ChatWeSIS

Farjana Akter
farjana@uni-bremen.de
University of Bremen
Bremen, Germany

Prokash Karmokar
prokash@uni-bremen.de
University of Bremen
Bremen, Germany

Sharin Ferdous Shampa
sharin1@uni-bremen.de
University of Bremen
Bremen, Germany

Sarah Islam Momo
momos@uni-bremen.de
University of Bremen
Bremen, Germany

Agoua Germaine Stephanie Don
donag@uni-bremen.de
University of Bremen
Bremen, Germany

Yeasir Arafat Ratul
yeasirar@uni-bremen.de
University of Bremen
Bremen, Germany

Md. Mahmudul Hasan
mdmahmud@uni-bremen.de
University of Bremen
Bremen, Germany

Papa Yaw Forson
papayaw@uni-bremen.de
University of Bremen
Bremen, Germany

Hanif Effah Dadzie
hdadzie@uni-bremen.de
University of Bremen
Bremen, Germany

Tianci Chen
tchen@uni-bremen.de
University of Bremen
Bremen, Germany

Tanmoy Indu
tanmoyin@uni-bremen.de
University of Bremen
Bremen, Germany

Awais Khurshid
awais1@uni-bremen.de
University of Bremen
Bremen, Germany

Zameer Khan
zkhan@uni-bremen.de
University of Bremen
Bremen, Germany

Phillip Leder
lederph@uni-bremen.de
University of Bremen
Bremen, Germany

Abstract

Science gateways enhance access to scholarly data but often require advanced skills for effective use, posing challenges in data discovery and usability. To address this, we present ChatWeSIS, a conversational agent integrated into the WeSIS platform, leveraging a Retrieval-Augmented Generation (RAG) pipeline and multiple LLMs to enable intuitive, context-aware interactions. Our implementation combines LangChain for dynamic query processing, vector-based retrieval, and agentic task execution, aiming to streamline dataset discovery and user guidance. Through a user study with researchers of varying expertise, we evaluate ChatWeSIS's effectiveness, usability, and alignment with user workflows. Results indicate strengths in accessibility for novice users, while expert users highlight needs for deeper contextual understanding and response precision. Then technical limitations are discussed, such as resource constraints and integration challenges, and improvements are presented, including memory features, role-based personalization, and automated data updates. This work contributes insights into designing LLM-driven conversational agents for knowledge discovery and evaluation of data reusability for science.

1 Introduction

Science gateways provide access to large data sources as well as computational and visualization tools, but their effective use often requires advanced skills from users to navigate the available resources and extract relevant knowledge [8, 26]. The large volume, diverse formats, inadequate search tools, and complex connections within scholarly knowledge linked to research data often make it challenging for researchers to organize data for effective discovery and reuse [16, 19]. Furthermore, from the user's perspective, researchers have heterogeneous data needs that are difficult to classify and lack capabilities in searching for data due to unfamiliar data search practices in contrast to academic literature search practices. Consequently, domain users seek intuitive, user-friendly tools to support their research workflows, particularly in assessing data reusability and engaging with datasets more effectively. Inspecting data, exploring its content in depth, and situating it within broader scholarly contexts are critical aspects of this process [19]. Conversational agents, including chatbots, have emerged as a promising solution for facilitating such interactions.

Recent work has focused on developing chatbots to assist users in querying and navigating open datasets more efficiently [34]. As a subset of conversational agents, chatbots enable human-computer interaction (HCI) through textual communication, providing a more intuitive and user-friendly approach to data discovery. By integrating these tools into science gateways, researchers can refine search queries, gain contextual insights, and evaluate dataset relevance without requiring extensive technical expertise.

Advancements in artificial intelligence (AI) are significantly transforming social science research [17]. Particularly large language models (LLMs) have further expanded the capabilities of conversational agents, enabling more sophisticated interactions and deeper engagement with data. The integration of LLM-based

chatbots into domain-specific science gateways enhances not only the accessibility of research data but also the discovery of complex relationships within datasets, ultimately improving knowledge extraction and reuse [14, 34].

Additionally, evaluating the real-world impact and user acceptance of LLM-driven chatbot systems is essential for validating their utility [15]. Prior studies have explored chatbots in diverse applications, such as improving geospatial data access in geoportals, enhancing pedagogical support in Data Science education, and facilitating dataset discovery through graph-based retrieval, demonstrating their effectiveness in streamlining information access and user interaction [9, 14, 34]. However, many existing implementations fail to account for the socio-technical considerations of chatbot interaction, particularly the misalignment between user expectations and system capabilities. Users often hold inaccurate assumptions about machine intelligence, system capabilities, and its goals, leading to mismatches in how these systems are perceived and utilized [22, 26]. Understanding how researchers interact with LLM-driven conversational agents in practical research workflows is therefore crucial for designing chatbots that are not only technically effective but also widely adopted in scientific communities.

In this paper, we introduce **ChatWeSIS**, a conversational agent designed to improve user interaction with the **WeSIS** science gateway – "a comprehensive, interactive web-based information system on global dynamics of social policy" [2] – addressing key challenges in data discovery and knowledge retrieval. Unlike conventional search interfaces, ChatWeSIS leverages multiple LLMs within a **Retrieval-Augmented Generation (RAG)** pipeline to enable more intuitive, contextual, and user-friendly engagement with research datasets [20]. Our implementation integrates **LangChain** for query processing [32], vector-based retrieval to enhance query relevance, and agent-based task execution for structured information extraction. By embedding this chatbot into the WeSIS platform, we aim to assist researchers in refining search queries, assessing dataset reusability, and navigating complex information landscapes more efficiently. Furthermore, we conduct a user study with researchers of varying familiarity with WeSIS to evaluate the system's effectiveness in improving knowledge discovery and usability, offering insights into user expectations, system limitations, and potential future developments.

The paper is structured as follows. In Section 2 we review similar approaches in recent research

2 Related Work

Several studies have explored conversational agents for data search, knowledge discovery, and domain-specific assistance. While prior research has demonstrated their potential, challenges remain in handling ambiguous queries, adapting to different expertise levels, and evaluating usability beyond technical accuracy.

2.1 Conversational Agents for Data Discovery

Conversational agents have been proposed to improve data search and retrieval by translating natural language into structured queries. Cantador et al. [8] introduced a chatbot for open government data, outperforming traditional keyword search but struggling with scalability. Fan et al. [14] developed DataChat, which uses a scholarly knowledge graph for dataset search but faced issues with complex query interpretation. Similarly, Oruche et al. [26] integrated sparse retrieval and intent classification in their *Vidura advisor design framework* (VADF), improving precision but showing sensitivity to query variations.

Beyond direct search, Zhang et al. [38] proposed a *Domain-Specific Topic Model* (DSTM) to uncover latent relationships between research topics, datasets, and tools. While promising, it failed to outperform standard retrieval models and lacked real-time interaction capabilities. Reis et al. [30] explored Flowise, a chatbot-enhanced system for biomedical database discovery using a RAG pipeline but identified issues with synonym handling in search. These studies highlight the need for adaptive, context-aware search mechanisms capable of handling structured and unstructured data sources.

2.2 Task-Oriented Chatbots in Domain-Specific Applications

Beyond data retrieval, chatbots have been applied in specialized domains to assist with research workflows. Vahidnia [34] developed a geoportal chatbot using deep transfer learning but noted intent confusion and terminology challenges. Carlander-Reuterfelt et al. [9] introduced JAICOB, an educational chatbot, but its flow-based design limited flexibility. Vekaria et al. [35] proposed OnTimeRecommend, a chatbot-enhanced recommender for science gateways, though its evaluation focused on algorithmic metrics rather than user interaction quality.

A broader challenge is the lack of user-centered evaluation frameworks, particularly for expert domain users. Følstad et al. [15] emphasize that chatbot assessments often prioritize technical accuracy over usability, trust, and adoption—factors critical for professionals. Without evaluating alignment with expert workflows, even functionally correct systems may fail to gain traction. This underscores the need for holistic evaluation metrics that capture engagement, efficiency, and long-term usability in expert settings.

While prior research has demonstrated the value of chatbots in structured data retrieval and domain-specific assistance, several gaps remain unaddressed. First, existing solutions often exhibit limited adaptability to diverse user expertise levels, making it difficult for both novice and expert users to extract relevant information efficiently [22]. Second, many chatbots lack robust retrieval mechanisms that integrate unstructured textual data, relying instead on rigid query structures that constrain flexibility in exploration [3, 8, 14, 26, 34, 37, 39]. Third, studies evaluating chatbot effectiveness tend to focus on technical accuracy rather than user experience, leaving socio-technical considerations—such as user expectations, trust, and adoption—underexplored [15].

3 Methodology

To address these gaps, we introduce **ChatWeSIS**, a conversational agent embedded in the WeSIS science gateway. Our approach builds on prior work by integrating a Retrieval-Augmented Generation (RAG) pipeline [20], combining structured vector-based retrieval with LangChain to enable dynamic, context-aware responses. Additionally, we conduct a user study with researchers of varying expertise levels with WeSIS to assess not only the chatbot's accuracy but also its usability, perceived usefulness, and interaction

quality for social science researchers. Through this, we contribute empirical insights into designing LLM-powered research assistants that bridge the gap between structured retrieval and user-centric interaction.

3.1 Technical Framework

3.1.1 Technical Considerations. Since our goal was a working prototype for a user study and not a fully-fledged finished chatbot, some aspects of our implementation are oriented towards a quick development cycle and need to be addressed going forward.

We used **Ollama** to host our LLMs since we had limited resources and needed multiple different LLMs for different tasks. This results in degraded performance, since Ollama is constantly loading and unloading the LLMs depending on need. We chose this approach to make sure we have full control over what data is being processed where.

