

effectiveness.

Overall, it can be said that agents are aware that the newly integrated CS system is still in its early stages and has a big potential. Many agents hope for targeted improvements, such as extending the database with older policies, refining the interface, and optimizing search accuracy. They are optimistic that with time, by overcoming the limitations and adding some new functionalities, such as voice-based queries and the ability to ask follow-up questions, the CS's reputation will be increased and it will be a significantly valuable tool for customer service operations in the long term.

5.2. Discussion of the Survey Results

Evaluation of the metrics The findings from the survey showed that the evaluation of the conversational search reveals mixed opinions on different metrics. One of the key strengths of the CS was its perceived ease of use. Most of the customer agents agreed that the CS is user-friendly and effectively helps them locate information without requiring additional support. Its interface is easy to navigate and requires minimal effort, suggesting that CS has successfully met agents' usability expectations.

For the metrics answer relevance and quality, agents generally agreed that CS provides relevant, complete, structured, and accurate responses. However, they noted that sometimes there are unnecessary details in the answers. Moreover, for the metrics answer faithfulness and satisfaction, a slightly above-average mean was found. Most agents agreed that CS meets their expectations, but they indicated that there are occasionally irregularities in CS's responses matching the retrieved context.

When it comes to context relevance and perceived usefulness, agents showed moderate agreement, with context relevance having the lowest mean among the evaluated metrics, followed by perceived usefulness as the second lowest. Agents weren't 100% convinced that the CS allowed them to complete their tasks faster, implying that there is room for improving its impact on user productivity. Agents mentioned in their interviews that speed plays a critical factor during phone calls with customers. Thus, when agents don't think they are faster with the CS, they do not tend to use it. This also aligns with the statement of a customer service agent who mentioned that they use the CS less for phone calls and more for post-processing where speed is not a critical factor.

Agents' evaluation for metric performance was mixed. While some agents found CS stable and coherent, others expressed concerns about its ability to quickly handle unusual requests and deliver consistent, relevant answers. This indicates that while CS performs well under standard conditions, its reliability under more complex scenarios needs to be improved.

Regarding the business value, most of the agents realized that the CS has the potential to save time and resources that would justify it as a business tool, however, some were neutral or hesitant about its long-term justification.

The metric openness to new technologies had the highest mean among the metrics, indicating that the agents are open to new technologies and most agreed that tools like CS can expand their skills and professional development. Despite this enthusiasm, feedback

on replaceability and necessity showed skepticism about the capability of CS to completely replace the traditional keyword search. Many agents considered CS as a useful extension but not a vital tool, which shows uncertainty about its positioning in existing workflows. This suggests that the role of CS will remain complementary unless further improvements are made.

Overall, it can be said that the mean values suggest a moderate to slightly positive trend. With time, improvements to the tool, and increased familiarity of the customer agents with it, could lead to higher mean values.

Correlation Analysis Results The findings also highlighted several correlations that helped us to understand the factors influencing the adoption of the CS. Of 11 metrics, 7 were found to influence agents' choice of CS. The metrics ease of use, answer faithfulness, satisfaction, perceived usefulness, business value, replaceability/necessity, and openness to new technologies were the metrics that had a significant positive correlation with choosing the CS.

Moreover, we discovered that when the agents are more open to new technologies, they are also likely to use the CS for open-ended questions. During the interviews, some agents were hesitant to use the CS for an open-ended question requiring a detailed answer, as they valued their familiarity with the old tool more and trusted it more to find the correct information. Additionally, the negative correlation between CS usage duration (i.e., how long they have been using CS) and choosing CS for open-ended questions suggests that long-term users may become less reliant on CS for open-ended queries. This could indicate that as users gain more experience with CS, they develop a better understanding of its strengths and limitations, leading them to select traditional keyword search methods for certain tasks. Since even experienced users ultimately shift away from using CS for open-ended questions, some agents' hesitations and lack of trust may not be entirely unjustified. However, we believe that rather than relying solely on assumptions, agents should have tried the tool themselves and evaluated its strengths and limitations firsthand before coming up with a conclusion.

While CS was not preferred for queries that require detailed answers, it was favored for short questions. The findings showed that when performance, answer faithfulness, and quality decrease, the usage of CS for short questions increases. Although this seems contradictory at first, the reason for it might be that the CS might still be faster and more convenient than the keyword search for retrieving basic information. In contrast, longer or open-ended queries might require more accuracy, leading users to prefer traditional keyword search. For short, simple questions, users might not expect a perfect response, or they just want to validate the information they already have in their minds.

Log Analysis Results The results of the log analysis showed that there is a misalignment between the logs results and the results from the surveys/interviews. This might be due to several factors. This supports the findings from the literature that interaction data and user perceptions can be misaligned, emphasizing the need for both objective measures (like interaction logs) and subjective measures (like surveys): "Interviews and questionnaires are indispensable to measure the users' judgments of the system, as interaction data will not

necessarily reflect the users' perceptions" [90, 91, 92]. User perception might be different than the logs because of cognitive biases, such as the peak-end rule. This rule suggests that users' memory of interaction is disproportionately shaped by the most intense moment (peak) and the final impression, rather than the entire experience [90, 93]. As a result, users may overestimate or underestimate the overall quality of an interaction. Moreover, the Think Aloud method, which requires users to verbalize their thoughts during or after an interaction, can be influenced by memory biases [90]. Users may remember their conversation better or worse than it actually was, leading to discrepancies between objective interaction data and subjective self-reports.

However, it is also possible that the misalignment was due to the sample size for the survey or possible bias in who took part in the surveys or interviews. If only certain agents completed the survey (e.g., those who were more engaged or those with more neutral opinions), the results would not fully reflect the experience of all agents. Those who were highly dissatisfied may not have participated. Moreover, another possible explanation is recency bias in the surveys and interviews [94]. The logs include ratings from February to November 2024, meaning agents rated the system while it was still new and potentially improving. The surveys and interviews were conducted at the end of November 2024 and the beginning of December 2024. By December, there might have been a more favorable impression after agents had more time to adapt to the tool and any improvements had been implemented. Thus, agents may have weighed their more recent experiences more heavily. Additionally, in Table 3.1, it was seen that 92,31% of the interview respondents were using the CS for less than 3 months, and their experience was based on the later, improved version of the system. Thus, responses were biased toward newer users.

6. Conclusion

6.1. Summary

With the introduction and integration of AI in search engines, how we search for information has changed. Unlike traditional keyword search engines, which mainly rely on keyword matching, these advanced systems understand the context and intent of user queries, providing more relevant and customized results [85]. It is important to understand the factors that lead users to adopt this new way of information search. For this purpose, we conducted quantitative and qualitative analysis through interviews and surveys with customer service agents to investigate their adoption of a newly integrated conversational search tool. The CS was integrated into the existing knowledge management software in the context of an European insurance company. It was not a substitute for the existing keyword search but a complementary tool. We identified scenarios and reasons for agents' preference for conversational search over traditional keyword search. We found ways to evaluate conversational search and assess the benefits and challenges of using conversational search. The detailed findings presented in the previous chapters will now be utilized to address the stated research questions.

6.2. Findings

6.2.1. RQ1: What factors influence customer service agents' choice between the conversational search and the traditional keyword search?

The findings from the interviews and correlation analysis showed that multiple factors influence agents' choice of search tool.

- **Familiarity with the Tool and Confidence in the Existing Knowledge**
  During the interviews, it was discovered that agents' familiarity with the tool and their trust and confidence in their existing knowledge play a very important role in their choice of search tool. Agents stated that if they don't know where to find the information and aren't familiar with the topic, then they would start searching by using the CS. However, if the agents are familiar with the topic and know where to find certain information by using the keyword search, then they prefer using the keyword search because they feel more comfortable with their old habits.

- **Trust**
  Agents' trust in the tool was another factor in their choice. It was seen that some agents

are hesitant to fully trust the CS due to it being a new tool. Especially for open-ended and long queries, there was a lack of trust in the accuracy of the CS's responses. On the other hand, keyword search is seen as a more reliable option since the information there is verified. Currently, even if agents use the CS, they also tend to cross-check the results using the keyword search.

- **Type of Query**
  The query type is also a factor that determines the agents' search tool preference. While the vast majority of agents preferred CS for complex, close-ended, and short questions, there was no strong preference for one tool over the other when it came to simple, open-ended, long, or procedural questions. For these four questions, agents tend to rely on what they already know and trust rather than solely evaluating the tool's effectiveness. The reason for choosing CS for complex, close-ended, and short questions was to get direct and precise answers quickly without going through multiple steps and vast documents using the keyword search.

- **Speed and Time Pressure**
  Speed and time pressure also played a role in agents' preferences. Some agents mentioned that they prefer the keyword search for live customer calls where they are under time pressure. They believe that they are faster using the keyword search due to reasons such as their familiarity with it and the fact that they don't have to write a full sentence but only one keyword. They further mentioned that it takes more time for them to formulate and type the question to CS. However, there were also agents stating that even if it takes more time for them to formulate and type the question to CS, the CS would provide the correct result faster as it provides multiple suggestions that are relevant. Moreover, agents stated that while on a phone call that requires speed, they prefer the keyword search if they have an idea where to find the answer. If they are not familiar with the topic or if they already know the answer but only want to validate, then they use the CS.

