Exploring Customer Service Agent Preferences for Conversational and Keyword-Based Information Retrieval

Nektarios Machner[a], Yaren Mändle[b] and Florian Matthes[c]
Department of Computer Science, Technical University of Munich, Munich, Germany

Keywords: Conversational Search, Knowledge Discovery, Knowledge Management, Information Retrieval, Customer Service.

Abstract: Effective knowledge discovery and information retrieval drive organizational innovation and competitive advantage. To support this, organizations have long used knowledge management systems that historically have relied on keyword-based search. The rise of artificial intelligence (AI), most notably large language models (LLMs), has enabled conversational search (CS) interfaces that understand natural-language queries, synthesize information from multiple sources, and generate answers. This study investigates the factors that influence customer service agents' preferences for conversational search versus traditional keyword-based search within an internal knowledge management system. Set in a large European insurance company, we employ a mixed-methods empirical approach, integrating semi-structured interviews (n = 13), a structured survey (n = 17), and log-file analysis of 508 real-world queries. Our research explores which factors drive agents' choice between the two search approaches, and examines the practical strengths and limitations of each approach. Our findings reveal that agents choose keyword search when they are confident of where to look and conversational search when they need natural-language guidance, with trust and time constraints further tipping the balance. This complementarity suggests hybrid interfaces, blending ease of use, reliable results, and flexible query handling, best support agents' workflows.

1 INTRODUCTION

Knowledge is a critical asset for modern organizations: "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" (Fowler, 2000). To facilitate effective knowledge sharing, many organizations deploy knowledge management systems that enable employees to retrieve information on demand.

Historically, the search process depended on keyword-based search, requiring users to select precise terms or Boolean operators to locate documents. While effective for well-defined queries, this approach often struggles with synonyms, polysemy, and contextual nuances. In response to the limitations of traditional keyword-based retrieval, the emergence of artificial intelligence (AI), particularly large language models (LLMs), has enabled conversational search interfaces that transform how employees interact with knowledge management systems. By understanding queries in natural language, LLM-based conversational search allows employees to simply describe their information needs in everyday terms, rather than crafting precise keyword strings. The model then aggregates relevant passages across multiple retrieved documents and generates an answer, while highlighting implicit connections in the data and further offering the option for clarifying follow-up queries.

However, successful deployment of AI-based systems not only depends on technology but also on employee acceptance. While there is extensive research on customer-facing conversational interfaces, far less attention has been paid to how employees integrate AI-based conversational search into their existing internal knowledge management workflows.

In this study, we investigate within the context of a large European insurance company the adoption of LLM-based conversational search in knowledge discovery and information retrieval compared with traditional keyword-based search. To this end, we defined

[a] https://orcid.org/0009-0001-8359-6668
[b] https://orcid.org/0009-0007-1087-5707
[c] https://orcid.org/0000-0002-6667-5452

Machner, N., Mändle, Y. and Matthes, F.
Exploring Customer Service Agent Preferences for Conversational and Keyword-Based Information Retrieval.
DOI: 10.5220/0013721400004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 349-356

the following research questions (RQs):

- **RQ1:** Which factors influence customer service agents' choice between the conversational search and keyword-based search?
- **RQ2:** What are the strengths and limitations of LLM-based conversational search versus traditional keyword-based search when integrated into existing knowledge management systems?

2 RELATED WORK

As we investigate the adoption of conversational search versus traditional keyword-based search in customer service knowledge management, it is important to contextualize our work within existing studies on these retrieval modalities across various domains.