3.1.2 ***How does it work? The chatbot consists of several layers:***

(1) The UI on the WeSIS website:
(2) The Python backend running the LangChain:
(3) The Python backend to upload context to the LLMs:
(4) The vector store containing contextual data:
(5) The Ollama service running the LLMs:

Chatbot UI

The integration of chatbot interfaces into WeSIS significantly improves user engagement and accessibility. ChatWeSIS provides an intuitive UI for user interaction with the WeSIS platform. The primary objectives of the chatbot UI are: Ensuring chat history persistence across page reloads. Maintaining chatbot window state across the navigation. Offering a user-friendly and visually appealing interface.

Frontend Architecture. The chatbot UI is developed using React.js and styled with CSS. The component structure is as follows:

(1) **Chatbot Component:** Manages chat window state and user interactions.
(2) **Message Component:** Displays user and bot messages dynamically.
(3) **Input Field:** Captures user messages and sends them to the backend.
(4) **Session Storage Mechanism:** Stores messages and UI state to ensure persistence.

The chat window includes: A header displaying the chatbot name.
A message area that scrolls automatically to the latest message.
An input field where users can type messages.
A send button to submit queries to the chatbot.

Message Handling and User Experience. ChatWeSIS supports an interactive chat experience with the following enhancements:

Welcome Message: Upon opening the chat window, users are greeted with a predefined welcome message.
Message Formatting: The chatbot UI supports clickable links and basic text formatting (bold, italics).
Auto-scrolling: The message area automatically scrolls to the latest message for better readability.
Typing Indicator: A visual indicator is shown while the bot is generating a response, making the conversation feel more natural.
Expanding Text Box: The input field dynamically expands as the user types longer messages, ensuring better visibility and usability.

Chat Window Persistence. To enhance the user experience, the chatbot maintains its state even when the user navigates to a different page. This is achieved by:

Session Storage: The chatbot state (open or closed) is stored using `sessionStorage`, ensuring that it remains in the same state when users navigate between pages.
Conversation History Retention: Messages exchanged between the user and the chatbot are stored temporarily, allowing users to revisit previous interactions.
Seamless Integration with Website Navigation: The chatbot remains accessible in a fixed position, allowing users to interact with it without losing context.

LangChain

LangChain is a powerful framework that is designed to develop Large-Language Model (LLM) powered applications. Its architecture is modular and flexible, which enables developers to build, configure, update, and deploy applications to cater to custom requirements [23]. Such modularity and flexibility make it ideal for building conversational applications.

LangChain allows the development of chains (customs pipelines) that can be configured with various components (modular abstractions) to develop intelligent systems [32]. We utilized components to build chains that can process and execute different tasks. Some of the most vital components used are:

Prompts: This is the input passed to the LLM. Prompts are configured with additional information/instructions together with the user's query and then passed to the LLM in order to generate a better response [32].

Chat Models: A language model where the inputs and outputs are chat messages. In contrast, the output of an LLM is plain text. We utilize **ChatOLLama** to generate chat message outputs.

Chains: Are utilized for sequential calls of tasks [32]. There are multiple chains with different purposes offered by LangChain, we utilized two prominent chains: **create_retrieval_chain** which takes user input, retrieves relevant documents, and then passes the combination to the LLM for a response. Additionally, the **create_stuff_documents_chain** takes and processes documents into a prompt and then passes it to the LLM for a response.

Text Embedding Models: Embedding models help create numerical (vector) representations of texts [32]. This enables the semantic searching of embedded data in vector space. We utilized the **OllamaEmbeddings** class to generate document embeddings.

Vector Stores: Are utilized for embedded data to be stored and searched over [32]. Embedded data is stored as embedded vectors and when a query is passed, it is embedded as well, which then results in a similarity search between the embedded vector and query.

Retrievers: Come in handy for retrieving relevant data or information [23]. LangChain offers many retrievers, one of which we

utilized is **ContextualCompressionRetriever** which works by adding an extra post-processing step after retrieving documents to select relevant information only.

Output Parsers: For the output from the language models to be more suitable, parsers are utilized to transform the output.

Streamlit

For document upload and management, we utilized Streamlit, a Python framework used for building interactive web applications. Users have the ability to upload documents via an interface, which would then after being processed via LangChain components, be utilized as context for the LLM.

Dataflow

System Architecture. ChatWeSIS leverages LangChain's modular framework to effectively process user queries. The process begins when a user passes a natural-language query to the chatbot interface. The LangChain router then determines the appropriate chain for handling the query: either a specialized Agent or a Retrieval-Augmented Generation (RAG) pipeline.

Agents handle specific processing of tasks such as link generation while the RAG pipeline handles more complex and open-ended queries by retrieving relevant information from vector stores containing document embeddings. After data is processed and retrieved from the appropriate chain, it is then forwarded to an LLM for a coherent and contextual response. Finally, via a LangChain orchestrator, the response is forwarded to the chatbot interface where it is displayed to the user.

Agentic Chain. Figure 2 describes the interaction flow within ChatWeSIS for generating links based on natural language user queries. This chain integrates various key components including a Large Language Model (LLM), a database, structured output parsing, and a runnable orchestrator, which work seamlessly together to generate links.

Firstly, the natural-language query is passed to a runnable sequence, which triggers a runnable lambda function. This lambda function formats the query into a structured prompt using **ChatPromptTemplate** and then triggers the LLM (**ChatOllama**) to extract fields (such as country name, indicator name, start year, and end year). The LLM's raw output is then converted into structured JSON data via the **StructuredOutputParser**. These fields are then used for database lookups to retrieve the country code and indicator ID using fuzzy matching techniques. A helper function is then leveraged to construct a URL that includes these identifiers from the database. Within the runnable sequence, the link is formatted into a structured message using a **PromptTemplate**. This message is transformed into a user-friendly, contextual, and conversational response using the LLM. Finally, the runnable sequence delivers this response which is informative and understandable to the chatbot interface.

RAG Chain. Figure 3 outlines the workflow of the Retrieval-Augmented Generation (RAG) pipeline within ChatWeSIS. The RAG approach allows the look-up of external knowledge bases. In our scenario, after receiving a natural-language query, it passes through a **CompressionRetriever** where a similarity search against a FAISS vector store is conducted to identify and retrieve the top 15 relevant external documents. Additionally, a compression pipeline consisting of **EmbeddingsRedundantFilter** and **LongContextReorder** is utilized to refine these documents by reducing redundancy and prioritizing the most relevant information. This refined context along with the original query from the user is then passed to the LLM (**ChatOllama**) to generate a user-friendly, contextual, and conversational response. This final response is then forwarded to the user.

3.2 User Research

This study investigates user needs and evaluates the effectiveness of the ChatWeSIS prototype in improving user navigation, knowledge discovery, and resource accessibility on the WeSIS platform. The primary objective of this research was to determine whether the introduction of a chatbot could enhance user engagement by streamlining interactions with the platform's features and data resources. To ensure the chatbot prototype met the specific requirements of the WeSIS platform's users, a structured research methodology was employed.

A preliminary platform review was conducted as the first step in the study to evaluate the existing functionalities of the WeSIS platform. This review provided valuable insights into the platform's structure, core features, and potential usability challenges. The findings helped identify areas where users struggled, which subsequently informed the development of the user research methodology. These insights were crucial in creating the requirement analysis survey, which captured user experiences and expectations in a systematic way. This approach ensured that the chatbot's functionalities were aligned with user needs and contributed to enhancing the platform's usability.

Requirement Analysis

The user research phase began with a requirement analysis, which involved a platform review followed by a survey. The purpose of this process was to identify the main challenges users faced when navigating the WeSIS platform and to understand their expectations for the chatbot. Previous studies have emphasized the importance of conducting a systematic requirement analysis to align digital tools with user needs [31, 34]. The pre-survey collected insights into user behaviors, difficulties, and preferences, which were used to refine the chatbot's design and functionality [14]. To ensure diverse participation, both online and offline survey methods were utilized. Online surveys are widely recognized as effective tools for collecting user feedback due to their accessibility and anonymity [9]. The online survey was conducted using LimeSurvey [21], allowing for anonymous responses in compliance with GDPR regulations [13], ensuring participant privacy and encouraging honest feedback. The importance of GDPR compliance in user research has been highlighted in previous studies to ensure ethical data handling and privacy protection [1]. The offline survey also allowed participants to indicate their willingness to participate in follow-up user studies, helping to recruit users for further testing of the chatbot prototype [5].

