OSNABRÜCK UNIVERSITY

BACHELOR THESIS

Assessing Answer Accuracy, Hallucination, and Document Relevance in a RAG-Based Chatbot at the Osnabrück University

---

Author:
Marvin Ives René WURCH
Matriculation number: 991576

Supervisors:
Dr. phil. Tobias THELEN
M. Sc. Yesid Cano CASTRO

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Institute of Cognitive Science
School of Human Sciences

January 29, 2025

Eigenständigkeitserklärung/Declaration of Authorship

Ich, Marvin Ives René WURCH(Immatrikulationsnummer: 991576), erkläre, dass diese Arbeit mit dem Titel „Assessing Answer Accuracy, Hallucination, and Document Relevance in a RAG-Based Chatbot at the Osnabrück University" und die darin dargestellte Arbeit meine eigene ist.

Ich bestätige, dass die Inhalte der hier vorgelegten Leistung meinen eigenen Wissensstand, mein eigenes Verständnis und meine eigene Auffassung zum bearbeiteten Thema wiedergeben. Falls KI-Tools eingesetzt wurden, habe ich deren Einsatzweise und -zweck transparent angegeben. Darüber hinaus habe ich alle meine Quellen akademischen Standards entsprechend ausgewiesen. Ich bin bereit und fähig, die hier erläuterten Inhalte zu erklären und die entwickelten Standpunkte zu vertreten. Die vorliegende Leistung wurde weder zum Teil noch vollständig für einen anderen Kurs an dieser oder einer anderen Universität eingereicht.

I, Marvin Ives René WURCH(Matriculation number: 991576), declare that this thesis titled, "Assessing Answer Accuracy, Hallucination, and Document Relevance in a RAG-Based Chatbot at the Osnabrück University" and the work presented in it are my own.

I confirm that the content of this thesis represents my own knowledge, my own understanding and my own perspective on the topic. In case artificial intelligence tools were used, their way and purpose of usage has been made transparent. Moreover, I have cited all my sources in accordance with academic standards. I am ready and able to explain and defend the positions developed in this thesis. This thesis has not been submitted, either in part or whole, at this or any other university

Signed:

Date:

OSNABRÜCK UNIVERSITY

Abstract

School of Human Sciences
Institute of Cognitive Science

Bachelor of Science

Assessing Answer Accuracy, Hallucination, and Document Relevance in a RAG-Based Chatbot at the Osnabrück University

by Marvin Ives René WURCH

This thesis evaluates the performance of a Retrieval-Augmented Generation (RAG)-based chatbot designed to assist students and prospective students at Osnabrück University. The primary aim is to develop a domain-specific evaluation framework to assess answer accuracy, hallucination rates, and document relevance in bilingual contexts (German and English). A mixed-methods approach was employed, combining human evaluations with automated metrics to assess chatbot responses, optionally against human-provided reference answers. Data collection involved bilingual surveys to gather authentic user questions and generate ground-truth answers, resulting in a domain-specific dataset for future applications. A systematic assessment of the chatbot's output followed this data collection.

Key findings reveal the chatbot's strengths in coherence, clarity, and fluency, with a low occurrence of hallucinations. However, variability in answer accuracy and context quality indicates challenges in retrieving domain-specific information. Human evaluators and automated metrics demonstrated moderate agreement, with advanced methods such as the LLM-as-a-Judge framework yielding better correlations when reference answers were available. Indicating that automatic metrics should be used with caution as a stand-in for human evaluators. The study also highlights limitations in inter-annotator agreement and minor performance differences across languages.

The dataset and code are publicly available on the official GitHub repository of this thesis: `https://github.com/MarvinIRW/Assessing-Answer-Accuracy-Hallucination-and-Document-Relevance-in-virtUOS-Chatbot`

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
API	**A**pplication **P**rogramming **I**nterface
BART	**B**idirectional and **A**uto-**R**egressive **T**ransformer
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
BLEU	**B**i**L**ingual **E**valuation **U**nderstudy
CRAG	**C**orrective **R**etrieval-**A**ugmented **G**eneration
LLM	**L**arge **L**anguage **M**odel
ML	**M**achine **L**earning
NLG	**N**atural **L**anguage **G**eneration
NLP	**N**atural **L**anguage **P**rocessing
QA	**Q**uestion **A**nswering
RAG	**R**etrieval-**A**ugmented **G**eneration
ROUGE	**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

Chapter 1

Introduction

1.1 The Problem

Recent advancements in the field of pre-trained large language models (LLMs) have accelerated the use of artificial intelligence (AI) chatbots across various sectors. The deployment of chatbots has promising results in business integration as they provide a wide variety of advantages, including 24/7 availability, reduced costs, quick response times, and scalability (Camilleri and Troise, 2023; Stoilova, 2021). It is reasonable to assume that integrating a similar system in a university setting would provide comparable benefits.

However, although LLMs excel across a broad range of topics, they face challenges in effectively answering questions related to knowledge-intensive or domain-specific tasks (Wang et al., 2023a, Section 2.2). RAG systems can overcome this knowledge gap by incorporating external information into the response-generation process (Lewis et al., 2020). Nevertheless, critically evaluating the generated answers remains essential, as inaccuracies can persist.

These challenges are particularly relevant to Osnabrück University. An AI-driven chatbot offers clear potential, such as providing prospective students with information, guiding enrolled students through administrative tasks, or assisting international students with language or visa queries. However, ensuring the quality of its answers is important. Erroneous or misleading information about curricula, fees, deadlines, or official regulations is not merely inconvenient, but can damage the university's reputation. Moreover, it is easier to evaluate chatbot performance when the system is already in production (Barnett et al., 2024); however, responsible deployment in a university setting demands a robust evaluation framework *before* widespread adoption. This paradox underscores why an evaluation framework tailored to Osnabrück University's needs is vital.

Consequently, this thesis seeks to develop and implement a domain-specific evaluation system for a RAG-based chatbot aimed to be deployed at the Osnabrück University. Drawing on the literature that identifies unique risks and mitigation strategies for LLMs, this thesis focuses on three critical performance dimensions:

1. **Answer Accuracy** – How well does the chatbot address each query, stays true to the question's intent, and resolve the query in a relevant manner?

2. **Hallucination** – Does the chatbot introduce information unsupported by retrieved material or widely recognized facts?

3. **Document Relevance** – How effectively does the chatbot select the most pertinent university web pages or official sources when formulating answers?

1.2 Research Question

To explore these dimensions systematically, this thesis addresses the following overarching research questions:

> **To what extent does the RAG-based chatbot, intended for deployment at Osnabrück University, provide hallucination-free, accurate, and coherent answers?**
>
> **Can these dimensions be evaluated automatically?**

These main questions are further divided into the following sub-questions:

1. **Answer Accuracy/Relevance**
   - Are the answers correct, complete, and relevant?
2. **Answer Hallucination**
   - To what extent does the chatbot generate information not supported by retrieved or commonly known facts?
3. **Context Relevance**
   - How appropriate are the documents that the chatbot retrieves for a given question?
4. **Answer Coherence**
   - How coherent, clear, and fluent is the language usage of the chatbot?
5. **Language Comparison**
   - Does the chatbot perform differently in German vs. English?

In pursuit of these questions, this thesis adopts a multifaceted research methodology. First, authentic user queries were gathered through a bilingual survey that targeted prospective, enrolled, and international students. Next, a separate survey produced human-written reference answers for a subset of these questions, which served as ground-truth benchmarks. The RAG-based chatbot was then queried systematically, generating both German and English outputs. Finally, a third survey and additional automated metrics assessed the quality of the chatbot's answers.

By building this tailored evaluation framework, this thesis not only examines the chatbot's current capabilities, but also lays the groundwork for iterative improvements. The methods and metrics developed in this thesis enable the ongoing monitoring of performance, support continuous refinements to the system, and foster trust in the university's digital services.

1.3 Structure

- Chapter 2, **Literature Review:** Maps the evolution of chatbots, focusing on LLM-powered systems and the phenomenon of hallucinations. Reviews RAG as a technique to enhance factual correctness and discusses existing evaluation metrics.

- **Chapter 3, The Chatbot in Question - askUOS:** Examines the chatbot's architecture, focusing on its use of the LangChain framework, its bilingual capabilities, and its integration of tools specifically designed to address university-related queries.
- **Chapter 4, Methods:** Details the development process of the reference dataset and outlines the multistaged evaluation methodology designed to comprehensively assess the chatbot's performance.
- **Chapter 5, Results:** Presents findings on the chatbot's performance and analyzes differences between German and English responses.
- **Chapter 6, Discussion:** Interprets the results in the context of prior research and outlines recommendations for improvement. Reflects on the thesis limitations and proposes directions for future work in evaluating RAG-based chatbots in specialized domains.
- **Chapter 7, Conclusion:** Synthesizes the thesis by emphasizing key findings.

1.4 AI Usage in This Thesis

The use of AI tools has become widespread in everyday tasks, including academic research and writing (Khalifa and Albadawy, 2024), while also having associated risks (Messeri and Crockett, 2024). This thesis acknowledges the role of AI in its development, highlighting that ChatGPT was used to refine sections, correct grammatical errors, and provide alternative formulations for clearer communication. Additionally, it assisted in debugging the code for the dataset and evaluation, and in brainstorming improvements to the structural layout of the document. All of these uses were conducted under careful human oversight and control.

It is worth noting that the widespread nature of AI nowadays extends beyond specialized tools, such as ChatGPT. Commonly used search engines such as Google also integrate AI technologies to enhance search results and user experience. Thus, distinguishing between *AI usage* and *non-AI usage* is becoming increasingly impractical. Instead, the researcher takes the perspective that reasonable usage of AI tools is both practical and necessary for modern academic work. This approach ensures efficiency while maintaining a focus on critical evaluation and intellectual contribution, as human expertise and oversight remains fundamental to the academic process.

To address the identified challenges and design a robust evaluation framework for the proposed chatbot, it is crucial to explore the evolution and current capabilities of chatbots, particularly those powered by LLMs. The next chapter provides a detailed review of relevant literature, focusing on the challenges of hallucinations and the potential of RAG to address these limitations.

Chapter 2

Literature Review

This literature review comprehensively examines the development and challenges of chatbots, particularly those powered by LLMs. The chapter begins by mapping the evolution of chatbots and highlighting their increasing sophistication and integration into various sectors, including education. It then delves into the specific challenges associated with LLM-powered chatbots, focusing on hallucinations. This review explores different types, implications, and mitigation strategies of hallucinations. Subsequently, RAG is introduced as a promising approach to enhance the quality of LLM outputs by incorporating external knowledge sources during the generation process. The mechanisms, benefits, challenges, and recent developments of RAG are discussed in detail. The chapter also examines evaluation metrics for LLMs, emphasizing methods for assessing the effectiveness of RAG systems. Finally, it identifies the existing research gap in evaluating RAG-based chatbots within specialized settings and outlines this thesis's contributions to addressing this gap.

2.1 Introduction to Chatbots

Chatbots, a mixture of *chat* and *robot*, are computer programs designed to simulate human conversation through text or voice interactions (Wang and Petrina, 2013). Initially developed as text-based dialog systems, chatbots have evolved significantly over the decades by integrating advanced technologies such as AI, Natural Language Processing (NLP), and Machine Learning to enhance their conversational capabilities and user experience (Al-Amin et al., 2024; Balcombe, 2023).

The primary function of chatbots is to interact with users in a natural and intuitive manner by providing information, assistance, or entertainment. They have been adopted across various sectors, including customer service, healthcare, education, and business, because of their ability to offer 24/7 availability, instant responses, and scalability (Zumstein and Hundertmark, 2017; Camilleri and Troise, 2023; Stoilova, 2021).

In educational institutions, chatbots assist in providing administrative support (Elnozahy et al., 2019; Hien et al., 2018), enhancing the learning experience (Lin and Tsai, 2019; Sandu and Gide, 2019), assessing student learning ability (Durall and Kapros, 2020), and improving academic research experience (Mckie and Narayan, 2019).

The evolution of chatbots mirrors the advancements in computational linguistics and AI technologies. This progression began with simple rule-based systems, such as ELIZA, which relied on pattern matching (Weizenbaum, 1966). Later, chatbots, such as ALICE, introduced the use of Artificial Intelligence Mark-up Language, employing a tree-based database structure for improved dialogue management (AbuShawar and Atwell, 2015). Today's sophisticated AI-driven conversational agents, such as ChatGPT (OpenAI et al., 2024a), Gemini (Gemini Team et al., 2024), and Claude (Anthropic, 2024) leverage the latest achievements in LLMs, enabling interactions that even integrate images and videos. This technological evolution has transformed human engagement with machines significantly.

2.2 Challenges in LLM-Powered Chatbots

As LLMs are now the backbone of chatbots, this thesis takes a closer look at them. LLMs offer a variety of strengths and potentials. They are equipped with a vast amount of world knowledge, can capture an enormous quantity of factual information (Petroni et al., 2019; Cohen et al., 2023), and can understand and produce highly coherent text (Hariri, 2024).

The potential of LLMs must be evaluated alongside the challenges they present, including among others the reinforcement and amplification of biases inherent in their training data, limitations in handling long contexts, security vulnerabilities, implications for social connections, potential over-reliance in education, and non-determinism in outputs. First, Hayes, Yax, and Palminteri (2024) and Lee (2023) demonstrated that LLMs can reinforce existing biases by exhibiting value biases similar to humans and by unintentionally self-reinforcing biases through generated text feeding into future training data. Second, LLMs face challenges in terms of effectively utilizing long context lengths. Liu et al. (2023a) found that LLMs struggle to retrieve and use information from the middle of long inputs, and Shi et al. (2023) showed that irrelevant information can distract LLMs, degrading performance in tasks like mathematical reasoning. In addition, Mirzadeh et al. (2024) suggested that current LLMs lack genuine logical reasoning abilities and rely heavily on pattern matching. Moreover, information security concerns have been raised. Yang et al. (2023) highlighted significant security threats to chatbots powered by LLMs, such as malicious input and data breaches, compromising user data and privacy. In terms of social connection, Folk, Yu, and Dunn (2024) found that, while chatbots can provide feelings of social connection, they may not fully replicate human interactions, especially when emulating human-like behaviors too closely. Another concern is the potential overreliance on LLMs in education. Wu (2024) cautioned that excessive reliance on LLMs by students and teachers can undermine the development of critical thinking and problem-solving skills. Additionally, it risks replacing human-driven creativity and personalization in teaching, which are essential for addressing diverse student needs and fostering innovative learning environments. Finally, non-determinism in LLM outputs poses a challenge. Ouyang et al. (2024) revealed that models like ChatGPT can produce significantly different code outputs when given the same prompt, affecting correctness and consistency, highlighting that setting the temperature parameter to zero does not guarantee deterministic outputs.

Apart from these challenges, concerns about the factuality and faithfulness of LLM outputs rise.

2.2.1 Hallucinations

The phenomenon of hallucinations in LLMs is linked to challenges across data, training, and inference stages (Huang et al., 2023; Ji et al., 2024). This issue arises, in part, because LLMs lack genuine understanding of real-world facts. They generate responses based on patterns in extensive training datasets, implicitly encoding information within their parameters rather than storing explicit factual knowledge, as traditional knowledge bases do (AlKhamissi et al., 2022; Hu et al., 2023). As Xu, Jain, and Kankanhalli (2024) have shown, hallucinations are not a temporary issue, but a fundamental characteristic of LLMs.

2.2.2 Definition of Hallucinations

In the context of LLMs, hallucinations are described as instances where the model generates content *"that seem syntactically sound, fluent, and natural but are factually incorrect, nonsensical, or unfaithful to the provided source input"* (Varshney et al., 2023). These manifest as factual errors, logical inconsistencies, or irrelevant information that cannot be verified (Huang et al., 2023; Ji et al., 2024; Li et al., 2024). The term draws an analogy to hallucinations in human psychology, in which perceptions do not align with reality (Macpherson and Platchias, 2013).

2.2.3 Types of Hallucinations

Early research categorized hallucinations into two main types:

- **Intrinsic Hallucinations:** Occur when the generated output does not accurately reflect or is inconsistent with the input, indicating a deviation within the context of the provided information (Ji et al., 2024; Maynez et al., 2020).
- **Extrinsic Hallucinations:** Arise when the output includes information that cannot be verified against the source content, introducing potentially incorrect external information (Ji et al., 2024; Maynez et al., 2020).

Recent studies have proposed more nuanced classifications of hallucinations in LLMs. Zhang et al. (2023) introduced a different categorization due to the versatility of LLMs, identifying the following types:

- **Input-Conflicting Hallucination:** When the generated content diverges from the users input. For example, the generated content may misinterpret or contradict the user's specific instructions or queries.
- **Context-Conflicting Hallucination:** LLMs can show self-contradictions when producing long or multi-turn responses.
- **Fact-Conflicting Hallucination:** Conflicts with established world knowledge, leading to incorrect or misleading outputs.

Zhang et al. (2023) further argue that current literature focuses on fact-conflicting hallucinations. This shift in emphasis is possibly due to the extensive study of the first two types of hallucinations (Ji et al., 2024) and the greater challenge it poses to LLMs due to the absence of an authoritative knowledge source as a reference.

Li et al. (2024) split factuality hallucinations into a fine-grained categorization:

- **Entity-Error Hallucination:** The model generates incorrect entities, such as wrong names, dates, locations, or objects that contradict established world knowledge.
- **Relation-Error Hallucination:** Involves generating incorrect relationships between entities, such as inaccurate quantitative or chronological connections.
- **Incompleteness Hallucination:** Occurs when the model provides incomplete outputs, especially in responses requiring lengthy or list-based answers.

- **Outdatedness Hallucination:** Arises when the model generates information that was accurate in the past but is no longer correct, often due to time-limited training data.
- **Overclaim Hallucination:** Refers to statements where the model expresses information beyond the scope of known factual knowledge.
- **Unverifiability Hallucination:** Occurs when the generated information cannot be verified against existing sources.

Mishra et al. (2024) further expanded on this by developing a fine-grained taxonomy that captures the specific nature and sources of hallucinations, including contradictions at different levels, subjective opinions, and unverifiable statements. They demonstrated that this detailed classification, distinguishing hallucinations from simplistic binary differentiations, such as factual or not factual, enabled more precise detection and correction strategies.

2.2.4 Challenges and Implications

Hallucinations present significant challenges in the deployment of LLM-powered chatbots, particularly in domains where factual accuracy is critical such as education, healthcare, and legal services (Al-Amin et al., 2024). The vast amount of training data used for LLMs can include outdated, biased, or incorrect information, aggravating the hallucination problem (Huang et al., 2023; Zhang et al., 2023).

Hallucinated text often appears fluent and natural even though it is unfaithful to the source material. It seems contextually grounded, yet the underlying context is difficult to identify or verify. Like psychological hallucinations, which can be challenging to distinguish from actual perceptions, hallucinated text can be difficult to detect at first glance because of its convincing surface-level presentation (Ji et al., 2024). This raises concerns about misinformation, user trust, and potential real-world consequences when incorrect information is accepted as true (Pan et al., 2023; Chen and Shu, 2023; Si et al., 2024).

It is worth noting that the literature sometimes has different opinions on the definitions of *faithfulness* and *factuality* with respect to hallucinations. *Faithfulness* is remaining consistent and true to the given source. *Factuality* denotes the quality of being real or grounded in fact. The terms *factuality* and *faithfulness* can overlap or diverge, depending on what is considered the fact. Maynez et al. (2020) distinguish *factuality* from *faithfulness* by defining the fact as world knowledge. On the other hand, Dong et al. (2020) used the source input as the fact to assess factual correctness, making *factuality* and *faithfulness* indistinguishable. This thesis follows the definition of Maynez et al. (2020), as the researcher believes that differentiating between source knowledge and world knowledge offers greater clarity.

2.2.5 Mitigation Strategies

Addressing hallucinations in LLMs is an active area of research, with several proposed mitigation strategies:

- **Improving Training Data Quality:** Curating and updating training datasets to ensure accuracy and relevance can reduce the occurrence of hallucinations due to outdated or incorrect information (Huang et al., 2023).

- **Reinforcement Learning from Human Feedback:** Incorporating human evaluations into the training process helps align the model's outputs with human judgments of correctness and relevance (Ouyang et al., 2022).
- **Fine-Grained Detection and Editing Systems:** Implementing systems that can identify specific types of hallucinations enables targeted corrections and improves overall reliability (Mishra et al., 2024; Varshney et al., 2023).
- **Self-Reflection:** Crafting prompts and a response loop that guides the model towards generating more accurate and fact-based responses can help mitigate hallucinations (Ji et al., 2023).
- **Fine-Tuning:** Adjusting model parameters with specific objectives, such as instruction-based or preference-based methods, to enhance performance on targeted tasks (Tian et al., 2023; Chung et al., 2022)
- **Retrieval-Augmented Generation:** Providing extra context alongside the user input via a retriever with an external database (Lewis et al., 2020; Shuster et al., 2021).

2.2.6 Conclusion

Hallucinations in LLM-powered chatbots pose significant challenges to their reliability and trustworthiness. Understanding hallucination types and causes is critical for developing mitigation strategies. As chatbots become more integrated into critical applications, decreasing hallucinations becomes vital for preventing the spread of misinformation and maintaining user trust. Incorporating RAG into LLMs has been shown to benefit the faithfulness of the system (Ovadia et al., 2024).

2.3 Retrieval-Augmented Generation (RAG)

RAG enhances LLMs by integrating external knowledge sources during the generation process. By retrieving relevant information from databases or corpora, RAG systems provide LLMs with up-to-date and accurate contexts, reducing the likelihood of hallucinations (Shuster et al., 2021; Huang and Huang, 2024; Gao et al., 2024; Mallen et al., 2023; Asai et al., 2024). This chapter explores the principles of RAG, its components, and how it addresses the limitations of LLM-powered chatbots discussed in the previous section.

2.3.1 Mechanism of RAG

RAG systems typically consist of two main components: a *retriever* and a *generator*. The retriever searches an external knowledge base or corpus to find relevant documents or passages based on the user's input or the current generation context. The generator then conditions its output on both the user's input and the retrieved information.

The retrieval process can occur once at the start (single-time retrieval) or iteratively throughout the generation process (multitime retrieval) (Ram et al., 2023). Some advanced RAG systems employ active retrieval strategies, where the model decides when and what to retrieve as it generates each token or sentence, showing better results in multiple tasks and datasets (Jiang et al., 2023).

While early methods depended on specialized LLMs for retrieval-augmented language modeling (Khandelwal et al., 2020; Borgeaud et al., 2022), recent research has shown that appending retrieved documents to an LLM's input can be effective as well (Khattab et al., 2023; Ram et al., 2023), enabling the use of retrieval-augmented strategies with LLMs accessible only via Application Programming Interfaces (APIs).

2.3.2 Benefits of RAG in Addressing Hallucinations

By incorporating external information, RAG systems can significantly reduce the occurrence of hallucinations in LLM outputs. The reliance on up-to-date and verifiable knowledge sources helps to ground the generated responses, ensuring that they are consistent with real-world facts, for example, by providing citations (Gao et al., 2023). Studies have shown that RAG models outperform traditional LLMs in knowledge-intensive tasks, providing more accurate and hallucination-free responses (Shuster et al., 2021; Zhao et al., 2024; Ovadia et al., 2024; Mallen et al., 2023).

For instance, Shuster et al. (2021) demonstrated that integrating retrieval mechanisms into dialogue models reduced hallucinations by over 60% compared to standard models. Similarly, Ovadia et al. (2024) found that RAG outperformed unsupervised fine-tuning in knowledge-intensive tasks, suggesting that retrieval augmentation effectively updates and expands LLM knowledge.

2.3.3 Challenges and Limitations of RAG

Despite its advantages, RAG introduces several challenges, in addition to those discussed earlier for LLMs. One major concern is the dependency on the quality and relevance of retrieved documents. If the retriever fetches incorrect or irrelevant information, it can lead to degraded performance or introduce new inaccuracies (Yan et al., 2024; Cuconasu et al., 2024). Surprisingly, Cuconasu et al. (2024) found that adding random documents to the prompt can increase accuracy by more than 30%. Although they conducted their experiments using models smaller than those commonly used today, these findings emphasize the importance of creating specialized strategies for integrating retrieval with language generation models, underscoring the need for continued research on the robustness of RAG.

Furthermore, incorporating retrieval mechanisms increases the computational overhead and system complexity, potentially affecting inference speed, scalability, and cost (Huang and Huang, 2024; Zhao et al., 2024).

2.3.4 Recent Developments in RAG

Recent studies have focused on improving the components of RAG systems to address these challenges.

Mallen et al. (2023) investigated LLMs parametric knowledge, finding that while these models are proficient at recalling well-known or frequently discussed facts, they struggle with less common or niche information. To overcome this limitation, they developed an adaptive retrieval method that dynamically retrieves non-parametric knowledge only when the LLM is deemed unlikely to have sufficient information. This approach not only improves the response accuracy for less common queries but also reduces inference time.

Active retrieval methods, such as the Forward-Looking Active Retrieval augmented generation proposed by Jiang et al. (2023), allow the model to iteratively retrieve information during the generation process, thereby enhancing the relevance and factual accuracy of long-form text generation.

Other studies have explored query rewriting techniques to improve retrieval quality. Ma et al. (2023) proposed a Rewrite-Retrieve-Read framework, where the model refines the search query to bridge the gap between the user's input and the required knowledge for effective retrieval.

A recent study by Yan et al. (2024) introduced Corrective Retrieval Augmented Generation (CRAG), which enhances RAG robustness through a lightweight evaluator that assesses the retrieval quality. CRAG triggers corrective actions, refining or replacing the retrieved content with web search results as needed. This adaptive approach ensures only relevant information contributes to generation, reducing hallucinations and improving response accuracy.

Adaptive-RAG, introduced by Jeong et al. (2024), proposes a dynamic framework for retrieval-augmented LLMs that adapts retrieval strategies based on the query complexity. This model assesses whether a given query requires a simple, single-step retrieval, multiple iterative steps, or no retrieval at all. By employing a classifier trained to identify query complexity, Adaptive-RAG achieves a balanced approach, reducing the computational overhead for simple queries while ensuring thorough multistep processing for more complex questions. This adaptive methodology improves retrieval efficiency and accuracy compared to rigid strategies, and has been shown to significantly outperform static models on diverse open-domain question-answer (QA) tasks.

2.3.5 Conclusion

Retrieval-Augmented Generation presents a promising approach for enhancing the reliability of LLM-powered chatbots by addressing the issue of hallucinations. By leveraging external knowledge sources, RAG systems can produce grounded and verifiable responses, particularly in knowledge-intensive applications. However, challenges remain in optimizing retrieval mechanisms, integrating retrieved information effectively, and managing the computational overhead.

2.4 Evaluation of Large Language Models

Guo et al. (2023) published an extensive servery about the evaluation of LLMs on many different levels, including capability, alignment and safety evaluation. As this thesis focuses on evaluating quality and document relevance, only relevant methods are discussed, leaving broader aspects of LLM evaluation beyond its scope.

2.4.1 Evaluation Methods

Traditional evaluation metrics such as Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which rely on n-gram overlap, have been shown to correlate poorly with human judgments regarding hallucinations (Maynez et al., 2020; Dhingra et al., 2019). Therefore, additional metrics have been developed to evaluate generated text.

These evaluation methods can be grouped into three categories: lexical or rule-based, model-based, and human evaluation.

Lexical or Rule-Based

Lexical or rule-based methods rely on surface-level features and predefined rules to assess the quality of generated text against a reference text.

Word Overlap: In addition to exact matches, early evaluation metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) were developed to measure lexical overlap between the generated text and a reference. However, while these metrics are easy to implement, they struggle to reliably detect hallucinations because they cannot capture semantic discrepancies or factual inaccuracies (Maynez et al., 2020; Dhingra et al., 2019). That being said, ROUGE is one of the most popular metrics reported in NLP (Gehrmann, Clark, and Sellam, 2022; Grusky, 2023). QUIP-Score (Weller et al., 2024) builds on the idea of exact match, but enhances grounding in a reference text by quantifying the amount of generated text that directly quotes known sources.

