Investigating the Adoption of Conversational Search by Customer Service Agents

Yaren Maendle

March 03, 2025, Final Presentation Master's Thesis

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

1. Background and Motivation
2. Research Questions
3. Methodology
4. Results
5. Conclusion

Background

Initial situation	• Clerks search too long and inefficiently for information via the full-text search • Newly hired clerks require even more effort to find relevant information
Problem/Consequences	• Inefficient process • Dissatisfaction / frustration among clerks and customers when workload is high
Solution	• AI-based chatbot for intelligent and optimized search for relevant information / work instructions • Training of AI through test fields

Outline

1. Background and Motivation
2. Research Questions
3. Methodology
4. Results
5. Conclusion

Research Questions

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

2. How can an LLM-based conversational agent be evaluated?

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

Outline

1. Background and Motivation
2. Research Questions
3. Methodology
4. Results
5. Conclusion

Interview Design

Introduction

Individual Characteristics: age, gender, work experience in the company, chatbot experience, frequency, CS usage period

Seven Scenarios: simple, complex, open-ended, close-ended, long, short, procedural

Additional Questions: factors influencing the choice, strengths/limitations, additional features

Survey Design

Perceived Ease of Use: Minimal effort, ease of use (Davis, 1989)

Performance: Completeness, promptness, appropriateness (Peras, 2018)

Answer Faithfulness: Alignment with retrieved context (Ares, 2023)

Answer Relevance: Correspondance to the question (Ares, 2023)

Context Relevance: Relevance of the retrieved context (James et al., 2023)

Satisfaction: Expectations, emotions, prior experience (Oliver, 1981)

Survey Design

Perceived Usefulness: Belief in performance enhancement (Davis, 1989)

Quality: Perceived service superiority (Oghuma et al., 2015)

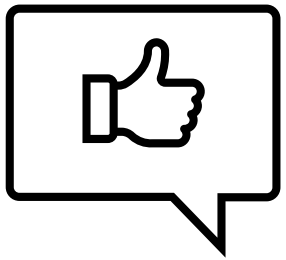
Business Value: Effectiveness vs. cost (Peras, 2018)

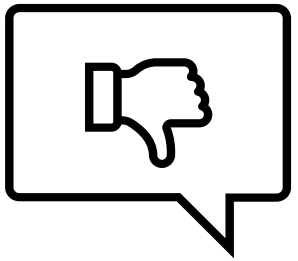
Openness to New Technologies (Mcknight, 2011)

Replaceability and Necessity of CS: Willingness to adopt technology

Analysis of Customer Agent Logs

- **Data Source:** JSON logs from a data labeling tool (February-November 2024)
- **Logged Information:** Queries, timestamps, CS answers, document suggestions, agent ratings
- **Rating System:**
- Answers: 1-5 scale (5 = satisfied)
- Document Suggestions: Thumbs up ("good") / Thumbs down ("bad")
- Manual categorization of queries