- **Openness to New Technologies**
  By self-assessment, agents who stated that they are open to new technologies to support their professional development tend to choose conversational search over traditional keyword search.

- **CS Evaluation Metrics in a General Context**
  According to the correlation analysis, agents' preference for conversational search was influenced by perceived ease of use, answer faithfulness, satisfaction, perceived usefulness, and business value. Agents were more likely to prefer CS if they found the interface easy to understand and navigate, believed its answers generally matched the retrieved information, were satisfied with its functionality, and felt that CS saved time and resources compared to the keyword search. These insights highlight the need for continued improvements in those metrics to further encourage agents to transition from keyword search to conversational search.

- **CS Evaluation Metrics Across Different Question Types**
  The impact of evaluation metrics also varied depending on the type of question. For simple questions, agents who assessed the answers of the CS as faithful preferred CS over the keyword search. When dealing with complex questions, multiple factors came into play, including ease of use, perceived usefulness, quality of responses, and business value. Similarly, for open-ended questions, multiple factors determined whether agents opted for CS: ease of use, answer faithfulness, quality, and business value. An interesting trend emerged with short questions. Even when CS's performance, faithfulness, and quality ratings declined, agents still tended to prefer CS over keyword search. This suggests that for short queries, convenience and speed may outweigh concerns about answer accuracy. Meanwhile, for procedural questions, the primary factor influencing agent preference was how easy they found the system to use. This indicates that when step-by-step guidance is needed, a user-friendly interface is crucial.

6.2.2. RQ2: How can an LLM-based conversational agent be evaluated?

First, through qualitative interviews, we gathered deeper insights into agents' perceptions of CS, its strengths, limitations, and additional functionalities they would find useful.

Furthermore, after analyzing the existing literature, we identified 11 metrics to evaluate an LLM-based conversational agent. Each metric was measured with one or more statements that were added to the survey as Likert-scale questions. Mean and standard deviations for each metric and statement were calculated. The identified metrics were perceived ease of use, performance, answer faithfulness, answer relevance, context relevance, satisfaction, perceived usefulness, quality, business value, openness to new technologies, and replaceability and necessity of CS.

Moreover, to gain deeper insights we conducted a correlation analysis. It was seen that many metrics were positively correlated with each other. The most significant correlations were between ease of use and satisfaction, ease of use and quality, answer relevance and perceived usefulness, perceived usefulness and business value, and perceived usefulness and replaceability and necessity of CS. With the help of correlation analysis, we also identified that metrics ease of use, answer faithfulness, satisfaction, perceived usefulness and business value

Additionally, we analyzed interaction logs, including agent ratings and thumbs-up/thumbs-down feedback on CS responses, to assess real-world usage patterns. However, findings showed some misalignment between interview responses and log data, suggesting differences between stated perceptions and actual behaviors.

Thus, the research question was addressed through a multi-method approach to evaluate the effectiveness and adoption of the LLM-based conversational agent, by combining interviews, survey data, correlation analysis, and interaction logs. The key takeaways can be summarized as:

- **Qualitative Insights:** Interviews provided deeper understanding of agents' experiences, trust issues, and desired functionalities, complementing quantitative data.

- **Evaluation Framework with Key Metrics:** The study identified 11 key metrics (e.g., ease of use, performance, satisfaction) for evaluating the CS, measured using Likert-scale survey questions. Survey questions provided a structured way to measure CS's effectiveness.
- **Correlation Analysis:** Correlation analysis showed that several metrics were positively correlated. This highlights that improving one factor can enhance overall adoption. Moreover, correlation analysis helped identifying the metrics that played a bigger role in agents' choice of using the CS instead of the traditional keyword search. Furthermore, correlation analysis can determine the correlation between the individual characteristics of users and the adoption. Overall, it is an important step to gather more insights on the users' adoption.
- **Analyzing the Difference Between Perceptions and Behavior:** Log data analysis showed that agents' actual use of CS did not always align with their stated preferences. This suggests a gap between what users claim and how they interact with the system in practice, highlighting the need for both subjective and objective evaluation methods.

6.2.3. RQ3: What are the benefits and challenges of adopting LLMs in existing knowledge management systems after integration?

The adoption of LLM-based conversational search within existing knowledge management systems has both benefits and challenges. One of the key benefits is time efficiency. While using the keyword search, agents have to sift through multiple links, and they easily get lost between too many documents. CS eases the process of finding information by providing only relevant links and answers that directly address the query. Moreover, for complex questions that require comprehensive results addressing various aspects of a query at the same time, CS is particularly helpful. Otherwise, agents had to search separately for different areas. Furthermore, the formulated answers that LLM generates are particularly helpful, making it easier for users to apply the retrieved information directly, instead of formulating it by themselves.

However, despite the advantages, there are also several challenges that hinder the adoption. A major barrier is user resistance to change, as many individuals are deeply accustomed to the traditional keyword search and hesitant to shift to a new system. Additionally, trust issues play a significant role. Users may question the accuracy and reliability of AI-generated responses, preferring familiar search methods where they have control over the sources they select. Lastly, openness to new technologies varies among users, with some agents reluctant to embrace AI-driven tools due to skepticism.

Consequently, we can say that while LLM-based CS offers a more practical, efficient, and comprehensive approach to knowledge retrieval, its successful adoption depends on overcoming user resistance, building trust, and encouraging openness to adopting new technologies.

6.3. Limitations and Future Work

The limitations of the study include:

- **Sample Size:** The study's findings may not be fully generalizable due to a limited sample of customer service agents. A larger, sample could lead to different insights.
- **Scope of the Study:** The research focused specifically on conversational search within knowledge management, which may limit its applicability to other types of conversational AI used in different contexts.
- **Survey and Interview Instrument:** The interview and survey questions may not have captured all relevant aspects of user preferences, experiences, and potential concerns, leading to gaps in understanding agent perceptions.
- **Early Adoption Phase:** Since the integrated CS is still in a test phase, agent preferences, and attitudes may evolve over time as they become more familiar with the system. Initial skepticism or reluctance may decrease with increased usage.

Our suggestions for future research include:

- **Expanding the Sample Size:**: Future studies should include a larger group of users to enhance the generalizability of the findings and ensure a more comprehensive representation of different user perspectives.
- **Exploring Other Conversational AI Technologies:** Broadening the scope to include various types of CS beyond CS in knowledge management could provide deeper insights into user preferences and adoption patterns across different applications.
- **Broader Interview and Survey Instrument:** More detailed interview and survey instruments may be created to capture the full range of user preferences and experiences with conversational search.
- **Conducting Follow-up Studies:** Follow-up research should track how user attitudes and preferences evolve over time as agents gain more experience with CS and as the technology improves, providing a better understanding of long-term adoption.

A. Customer Agents Survey

A.1. Customer Agents Survey Questions in English

1. Please give your feedback on CS's perceived ease of use.
CS helps me find the information I am looking for without needing additional support.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS's user interface is easy to understand and requires minimal effort to use.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

2. Please give your feedback on CS's performance.
CS remains stable and functional when faced with unusual requests.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS's answers are consistent and logically connected to my questions.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS efficiently delivers fast and relevant answers without unnecessary steps or delays.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

3. Please give your feedback on CS's answer faithfulness.
I noticed cases where the response didn't match the context retrieved.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
In most cases, CS makes statements that are supported by the information retrieved.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

4. Please give your feedback on CS's answer relevance.
CS's answers are directly relevant to the questions I asked.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS provides complete answers without leaving out important information.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS's answers are free of unnecessary details and focus only on what is being asked.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

5. Please give your feedback on CS's context relevance.
I feel that CS only uses information that is relevant to my request.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

6. Please give your feedback on CS's satisfaction.
CS met or exceeded my expectations in terms of functionality and performance.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

7. Please give your feedback on CS's perceived usefulness.
Using CS allows me to complete tasks faster and improves my work performance.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

8. Please give your feedback on CS's quality.
CS consistently provides accurate, correct, and reliable information.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS's answers are structured to make it easy to understand and follow the information provided.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

9. Please give your feedback on CS's business value.
CS saves time and resources compared to today's keyword search.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS's performance justifies its use as a business tool.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral

[ ] Agree
[ ] Strongly Agree

10. Please give your feedback on CS's openness to new technologies.
I am always open to trying new technologies as I believe they will expand my skills and support my professional development.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

11. Please give your feedback on replaceability and necessity of CS.
Today's keyword search can be replaced by CS.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree
CS is a useful extension but not absolutely necessary.
[ ] Strongly Disagree
[ ] Disagree
[ ] Neutral
[ ] Agree
[ ] Strongly Agree

12. Do you have any further feedback about the CS?

A.2. Customer Agents Survey Questions in Original Language German

1. Bitte geben Sie ihr Feedback zur konversationelle Suches (KS) Benutzerfreundlichkeit
KS hilft mir, die gesuchte Information zu finden, ohne zusätzliche Unterstützung zu benötigen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
Die Benutzeroberfläche von KS ist leicht zu verstehen und erfordert nur minimalen Aufwand in der Nutzung.