Model-Based

Model-based methods utilize neural network models, including specialized evaluation models and LLMs, to assess the quality of the generated text. By leveraging deep learning techniques, these methods aim to capture semantic nuances and factual consistency that traditional lexical or rule-based methods may overlook.

Semantic and Text-Based Alignment: This category includes methods that assess the alignment between generated text and reference text based on semantic similarity and textual coherence. Key approaches are BERTScore (Zhang et al., 2020), BLEURT (Sellam, Das, and Parikh, 2020), and BARTScore (Yuan, Neubig, and Liu, 2021).

BERTScore employs contextual embeddings from a bidirectional encoder representations from transformers model (BERT) (Devlin et al., 2019) to measure the semantic similarity between generated and reference texts at the token level, providing a nuanced analysis beyond surface-level word overlap (Zhang et al., 2020). BLEURT builds upon this by pre-training on synthetic sentence pairs and fine-tuning on human-annotated ratings, enhancing its ability to capture subtle shifts in text quality and fluency (Sellam, Das, and Parikh, 2020). BARTScore utilizes a bidirectional and auto-regressive transformer model (BART) (Lewis et al., 2019) to evaluate the generated text through a scoring system that considers aspects such as translation accuracy and fluency, making it adaptable across various text evaluation tasks (Yuan, Neubig, and Liu, 2021).

Question-Answering-Based: QA-based methods evaluate models by assessing if answers derived from both the source text and the generated text remain consistent in response to specific questions. This process involves generating questions from the system text and using a QA model to find corresponding answers in the source text. The similarity between answers indicates the degree of faithfulness of the generated text (Durmus, He, and Diab, 2020; Honovich et al., 2021).

Internal State: For LLMs with access to their internal states, various methods are used to quantify the model's confidence in its outputs. Techniques include measuring conditional entropy to detect moments of high uncertainty (Poel, Cotterell, and Meister, 2022), using length-normalized sequence log-probability (Guerreiro, Voita, and Martins, 2023), and obtaining a probability score by examining logit output values (Varshney et al., 2023). Another approach, GPTScore, leverages the generative pre-trained model's internal token log-probabilities and evaluates text quality by calculating the weighted log-probability of each token, given the prior context and a task-specific prompt. This method uses prompt engineering to define evaluation protocols, enabling task-dependent, customizable scoring without extra training (Fu et al., 2023). Detecting hallucinations during the generation process is particularly advantageous, as they tend to propagate during the generation process. Thus, identifying and correcting them in real time can enhance the overall output quality (Varshney et al., 2023).

LLM-as-a-Judge: LLM-as-a-Judge methods leverage LLMs to evaluate the quality of generated text by following specific task-oriented prompts. Chiang and Lee (2023) and Wang et al. (2023b) were one of the first studies to apply this approach, they investigated the potential of treating ChatGPT as a human evaluator. By providing task- and aspect-specific prompts, they extracted scores that showed high correlation with human judgments, demonstrating ChatGPT's potential effectiveness as a natural language generation (NLG) evaluation metric. Building on this, G-EVAL introduced a framework using GPT-4 with chain-of-thought prompting and probability-weighted scoring, enhancing alignment with human evaluations in tasks like summarization and dialogue generation (Liu et al., 2023b). Zheng et al. (2023) further underlines the promising results of LLMs as a judge, showing an agreement rate of over 80% with humans in comparing generated text, comparable to the level of agreement seen among human experts.

LLM evaluation offers several benefits: it is reasonably reproducible, as the results can be controlled by model version, random seed, and hyperparameters. Additionally, each evaluation sample is processed independently, avoiding the cumulative biases human evaluators might develop across multiple samples. Furthermore, LLM evaluation is cost-effective and significantly faster than human evaluation, thereby reducing both time and financial costs. It also prevents human evaluators from being exposed to potentially harmful or objectionable content, a concern in sensitive NLP tasks (Chiang and Lee, 2023; Zheng et al., 2023).

On the downside, LLMs can show, in addition to the already discussed biases, a position bias when comparing two candidates, often preferring the first-presented response (Wang et al., 2023c; Zheng et al., 2023), and a verbosity bias, when LLMs favor longer responses, even if the LLM only repeats the already stated information (Zheng et al., 2023). In addition, Zheng et al. (2023) observed that LLM judges may exhibit self-preference bias, often favoring answers generated by the same model. Building on this, Panickssery, Bowman, and Feng (2024) provided initial evidence that this bias is linked to the model's capability to recognize its own outputs, which correlates with a stronger preference for its own answers than those of other models or humans.

In addition, refer to the recent survey by Gu et al. (2025) for a detailed breakdown of this method.

Human Evaluation

Although most papers regarding the evaluation of LLMs use human evaluation as the gold standard, there is a lack of standardization and significant challenges affecting the reliability and reproducibility of these evaluations (Belz et al., 2023;

Gehrmann, Clark, and Sellam, 2022; Krishna et al., 2023). Human evaluation is considered essential for assessing the quality of outputs generated by LLMs; however, variability in evaluation protocols, criteria, and insufficient reporting across studies make it difficult to compare results or replicate findings.

Belz et al. (2023) conducted a comprehensive analysis and found that the majority of human evaluations in NLP are not repeatable or reproducible due to missing information, unresponsive authors, and experimental flaws. They reported that only 13% of the papers they examined had sufficiently low barriers to reproduction and enough obtainable information to be considered for reproduction. Furthermore, all but one of the experiments selected for reproduction had flaws that questioned the meaningfulness of conducting a reproduction. This highlights the need for better reporting practices and the standardization of human evaluation studies. Belz, Mille, and Howcroft (2020) suggested a classification system and further refined this approach with the Human Evaluation Datasheet (Shimorina and Belz, 2022).

Subjectivity in human evaluations can lead to low inter-annotator agreement, thereby impacting the consistency of the evaluations (Gehrmann, Clark, and Sellam, 2022; Krishna et al., 2023). Recent efforts have aimed to improve the consistency and reliability of human evaluation. Krishna et al. (2023) found that fine-grained annotations yield higher inter-annotator agreement than coarse-grained annotations. This suggests that providing annotators with detailed guidelines and specific criteria enhances the quality and consistency of human evaluations.

Despite advancements in automated and model-based evaluation metrics, human judgment remains essential for capturing the nuances of LLM outputs, particularly in terms of hallucinations and accuracy. Incorporating best practices, standardized protocols, and detailed reporting can mitigate some challenges and ensure more reliable and valuable human evaluations (Belz et al., 2023; Gehrmann, Clark, and Sellam, 2022).

2.4.2 Evaluation Metrics for Retrieval-Augmented Generation Systems

Evaluating RAG systems requires specialized methods due to their unique integration of retrieval mechanisms with generative language models. Traditional evaluation metrics often fail to capture the nuances of document relevance that are specific to RAG architectures. Therefore, tailored evaluation frameworks have been developed to assess the performance and reliability of RAG applications.

Saad-Falcon et al. (2024) introduced ARES, an automated evaluation framework designed to assess RAG systems along the dimensions of context relevance, answer faithfulness, and answer relevance. ARES utilizes synthetic data generation and fine-tunes lightweight language model judges to evaluate individual RAG components, employing prediction-powered inference to enhance evaluation accuracy with a small set of human-annotated data.

Similarly, Es et al. (2023) proposed RAGAS, a reference-free evaluation framework that focuses on faithfulness, answer relevance, and context relevance. RAGAS employs language model-prompting strategies to assess these aspects without relying on ground-truth annotations, thereby facilitating rapid evaluation cycles and adaptability to various contexts.

Sivasothy et al. (2024) developed RAGProbe, an automated approach that generates domain-specific question-answer pairs using a set of evaluation scenarios. RAGProbe aims to expose failure points in RAG pipelines by providing a schema for different types of question-answer pairs, enabling continuous monitoring and improvement.

Zhu et al. (2024) introduced RAGEval, a scenario-specific evaluation framework for assessing RAG systems across domains like finance, law, and medicine. Using a schema-based pipeline, RAGEval first generates documents encapsulating key scenario-specific elements, then creates question-answer pairs and optimizes these for evaluation. The framework assesses models on metrics such as Recall, Effective Information Rate, Completeness (coverage of essential information), Hallucination (inaccuracies or contradictions), and Irrelevancy (unrelated content). While RAGEval relies heavily on LLMs, it produces artificial datasets that avoid privacy and intellectual property concerns, achieving reliable, scenario-focused evaluations.

Moreover, Niu et al. (2024) introduced RAGTruth, a high-quality hallucination detection dataset for RAG models. RAGTruth includes nearly 18,000 responses annotated for word-level hallucinations generated across multiple LLMs within RAG tasks such as question answering and summarization. The dataset labels hallucinations into four different types, allowing for detailed analysis. RAGTruth has enabled the fine-tuning of smaller LLMs for effective hallucination detection, which has shown performance competitive with advanced models such as GPT-4, thus demonstrating its potential to improve LLM trustworthiness through targeted fine-tuning.

By focusing on critical aspects, such as document relevance and hallucination reduction, these methods address the specific challenges inherent to RAG systems.

2.5 Research Gap and Thesis Contribution

Despite the advancements in LLMs and RAG systems, there is a noticeable gap in the literature concerning the evaluation of RAG-based chatbots within specialized and smaller-scale settings, such as university environments. Most existing studies focus on general-purpose chatbots or large-scale applications and often neglect the unique challenges and requirements of deploying chatbots in specific institutional settings (Lewis et al., 2020; Shuster et al., 2021; Gao et al., 2024).

Current evaluation frameworks and metrics are generally designed for broad applications, and may not adequately address the nuances of domain-specific information needs (Saad-Falcon et al., 2024; Es et al., 2023). Moreover, there is a lack of tailored methodologies for assessing hallucination rates, answer accuracy, coherence, and document retrieval in RAG-based chatbots when applied to localized content, such as university-specific queries.

This thesis aims to fill this gap by developing a customized evaluation system for the RAG-based chatbot intened to be deployed at the Osnabrück University. The contributions of this thesis are threefold: **(1)** A domain-specific dataset reflecting genuine queries and reference answers. **(2)** A multi-method evaluation of a RAG-based chatbot's quality. **(3)** A reproducible evaluation procedure for assessing future iterations of the chatbot.

By addressing the specific challenges of evaluating a RAG-based chatbot in a university setting, this thesis contributes to the broader field of AI-driven conversational agents, enhancing understanding of their applicability and reliability in specialized domains.

2.6 Summary

This chapter has explored the historical development, challenges, and advancements of chatbots, with a particular focus on LLM-powered systems. The evolution of chatbots from rule-based systems to sophisticated LLMs has highlighted their increasing ability to simulate human-like interactions and address complex user queries.

A critical challenge discussed is the phenomenon of hallucinations in LLMs, where models produce outputs that are factually incorrect, nonsensical, or unfaithful to source material. The literature identifies various types and categorizations of hallucinations, ranging from intrinsic and extrinsic errors to finer-grained classifications, such as entity and relation errors. Mitigation strategies, including improved training data, reinforcement learning, and RAG, offer promising avenues for enhancing the accuracy and reliability of LLM outputs.

In particular, RAG is a promising approach to reduce hallucinations by integrating external knowledge during the generation process. However, the incorporation of retrieval mechanisms introduces additional challenges such as dependency on the quality of retrieved documents and increased system complexity. Recent advancements have sought to address these issues using adaptive retrieval strategies, query refinement techniques, and domain-specific evaluation frameworks.

Despite these advancements, there is no universally accepted method for evaluating the quality of LLMs, particularly in specialized contexts such as university environments. Existing methods have strengths and limitations, and their applicability depends on the specific evaluation goals. This thesis will leverage these established methods to evaluate the Osnabrück University chatbot. The specifics of this chatbot are introduced in the next chapter.

Chapter 3

The Chatbot in Question - askUOS

Before outlining the methods used to evaluate the chatbot, this chapter provides a detailed look at how the chatbot is implemented, focusing in particular on the `CampusManagementOpenAIToolsAgent` class. While the project uses StreamLit for the user interface, the discussion centers on the underlying logic and architecture of the chatbot.

Since the chatbot is still in development at the time of writing, this thesis forked the official GitHub page[1] into the GitHub page of this thesis on the 8th of January 2025 to have a consistent point for evaluation. Thus, the chatbot implementation available at `Thesis GitHub/chatbot` is discussed in this thesis.

3.1 Overview of the Chatbot Architecture

At a high level, the chatbot is built on the *LangChain* framework, which allows LLMs to reason about user queries, decide whether to call external *tools*, and synthesize final responses. Figure 3.1 provides a conceptual overview:

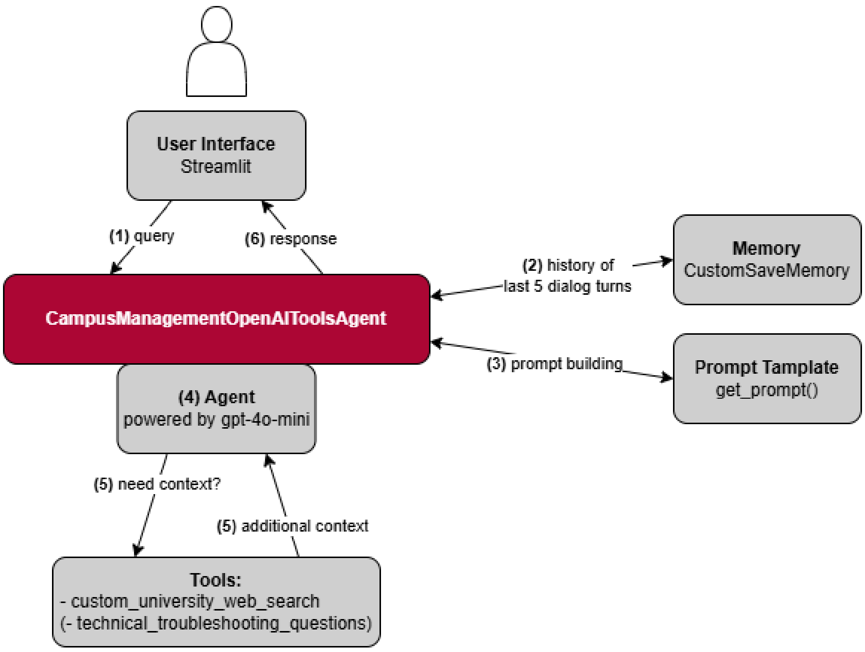


FIGURE 3.1: High level workflow overview of askUOS

[1]Official askUOS GitHub: `https://github.com/virtUOS/askUOS`

The *agent* is central to this architecture. In LangChain terms, the agent orchestrates how the LLM interacts with external tools, memory, and user input. In this project, the primary agent is implemented in the `CampusManagementOpenAIToolsAgent` class.

1. **User Query**: The user asks a question.
2. **Memory**: The history of the past dialogue turns, if available, is retrieved.
3. **Agent + Prompt**: The user query, conversation history, and system prompt are combined into the prompt text for the agent (see Appendix Fig. B.1).
4. **LLM Reasoning**: The LLM composes its final response, potentially multiple times calling tools (e.g., to scrape or retrieve data) until it is confident in the answer.
5. **Tools**: If the LLM/agent decides it needs additional information, it can call:
   - A **technical troubleshooting question** system is powered by the vector store ChromaDB to look up locally stored documents. Although the database exists, it is relatively limited in scope and can be considered a placeholder for future development. Therefore, it is not discussed further in this thesis.
   - A **web search tool** to scrape the Osnabrück University's website for up-to-date information.
6. **Response**: The chatbot provides a final text answer, including references or source links, to the user interface.

3.2 The `CampusManagementOpenAIToolsAgent` Class

3.2.1 Purpose and Role

The `CampusManagementOpenAIToolsAgent` class is designed to answer questions related to Osnabrück University. It leverages a combination of:

Prompt Templates to guide the conversation. *Memory* to keep track of recent user interactions. *Tools* to retrieve up-to-date or locally embedded information. *OpenAI GPT-4o-mini*(OpenAI et al., 2024b) for language processing and reasoning.

3.2.2 Class Structure

Below is a high-level summary of the class. It is defined as a *Pydantic model* (extending `BaseModel`) and uses a *singleton* approach to ensure there is only one active agent instance at a time:

```
class CampusManagementOpenAIToolsAgent(BaseModel):
    prompt: ChatPromptTemplate
    prompt_length: int
    language: Optional[str]
    llm: ChatOpenAI
    tools: List[BaseTool]
    memory: BaseMemory
    _agent_executor: AgentExecutor = PrivateAttr(default=None)
    ...
```

- `prompt`: A `ChatPromptTemplate` generated by the helper function `get_prompt`. This template includes a system message defining the chatbot's role and constraints (e.g., answer only university-related questions).
- `prompt_length`: An integer representing the approximate token length of the prompt. The chatbot uses this to estimate the remaining context space.
- `language`: Stores the language (German or English) in which the agent should respond, pulled from the global settings or the users preference.
- `llm`: An instance of `ChatOpenAI`, configured with streaming callbacks and a particular model name (e.g., GPT-4o-mini).
- `tools`: A list of `BaseTool` objects available to the agent.
- `memory`: A specialized *conversation memory* (in this case `CustomSaveMemory`) that remembers the last few turns of the conversation.

The `_agent_executor` attribute is where the agent's logic (the *chain of calls* involving the LLM and tools) is compiled by LangChain.

Initialization and Singleton Pattern

```
def __new__(cls, *args, **kwargs):
    # Ensures only one instance unless the language changes
    if cls._instance is None:
        cls._instance = super(CampusManagementOpenAIToolsAgent, cls).__new__(cls)
    elif hasattr(cls._instance, "language") and cls._instance.language != settings.language:
        cls._instance = super(CampusManagementOpenAIToolsAgent, cls).__new__(cls)
    return cls._instance
```

- **Singleton**: If an instance of `CampusManagementOpenAIToolsAgent` already exists, the code returns that instance instead of creating a new one.
- **Language Check**: The only exception occurs if the chatbot's language changes (German $\leftrightarrow$ English). In that situation, a new instance is created so that the prompt and memory can be reset.

Agent Executor Creation

After instantiation, `__init__()` calls a private method `_create_agent_executor()`:

```
def _create_agent_executor(self):
    llm_with_tools = self.llm.bind_functions([*self.tools])
    agent = (
        {
            "input": lambda x: x["input"],
            "chat_history": lambda x: x["chat_history"],
            "agent_scratchpad": lambda x:
                format_to_openai_function_messages(
```

```
                x["intermediate_steps"]
            ),
        }
        | self.prompt
        | llm_with_tools
        | parse
    )
    self._agent_executor = AgentExecutor(
        agent=agent,
        tools=self.tools,
        memory=self.memory,
        handle_parsing_errors=True,
        max_execution_time=60,
        ...
    )
```

- `bind_functions`: This instructs the LLM to handle `function_calls` for each tool, such as the web search tool.
- `agent` pipeline: A custom pipeline merges (a) the user input, (b) the conversation history, (c) the system prompt, and (d) the `parse` function that intercepts the LLM's JSON-based `function_call`.
- `AgentExecutor`: Orchestrates the final chain of reasoning, deciding whether to invoke a tool or provide a final user-facing answer.

Handling User Queries

When the application calls `CampusManagementOpenAIToolsAgent` like a function (i.e., `agent_executor("some question")`), the `__call__()` method triggers the underlying LangChain pipeline:

```
def __call__(self, input: str):
    config = {"callbacks": [CallbackHandlerStreaming()]}
    response = self._agent_executor.invoke({"input": input}, config=
        config)
    return response
```

- **Streaming Callback**: A custom `CallbackHandlerStreaming` intercepts the token by token output from the LLM, allowing partial responses to appear in real time.
- **Tool Invocation**: If the LLM's final message includes a `function_call` (e.g., `custom_university_web_search`), the agent automatically runs the appropriate tool with the JSON arguments parsed by `parse()`.
- **Return**: The final output is returned as a dictionary. The top-level key `output` is the user-facing text.

3.3 Supporting Components

Although `CampusManagementOpenAIToolsAgent` is the core orchestrator, it relies on several auxiliary modules:

1. **Prompt Templates** (`prompt.py`, `prompt_text.py`): Provide the system instructions that shape how the LLM responds (e.g., bilingual support, restricting answers to the Osnabrück University domain).

2. **Custom Web Search** (`search_web_tool.py`): Acts as a tool for the Agent and scrapes the university's website via Selenium. At the time of writing, the university website is mostly available in German. Thus, if the agent decides to use this tool, it is instructed to generate a *German* query for the university websites search algorithm. This ensures that the chatbot has access to the information available on the website. For more details on how the web search tool operates, see Section 3.3.1.

3. **Tool Helper** (`tool_helpers.py`): Via the class `VisitedLinks` the custom web search keeps track of and can list the visited links to the user.

4. **LLM Helper** (`agent_helpers.py`): Ensures a single shared `ChatOpenAI` object, specifying model name, temperature, token streaming, and optional caching.

5. **Configuration and Logging** (`core_config.py`, `chatbot_logger.py`): Centralize environment variables (e.g., model name, search URL, language) and define a logging system for debugging and data analysis.

3.3.1 The Web Search Tool

Although the *web search tool* was introduced briefly above, its functionality is central to askUOS's ability to retrieve information from the Osnabrück University's website and acts as the retrieval component in this RAG-based chatbot. The tool is defined in `search_web_tool.py` and leverages the following components:

1. **Selenium Setup:** The tool initializes a headless Firefox browser via the GeckoDriver (specified by `service` in `config.yaml`). When an agent query, already a German search query based on the user question, is passed to this tool:

    - The query is first *decoded* or reformatted into a German search string, by the helper function `decode_string` in `tool_helpers.py`.
    - The headless browser navigates to the on-site Google search page of Osnabrück University (`search_url`).

2. **Link Extraction:** After Selenium loads the search results, the HTML is parsed with BeautifulSoup to find anchor tags. The tool:

    - Collects up to `MAX_NUM_LINKS` from the search results.
    - Keeps track of visited links via a singleton `VisitedLinks` class, avoiding repeated scrapes of the same URLs.

3. **Asynchronous Fetching and Summarization:** For each link:

    - The code uses `aiohttp` to retrieve page contents (HTML or PDF) concurrently, accelerating the search process.
    - If the total combined text plus prompt tokens exceed the LLM's context window, the tool triggers an *internal summarization* using LangChain's `load_summarize_chain`. This process enables efficient handling of extensive text excerpts by summarizing the extracted websites individually, in reverse order, until the context window is no longer exceeded. The summarized content is then returned by the tool.

4. **Return to the Agent:** Once text from each selected link is fetched (and possibly summarized), the tool:
   - Joins the retrieved content into a single string.
   - Appends source information ("Information taken from: `<URL>`") within the text, allowing the chatbot to display references in the user interface.
   - Passes it back to the `CampusManagementOpenAIToolsAgent` for final reasoning and answer composition.

By combining Selenium-based scraping, asynchronous network requests, and dynamic summarization, the *web search tool* empowers the chatbot to provide up-to-date responses. While this approach hinges on the reliability and structure of the university's search results, it offers a flexible means for askUOS to retrieve live information that is not stored in the internal vector database.

3.4 Current askUOS Settings

In the present state of the code, the askUOS chatbot runs with the following core settings (primarily defined in `config.yaml` and loaded by `core_config.py`):

- **LLM Model:** `gpt-4o-mini` with a `context_window` of 20,000 tokens. This token limit allows the system to handle large volumes of text efficiently while balancing performance and cost. Although the OpenAI models support up to 128,000 tokens, this configuration avoids fully utilizing their maximum capacity to optimize resource usage.
- **Language:** By default set to `Deutsch`. However, the user can toggle between `English` and `Deutsch` via the radio button in the Streamlit UI (`language.py`). Once changed, the `settings.language` field updates, and the `CampusManagementOpenAIToolsAgent` is re-instantiated (due to its singleton logic checking for a different language).
- **Web Search Configuration:**
  - `search_url` is the Osnabrück University on-site Google search URL:

    `https://www.uni-osnabrueck.de/universitaet/organisation/zentrale-verwaltung/google-suche/?q=`
  - `MAX_NUM_LINKS` is set to *4*. This is rather arbitrary and could easily be changed. With this setting, the tool will scrape the first 4 sites appearing in the search results.

  This setup allows the chatbot to perform real-time queries on the university's website.
- **Memory Strategy:** The chatbot uses a `ConversationBufferWindowMemory` with `k=5`, retaining the last five message exchanges. Longer sessions may lose older context unless it is re-summarized (not yet implemented) or reintroduced by the user.

3.5 Flow of a Typical User Query

Putting all of the above together, a typical user query flows as follows:

1. **User Input**: The user types a question.
2. **Agent Invocation**: The application calls `CampusManagementOpenAIToolsAgent`.
3. **Prompt Construction**:
   - The agent merges the system message, chat history, and user question into a single prompt.
   - The agent also sets up a small internal pipeline so that any `function_call` from the agent is routed to the correct tool.
4. **Agent Reasoning**: The agent sees the text; if it needs external information, it calls `custom_university_web_search` with a query (in German as required by the system prompt).
5. **Web Search Tool**:
   - Uses Selenium to retrieve first four pages returned by the search query from Osnabrück University's website.
   - Summarizes them if the total tokens are over 20,000.
   - Returns the retrieved context to the agent.
6. **Answer Generation**:
   - The agent processes the returned text, possibly referencing it in the final answer.
   - Once it has enough information, it returns a JSON response with a `Response` function call, effectively finalizing the user-facing answer.
7. **Response**: The user sees the answer in the interface, optionally with links to pages visited during the search, as shown in Figure 3.2.

3.6 Summary

The `CampusManagementOpenAIToolsAgent` class lies at the center of this RAG-based chatbot design. It encapsulates:

- **Initialization** and **singleton logic**, ensuring consistent configuration.
- A **LangChain agent executor** pipeline, which seamlessly routes user input to the LLM and decides whether to invoke helpful tools.
- **Memory management**, allowing short-term conversation context to persist while avoiding excessive token usage.
- **Prompt management**, relying on system instructions that enforce domain constraints (only university-related topics) and bilingual capabilities.

By separating the chatbot's functionality (agent, tools, memory) from the front-end interface, the system remains flexible and scalable. Additional features, such as more robust caching, additional specialized tools, and local summarization models, can be integrated without disrupting the main agent logic.

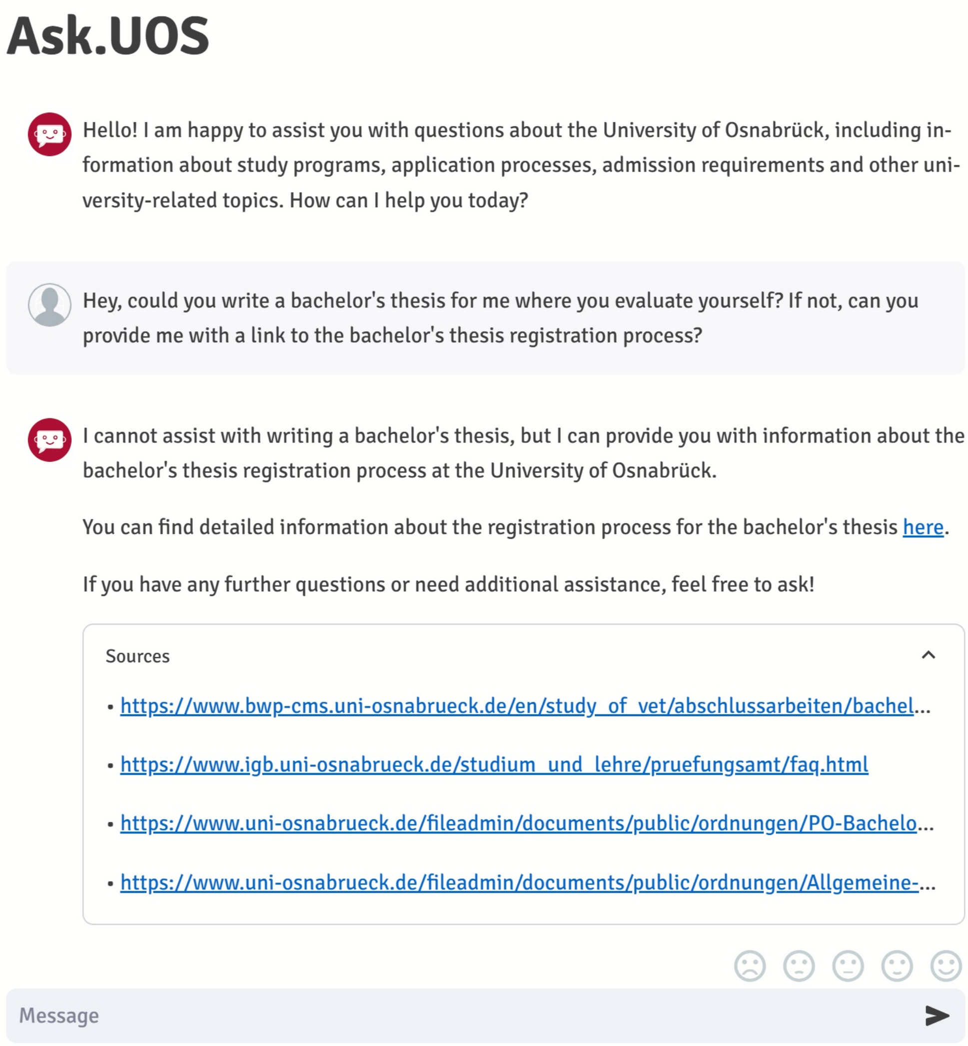


FIGURE 3.2: An example conversation with the chatbot askUOS.