Survey results highlighted several key findings. Users exhibited varying levels of familiarity with the WeSIS platform, ranging from

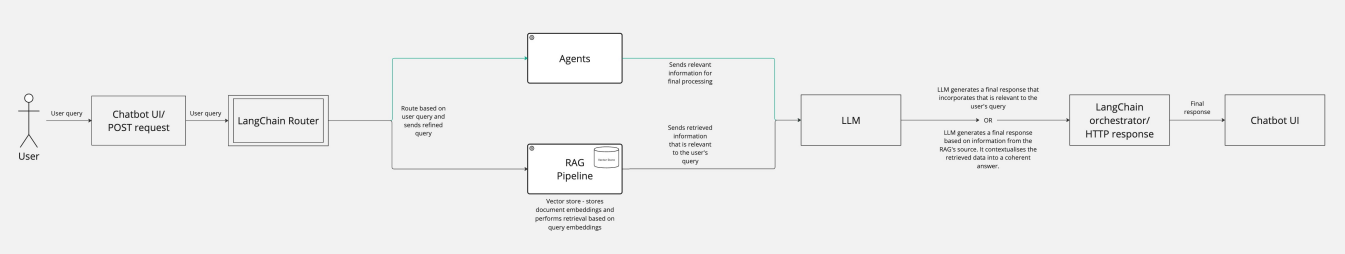

Figure 1: System Architecture

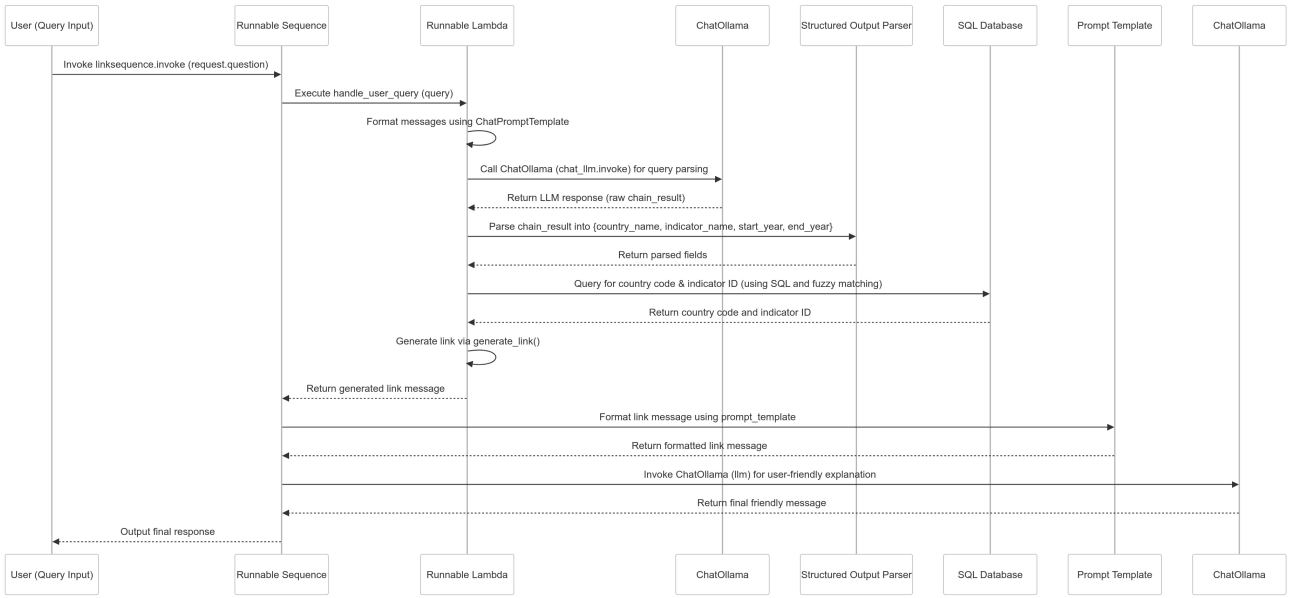

Figure 2: Agentic Chain

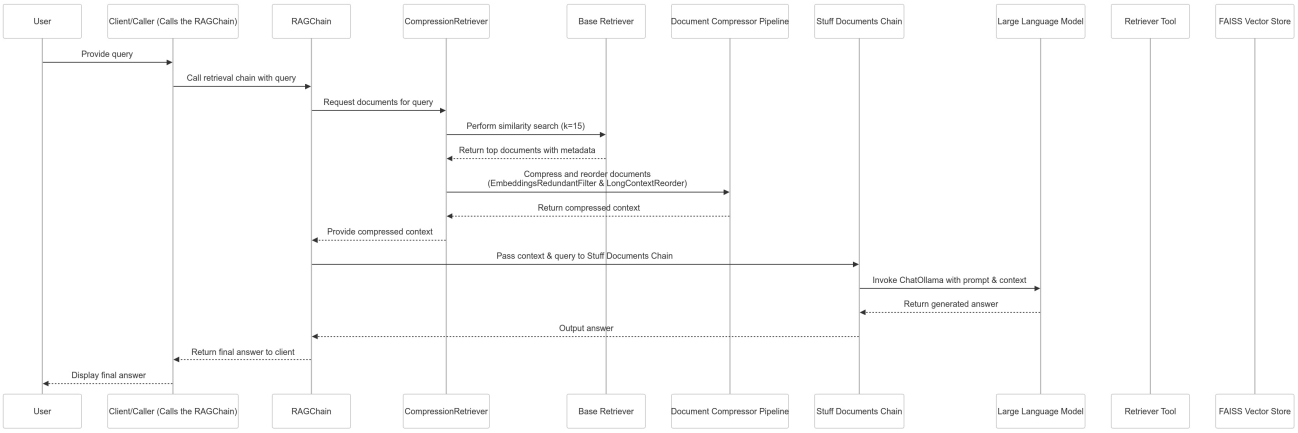

Figure 3: RAG Chain

beginners to experts [8]. Some accessed the platform daily, while others used it infrequently or had never used it. This variation indicated that the chatbot should support both novice and experienced users, offering basic guidance for beginners and efficient navigation for more advanced users [28]. Research on user experience has demonstrated that personalization in digital platforms significantly improves user engagement [24].

Participants reported performing tasks such as uploading and downloading data, navigating datasets, and searching for specific indicators. However, many encountered challenges in these tasks. Common issues included difficulty finding specific datasets, slow-loading network data, unclear metadata, and the inability to plot certain indicators [34]. Usability studies have shown that unclear metadata and data retrieval inefficiencies often result in user frustration [14]. Users also struggled with long and complex error logs, making it harder to troubleshoot problems, which aligns with research findings that emphasize the need for clear and structured error reporting [31]. Additionally, some found it difficult to get an overview of available indicators or locate related indicators unless browsing by topic [32].

One of the most frequent concerns was the lack of guidance and tutorials on how to use the platform. Some users, particularly those without a background in social sciences, found the data difficult to understand [1]. This suggested that the chatbot should not only assist with navigation but also provide explanations and educational support to help users interpret the available data. Research in human-computer interaction (HCI) suggests that digital assistants that incorporate educational support enhance user comprehension and engagement [8]. Participants ranked "Guiding users through datasets and indicators" as the most useful chatbot feature, followed by "assisting with navigation" and "providing explanations for indicators" [24]. Other helpful features included offering tutorials and helping users find external resources such as publications and WeSISpedia entries [9]. These findings indicated that the chatbot should focus on improving dataset discovery, navigation, and explanatory support.

There was no single preferred response style for the chatbot. Some users wanted short, direct answers, while others preferred detailed explanations [31]. Many suggested that the chatbot should offer both options, allowing users to request additional details when necessary. This highlighted the need for a flexible response system that could adapt to different user needs [5]. Research in conversational AI suggests that adaptable response strategies lead to more effective chatbot interactions [28].

While many participants believed that a chatbot could improve their experience with WeSIS, some expressed doubts about whether it could address deeper platform issues, such as inconsistencies in data uploads and a lack of coordination among data contributors [1]. This suggested that while the chatbot could enhance user engagement and accessibility, broader improvements to the platform itself might also be necessary. Users also suggested that the chatbot should be able to recommend relevant indicators based on specific research questions [14]. Additionally, they emphasized the importance of providing structured, reliable answers while avoiding vague or overly technical responses [34]. Previous studies have found that recommendation systems improve the discoverability of relevant information in digital knowledge repositories [9].

These insights were used to fine-tune the chatbot prototype and helped to prioritize improvements in navigation support, data discovery, user guidance, and response clarity. By addressing these areas, the chatbot could be made more intuitive and useful for a diverse range of users [32].

4 User Study Methodology

The methodology employed in this study was designed to systematically assess the effectiveness of the ChatWeSIS prototype in enhancing user engagement and resource accessibility on the WeSIS platform. The study involved a combination of qualitative and quantitative data collection methods, including task-based interactions, a usability questionnaire, and open-ended interviews. This mixed-methods approach aligns with established research in chatbot usability and human-computer interaction (HCI) [18, 36].

4.1 User Study Design

Following the requirement analysis, a user study was conducted to evaluate the usability and effectiveness of ChatWeSIS in supporting navigation, resource discovery, and user engagement on the WeSIS platform. The main goal of this study was to determine whether the chatbot effectively helped users access and utilize the platform's resources. Ten participants were recruited for the study. Participants were selected based on their responses from the requirement analysis survey, and targeted outreach was conducted to ensure a diverse and representative sample. Participation was voluntary, and all data collection followed GDPR guidelines [13], ensuring participant anonymity.

4.1.1 Participants. As detailed in Table 1, the study involved ten participants (n=10) with diverse academic backgrounds, familiarity with the WeSIS platform, and varying levels of experience with chatbots. The participants' educational qualifications ranged from high school to PhD degrees, with fields of study including political science, sociology, computer science, and geography.

Out of the ten participants, seven had prior experience with the WeSIS platform, with familiarity levels ranging from beginner to expert. Three participants had no previous experience with WeSIS, offering a valuable perspective from those less familiar with the platform.

Some participants had used chatbots for research before our study, whereas others had not. Similarly, the frequency of general chatbot usage varied: three participants reported using a chatbot rarely, three participants used it sometimes, one participant used it often, and three participants had never used a chatbot.

This mix of participants with different levels of experience and backgrounds provided a well-rounded view of how familiarity with both the WeSIS platform and chatbots influenced their interactions with the ChatWeSIS prototype.