- Information access via conversational agents and traditional keyword search was compared by (Preininger et al., 2021) within a widely used pharmacologic knowledge base. They found that for certain topics, users accessed information more frequently with the conversational agent, while other topics saw higher access rates under the keyword-based approach. However, their study did not explore why users chose one method over the other, nor did it assess the usability or user satisfaction associated with each search approach.
- A conversational search system for exploring scholarly publications using a knowledge graph was developed by (Schneider and Matthes, 2024). They evaluated it through a user study with 40 participants, comparing it to a traditional graphical interface with keyword search. Their results indicate that the conversational interface enables more effective discovery of research publications. However, unlike our study, their work focused on individual private users in an academic search context, rather than on enterprise users in a knowledge management setting.
- A survey of college students by (Sakirin and Ben Said, 2023) compared ChatGPT-powered conversational interfaces with traditional keyword search. Using descriptive and inferential statistics, they found that most participants favored the ChatGPT interface for its convenience and efficiency. In contrast to our study, their work examined individual private users instead of enterprise users operating within a knowledge management environment.
- A user study conducted by (Liu et al., 2021) compared interaction behaviors and explicit feedback when searching legal cases via a traditional keyword system versus a conversational search interface. They tracked both search interactions and outcome metrics, finding that participants achieved higher retrieval performance with the conversational system. While their results underscore the potential of conversational agents for improving search effectiveness, their work is situated in a legal case retrieval context, whereas our study focuses on customer service agents navigating an insurance-domain knowledge management system.
- Traditional keyword search versus LLM-based search for image geolocation tasks was evaluated by (Wazzan et al., 2024), asking users to pinpoint where an image was taken. They examined both task performance and how users adjusted their query strategies, finding that keyword search yielded more accurate location predictions than the LLM-based approach. While their work highlights differences in tool effectiveness and user behavior, it differs from ours in that we compare these search modalities within an organizational knowledge-management setting and assess the LLM-based system using metrics such as perceived ease of use and answer relevance rather than geolocation accuracy.
- Two online experiments run by (Spatharioti et al., 2023) compared traditional keyword search and an LLM-based tool for consumer-product research. They found that the LLM interface enabled faster task completion with fewer but more complex queries, although participants sometimes overrelied on incorrect model outputs. Unlike their lab-style experiments, our study examines customer service agents in an organizational insurance context using interviews, surveys, and log analysis.

3 METHODOLOGY

3.1 Case Study Context

As this study is conducted within the context of our case study company, relevant context information is disclosed below:

- All data is collected within the context of a large European insurance company, more specifically within the German branch of customer care.
- Their in-house developed knowledge management software serves as the central help system for internal use supporting employees across

all divisions in case processing. It has been in use since 2008 and consists of over 168,000 pages across approximately 5,600 documents with roughly 40,000 active users.

- The software was extended in 2023 by adding conversational search to support customer service agents in efficiently finding information to save time and effort, especially when handling customer service requests simultaneously.
- The conversational search was implemented in addition to the traditional keyword-based search and not as a substitute. Customer service agents are free to choose which system to use based on their needs.

3.2 Search Workflow & IT Architecture

To better understand the differences between the two solution approaches, a typical workflow and the underlying architecture of each will be briefly highlighted below.

3.2.1 Conversational Search

Workflow

- When a customer service agent submits a query, a retriever model first fetches documents accessible to the user's authentication group. A ranking model then prioritizes the most relevant results. Finally, an OpenAI GPT-4-based LLM generates a response, including references to the top three retrieved documents (OpenAI, 2024).
- Customer service agents rate answers on a 1–5 scale, where 5 indicates high satisfaction, and mark each retrieved document as relevant or not using thumbs up/down. This feedback is used to improve the conversational search pipeline.

Architecture

The conversational search assistant consists of multiple components and follows a typical Retrieval-Augmented Generation (RAG) pipeline architecture.

- A centralized authentication service is used to manage users and their access rights.
- All documents of the knowledge management system are stored in the file storage and embedded in a vector database.
- Retriever models are used to search the vector database and retrieve relevant documents.
- Ranker models are used to rank the retrieved documents based on relevance to the query.
- An LLM (GPT-4) is used to generate answers from the retrieved results (OpenAI, 2024).
- Feedback from the customer service agents is stored in a database, where it is later reviewed by experts and used to generate a fine-tuning dataset to continuously refine and improve the three types of models used throughout the pipeline.