This chapter detailed the architecture and implementation of the askUOS chatbot, emphasizing its RAG-based design and reliance on LangChain for managing queries and retrieving context. These elements underpin its performance and highlight the importance of a systematic evaluation approach to assess aspects like hallucination rates, answer accuracy, coherence, and document relevance, which will be explored in detail in the following chapter.

Chapter 4

Methods

As presented in Chapter 2, commonly used methods for evaluating LLM-powered chatbots, as the one described in Chapter 3, involve comparing system-generated text to a reference text. This chapter builds on that foundation by outlining the methods employed to evaluate the Osnabrück University chatbot. It emphasizes creating a robust dataset and systematically evaluating the chatbot's answers to genuine user questions.

This chapter ensures transparency, reproducibility, and ethical rigour. By leveraging authentic user-generated questions and carefully curated reference answers, this thesis establishes a domain-specific dataset that reflects the real-world information needs of the chatbot's target audience. The evaluation process provides insights into the chatbot's performance, including dimensions such as answer accuracy, hallucination rate, and document retrieval quality, while addressing the challenges and limitations discussed in the literature review.

4.1 Overview

Evaluating a domain-specific chatbot requires two key resources: (**1**) a dataset consisting of authentic user-generated questions that reflect the target audience's real information needs and high-quality reference answers aligned with Osnabrück University's official information sources and (**2**) the metrics to evaluate the chatbot answers, optionally against the reference answers.

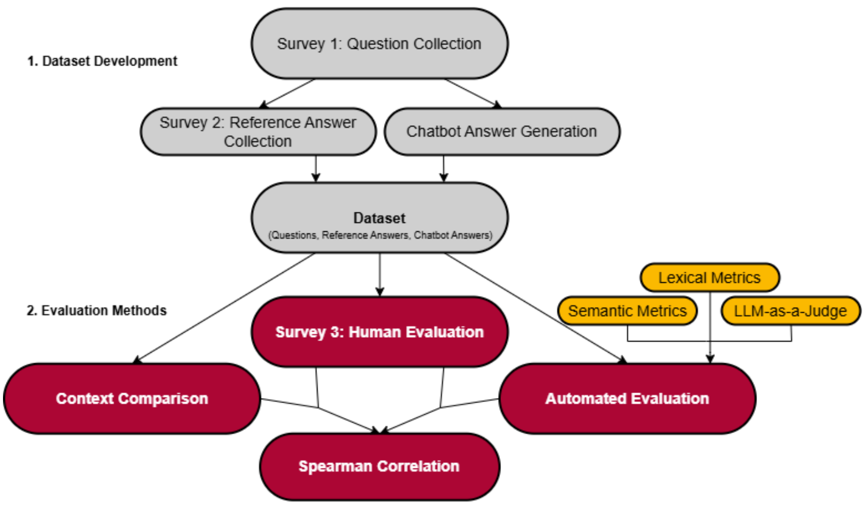


FIGURE 4.1: The outline of the methods chapter, starting with dataset development and ending with the correlation of the evaluation methods to the human gold standard.

To achieve this, this chapter begins by introducing the steps involved in gathering a domain-specific dataset. It then discusses the various evaluation methods employed and concludes by addressing the limitations and ethical standards. This flow is illustrated in Figure 4.1.

4.2 Dataset Creation

This thesis deliberately avoided using LLMs for data generation to minimize biases and hallucinations. Instead, human participants were engaged in the data-collection process to ensure authenticity and diversity in the dataset, to reflect genuine user inquiries.

Data collection was divided into two surveys—one for user questions and another for reference answers—to reduce participants' burden and improve data quality (Andreadis and Kartsounidou, 2020). This separation also helps minimize distortions, such as preventing participants from creating overly simple questions that are easy to answer in order to reduce their workload.

4.2.1 Survey 1: Question Generation

In the first stage, authentic user questions were collected. Insights from the chatbot's developer, Yesid Cano Castro, suggested focusing on prospective, enrolled, and international students as the main target user groups. While other groups may be relevant, these three were chosen as a reasonable starting point.

Initial Design and Adjustments

The initial survey design assigned participants hypothetical roles—prospective, enrolled, and international students —to encourage diverse question types. However, feedback from a pre-survey (see Appendix A.1) revealed that participants struggled to fully immerse themselves in these roles, reducing the authenticity of the questions generated. Additionally, generating multiple questions for each role led to survey fatigue.

Through collaboration with a local Gymnasium, namely the Besselgymnasium Minden[1], data collection became possible for prospective students, while enrolled and international students were reached through the university's communication channels managed by the researcher. Rather than hypothetical roles, participants identified their actual status (prospective student, enrolled student, international student, or other) and generated questions based on their real information needs.

Survey Structure and Bilingual Format

The survey, implemented in LimeSurvey, comprised three parts:

1. **Demographic Data Collection:** Gathered background information and categorized participants into target groups.
2. **Question Generation Task:** Participants generated questions for the chatbot based on a realistic scenario.
3. **Additional Information (Optional):** Participants could provide an email for subject hours, stored separately to maintain anonymity.

[1]See school Website: https://www.besselgymnasium.de/

The survey was offered in German and English to include both local and international participants, with carefully translated instructions to ensure consistency and clarity.

Refer to Appendix A.2.1 for the complete survey structure and participant instructions.

Survey Design and Implementation

The design balanced the need for diverse demographic data with a focus on realism in the question-generation task.

Demographic Data Collection While it would have been possible to inquire about a broader range of demographic factors, such as socioeconomic status, ethnicity, or race, as Appelbaum et al. (2018) suggests and Hughes et al. (2022) formulates, this thesis intentionally focused on a limited set of demographic questions. This choice aimed to maintain participants comfort and privacy, minimize survey length, and concentrate on variables relevant to participants engagement with the university. The survey collected the following demographic data:

- **Age Group:** Allows contextualizing participants life stages and proximity to academic decisions.
- **Gender:** Included to acknowledge diversity in the participant pool and offering a 'other, namely' option.
- **Highest Education Level:** Identifies the participant's academic background, potentially aiding in understanding their perspective on academic inquiries.
- **Current Educational Status:** Critical for mapping participants into the main user groups: prospective, enrolled, international students and others.
- **Current Study Program (Optional):** Provides extra context for enrolled participants.
- **Familiarity with Chatbots:** Could be used to analyze whether participants' prior experience influences the outcome of the evaluation.

All demographic questions offered an "I prefer not to answer" option to respect privacy and anonymity.

Question Generation Task After demographics, participants encountered the scenario shown in Figure 4.2.

By anchoring the task in a real browsing scenario, the survey encouraged practical, topic-specific queries rather than artificial ones. Asking for a minimum of seven questions ensures a rich dataset, while an upper limit of 15 prevents fatigue. Optional question fields beyond the seventh were revealed progressively, allowing more engaged participants to contribute additional questions without forcing all participants to provide more than seven.

Additional Information (Optional Email Address) An optional field was provided for participants to enter their university email if they wished to receive subject-hour credit. These email addresses were stored separately and later deleted, ensuring dataset anonymity while incentivizing participation from university groups.

Task Description of Survey 1

Scenario:
Imagine you are browsing the official Osnabrück University website to seek information about the university. On any page of this Website, you notice a chatbot displayed in a dedicated chat window with the message:
"How can I help you?"
This chatbot is specifically designed to assist users of the university website by providing helpful and accurate answers to questions about the university. It covers topics such as academic programs, admissions, campus services, student life, and other university-related information. Your task is to generate questions you might ask this chatbot based on your own information needs and interests.
Task:

- **Reflect on Your Needs:** Consider the specific information you might seek on the university website or topics that interest you.
- **Be Genuine:** Provide authentic questions that reflect your interests, concerns, curiosity or needs about the university.
- **Ensure Clarity:** Formulate your questions concisely and clearly, avoiding ambiguity.
- **Generate Questions:** Think about the questions you have or might have about Osnabrück University. Please provide at least seven independent questions (maximum 15) that you would ask the chatbot.

FIGURE 4.2: Instructions provided to participants for Survey 1: Question Generation Task

Participants

A total of 102 participants were recruited for the study, with 51 completing all survey sections (defined as reaching the final survey page). The following demographic data provides an overview of the participant pool of complete surveys:

The plotted participant demographic data can be seen in Appendix A.2.2.

Distribution: Most participants belonged to the 18–24 age group (n = 24), followed by 25–34 (n=14). A smaller subset included individuals under 18 (n=11), while one participant each belonged to the 35–44 age group or chose not to disclose their age.

Gender Distribution: The sample consisted predominantly of female participants (n=35), with fewer male participants (n=14). Additionally, one participant identified as "other," and one chose not to disclose their gender.

Education Level: Most participants had a university entrance qualification, such as a high school diploma or Abitur (n=26). Other qualifications included secondary school diplomas (n=10), bachelor's degrees (n=9), master's degrees (n=3), and vocational or technical training (n=2). One participant chose not to disclose their educational background.

Educational Role: The largest group of participants consisted of enrolled students (n = 30). Other categories included individuals identifying as "other" (n=8), prospective students (n=7), and international students (n=5). One participant chose not to disclose their educational status.

Chatbot Familiarity: Regarding familiarity with chatbots, 21 participants reported occasional use, while 12 indicated a neutral familiarity and 11 were very familiar with chatbots. Fewer participants reported limited experience (n=4) or no prior usage (n=2). One participant did not respond to this question.

This demographic breakdown highlights a diverse participant pool, encompassing a range of age groups, educational backgrounds, and levels of chatbot familiarity, which enhances the representativeness of the dataset.

4.2.2 Data Processing of the Collected Questions

Before integrating the collected user questions into the second survey, aimed at gathering reference answers, several data preparation steps were performed to ensure usability, balance, and integrity of the dataset. The relevant steps are discussed here. The full data processing can be seen in the `Thesis GitHub/code/dataset` Notebooks marked 0–3.

Role Mapping and Standardization

Participants' reported educational statuses were mapped into a simplified `role` column ('prospective', 'enrolled', 'international', 'other'). Anomalies, such as one participant's survey data that reported an English survey interface but had German questions, were corrected.

Question Quality Control and Filtering

While processing the question data, additional considerations were made to preserve the authenticity and practicality of the dataset. While the researcher considered correcting spelling and grammatical errors, such interventions were ultimately avoided. Retaining such imperfections ensures that the resulting dataset remains closer to naturally occurring user input, reflecting the types of queries an actual chatbot might encounter.

Additionally, questions containing placeholder variables (e.g., *"When can I enroll for [STUDY COURSE]?"*) and invalid questions (e.g., *"No further question "*) were manually excluded. Although placeholder-based queries have the potential to expand the range of questions in future research, where placeholders can be programmatically replaced with valid university-related terms, they do not reflect naturally occurring user inquiries in their current form. Furthermore, one participant (ID 98) who submitted completely irrelevant content and whose demographic data was logically inconsistent was excluded. Removing these questions helped maintain a dataset closely aligned with authentic user inquiries without altering them.

In contrast, user queries that do not relate directly to the university or would need some clarification to be answered directly (e.g., *"What career opportunities are there after completing my program?"*), were deliberately kept. These questions are valuable for evaluating the chatbot's ability to handle these inquiries, either by politely refusing to answer or by asking the user more questions to specify. Including these queries ensures that the dataset tests not only straightforward questions but also whether the chatbot complies with its system prompt.

Language Alignment and Translation

While the first survey was bilingual, the subsequent answer-generation survey was administered exclusively in German to align with the university's main language resources. Any English questions from Survey 1 were automatically translated into German and vice versa, following careful instructions to maintain accuracy and consistency, and then verified by the researcher. This resulted in a complete bilingual question dataset.

Balancing and Randomization by Role

To ensure a fair and representative selection of questions in the next survey, the final set of cleaned and translated questions was randomized and interleaved by role. Whenever possible, a prospective question row was followed by an enrolled and then international role question.

Automated Survey Generation

Using the below described Survey 2 as a template, the dataset was transformed into surveys with different questions suitable for publication in LimeSurvey. The cleaned and role-balanced questions were arranged into sets of three, ensuring inclusion of one question from each participant role (prospective, enrolled, international) whenever enough questions of that group were available. This balanced approach enabled fair representation and diversity in the questions presented to answer providers.

A Python script automated survey creation by embedding questions into placeholders within a predefined LimeSurvey template. For each batch of three questions, placeholders were replaced with corresponding question texts and their unique question ID. Additionally, survey titles were dynamically updated with a sequential numbering system to facilitate organization and identification.

The final surveys were saved as individual `.lss` files, ready for direct upload to LimeSurvey. This workflow streamlined the survey creation process, ensuring consistency, accuracy, and alignment with the requirements of the subsequent Survey 2.

In summary, the data processing phase transformed a raw, mixed multilingual, wide-format dataset of varying quality into a clean, balanced, and role-distributed bilingual corpus of questions and generated multiple surveys based on the placeholder survey described below.

4.2.3 Survey 2: Reference Answer Creation

Survey 2 was designed to create a dataset of reference answers to evaluate the chatbot's answers against. This survey targeted participants familiar with the university's website and information systems. Participants were tasked to answer questions, ensuring that the responses were accurate, helpful, and based on the latest information available on the university website.

Refer to Appendix A.3.1 for the complete survey structure and participant instructions.

Survey Structure and Monolingual Format

The survey, implemented in LimeSurvey, comprised three parts:

1. **Demographic Data Collection:** Gathered background information.
2. **Question Generation Task:** Participants generated answers to questions gathered from Survey 1.
3. **Additional Information (Optional):** Participants could provide an email for subject hours or entering a lottery, stored separately to maintain anonymity.

Contrary to Survey 1, Survey 2 was only available in German. With information obtained from Osnabrück University's Communication and Marketing team[2], the website offers roughly 1,000 active German pages compared to 300 in English, making German the resource-rich environment (as of November 22, 2024). This alignment allowed answer providers to use the most comprehensive and current source materials, consistent with the chatbot's retrieval process described in Section 3.3.1.

Survey Design and Implementation

Demographic Questions The demographic section was identical to that of Survey 1, collecting basic information such as age, gender, educational background, and familiarity with chatbots. Because the design and rationale for these questions remained unchanged, they are not further discussed here.

Answer Generation Task Participants were presented with detailed instructions with a focus on providing accurate, detailed, and structured answers as if they were responding on behalf of the chatbot. The instructions given to participants emphasized several critical aspects guided by the system prompt of the chatbot (See Appendix Figure B.1) so ensure a fair comparison.

To clarify expectations, the researcher provided a hand-crafted example question and answer in the instructions. Participants could revisit the task description at any point during the survey via a collapsible text field. The full task description text, translated into English, can be seen in the Appendix A.7.

After the instruction, each question was displayed individually, accompanied by a direct link to the university's search page to aid the participants in locating relevant information.

Furthermore, the participants were instructed to enter the URLs used to generate the answer directly under the provided answer.

Recruitment and Participation Incentives

University related channels were used to gather participants. Interested participants received a unique link and token to one of the surveys. This process ensured that each question received one reference answer.

Initially, participation rates were below expectations. To address this, a lottery offering four digital vouchers worth 15€ each was introduced one week after the launch of the survey. The lottery was open to participants aged 18 or older and served as an additional incentive alongside the optional credit of 0.5 participant hours.

Participants

A total of 13 participants were recruited for this survey, with 11 completing all survey sections. The following demographic data summarizes this participant group:

The plotted participant demographic data can be seen in Appendix A.3.2.

Age Distribution: Most participants were aged 18–24 (n=10), with one participant in the 25–34 age group.

Gender Distribution: The sample included six female participants and five male participants.

[2]See communication and marketing website: `https://www.uni-osnabrueck.de/kommunikation/kommunikation-und-marketing-angebot-und-aufgaben/leitung-und-team/`

Education Level: The majority of participants had a university entrance qualification, such as a high school diploma or Abitur (n=9), while two participants held a bachelor's degree or equivalent.

Educational Status: All participants were enrolled students (n=11).

Chatbot Familiarity: Participants were evenly split between those occasionally using chatbots (n=5) and those very familiar with chatbots (n=5), with one participant indicating a neutral familiarity level.

This participant group represents a smaller, more targeted sample, reflecting individuals with a higher likelihood of engaging with the chatbot system due to their familiarity and educational context.

4.2.4 Data Processing of the Collected Answers

The processing of answers collected from Survey 2 shared many steps with the question data processing outlined in Section 4.2.2. For specific details, refer to `Thesis GitHub/code/dataset` Notebooks marked 4–5. Below, only the additional or distinct processing steps are summarized.

Language Alignment and Translation

To enable insights into potential bilingual differences in chatbot performance, all German answers obtained from Survey 2 were translated into English. This step aimed to facilitate a direct comparison between the chatbot's answer generation capabilities in German and English. While a dedicated survey to collect human-written English answers would have been ideal, this approach was deemed necessary due to time constraints, the limited availability of English-language content on the university's website, and the need for a fair comparison of the chatbot's and human participants' access to domain-specific information.

Automated tools were used to perform the translations, following explicit instructions to ensure consistency and accuracy. The researcher subsequently verified the translations to ensure alignment with the original German answers and to mitigate the risk of misinterpretation. Although machine translation has inherent limitations, such as potential loss of nuance or contextual shifts, this process was deemed sufficient for the scope of this thesis.

By incorporating both German and English versions of the answers, the dataset allows for a robust evaluation of the chatbot's bilingual performance. This dual-language dataset also provides a unique opportunity to explore the impact of linguistic differences on evaluation metrics, contributing to a broader understanding of the chatbot's effectiveness in multilingual settings.

4.2.5 Chatbot Answer Generation

The final stage of dataset creation involved systematically gathering the chatbot's actual answers, in both German and English, for the questions in the dataset. This process ensured that each user-generated query with a reference answer was answered under controlled and reproducible conditions, mirroring how a real user would interact with the chatbot. All steps discussed below are implemented in Notebook 6 in the `Thesis GitHub/code/dataset`.

Agent Setup and Warm-Up A dedicated instance of the `CampusManagementOpenAIToolsAgent` class (see Section 3.2) was initialized twice, once with the language set to *German* and another with the language set to *English*. Before generating real answers, each agent was given a small set of 'warm-up' queries to handle model initialization overhead in advance. This step mitigated delays or inconsistencies that might arise from on-demand model loading during the actual generations.

Memory Reset for Each Query Each agent's conversation *memory* was cleared before every query. This ensured that no context from previous queries could carry over and influence subsequent answers. By resetting the memory to the baseline state, each question received an independently computed response, thereby avoiding cross-question leakage.

Generating and Storing Answers After preparing the agents, each user question from the dataset was passed to either the German or English agent, depending on the question language. The corresponding chatbot answers were recorded in a CSV file. For each row (i.e., each question-answer pair), the pipeline stored:

- The **chatbot's final text** output in German or English.
- The **visited links** reported by the chatbot's `visited_links()` method, indicating which URLs were used (or at least claimed to have been used) during the answer generation process.

In practice, there were rare instances in which the chatbot returned a link not included in the actual context retrieval step. To address this, a separate `used_context` and `visited_links` log was maintained in the final dataset so that discrepancies could be identified and filtered out as needed.

Collecting Metadata from LangSmith Once the answers were generated, additional information about each query's *chain of reasoning* and *tool usage* was retrieved from the external logging platform LangSmith[3]. By referencing a unique session start time, all root runs (i.e., top-level chatbot queries) after that timestamp were pulled into a metadata DataFrame. This dataset clarified precisely:

- Which questions triggered a **web search tool** call.
- The **latency** and **start/end time** of each query.
- The actual **content** the web search tool retrieved when invoked by the agent.

This step helped ensure that the final evaluation set included not only the chatbot's textual outputs but also the actual context and links the agent leveraged for each answer.

Final Merging and Export Lastly, the question-answer CSV and the LangSmith-derived metadata CSV were merged on the unique question text (both German and English). This created a comprehensive dataset containing the following:

1. The question in German and English.
2. The chatbot's corresponding answer in each language.

[3]See LangSmith Website: https://www.langchain.com/langsmith

3. The visited links reported by the chatbot.
4. The actual search context used by the agent.
5. Timestamps, latencies, and tool usage indicators for further analysis.

By combining the user queries, chatbot outputs, and precise metadata on how each response was generated, subsequent analyses could accurately pinpoint any gaps in the chatbot's performance or deviations from the official sources.

4.2.6 Dataset Merging

To create a unified dataset for evaluating the chatbot's performance, the three distinct datasets—questions from Survey 1, answers from Survey 2, and the chatbot-generated answers—were merged into a single comprehensive dataset. This merging process ensures that each question is paired with both the chatbot's and human-provided answers for direct comparison, in addition the the participants' demographic data for potential future analysis.

Two versions of the merged dataset were created:

- A comprehensive dataset containing all columns and metadata from the two surveys and questions, human answers and chatbot answers both in English and German.
- A condensed dataset with only the essential columns for evaluation, including the question text, the original question language, the chatbot answer, the human answer, associated source links for the chatbot and humans, and the corresponding question and participant IDs. Once for English and once for German entries.

The exact implementation can be found in `Thesis GitHub/code/dataset` Notebook 7. By maintaining both comprehensive and condensed versions, the dataset is flexible for the various analytical needs, described in the next section.

4.3 Evaluation Metrics and Tools

At this stage, all user questions have been gathered (Survey 1), carefully sourced reference answers have been collected (Survey 2), and chatbot answers have been generated with accompanying metadata (Chatbot Answer Generation). Thus, the final dataset allows the chatbot's responses to be evaluated and compared against human-provided references.

The structure of this section first presents the human evaluation, serving as a baseline to compare other metrics against. Subsequently, it moves to the various automated metrics. Afterward, a comparison of referenced links (human vs. chatbot) is discussed, followed by a meta-level correlation analysis of automatic vs. human evaluation.

4.3.1 Survey 3: Chatbot Answer Evaluation

Building on the previous Survey 1 and the generated chatbot answers, the third survey was designed to collect human evaluations of the chatbot's answers. This survey asked participants to evaluate both the quality of the chatbot's responses and the context used by the chatbot. The resulting human evaluations served as an authoritative baseline for assessing hallucination rates, answer accuracy, user satisfaction, coherence/clarity/fluency, and the quality of the provided context.

Refer to Appendix A.4.1 for the full survey structure and participant instructions.

Survey Structure and Bilingual Format

This evaluation survey was conducted in German and English to ensure that participants could assess the QA-pairs generated in both languages. To accommodate participants' linguistic preferences and the bilingual nature of the dataset, the instructions, rating dimensions and rating scales were presented in both languages. Upon starting the survey, participants were asked to confirm that they could understand German and English, ensuring that they could fully comprehend both the question text and the chatbot's answer, which might be in German or English.

Demographic Data The demographic questions were identical to Surveys 1 and 2.

Survey Design and Implementation

Survey 3 was divided into three main sections:

1. **Demographics and Consent:** Collected participant demographics and asked for consent to proceed. Participants who were not proficient in *both* German and English were screened out via a quota.

2. **Evaluation of Chatbot Answers:** This core section displayed six to seven randomly selected QA-pairs, accompanied by the links the chatbot used. Each pair was rated according to five dimensions described below, an overall impression dimension and concluded with an optional open-text field for additional remarks.

3. **Incentives:** The survey ended with an option to receive 0.5 participation hours or enter a voucher lottery, similar to Survey 2.

By randomizing which QA-pairs participants saw and balancing the display of German and English items, the survey minimized fatigue and maintained coverage of the entire dataset. This approach also helped mitigate potential order effects since each participant's rating sequence was unique.

Rating Dimensions

Drawing on the literature review in Chapter 2 and the goal of the thesis in mind, five evaluation dimensions were identified. While some prior research recommends more fine-grained categorizations (Krishna et al., 2023; Mishra et al., 2024), the goal here was to keep the survey feasible for non-expert annotators while still collecting information highly relevant to the research questions:

1. **Hallucination**
   Definition: Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.

   This single dimension covers both *factuality* (correctness of real-world or well-known facts) and *faithfulness* (alignment with the provided context).

Motivation: Hallucination is central to the thesis. Although more fine-grained classifications (e.g., Li et al., 2024; Mishra et al., 2024) offer deeper diagnostic insights, they can overwhelm average participants and require extensive training. A single broad hallucination dimension thus balances depth and survey feasibility.

2. **Answer Accuracy**
   Definition: The degree to which an answer accurately addresses the user's question by providing correct, complete, and relevant information that matches the intent of the question. Factual accuracy (no hallucination) is necessary, but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.

   This category merges both *completeness* and *relevance*, following the literature's emphasis on ensuring that the chatbot directly addresses the user's question with *correct*, *complete* and *relevant* information (Chang et al., 2023, Section 5.2).
   Motivation: While strongly tied to hallucination, answer accuracy also includes aspects of whether the chatbot *gets* the question intent, covers all important points, and remains relevant to what the user asked.

3. **User Satisfaction**
   Definition: Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.

   Reflects a more subjective dimension: How satisfied the participant, as a hypothetical user, feels about the answer. Although it can overlap with *answer accuracy*, user satisfaction can also capture the overall helpfulness or personal preference factors.
   Motivation: While not commonly evaluated this is an interesting dimension to see if there is a difference in the participants evaluation of accuracy vs the perceived helpfulness of the chatbot. Keeping a separate user satisfaction category helps disentangle subjective impressions from purely factual considerations.

4. **Coherence, Clarity, and Fluency**
   Definition: Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.

   Evaluates how well structured, grammatically correct, and comprehensible the answer is.
   Motivation: While modern large language models rarely struggle with basic fluency, the dimension remains relevant for ensuring the text is not overly confusing or disjointed.

5. **Context Quality**
   Definition: Assesses the relevance and completeness of the context in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response.

   Assesses how relevant, sufficient, or well-targeted the linked sources are for generating the answer. Participants were shown the list of URLs that the chatbot used and were asked to judge whether these links would support a good answer. If no context was provided, participants rated whether that omission

negatively impacted the answer quality.
Motivation: The thesis heavily focuses on a RAG chatbot, thus the context quality is a major interest. The researcher chose not to provide the raw text that the chatbot actually used as context, as it would have been too lengthy, unclear, and burdensome for participants. Instead, the researcher displayed the website links, allowing participants to click on the URLs and assess their relevance and sufficiency in addressing the question in a more human-friendly manner.

For each dimension, participants chose a rating on a 5-point Likert scale (1 = Very Bad, 5 = Very Good). If the chatbot declined to answer or no context was supplied, participants were asked to interpret how that impacted each dimension. The definitions and guidance for rating on the Likert scale can be found in the task description in Appendix A.4.1, pages 7–8.