[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

2. Bitte geben Sie Ihr Feedback zur Leistung von KS.
KS bleibt stabil und funktionsfähig, wenn sie mit ungewöhnlichen Anfragen konfrontiert wird.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
Die Antworten von KS sind konsistent und logisch mit meinen Fragen verbunden.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
KS liefert effizient schnelle und relevante Antworten ohne unnötige Schritte oder Verzögerungen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

3. Bitte geben Sie Ihr Feedback zur Antworttreue von KS.
Ich habe Fälle bemerkt, in denen die Antwort nicht mit dem abgerufenen Kontext übereinstimmte.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
In den meisten Fällen macht KS Aussagen, die durch die abgerufenen Informationen gestützt werden.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu

[ ] Stimme voll und ganz zu

4. Bitte geben Sie Ihr Feedback zur Antwortrelevanz von KS.
Die Antworten von KS sind direkt relevant für die von mir gestellten
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
KS liefert vollständige Antworten, ohne wichtige Informationen auszulassen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
Die Antworten sind frei von unnötigen Details und konzentrieren sich nur auf das Gefragte.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

5. Bitte geben Sie Ihr Feedback zur Kontextrelevanz von KS.
Ich habe das Gefühl, dass KS nur Informationen verwendet, die für meine Anfrage relevant sind.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

6. Bitte geben Sie Ihr Feedback zu Ihrer Zufriedenheit mit KS.
KS hat meine Erwartungen in Bezug auf Funktionalität und Leistung erfüllt oder übertroffen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

7. Bitte geben Sie Ihr Feedback zu Ihrem wahrgenommenen Nutzen von KS.

Die Nutzung von KS ermöglicht es mir, Aufgaben schneller zu erledigen und verbessert meine Arbeitsleistung.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

8. Bitte geben Sie Ihr Feedback zur Qualität von KS.
KSliefert konsistent genaue, korrekte und zuverlässige Informationen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
Die Antworten sind so strukturiert, dass es leicht ist, die bereitgestellten Informationen zu verstehen und zu folgen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

9. Bitte geben Sie Ihr Feedback zum Geschätswert von KS.
KS spart im Vergleich zur heutigen Stichwortsuche Zeit und Ressourcen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
Die Leistung von KS rechtfertigt ihren Einsatz als Geschäftsinstrument.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

10. Bitte geben Sie Ihr Feedback zu Ihrer Offenheit für neue Technologien.
Ich bin stets offen, neue Technologien auszuprobieren, da ich davon überzeugt bin, dass sie meine Fähigkeiten erweitern und mich in meiner beruflichen Entwicklung unterstützen.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu

[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

11. Bitte geben Sie Ihr Feedback zur Ersatzbarkeit und Notwendigkeit von KS.
Die heutigen Stichwortsuche kann durch KS ersetzt werden.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu
KS ist eine sinnvolle Erweiterung aber nicht zwingend notwendig.
[ ] Stimme überhaupt nicht zu
[ ] Stimme nicht zu
[ ] Stimme weder zu noch nicht zu
[ ] Stimme zu
[ ] Stimme voll und ganz zu

12. Haben Sie noch weiteres Feedback über KS?

A.3. Customer Agents Survey Descriptive Statistics

Evaluated Metric	Mean	SD	Skew
Perceived Ease of Use	4.24	0.39	-0.66
Performance	3.55	1.13	-0.12
Answer Faithfulness	3.56	0.70	0.22
Answer Relevance	3.65	0.71	-0.77
Context Relevance	3.29	0.88	0.05
Satisfaction	3.59	0.71	-0.35
Perceived Usefulness	3.47	1.57	-0.22
Quality	3.69	0.8	-0.28
Business Value	3.71	0.89	-0.40
Openness to New Technologies	4.35	0.853	-1.64
Replaceability and necessity of CS	3.09	0.94	0.35

Table A.1.: Descriptive Statistics of Evaluation Metrics.

Evaluated Metric	Statement	Mean	SD
Perceived Ease of Use	1. CS helps me find the information I am looking for without needing additional support.	4.18	0.53
	2. CS's user interface is easy to understand and requires minimal effort to use.	4.29	0.77
Performance	1. CS remains stable and functional when faced with unusual requests.	3.41	1.18
	2. CS's answers are consistent and logically connected to my questions.	3.71	0.85
	3. CS efficiently delivers fast and relevant answers without unnecessary steps or delays.	3.71	1.21
Answer Faithfulness	1. I noticed cases where the response didn't match the context retrieved.	2.94	1.09
	2. In most cases, CS makes statements that are supported by the information retrieved.	4.06	0.66
Answer Relevance	1. CS's answers are directly relevant to the questions I asked.	3.82	0.64
	2. CS provides complete answers without leaving out important information.	3.59	0.80
	3. CS's answers are free of unnecessary details and focus only on what is being asked.	3.53	1.01
Context Relevance	1. I feel that CS only uses information that is relevant to my request.	3.29	0.85
Satisfaction	1. CS met or exceeded my expectations in terms of functionality and performance.	3.59	0.71
Perceived Usefulness	1. Using CS allows me to complete tasks faster and improves my work performance	3.47	1.12
Quality	1. CS consistently provides accurate, correct, and reliable information.	3.59	0.94
	2. CS's answers are structured to make it easy to understand and follow the information provided.	3.71	0.92
Business Value	1. CS saves time and resources compared to today's keyword search.	3.53	0.94
	2. CS's performance justifies its use as a business tool.	4	0.912
Openness to New Technologies	1. I am always open to trying new technologies as I believe they will expand my skills and support my professional development.	3.88	0.86
Replaceability and necessity of CS	1. Today's keyword search can be replaced by CS.	3.0	1.06
	2. CS is a useful extension but not absolutely necessary.	2.82	1.07

Table A.2.: Survey Questions Mapped to Metrics with Descriptive Statistics for Each Statement.

A.4. Correlation Analysis Results

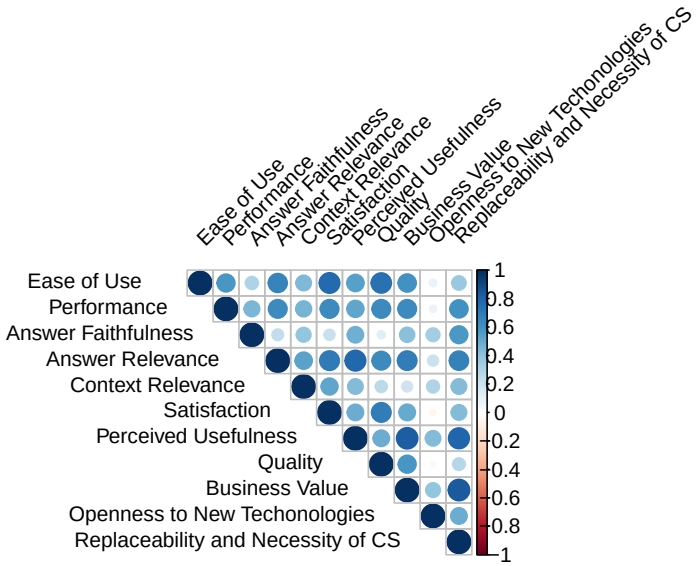

Figure A.1.: Correlation Results Between Evaluation Metrics.

B. Search Tool Preferences Across Different Query Types

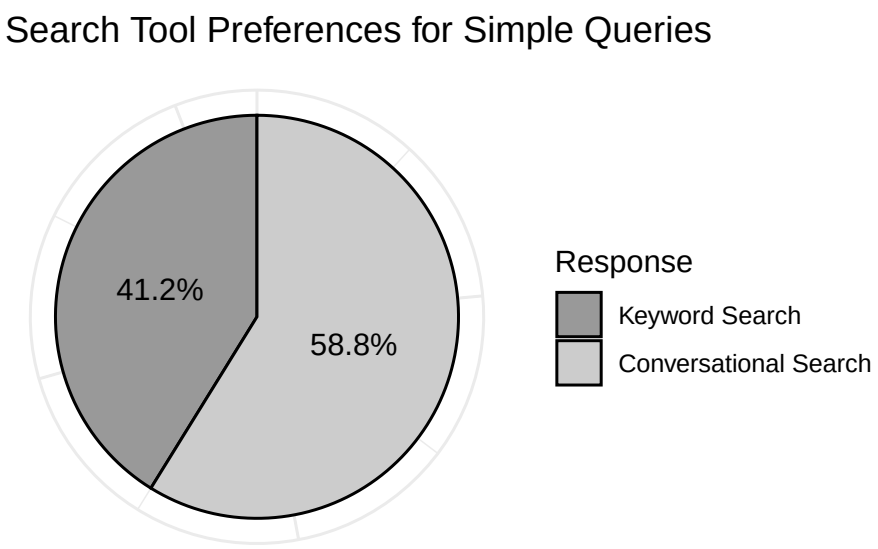

Figure B.1.: Search Tool Preferences for Simple Queries (n=17).