4.1.2 Study Team Roles. Each participant was assisted by two individuals: an interviewer and a moderator. The interviewer guided the participant through the study and asked questions related to the tasks and the chatbot's functionality. The moderator, on the other hand, focused on ensuring the smooth technical operation of the study and assisted with the chatbot's interaction. This role

Table 1: Demographics and background data about the participants

Participant	Highest Degree	Field of Degree	WeSIS Familiarity	Used Chatbot for Research	Frequency of Chatbot Usage
1	Master	Political Science	Advanced	No	Rarely
2	Master	Computer Science	Expert	Yes	Rarely
3	PhD	Political Science	Beginner	No	Never
4	PhD	Sociology	Beginner	No	Never
5	PhD	Sociology	Expert	Yes	Often
6	PhD	Political Science	Expert	Yes	Sometimes
7	Master	Political Science	Intermediate	No	Sometimes
8	High School	Geography	Not applicable	No	Sometimes
9	PhD	Political Science	Advanced	No	Rarely
10	PhD	Sociology	Beginner	No	Never

division helped maintain the participant's focus and minimized any distractions during the study.

4.1.3 Study Phases. Table 2 depicts the three phases of the user study, each aimed at evaluating different aspects of the chatbot's functionality.

Phase 1: Task-Based Interaction. In the first phase, participants were given three structured tasks based on specific scenarios. These tasks required participants to interact with ChatWeSIS to complete specific research-related activities on the WeSIS platform. The goal was to assess how well the chatbot assisted participants in navigating the platform. Due to the open-ended nature of the task questions, the time required for Phase 1 could not be controlled. To capture participant engagement, all interactions were recorded using Zoom with consent, and participants were asked to think aloud during this phase [12].

Phase 2: Chatbot Usability Questionnaire (CUQ). The second phase involved participants completing the CUQ, a standardized tool for evaluating chatbot usability [18].

The CUQ results in a score that is calculated by assigning a value between 1 and 5 to each response, summing positive and negative question scores separately, adjusting them, and scaling the final result to a percentage out of 100:

$$\text{CUQ Score} = \left(\frac{(\sum Q_{\text{odd}} - 8) + (40 - \sum Q_{\text{even}})}{64} \right) \times 100$$

We modified the CUQ by removing half of the paired questions to avoid *survey-taking fatigue* [29, 36] and adjusted the score calculation by subtracting 4 from the sum of odd-numbered questions and subtracting the sum of even-numbered questions from 20 before computing the final score:

$$\text{CUQ Score}_{\text{adjusted}} = \left(\frac{(\sum Q_{\text{odd}} - 4) + (20 - \sum Q_{\text{even}})}{32} \right) \times 100$$

Phase 3: Open-Ended Interviews. The final phase consisted of qualitative interviews, capturing user experiences, challenges, and suggestions for improvement. These insights complemented the quantitative CUQ findings, providing a holistic evaluation.

4.1.4 Data Collection and Analysis. All recorded interactions were transcribed for systematic documentation. The transcripts, along with CUQ responses, were used for further analysis. Additional details can be found in Section 7.

5 Results

Quantitatively, CUQ scores (see Figure 4 and Table 3) showed modest differences between expert and novice users: experts (Participants 1, 2, 5, 6, 9) averaged *73.1* (range: 56.2–84.4), while novices (Participants 3, 4, 7, 10) averaged *76.6*. This suggests comparable perceived usability. However, qualitative analysis revealed stark contrasts: novices praised the chatbot's accessibility and guidance, whereas experts critiqued response precision, contextual limitations, and inadequate support for advanced tasks like dataset exploration or multi-query workflows.

5.1 Comprehensive Analysis Overview for Phase 1 Transcription

For Phase 1, we have performed an in-depth analysis of the transcripts using three different approaches: **Sentiment Analysis**, **Discourse Analysis**, and **Thematic Analysis**. Along with the analysis, we also added and compared the results with the CUQ score. The CUQ data is summarized in Table 3 and visualized in Figure 4.

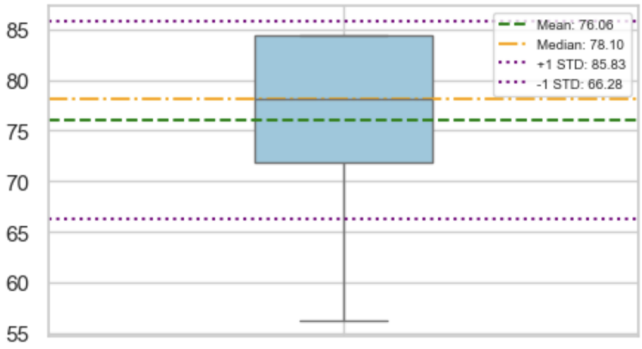

Figure 4: CUQ Score Distribution

Thematic analysis (TA) is a qualitative method for identifying, analyzing, and interpreting patterns of meaning (themes) within textual data [6]. Unlike methodologies bound to specific theoretical frameworks, TA offers flexibility in application across research paradigms, enabling both inductive (data-driven) and deductive (theory-driven) approaches [11]. Themes are derived through iterative coding processes, where codes capture salient features of

Table 2: User Study Phases

Phase	Description
Task-Based Evaluation	Participants completed three tasks designed to assess the chatbot's ability to assist with platform navigation, data retrieval, and user guidance. The tasks included: • **General Information Retrieval:** Locate and open the ChatWeSIS chatbot, ask about WeSIS, its key features, and available tutorials. • **Exploring Social Policy Indicators:** Request and compare data on the indicator "Actors responsible for financing at introduction" for Denmark and Ireland (1921–2000), and check related publications. • **Creating Indicator Pages:** Ask for instructions on creating a new indicator page and uploading files, verifying if the chatbot provides relevant resources.
CUQ Questionnaire	Participants rated chatbot usability using the Chatbot Usability Questionnaire (CUQ), a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). The questionnaire covered aspects such as ease of navigation, engagement, clarity of guidance, and user experience [18].
Interview	A semi-structured interview captured qualitative feedback on chatbot performance, usability challenges, and suggested improvements. Participants were asked questions such as: • What did you like most about using the chatbot? • What challenges or issues did you face? • How could the chatbot be improved? • How easy or difficult was it to find information with the chatbot's help? • Do you have any other comments?

Table 3: CUQ Scores

	Question								
Participant	1	2	3	4	5	6	7	8	CUQ $\text{Score}_{\text{adjusted}}$
1	4	1	4	3	3	2	3	1	71.9
2	4	2	3	3	2	3	4	3	56.2
3	4	2	4	2	4	3	4	2	71.9
4	5	1	5	4	4	1	4	1	84.4
5	5	1	5	2	1	5	4	1	68.8
6	5	1	5	2	4	2	5	3	84.4
7	5	2	5	1	4	2	4	2	84.4
8	4	2	5	3	5	1	3	2	78.1
9	4	1	4	1	4	1	4	2	84.4
10	4	2	3	2	3	2	3	2	65.6

the data, and themes consolidate these codes into broader patterns anchored by a central organizing concept. TA emphasizes rigor through systematic procedures, including theme review against coded data and datasets, ensuring analytical coherence [7].

In our study, TA enables systematic categorization of user feedback (e.g., challenges in navigation, expectations for chatbot functionality) into themes such as engagement, usability barriers, and response clarity. This approach aligns with our goal to uncover recurring patterns in user experiences, particularly across varying expertise levels (novice vs. expert users). By prioritizing themes like user confusion and information quality, TA provides actionable insights into design improvements and user needs.

Discourse analysis (DA) examines language use in social contexts, focusing on how meaning is constructed through spoken or written interactions [27]. DA encompasses diverse approaches, including conversation analysis (turn-taking, adjacency pairs), critical discourse analysis (power dynamics, ideology), and genre analysis (textual structures) [27]. It interrogates linguistic patterns, pragmatic intentions, and sociocultural influences, emphasizing the relationship between language and context [10].

For ChatWeSIS, DA facilitates exploration of how users interact with the chatbot—examining conversational norms (e.g., query phrasing, follow-up requests), pragmatic misalignments (e.g., mismatched user expectations), and sociotechnical dynamics (e.g., trust in AI responses). By analyzing discourse features such as politeness strategies, conversational implicatures, and narrative coherence, DA reveals interactional challenges (e.g., frustration due to repetitive queries) and systemic limitations (e.g., lack of contextual memory). This complements TA by situating thematic findings within the broader context of human-AI communication dynamics.

Each of these analyses is integrated by different models. The summary of the models and classification is below:

Sentiment Analysis: We utilised a couple of models for the sentiment, thematic, and discourse analysis. For the sentiment classification, the nlptown/bert-base-multilingual-uncased-sentiment model was used [33]. This is a multilingual BERT-based model that was fine-tuned for sentiment analysis on product reviews in six languages and classifies sentiment on a 5-point scale, ranging from 1 (Very Negative) to 5 (Very Positive). This model is intended for direct use in sentiment analysis or further fine-tuning.

Thematic Analysis: The facebook/bart-large-mnli [4] model was employed for the thematic classification using a zero-shot classification approach. This method, proposed by Yin et al., leverages pre-trained NLI models to classify sequences without specific training examples for each theme. The sequence to be classified is posed

as an NLI premise, and a hypothesis is constructed from each candidate label. The probability of entailment is then used as the probability of the label being true. The themes used for classification were: Engagement and Satisfaction, User Confusion and Overload, Overall User Experience, Quality of Information, Clarity in Communication, and Ease of Use and Navigation. Only the top listed theme holding the top matching value is counted. This approach is effective, particularly with larger pre-trained models like BART.

Discourse Analysis: The discourse analysis was done using the facebook/bart-large-mnli model via the pipeline (zero-shot classification) [4]. Similar to thematic classification, this zero-shot method allowed for the classification of the tone or intent of the message based on predefined labels. The labels used for discourse classification were: Confusion, Suggestion, Frustration, Appreciation, and Criticism. The zero-shot-classification pipeline in Hugging Face Transformers enables the classification of sequences into any specified class names.

We have divided our analysis into four parts with a combination of Sentiment and Themetic Analysis.