Implementation Details

Similar to Surveys 1 and 2, Survey 3 was developed in LimeSurvey. Key implementation notes include:

- **Randomization and Quotas:** The built-in `Relevance Equation` feature was used to allocate users into one of 10 groups with six to seven QA-pairs from the total dataset displayed in a random order. A hidden equation question controlled random group assignment to ensure coverage across participants.
- **Bilingual Question–Answer Presentation:** Each item displayed the question in the language it was asked (German or English) and the chatbot's corresponding answer.
- **Context Links:** Under the heading "Context used" a hyperlinked list of the URLs was shown. Participants were encouraged to click on them and scan the content. They were also instructed not to rely on outside websites beyond these URLs.
- **Optional Free-Text Comment:** For each QA-pair, participants could leave a short comment. This was not mandatory, but encouraged participants to note any additional observations they made during the evaluation.

Aimed Participation Goal

The survey targeted a minimum of three evaluations per question-answer pair in order to achieve robust results. This threshold aligns with the literature on human evaluations of LLM-generated text (Chiang and Lee, 2023; Goyal, Li, and Durrett, 2023; Fabbri et al., 2021; Gao and Wan, 2022).

According to Ferrante and Saltalamacchia (2006) formula for *The Coupon Collector's Problem with Multiple Collections with Equal Probabilities*, the survey would need 60 participants to ensure three evaluators per question-answer pair with random allocation. To address potential imbalances in group assignments, the researcher could adjust the random allocation if any group fell short of the target number of three evaluators. Consequently, the target participation number was set at 60, with a minimum threshold of 30.

The approach of one large survey with different random groups was used because of the poor participation results using invite links and tokens from Survey 2. The approach from Survey 1, where participants could click on a link and start without first obtaining an invite link from the researcher, yielded more participants. Thus, the same approach was used here, with the slight limitation that the researcher had to monitor group allocations.

Participants

A total of 69 participants were recruited for the third survey, with 30 completing all survey sections. The following demographic data provides an overview of this participant group:

The plotted participant demographic data can be seen in Appendix A.4.2.

Age Distribution: The majority of participants were aged 18–24 (n=19), followed by 25–34 (n=10), with one participant in the 55–64 age group.

Gender Distribution: The sample consisted of 18 female participants and 12 male participants.

Education Level: The majority of participants possessed a university entrance qualification (n=18). Other qualifications included bachelor's degrees (n=8), vocational or technical training (n=2), master's degrees (n=1), and secondary school diplomas (n=1).

Educational Status: The largest group of participants were enrolled students (n=22). Additional categories included individuals identifying as "other" (n=4), international students (n=2), one prospective student, and one participant who chose not to disclose their educational status.

Chatbot Familiarity: Participants documented varying levels of familiarity with chatbots: 15 used chatbots occasionally, 6 had neutral familiarity, 4 were very familiar, 3 reported limited experience, and 2 had no prior chatbot experience.

Survey 3 yields a valuable *human-evaluation* baseline across multiple quality dimensions, the minimum participation number was achieved. Its outcomes, combined with the automatic metrics presented in later sections of this chapter, allow for a rich comparison of how well the chatbot performs in real-world settings.

4.3.2 Rationale for Selecting the Evaluation Metrics

The selection of BLEU, ROUGE, BERTScore, BARTScore, BLEURT, and the LLM-as-a-Judge approach in this thesis reflects both practical and conceptual considerations:

- **Common Metrics in Literature**

  BLEU and ROUGE are among the most widely adopted lexical metrics for comparing system outputs to reference texts. Though limited in detecting factuality or hallucinations, they help situate the results relative to established baselines. BERTScore, BARTScore, and BLEURT extend this analysis by focusing on semantic alignment and fluency, offering deeper insight into whether the chatbot's paraphrased or differently structured answers still match the reference answer's meaning.

- **Time and Resource Constraints**

  While QA-based evaluations and specialized RAG metrics (see Sections 2.4.1 & 2.4.2) could yield more detailed analyses of factuality and faithfulness, implementing such frameworks often require constructing additional annotated data and designing complex validation sets (Saad-Falcon et al., 2024; Es et al., 2023). Implementing such frameworks for all metrics in a single, domain-specific study exceeded the time and resource scope of this thesis.

- **Internal State Inaccessibility**

  Methods that inspect the model's internal probability distributions (e.g., analyzing token log-probabilities) were infeasible given the use of a closed-source API in the chatbot (`chatgpt4o-mini`). Without direct access to model internals, techniques like conditional entropy or probability-based hallucination detection could not be applied.

- **RAG-Specific Frameworks Outside Project Scope**

  Tools such as RAGProbe (Sivasothy et al., 2024) and RAGEval (Zhu et al., 2024) focus on scenario-specific retrieval pipeline failures or synthetic data generation, which diverge from this study's narrower examination of genuine single-turn user queries. RAGEval, for instance, targets broad domains like finance or law and relies on a more elaborate schema-based pipeline, whereas this thesis centers on real queries from Osnabrück University. Adapting these frameworks would add substantial methodological complexity without necessarily yielding additional insights under the present constraints.

Therefore, limiting the set of metrics enabled a focused investigation into how well general-purpose approaches (lexical overlap, semantic similarity, LLM-as-a-Judge) align with human evaluations in a specialized, university-specific domain. This streamlined evaluation strategy allowed for a practical demonstration of metric performance without exceeding the thesis's timeline or resources.

The following implementations can be found in the `Thesis GitHub/code/eval` folder.

4.3.3 Lexical Metrics

This thesis utilized two well-established metrics, BLEU and ROUGE, to evaluate the lexical similarity between the chatbot's generated text and human-provided reference answers. Although originally designed for tasks such as machine translation and summarization, their widespread adoption and straightforward interpretability make both metrics useful for measuring surface-level overlap in question answering contexts.

BLEU

BLEU (Papineni et al., 2002) measures how closely a machine-generated text matches one or more human reference texts by calculating n-gram precision and applying a brevity penalty to discourage overly short outputs. In this thesis, *sentence-level* BLEU scores were computed to capture the performance of the individual chatbot answers.

Implementation To ensure methodological consistency and reproducibility, this thesis adopted the `SacreBLEU` library (Post, 2018), which addresses various tokenization and smoothing inconsistencies common in earlier BLEU implementations. The steps include:

1. **Sentence-Level Computation:** For each answer–reference pair, `SacreBLEU` calculates BLEU scores, including 1- to 4-gram precision, brevity penalty, and other summary statistics.

2. **Configuration Documentation:** The SacreBLEU signature and settings were recorded to ensure full transparency:

```
nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.4.3
```

ROUGE

Initially developed for summarization tasks, ROUGE (Lin, 2004) has become one of the most widely used metrics in NLP (Gehrmann, Clark, and Sellam, 2022; Grusky, 2023). It examines overlap between generated text and reference text at various granularities, emphasizing *recall* of n-grams, subsequences, or skip-bigrams.

ROUGE Variants This thesis focuses on the most promising ROUGE variants, taking inspiration from the discussion by Gehrmann, Clark, and Sellam (2022), which highlights that while ROUGE-1, ROUGE-2, and ROUGE-L are the most commonly reported, they are not necessarily the most effective variants. Thus the following variants were calculated:

- **ROUGE-1 to ROUGE-4** (unigram overlap to 4-word sequence overlap)
- **ROUGE-L** (longest common subsequence)
- **ROUGE-SU4** (skip-bigrams, allowing a skip of up to 4 words)
- **ROUGE-W-1.2** (weighted longest common subsequence)

Implementation and Reproducibility In line with the reproducibility concerns expressed by Grusky (2023), this thesis employed SacreROUGE (Deutsch and Roth, 2020), the only verified python implementation that matches the original ROUGE definitions. The following points highlight this setup:

- No stemming or stopword removal was applied, as ROUGE's default WordNet based stemming is not readily available for German.
- Each answer pair (chatbot vs. human) was evaluated with ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-SU4, and ROUGE-W-1.2. Resulting in the following metric initialization:

```
rouge_metric = Rouge(
    max_ngram=4,
    use_porter_stemmer=False,
    remove_stopwords=False,
    compute_rouge_l=True,
    skip_bigram_gap_length=4,
    wlcs_weight=1.2
)
```

In summary, BLEU and ROUGE offer complementary insights into *how literally* the chatbot's answers match reference texts. Although these metrics are efficient for large-scale lexical evaluations, they are augmented in this thesis by semantic metrics that better capture paraphrasing, factual correctness, and deeper contextual alignment.

4.3.4 Semantic Metrics

While BLEU and ROUGE focus on surface-level n-gram overlap, a more nuanced assessment of answer quality requires capturing deeper semantic alignment, contextual correctness, and generative fluency. To this end, this thesis employs three semantic-oriented metrics, BERTScore, BARTScore, and BLEURT, each offering complementary insights into the chatbot's performance beyond strict lexical similarity.

BERTScore

BERTScore (Zhang et al., 2020) computes semantic similarity between candidate and reference texts by comparing contextual embeddings from a pretrained transformer (e.g., BERT). Unlike purely lexical metrics, BERTScore is robust to paraphrasing and reordering, making it particularly suitable for *question answering* scenarios in which the chatbot may express correct content in varied forms.

Implementation This thesis followed the official `bert-score` GitHub implementation (Zhang et al., 2020). Key steps include:

1. **Model Selection:** `bert-base-multilingual-cased` for German and `roberta-large` for English, following the authors recommendations.

2. **Metric Computation:** BERTScore calculated precision, recall, and F1 at the token level per QA-pair.

3. **Hash Codes:** For reproducibility, the BERTScore signatures were documented:

   - **English**:

     ```
     roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.47.1)
     ```

   - **German**:

     ```
     bert-base-multilingual-cased_L9_no-idf_version=0.3.12(
         hug_trans=4.47.1)
     ```

BARTScore

BARTScore (Yuan, Neubig, and Liu, 2021) adopts a *generative* perspective: it measures the likelihood of producing one text from another using transformer models like BART. Beyond lexical overlap, this approach evaluates fluency, coherence, and partial factual alignment by comparing forward and backward generation probabilities.

Implementation To ensure reproducibility and adherence to the original methodology, this thesis employs two versions of BARTScore: the standard BARTScore implementation used by Yuan, Neubig, and Liu (2021) and a multilingual version. The implementation details are outlined as follows:

1. **Model Selection:** Two pre-trained models were utilized: the `facebook/bart-large-cnn` checkpoint, and the multilingual model `facebook/mbart-large-50-many-to-many-mmt`. Both models were used to evaluate English and German texts, allowing a direct comparison of their effectiveness in multilingual evaluation.

2. **Score Computation:** BARTScore's *F-Score* was computed for each chatbot-reference answer pair. Because the original paper by Yuan, Neubig, and Liu (2021) did not explicitly state whether the harmonic or arithmetic mean was used, both calculations were performed in this thesis to ensure a comprehensive approach. The harmonic mean accounts for the balance between precision (from reference to chatbot-generated text) and recall (from chatbot-generated text to reference), whereas the arithmetic mean provides a simpler averaging method. The results for both means were recorded.

3. **Reproducibility:** The full implementation can be found in Thesis GitHub/-code/eval/BARTscore_eval.ipynb.

Possibility for Faithfulness Evaluation BARTScore has the potential to assess the faithfulness of generated text by calculating its score relative to the source text (e.g., the content retrieved from cited URLs). However, this application was deemed infeasible in this study due to practical constraints. The source text, often significantly longer than chatbot-generated answers, would exceed the token limit of 1024 tokens, rendering the faithfulness evaluation unreliable.

BLEURT

BLEURT (Sellam, Das, and Parikh, 2020) is a learned metric that uses a transformer model fine-tuned on human ratings, focusing on semantic adequacy and fluency. Unlike overlap-based scores, BLEURT captures subtle differences by penalizing dissimilar outputs while rewarding paraphrases that preserve the reference's meaning.

Implementation Following the official BLEURT library (Sellam, Das, and Parikh, 2020), the workflow includes:

1. **Checkpoint:** `BLEURT-20` for sentence-level scoring, tested on both English and German subsets.

2. **Pairwise Comparison:** Each chatbot answer was paired with its reference text, yielding a numeric score that typically ranges between 0 and 1.

3. **Reproducibility:** The full implementation can be found in Thesis GitHub/-code/eval/BLEURT_eval.ipynb.

Interpretation and Language Coverage BLEURT primarily assesses how well a generated answer expresses the same meaning as the reference, with an inherent tilt toward measuring fluency. While trained mostly for English, BLEURT-20 supports a limited set of additional languages (including German).

Synthesis of Semantic Metrics

Taken together, BERTScore, BARTScore, and BLEURT add robust *semantic* insights to the evaluation, complementing the lexical perspectives of BLEU and ROUGE. While each metric has distinct strengths—embedding-based alignment (BERTScore), generative likelihood (BARTScore), and learned adequacy/fluency (BLEURT)—they converge on a shared goal: identifying how *meaningful*, *coherent*, and *faithful* the chatbot's answers are relative to human references.

4.3.5 LLM-as-a-Judge

In addition to relying on human evaluators and the automated metrics discussed above, this thesis also employed an LLM to automatically evaluate the chatbot's answers on the same five dimensions used by the participants of Survey 3 (see Section 4.3.1). Using these *identical definitions*, the 'LLM-as-a-Judge' framework enables a direct comparison between human and model-based evaluations. This approach extends upon prior work from Liu et al. (2023b), Zheng et al. (2023), Chiang and Lee (2023), Wang et al. (2023b), and Fu et al. (2023).

Approach and Dimensions

Rather than reintroducing the five evaluation dimensions, the LLM-based method adopts the same operational definitions and rating guidelines. This alignment ensures that automated scores can be compared against the human scores on a one-to-one basis, facilitating an assessment of how reliably an LLM can replicate human judgments.

Each dimension was scored on a 0–4 scale (0 = Very Bad, 4 = Very Good), with short justifications for each score. Leng, Uhlenhuth, and Polyzotis (2023) blog post influenced the decision to use a 0-4 scale by demonstrating that low-precision scales like 0-3 or 0-4 improve grading consistency, explainability, and alignment with human evaluations in LLM-based assessments. Additionally, this range can be easily compared against the collected 1–5 Likert scale from human evaluations.

For certain evaluations, reference answers from Survey 2 were included to determine whether their presence improves alignment between the LLM's and human evaluations.

Configurations and Reference-Answer Options

To explore whether providing the human *gold standard* reference answer helps the LLM to evaluate the answers, the thesis tested four configurations on both languages. All configurations were run using the same model version, `gpt-4o-2024-08-06`:

1. **Together, no reference:** A single API call per QA-pair, evaluating all five dimensions simultaneously, with no reference answer provided.

2. **Together, with reference:** A single API call per QA-pair, but this time including the human-written reference answer in the prompt.

3. **Separate, no reference:** Five API calls per QA-pair (one for each dimension), without a reference answer.

4. **Separate, with reference:** Five API calls per QA-pair, providing a reference answer in each dimension-specific prompt.

In the *together* configurations, the LLM receives a single prompt with instructions to rate all five dimensions simultaneously, whereas in the *separate* configurations, each dimension is prompted individually for finer control and potentially more focused explanations. The system prompts used are presented in Appendix B.2

Implementation Details and Prompting

The evaluations leverage the OpenAI API's *structured output* guidelines[4] to keep the returned scores consistent across dimensions. Incorporating delimiters to distinctly highlight specific sections of the input and using chain of thought to enhance results were used [5]. Each QA-pair with its context was injected into the prompt, followed by optional reference answer if the configuration required it. The LLM's textual justification was then stored along with the numeric scores.

One difference between this method and the human evaluation lies in the context provided to the evaluator. In the human evaluation, the researcher chose to show the links used by the chatbot. By contrast, this method provides the evaluator with the exact context that the `custom_university_web_search` tool returned to the agent during the answering process.

Overall Score

To ensure simplicity and uniform evaluation across all dimensions, an equal-weight scoring system was implemented. This approach reflects a balanced perspective, emphasizing the equal importance of all dimensions in assessing the chatbot's performance. The overall score was calculated as the average of the individual scores.

In spite of the challenges discussed in Section 4.4.3, the *LLM-as-a-Judge* method offers a flexible, explainable, and domain-agnostic approach to automated evaluation. Aligning its dimensions and definitions with Survey 3's human-based approach allowed this thesis to directly measure whether an LLM can serve as a credible stand-in for human evaluators.

4.3.6 Context Comparison

A key objective of this thesis is to assess the chatbot's retrieval strategy, partly done in the Survey 3 and LLM-as-a-Judge, additionally one more metric is calculated: whether the website links the chatbot uses to answer questions match the links chosen by humans. To this end, the merged dataset was used to gather the participant and chatbot links. The code found in `Thesis GitHub/code/eval/context_links_compare.ipynb` extracts URL lists and calculates the F1-Score:

- **True Positives (TP):** Links present in both `human_links` and `chatbot_links`, indicating overlap between the human *gold standard* and the chatbot's selection.
- **False Positives (FP):** Links that appear only in `chatbot_links` (i.e., extra or potentially irrelevant links the chatbot used but humans did not).
- **False Negatives (FN):** Links that appear only in `human_links` (i.e., the chatbot missed links the human considered important).

From these counts, the script computes:

1. **Precision**: The fraction of chatbot-used links that the human also used

$$\text{Precision} = \frac{TP}{TP + FP}$$

[4]See OpenAI's website: https://platform.openai.com/docs/guides/structured-outputs
[5]See OpenAI's website: https://platform.openai.com/docs/guides/prompt-engineering

2. **Recall**: The fraction of human-chosen links correctly identified by the chatbot

$$\text{Recall} = \frac{TP}{TP + FN}$$

3. **F1-Score**: The harmonic mean of precision and recall

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

   A special case occurs when neither the human nor the chatbot selects any links. According to the formula, this would result in an F1-score of 0. However, in this specific instance, it should be regarded as a positive outcome since the chatbot correctly avoided retrieving any links. Therefore, in this case the F1-score was manually set to 1.

This setup facilitates both per-question and aggregated analyses. Per-question metrics highlight specific cases of missed (FN) or extra (FP) sources, whereas averaged metrics reveal overall trends in how reliably the chatbot's retrieval aligns with human-chosen references.

4.3.7 Evaluating the Evaluation Methods

In addition to collecting and analyzing human ratings and automated metric scores, this thesis also examines the *consistency* of the human evaluations and their *alignment* with automatic methods. Three key steps guided this meta-evaluation:

1. **Inter-Annotator Agreement (Krippendorff's Alpha).** Multiple participants rated each QA-pair across the five introduced dimensions plus an overall impression dimension. To assess the consistency of participants' ratings across dimensions, Krippendorff's alpha was calculated (Krippendorff, 1980). Unlike simpler statistics (e.g., Cohen's kappa), Krippendorff's alpha accommodates varying numbers of raters and ordinal data, making it suitable for capturing rating consistency in multi-annotator settings.

2. **Aggregation of Human Ratings.** The ratings from Survey 3 were aggregated by taking the *average score per QA-pair and dimension*. This yields a single representative human evaluation for each dimension of each QA-pair, mitigating noise in individual responses and simplifying downstream comparisons.

   This thesis follows Prabhakaran, Davani, and Díaz (2021) three recommendations for releasing human-annotated data. In addition to the aggregated human ratings, annotator-level labels (found at Thesis GitHub/data/human_eval), annotator demographic information (see Appendix A.4.2), as well as the recruitment strategy (see Section 4.4.1) were disclosed.

3. **Correlation with Automated Metrics.** Finally, the aggregated human ratings were compared to the outputs of the automatic metrics (lexical and semantic) and the LLM-as-a-Judge scores. *Spearman's rank correlation* (Zar, 2005) was employed to capture monotonic relationships even if the distributions are not strictly linear, as common in the literature (Liu et al., 2023b; Wang et al., 2023b; Fu et al., 2023). A higher Spearman correlation indicates stronger alignment between the automated metric and the aggregated human judgment, providing an index of how accurately each metric aligns with human evaluations of answer quality.

This multistage approach—inter-annotator agreement, rating aggregation, and cross-metric correlation—provides insights into the agreement of the human evaluations themselves and the extent to which automated metrics can approximate human judgments in practice.

4.4 Limitations and Ethical Considerations

This section summarizes key constraints and ethical considerations in data collection and evaluation. It focuses on how participants, sampling, and measurement tools might influence the credibility and applicability of the results.

4.4.1 Survey-Based Data Collection

Convenience Sampling and Self-Selection

All data collection surveys relied on convenience sampling via university mailing lists, social media, and secondary school collaborations. Despite enabling efficient recruitment, convenience sampling limits the representativeness of the participant pool; individuals most interested in the chatbot or drawn to the offered incentives (e.g., participant-hour credits and prize drawings) may be overrepresented. This self-selection bias can affect the variety and difficulty of user-submitted questions (Survey 1), the thoroughness of reference answers (Survey 2), or the consistency of ratings (Survey 3).

Language Nuances and Translations

Surveys were conducted in both German and English (with the exception of Survey 2, which was solely German). Although bilingual formats increase accessibility, they also introduce potential linguistic discrepancies between original items and translated versions. Variations in instruction clarity or subtle differences in phrasing may affect how participants perceive and rate chatbot answers.

Participant Fatigue in Survey 3

Because each participant was asked to rate multiple QA-pairs on five different dimensions, *respondent fatigue* could have occurred. Although forced responses minimized missing data, tired participants may have provided less careful evaluations and subjective terms (e.g., "hallucination" or "accuracy") left room for personal interpretation. Consequently, the reliability of these dimensions can be partially constrained by participant engagement.

Ethical Compliance

All surveys adhered to ethical standards:

- **Informed Consent and Voluntary Participation:** Participants were provided with an overview of the study's objectives and informed that they could withdraw at any time without any repercussions.
- **Anonymity and Data Protection:** Personal information was minimized, with email contact for incentives stored separately and deleted after use.

- **Bilingual Proficiency for Survey 3:** Non-bilingual individuals were excluded to ensure fair evaluation of both German and English question-answer pairs.

While these measures helped mitigate ethical risks and protect privacy, using incentives and open invitations necessarily entailed some level of sampling bias.

4.4.2 Limitations of Automated Evaluation Metrics

Lexical Metrics (BLEU, ROUGE)

BLEU and ROUGE do not account for semantic correctness or factual completeness. Short or paraphrased, yet *valid*, answers may receive artificially low scores, whereas verbatim overlap with a reference can be rewarded even if the underlying content omits important details.

Semantic Metrics (BERTScore, BARTScore, BLEURT)

BERTScore relies on pre-trained multilingual embeddings, which do not necessarily capture the context specific to Osnabrück University. BARTScore, on the other hand, requires significantly higher computational resources when applied to multilingual setups. BLEURT has demonstrated strong correlations with human evaluations of adequacy and fluency; however, its performance can be compromised by noisy scores if not fine-tuned on relevant in-domain data.

Additionally, token limits in these models may truncate lengthy outputs, leading to the omission of portions of answers during the scoring process, which can affect the reliability of the evaluations.

4.4.3 LLM-as-a-Judge Constraints

Finally, an additional evaluation used the GPT-4o-based *LLM-as-a-Judge* approach. This entails:

Potential Biases: Known issues such as position bias or verbosity bias can shift numeric scores in systematic ways.

Prompting Language and Scale: Both German and English QA-pairs were evaluated with English instructions. Future work might compare the use of German prompts for German QA-pairs to improve the evaluation quality.

Score Granularity: A simple 0–4 scale was used for each dimension. While narrower scales or chain-of-thought methods may enhance rater consistency, the optimal scoring strategy remains unclear.

Model Choice: Different large language models or more advanced prompt engineering approaches could yield different judgments, influencing the perceived performance of the chatbot.

Scaling Costs: Evaluating each dimension separately increases API calls and costs, limiting large-scale or repeated evaluations.

4.5 Summary

This chapter outlined the methodological approach employed to evaluate the Osnabrück University chatbot. A robust, bilingual dataset was created through surveys targeting authentic user groups, ensuring the inclusion of genuine, diverse questions and reference answers. The subsequent evaluation used a combination of human judgments and automated metrics, including lexical, semantic, and LLM-based scoring, to assess the chatbot's performance across key dimensions. The careful design of surveys, processing pipelines, and metric implementations ensured transparency, reproducibility, and ethical rigor.

The chapter concluded by addressing limitations and considerations, including potential biases in sampling, language nuances, and the constraints of automated evaluation tools. These factors provide critical context for interpreting the results.

The following chapter presents these results, highlighting key findings and offering insights into the chatbot's strengths, weaknesses, and areas for improvement.

Chapter 5

Results

5.1 Overview

This chapter presents the findings of the methods described in Chapter 4. Section 5.2 begins with the human evaluation results and provides a baseline for the chatbot's performance. The subsequent sections compare automated metrics with human judgments, highlight differences across languages, and analyze the chatbot's context selection.

5.1.1 Summary of Collected Data

The multistage data collection process yielded three principal datasets:

- **Question Survey (Survey 1):** Conducted with 51 participants, resulting in 337 unique questions. Of these, 203 originated from enrolled students, 52 from prospective students, 49 categorized to 'other' and 33 from international students.
- **Reference Answers (Survey 2):** From 11 enrolled students, 33 German reference answers were collected, addressing 11 questions each for prospective, enrolled, and international students. These answers were then translated into English, resulting in a total of 66 QA-pairs.
- **Evaluation Survey (Survey 3):** Involved 30 participants who collectively provided evaluation ratings on the chatbot 66 QA-pairs. Each pair was evaluated by three different annotators, ensuring multiple independent judgments per QA-pair.

All results and the datasets can be accessed in the `Thesis GitHub/data` folder.

5.1.2 Automated Metrics

To complement human evaluations, 16 different automated methods were applied to each QA-pair in both German and English. These methods included seven variants of ROUGE, BLEU, BERTScore, BARTScore with two different models (`bart-large-cnn` and `facebook/mbart-large-50-many-to-many-mmt`), BLEURT, and LLM-as-a-Judge in four configurations: together with no reference answer, together with reference, separated with no reference, and separated with reference.

5.2 Human Evaluation Results

This section presents the human participants ratings of the QA-pairs, serving as a baseline or *gold standard* for subsequent comparisons with automated metrics. The following subsections first report inter-annotator agreement using Krippendorff's α, then provide descriptive statistics for each rated dimension, ending in a brief overview of the rating distributions.

5.2.1 Inter-Annotator Agreement

The consistency of human ratings across the introduced dimensions (see Section 4.3.1 was assessed using Krippendorff's α. Table 5.1 summarizes the inter-annotator agreement values for the overall dataset as well as for the German and English subsets. The agreement levels varied across dimensions and languages, reflecting the subjective nature of the evaluation tasks.

TABLE 5.1: Krippendorff's α for the six rated dimensions, computed over the whole dataset, and separately for each language subset. Higher values indicate greater consistency among raters.

Dimension	**Overall**	**German**	**English**
Hallucination	0.215	0.208	0.256
Answer Accuracy	0.097	0.105	0.061
User Satisfaction	0.108	0.087	0.120
Coherence, Clarity, and Fluency	0.340	0.304	0.411
Context Quality	0.191	0.215	0.159
Overall	0.190	0.277	0.163

The highest agreement was observed for the dimension of *coherence, clarity, and fluency*, particularly in the English dataset (Krippendorff's α = 0.411), suggesting moderate consensus on the linguistic quality of responses. In contrast, the dimensions of *answer accuracy* and *user satisfaction* showed lower agreement levels, underscoring the inherent subjectivity and variability of human evaluations.

Overall, the Krippendorff's α values are relatively low (ranging from roughly 0.06 to 0.41 across different dimensions). These values indicate that while the participants somewhat agreed on the broad notion of *coherence/clarity/fluency* (highest α), there was more variability in how they judged dimensions, such as *answer accuracy* and *user satisfaction*. Despite these modest agreement scores, this thesis continues to use the averaged human ratings for each QA-pair.