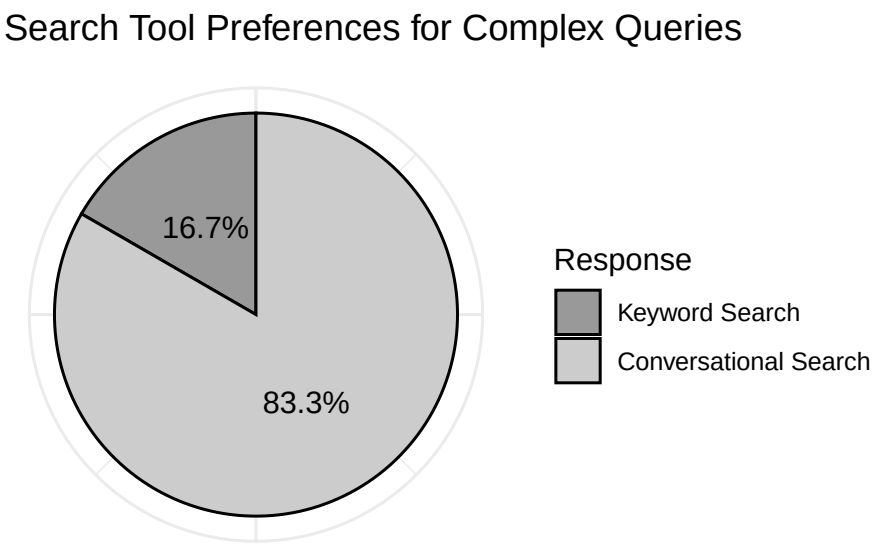

Figure B.2.: Search Tool Preferences for Complex Queries (n=17).

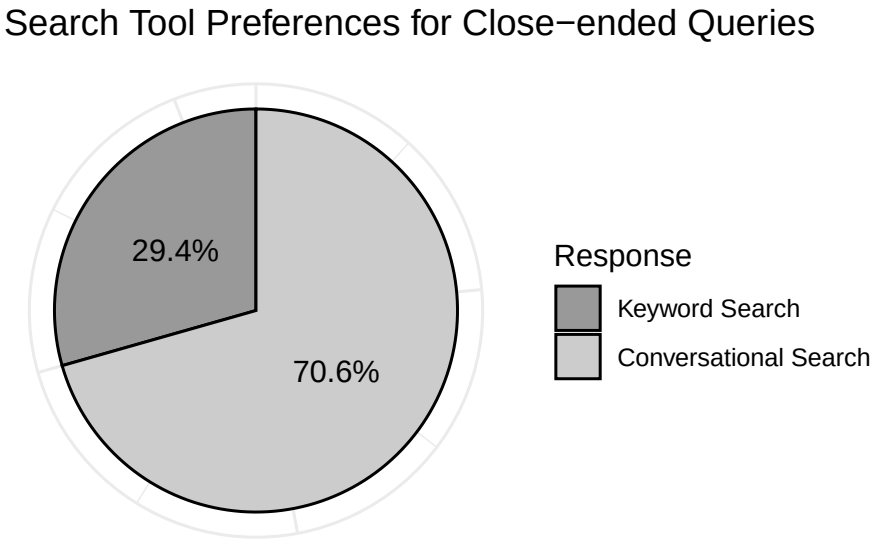

Figure B.3.: Search Tool Preferences for Closed Queries (n=17).

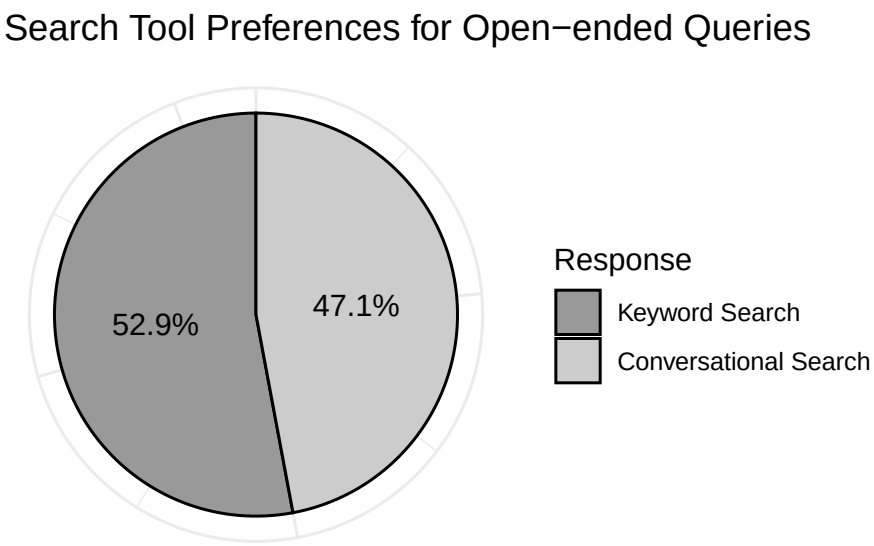

Figure B.4.: Search Tool Preferences for Open Queries (n=17).

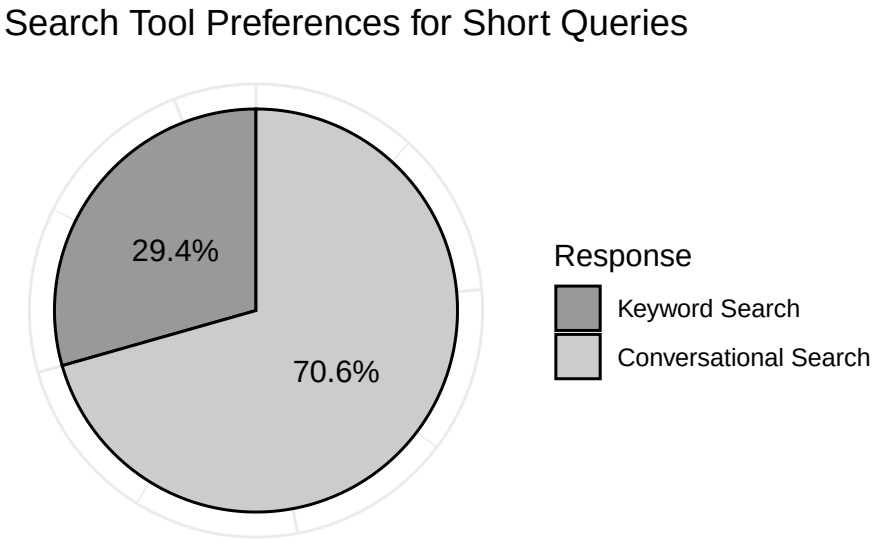

Figure B.5.: Search Tool Preferences for Short Queries (n=17).

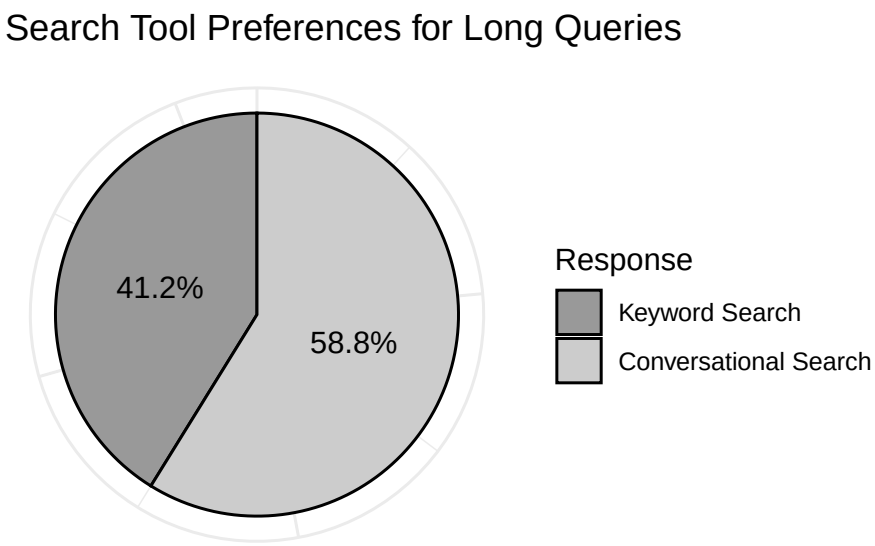

Figure B.6.: Search Tool Preferences for Long Queries(n=17).