5.1.1 ***High expectations and critical feedback among expert users.*** Expert participants of the interview, especially Participants 2 and 5, expressed their expectations about performance and precision. Participant 2 scored the lowest in CUQ score (56.2) (See Table 3) despite being familiar with the WeSIS. They showed the most activity in the Ease of Use and Navigation (11 mentions) and Engagement and Satisfaction (10 mentions) themes (see Figure 5). . Besides, they expressed some negative sentiments regarding User Experience and Information Quality. Their expectation was not fulfilled because of mismatches between their task-specific goals and the answer of the chatbot. Participant 5 (Expert, PhD), a frequent chatbot user, had a moderate CUQ score of 68.8 . They reported the highest about User Confusion and Overload, and Overall User Experience for 16 and 26 times, respectively and talked about Ease of Use and Clarity in Communication very little. This indicates dissatisfaction and a lack of conceptual clarity. This highlights the expectation of deep conceptual understanding and tailored data presentation by expert users.

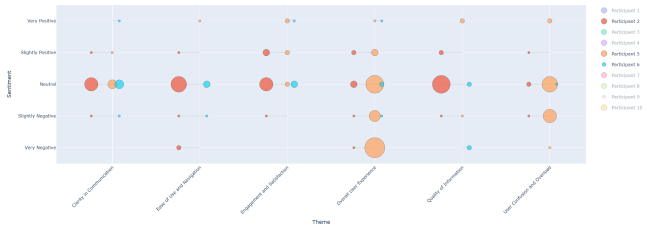

Figure 5: Bubble chart showing expert participant(2,5,6) sentiments across themes.

On the other hand, Participant 6 showed a high CUQ score of 84.4 but didn't express much in thematic feedback. Their feedback shows little confusion and low overall interaction. Their profile shows little confusion and low overall interaction, suggesting passive satisfaction. The lack of expressive sentiment or detailed feedback implies that this user either did not explore deeply or found the platform intuitive.

5.1.2 ***Beginners and Intermediates: Usability Strengths but gaps in depth.*** This group of participants responded mostly positively to the interface but exhibited signs of cognitive overload and frustration in comprehension. Participant 3, a PhD holder but with beginner-level familiarity, for example, with CUQ score of 71.9 (See Table 3), gave notable feedback in Engagement and Satisfaction (12) and Overall User Experience (11)(see Figure 11), which indicates a curiosity-driven experience tempered by barriers to understanding.

Figure 6: Bubble chart showing participant(3,4,7,10) sentiments across themes.

With a CUQ score of 84.4, Participant 4 (PhD, Beginner) showed high counts in two themes. First, Clarity in Communication (16), where they sounded neutral 9 times, gave very positive and slightly positive feedback 4 and 1 time, respectively. Second, they expressed positive sentiments 10 times out of 17 in total mentions under the Engagement and Satisfaction theme. But still they mentioned user confusion twice, suggesting that even engaged users felt ambiguity without guidance.

Participant 7 (Master's, Intermediate) represents an intermediate user. They reported positive and neutral sentiments spread over the themes. The CUQ score of 84.4 is good and showed high involvement in Clarity in Communication (15) and Quality of Information (12)(see Figure 11).

Beginner but highly active Participant 10 scored 65.6 in the CUQ scale. This interviewee talked most about Engagement and Satisfaction and Overall User Experience 21 and 16 times respectively. Even though the participant talked about User Confusion 16 times, in most of the cases, they sounded neutral about it. These patterns show that new users are motivated to explore. This user would benefit from a chatbot offering stepwise exposure, micro-interactions, and layered data responses.

5.1.3 ***Balanced Viewpoints and Iterative Use Cases.*** In this group, we have two users with moderate CUQ scores (71.9) and balanced sentiment feedback, both positive and negative. This group includes Participant 1 and Participant 3.

Participant 1 is an advanced WeSIS user, but the person does not often use a chatbot. This participant provided almost equal feedback where 15 were Slightly Positive, 9 Very Positive, 12 Slightly Negative, and 11 Very Negative(see Figure 12). Moreover, this person has indicated improvement over two themes, which are User Confusion and Overload and Quality of Information, by giving a very negative opinion. This participant also suggests minor improvement

in the section of Overall User Experience, which could enhance the chatbot performance. On the other hand, Participant 3 has 16

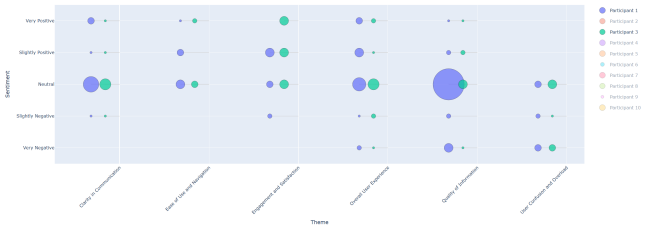

Figure 7: Bubble chart showing sentiments across themes for participant(1,3).

positive and 12 negative sentiments in total. Instead of having a moderate CUQ score, this person also wants improvement in User Confusion and Overload and Quality of Information. These two participants have also detected that the interaction with the chatbot is very easy by giving positive sentiments under the Engagement and Satisfaction theme.

5.1.4 ***Light Users and Missed Potential.*** In this category, Participant 9 was a light chatbot user but had deeper knowledge of WeSIS. This person shares the least amount of feedback among all the themes. This person has an equal number of positive (3) and negative (3) responses in total. Despite having a high CUQ score (84.4), this person has a lack of interaction and engagement with the chatbot. This person has only 8 sentiments, where 2 of them are very negative across the Clarity in Communication and Ease of Use and Navigation theme. This engagement might suggest that this person wants improvement in navigation and clarity of information because, as a rare user, Person 9 does not want to spend more time with the chatbot.

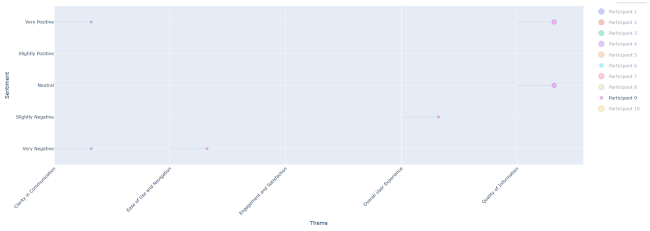

Figure 8: Bubble chart showing sentiments across themes for Light User (participant 9).

Overall Discourse Analysis of Phase 1 Transcription. In this section, we divided discourse analysis into 4 categories for all participants. Below, we describe the analysis of each category according to the analysis.

Suggestion: The participants offered some ideas for improving the system's functionality and user experience, which can be seen in the bar chart under the suggestion section (see Figure 14). Some participants discussed the desirability of the system, remembering previous user inputs to avoid repetition. Others proposed the addition of new features, such as the ability to save settings for

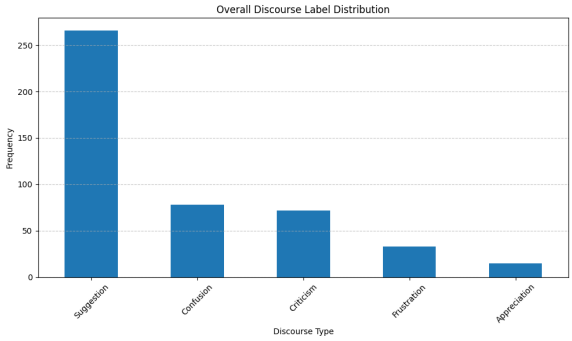

Figure 9: Bar chart showing Discourse analysis for all participants.

future use. Questions were raised about the clarity of the user interface, which implies suggestions for clearer labeling and button placement. As seen from the graph, most of the feedback was suggestions for improvement, hence, "suggestion" having the highest frequency.

Confusion: Across several interactions, participants expressed a lack of clarity regarding the system's behavior and their next steps. Some participants indicated uncertainty about what would happen after providing information or clicking a button. A participant expressed surprise at being asked for information they believed they had already provided.

Criticism: Confusion and Criticism have the same count number on the graph (see Figure 14). Participants provided negative feedback on various aspects of the system. Criticisms were made about the time it took for the system to respond and how unhelpful some of the answers provided were. The lack of a feature to go back and correct mistakes in previous steps was also highlighted as a negative aspect.

Frustration: Feelings of frustration were evident in the interactions. The experience of repeatedly receiving unhelpful answers led to expressions of frustration. The inability to navigate back to previous steps to make corrections, as mentioned under the criticism, was explicitly stated as frustrating by one of the participants.

Appreciation: There were few instances of participants expressing appreciation for the system or the interaction. This is shown in the low frequency of the "Appreciation" label on the graph. The low number of direct expressions of appreciation suggests that positive sentiments were not a dominant feature. The focus of the participants seemed to be more on identifying areas for improvement or expressing difficulties they encountered.

The overall discourse label distribution indicates that "Suggestions"(over 250 counts) were the most frequent type of feedback. On the other hand, "Appreciation" was the least frequent discourse type, which reflects the scarcity of explicit positive feedback observed in the participant interactions. While "Confusion" and "Criticism" also occurred notably, the primary trend suggests a user focus on proposing enhancements and, to a lesser extent, expressing difficulties rather than offering direct appreciation.

5.2 Analysis of Open-Ended Questions (Phase 3):

To evaluate the effectiveness of the Chat WeSIS, a series of 5 open-ended questions (see Table **??**) were asked to the participants to note their perceptions, hurdles, and suggestions. The responses are then thematically coded alongside sentiment levels, CUQ scores, and participant background to give a comprehensive overview. This section presents insights of individual and cross-participant responses across 5 questions.