5.2.2 Dimension-by-Dimension Averages

To gain an overview of how participants rated the chatbot across *hallucination*, *answer accuracy*, *user satisfaction*, *coherence/clarity/fluency*, *context quality*, and an *overall* dimensions, Table C.1 in the Appendix provides descriptive statistics (mean ± standard deviation). Figure 5.1 then visualizes these means providing an overall snapshot of the chatbot's performance, with higher scores indicating better evaluated quality.

Alongside the *overall* rating provided by participants, an additional dimension, *Overall (Mean)*, is included. This represents the average of the five evaluated dimensions.

Hallucination: The chatbot received an average score of 4.24 across all QA-pairs, with slightly higher scores in the German subset (4.31) compared to the English subset (4.17).

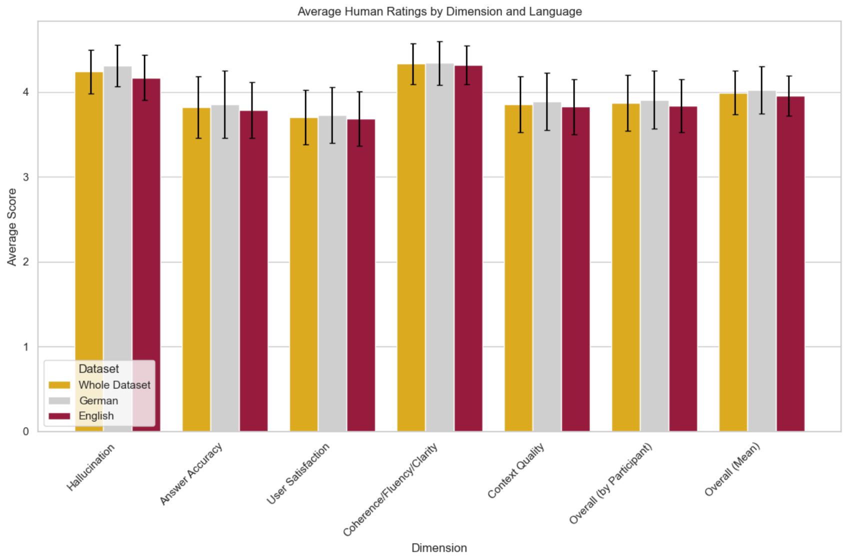


FIGURE 5.1: Mean ratings (± SD) for each dimension (1–5 scale). Bars compare the entire dataset, as well as separate subsets for German and English. Error bars indicate the dimension-wise standard deviations.

Answer Accuracy: The average score for answer accuracy was 3.82, indicating variability across dimensions. The German subset scored slightly higher (3.86) than the English subset (3.79).

User Satisfaction: The user satisfaction scores averaged 3.71, with minimal variation between German (3.73) and English (3.69).

Coherence, Clarity, and Fluency: This dimension received the highest average score of 4.33 across all QA-pairs, with consistent performance in both German (4.34) and English (4.32).

Context Quality: Context quality averaged 3.86, with German QA-pairs scoring slightly higher (3.89) compared to English (3.83).

Overall (by Participant): The overall impression of the participants across the whole dataset achieved a score of 3.87 with the German subset scoring a bit higher (3.91) in comparison to the English subset (3.84)

Overall (Mean): Across all dimensions and QA-pairs, the overall mean score was 3.99, highlighting consistent performance across both language subsets, with German scoring 4.03 and English 3.96.

The results indicate that while the chatbot consistently received high marks overall and especially in *coherence, clarity and fluency*, there remains room for improvement in dimensions such as *answer accuracy*, *user satisfaction* and *context quality*.

5.2.3 Rating Distributions

Figure 5.2 illustrates the distribution of participant ratings across all evaluation dimensions. The stacked bar chart shows the proportion of ratings on the Likert scale ranging from 1 (Very Bad) to 5 (Very Good) for each dimension. This visual representation provides insights into the variability of participant scores and highlights differences in rating patterns across dimensions.

The results for the combined dataset (marked as *ALL* in Figure) reveal that:

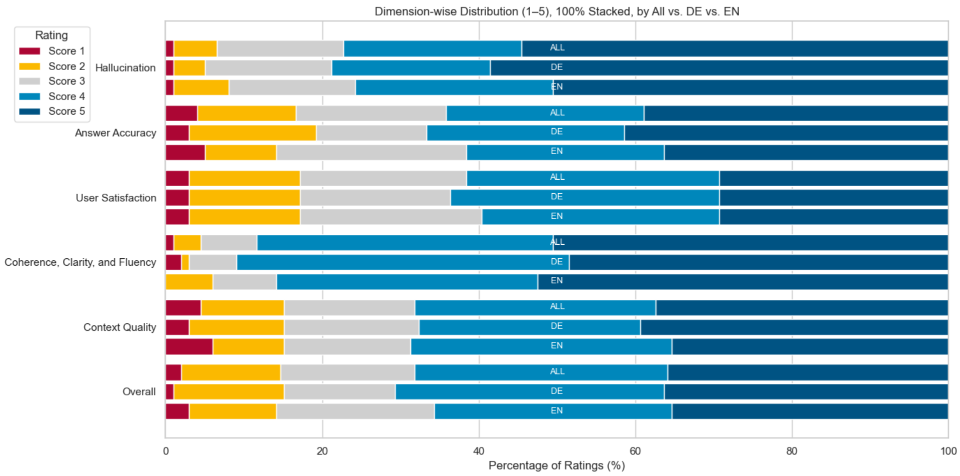


FIGURE 5.2: Proportion of ratings for each dimension (1 = Very Bad, 5 = Very Good). Bars represent the percentage of times each dimension received a particular rating. This helps differentiate whether a dimension's mean is driven by consistently moderate ratings versus a mix of high and low ratings.
See Table C.2 in the Appendix for the exact values.

Hallucination: Over half (54.55%) of the ratings fall into the highest category (5), suggesting that the chatbot often avoids generating hallucinatory content. However, lower ratings (1 or 2) accounted for approximately 6.57%, indicating occasional failures in generating factual and faithful content.

Answer Accuracy: Ratings are more evenly distributed, with 38.89% receiving the highest score (5). While 25.25% of ratings fall in the second-highest category (4), the presence of lower ratings (16.67% at 1 or 2) indicates variability in the chatbot's ability to fully address user questions.

User Satisfaction: Although the majority of ratings (61.61%) are in the upper categories (4 and 5), a notable 17.17% of responses were rated as 1 or 2, reflecting room for improvement in the chatbot's impression on users.

Coherence, Clarity, and Fluency: This dimension exhibits the highest concentration of positive ratings, with 50.51% rated as 5 and an additional 37.88% rated as 4. Only 4.55% of the responses received a score of 1 or 2, confirming the chatbot's strong performance in producing fluent and comprehensible answers.

Context Quality: While 37.37% of ratings are in the highest category (5), lower scores (1 or 2) account for 15.16%. This suggests that while the chatbot often uses relevant sources, its retrieval performance occasionally falls short.

Overall Impression: The ratings for overall quality closely mirror those of context quality, with 68.18% rated as 4 or 5. A small proportion (14.65%) of responses received scores of 1 or 2, indicating some variability in overall answer quality.

The data indicate strong performance in *coherence, clarity, and fluency*, while dimensions like *answer accuracy*, *user satisfaction*, and *context quality* exhibit greater variability. This variability underscores areas where the chatbot's performance can be enhanced to provide a consistently high-quality user experience.

Analyzing the English and German QA-pairs alone, as in Figure 5.2 (*DE* & *EN*), shows a comparable picture with slightly better score for the German subset.

In sum, the human evaluations indicate that the chatbot's language quality (*coherence/clarity/fluency*) and relatively few perceived *hallucinations* are the strongest aspects, while *answer accuracy*, *user satisfaction* and *context quality* appear to vary more. Although the Inter-annotator agreement was modest to low, the aggregated means and distributions revealed a generally positive but not flawless chatbot. These human ratings form the foundation for subsequent sections, where automated metrics and an LLM-based judge are compared against this *gold standard*.

5.3 Automated Metrics

5.3.1 Lexical Metrics: BLEU and ROUGE

Lexical similarity metrics, specifically BLEU and ROUGE, were employed to evaluate the surface-level overlap between the chatbot's generated text and the human-provided reference answers and were correlated with the human evaluations.

Results

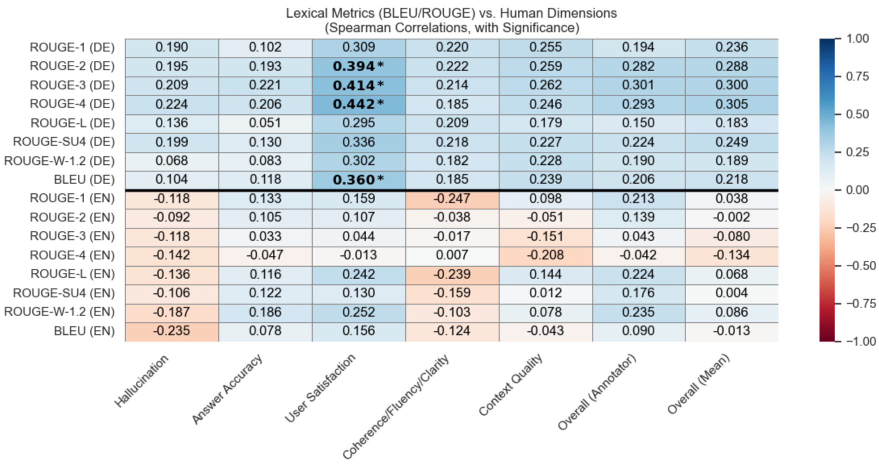


FIGURE 5.3: Heatmap showcasing Spearman correlations(r) between lexical metrics (BLEU/ROUGE) and human evaluation dimensions, with significance levels ($p < 0.05$) indicated by bold text and asterisks. Positive correlations are shown in blue, and negative correlations in red, with intensity reflecting correlation strength.

The analysis revealed moderate correlations between lexical metrics and human evaluations, with results varying by dimension and language subset. Figure 5.3 illustrates the Spearman correlations between BLEU, ROUGE, and the human evaluation dimensions.

BLEU: BLEU showed a significant weak positive correlation with the dimension *user satisfaction* in the German dataset ($r = 0.360$, $p = 0.04$). However, correlations with other dimensions were not statistically significant. In the English dataset, BLEU did not exhibit any significant correlations with human-rated dimensions.

ROUGE: ROUGE metrics demonstrated stronger correlations in the German subset compared to English. Notably, ROUGE-4 correlated moderately positive with *user satisfaction* ($r = 0.442$, $p = 0.01$), indicating that higher n-gram overlap aligns with higher satisfaction levels from human evaluators. ROUGE-2 and ROUGE-3 also showed significant correlations with *user satisfaction* ($r = 0.394$, $p = 0.02$ and $r = 0.414$, $p = 0.02$, respectively). Other dimensions exhibited weaker and non-significant correlations across both languages.

The results highlight that lexical metrics are more predictive of user-perceived satisfaction in the German dataset than in the English dataset. This discrepancy could reflect differences in linguistic structure or the translation alignment of reference answers. Despite their limitations in capturing semantic equivalence or factual correctness, BLEU and ROUGE remain useful for identifying lexical similarity trends. However, their lower correlations with dimensions such as *hallucination* underscore the need for complementary semantic and LLM-based metrics to assess chatbot performance that aligns with human judgement.

5.3.2 Semantic Metrics: BERTScore, BARTScore and BLEURT

Semantic-oriented metrics were applied to assess the deeper contextual and semantic alignment between the chatbot's generated answers and human-provided reference answers. This section discusses the results for BERTScore, BARTScore, and BLEURT, highlighting their performance and correlations with human evaluation dimensions. Results are visualized in Figure 5.4.

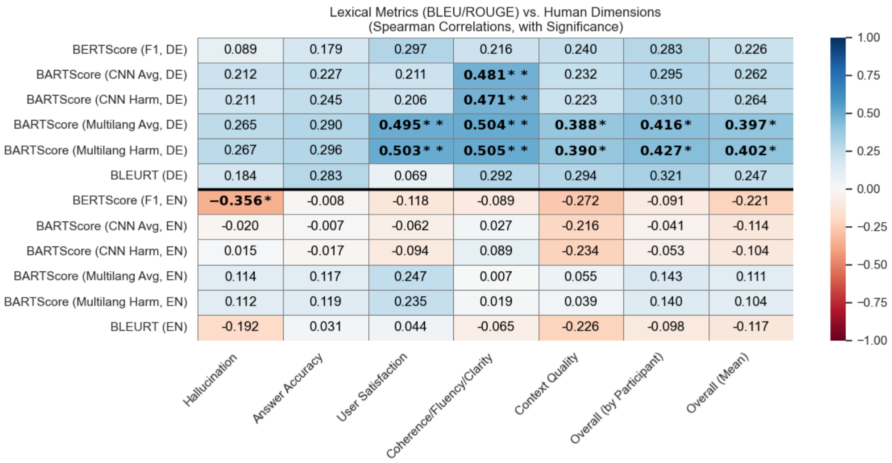


FIGURE 5.4: Spearman correlation (r) between each semantic metric (BERTScore, BARTScore, BLEURT) and the human-rated dimensions (Hallucination, Answer Accuracy, User Satisfaction, Coherence, Context Quality, Overall). Cells are bolded and marked with $*$ to indicate $p < 0.05$, $**$ to indicate $p < 0.01$.

BERTScore

In the German dataset, BERTScore showed weak positive correlations with dimensions such as *user satisfaction* ($r = 0.297$, $p = 0.093$) and *overall (by Participant)*

($r = 0.283$, $p = 0.11$), although these were not statistically significant. In the English dataset, BERTScore exhibited a significant weak negative correlation with *hallucination* ($r = -0.356$, $p = 0.042$), indicating an inverse relationship, where lower semantic alignment corresponded to higher hallucination scores, thus fewer hallucinations.

BARTScore

The German dataset demonstrated significant moderate positive correlations with human evaluations with the multilingual model, particularly in the dimensions *user satisfaction* ($r = 0.503$, $p = 0.003$), *coherence/fluency/clarity* ($r = 0.505$, $p = 0.003$), and both *overall* (by participant: $r = 0.427$, $p = 0.013$ and mean: $r = 0.402$, $p = 0.02$). These findings indicate that BARTScore effectively captures semantic coherence and user-aligned quality in generated answers. In contrast, correlations in the English dataset were weaker and not statistically significant across all dimensions, suggesting variability in performance across languages.

The different F-score calculations (average and harmonic) did not influence the correlation in a meaningful way.

BLEURT

BLEURT was used to assess sentence-level semantic adequacy and fluency. In the German dataset, BLEURT showed a marginal positive correlation with *coherence/fluency/clarity* ($r = 0.292$, $p = 0.099$) and *context quality* ($r = 0.294$, $p = 0.097$), although these correlations were not statistically significant. The English dataset demonstrated weaker correlations, with no dimension showing significant alignment with BLEURT scores.

Overall, semantic metrics, particularly BARTScore, demonstrated significant correlation with human evaluations in the German dataset, indicating their effectiveness in capturing semantic coherence and general quality. However, performance variability across languages and metrics underscores the need for domain-specific fine-tuning and further exploration of semantic evaluation methods.

5.3.3 LLM-as-a-Judge

While lexical and embedding-based metrics offer indirect measures of answer quality, an alternative is to have a LLM evaluate each QA-pair on the same dimensions used by human evaluators. As discussed previously, this thesis tested four prompt configurations of the LLM-based judge (see Section 4.3.5).

Overview of LLM-as-a-Judge Data

Table C.5 in the Appendix summarizes basic statistics (mean, standard deviation) for the LLM-given scores under each configuration. Generally, the LLM assigned moderately high marks (3–4 on a 0–4 scale) in most cases, particularly on dimensions such as *hallucination* and *coherence/clarity/fluency*. However, there was notable variability and lower scores (dropping below 3) when it had access to the human reference, particularly in *answer accuracy*, *user satisfaction*, and *context quality*. Suggesting that including the reference sometimes led the LLM to penalize the chatbot more harshly if it perceived discrepancies.

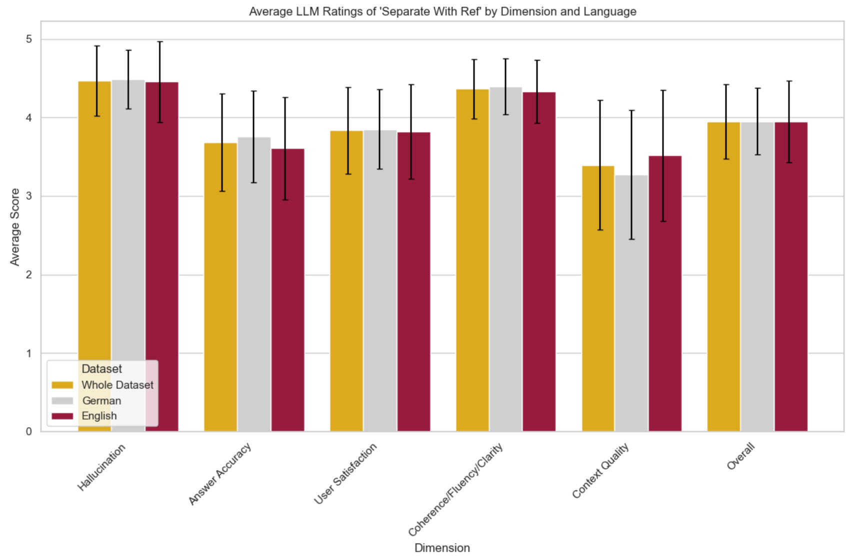


FIGURE 5.5: Mean ratings (± SD) for each dimension (shifted by one on a 0–4 scale). Bars compare the entire dataset, as well as separate subsets for German and English. Error bars indicate the dimension-wise standard deviations.

Figure 5.5 shows the average scores, shifted by one to be visually comparable to the human ratings, for one of the best LLm-as-a-Judge configurations, *Separate with Ref*.

Notably, in the *Together No Ref* configuration for the *coherence/clarity/fluency* dimension within the English dataset, the LLM assigned the maximum score to every QA-pair.

Correlation with Human Scores

The LLM generated a *score* for each dimension, which the researcher correlated with the matching averaged human dimension. Additionally, the overall score from the LLM (see Section 4.3.5) was compared to both *overall (by Participant)* and *overall (Mean)*.

Figure 5.6 lists these correlations, showing only the one-to-one matches (e.g., LLM *hallucination* score vs. human *hallucination* score).

Observations and Patterns

- **Influence of Reference:** Across all but one instance, adding the reference answer (*with ref*) appears to increase correlation with human ratings. For example, *separate with reference* in German yields significant weak to moderate positive correlations ($r \approx 0.38$–0.55) on multiple dimensions, whereas *separate no reference* has fewer and lower significant correlations.

- **Separate vs. Together Prompts:** Particularly in the *English* subset, the *separate* approach often outperforms the *together* approach on many dimensions. This suggests that prompting the LLM individually for each dimension may yield more focused and reliable scores.

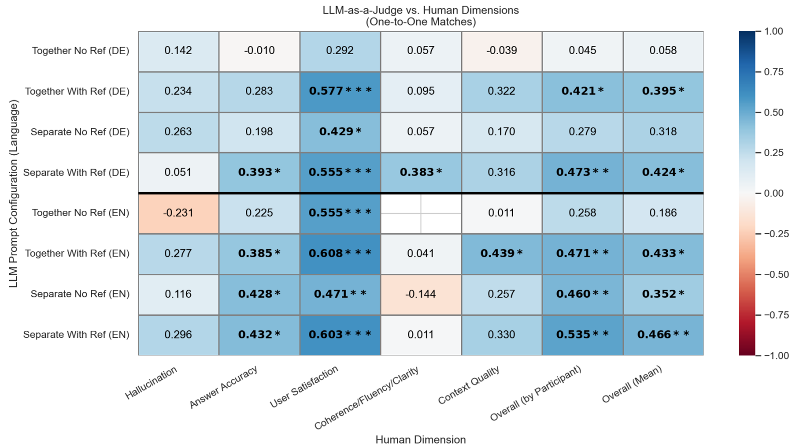


FIGURE 5.6: Spearman correlation (r) between LLM-based judge scores and human dimensions under four prompt configurations (*together* vs. *separate* prompts; *with ref* vs. *no ref*), across German (DE) and English (EN). Statistically significant correlations ($p < 0.05$) are in bold; $*$, $**$, $***$ denote $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

- **Dimension Differences:** Among the five main dimensions, *user satisfaction* most frequently shows robust positive correlations (often $r > 0.50$, $p < 0.01$). In contrast, *hallucination* shows less consistent correlations and no significant alignment, suggesting that the LLM judge may face challenges in detecting factual inaccuracies.

- **Overall Scores:** Correlations for the LLM's *overall score* with both *overall (by Participant)* and *overall (Mean)* are moderately high under *with reference* conditions, especially in English. This suggests that when exposed to both the chatbot answer and the human reference, the LLM can approximate a global quality rating that aligns moderately well with human evaluations.

In summary, providing the reference answer and prompting each dimension separately tend to boost alignment with human judgments, echoing prior findings that LLMs benefit from more specific prompts and domain context. However, as with other automated metrics, correlations vary across dimensions and languages, pointing to inherent challenges in fully emulating human evaluations with an LLM-based judge.

5.4 Context Quality Analysis

In addition to evaluating textual aspects of chatbot answers, this study also examined how effectively the chatbot retrieves relevant websites as context relative to human-provided links. This section analyzes the overlap between links used by the chatbot and those cited by humans, followed by a correlation test between link overlap and the human-rated *context quality* dimension.

5.4.1 Link Usage and Overlap

Chatbot vs. Human Link Counts. Over the whole dataset, human annotators cited roughly 2 links per question, whereas the chatbot produced approximately 2.6 links in English and 2.3 links in German. Notably, the chatbot defaulted to retrieving four links if the agent decided to retrieve context, as hard-coded in its custom web search tool; however, a significant fraction of its responses used *no* links (33% of English answers, 42% of German answers). In contrast, humans used no links only about 15% of the time. Moreover, approximately 6% of human answers exceeded four links, indicating that humans occasionally referenced a broader set of resources than the chatbot's default retrieval limit.

Link Overlap via F1-Score To quantify how closely the chatbot's chosen links match the links used by humans, this thesis calculated a per-question F1-score based on true positives (same links cited by both), false positives (chatbot-only links), and false negatives (human-only links). On average, the chatbot achieved an F1 of $0.19 - 0.20$, indicating modest to low overlap, as seen in Tabel 5.2

TABLE 5.2: Context links comparison between the link usage of the chatbot vs the link usage of human participants for the English and German Dataset

Metric	English Dataset	German Dataset
Average Recall	0.202	0.169
Average Precision	0.098	0.083
Macro level F1-Score	0.186	0.195

These relatively low scores indicate that the chatbot often references links that differ from those deemed most relevant by humans. Additionally, it often omits human-cited links. Moreover, in 2 cases in the English dataset and 3 in the German dataset, neither the human nor the chatbot provided links.

5.4.2 Correlation with Context Quality Dimension

To explore whether better link overlap translates into higher *context quality* in human evaluation, the Spearman's correlation between each QA-pairs F1-link-overlap and its human *context quality* rating was calculated. In English, the correlation was $r = 0.133$ ($p = 0.460$), and in German it was $r = 0.304$ ($p = 0.085$). Although the correlation in German is a bit higher, it still does not reach conventional significance ($p < 0.05$).

Hence, while there is a mild tendency for better overlap (F1) to coincide with higher context quality scores, the relationship is not particularly strong or statistically conclusive.

5.5 Summary

Overall, the evaluations reveal that while the chatbot exhibits strengths in linguistic fluency and low hallucination rates, its *answer accuracy* and context usage remain inconsistent. Traditional lexical and semantic metrics partially track human evaluation of quality, but their correlations vary widely by language and dimension, highlighting the complexity of automated evaluation in this domain. The LLM-as-a-Judge method offers more targeted alignment with human judgments when provided with reference answers, although further refinements are needed to capture nuanced aspects like *hallucination* and *answer accuracy*. Finally, the chatbot's link retrieval strategy diverges notably from human-cited sources, yet only weakly correlates with human ratings of *context quality*.

These findings collectively inform the final discussion in Chapter 6, which addresses possible procedures for improving the chatbot's performance, limitations of the results, and directions for future research.

Chapter 6

Discussion

6.1 Introduction

Building on the results presented in Chapter 5, this chapter provides an in-depth interpretation of the chatbot's performance within Osnabrück University's context. By reflecting on prior research, the subsequent sections evaluate whether the chatbot met its core objectives of delivering accurate, low-hallucination answers grounded in relevant university documents.

This chapter begins by linking the main observations to the research questions and then outlines broader limitations, considerations for potential improvements of the chatbot, and future research avenues.

6.2 Analysis of Findings in Light of the Research Questions

This thesis aimed to test a domain-specific, RAG chatbot (askUOS) to address queries from prospective, enrolled, and international students at Osnabrück University. The evaluation focused on multiple performance dimensions, including *hallucination*, *answer accuracy*, *user satisfaction*, *coherence, clarity and fluency* and *context quality*, while also exploring the utility of automated metrics. The findings, summarized below, reflect both the chatbot's strengths and areas in need of further refinement.

6.2.1 Overall Performance

Human evaluators gave notably high ratings for *coherence, clarity, and fluency*, averaging above 4.3 on a 5-point scale. This outcome parallels existing literature indicating that LLMs often excel in producing text that is syntactically well-formed and stylistically convincing (Hariri, 2024). Reassuringly, *hallucinations* were also rated favorably (above 4.2), suggesting that the RAG framework was largely successful at producing factual and faithful content.

In spite of these strengths, *answer accuracy* received lower ratings (just under 3.9) and displayed considerable variance; indicating that some answers lacked precise or complete domain-specific detail, pointing to occasional retrieval gaps or system uncertainty regarding whether a given question fell within its target scope (e.g. question ID 9's English answer). Moreover, high fluency may mask inaccuracies (Li et al., 2024; Varshney et al., 2023). This suggests that retrieval inconsistencies or limitations in grasping questions intent may influence performance.

Inter-annotator agreement was low across most dimensions, underlining the inherent subjectivity in human judgments of the chatbot outputs (Belz et al., 2023; Gehrmann, Clark, and Sellam, 2022).

6.2.2 Answer Accuracy and Relevance

RQ: ***"Are the answers correct, complete, and relevant?"***

Although human participants generally perceived the chatbot's answers as sufficiently relevant and accurate, its average *answer accuracy* score fell below 3.9 on a 5-point scale. The breadth of potential university-related questions, as well as the variable depth of official web resources, may influence both the completeness and exactness of the chatbot's responses. The results confirm prior findings that domain-specific retrieval is not always straightforward: relevant information may be scattered across multiple pages, potentially causing the chatbot to omit details that human reference answers included (Cuconasu et al., 2024).

Interestingly, *user satisfaction* mirrored this moderate performance in answer accuracy. The chatbot sometimes declined to answer questions even if they were within its intended domain, which might have contributed to these scores. These observations suggest opportunities for further refinement in retrieval processes and response strategies in the system prompt.

6.2.3 Hallucination

RQ: ***"To what extent does the chatbot generate information unsupported by retrieved or widely recognized facts?"***

Evaluators found few instances of hallucinations, as indicated by an average *hallucination* rating of over 4.2/5. This outcome indicates that integrating university-specific documents into the prompting process helps ground the chatbot's responses, thereby reducing unfaithful outputs (Shuster et al., 2021; Huang et al., 2023). Some inaccuracies persisted in the chatbot's responses, sometimes when it was overly cautious or left its statements vaguely incomplete (e.g. question ID 212's answers, question ID 73's answers). This highlights that not all errors are obvious fabrications, and a more nuanced definition is required to understand these inaccuracies fully (Mishra et al., 2024; Zhang et al., 2023; Li et al., 2024). These findings confirm that while RAG-based systems can mitigate overt hallucinations, they require ongoing refinement to improve factual coverage.