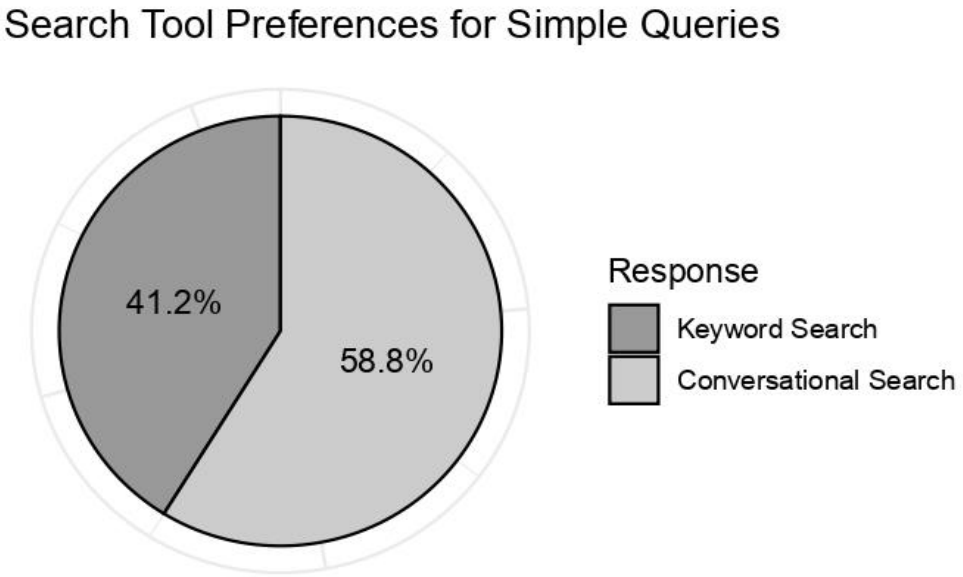
Analysis of Customer Agent Logs

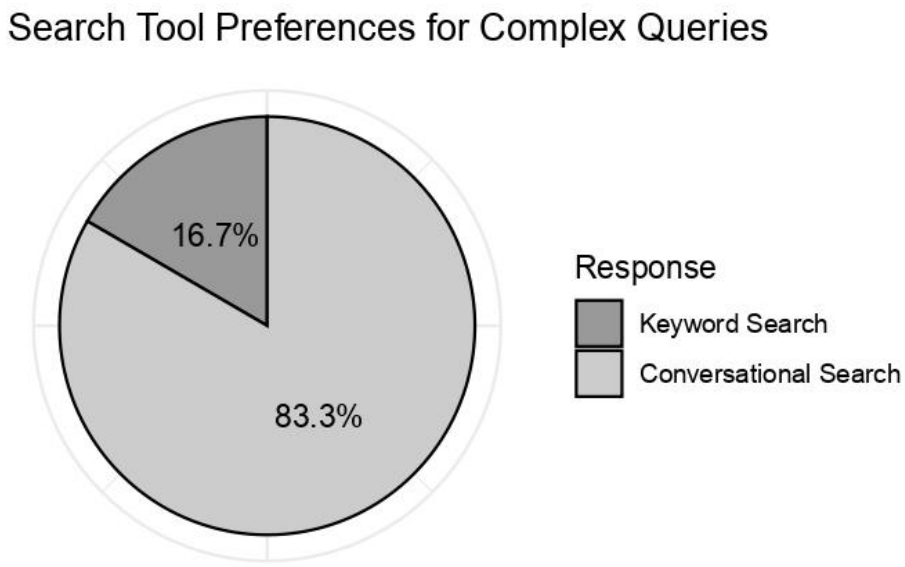

```
"result": [{
        "id": "E2u4YMMmEp",
        "type": "rating",
        "value": {
            "rating": 5
        },
        "origin": "manual",
        "to_name": "answer",
        "from_name": "rating"
    }, {
        "id": "AiHiMs3tw8",
        "type": "ranker",
        "value": {
            "ranker": {
                "_": [],
                "bad": ["1557f46c88081c1ae66db2a140c6b6c8", "613a2f5dc1ee47dc923e799467cd287f", "7cc2ac9fafcb1bd7553ab2d906a7e9ec"],
                "good": []
            }
        },
"predictions": [],
"data": {
    "stars": null,
    "answer": "Nein, E-Scooter sind nicht im Privatschutz versicherbar.",
    "question": "Sind E-Scooter in Privatschutz versicherbar?",
    "documents": [{
            "id": "1557f46c88081c1ae66db2a140c6b6c8",
            "html": "<a href=\"https:\/\/awiso.allianz.de.awin\/#\/docat\/PH-ABC\/Gebrauch%20bestimmter%20nicht-versi
            "title": "PH-ABC ; Fachlicher Inhalt ; Fahrzeuge ; Landfahrzeuge ; Gebrauch bestimmter nicht-versicherung
        }, {
```