Q1: What did you like most about using the chatbot? As the first question of open-ended questions participants were asked to note down the aspects that they preferred most when using the chatbot. The answers revealed that the participant voiced positive sentiments in general. They talked about ease of use and navigation, quality of information, and engagement and satisfaction. The sentiment analysis confirmed a dominant trend of slightly positive feedback. Participants 1, 2, 6, 8, and 9 specifically highlighted the chatbot's existence, encapsulation of data and beginner friendliness as key strengths.

For instance, Participant 1, an advanced WeSIS user, noted, " I liked most was its usability to condense a lot of information." Participant 6, an expert user with a high CUQ score of 84.4 (See Table 3), said, "I think in general, its existence is great. I think for a start, for somebody who starts, it's very helpful," reflecting a strong appreciation for the chatbot's necessity.

Even users who do not quite familiar with WeSIS, for example, participants 3,4, and 10 conveyed either neutral or positive sentiments in general(see Figure **??**). They talked about 'ease of use', 'navigation and satisfaction', 'quality of information ', and 'overall user experience'. The results imply that irrespective of participants' background or experience the chatbot became successful at giving a friendly and practical experience, particularly for surface-level information.

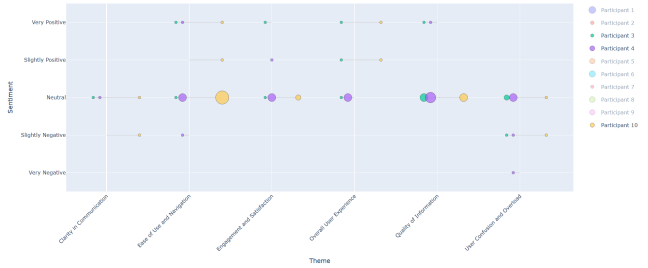

Figure 10: Bubble chart showing participant(3,4,10) sentiments across themes.

Q2: What challenges or issues did you face when using the chatbot? When the participants were asked to speak about their frustration and dislike regarding the chatbot, they explained a range of answers with mixed sentiments from Slightly Negative to Neutral, with a few positive mentions(see Figure 11). The most common challenges reported were related to user confusion and information overload and clarity of information.

Participants frequently pointed out issues with interpreting data or receiving overly generic replies to more specific or complex queries. For example, Participant 7, an expert user, remarked, "I was asking about the specific structure. As someone who was also uploading data, I found that it referred to the structure of how the data was uploaded and how the projects were constructed, rather than addressing the questions I actually asked." Also, Participant 1 said, "I felt that sometimes it didn't fully understand what I was trying to convey. Our interaction showed this, as I struggled to retrieve specific data, and it couldn't provide some links I requested."

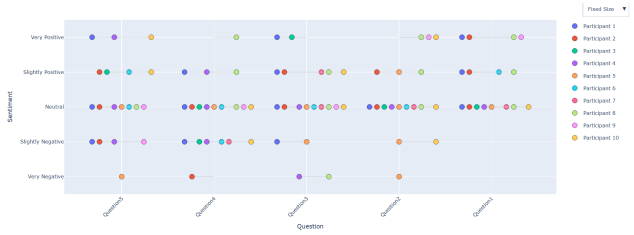

Figure 11: Bubble chart showing participants sentiments across themes.

Showing a key limitation in the chatbot's ability to process subtle, area-specific inquiries. Even though the interface of the chatbot was user-friendly, these comments unfold a noticeable gap between common usability and the capacity for task-oriented work, which is especially vital for experienced researchers and data-driven users looking for more than surface-level interactions.

Q3: How could the chatbot be improved? Participants gave a range of valuable suggestions, from appreciation for the fundamental features to areas requiring improvement. The sentiments range primarily from neutral to slightly negative (see Figure 11), which reflects a tone of realistic feedback rather than dissatisfaction. Thematically, responses were dominated by concerns related to user confusion and overload, overall user experience, and Clarity in communication(see Figure 12). Participants advocated for improvements in the chatbot's contextual accuracy and interactive capability. For example, Participant 1 (Advanced) suggested that

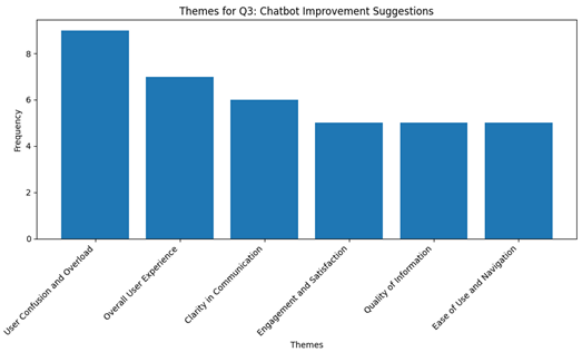

Figure 12: Bar chart for question 3 across the themes.

the chatbot should be able to provide more specific answers. They said, "I found what I was looking for in terms of general indicators, like being able to identify the name of the indicator. However, I had trouble finding specific data points. This might not be the primary role of a chatbot." Reflecting the same tone, Participant 3 (Beginner) said "Sometimes it's just too much information, so just more specific or to say okay here is the information." Similarly, Participants 5 (Expert), 7 (Intermediate), and 8 (not mentioned) expected the chatbot should have memory and remember the previous conversations. The discussed insights emphasize the different expectations across participants. The chatbot should transform from a functional assistant into an exploratory context-aware tool and should be capable of adapting to diverse user needs to facilitate engagement with the WeSIS platform.

Q4: How easy or difficult was it to find information on the platform with the chatbot's help?? The sentiments of the participants in response to this question were mostly slightly positive to very positive. That means the chatbot was able to make it easier to access the information on the WeSIS platform. Besides, the themes that were most discussed are Ease of Use and Navigation, and Overall User Experience.

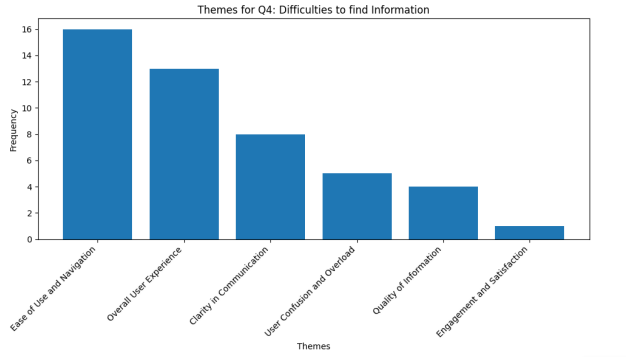

Figure 13: Bar chart for question 4 across the themes.

Participants 6 and 8, with CUQ socre of 84.4 and 78.1 (See Table 3), noted that the chatbot was very easy to find information. Participant 6 said ' It was it was rather easy to get the general information. I would say not perfectly easy, but rather easy.' Paricipant 8 agreed with this but faced difficulties with detailed analytical questions, they voiced 'I'd say. General information was very easy. And detailed analytic information rather difficult.' On the other Hand, Participant 2 focused on interface design, pointing out that long explanations are difficult to read in a small window. They said: "For the answers, that are a bit longer. I have to scroll too much. Participant 1 also echos for a more readable, navigable, user-friendly interface. Other participants mentioned about language barrier, ambiguity, and lack of clarity.

Q5: Do you have any other comments? As the final question of the open-ended questionnaire, the interviewers asked the participants if they had any other comments. In response to that, 4 participants informed the interviewers that they have nothing to say, or they have already said what they wanted to say, but others offered thoughtful suggestions. Overall, the trend of sentiments ranges from neutral to slightly positive, with a constructive tone of suggesting more refinement.

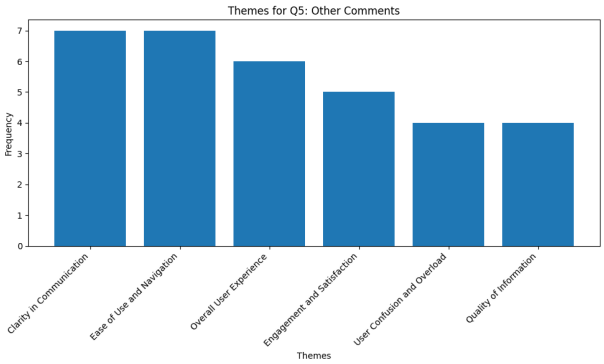

Figure 14: Bar chart for question 5 across the themes.

The answers to this question brought on comments related to Clarity in Communication, Ease of Use and Navigation, and Overall User Experience themes. Participants 1 and 9 suggested more clear and transparent guidance. Participant 2 advocated involving the social scientist to improve the quality of the answers, and Participant 4 advised gathering more user experiences. Participant 6 also emphasized the importance of making the chatbot accessible to non-research audiences by simplifying terminology and explanations. The feedback reveals that while the prototype performs well in usability and accessibility, users are looking for more depth, linguistic flexibility, and instructional clarity.

<

6 Discussion

Our study of ChatWeSIS showed that quantitative scores (like CUQ) and qualitative feedback tell different stories. While CUQ scores were similar for experts (73.1) and novices (76.6), user interviews revealed very different needs. Experts struggled with the chatbot's inability to answer complex questions, like comparing datasets across countries, while novices found it helpful for basic tasks like finding tutorials. This gap highlights a key problem: standardized scores alone cannot fully measure chatbot usability, especially when users need systems to adapt to specific tasks [18]. The small number of participants (n=10) also limits our findings. While past research argues 5–8 users are enough to find most usability issues in traditional software [25], chatbots—with their conversational complexity—likely need larger studies to uncover all challenges. Holmes et al. [18] suggest that chatbots' conversational complexity may demand a larger sample size of around 20 to 30 participants for stable interpretation of the CUQ score. Our study's participant pool, though aligned with qualitative depth goals, underscores the need for caution in generalizing quantitative metrics. Future evaluations should prioritize scaling participant diversity to better capture chatbot-specific usability challenges.