6.2.4 Context Quality

RQ: ***"How appropriate are the documents the chatbot retrieves for a given question?"***

The evaluation survey included a *context quality* rating, capturing whether retrieved documents aided in forming an accurate response. Although human evaluators assigned moderate-to-high scores (just under 3.9), the chatbot's cited URLs did not significantly overlap with the so-called *gold standard* human-selected links (macro-F1 $\approx$ 0.19–0.20). Breaking this further, recall rates were around 17–20%, while precision scores were less than 10%, indicating that the chatbot often *missed* links deemed valuable by humans and *added* links considered unnecessary. Despite this, participants did not strongly penalize the chatbot for referencing alternative pages, indicating that information could be found on different webpages of the university or perceived context quality is more influenced by the utility of the final answer than strict link alignment.

Roughly 37% of the chatbot's outputs omitted any references altogether, in comparison with 15% in human reference answers. It is unclear whether these omissions stem from the chatbot deeming the questions answerable without retrieval or considered them out of scope.

6.2.5 Answer Coherence

*RQ: **"How coherent, clear, and fluent is the chatbot's language usage?"***

Across both German and English responses, evaluators rated *coherence/clarity/fluency* the highest dimension. This finding reaffirms known strengths of advanced LLMs in producing grammatically well-formed text that reads naturally. From a practical standpoint, fluency and clarity may enhance user trust, even if subtle inaccuracies persist.

6.2.6 Language Comparison

*RQ: **"Does the chatbot perform differently in German vs. English?"***

Although evaluators' overall scores were broadly similar across languages with German scoring slightly higher in all dimensions, subtle distinctions emerged in how automated metrics correlated with human judgments. For example, BARTScore showed stronger correlations in the German subset compared to the English. This may reflect differences in the training data's coverage or quality for each language.

Surprisingly, the German subset was found to utilized the retrieval mechanism less often than the English subset (42.42% vs. 33.33% no retrieval usage), despite its better scores. Emphasizing a need to explore retrieval behavior across languages. These findings highlight the importance of considering multilingual variations when designing and deploying chatbots.

6.2.7 Automated Metrics and the LLM-as-a-Judge Approach

*RQ: **"Can these dimensions be evaluated automatically?"***

Lexical metrics like BLEU and ROUGE correlated only weakly with human evaluations, reaffirming concerns that simplistic n-gram overlaps cannot reliably track hallucination (Maynez et al., 2020) and other dimensions. More encouragingly, the *LLM-as-a-Judge* framework yielded moderate to strong correlations with human evaluators—especially when the model was provided with the reference answer and prompted to assess each dimension separately. This result aligns with emerging evidence that specialized prompting can turn LLMs into competent evaluators of each other's outputs (Chiang and Lee, 2023; Liu et al., 2023b).

While human evaluation remains indispensable, automated methods, particularly LLM-as-a-Judge, offer scalable supplements. However, the correlation with the human *hallucination* dimension was relatively low, indicating that subtle factual inaccuracies may still slip through the automated evaluation net (Ji et al., 2024). Further fine-grained prompt engineering or using different models may enhance the reliability of this method. In addition, self-reference bias is a potential concern when using the same model family for both generation and evaluation (Zheng et al., 2023; Panickssery, Bowman, and Feng, 2024).

6.2.8 Synthesis and Preliminary Conclusions

The findings reinforce the notion that retrieval-augmented systems can mitigate overt hallucinations and deliver accurate, relevant, coherent, and fluent answers, but they do not guarantee optimal information coverage or document matching. For a university context, this duality implies that students might experience high-quality language and minimal blatant errors, though certain finer details could still be overlooked.

Moreover, these results highlight important methodological insights. On the one hand, purely lexical metrics struggle to capture nuanced aspects in the evaluations. On the other hand, more advanced model-based evaluators, such as the *LLM-as-a-Judge* approach, promise cost-effective scalability yet still need reference answers to pinpoint subtle differences. Ultimately, ensuring robust, domain-specific performance likely demands an iterative, hybrid evaluation process that leverages both automated metrics and carefully structured human reviews.

The following section addresses broader limitations that shape the interpretation of these results, providing context for how they might be generalized or refined in future iterations.

6.3 Limitations in Interpreting the Results

Beyond the constraints embedded in data collection and metric choice discussed in the Methods Chapter, several broader issues limit the generalizability and applicability of the findings.

6.3.1 Sample Size and Coverage of QA-Pairs

Although the dataset contained 66 total QA-pairs (33 per language), this volume may not reflect the diverse types of inquiries encountered in practice. The modest number of items also restricts statistical power, making it unreliable to generalize the observed correlations to all conceivable university queries.

The limited sample size and large number of correlations tested (16 metrics, 7 dimensions) increase the likelihood of false positives. Reported p-values should be read with caution; more robust or confirmatory studies would require larger datasets and possibly pre-registered analysis plans.

6.3.2 Inter-Annotator Agreement and Subjectivity

Krippendorff's α values were low across most dimensions, signaling that different evaluators often disagreed on what constitutes a good or bad response. Such subjectivity complicates definitive conclusions and suggests the need for further standardization or professional evaluators, consistent with broader calls for improved human evaluation in NLP (Belz et al., 2023).

Additionally, the five rating dimensions may inadvertently overlap or omit crucial qualitative aspects. Future evaluations could adopt finer-grained dimensions to uncover more nuanced insights.

6.3.3 Reference Answers and Website Dynamics

Survey 2 reference answers were collected 1–2 weeks before the chatbot's answers were generated, during which time university web pages could have changed. This temporal gap might introduce mismatches between the chatbot's retrieved context and the *ground-truth* references. Moreover, there was no verification to ensure that Survey 2 participants thoroughly consulted the official website; therefore, their reference answers may contain gaps or inaccuracies. These issues reduce the precision of *reference vs. chatbot* comparisons.

6.3.4 Single-Turn Interaction

All evaluations focused on isolated queries rather than ongoing conversations. In real-world use, many university inquiries require clarifications or multi-step guidance (e.g., complicated enrollment procedures). The model's performance under multi-turn conditions remains untested, which limits the conclusions that can be drawn about its overall dialogue capacity. While this is partly addressed by not filtering out potential questions designed to test the chatbot's ability to ask for clarification, it does not fully resolve the issue.

6.3.5 Domain and Language Factors

Osnabrück University's website offers more extensive German resources than English ones. Furthermore, the English reference answers were translated via machine translation. Although they were additionally verified by the researcher, both factors may have reduced the chatbot's English performance or introduced translation artifacts into the reference dataset. Future research may build a stronger domain-specific dataset with a complete separation of English and German data generation.

6.3.6 Latency and Real-World Feasibility

This thesis did not analyze the latency of responses. However, in extreme cases, some queries took up to 50 seconds, particularly when the chatbot utilized the web search tool, exceeded the context window, and had to summarize the retrieved context. Such delays can significantly impact user satisfaction in practical deployments. Nonetheless, the latency data has been preserved in the dataset, allowing for further analysis to identify patterns or optimizations in future work.

6.3.7 Omitted Alternative Evaluations

Due to time constraints, approaches such as QA-based approaches (see Section 2.4.1) or specialized RAG evaluation methods (see Section 2.4.2) have not been tested. Moreover, the human and LLM provided comments on the evaluation scores were not analysed.

Additional methodological diversity might yield richer insights into how the chatbot handles questions and what new errors might surface under different evaluation angles.

In general, the results of this thesis should be interpreted with the understanding that they are constrained to the specific and narrowly defined cohort of evaluators (see A.4.2), question providers (see A.2.2), and reference answer creators (see A.3.2) involved in the surveys. While the chatbot demonstrates strong performance within this context—showing fluency and a general absence of hallucinations—broader interpretative caution is warranted. The findings do not indicate perfection nor guarantee generalizability beyond this scope.

In conclusion, low agreement among human evaluators, limited QA-pair coverage, and the focus on single-turn testing all limit the ability to conclusively confirm the reliability of the chatbot. Follow-up studies incorporating multi-turn contexts, more diverse or larger data samples, and adding different evaluation metrics could address these limitations, providing a more comprehensive understanding of the chatbot's readiness for real-world deployment.

6.4 Future Research and Practical Implications

6.4.1 Refining the Retrieval Pipeline

One way for improvement is a more sophisticated retriever to ensure better coverage of official pages and reduce the mismatch between chatbot and human-cited links. This might involve embedding-based indexing (Khattab et al., 2023), multi-turn active retrieval (Jiang et al., 2023), or domain-specific query rewriting, which the chatbot already partially employs (Ma et al., 2023). Subsequent evaluations should examine whether improved retrieval leads to measurable gains in context quality and answer accuracy.

6.4.2 Further Analysis

Further analysis of the use of LLMs-as-a-Judges could provide valuable insights, given their promising potential. For instance, researchers could experiment with different prompts, align the language of prompts with that of the QA-pairs, explore alternative scoring scales, and test various models to determine their effectiveness in this domain-specific environment. Additionally Figure C.1 in the Appendix shows the correlation between *all* LLM-as-a-Judge dimensions vs human evaluation dimensions. Further analysis if different dimensions correlate would be an interesting approach.

Additionally, although the QA-pairs were equally role balanced, further analysis of whether the original role of the question provider influenced the score and its correlation with automated metrics would be of interest.

6.4.3 Automated Monitoring and Fine-Tuning

Given the encouraging correlations for some automated metrics, for example BARTScore in German, and the LLM-as-a-Judge approach, it may be feasible to integrate continuous monitoring of new chatbot queries. Additionally, the system could be periodically fine-tuned using user feedback or higher-quality domain data (Ouyang et al., 2022; Tian et al., 2023). Although this is not feasible with the current implementation of the chatbot (closed GPT4o-mini API), future iterations may rely on local or open models. For instance, the open-source model DeepSeek (DeepSeek-AI et al., 2025; DeepSeek-AI et al., 2024) appears to be a promising candidate, despite its recent release and lack of independent testing.

6.4.4 Other Automatic Evaluation and Deeper Hallucination Analysis

While strong hallucinations were rare, systematic analysis is needed to pinpoint *subtle* factual inaccuracies and contradictory claims. More fine-grained annotation of errors could reveal whether certain question types are more prone to hallucinations (Mishra et al., 2024; Varshney et al., 2023). Additionally, other methods for automatic evaluation, as discussed in Section 2.4.1, could be used with the generated dataset to gain a more in-depth understanding of the chatbot quality.

6.4.5 Wider University Integration

Beyond evaluation surveys, real-world usage metrics (e.g., conversation logs, user follow-up queries, drop-off rates, A/B studies) could shed light on how well the chatbot meets authentic users information needs. Data from these interactions might be used to refine both the retrieval pipeline and the model prompts, ultimately guiding a more robust deployment at Osnabrück University.

6.5 Summary

In summary, the RAG-based chatbot for Osnabrück University exhibited a generally positive picture with strengths in generating coherent, fluent and clear answers with relatively few hallucinations. However, the partial misalignment of the system with human-cited reference links and the moderate variability in *answer accuracy* highlight persistent challenges in domain-specific retrieval and response generation. Automated metrics and an LLM judge generally captured certain aspects of quality, but varied in correlation across languages and dimensions, reiterating the importance of *multi-method* evaluations.

Looking ahead, improvements to retrieval pipelines, systematic integration of user feedback, and more fine-grained hallucination analysis could further elevate the chatbots reliability. As universities increasingly adopt AI-driven assistance tools, ensuring accurate and relevant answers, keeping hallucinations low, and having good contextual retrieval remains pivotal. The findings of this thesis provide a benchmark for further optimization, a framework for continued evaluation of future iterations, and a solid dataset to build evaluation on.

Chapter 7

Conclusion

This thesis aimed to investigate how effectively a RAG chatbot can provide accurate, hallucination-free answers for Osnabrück University by retrieving relevant context. By combining human and automated evaluations under bilingual conditions, this thesis yielded three central insights.

1. **Fluency and Rare Hallucinations**

   The chatbot generally produced coherent, clear, and fluent responses while displaying few instances of hallucination. This affirms that RAG-based approaches can mitigate factual distortions, frequently observed in large language models, by grounding outputs in source documents.

2. **Variability in Accuracy**

   Despite high linguistic fluency, *answer accuracy* and *context quality* exhibited higher variance. The chatbot did not always retrieve or synthesize perfectly aligned information for some queries. Low link overlap with human-cited references, yet moderate to good acceptance of retrieved context, implies that the presence of *some* verifiable information may suffice for the chatbot, even if it differs from the gold-standard source.

3. **Evaluation Complexity**

   Human ratings showed low inter-annotator agreement, and commonly used automated metrics did not consistently reflect human evaluations. While the *LLM-as-a-Judge* framework strengthened correlations with human judgments–particularly when the reference answer was accessible–it struggled to identify more subtle factual discrepancies. This underscores the importance of *multi-faceted, hybrid* evaluation methods that combine automated checks with targeted human assessments.

From a practical standpoint, these findings demonstrate that a basic RAG-based system can offer an effective starting point for institution-specific chatbots. However, continuous refinement is essential. Integrating real-world user feedback, refining retrieval mechanisms and system prompts, and validating subtle factual claims against domain-specific knowledge sources are key steps to boost *answer accuracy* and *context quality*, ultimately gaining user trust.

In addition, further studies could benefit from:

- **Expanded Datasets and Bilingual Evaluation**

  A larger, more diverse dataset of question-answer pairs that reflects the full breadth of the university's information needs—including multi-turn conversations—would enhance both the reliability and the external validity of future findings.

- **Refined and Specialized Evaluation**

  Advanced techniques such as QA-based checks, specialized RAG evaluation protocols, more granular evaluation dimensions and iterative human-in-the-loop feedback can shed light on subtle quality differences.

In conclusion, while the chatbot demonstrates clear promise in handling Osnabrück University–related queries, ensuring continual accuracy and relevance in a complex, evolving information ecosystem remains a challenge. Nonetheless, the results confirm that RAG-based chatbots can serve as a reliable foundation for user-centered institutional systems

Bibliography

AbuShawar, Bayan and Eric Atwell (Dec. 27, 2015). "ALICE Chatbot: Trials and Outputs". In: *Computación y Sistemas* 19.4. ISSN: 2007-9737, 1405-5546. DOI: 10.13053/cys-19-4-2326. URL: http://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2326 (visited on 10/27/2024).

Al-Amin, Md et al. (Feb. 4, 2024). *History of generative Artificial Intelligence (AI) chatbots: past, present, and future development*. arXiv: 2402.05122. URL: http://arxiv.org/abs/2402.05122 (visited on 10/26/2024).

AlKhamissi, Badr et al. (Apr. 12, 2022). *A Review on Language Models as Knowledge Bases*. arXiv: 2204.06031. URL: http://arxiv.org/abs/2204.06031 (visited on 10/30/2024).

Andreadis, Ioannis and Evangelia Kartsounidou (Apr. 10, 2020). "The Impact of Splitting a Long Online Questionnaire on Data Quality". In: *Survey Research Methods* 14.1. Number: 1, pp. 31–42. ISSN: 1864-3361. DOI: 10.18148/srm/2020.v14i1.7294. URL: https://ojs.ub.uni-konstanz.de/srm/article/view/7294 (visited on 11/20/2024).

Anthropic, AI (2024). *Claude 3.5 Sonnet Model Card Addendum*. URL: https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf (visited on 10/27/2024).

Appelbaum, Mark et al. (Jan. 2018). "Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report." In: *American Psychologist* 73.1, pp. 3–25. ISSN: 1935-990X, 0003-066X. DOI: 10.1037/amp0000191. URL: https://doi.apa.org/doi/10.1037/amp0000191 (visited on 11/20/2024).

Asai, Akari et al. (Mar. 5, 2024). *Reliable, Adaptable, and Attributable Language Models with Retrieval*. arXiv: 2403.03187. URL: http://arxiv.org/abs/2403.03187 (visited on 10/29/2024).

Balcombe, Luke (Dec. 2023). "AI Chatbots in Digital Mental Health". In: *Informatics* 10.4. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 82. ISSN: 2227-9709. DOI: 10.3390/informatics10040082. URL: https://www.mdpi.com/2227-9709/10/4/82 (visited on 10/26/2024).

Banerjee, Satanjeev and Alon Lavie (June 2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein et al. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://aclanthology.org/W05-0909 (visited on 11/08/2024).

Barnett, Scott et al. (Jan. 11, 2024). *Seven Failure Points When Engineering a Retrieval Augmented Generation System*. DOI: 10.48550/arXiv.2401.05856. arXiv: 2401.05856[cs]. URL: http://arxiv.org/abs/2401.05856 (visited on 01/15/2025).

Belz, Anya, Simon Mille, and David M. Howcroft (Dec. 2020). "Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing". In: *Proceedings of the 13th International Conference on Natural Language Generation*. INLG 2020. Ed.

by Brian Davis et al. Dublin, Ireland: Association for Computational Linguistics, pp. 183–194. DOI: 10.18653/v1/2020.inlg-1.24. URL: https://aclanthology.org/2020.inlg-1.24 (visited on 11/11/2024).
Belz, Anya et al. (Aug. 7, 2023). *Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP*. arXiv: 2305.01633. URL: http://arxiv.org/abs/2305.01633 (visited on 11/11/2024).
Borgeaud, Sebastian et al. (Feb. 7, 2022). *Improving language models by retrieving from trillions of tokens*. DOI: 10.48550/arXiv.2112.04426. arXiv: 2112.04426[cs]. URL: http://arxiv.org/abs/2112.04426 (visited on 01/16/2025).
Camilleri, Mark Anthony and Ciro Troise (Mar. 1, 2023). "Live support by chatbots with artificial intelligence: A future research agenda". In: *Service Business* 17.1, pp. 61–80. ISSN: 1862-8508. DOI: 10.1007/s11628-022-00513-9. URL: https://doi.org/10.1007/s11628-022-00513-9 (visited on 10/09/2024).
Chang, Yupeng et al. (Dec. 29, 2023). *A Survey on Evaluation of Large Language Models*. arXiv: 2307.03109. URL: http://arxiv.org/abs/2307.03109 (visited on 10/12/2024).
Chen, Canyu and Kai Shu (Nov. 9, 2023). *Combating Misinformation in the Age of LLMs: Opportunities and Challenges*. arXiv: 2311.05656. URL: http://arxiv.org/abs/2311.05656 (visited on 10/31/2024).
Chiang, Cheng-Han and Hung-yi Lee (May 3, 2023). *Can Large Language Models Be an Alternative to Human Evaluations?* arXiv: 2305.01937. URL: http://arxiv.org/abs/2305.01937 (visited on 11/06/2024).
Chung, Hyung Won et al. (Dec. 6, 2022). *Scaling Instruction-Finetuned Language Models*. arXiv: 2210.11416. URL: http://arxiv.org/abs/2210.11416 (visited on 11/03/2024).
Cohen, Roi et al. (Jan. 30, 2023). *Crawling the Internal Knowledge-Base of Language Models*. DOI: 10.48550/arXiv.2301.12810. arXiv: 2301.12810[cs]. URL: http://arxiv.org/abs/2301.12810 (visited on 01/16/2025).
Cuconasu, Florin et al. (May 1, 2024). *The Power of Noise: Redefining Retrieval for RAG Systems*. arXiv: 2401.14887. URL: http://arxiv.org/abs/2401.14887 (visited on 11/05/2024).
DeepSeek-AI et al. (Dec. 27, 2024). *DeepSeek-V3 Technical Report*. DOI: 10.48550/arXiv.2412.19437. arXiv: 2412.19437[cs]. URL: http://arxiv.org/abs/2412.19437 (visited on 01/28/2025).
DeepSeek-AI et al. (Jan. 22, 2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. DOI: 10.48550/arXiv.2501.12948. arXiv: 2501.12948[cs]. URL: http://arxiv.org/abs/2501.12948 (visited on 01/28/2025).
Deutsch, Daniel and Dan Roth (Nov. 2020). "SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics". In: *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. NLPOSS 2020. Ed. by Eunjeong L. Park et al. Online: Association for Computational Linguistics, pp. 120–125. DOI: 10.18653/v1/2020.nlposs-1.17. URL: https://aclanthology.org/2020.nlposs-1.17 (visited on 12/27/2024).
Devlin, Jacob et al. (May 24, 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 01/16/2025).
Dhingra, Bhuwan et al. (June 3, 2019). *Handling Divergent Reference Texts when Evaluating Table-to-Text Generation*. arXiv: 1906.01081. URL: http://arxiv.org/abs/1906.01081 (visited on 11/08/2024).

Dong, Yue et al. (Oct. 6, 2020). *Multi-Fact Correction in Abstractive Text Summarization*. arXiv: 2010.02443. URL: http://arxiv.org/abs/2010.02443 (visited on 11/01/2024).

Durall, Eva and Evangelos Kapros (2020). "Co-design for a Competency Self-assessment Chatbot and Survey in Science Education". In: *Learning and Collaboration Technologies. Human and Technology Ecosystems*. Ed. by Panayiotis Zaphiris and Andri Ioannou. Cham: Springer International Publishing, pp. 13–24. ISBN: 978-3-030-50506-6. DOI: 10.1007/978-3-030-50506-6_2.

Durmus, Esin, He He, and Mona Diab (July 2020). "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 5055–5070. DOI: 10.18653/v1/2020.acl-main.454. URL: https://aclanthology.org/2020.acl-main.454 (visited on 11/08/2024).

Elnozahy, Walaa A. et al. (2019). "Question Answering System to Support University Students' Orientation, Recruitment and Retention". In: *Procedia Computer Science* 164, pp. 56–63. ISSN: 18770509. DOI: 10.1016/j.procs.2019.12.154. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877050919321933 (visited on 10/26/2024).

Es, Shahul et al. (Sept. 26, 2023). *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. arXiv: 2309.15217. URL: http://arxiv.org/abs/2309.15217 (visited on 10/10/2024).

Fabbri, Alexander R. et al. (Apr. 26, 2021). "SummEval: Re-evaluating Summarization Evaluation". In: *Transactions of the Association for Computational Linguistics* 9, pp. 391–409. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00373. URL: https://doi.org/10.1162/tacl_a_00373 (visited on 01/07/2025).

Ferrante, M and M Saltalamacchia (2006). "The Coupon Collector's Problem". In: 1.

Folk, Dunigan, Stephanie Yu, and Elizabeth Dunn (May 10, 2024). "Can Chatbots Ever Provide More Social Connection Than Humans?" In: *Collabra: Psychology* 10.1. Ed. by Justin Matthews, p. 117083. ISSN: 2474-7394. DOI: 10.1525/collabra.117083. URL: https://doi.org/10.1525/collabra.117083 (visited on 10/29/2024).

Fu, Jinlan et al. (Feb. 13, 2023). *GPTScore: Evaluate as You Desire*. arXiv: 2302.04166. URL: http://arxiv.org/abs/2302.04166 (visited on 11/08/2024).

Gao, Mingqi and Xiaojun Wan (July 2022). "DialSummEval: Revisiting Summarization Evaluation for Dialogues". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 5693–5709. DOI: 10.18653/v1/2022.naacl-main.418. URL: https://aclanthology.org/2022.naacl-main.418/ (visited on 01/07/2025).

Gao, Tianyu et al. (Oct. 31, 2023). *Enabling Large Language Models to Generate Text with Citations*. arXiv: 2305.14627. URL: http://arxiv.org/abs/2305.14627 (visited on 11/02/2024).

Gao, Yunfan et al. (Mar. 27, 2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. DOI: 10.48550/arXiv.2312.10997. arXiv: 2312.10997. URL: http://arxiv.org/abs/2312.10997 (visited on 10/31/2024).

Gehrmann, Sebastian, Elizabeth Clark, and Thibault Sellam (Feb. 14, 2022). *Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated*

Text. arXiv: 2202.06935. URL: http://arxiv.org/abs/2202.06935 (visited on 11/11/2024).

Gemini Team et al. (Aug. 8, 2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. DOI: 10.48550/arXiv.2403.05530. arXiv: 2403.05530. URL: http://arxiv.org/abs/2403.05530 (visited on 10/27/2024).

Goyal, Tanya, Junyi Jessy Li, and Greg Durrett (May 23, 2023). *News Summarization and Evaluation in the Era of GPT-3*. DOI: 10.48550/arXiv.2209.12356. arXiv: 2209.12356[cs]. URL: http://arxiv.org/abs/2209.12356 (visited on 01/07/2025).

Grusky, Max (July 2023). "Rogue Scores". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1914–1934. DOI: 10.18653/v1/2023.acl-long.107. URL: https://aclanthology.org/2023.acl-long.107 (visited on 12/26/2024).

Gu, Jiawei et al. (Jan. 9, 2025). *A Survey on LLM-as-a-Judge*. DOI: 10.48550/arXiv.2411.15594. arXiv: 2411.15594[cs]. URL: http://arxiv.org/abs/2411.15594 (visited on 01/15/2025).

Guerreiro, Nuno M., Elena Voita, and André F. T. Martins (Mar. 5, 2023). *Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation*. arXiv: 2208.05309. URL: http://arxiv.org/abs/2208.05309 (visited on 11/08/2024).

Guo, Zishan et al. (Nov. 25, 2023). *Evaluating Large Language Models: A Comprehensive Survey*. arXiv: 2310.19736. URL: http://arxiv.org/abs/2310.19736 (visited on 11/07/2024).

Hariri, Walid (Aug. 3, 2024). *Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing*. DOI: 10.48550/arXiv.2304.02017. arXiv: 2304.02017[cs]. URL: http://arxiv.org/abs/2304.02017 (visited on 01/25/2025).

Hayes, William M., Nicolas Yax, and Stefano Palminteri (May 19, 2024). *Large Language Models are Biased Reinforcement Learners*. arXiv: 2405.11422. URL: http://arxiv.org/abs/2405.11422 (visited on 10/29/2024).

Hien, Ho Thao et al. (2018). "Intelligent Assistants in Higher-Education Environments: The FIT-EBot, a Chatbot for Administrative and Learning Support". In: *Proceedings of the Ninth International Symposium on Information and Communication Technology - SoICT 2018*. the Ninth International Symposium. Danang City, Viet Nam: ACM Press, pp. 69–76. ISBN: 978-1-4503-6539-0. DOI: 10.1145/3287921.3287937. URL: http://dl.acm.org/citation.cfm?doid=3287921.3287937 (visited on 10/26/2024).

Honovich, Or et al. (Sept. 9, 2021). Q^2 : $Evaluating Factual Consistency in Knowledge - Grounded Dialogues via Question Generation and Question Answering$. arXiv: 2104.08202. URL: http://arxiv.org/abs/2104.08202 (visited on 11/08/2024).

Hu, Xuming et al. (Oct. 8, 2023). *Do Large Language Models Know about Facts?* arXiv: 2310.05177. URL: http://arxiv.org/abs/2310.05177 (visited on 10/30/2024).

Huang, Lei et al. (Nov. 9, 2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. arXiv: 2311.05232. URL: http://arxiv.org/abs/2311.05232 (visited on 10/29/2024).

Huang, Yizheng and Jimmy Huang (Aug. 23, 2024). *A Survey on Retrieval-Augmented Text Generation for Large Language Models*. arXiv: 2404.10981. URL: http://arxiv.org/abs/2404.10981 (visited on 11/02/2024).

Hughes, Jennifer L. et al. (2022). “INVITED EDITORIAL: Guidance for Researchers When Using Inclusive Demographic Questions for Surveys: Improved and Updated Questions”. In: *Psi Chi Journal of Psychological Research* 27.4, pp. 232–255. ISSN: 23257342. DOI: 10.24839/2325-7342.JN27.4.232. URL: https://www.psichi.org/page/274JNWinter2022 (visited on 11/16/2024).

Jeong, Soyeong et al. (Mar. 28, 2024). *Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity*. arXiv: 2403.14403. URL: http://arxiv.org/abs/2403.14403 (visited on 11/06/2024).

Ji, Ziwei et al. (Oct. 10, 2023). *Towards Mitigating Hallucination in Large Language Models via Self-Reflection*. arXiv: 2310.06271. URL: http://arxiv.org/abs/2310.06271 (visited on 10/31/2024).

Ji, Ziwei et al. (July 14, 2024). *Survey of Hallucination in Natural Language Generation*. arXiv: 2202.03629. URL: http://arxiv.org/abs/2202.03629 (visited on 10/10/2024).