Outline

1. Background and Motivation
2. Research Questions
3. Methodology
4. Results
5. Conclusion

Interview Results: Scenarios: Simple and Complex Queries

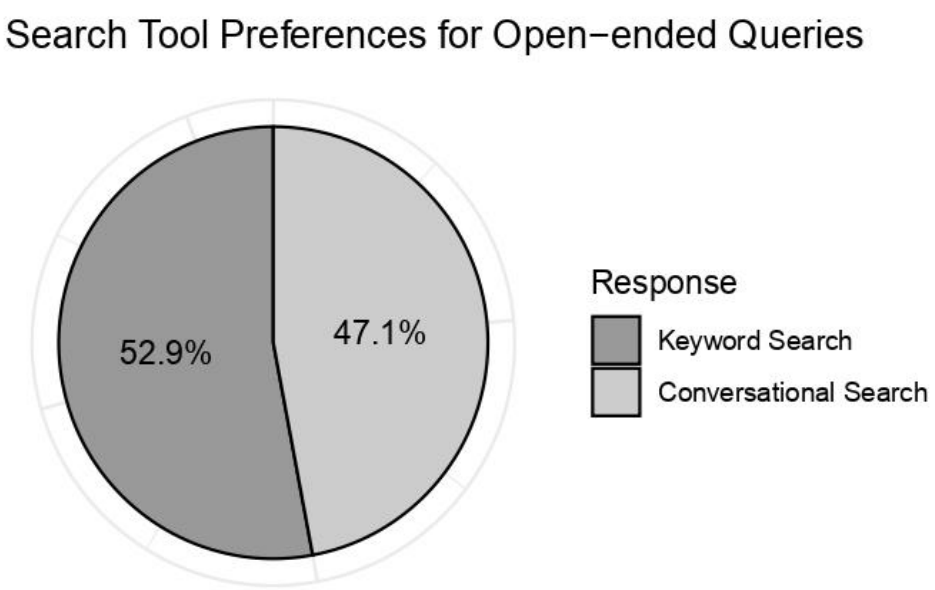


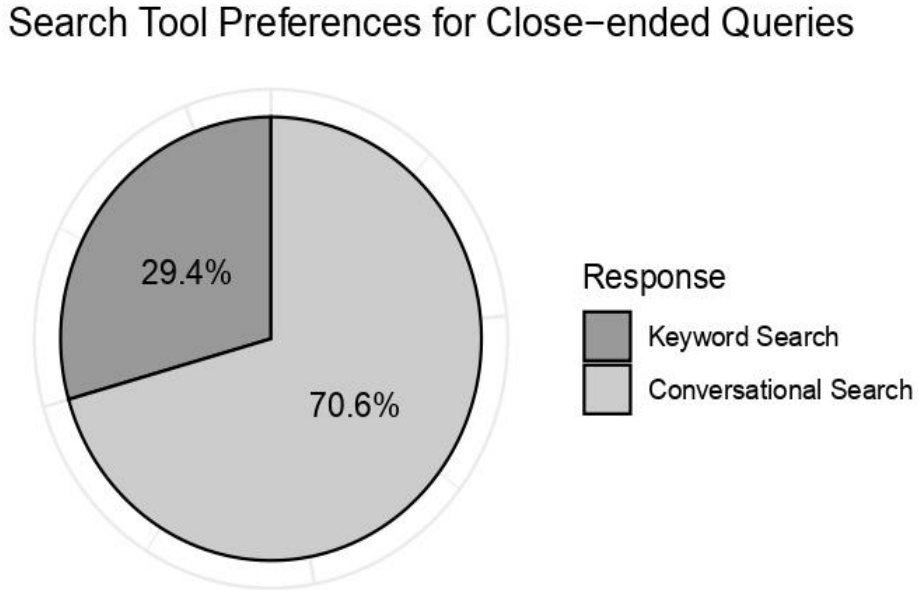


Main reason for the choice:
KS: familiarity with the topic
CS: not being familiar with the topic