6.1 Technical Considerations

Our study was conducted using Meta Llama 3.1 with 8 billion parameters due to resource constraints. While more advanced LLMs could provide better accuracy and improved responses, our prompt engineering is optimized for this specific model, meaning a direct replacement could degrade performance. Future improvements should consider upgrading to a larger LLM while ensuring response quality remains stable. Implementing a memory feature would enhance the chatbot's ability to retain context across interactions. LangChain offers a suitable plugin, LangGraph, which integrates well with our current implementation, though alternative solutions exist. Ensuring permanent availability of LLM resources with sufficient computational power is also essential for scalability. This would require either high-performance GPUs, which currently exceed the WeSIS project's resources, or a third-party provider capable of hosting the necessary infrastructure.

The chatbot should be fully integrated into the WeSIS platform rather than relying on multiple standalone applications. However, given that WeSIS is built on Ruby on Rails while our backend services are in Python, seamless integration may not be straightforward. At a minimum, key functionalities such as the backoffice for uploading contextual data should be embedded into WeSIS. Consolidating the separate backends into a single API could also improve efficiency by managing both LangChain processes and the vector store, reducing the number of independent services required.

Currently, the chatbot's contextual data is manually scraped from static pages and WeSISpedia before being uploaded to the vector store. Automating this process is crucial to ensure up-to-date information while still allowing for additional contextual data uploads. This is important for including information that is not explicitly available in WeSIS or WeSISpedia but is beneficial for chatbot responses, such as curated knowledge about data sources and processes commonly known to researchers but not documented elsewhere. Alternatively, documenting this information directly in WeSIS or WeSISpedia could be another viable approach.

At present, the chatbot does not differentiate between users, providing the same responses to all. However, some features, such as data uploads, are restricted to specific users, and other use cases may require tailored responses based on user roles. Recognizing user types (e.g., guest, regular user, admin) would allow for more relevant responses. Additionally, introducing a tiered response pattern could improve usability, where experienced users could indicate their familiarity with WeSIS and receive more data-focused responses, while less experienced users would receive additional guidance.

The chatbot currently lacks access to the actual database and instead relies on scraped indicator-related data to generate links and explanations. This is intentional to prevent LLM hallucinations, but enabling direct database access could enhance usability. However, such access would require refined prompt engineering to ensure the chatbot only returns data if it exists and provides a clear negative response otherwise. Additional safeguards would also be necessary to prevent misidentification of indicators with similar names, topics, or IDs. While this feature could significantly enhance research workflows by allowing direct data retrieval for comparisons and batch queries, it also poses risks if incorrect data is retrieved without verification.

Finally, as this chatbot is a prototype, it has several inherent limitations:

(1) Four-digit IDs may be misinterpreted as years rather than identifiers.
(2) Detection of multiple countries in queries is not functioning correctly.
(3) Due to the agent structure, the chatbot struggles to answer multiple questions in a single prompt, sometimes resulting in confusing or incomplete responses.

Addressing these challenges would significantly improve the chatbot's reliability, usability, and overall integration into the WeSIS research ecosystem.

6.2 User Study Findings and Implications for Future Development

The user study conducted on the ChatWeSIS prototype provides valuable insights into its effectiveness in assisting users with navigating the WeSIS platform. While the prototype showed positive results in supporting user navigation, several areas for enhancement were identified. This section outlines the key findings from the study and discusses the implications for the future development of ChatWeSIS, with a focus on improving clarity, resource accessibility, and personalization to better serve the needs of researchers.

6.2.1 Chatbot Usability and Interaction. The chatbot was generally seen as user-friendly and helpful, particularly for beginners. It provided a starting point for new users and was appreciated for its ability to direct users to relevant resources, such as links and YouTube videos. Some participants appreciated the chatbot's friendly tone and its ability to provide a basic understanding of the system, including guidance for uploading files and navigating the platform. However, there were also several weaknesses. Many participants found the chatbot's responses to be too broad or inaccurate. It often failed to directly answer specific questions or provide irrelevant data, such as redirecting users to external sources like the OECD database instead of providing WeSIS-specific data.

Several participants had difficulty finding the chatbot, navigating it, or understanding its functionality. Some felt that the chatbot was not intuitive enough and struggled with the design or location of certain features. Additionally, the chatbot did not retain conversation history, which made interactions less personalized and required users to repeat themselves. Some participants found the chatbot's responses to be too detailed, making it difficult to locate specific information. The chatbot often provided unnecessary information or failed to simplify complex concepts. Moreover, there was feedback regarding the chatbot's lack of clear instructions for using key features like uploading data. Some participants suggested that more proactive guidance, examples, and direct links to relevant pages would improve the experience.

6.2.2 Response Relevance and Efficiency. A key finding from the user study was that participants appreciated the information provided by the chatbot, although they occasionally felt that responses were more detailed than necessary, which could make finding specific information more challenging. Users preferred shorter, more direct responses, especially when the information provided was highly relevant to their immediate needs. To enhance the chatbot's

efficiency, it could be beneficial to program it to deliver concise, context-specific answers that align more closely with the user's current task [18].

In addition, the chatbot was generally effective in assisting users with platform navigation. However, there were moments when users had difficulty locating specific resources or datasets. This highlights an opportunity for the chatbot to improve its ability to recognize user tasks more effectively and provide even more tailored, task-oriented guidance. Interestingly, despite some of the prototype's responses not being entirely accurate, participants were still satisfied with their experience. This could be attributed to the participants' varying levels of familiarity with the WeSIS platform and their diverse academic backgrounds, which included fields such as Political Science, Computer Science, Sociology, and Geography. As a result, their expectations and understanding of the chatbot's performance varied.

In some cases, the accuracy of the chatbot's responses was not as critically assessed by participants, particularly those with less experience with the platform or those not experts in the specific domain of the task [29]. This suggests that user satisfaction may not always be directly linked to the accuracy of the chatbot's responses, especially for users with different academic backgrounds and varying familiarity with the platform.

6.2.3 Knowledge Discovery and Resource Accessibility. Participants reported that the chatbot was not always successful in helping them find the right resources on the WeSIS platform. In some cases, it directed them to irrelevant sections or failed to locate certain datasets. Many participants suggested that the chatbot should be more integrated with WeSISpedia to help users access information more efficiently. Improving the chatbot's ability to retrieve and organize platform resources is a key area for future development, as it would make the platform more useful and easier to navigate [36].

6.2.4 Technical Functionality and Suggestions for Improvement. Some technical issues were also highlighted during the study. The chatbot was not able to maintain continuity in conversations, meaning it could not remember previous interactions or provide follow-up questions. This limited its ability to offer a more personalized experience. Users suggested that the chatbot should have a memory feature to recall previous interactions, which would allow it to provide more tailored responses [12].

6.3 Key Recommendations

To improve the ChatWeSIS chatbot, several enhancements are recommended:

6.3.1 Technical Recommendations.

(1) **Upgrade Model and Memory Feature:** Move to a larger LLM for improved accuracy and integrate a memory system to retain context across interactions [32].
(2) **Infrastructure and Integration:** Ensure scalable resources (e.g., GPUs) and fully integrate the chatbot into WeSIS by consolidating backend services.
(3) **Automate Contextual Data Updates:** Automate the scraping and uploading of contextual data to keep information up-to-date, with manual uploads for additional data.
(4) **Role-Based Personalization:** Tailor responses based on user roles (guest, regular user, admin) and familiarity with the platform.
(5) **Direct Database Access:** Allow direct database access for data retrieval with safeguards to ensure accuracy and clarity in responses.
(6) **Address Prototype Limitations:** Fix issues with ID interpretation, multi-country detection, and handling multiple questions in one prompt to improve reliability.

6.3.2 User Experience Recommendations.

(1) **Simplify responses**: Focus on providing concise, context-specific answers that meet the user's immediate needs.
(2) **Enhance resource accessibility**: Integrate better with WeSISpedia and improve the chatbot's ability to find specific datasets and resources.
(3) **Add memory features**: Implement a memory system to recall previous interactions and offer more personalized responses.
(4) **Improve design and interactivity**: Make the chatbot's interface more intuitive, with features like speech bubbles, automatic prompts, and a streamlined design.
(5) **Expand knowledge**: Ensure the chatbot provides more in-depth knowledge of WeSIS-specific features and procedures, and proactively suggest relevant links and resources.
(6) **Provide clearer instructions**: Include more proactive guidance, examples, and direct links to key features such as uploading data.

7 Conclusion

The user study of the ChatWeSIS prototype, publicated by the Welfare Computational project team in March, 2025 in University of Bremen, Germany, provides a comprehensive evaluation of its effectiveness as a tool for assisting researchers in navigating the WeSIS platform. The study, involving ten participants with diverse academic backgrounds and varying levels of familiarity with WeSIS and chatbot usage, revealed both strengths and areas for improvement in the prototype's design and functionality. Through a three-phase methodology, task-based interaction, Chatbot Usability Questionnaire (CUQ), and open-ended interviews. The research highlighted the chatbot's potential as a user-friendly assistant while identifying critical limitations that must be addressed for future development.

The findings indicate that ChatWeSIS was generally well-received, particularly by beginners, for its ease of use, friendly interface, and ability to provide surface-level guidance, such as locating resources and offering basic navigation support. Participants with higher CUQ scores (e.g., Participants 4, 6, and 7, scoring 84.4) expressed satisfaction with its clarity and engagement, underscoring its value as an entry point for new users. However, expert users and those with specific research needs, such as Participants 2 and 5, reported lower satisfaction (CUQ scores of 56.2 and 68.8, respectively), citing the chatbot's inability to deliver precise, context-aware responses or handle complex, task-specific queries. Thematic and discourse analyses further revealed prevalent user confusion, criticism, and suggestions for improvement, with "Suggestions" being the most frequent feedback type, reflecting a proactive user desire to enhance the system.