Jiang, Zhengbao et al. (Oct. 22, 2023). *Active Retrieval Augmented Generation*. arXiv: 2305.06983. URL: http://arxiv.org/abs/2305.06983 (visited on 11/02/2024).

Khalifa, Mohamed and Mona Albadawy (2024). “Using artificial intelligence in academic writing and research: An essential productivity tool”. In: *Computer Methods and Programs in Biomedicine Update* 5, p. 100145. ISSN: 26669900. DOI: 10.1016/j.cmpbup.2024.100145. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666990024000120 (visited on 01/23/2025).

Khandelwal, Urvashi et al. (Feb. 15, 2020). *Generalization through Memorization: Nearest Neighbor Language Models*. DOI: 10.48550/arXiv.1911.00172. arXiv: 1911.00172[cs]. URL: http://arxiv.org/abs/1911.00172 (visited on 01/16/2025).

Khattab, Omar et al. (Jan. 23, 2023). *Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP*. DOI: 10.48550/arXiv.2212.14024. arXiv: 2212.14024[cs]. URL: http://arxiv.org/abs/2212.14024 (visited on 01/16/2025).

Krippendorff, Klaus (1980). “Content Analysis: An Introduction to Its Methodology”. In: URL: https://api.semanticscholar.org/CorpusID:62392461.

Krishna, Kalpesh et al. (Jan. 30, 2023). *LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization*. arXiv: 2301.13298. URL: http://arxiv.org/abs/2301.13298 (visited on 11/09/2024).

Lee, Minhyeok (June 12, 2023). *On the Amplification of Linguistic Bias through Unintentional Self-reinforcement Learning by Generative Language Models – A Perspective*. arXiv: 2306.07135. URL: http://arxiv.org/abs/2306.07135 (visited on 10/29/2024).

Leng, Quinn, Kasey Uhlenhuth, and Alkis Polyzotis (Sept. 12, 2023). *Best Practices for LLM Evaluation of RAG Applications*. Databricks. URL: https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG (visited on 01/03/2025).

Lewis, Mike et al. (Oct. 29, 2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. DOI: 10.48550/arXiv.1910.13461. arXiv: 1910.13461[cs]. URL: http://arxiv.org/abs/1910.13461 (visited on 01/16/2025).

Lewis, Patrick et al. (2020). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (visited on 10/09/2024).

Li, Junyi et al. (Jan. 6, 2024). *The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models*. arXiv: 2401.03205. URL: http://arxiv.org/abs/2401.03205 (visited on 10/29/2024).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://aclanthology.org/W04-1013 (visited on 11/08/2024).

Lin, Yu-Hong and Tienwei Tsai (Dec. 2019). "A Conversational Assistant on Mobile Devices for Primitive Learners of Computer Programming". In: *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. 2019 IEEE International Conference on Engineering, Technology and Education (TALE). ISSN: 2470-6698, pp. 1–4. DOI: 10.1109/TALE48000.2019.9226015. URL: https://ieeexplore.ieee.org/document/9226015/?arnumber=9226015 (visited on 10/26/2024).

Liu, Nelson F. et al. (Nov. 20, 2023a). *Lost in the Middle: How Language Models Use Long Contexts*. DOI: 10.48550/arXiv.2307.03172. arXiv: 2307.03172. URL: http://arxiv.org/abs/2307.03172 (visited on 10/31/2024).

Liu, Yang et al. (May 23, 2023b). *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. arXiv: 2303.16634. URL: http://arxiv.org/abs/2303.16634 (visited on 10/15/2024).

Ma, Xinbei et al. (Oct. 23, 2023). *Query Rewriting for Retrieval-Augmented Large Language Models*. arXiv: 2305.14283. URL: http://arxiv.org/abs/2305.14283 (visited on 11/02/2024).

Macpherson, Fiona and Dimitris Platchias (Aug. 30, 2013). *Hallucination: Philosophy and Psychology*. Google-Books-ID: _bwtAAAAQBAJ. MIT Press. 433 pp. ISBN: 978-0-262-01920-0.

Mallen, Alex et al. (July 2, 2023). *When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories*. arXiv: 2212.10511. URL: http://arxiv.org/abs/2212.10511 (visited on 11/04/2024).

Maynez, Joshua et al. (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173 (visited on 10/30/2024).

Mckie, Indra Ayu Susan and Bhuva Narayan (July 3, 2019). "Enhancing the Academic Library Experience with Chatbots: An Exploration of Research and Implications for Practice". In: *Journal of the Australian Library and Information Association* 68.3, pp. 268–277. ISSN: 2475-0158. DOI: 10.1080/24750158.2019.1611694. URL: https://doi.org/10.1080/24750158.2019.1611694 (visited on 10/26/2024).

Messeri, Lisa and M. J. Crockett (Mar. 2024). "Artificial intelligence and illusions of understanding in scientific research". In: *Nature* 627.8002. Publisher: Nature Publishing Group, pp. 49–58. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07146-0. URL: https://www.nature.com/articles/s41586-024-07146-0 (visited on 01/23/2025).

Mirzadeh, Iman et al. (Oct. 7, 2024). *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. arXiv: 2410.05229. URL: http://arxiv.org/abs/2410.05229 (visited on 11/05/2024).

Mishra, Abhika et al. (Aug. 12, 2024). *Fine-grained Hallucination Detection and Editing for Language Models*. arXiv: 2401.06855. URL: http://arxiv.org/abs/2401.06855 (visited on 10/29/2024).

Niu, Cheng et al. (May 17, 2024). *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*. arXiv: 2401.00396. URL: http://arxiv.org/abs/2401.00396 (visited on 11/01/2024).

OpenAI et al. (Mar. 4, 2024a). *GPT-4 Technical Report*. arXiv: 2303.08774. URL: http://arxiv.org/abs/2303.08774 (visited on 10/27/2024).

OpenAI et al. (Oct. 25, 2024b). *GPT-4o System Card*. version: 1. DOI: 10.48550/arXiv.2410.21276. arXiv: 2410.21276[cs]. URL: http://arxiv.org/abs/2410.21276 (visited on 01/04/2025).

Ouyang, Long et al. (Dec. 6, 2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf (visited on 10/31/2024).

Ouyang, Shuyin et al. (Oct. 17, 2024). *An Empirical Study of the Non-determinism of ChatGPT in Code Generation*. arXiv: 2308.02828. URL: http://arxiv.org/abs/2308.02828 (visited on 11/01/2024).

Ovadia, Oded et al. (Jan. 30, 2024). *Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs*. arXiv: 2312.05934. URL: http://arxiv.org/abs/2312.05934 (visited on 11/03/2024).

Pan, Yikang et al. (Oct. 26, 2023). *On the Risk of Misinformation Pollution with Large Language Models*. arXiv: 2305.13661. URL: http://arxiv.org/abs/2305.13661 (visited on 10/31/2024).

Panickssery, Arjun, Samuel R. Bowman, and Shi Feng (Apr. 15, 2024). *LLM Evaluators Recognize and Favor Their Own Generations*. arXiv: 2404.13076. URL: http://arxiv.org/abs/2404.13076 (visited on 11/09/2024).

Papineni, Kishore et al. (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. ACL 2002. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040 (visited on 11/08/2024).

Petroni, Fabio et al. (Sept. 4, 2019). *Language Models as Knowledge Bases?* DOI: 10.48550/arXiv.1909.01066. arXiv: 1909.01066[cs]. URL: http://arxiv.org/abs/1909.01066 (visited on 01/16/2025).

Poel, Liam van der, Ryan Cotterell, and Clara Meister (Oct. 29, 2022). *Mutual Information Alleviates Hallucinations in Abstractive Summarization*. arXiv: 2210.13210. URL: http://arxiv.org/abs/2210.13210 (visited on 11/08/2024).

Post, Matt (Sept. 12, 2018). *A Call for Clarity in Reporting BLEU Scores*. DOI: 10.48550/arXiv.1804.08771. arXiv: 1804.08771[cs]. URL: http://arxiv.org/abs/1804.08771 (visited on 12/28/2024).

Prabhakaran, Vinodkumar, Aida Mostafazadeh Davani, and Mark Díaz (Oct. 12, 2021). *On Releasing Annotator-Level Labels and Information in Datasets*. DOI: 10.48550/arXiv.2110.05699. arXiv: 2110.05699[cs]. URL: http://arxiv.org/abs/2110.05699 (visited on 01/22/2025).

Ram, Ori et al. (Aug. 1, 2023). *In-Context Retrieval-Augmented Language Models*. arXiv: 2302.00083. URL: http://arxiv.org/abs/2302.00083 (visited on 11/04/2024).

Saad-Falcon, Jon et al. (Mar. 31, 2024). *ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems*. arXiv: 2311.09476. URL: http://arxiv.org/abs/2311.09476 (visited on 11/02/2024).

Sandu, Nitirajsingh and Ergun Gide (Sept. 2019). "Adoption of AI-Chatbots to Enhance Student Learning Experience in Higher Education in India". In: *2019*

18th International Conference on Information Technology Based Higher Education and Training (ITHET). 2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET). ISSN: 2380-1603, pp. 1–5. DOI: 10.1109/ITHET46829.2019.8937382. URL: https://ieeexplore.ieee.org/document/8937382/?arnumber=8937382 (visited on 10/26/2024).

Sellam, Thibault, Dipanjan Das, and Ankur P. Parikh (May 21, 2020). *BLEURT: Learning Robust Metrics for Text Generation*. arXiv: 2004.04696. URL: http://arxiv.org/abs/2004.04696 (visited on 11/08/2024).

Shi, Freda et al. (June 6, 2023). *Large Language Models Can Be Easily Distracted by Irrelevant Context*. DOI: 10.48550/arXiv.2302.00093. arXiv: 2302.00093. URL: http://arxiv.org/abs/2302.00093 (visited on 11/01/2024).

Shimorina, Anastasia and Anya Belz (May 2022). "The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP". In: *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*. HumEval 2022. Ed. by Anya Belz et al. Dublin, Ireland: Association for Computational Linguistics, pp. 54–75. DOI: 10.18653/v1/2022.humeval-1.6. URL: https://aclanthology.org/2022.humeval-1.6 (visited on 11/11/2024).

Shuster, Kurt et al. (Apr. 15, 2021). *Retrieval Augmentation Reduces Hallucination in Conversation*. arXiv: 2104.07567. URL: http://arxiv.org/abs/2104.07567 (visited on 11/01/2024).

Si, Chenglei et al. (Apr. 1, 2024). *Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong*. arXiv: 2310.12558. URL: http://arxiv.org/abs/2310.12558 (visited on 10/31/2024).

Sivasothy, Shangeetha et al. (Sept. 24, 2024). *RAGProbe: An Automated Approach for Evaluating RAG Applications*. arXiv: 2409.19019. URL: http://arxiv.org/abs/2409.19019 (visited on 11/11/2024).

Stoilova, Elitza (Dec. 16, 2021). "AI chatbots as a customer service and support tool". In: *ROBONOMICS: The Journal of the Automated Economy* 2, pp. 21–21. ISSN: 2683-099X. URL: https://www.journal.robonomics.science/index.php/rj/article/view/21 (visited on 10/09/2024).

Tian, Katherine et al. (Nov. 14, 2023). *Fine-tuning Language Models for Factuality*. arXiv: 2311.08401. URL: http://arxiv.org/abs/2311.08401 (visited on 11/03/2024).

Varshney, Neeraj et al. (Aug. 12, 2023). *A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation*. arXiv: 2307.03987. URL: http://arxiv.org/abs/2307.03987 (visited on 11/08/2024).

Wang, Cunxiang et al. (Dec. 16, 2023a). *Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity*. arXiv: 2310.07521[cs]. URL: http://arxiv.org/abs/2310.07521 (visited on 10/11/2024).

Wang, Fei Yi and Stephen Petrina (2013). "Using Learning Analytics to Understand the Design of an Intelligent Language Tutor – Chatbot Lucy". In: *International Journal of Advanced Computer Science and Applications* 4.11. ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2013.041117. URL: http://thesai.org/Publications/ViewPaper?Volume=4&Issue=11&Code=IJACSA&SerialNo=17 (visited on 10/26/2024).

Wang, Jiaan et al. (Oct. 24, 2023b). *Is ChatGPT a Good NLG Evaluator? A Preliminary Study*. arXiv: 2303.04048. URL: http://arxiv.org/abs/2303.04048 (visited on 11/08/2024).

Wang, Peiyi et al. (Aug. 30, 2023c). *Large Language Models are not Fair Evaluators*. arXiv: 2305.17926. URL: http://arxiv.org/abs/2305.17926 (visited on 11/09/2024).

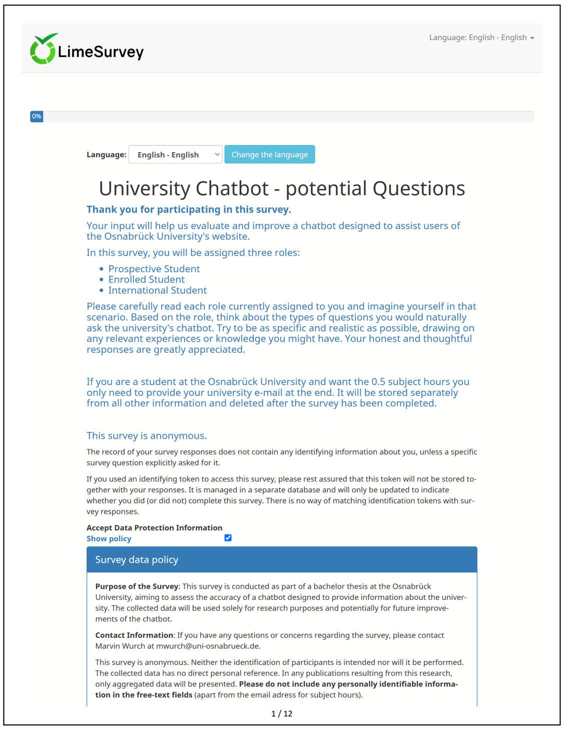
Weizenbaum, Joseph (Jan. 1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1, pp. 36–45. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/365153.365168. URL: https://dl.acm.org/doi/10.1145/365153.365168 (visited on 10/27/2024).
Weller, Orion et al. (Feb. 26, 2024). *"According to ...": Prompting Language Models Improves Quoting from Pre-Training Data*. arXiv: 2305.13252. URL: http://arxiv.org/abs/2305.13252 (visited on 10/26/2024).
Wu, Yuhan (Mar. 5, 2024). "Exploring the Influence of Large Language Models (LLMs) on English Learners and their Teachers". In: *Journal of Education, Humanities and Social Sciences* 27, pp. 530–535. DOI: 10.54097/zghke663.
Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli (Jan. 22, 2024). *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. arXiv: 2401.11817. URL: http://arxiv.org/abs/2401.11817 (visited on 10/30/2024).
Yan, Shi-Qi et al. (Oct. 7, 2024). *Corrective Retrieval Augmented Generation*. arXiv: 2401.15884. URL: http://arxiv.org/abs/2401.15884 (visited on 11/05/2024).
Yang, Jing et al. (Jan. 2023). "A Systematic Literature Review of Information Security in Chatbots". In: *Applied Sciences* 13.11. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 6355. ISSN: 2076-3417. DOI: 10.3390/app13116355. URL: https://www.mdpi.com/2076-3417/13/11/6355 (visited on 10/29/2024).
Yuan, Weizhe, Graham Neubig, and Pengfei Liu (Oct. 27, 2021). *BARTScore: Evaluating Generated Text as Text Generation*. arXiv: 2106.11520. URL: http://arxiv.org/abs/2106.11520 (visited on 11/08/2024).
Zar, Jerrold H. (2005). "Spearman Rank Correlation". In: *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd. ISBN: 978-0-470-01181-2. DOI: 10.1002/0470011815.b2a15150. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15150 (visited on 01/22/2025).
Zhang, Tianyi et al. (Feb. 24, 2020). *BERTScore: Evaluating Text Generation with BERT*. arXiv: 1904.09675. URL: http://arxiv.org/abs/1904.09675 (visited on 11/08/2024).
Zhang, Yue et al. (Sept. 24, 2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. arXiv: 2309.01219. URL: http://arxiv.org/abs/2309.01219 (visited on 10/29/2024).
Zhao, Penghao et al. (June 21, 2024). *Retrieval-Augmented Generation for AI-Generated Content: A Survey*. arXiv: 2402.19473. URL: http://arxiv.org/abs/2402.19473 (visited on 11/02/2024).
Zheng, Lianmin et al. (Dec. 24, 2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. arXiv: 2306.05685. URL: http://arxiv.org/abs/2306.05685 (visited on 10/10/2024).
Zhu, Kunlun et al. (Oct. 17, 2024). *RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework*. DOI: 10.48550/arXiv.2408.01262. arXiv: 2408.01262. URL: http://arxiv.org/abs/2408.01262 (visited on 11/06/2024).
Zumstein, Darius and Sophie Hundertmark (Nov. 15, 2017). "Chatbots – An Interactive Technology for Personalized Communication, Transactions and Services". In: *IADIS International Journal on WWW/Internet* 15, pp. 96–109.

Appendix A

Dataset

A.1 Pre-Survey Question Generation

The following pages present a Pre-Survey submission from a randomly selected participant, showcasing the type of questions collected and the survey structure. Participant feedback was obtained through a brief oral interview conducted after the survey.


LimeSurvey

Language: English - English

0%

Language: English - English | Change the language

University Chatbot - potential Questions

Thank you for participating in this survey.

Your input will help us evaluate and improve a chatbot designed to assist users of the Osnabrück University's website.

In this survey, you will be assigned three roles:

- Prospective Student
- Enrolled Student
- International Student

Please carefully read each role currently assigned to you and imagine yourself in that scenario. Based on the role, think about the types of questions you would naturally ask the university's chatbot. Try to be as specific and realistic as possible, drawing on any relevant experiences or knowledge you might have. Your honest and thoughtful responses are greatly appreciated.

If you are a student at the Osnabrück University and want the 0.5 subject hours you only need to provide your university e-mail at the end. It will be stored separately from all other information and deleted after the survey has been completed.

This survey is anonymous.

The record of your survey responses does not contain any identifying information about you, unless a specific survey question explicitly asked for it.

If you used an identifying token to access this survey, please rest assured that this token will not be stored together with your responses. It is managed in a separate database and will only be updated to indicate whether you did (or did not) complete this survey. There is no way of matching identification tokens with survey responses.

Accept Data Protection Information
Show policy

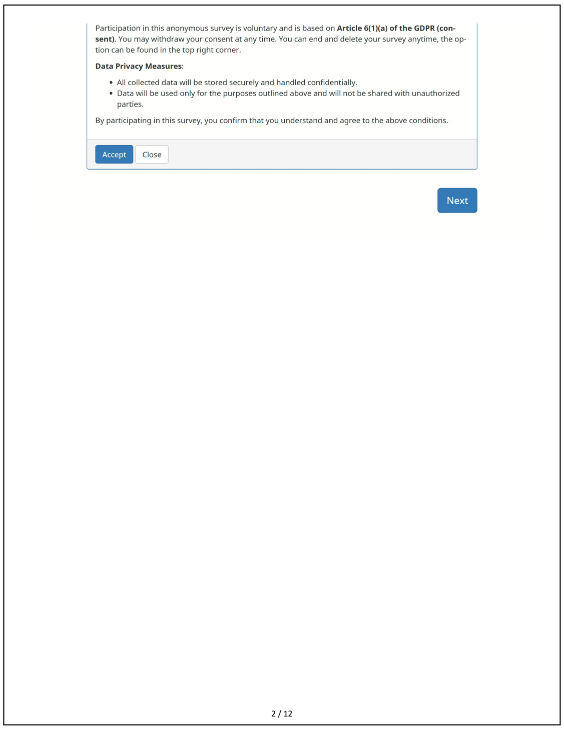
Survey data policy

Purpose of the Survey: This survey is conducted as part of a bachelor thesis at the Osnabrück University, aiming to assess the accuracy of a chatbot designed to provide information about the university. The collected data will be used solely for research purposes and potentially for future improvements of the chatbot.

Contact Information: If you have any questions or concerns regarding the survey, please contact Marvin Wurch at mwurch@uni-osnabrueck.de.

This survey is anonymous. Neither the identification of participants is intended nor will it be performed. The collected data has no direct personal reference. In any publications resulting from this research, only aggregated data will be presented. **Please do not include any personally identifiable information in the free-text fields** (apart from the email adress for subject hours).

1 / 12


Participation in this anonymous survey is voluntary and is based on **Article 6(1)(a) of the GDPR (consent)**. You may withdraw your consent at any time. You can end and delete your survey anytime, the option can be found in the top right corner.

Data Privacy Measures:

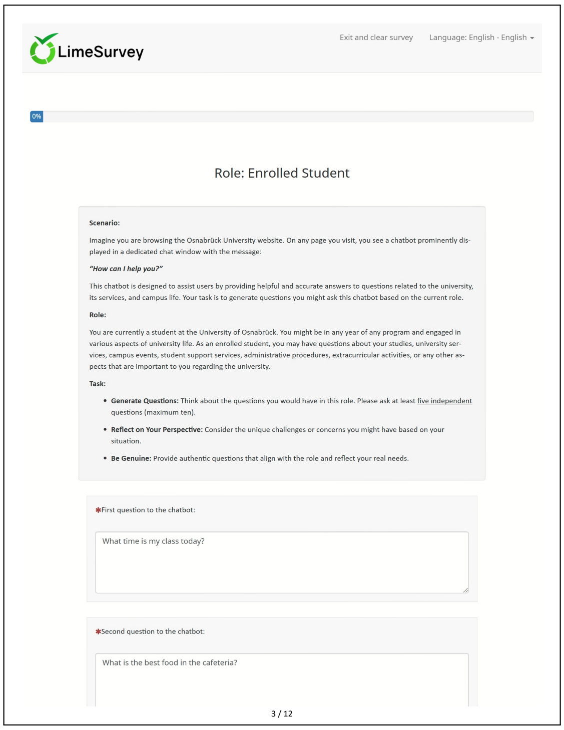
- All collected data will be stored securely and handled confidentially.
- Data will be used only for the purposes outlined above and will not be shared with unauthorized parties.

By participating in this survey, you confirm that you understand and agree to the above conditions.

Accept | Close

Next

2 / 12


LimeSurvey

Exit and clear survey | Language: English - English

0%

Role: Enrolled Student

Scenario:

Imagine you are browsing the Osnabrück University website. On any page you visit, you see a chatbot prominently displayed in a dedicated chat window with the message:

"How can I help you?"

This chatbot is designed to assist users by providing helpful and accurate answers to questions related to the university, its services, and campus life. Your task is to generate questions you might ask this chatbot based on the current role.

Role:

You are currently a student at the University of Osnabrück. You might be in any year of any program and engaged in various aspects of university life. As an enrolled student, you may have questions about your studies, university services, campus events, student support services, administrative procedures, extracurricular activities, or any other aspects that are important to you regarding the university.

Task:

- **Generate Questions:** Think about the questions you would have in this role. Please ask at least five independent questions (maximum ten).
- **Reflect on Your Perspective:** Consider the unique challenges or concerns you might have based on your situation.
- **Be Genuine:** Provide authentic questions that align with the role and reflect your real needs.

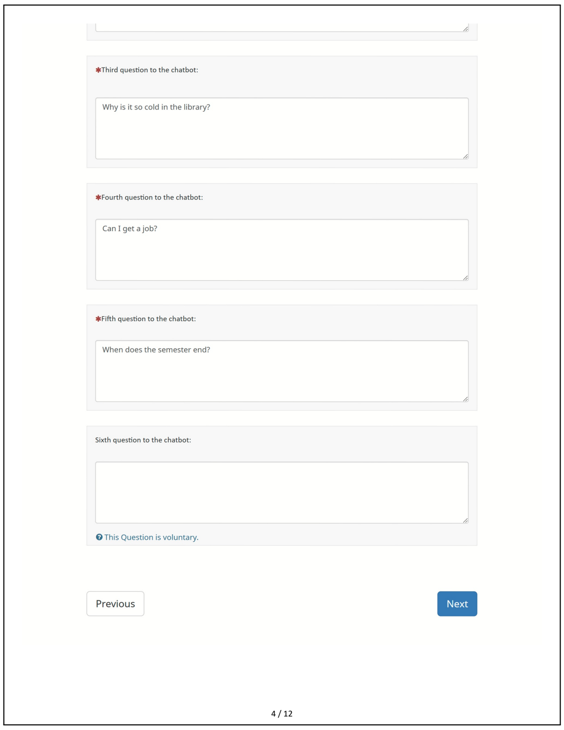
*First question to the chatbot:

What time is my class today?

*Second question to the chatbot:

What is the best food in the cafeteria?

3 / 12


*Third question to the chatbot:

Why is it so cold in the library?

*Fourth question to the chatbot:

Can I get a job?

*Fifth question to the chatbot:

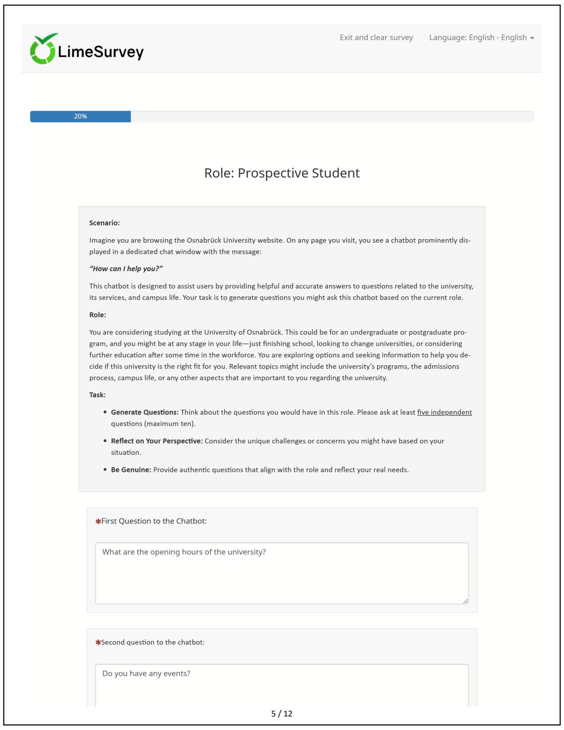
When does the semester end?

Sixth question to the chatbot:

This Question is voluntary.

Previous | Next

4 / 12


LimeSurvey
Exit and clear survey Language: English - English

20%

Role: Prospective Student

Scenario:

Imagine you are browsing the Osnabrück University website. On any page you visit, you see a chatbot prominently displayed in a dedicated chat window with the message:

"How can I help you?"

This chatbot is designed to assist users by providing helpful and accurate answers to questions related to the university, its services, and campus life. Your task is to generate questions you might ask this chatbot based on the current role.

Role:

You are considering studying at the University of Osnabrück. This could be for an undergraduate or postgraduate program, and you might be at any stage in your life—just finishing school, looking to change universities, or considering further education after some time in the workforce. You are exploring options and seeking information to help you decide if this university is the right fit for you. Relevant topics might include the university's programs, the admissions process, campus life, or any other aspects that are important to you regarding the university.

Task:

- **Generate Questions:** Think about the questions you would have in this role. Please ask at least five independent questions (maximum ten).
- **Reflect on Your Perspective:** Consider the unique challenges or concerns you might have based on your situation.
- **Be Genuine:** Provide authentic questions that align with the role and reflect your real needs.

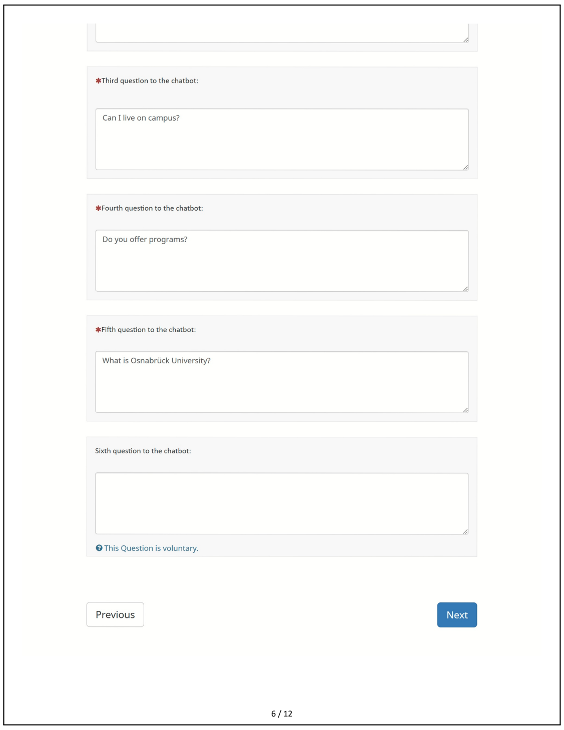
*First Question to the Chatbot:

What are the opening hours of the university?