Main reason for the choice:
CS: Seeing all relevant documents at once

Interview Results: Scenarios: Open-ended and Close-ended Queries

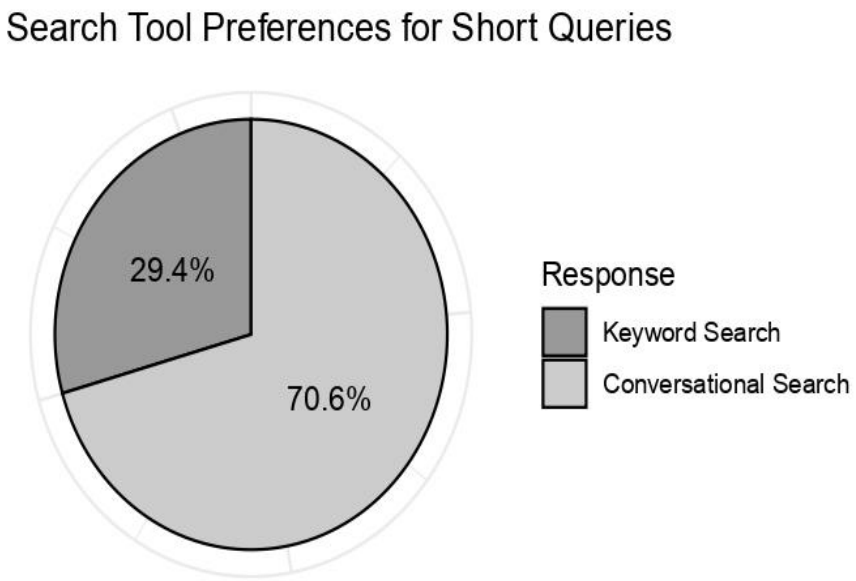


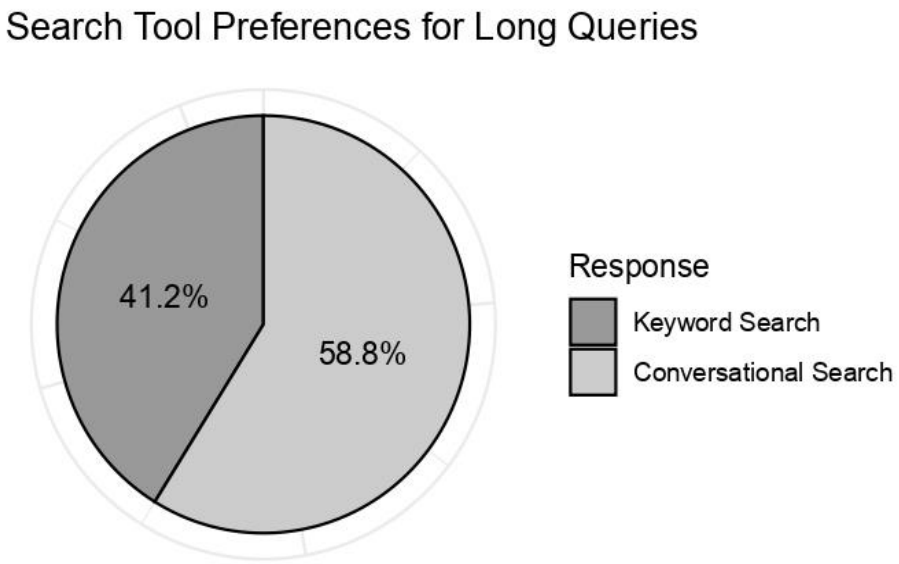


Main reason for the choice:
KS: familiarity with the topic, not trusting CS
CS: not being familiar with the topic

Main reason for the choice:
KS: familiarity with the topic
CS: efficient, quick, minimal effort

Interview Results: Scenarios: Short and Long Queries

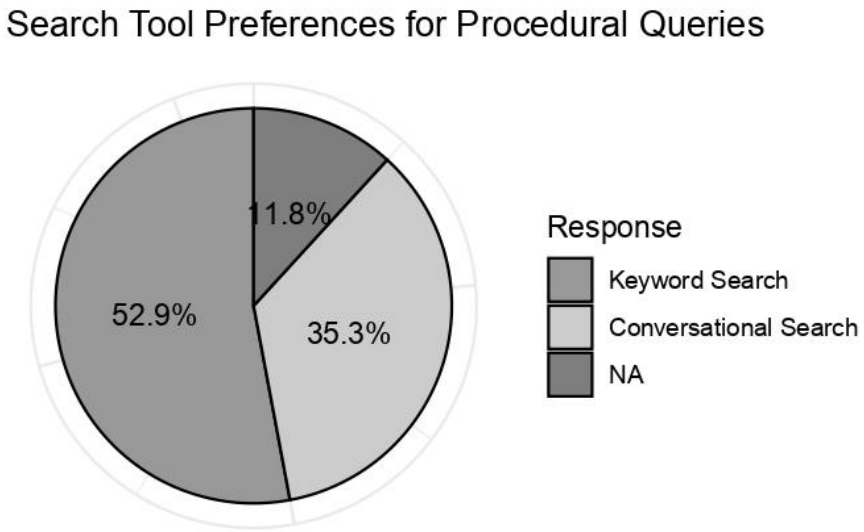




Main reason for the choice:
KS: familiarity with the topic
CS: efficient, quick, minimal effort

Main reason for the choice:
KS: not trusting CS
CS: if query formulated well enough, CS would give a precise answer