Key challenges included the chatbot's lack of memory, leading to repetitive interactions, and its tendency to provide overly broad or irrelevant responses, which frustrated users seeking detailed, WeSIS-specific data. Technical limitations, such as reliance on manually scraped data, inability to differentiate user roles, and lack of direct database access, compounded these issues, hindering the chatbot's scalability and utility for advanced research workflows. Participants across all experience levels suggested improvements, including better contextual understanding, memory retention, and a more navigable interface, emphasizing the need for a shift from a basic assistant to a robust, adaptive tool.

In conclusion, while the ChatWeSIS prototype demonstrates promise in enhancing accessibility and usability within the WeSIS platform, its current iteration falls short of meeting the diverse needs of its user base, particularly for experienced researchers requiring in-depth support. Future development should prioritize integrating memory features (e.g., via LangGraph), enabling direct database access with safeguards against inaccuracies, and tailoring responses based on user expertise. Additionally, automating data updates, consolidating backend services, and improving integration with the Ruby on Rails-based WeSIS platform will be crucial for scalability and efficiency. By addressing these technical and usability gaps, ChatWeSIS can evolve into a powerful tool that not only simplifies navigation but also empowers researchers with precise, actionable insights, aligning with the broader goals of the WeSIS ecosystem.

Appendix

Tasks, CUQ, and Interview Questions PDF

A detailed PDF containing all the tasks, CUQ questions, and open-ended interview questions is available. You can access the document via the following link:

https://nc.uni-bremen.de/index.php/s/abJKJmjQLqCkBzo

References

[1] [n. d.]. *Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022, Revised Selected Papers | SpringerLink*. https://link.springer.com/book/10.1007/978-3-031-25581-6

[2] [n. d.]. *WeSIS - About*. https://wesis.org/static_pages/about_crc

[3] Eleni Adamopoulou and Lefteris Moussiades. 2020. An Overview of Chatbot Technology. In *Artificial Intelligence Applications and Innovations* (Cham, 2020), Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis (Eds.). Springer International Publishing, 373–383. doi:10.1007/978-3-030-49186-4_31

[4] Facebook AI. 2020. bart-large-mnli. https://huggingface.co/facebook/bart-large-mnli. Accessed: 2025-03-29.

[5] Addi Ait-Mlouk and Lili Jiang. 2020. KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data. PP (2020), 1–1. doi:10.1109/ACCESS.2020.3016142

[6] Virginia Braun, , and Victoria Clarke. 2006. Using thematic analysis in psychology. 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[7] Virginia Braun and Victoria Clarke. 2013. Successful Qualitative Research : A Practical Guide for Beginners. (2013), 1–400. https://www.torrossa.com/en/resources/an/5017629 Publisher: SAGE Publications Ltd.

[8] Iván Cantador, Jesús Viejo-Tardío, María E. Cortés-Cediel, and Manuel Pedro Rodríguez Bolívar. 2021. A Chatbot for Searching and Exploring Open Data: Implementation and Evaluation in E-Government. In *Proceedings of the 22nd Annual International Conference on Digital Government Research* (New York, NY, USA, 2021-06-09) *(dg.o '21)*. Association for Computing Machinery, 168–179. doi:10.1145/3463677.3463681

[9] Daniel Carlander-Reuterfelt, Álvaro Carrera, Carlos A. Iglesias, Óscar Araque, Juan Fernando Sánchez Rada, and Sergio Muñoz. 2020. JAICOB: A Data Science Chatbot. 8 (2020), 180672–180680. doi:10.1109/ACCESS.2020.3024795 Conference Name: IEEE Access.

[10] Patricia L. Carrell. 1982. Cohesion Is Not Coherence. 16, 4 (1982), 479–488. doi:10.2307/3586466 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/3586466.

[11] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. 12, 3 (2017), 297–298. doi:10.1080/17439760.2016.1262613

[12] K. Anders Ericsson and Herbert A. Simon. 1980. Verbal reports as data. 87, 3 (1980), 215–251. doi:10.1037/0033-295X.87.3.215 Place: US Publisher: American Psychological Association.

[13] European Union. 2018. General Data Protection Regulation (GDPR). https://gdpr.eu/ Accessed: 2025-03-30.

[14] Lizhou Fan, Sara Lafia, Lingyao Li, Fangyuan Yang, and Libby Hemphill. 2023. DataChat: Prototyping a Conversational Agent for Dataset Search and Visualization. 60, 1 (2023), 586–591. doi:10.1002/pra2.820

[15] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, Rebecca Wald, Fabio Catania, Raphael Meyer von Wolff, Sebastian Hobert, and Ewa Luger. 2021. Future directions for chatbot research: an interdisciplinary research agenda. 103, 12 (2021), 2915–2942. doi:10.1007/s00607-021-01016-7

[16] Kathleen Gregory, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2020. Lost or found? Discovering data needed for research. (2020). doi:10.1162/99608f92.e38165eb arXiv:1909.00464 [cs]

[17] Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. AI and the transformation of social science research. 380, 6650 (2023), 1108–1109. doi:10.1126/science.adi1778 Publisher: American Association for the Advancement of Science.

[18] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (New York, NY, USA, 2019-09-10) *(ECCE '19)*. Association for Computing Machinery, 207–214. doi:10.1145/3335082.3335094

[19] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2021. Talking datasets – Understanding data sensemaking behaviours. 146 (2021), 102562. doi:10.1016/j.ijhcs.2020.102562

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems* (2020), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[21] LimeSurvey Team. n.d.. LimeSurvey: Free Online Survey Tool. https://www.limesurvey.org/ Accessed: 2025-03-30.

[22] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016-05-07) *(CHI '16)*. Association for Computing Machinery, 5286–5297. doi:10.1145/2858036.2858288

[23] Vasilios Mavroudis. 2024. LangChain. (2024). https://www.academia.edu/125395559/LangChain

[24] Thai Ha Nguyen, Lena Waizenegger, and Angsana A. Techatassanasoontorn. 2022. "Don't Neglect the User!" – Identifying Types of Human-Chatbot Interactions and their Associated Characteristics. 24, 3 (2022), 797–838. doi:10.1007/s10796-021-10212-x

[25] Jakob Nielsen and Thomas K Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 206–213.

[26] Roland Oruche, Xiyao Cheng, Zian Zeng, Audrey Vazzana, MD Ashraful Goni, Bruce Wang Shibo, Sai Keerthana Goruganthu, Kerk Kee, and Prasad Calyam. 2025. Chatbot Dialog Design for Improved Human Performance in Domain Knowledge Discovery. (2025), 1–16. doi:10.1109/THMS.2024.3514742 Conference Name: IEEE Transactions on Human-Machine Systems.

[27] Brian Paltridge. 2021. *Discourse analysis*. Springer.

[28] Dijana Peras. 2025. Chatbot User Experience: Design and Evaluation. In *HCI International 2024 – Late Breaking Papers* (Cham, 2025), Adela Coman, Simona Vasilache, Fiona Fui-Hoon Nah, Keng Leng Siau, June Wei, and George Margetis (Eds.). Springer Nature Switzerland, 77–93. doi:10.1007/978-3-031-76806-4_6

[29] Stephen R. Porter, Michael E. Whitcomb, and William H. Weitzer. 2004. Multiple surveys of students and survey fatigue. 2004, 121 (2004), 63–73. doi:10.1002/ir.101 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ir.101.

[30] João António Reis, João Rafael Almeida, Tiago Melo Almeida, and José Luís Oliveira. 2024. Using Flowise to Streamline Biomedical Data Discovery and Analysis. In *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)* (2024-06). 695–700. doi:10.1109/MELECON56669.2024.10608738 ISSN: 2158-8481.

[31] Geovana Ramos Sousa Silva and Edna Dias Canedo. 2024. Towards User-Centric Guidelines for Chatbot Conversational Design. 40, 2 (2024), 98–120. doi:10.1080/10447318.2022.2118244 arXiv:2301.06474 [cs]

[32] Oguzhan Topsakal and T. Cetin Akinci. 2023. Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. 1 (2023), 1050–1056. doi:10.59287/icaens.1127

[33] NLP Town. 2020. bert-base-multilingual-uncased-sentiment. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. Accessed: 2025-03-29.

[34] Mohammad H. Vahidnia. 2024. Empowering geoportals HCI with task-oriented chatbots through NLP and deep transfer learning. 8, 4 (2024), 608–648. doi:10.1080/20964471.2024.2403166 Publisher: Taylor & Francis.

[35] Komal Vekaria, Prasad Calyam, Sai Swathi Sivarathri, Songjie Wang, Yuanxun Zhang, Ashish Pandey, Cong Chen, Dong Xu, Trupti Joshi, and Satish Nair. 2021. Recommender-as-a-service with chatbot guided domain-science knowledge discovery in a science gateway. 33, 19 (2021), e6080. doi:10.1002/cpe.6080 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.6080.

[36] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. 27, 3 (2020), 15:1–15:37. doi:10.1145/3381804

[37] Okan Yetişensoy and Hıdır Karaduman. 2024. The effect of AI-powered chatbots in social studies education. 29, 13 (2024), 17035–17069. doi:10.1007/s10639-024-12485-6

[38] Yuanxun Zhang, Prasad Calyam, Trupti Joshi, Satish Nair, and Dong Xu. 2023. Domain-Specific Topic Model for Knowledge Discovery in Computational and Data-Intensive Scientific Communities. 35, 2 (2023), 1402–1420. doi:10.1109/TKDE.2021.3093350 Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[39] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2024. Large Language Models for Information Retrieval: A Survey. doi:10.48550/arXiv.2308.07107 arXiv:2308.07107 [cs]