3.2.2 Keyword-Based Search

Workflow

- A customer service agent can type keywords into an input field and filter by metadata. The search engine checks for the presence of a keyword within all of its indexed documents and evaluates its frequency. The algorithm returns relevant documents to the customer service agent.

Architecture

- Users are authenticated by their email address and global identifiers to determine which documents they are allowed to access.
- Documents are stored in JSON format with a unique identifier, mandatory fields such as title, and metadata.
- All search queries are processed by the search engine in Python.
- Search results are ranked by assigning weights that consider matches in titles, headings, body text, the document type, and term frequency.
- An auto-complete function suggests next words during typing, and filters can narrow results based on metadata.

3.3 Literature Review

To design our interview and survey questions, we first conducted a literature review to identify relevant evaluation criteria for comparing conversational and keyword-based search. We consolidated our findings and adapted multiple metrics based on prior systematic literature reviews on conversational agent adoption (Ling et al., 2021; Lewandowski et al., 2021). These works identify a range of factors, including user-related, agent-related, and attitude-based dimensions (Ling et al., 2021), as well as organizational, technical, and environmental drivers (Lewandowski et al., 2021).

Unlike these reviews, our work empirically examines customer service agents' preferences between conversational and keyword-based search within an insurance company's knowledge management system.

3.4 Interview Design

We conducted semi-structured interviews with customer service agents who had used both the traditional keyword-based search and the more recently implemented conversational search. Thirteen agents participated, varying in age, gender, and professional experience. To understand when agents preferred one approach over the other, we defined seven scenarios, each corresponding to a search query type adapted from real log data. For each scenario, interviewees indicated their preferred search method and explained why. Since query types can overlap, a single query may belong to multiple categories. Each category is briefly defined (Def.) and illustrated with an example (Ex.) below.

Simple Query

- Def.: Requires manual lookup in a single document of the knowledge base by the agent.
- Ex.: Are bikes insurable under private insurance?

Complex Query

- Def.: Requires more intensive research, such as looking up multiple documents or entries in the knowledge base.
- Ex.: How can I insure a minor policyholder?

Close-Ended Query

- Def.: Polar questions answered with 'yes' or 'no'.
- Ex.: Is Parkinson's disease a chronic illness?

Open-Ended Query

- Def.: Requires more detailed and extensive answers.
- Ex.: What do I need to consider as a buyer or seller during a change of ownership?

Short Query

- Def.: Search queries containing no more than about ten words.
- Ex.: How long is the immediate coverage valid?

Long Query

- Def.: Search queries containing more than ten words.
- Ex.: Are damages caused by my pet, such as bite injuries or property damage, covered under liability insurance?

Procedural Query

- Def.: Requires guidance or a description of how to perform a specific task step-by-step.
- Ex.: How do I withdraw a balance?

After going through all scenarios, we asked the interviewees more generally which factors influence their choice between conversational search and keyword-based search, and what they believe their respective strengths and limitations are.

3.5 Survey Design

After collecting qualitative feedback through the semi-structured interviews, we sent out an online survey to all interviewees and further customer service agents to also collect quantitative feedback. From the original 13 interviewees, 12 also filled out the survey, as well as an additional 5 customer service agents who could not participate in our interviews, for a total of 17 survey participants. The survey consisted of 20 statements mapped to 11 evaluation metrics, each rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

3.6 Log Analysis

To gather feedback on the use of conversational search, the case study company integrated a feedback mechanism into the software that logs evaluation data in JSON format. These logs include the original search queries, the system-generated answers, the three retrieved documents, and customer service agents' ratings of both the answers and the document suggestions. Answers are rated on a 1–5 scale, with higher scores indicating greater satisfaction. Retrieved documents are rated with a thumbs up (good) or thumbs down (bad). We analyzed logs collected between February and November 2024 and compared the findings with insights from the interviews and survey.

4 RESULTS & DISCUSSION

4.1 Interview Results

We interviewed 13 customer service agents and asked them about their preferences regarding keyword-based and conversational search for seven different scenarios. The findings regarding each scenario are presented and discussed individually below.