Interview Results: Scenarios: Procedural Queries



Main reason for the choice:
KS: familiarity with the topic
CS: not being familiar with the topic

Interview Results: Additional Questions

Factors Affecting the Choice:

- Familiarity with the tool
- Confidence in existing knowledge
- Trust
- Type of query
- Speed and time pressure

Strengths of the Keyword Search:

- Familiarity
- Broadness
- Reliability
- Speed

Limitations of the Keyword Search:

- Necessity of knowing the exact term to use
- Reviewing large number of results
- Lack of ability to interpret the context
- Hard to navigate

Interview Results: Additional Questions

Strengths of the CS:

- Ease of use and efficiency
- Well-formulated summaries and answers
- Comprehensive results that address various aspects of a query

Limitations of the CS:

- Slow response time for detailed queries
- Older policies not included
- Writing in full sentences instead of keywords
- Accuracy issues

Additional Feature Wishes:

- Voice interaction
- Multiple chat sessions/ category-specific chats
- Improved user interface design
- Confidence levels for response accuracy
- Dynamic follow-up questions

Correlation Results Between Individual Characteristics and Choices (n=12)

	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

- Openness to new technologies → Increased use of CS

Note:
- Closer to +1 → Strong positive relationship
- Closer to -1 → Strong negative relationship
- *p < 0.05, **p < 0.01, ***p < 0.001

Correlation Results Between Choice Cases and Individual Characteristics (n=12)

	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

- Openness to new technologies → preferring CS for open-ended questions
- CS usage duration → not preferring CS for open-ended questions

Note:
- Closer to +1 → Strong positive relationship
- Closer to -1 → Strong negative relationship
- *p < 0.05, **p < 0.01, ***p < 0.001

Correlation Results Between the Evaluation Metrics (n=17)

	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											

- Perceived Usefulness - Business Value
- Openness to New Technologies - Replaceability/Necessity of CS
- Perceived Usefulness - Replaceability/Necessity of CS
- Answer Relevance - Perceived Usefulness
- Ease of Use - Satisfaction
- Ease of Use - Quality

Note:
- Closer to +1 → Strong positive relationship
- Closer to -1 → Strong negative relationship
- *p < 0.05, **p < 0.01, ***p < 0.001

Correlation Results Between Choices and Evaluation Metrics (n=17)

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*

- ease of use, answer faithfulness, satisfaction, perceived usefulness, business value, replaceability and necessity of CS → Increased use of CS

Note:
- Closer to +1 → Strong positive relationship
- Closer to -1 → Strong negative relationship
- *p < 0.05, **p < 0.01, ***p < 0.001

Correlation Results Between Choice Cases and Evaluation Metrics (n=17)

	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

- Simple Question: answer faithfulness → preferring CS
- Complex Question: ease of use, perceives usefulness, business value → preferring CS
- Open-ended Question: ease of use, answer faithfulness, quality, business value → preferring CS
- Short Question: performance, answer faithfulness, quality → preferring CS
- Procedural Question: ease of use → preferring CS

Note:
- Closer to +1 → Strong positive relationship
- Closer to -1 → Strong negative relationship
- *p < 0.05, **p < 0.01, ***p < 0.001

Survey Results

Evaluated Metric	Mean	SD
Perceived Ease of Use	4.24	0.39
Performance	3.55	1.13
Answer Faithfulness	3.56	0.70
Answer Relevance	3.65	0.71
Context Relevance	3.29	0.88
Satisfaction	3.59	0.71
Perceived Usefulness	3.47	1.57
Quality	3.69	0.8
Business Value	3.71	0.89
Replaceability and necessity of CS	3.09	0.94

Overall moderate to slightly positive perception of CS

Findings from Customer Agent Log Analysis

Rating	Number of Answers
1 (not satisfied)	251 (49.9%)
2	22 (4.37%)
3	39 (7.75%)
4	10 (1.99%)
5 (satisfied)	181 (35.98%)

Rating	Number of Answers
All 3 documents thumbs down	292 (58.99%)
At least one document thumbs up	203 (41.01%)

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%)

Misalignment with survey results:

- Memory bias
- Sample size
- Recency bias

Outline

1. Background and Motivation
2. Research Questions
3. Methodology
4. Results
5. Conclusion

Conclusion

Key Results:

- CS is favored for complex, close-ended and short questions with a very low sd.
- Other than the query type, familiarity and trust play a big role. Users hesitate to shift to a new system.
- Interaction data and user perceptions can be misaligned, emphasizing the need for both objective measures (like interaction logs) and subjective measures (like surveys).