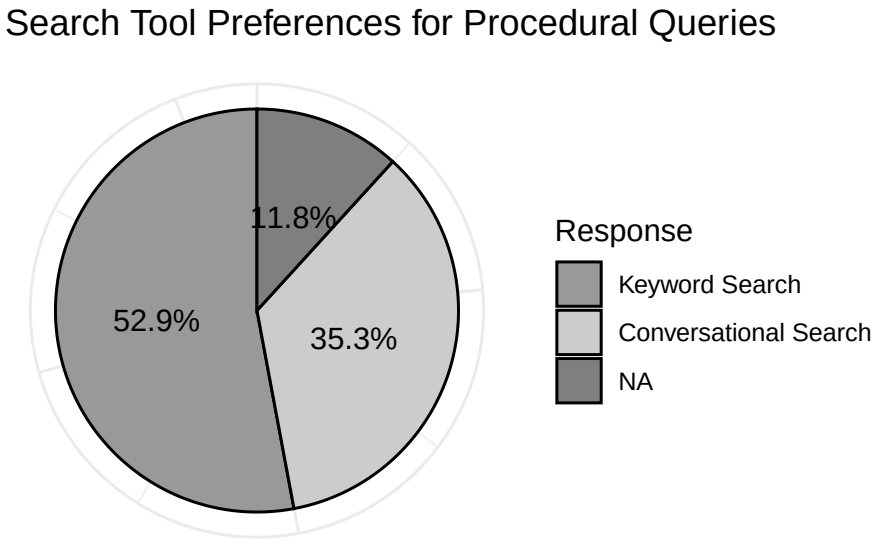

Figure B.7.: Search Tool Preferences for Procedural Queries(n=17).

C. Transcript Example

Note: The interviews were conducted and transcribed in their original language German. For maintaining the anonymity, the previous part of the interview where individual characteristics were gathered, is not included in the transcript.

Q: Jetzt, haben wir ein paar Szenarien und hier geht es darum zu verstehen, in welchen Fällen du die konversationelle Suche bevorzugen würdest und in welchen Fällen heutige Stichwortsuche. Und hier haben wir als Beispiel eine einfache Kundenabfrage und einfach bedeutet jetzt, man schaut in einem Dokument und man findet schon die Antwort da. Und als Beispiel hier haben wir die Abfrage, sind E -Scooter in Privatschutz versicherbar? Und wir nehmen hier an, dass du die Antwort nicht kennst. Für diese Frage, welches Suchwerkzeug würdest du lieber benutzen und warum?

A: Wenn das so ist, wie du es gerade gestellt hast, eine einfache Sache, wo ich weiß, dass ich es finde, dann vertraue ich immer auf herkömmliche Quellen. Also ich würde immer erst auf meine Quelle, die ich ursprünglich schon vor der konversationellen Suche hatte, schauen. Weil ich sicher bin, dass diese Antwort zu 99% richtig ist.

Q: Und jetzt, haben wir eine komplexe Abfrage und das erfordert eine intensivere Recherche in mehrere Dokumente zu schauen. Und Beispiel wäre, wie versichere ich einen minderjährigen Versicherungsnehmer? Und in diesem Fall, welches würdest du bevorzugen und warum?

A: Ja, darf ich nicht wüsste, also dass ich so unwissentlich bin, weil ich weiß die Antwort leider, würde ich in der Tat dann die konversationelle Suche benutzen, weil ich auch gerade da dann eine Anzeige haben möchte von unterschiedlichen Quellen, also beziehungsweise die Suche würde dann schneller gehen, weil die konversationelle Suche ja bestimmte Punkte dann aufzeigt, zum Beispiel bestimmte Themen, aus die man auswählen kann und dann würde ich mal oberflächlich angucken und dann aber noch die am Anfang ausgeblendete konversationelle Suche Antwort angucken. Also ich gucke mir ganz schnell die Themen an, ob das überhaupt zu meiner Frage passt, weil man kann an diesen Auswahlkriterien schon sehen, aha, das hat nichts mit meiner Frage zu tun, das ist ein anderes Thema, das ist eine andere Branche. Dann gucke ich ganz schnell einmal auf die Antwort und wenn die Antwort passt dann suche ich noch mal in den Lösungsweg die mir vorher gezeigt werden. Und wenn es nicht passt, dann gehe ich auf meine eigenen Recherchen suche.

Q: Okay, und als nächstes, wir haben eine geschlossene Abfrage und das bedeutet, die Antwort ist ja oder nein. Und das Beispiel wäre, ist Morbus Addison eine Masterkrankheit?

Und welches Werkzeug würdest du auswählen?

A: Das ist jetzt schwer. Da würde ich die konversationelle Suche, weil ich den Link schneller öffnen kann. Bei konversationeller Suche weiß ich, da ist die Eingabetaste. Wo ich den Text eingeben kann. Und auch Text als Antwort bekomme.

Q: Und jetzt haben wir eine offene Abfrage und hier es erfordert eine ausführlichere und eine detailliertere Antwort, also nicht nur jetzt ja oder nein, sondern so eine detailliertere Antwort. Und Beispiel ist jetzt, was muss ich als Erwerber/ Veräußerer bei einem Besitzwechsel beachten? Und jetzt meine Frage ist wieder, welches Suchwerkzeug du hier lieber benutzen würdest und warum?

A: Da benutze ich Stichwortsuche, weil das eine Frage ist die ich ganz häufig verwende und wenn ich ganz häufig eine Anfrage habe wo ich einfach mich selber bestätigen möchte dann gehe ich meine ursprüngliche Quelle und da gibt es einen Begriff, der nennt sich Rund um den Eigentums Wechsel. Und da weiß ich immer, auf welche Seite ich ungefähr blättern muss, auf welcher Seite so die Lösungsvorschläge für die konkrete Situation sind.

Q: Okay. Und die nächste Metrik ist die Länge der Abfrage. Als Beispiel haben wir hier eine kurze Abfrage, die weniger als acht Wörter ist. Als Beispiel, für wie lange gilt der Sofortschutz? Was würdest du hier bevorzugen?

A: Ich würde die konversationelle Suche benutzen, weil ich da eben halt in kurzen Wörtern schreiben kann welche Regel besteht für Sofortschutz in Privatschutz und dann hoffe ich, dass es mir gleich die Felder angibt.

Q: Okay, und wir haben jetzt eine lange Abfrage. Beispiel, sind Schäden durch mein Haustier wie zum Beispiel Bissverletzungen oder Sachschäden in der Haftpflichtversicherung mitversichert? Und hier, welches Werkzeug würdest du benutzen?

A: Da würde ich tatsächlich Stichwortsuche benutzen, weil ich wie bei ChatGPT noch nicht alle Antworten vertraue, wenn da mehrere Komplexe reinfolgen sind. Wenn die Antwort aus mehreren Prüfteilen kommt, man sagt heutzutage man prüft causal adäquat. Man prüft ist dies Ereignis überhaupt ein Versicherungspunkt und ja dann versichern wir das? Ich denke KI kann das noch nicht.

Q: Okay. Und unser letztes Szenario ist eine prozedurale Abfrage und das bedeutet als Antwort, man braucht so eine Schritt für Schritt Ausführung, wie man etwas machen kann. Zum Beispiel, wie zahle ich ein Guthaben wieder aus? Und hier, welcher würdest du bevorzugen und warum?

A: Schwere Frage ich habe es noch nicht so gehabt. Es ist schwer, weil ich. es weiß. Ich muss

mal kurz überlegen, wie ich jemanden denken würde, der es nicht wüsste. Also ich stelle mir jetzt gerade vor, ich würde es nicht mehr wissen. Ich würde tatsächlich im ersten Schritt in der konversationellen Suche gucken, mit der Hoffnung, dass mir bei konversationeller Suche Lösungsdateien angeboten werden, wo ich vielleicht oberflächlich schon sehen kann, das trifft genau das Thema. Und wenn es das Thema trifft, gehe ich in diesen jeweiligen Lösungspunkt rein.

Q: Jetz habe ich ein Paar zusätzliche Fragen. Welche Faktoren beeinflussen deine Wahl zwischen der konversationellen Suche und der heutigen Stichwortsuche?

A: Die Erfahrung mit welcher Treffsicherheit mir die Lösung ansteht und angezeigt wird. Wenn ich weiß, dass meine Frage an einer bestimmten Stelle steht oder ich schon mal in der Vergangenheit gefunden habe, dann würde ich nach der Stichwortsuche gehen, weil ich dann die Erwartung habe, mir wird das angezeigt, was ich erwarte. Ist es ein neues Thema, was ich nicht so häufig habe oder nicht so häufig aufrufe, was ganz neues, auch mit Stichwort -Zufriedenheit besucht werden müsste, dann würde ich über die konversationalle Suche reingehen, weil ich der Meinung bin, dass die konversationelle Suche eventuell durch diese Antwort -Kacheln vielleicht die Antwort schneller finden könnte.

Q: Okay. Und deiner Meinung nach, was sind die Stärken und Einschränkungen von der konversationellen Suche?