4.1.1 Simple Query

For simple queries, 7 out of 13 agents preferred conversational search, noting its ability to distinguish between policy types (e.g., liability vs. property insurance) and to pinpoint relevant coverage details. When

agents are uncertain of the precise keywords to use, they consider conversational search more effective, as it supports queries in natural language rather than requiring the exact terms. The 6 agents who preferred keyword search stated speed and familiarity with the tool as their reasons, especially during ongoing calls with customers, as typing a few keywords is more efficient than typing a full question. Agents familiar with the tool also generally know where to find the necessary information quickly and directly as they have consulted the documentation multiple times.

4.1.2 Complex Query

For complex queries, all 13 agents preferred conversational search over keyword-based search. As complex queries require looking up multiple documents, agents value conversational search for its ability to quickly aggregate relevant information from multiple sources, enabling them to review related topics and assess their relevance. By contrast, using keyword-based search for such queries is regarded as tedious and time-consuming, as it requires locating multiple documents and manually evaluating each one.

4.1.3 Close-Ended Query

For close-ended queries, 10 agents preferred conversational search, while the remaining 3 opted for keyword-based search. The main reason for CS is its ability to directly provide a 'yes' or 'no' answer without having to look up documents, thereby saving time and effort. The agents favoring keyword-based search were all sufficiently familiar with the knowledge base to already know which documents to look for, suggesting their preference stemmed from familiarity with the system rather than keyword-based search being better suited to the task.

4.1.4 Open-Ended Query

For open-ended queries, 7 out of 13 agents preferred keyword-based search, stating that conversational search provides only summarized answers, whereas they require more detailed information. They either know where to find it due to their familiarity with the knowledge base or are willing to look it up manually, considering the inquiries important enough to justify the extra effort. The agents preferring conversational search explained that, lacking precise keywords, they would not know what to look for and therefore would use the conversational search and its document suggestions as a starting point for deeper exploration into the knowledge base.

4.1.5 Short Query

For short queries, 10 out of 13 agents favor conversational search, with the main reason being its ability to quickly and directly produce an answer to the search query. Agents found it particularly effective for broad or general topics, trusting the system to provide accurate responses. Conversely, the 3 agents who opted for keyword search did so out of habit and confidence in their existing retrieval strategies for quick lookups.

4.1.6 Long Query

For long queries, 7 of the 13 agents favored conversational search while 6 preferred keyword search. Those who chose CS believed that, when the query is well constructed, it results in precise answers, especially for newer policies with predefined responses, making it easier and faster to locate needed information. Nevertheless, some agents noted that they still revert to keyword search for more complex issues to verify the accuracy of the CS results.

4.1.7 Procedural Query

For procedural queries, 7 of the 13 agents preferred conversational search, 4 chose keyword search, and 2 were undecided. Those favoring CS appreciated that it delivers focused, relevant results without the broad, unfocused listings typical of keyword search, which require clicking through multiple links to assess relevance. Many agents also reported encountering such procedural questions for the first time and not knowing where to begin with keyword search; in these cases, they found CS faster and more intuitive for matching information to their query before diving into the full document. Conversely, agents unfamiliar with using CS for procedural tasks expressed skepticism about its accuracy and therefore preferred the reliability of keyword search.

4.1.8 Influencing Factors

After the predefined scenarios, we asked agents about their general preferences between conversational and keyword-based search. Their choice largely depends on familiarity with the knowledge base and confidence in locating information. Agents who know where to look tend to favor keyword search for its speed and reliability, while those less certain about a topic or its structure prefer conversational search for its guided, natural-language interface.

Query complexity also shapes preferences. For broad or complex questions, especially those involving multiple documents or recent policies, agents value CS's ability to surface relevant passages and

suggest follow-up prompts. In contrast, for routine or well-defined queries, particularly under time pressure (e.g., live calls), keyword search remains the go-to option. Agents also report switching back to keyword search to verify CS responses on critical or unfamiliar issues.