Limitations:

- Sample Size
- Scope of the Study
- Survey and Interview Instruments
- Early Adoption Phase

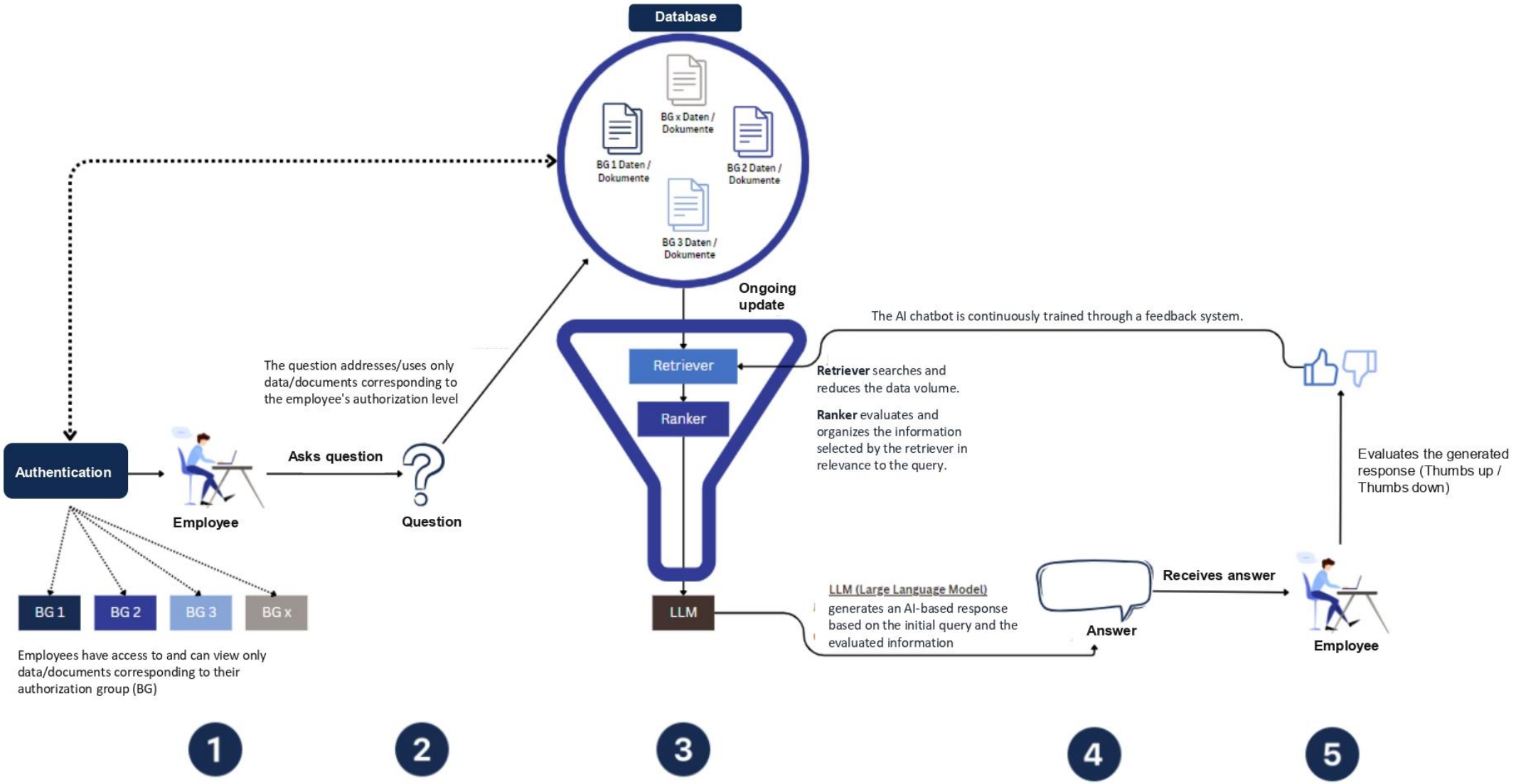
Yaren Maendle

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

Additional Slides: Process Goal



Additional Slides: Profile of Respondents

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%		

Additional Slides: Interview Scenarios

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?