A: Die Stärken sind, dass es wirklich gute Antworten haben kann. Also es kann hilfreich sein. Schwachstellen sind, dass man auch hier immer noch kritisch auf die Antworten gucken muss, inwiefern sie richtig sind. Und was ich zum Beispiel über andere intelligente KI-Programme finde, ist, dass man eine Antwort nicht mit einer weiteren Antwort dynamisieren kann. Zum Beispiel wenn ich frage, wie klappt eine Doppelversicherung und das System findet die Antwort nicht sofort, dass ich sagen kann, auf die vorherige Frage, wie lautet da die Antwort in Bezug auf das Haftpflichtversicherung, sodass er dann darauf einsteigt, die erste Frage, achten wir mal auf das Thema Klärung der Doppelversicherung, die zweite Frage, sagt er was von Haftpflichtversicherung, dass er sie sofort mit zusammenfasst. Für die konversationelle Suche muss man die Frage endlich neu stellen, wenn eine Antwort fertiggestellt ist?

Q: Okay, und deiner Meinung nach, was sind die Stärken und Einschränkungen der heutigen Stichwortsuche?

A: Das Problem bei der Stichwortsuche ist das man immer nur nach einem Wort gucken kann. Das bedeutet man soll schon ein genau synonymes oder ein gewisses Wort wählen um die Lösung zu haben. Das ist ein Nachteil. Der Vorteil ist, wenn ich das Wort kenne und ich weiß, dass es da steht, kann ich mit diesem Wort sofort auf das Ziel kommen und dann wird mir das Ziel angezeigt. Also die Zielanteile, die ich erwarte.

Q: Okay. Und gibt es zusätzliche Funktionen oder Verbesserungen, die du dich für die konversationelle Suche wünschen würdest?

A: Ich würde wünschen das man dort eine Sprachbenutzung hinzufügen kann. Das eintippen von Fragen Dauer länger als wenn ich sagen würde wie ein Handy "Hey...". Das Eintippen dauert langer als das man es ausspricht. Dann hatte ich als Vorschlag noch, dass man dynamisiert und dass man auf vorherige Frage Bezug nehmen kann. Die konversationelle Suche vorherige Frage noch mitintegrieren kann und auch Folgefrage. Das geht z.B. bei ChatGPT. Wenn man da was fragt und sagt, das ist nicht die Antwort, dann kann man das noch ergänzend reinbringen, damit die Antwort doch nochmal neu kontrolliert und neu ausgegeben wird. Und vielleicht wirkt die konversationelle Suche immer noch so ein bisschen starr. Es ist jetzt ganz neu aber es ist ganz starr wie damals Yahoo Browser aussah. Und darüber denkt man heute oh Gott das war langweilig. Das hat ja keine interessiert also eine Werbung da drin fand schlecht aus, wie so eine Briefmarke auf einer Postkarte, wo man sagt, das passt nicht. Und heutzutage mit der Zeit kommt das ja alles und modernisiert sich Schritt zu Schritt von Version zu Version. Die konversationelle Suche wirkt aktuell für mich noch optisch langweilig.
Was noch gut wäre, dass man bestimmte Hits oder Treffer in Prozenten ausdrücken kann z.B. wie Richtig könnte diese Antwort sein. 90 Prozent, 80 Prozent. Man könnte die Fragen zum Beispiel auch farblich darstellen. Sobald eine Antwort größer als 90 Prozent ist, dann ist das eine gute Antwort. Antwortfenster zum Beispiel in einem gewissen gleichen Grün darstellen und wenn die Antworten schlechter werden, wo das System erkennt, ja, die Trefferquote liegt da nur bei 60 oder 40 Prozent, könnte man das farblich auch so ein bisschen hervorheben, dass man gleich weiß, da braucht man sich eigentlich gar nicht mehr so richtig mit drum beschäftigen.

Q: Und eigentlich das war's. Ich habe alle meine Fragen gestellt. Hast du noch etwas, das du noch ergänzen möchtest?

A: Aktuell finde ich die konversationelle Suche sehr schwächlich. Die Antworten sind leider alle nicht so perfekt wie man sie haben möchte. Es ist manchmal sehr schwer eine Antwort zu beurteilen, wenn man sie kritisch einmal liest. Könnte das richtig sein? Könnte es nicht richtig sein? Weil manchmal ist ja noch sehr pauschal die Antwort oft, noch nicht so zielgerecht, wie zum Beispiel ChatGPT, was schon seit Jahren läuft. Also es ist meinen Augen leider noch sehr anfällig für Fehler. Die Vertrauensbasis dort liegt bei mir bei 50 Prozent zu 50 Prozent. Also 50 -50, so lese ich eine Antwort. Also ich bin immer wieder Misstrauisch, wenn ich eine Antwort lese und sehr Misstrauisch. Wo ich für ChatGPT ein Vertrauenswert von 70% habe.

Q: Ich habe auch noch eine Umfrage, die du selber ausfüllen kannst, ich schicke mal den Link. Es wäre super, wenn du das auch noch ausfüllen kannst und ansonsten vielen lieben Dank, dass du Zeit genommen hast.

List of Acronyms

AIDUA	Artificially Intelligent Device Use Acceptance
AI	Artificial Intelligence
ARES	Automated Retrieval-Enhanced Scoring
B2C	Business-to-Consumer
CA	Conversational Agent
CS	Conversational Search
EKA	Enterprise Knowledge Assistant
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
HIS	Hybrid Intelligence System
JSON	JavaScript Object Notation
LLM	Large Language Model
RAGAS	Retrieval-Augmented Generation Assessment
TAM	Technology Acceptance Model
UTAUT	Unified Theory of Acceptance and Use of Technology

List of Figures

List of Tables

Bibliography

[1] A. Fowler. "The role of AI-based technology in support of the knowledge management value activity cycle". In: *The Journal of Strategic Information Systems* 9 (Sept. 2000), pp. 107–128. DOI: 10.1016/S0963-8687(00)00041-X.

[2] J. Liebowitz. "Knowledge management and its link to artificial intelligence". en. In: *Expert Syst. Appl.* 20.1 (2001), pp. 1–6.

[3] P. Tyndale. "A taxonomy of knowledge management software tools: origins and applications". en. In: *Eval. Program Plann.* 25.2 (2002), pp. 183–190.

[4] J. Liebowitz and T. Beckman. *Knowledge Organizations: What Every Manager Should Know*. CRC Press, 2020.

[5] F. Brachten, T. Kissmer, and S. Stieglitz. "The acceptance of chatbots in an enterprise context – A survey study". In: *International Journal of Information Management* 60 (Oct. 2021), p. 102375. DOI: 10.1016/j.ijinfomgt.2021.102375.

[6] C. Wiethof, M. Poser, and E. Bittner. "Design and evaluation of an employee-facing conversational agent in online customer service". In: *Proceedings of the XYZ Conference*. 2022.

[7] T. Lewandowski, J. Delling, C. Grotherr, and T. Böhmann. "State-of-the-Art Analysis of Adopting AI-based Conversational Agents in Organizations: A Systematic Literature Review". In: *PACIS 2021 Proceedings*. 167. 2021. URL: https://aisel.aisnet.org/pacis2021/167.

[8] I. Nonaka and H. Takeuchi. "The knowledge-creating company: How Japanese companies create the dynamics of innovation". In: *Oxford University Press* (1995).

[9] E. Bolisani and C. Bratianu. *Emergent knowledge strategies*. en. 1st ed. Knowledge Management and Organizational Learning. Cham, Switzerland: Springer International Publishing, July 2017.

[10] A. J. Ayer. "The right to be true". In: *R. Neta & D. Pritchard (Eds.) Arguing about knowledge* (2009), pp. 11–13.

[11] A. S. Bollinger and R. D. Smith. "Managing organizational knowledge as a strategic asset". en. In: *J. Knowl. Manag.* 5.1 (Mar. 2001), pp. 8–18.

[12] T. S. Xue. "A Literature Review on Knowledge Management in Organizations". In: *Research in Business and Management* 4 (Feb. 2017), p. 30. DOI: 10.5296/rbm.v4i1.10786.

[13] T. Davenport and L. Prusak. *Working Knowledge: How Organizations Manage What They Know*. Vol. 1. Jan. 1998. DOI: 10.1145/348772.348775.

[14] F. Pangil and A. Nasurddin. "Knowledge and the Importance of Knowledge Sharing in Organizations". In: (2013). URL: https://api.semanticscholar.org/CorpusID:52063885.

[15] R. Stevens, M. Joshua, and C. Sondra. "Waves of Knowledge Management: The Flow between Explicit and Tacit Knowledge". In: *American Journal of Economics and Business Administration* 2 (Jan. 2010).

[16] T. Calo. "Talent Management in the Era of the Aging Workforce: The Critical Role of Knowledge Transfer". In: *Public Personnel Management* 37 (Dec. 2008), pp. 403–416. DOI: 10.1177/009102600803700403.

[17] R. A. Noe. *Employee Training and Development*. en. 5th ed. Maidenhead, England: McGraw Hill Higher Education, Dec. 2009.