Trust and usability further influence adoption. Some agents hesitate to rely on CS until its accuracy and document coverage, especially for older policies, improve. Keyword search, by contrast, benefits from long-standing trust in its precision. Agents who find CS intuitive and are open to new tools are more likely to adopt it, highlighting the importance of clear trust indicators, comprehensive document inclusion, and seamless workflows to encourage broader use.

4.2 Survey Results

The aggregated survey results from 17 participants, including the mean and standard deviation for each evaluation metric, are presented and discussed individually below.

Perceived Ease of Use (Davis, 1989)
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 low variability, meaning that most agents rated the system similarly.

Performance (Peras, 2018)
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 (Saad-Falcon et al., 2024)
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. A standard deviation of 0.70 indicates moderate variability, suggesting that while some agents find the system's answers faithful, others hold differing views.

Answer Relevance (Saad-Falcon et al., 2024)
The answer relevance metric has a mean score of 3.65, suggesting that customer service agents generally agree that CS provides relevant answers. However, the moderate standard deviation of 0.71 indicates some variability, with some agents differing in their perception of answer relevance.

Context Relevance (Es et al., 2024)
The context relevance metric has a mean score of 3.29, reflecting a mostly neutral perception. The standard deviation of 0.88 shows notable variability in perceptions, suggesting mixed views among agents.

Satisfaction (Oliver, 1981)
The satisfaction metric has a mean score of 3.59, implying a slightly positive perception regarding satisfaction. The moderate standard deviation of 0.71 indicates moderate variability, suggesting that while many agents view CS positively, opinions are not uniform.

Perceived Usefulness (Davis, 1989)
The mean score of the perceived usefulness metric is 3.47, indicating a moderate perception. The standard deviation of 1.57 is very high, reflecting significant variability in responses and, thus, differences in how useful agents perceive conversational search.

Quality (Oghuma et al., 2015)
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 (Peras, 2018)
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 is 0.89, representing moderate variability. This indicates that the agents have diverse opinions regarding the system's value to the organization.

Openness to New Technologies (Mcknight et al., 2011)
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
The metric "Replaceability and Necessity of CS" has a mean score of 3.27, suggesting a moderate agreement on the system's necessity and replaceability. The standard deviation of 0.95 indicates moderate variability, reflecting differing views on the matter.

4.3 Log Analysis Results

The log files we analyzed contained a total of 508 queries spanning the time frame from February 2024 to November 2024. We manually categorized a sample size of 400 queries into the seven types we previously defined for our interviews, whereas each query could belong to multiple categories. Table 1 shows an overview of the categorization and the frequency of search queries.

Table 1: Categorization of Search Queries in the Logs.

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%)

As can be seen in the table, a large amount of queries were classified as simple, close-ended, or short, whereas only seven queries were procedural. This aligns with the interview results that conversational search is preferred more for short and close-ended queries. 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 not complete sentences.

Next, we examined customer service agents' evaluations of the answers generated by the LLM. Of the 508 total queries, agents rated 503 of them. Table 2 presents a summary of these ratings.

Table 2: User Ratings for Answers Generated by the LLM.

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

On a scale of 1 to 5, with 5 meaning the agent was highly satisfied with the answer, slightly more than a third of the answers received the highest possible rating. Interestingly, roughly half the answers received the worst possible rating. Overall, the distribution suggests that answers were either fully satisfying or not satisfying at all, leaving little middle ground in between.

Finally, we assessed the agents' ratings of the retrieved documents that they rated with either a thumbs up or a thumbs down. Note that if at least one document was evaluated with a thumbs up, the feedback system indicated a success rate of 100% for that query, as the agents could find the answer to their questions in one of the documents. Table 3 shows a summary of the document ratings.

Table 3: User Ratings for Document Relevance.

Rating	Number of Answers
Thumbs Up	203 (41.01%)
Thumbs Down	292 (58.99%)

After analyzing the logs, we find that the ratings and thumbs-up/thumbs-down results do not fully align with the interview or survey results. For five out of seven scenarios we examined in the interviews, the majority of the agents preferred conversational search over keyword-based 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 conversational search 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 conversational search system improves their task efficiency and work performance. While the mean scores from the survey results showed an overall moderately positive perception of conversational search, 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.