Additional Slides: Survey Questions

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.

Additional Slides: Survey Questions

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.

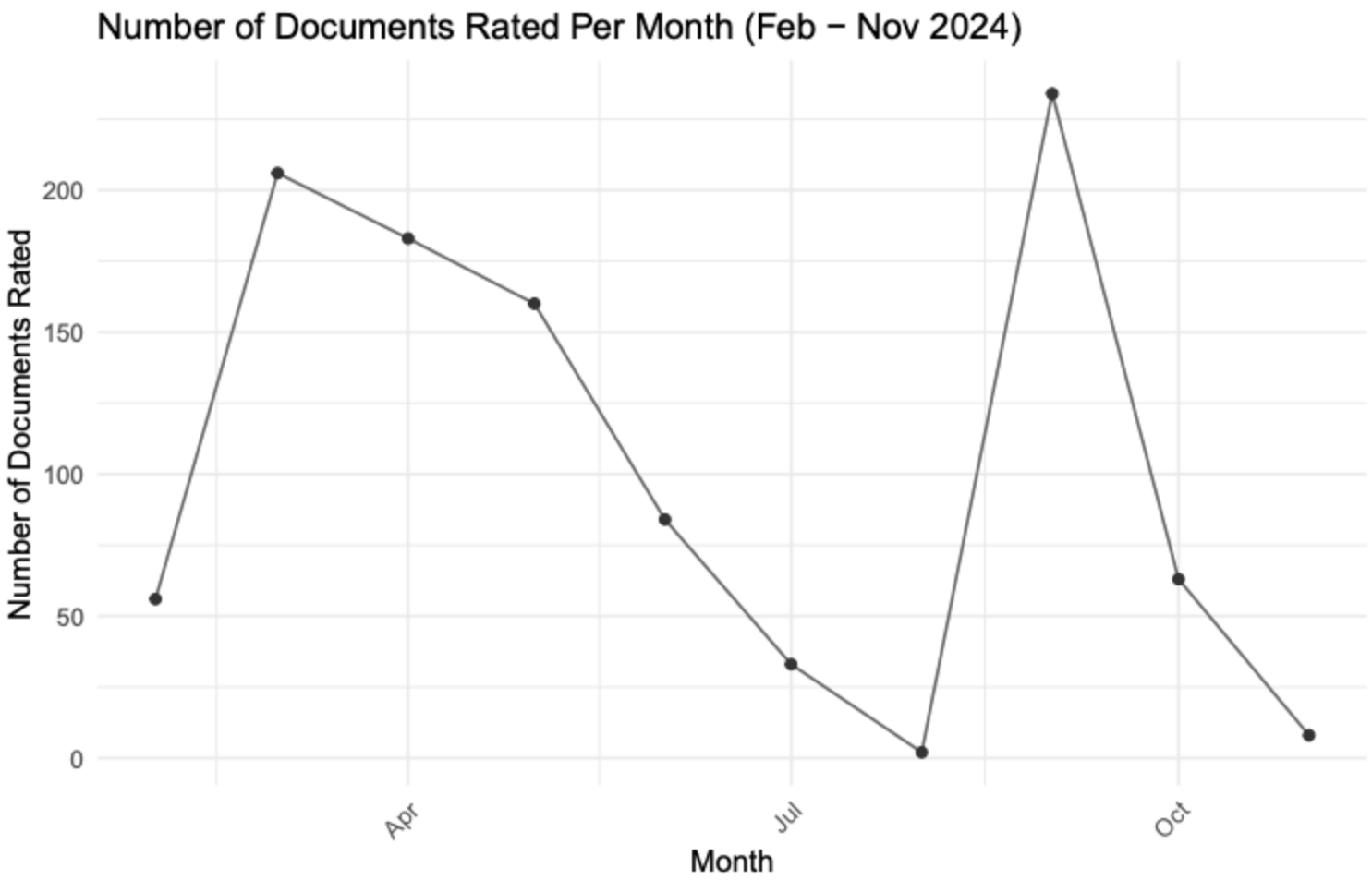
Quality
Definition: User perception of the superiority of a service [82].
QUA1 CS consistently provides accurate, correct, and reliable information.
QUA2 CS's answers are structured to make it easy to understand and follow the information provided.

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.

Additional Slides: Log Analysis with Time



Additional Slides: Log Analysis with Time