[18] J. P. Soto, A. Vizcaíno, J. Portillo, and M. Piattini. "Modelling a Knowledge Management System Architecture with INGENIAS Methodology". In: *Proceedings of the 15th International Conference on Computing, CIC 2006*. 15th International Conference on Computing, CIC 2006 ; Conference date: 21-11-2006 Through 24-11-2006. 2006, pp. 167–173. ISBN: 0769527086. DOI: 10.1109/CIC.2006.45.

[19] B. Hooff and M. Huysman. "Managing Knowledge Sharing: Emergent and Engineering Approaches". In: *Information & Management* 46 (Jan. 2009), pp. 1–8. DOI: 10.1016/j.im.2008.09.002.

[20] M. Ipe. "Knowledge Sharing in Organizations: A Conceptual Framework". In: *Human Resource Development Review* 2 (Dec. 2003), pp. 337–359. DOI: 10.1177/1534484303257985.

[21] M. Lytras, N. Pouloudi, and A. Poulymenakou. "Knowledge management convergence: Expanding learning frontiers". In: *Journal of Knowledge Management* 6 (Mar. 2002), pp. 40–51. DOI: 10.1108/13673270210417682.

[22] C. O'Dell and T. Davenport. "Application of AI for Knowledge Management". In: *CIOReview* (2019). URL: https://knowledgemanagement.cioreview.com/cxoinsight/application-of-ai-for-knowledge-management-nid-30328-cid-132.html.

[23] S.-C. Chong, K. Y. Wong, and B. Lin. "Criteria for Measuring KM Performance Outcomes in Organisations". In: *Industrial Management and Data Systems* 106 (Aug. 2006), pp. 917–936. DOI: 10.1108/02635570610688850.

[24] M. Shongwe. "An Analysis of Knowledge Management Lifecycle Frameworks: Towards a Unified Framework". In: *Electronic Journal of Knowledge Management* (Jan. 2016), pp. 140–153.

[25] A. Garavelli, M. Gorgoglione, and B. Scozzi. "Managing Knowledge Transfer by Knowledge Technologies". In: *Technovation* 22 (May 2002), pp. 269–279. DOI: 10.1016/S0166-4972(01)00009-8.

[26] M. Alavi and D. Leidner. "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues". In: *MIS quarterly* 1 (Mar. 2001), pp. 107–. DOI: 10.2307/3250961.

[27] S. M. Jasimuddin. "An Integration of Knowledge Transfer and Knowledge Storage: An Holistic Approach". In: 2005. URL: https://api.semanticscholar.org/CorpusID:3201106.

[28] S. S.-L. Tan, H.-H. Teo, B. C. Y. Tan, and K. K. Wei. "Developing a Preliminary Framework for Knowledge Management in Organizations". In: 1998. URL: https://api.semanticscholar.org/CorpusID:166880605.

[29] E. Ode and R. Ayavoo. "The mediating role of knowledge application in the relationship between knowledge management practices and firm innovation". In: *Journal of Innovation & Knowledge* 5 (Sept. 2019). DOI: 10.1016/j.jik.2019.08.002.

[30] H. Boateng and F. G. Agyemang. "The effects of knowledge sharing and knowledge application on service recovery performance". In: *Business Information Review* 32.2 (2015), pp. 119–126. DOI: 10.1177/0266382115587852.

[31] M. Song, H. Van Der Bij, and M. Weggeman. "Determinants of the Level of Knowledge Application: A Knowledge-Based and Information-Processing Perspective*". In: *Journal of Product Innovation Management* 22.5 (2005), pp. 430–444. DOI: https://doi.org/10.1111/j.1540-5885.2005.00139.x.

[32] S. Allameh, S. Zare, and S. Davoodi. "Examining the Impact of KM Enablers on Knowledge Management Processes". In: *Procedia CS* 3 (Dec. 2011), pp. 1211–1223. DOI: 10.1016/j.procs.2010.12.196.

[33] H. Tsoukas and N. Mylonopoulos. "Introduction: Knowledge construction and creation in organizations*". en. In: *Br. J. Manag.* 15.S1 (Mar. 2004), S1–S8.

[34] J. van Aalst. "Distinguishing knowledge-sharing, knowledge-construction, and knowledge-creation discourses". en. In: *Int. J. Comput. Support. Collab. Learn.* 4.3 (June 2009), pp. 259–287.

[35] I. Nonaka. "A dynamic theory of organizational knowledge creation". en. In: *Organ. Sci.* 5.1 (Feb. 1994), pp. 14–37.

[36] W. Hu. "Framework of knowledge acquisition and sharing in multiple projects for contractors". In: *2008 International Symposium on Knowledge Acquisition and Modeling*. Wuhan, China: IEEE, Dec. 2008.

[37] G. Huber. "Organizational Learning: The Contributing Processes and the Literatures". In: *Organization Schiences* 2 (Feb. 1991). DOI: 10.1287/orsc.2.1.88.

[38] K. Metaxiotis, K. Ergazakis, E. Samouilidis, and J. Psarras. "Decision support through knowledge management: The role of the artificial intelligence". In: *Inf. Manag. Comput. Security* 11 (Dec. 2003), pp. 216–221. DOI: 10.1108/09685220310500126.

[39] A. Agrawal, J. Gans, and A. Goldfarb. "What to Expect from Artificial Intelligence Technology". In: Jan. 2018, pp. 23–36. ISBN: 9780262345354. DOI: 10.7551/mitpress/11645.003.0008.

[40] A. Simmons and S. Chappell. "Artificial intelligence-definition and practice". In: *IEEE Journal of Oceanic Engineering* 13.2 (1988), pp. 14–42. DOI: 10.1109/48.551.

[41] High Level Expert Group on Artificial Intelligence. *A definition of AI: Main capabilities and disciplines*. Accessed: 2024-05-30. 2019. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341.

[42] B. Wirtz, J. Weyerer, and C. Geyer. "Artificial Intelligence and the Public Sector—Applications and Challenges". In: *International Journal of Public Administration* 42.7 (2018), pp. 596–615. DOI: 10.1080/01900692.2018.1498103.

[43] AI 4 Belgium Coalition. *AI 4 Belgium Report*. Accessed: 2024-05-30. 2019. URL: https://www.ai4belgium.be/report.

[44] P. Mikalef and M. Gupta. "Artificial Intelligence Capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance". In: *Information & Management* 58 (Jan. 2021), p. 103434. DOI: 10.1016/j.im.2021.103434.

[45] F. Olan, E. Arakpogun, J. Suklan, F. Nakpodia, N. Damij, and U. Jayawickrama. "Artificial Intelligence and Knowledge Sharing: Contributing Factors to Organizational Performance". In: *Journal of Business Research* 145 (Mar. 2022). DOI: 10.1016/j.jbusres.2022.03.008.

[46] M. H. Jarrahi, D. Askay, A. Eshraghi, and P. Smith. "Artificial Intelligence and Knowledge Management: A Partnership Between Human and AI". In: *Business Horizons* 66 (Mar. 2022). DOI: 10.1016/j.bushor.2022.03.002.

[47] K. Sowa, A. Przegalinska, and L. Ciechanowski. "Cobots in knowledge work Human -AI collaboration in managerial professions". In: *Journal of Business Research* 125 (Mar. 2021), pp. 135–142. DOI: 10.1016/j.jbusres.2020.11.038.

[48] T. Sakirin and R. Ben Said. "User preferences for ChatGPT-powered conversational interfaces versus traditional methods". In: *Mesopotamian Journal of Computer Science* 8.1 (Jan. 2023), pp. 24–31.

[49] L. Nicolescu and M. T. Tudorache. "Human-computer interaction in customer service: The experience with AI chatbots—A systematic literature review". In: *Electronics* 11.10 (May 2022), p. 1579. DOI: 10.3390/electronics11101579.

[50] A. Bin Sawad, B. Narayan, A. Alnefaie, A. Maqbool, I. Mckie, J. Smith, B. Yuksel, D. Puthal, M. Prasad, and A. B. Kocaballi. "A systematic review on healthcare artificial intelligent conversational agents for chronic conditions". In: *Sensors* 22.7 (Mar. 2022), p. 2625. DOI: 10.3390/s22072625.

[51] T. Wambsganss, R. Winkler, M. Söllner, and J. M. Leimeister. "A conversational agent to improve response quality in course evaluations". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA: ACM, Apr. 2020. DOI: 10.1145/3334480.3383006.

[52] X. Wang, X. Lin, and B. Shao. "How does artificial intelligence create business agility? Evidence from chatbots". In: *International Journal of Information Management* 66 (Oct. 2022), p. 102535. DOI: 10.1016/j.ijinfomgt.2022.102535.

[53] Y. Xu, C.-H. Shieh, P. van Esch, and I.-L. Ling. "AI customer service: Task complexity, problem-solving ability, and usage intention". In: *Australasian Marketing Journal* 28.4 (2020), pp. 189–199.