5 CONCLUSION

Our mixed-methods investigation shows that customer service agents' choice between conversational and keyword-based search is driven primarily by their familiarity with the knowledge base and confidence in locating information. Agents who know where to look tend to default to keyword search, while those less certain rely on conversational search's natural-language guidance. Trust concerns, particularly for open-ended or complex queries, prompt some agents to cross-check conversational outputs with keyword results, and time pressure further influences preferences: conversational search excels at handling short, complex, or close-ended queries efficiently, whereas keyword search remains the go-to under live-call conditions when precise document retrieval is paramount. Adoption of conversational search also aligns with perceived ease of use, answer faithfulness, and time-saving benefits, and is stronger

among agents open to new technologies. These findings underscore the complementary strengths of both modalities and suggest that enhancing trust indicators, refining usability, and integrating hybrid search interfaces will better support agent workflows.

Limitations

Our study is subject to the following limitations:

- **Scope & Applicability:** This study is confined to our single case study company operating in the insurance domain and therefore may not generalize to other industries or organizations. Moreover, the integration of conversational search is still in its test phase at this company. Agent attitudes and preferences may change over time as they become more familiar with the system.
- **Sample Size:** The limited number of interviewees (n=13) and survey participants (n=17) may restrict the generalizability of our findings.

ACKNOWLEDGEMENTS

We would like to thank our case study company and its employees who participated in the interviews and made this study possible. Generative AI (GPT-4) was used in this study for the conversational search as described above. ChatGPT (https://chatgpt.com/) was used minimally for wording and phrasing of this paper, with full responsibility for the content, interpretation, and final version remaining with the authors.

REFERENCES

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Fowler, A. (2000). The role of ai-based technology in support of the knowledge management value activity cycle. *The Journal of Strategic Information Systems*, 9(2):107–128.

Lewandowski, T., Delling, J., Grotherr, C., and Böhmann, T. (2021). State-of-the-art analysis of adopting ai-based conversational agents in organizations: A systematic literature review. In *Proceedings of the 25th Pacific Asia Conference on Information Systems (PACIS 2021)*, page 167. Association for Information Systems.

Ling, E. C., Tussyadiah, I., Tuomi, A., Stienmetz, J., and Ioannou, A. (2021). Factors influencing users' adoption and use of conversational agents: A systematic review. *Psychol. Mark.*, 38(7):1031–1051.

Liu, B., Wu, Y., Liu, Y., Zhang, F., Shao, Y., Li, C., Zhang, M., and Ma, S. (2021). 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.

Mcknight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.*, 2(2).

Oghuma, A., Libaque-Saenz, C., Wong, S. F., and Chang, L. Y. (2015). An expectation-confirmation model of continuance intention to use mobile instant messaging. *Telematics and Informatics*, 33:34–47.

Oliver, R. L. (1981). Measurement and evaluation of satisfaction processes in retail settings. *Journal of Retailing*, 57(3):25–48.

OpenAI (2024). Gpt-4 technical report.

Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97.

Preininger, A. M., Rosario, B. L., Buchold, A. M., Heiland, J., Kutub, N., Bohanan, B. S., South, B., and Jackson, G. P. (2021). Differences in information accessed in a pharmacologic knowledge base using a conversational agent vs traditional search methods. *International Journal of Medical Informatics*, 153:104530.

Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Sakirin, T. and Ben Said, R. (2023). User preferences for chatgpt-powered conversational interfaces versus traditional methods. *Mesopotamian Journal of Computer Science*, 2023:24–31.

Schneider, P. and Matthes, F. (2024). Conversational exploratory search of scholarly publications using knowledge graphs. In Abbas, M. and Freihat, A. A., editors, *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 384–396, Trento. Association for Computational Linguistics.

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., and Hofman, J. M. (2023). Comparing traditional and llm-based search for consumer choice: A randomized experiment.

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