[54] M. Adam, M. Wessel, and A. Benlian. "AI-based chatbots in customer service and their effects on user compliance". In: *Electronic Markets* 31.2 (June 2021), pp. 427–445. DOI: 10.1007/s12525-020-00414-7.

[55] T. Reddy. "How chatbots can help reduce customer service costs by 30%". In: *IBM Watson Blog* (2017). URL: https://www.ibm.com/blogs/watson/2017/10/how-chatbots-reduce-customer-service-costs-by-30-percent/.

[56] M. Poser, S. Singh, and E. Bittner. "Hybrid service recovery: Design for seamless inquiry handovers between conversational agents and human service agents". In: *Proceedings of the 54th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2021.

[57] M. Ashfaq, J. Yun, S. Yu, and S. Loureiro. "I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents". In: *Telematics and Informatics* 54 (2020), p. 101473. DOI: 10.1016/j.tele.2020.101473.

[58] P. Reinhard, N. Neis, L. Kolb, D. Wischer, M. Li, A. Winkelmann, F. Teuteberg, U. Lechner, and J. M. Leimeister. "Augmenting frontline service employee onboarding via hybrid intelligence: Examining the effects of different degrees of human-GenAI interaction". In: *Proceedings of the XYZ Conference*. 2024.

[59] M. Song, X. Xing, Y. Duan, J. Cohen, and J. Mou. "Will artificial intelligence replace human customer service? The impact of communication quality and privacy risks on adoption intention". In: *Journal of Retailing and Consumer Services* 66 (2022), p. 102900. DOI: 10.1016/j.jretconser.2021.102900.

[60] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister. "Hybrid intelligence". In: *Business and Information Systems Engineering* 61.5 (2019), pp. 637–643.

[61] P. Hemmer, M. Schemmer, M. Vössing, and N. Kühl. "Human-AI complementarity in hybrid intelligence systems: A structured literature review". In: *Proceedings of the XYZ Conference*. 2021.

[62] E. Ritz, F. Donisi, E. Elshan, and R. Rietsche. "Artificial socialization? How artificial intelligence applications can shape a new era of employee onboarding practices". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2023.

[63] X. Pierre. "Levels of involvement and retention of agents in call centres: Improving well-being of employees for better socioeconomic performance". In: *Journal of Business Research* 64.7 (2011), pp. 735–740.

[64] Z. Gao and J. Jiang. "Evaluating human-AI hybrid conversational systems with chatbot message suggestions". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Virtual Event Queensland Australia: ACM, Oct. 2021.

[65] D. Peras. "Chatbot evaluation metrics: review paper". In: 2018, pp. 89–97.

[66] K. Kuligowska. "Commercial Chatbot: Performance Evaluation, Usability Metrics and Quality Standards of Embodied Conversational Agents". In: *Professionals Center for Business Research* 2 (Jan. 2015), pp. 1–16. DOI: 10.18483/PCBR.22.

[67] N. Radziwill and M. Benton. "Evaluating Quality of Chatbots and Intelligent Conversational Agents". In: (Apr. 2017). DOI: 10.48550/arXiv.1704.04579.

[68] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. "RAGAS: Automated evaluation of Retrieval Augmented Generation". In: (2023). eprint: 2309.15217 (cs.CL).

[69] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia. "ARES: An Automated evaluation framework for retrieval-augmented generation systems". In: (2023). eprint: 2311.09476 (cs.CL).

[70] R. De Cicco, S. Iacobucci, A. Aquino, F. Romana Alparone, and R. Palumbo. "Understanding users' acceptance of chatbots: An extended TAM approach". In: *Chatbot Research and Design*. Lecture notes in computer science. Cham: Springer International Publishing, 2022, pp. 3–22.

[71] F. D. Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology". In: *MIS Q* 13.3 (1989), p. 319.

[72] Venkatesh, Morris, Davis, and Davis. "User acceptance of information technology: Toward a unified view". In: *MIS Q* 27.3 (2003), p. 425.

[73] O. H. Chi, C. Chi, D. Gursoy, and R. Nunkoo. "Customers' acceptance of artificially intelligent service robots: The influence of trust and culture". In: *International Journal of Information Management* 70 (June 2023), p. 102623. DOI: 10.1016/j.ijinfomgt.2023.102623.

[74] D. Gursoy, O. H. Chi, L. Lu, and R. Nunkoo. "Consumers Acceptance of Artificially Intelligent Device Use in Service Delivery". In: *International Journal of Information Management* 49 (Apr. 2019), pp. 157–169. DOI: 10.1016/j.ijinfomgt.2019.03.008.

[75] A. M. Preininger, B. L. Rosario, A. M. Buchold, J. Heiland, N. Kutub, B. S. Bohanan, B. South, and G. P. Jackson. "Differences in information accessed in a pharmacologic knowledge base using a conversational agent vs traditional search methods". en. In: *Int. J. Med. Inform.* 153.104530 (Sept. 2021), p. 104530.

[76] B. Liu, Y. Wu, Y. Liu, F. Zhang, Y. Shao, C. Li, M. Zhang, and S. Ma. "Conversational vs traditional: Comparing search behavior and outcome in legal case retrieval". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021.

[77] A. Wazzan, S. MacNeil, and R. Souvenir. "Comparing traditional and LLM-based search for image geolocation". In: *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, 2024.

[78] S. Spatharioti, D. Rothschild, D. Goldstein, and J. Hofman. "Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment". In: (July 2023). DOI: 10.48550/arXiv.2307.03744.

[79] E. C. Ling, I. Tussyadiah, A. Tuomi, J. Stienmetz, and A. Ioannou. "Factors influencing users' adoption and use of conversational agents: A systematic review". en. In: *Psychol. Mark.* 38.7 (2021), pp. 1031–1051.

[80] L. Lu, R. Cai, and D. Gursoy. "Developing and Validating a Service Robot Integration Willingness Scale". In: *International Journal of Hospitality Management* 80 (July 2019). DOI: 10.1016/j.ijhm.2019.01.005.

[81] R. L. Oliver. "Measurement and evaluation of satisfaction processes in retail settings". In: *Journal of Retailing* 57.3 (1981), pp. 25–48.

[82] A. Oghuma, C. Libaque-Saenz, S. F. Wong, and L. Y. Chang. "An Expectation-Confirmation Model of Continuance Intention to Use Mobile Instant Messaging". In: *Telematics and Informatics* 33 (May 2015), pp. 34–47. DOI: 10.1016/j.tele.2015.05.006.

[83] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay. "Trust in a specific technology". en. In: *ACM Trans. Manag. Inf. Syst.* 2.2 (2011), pp. 1–25.

[84] C. Nordheim, A. Følstad, and C. Bjørkli. "An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study". In: *Interacting with Computers* 31 (May 2019), pp. 317–335. DOI: 10.1093/iwc/iwz022.

[85] V. K. Pham, T. Thi, and N. Duong. "A Study on Information Search Behavior Using AI-Powered Engines: Evidence From Chatbots on Online Shopping Platforms". In: *SAGE Open* 14 (Nov. 2024). DOI: 10.1177/21582440241300007.

[86] G. L. Polites. "The duality of habit in information technology acceptance". In: (2009).

[87] M. Alsharo, Y. Alnsour, and M. Alabdallah. "How habit affects continuous use: evidence from Jordan's national health information system". en. In: *Inform. Health Soc. Care* 45.1 (2020), pp. 43–56.

[88] W. Aslam, D. Siddiqui, I. Arif, and D.-K. Farhat. "Chatbots in the frontline: drivers of acceptance". In: *Kybernetes* 52 (Apr. 2022). DOI: 10.1108/K-11-2021-1119.

[89] L. Meyer-Waarden, G. Pavone, T. Poocharoentou, P. Prayatsup, M. Ratinaud, A. Tison, and S. Torné. "How Service Quality Influences Customer Acceptance and Usage of Chatbots?" In: *Journal of Service Management Research* 4 (Jan. 2020), pp. 35–51. DOI: 10.15358/2511-8676-2020-1-35.

[90] B. Weiss, I. Wechsung, C. Kühnel, and S. Möller. "Evaluating embodied conversational agents in multimodal interfaces". In: *Computational Cognitive Science* 1 (Dec. 2015). DOI: 10.1186/s40469-015-0006-9.

[91] A. Holzinger. "Usability engineering for software developers". In: *Communications of the ACM* 48.1 (2005), pp. 71–74.

[92] A. Naumann and I. Wechsung. "Developing usability methods for multimodal systems: The use of subjective and objective measures". In: *Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM)*. 2008, pp. 8–12.

[93] D. Kahneman, B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier. "When more pain is preferred to less: Adding a better end". In: *Psychological Science* 4 (1993), pp. 401–405.

[94] Y.-W. Cheung and D. Friedman. "Individual Learning in Normal Form Games: Some Laboratory Results". In: *Games and Economic Behavior* 19 (1997), pp. 46–76.