*Second question to the chatbot:

Do you have any events?

5 / 12


*Third question to the chatbot:

Can I live on campus?

*Fourth question to the chatbot:

Do you offer programs?

*Fifth question to the chatbot:

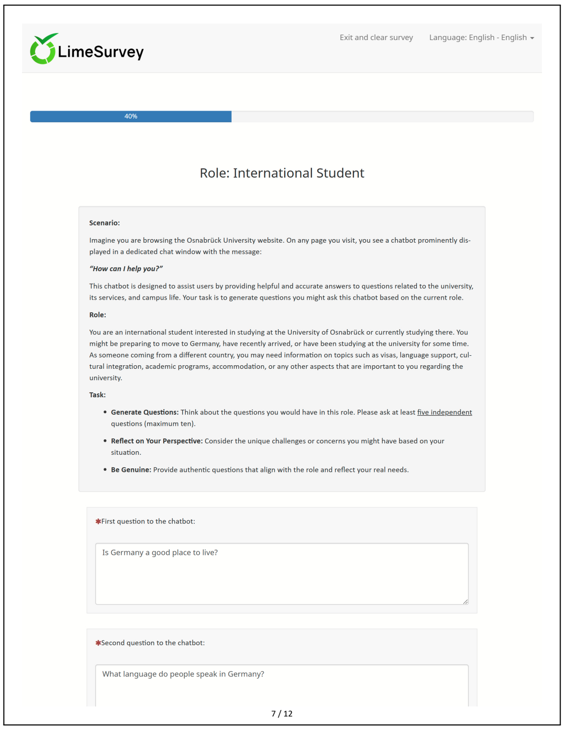
What is Osnabrück University?

Sixth question to the chatbot:

This Question is voluntary.

Previous Next

6 / 12


LimeSurvey
Exit and clear survey Language: English - English

40%

Role: International Student

Scenario:

Imagine you are browsing the Osnabrück University website. On any page you visit, you see a chatbot prominently displayed in a dedicated chat window with the message:

"How can I help you?"

This chatbot is designed to assist users by providing helpful and accurate answers to questions related to the university, its services, and campus life. Your task is to generate questions you might ask this chatbot based on the current role.

Role:

You are an international student interested in studying at the University of Osnabrück or currently studying there. You might be preparing to move to Germany, have recently arrived, or have been studying at the university for some time. As someone coming from a different country, you may need information on topics such as visas, language support, cultural integration, academic programs, accommodation, or any other aspects that are important to you regarding the university.

Task:

- **Generate Questions:** Think about the questions you would have in this role. Please ask at least five independent questions (maximum ten).
- **Reflect on Your Perspective:** Consider the unique challenges or concerns you might have based on your situation.
- **Be Genuine:** Provide authentic questions that align with the role and reflect your real needs.

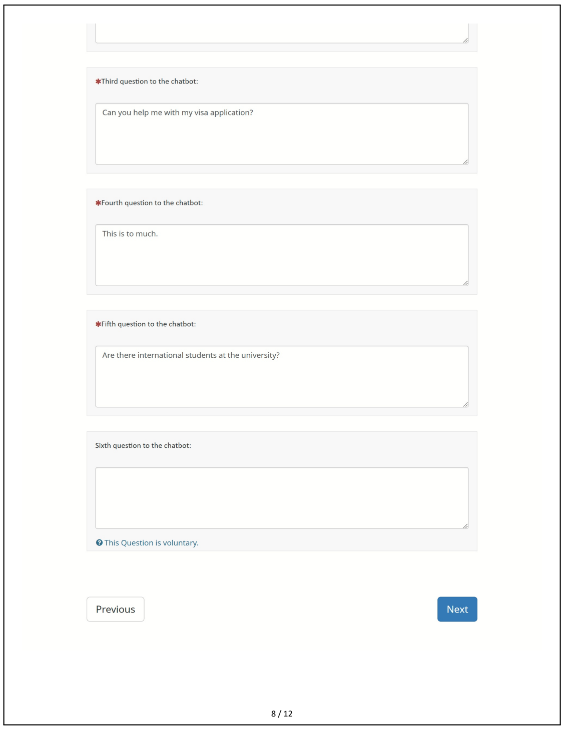
*First question to the chatbot:

Is Germany a good place to live?

*Second question to the chatbot:

What language do people speak in Germany?

7 / 12


*Third question to the chatbot:

Can you help me with my visa application?

*Fourth question to the chatbot:

This is to much.

*Fifth question to the chatbot:

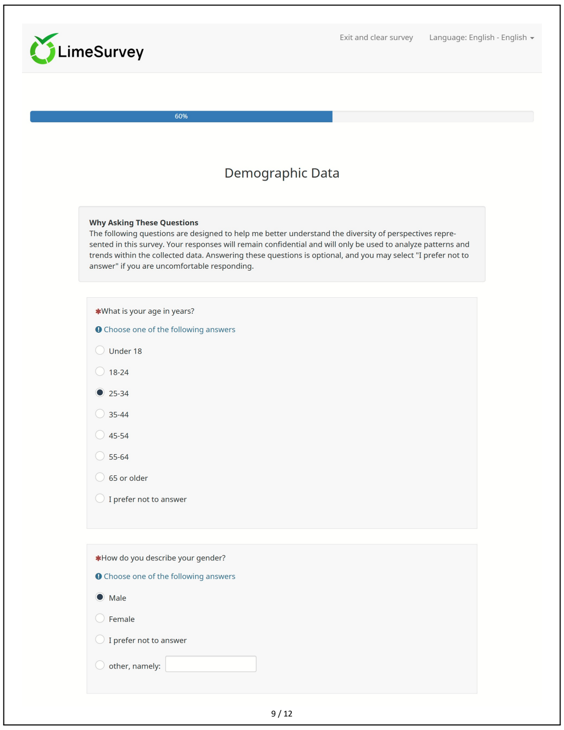
Are there international students at the university?

Sixth question to the chatbot:

This Question is voluntary.

Previous Next

8 / 12


LimeSurvey
Exit and clear survey Language: English - English
60%

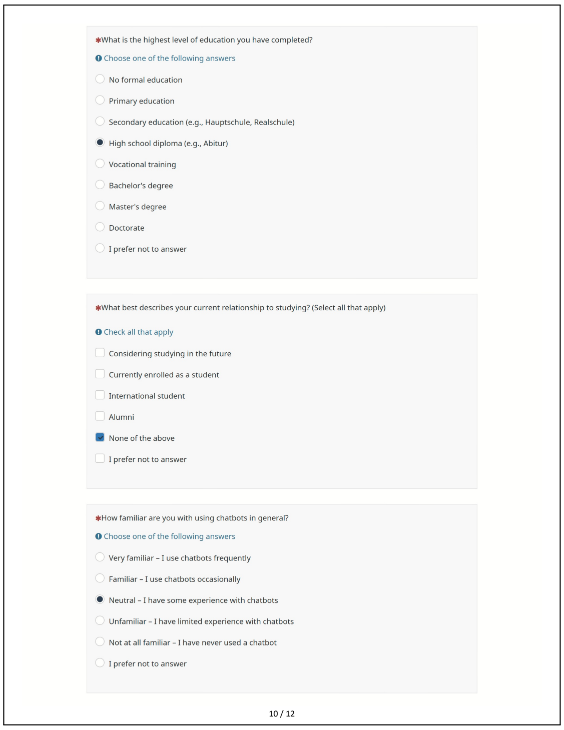
Demographic Data

Why Asking These Questions
The following questions are designed to help me better understand the diversity of perspectives represented in this survey. Your responses will remain confidential and will only be used to analyze patterns and trends within the collected data. Answering these questions is optional, and you may select "I prefer not to answer" if you are uncomfortable responding.

*What is your age in years?
Choose one of the following answers
- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 or older
- I prefer not to answer

*How do you describe your gender?
Choose one of the following answers
- Male
- Female
- I prefer not to answer
- other, namely:

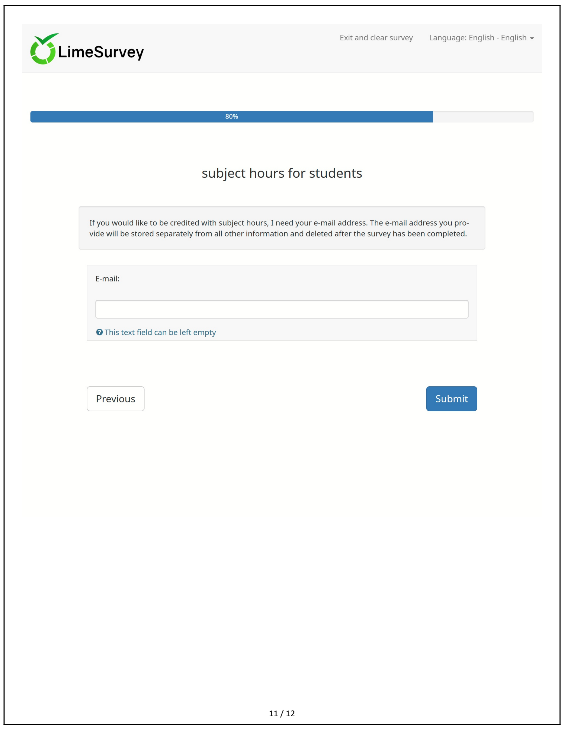
9 / 12


*What is the highest level of education you have completed?
Choose one of the following answers
- No formal education
- Primary education
- Secondary education (e.g., Hauptschule, Realschule)
- High school diploma (e.g., Abitur)
- Vocational training
- Bachelor's degree
- Master's degree
- Doctorate
- I prefer not to answer

*What best describes your current relationship to studying? (Select all that apply)
Check all that apply
- Considering studying in the future
- Currently enrolled as a student
- International student
- Alumni
- None of the above
- I prefer not to answer

*How familiar are you with using chatbots in general?
Choose one of the following answers
- Very familiar – I use chatbots frequently
- Familiar – I use chatbots occasionally
- Neutral – I have some experience with chatbots
- Unfamiliar – I have limited experience with chatbots
- Not at all familiar – I have never used a chatbot
- I prefer not to answer

10 / 12


LimeSurvey
Exit and clear survey Language: English - English
80%

subject hours for students

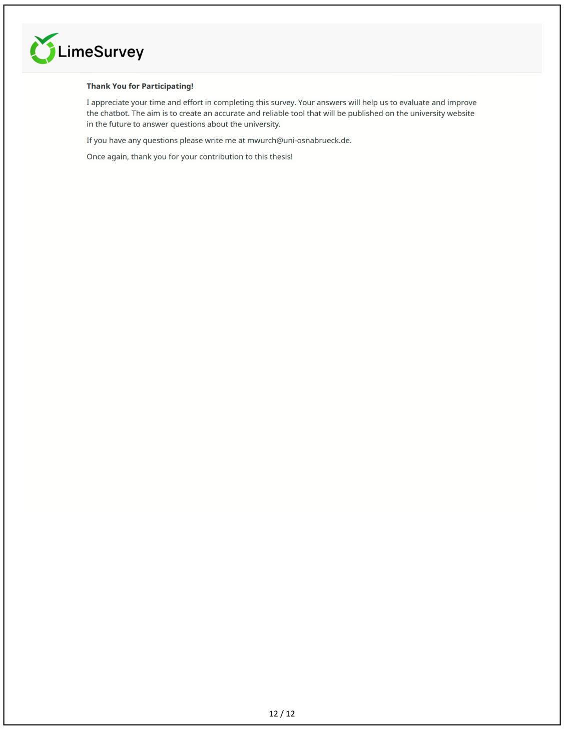
If you would like to be credited with subject hours, I need your e-mail address. The e-mail address you provide will be stored separately from all other information and deleted after the survey has been completed.

E-mail:

This text field can be left empty

Previous Submit

11 / 12


LimeSurvey

Thank You for Participating!
I appreciate your time and effort in completing this survey. Your answers will help us to evaluate and improve the chatbot. The aim is to create an accurate and reliable tool that will be published on the university website in the future to answer questions about the university.

If you have any questions please write me at mwurch@uni-osnabrueck.de.

Once again, thank you for your contribution to this thesis!

12 / 12

A.2 Question Generation Survey

A.2.1 Full Question Generation Survey

The following pages present the Question Generation Survey submission from a randomly selected participant (Participant 17). This example illustrates the types of questions generated during the survey and the complete survey structure. The participant's responses are displayed exactly as provided, without any modifications. For privacy reasons, the email address has been masked.

LimeSurvey

Exit and clear survey   Language: English - English

0%

Language: English - English   Change the language

University Chatbot evaluation Dataset - Questions

Thank you for participating in this survey.

Your input will help us evaluate and improve a chatbot designed to assist users of the Osnabrück University's website. The chatbot is currently under development by virtUOS and will potentially be integrated into the university's website and other university services in the future. Before this happens, the chatbot needs to be evaluated, and the first step is to create a domain-specific dataset.

Please carefully read the scenario and come up with different questions you would naturally ask the university's chatbot. Try to be as specific and realistic as possible, drawing on any relevant experiences, knowledge or needs you might have. Your honest and thoughtful responses are greatly appreciated.

If you are a student of Cognitive Science and want the **0.5 subject hours**, you only need to provide your university E-mail at the end. It will be stored separately from all other information and deleted after the survey has been completed.

This survey is anonymous.

The record of your survey responses does not contain any identifying information about you, unless a specific survey question explicitly asked for it.

If you used an identifying token to access this survey, please rest assured that this token will not be stored together with your responses. It is managed in a separate database and will only be updated to indicate whether you did (or did not) complete this survey. There is no way of matching identification tokens with survey responses.

Accept Data Privacy Information
Show policy ☑

Survey data policy

Purpose of the Survey: This survey is conducted as part of a bachelor thesis at the Osnabrück University, aiming to assess the accuracy of a chatbot designed to provide information about the university. The collected data will be used solely for research purposes and potentially for future improvements of the chatbot.

Contact Information: If you have any questions or concerns regarding the survey, please contact Marvin Wurch at mwurch@uni-osnabrueck.de.

This survey is anonymous. Neither the identification of participants is intended nor will it be performed. The collected data has no direct personal reference. **Please do not include any personally identifiable information in the free-text fields** (apart from the email adress for subject hours).

Participation in this anonymous survey is voluntary and is based on **Article 6(1)(a) of the GDPR (consent)**. You may withdraw your consent at any time. You can end and delete your survey anytime, the option can be found in the top right corner.

1 / 10

Data Privacy Measures:

- All collected data will be stored securely and handled confidentially.
- Data will be used only for the purposes outlined above and will not be shared with unauthorized parties.

By participating in this survey, you confirm that you understand and agree to the above conditions.

Accept   Close

Next

2 / 10

LimeSurvey

Exit and clear survey   Language: English - English

0%

Demographic Data

Purpose of These Questions
The following questions are designed to help better understand the diversity of perspectives represented in this survey. Your responses will remain confidential and will only be used to analyze patterns and trends within the collected data. While answering these questions is optional, providing your responses would be extremely helpful for gaining valuable insights. If you are uncomfortable responding, you may select "I prefer not to answer."

*Please select your age group:

Choose one of the following answers

- ○ Under 18
- ● 18-24
- ○ 25-34
- ○ 35-44
- ○ 45-54
- ○ 55-64
- ○ 65 or older
- ○ I prefer not to answer

*How do you describe your gender?

Choose one of the following answers

- ○ Male
- ● Female
- ○ I prefer not to answer
- ○ other, namely:

3 / 10

*What is the highest level of education you have completed?

Choose one of the following answers

- ○ No formal education
- ○ Secondary school diploma (e.g., intermediate school certificate)
- ○ Vocational or technical training
- ● University entrance qualification (e.g., high school diploma, Abitur)
- ○ Bachelor's degree or equivalent
- ○ Master's degree or equivalent
- ○ Higher academic qualification
- ○ None of the above
- ○ I prefer not to answer

*Which of the following best describes your current status?

Choose one of the following answers

- ○ School student (e.g., high school, secondary school)
- ○ School student intending to study at a university
- ○ Prospective student (not currently a school student, but planning to study)
- ● University student (enrolled in a university program)
- ○ International student (studying outside your home country)
- ○ Alumni (former university student)
- ○ None of the above
- ○ I prefer not to answer

What is your current study program or field of study?
(Please write the full name of your study program. If not applicable, leave the field blank)

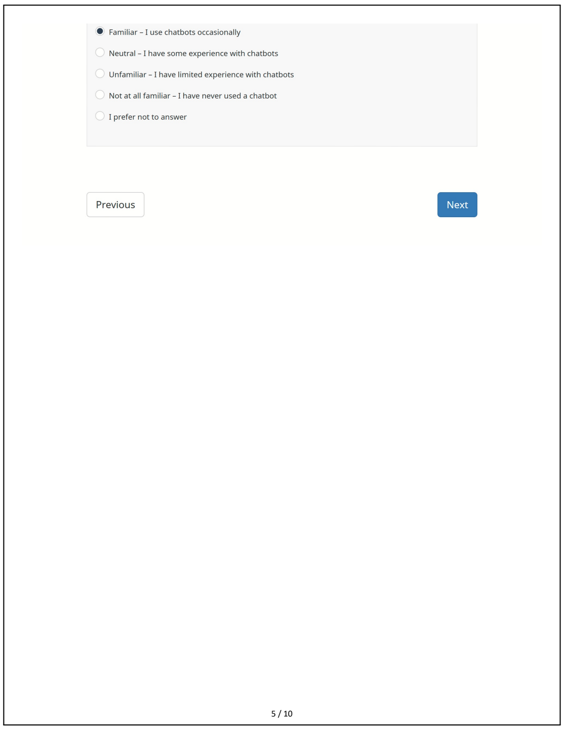
Cognitive Science

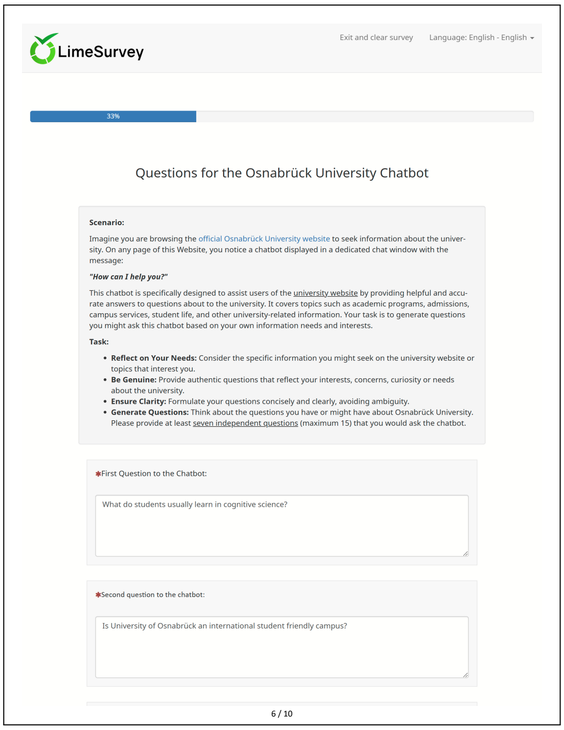
*How familiar are you with using chatbots in general?

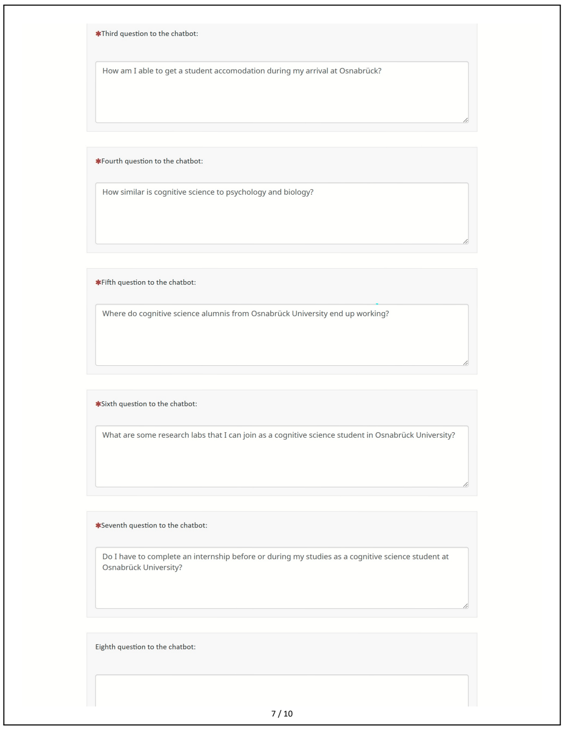
Choose one of the following answers

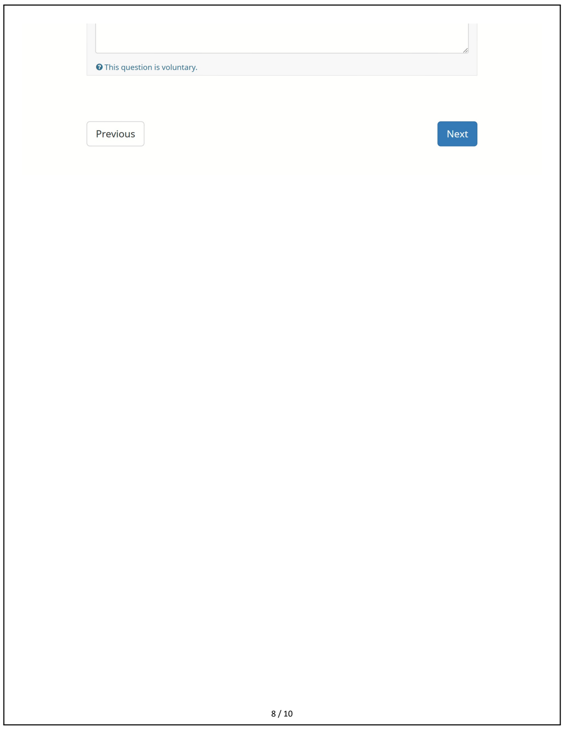
- ○ Very familiar – I use chatbots frequently

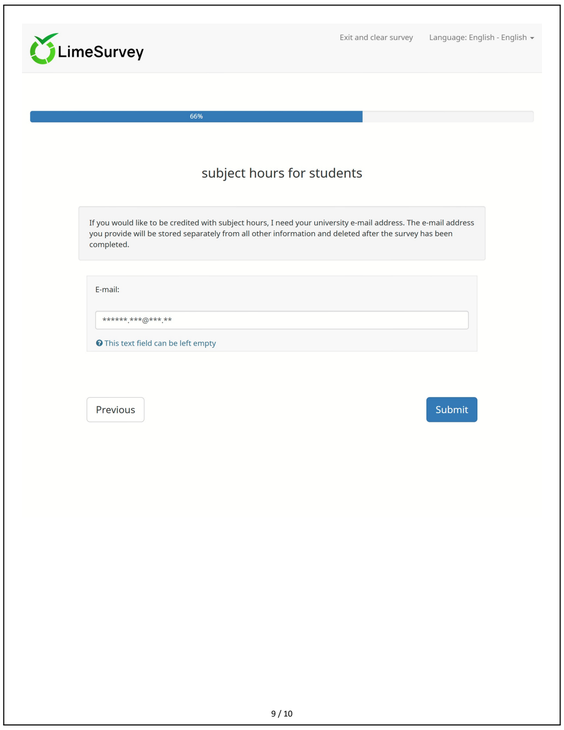
4 / 10


Familiar – I use chatbots occasionally
Neutral – I have some experience with chatbots
Unfamiliar – I have limited experience with chatbots
Not at all familiar – I have never used a chatbot
I prefer not to answer
Previous
Next
5 / 10


LimeSurvey
Exit and clear survey
Language: English - English
33%
Questions for the Osnabrück University Chatbot
Scenario:
Imagine you are browsing the official Osnabrück University website to seek information about the university. On any page of this Website, you notice a chatbot displayed in a dedicated chat window with the message:
"How can I help you?"
This chatbot is specifically designed to assist users of the university website by providing helpful and accurate answers to questions about to the university. It covers topics such as academic programs, admissions, campus services, student life, and other university-related information. Your task is to generate questions you might ask this chatbot based on your own information needs and interests.
Task:
Reflect on Your Needs: Consider the specific information you might seek on the university website or topics that interest you.
Be Genuine: Provide authentic questions that reflect your interests, concerns, curiosity or needs about the university.
Ensure Clarity: Formulate your questions concisely and clearly, avoiding ambiguity.
Generate Questions: Think about the questions you have or might have about Osnabrück University. Please provide at least seven independent questions (maximum 15) that you would ask the chatbot.
*First Question to the Chatbot:
What do students usually learn in cognitive science?
*Second question to the chatbot:
Is University of Osnabrück an international student friendly campus?
6 / 10


*Third question to the chatbot:
How am I able to get a student accomodation during my arrival at Osnabrück?
*Fourth question to the chatbot:
How similar is cognitive science to psychology and biology?
*Fifth question to the chatbot:
Where do cognitive science alumnis from Osnabrück University end up working?
*Sixth question to the chatbot:
What are some research labs that I can join as a cognitive science student in Osnabrück University?
*Seventh question to the chatbot:
Do I have to complete an internship before or during my studies as a cognitive science student at Osnabrück University?
Eighth question to the chatbot:
7 / 10


This question is voluntary.
Previous
Next
8 / 10


LimeSurvey

Exit and clear survey Language: English - English

66%

subject hours for students

If you would like to be credited with subject hours, I need your university e-mail address. The e-mail address you provide will be stored separately from all other information and deleted after the survey has been completed.

E-mail:

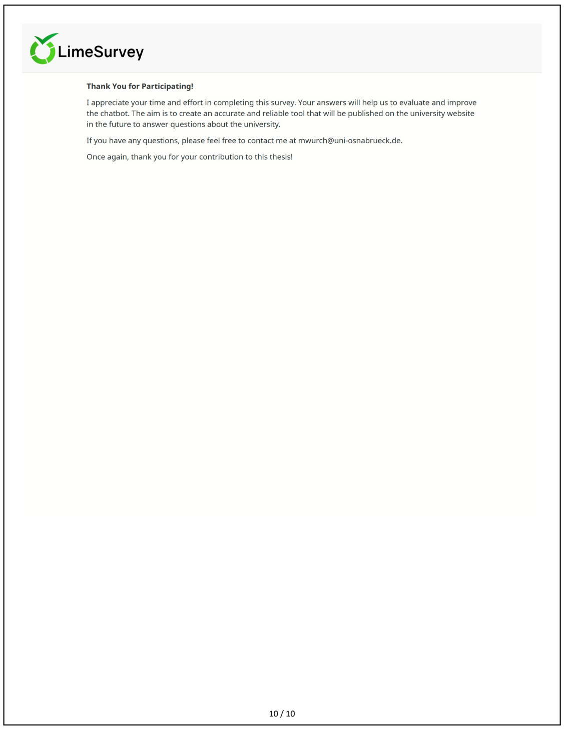
******.***@***.**

This text field can be left empty

Previous

Submit

9 / 10


LimeSurvey

Thank You for Participating!

I appreciate your time and effort in completing this survey. Your answers will help us to evaluate and improve the chatbot. The aim is to create an accurate and reliable tool that will be published on the university website in the future to answer questions about the university.

If you have any questions, please feel free to contact me at mwurch@uni-osnabrueck.de.

Once again, thank you for your contribution to this thesis!

10 / 10

A.2.2 Participant Data Plots

The following section presents the demographic data of the participants who completed Survey 1.

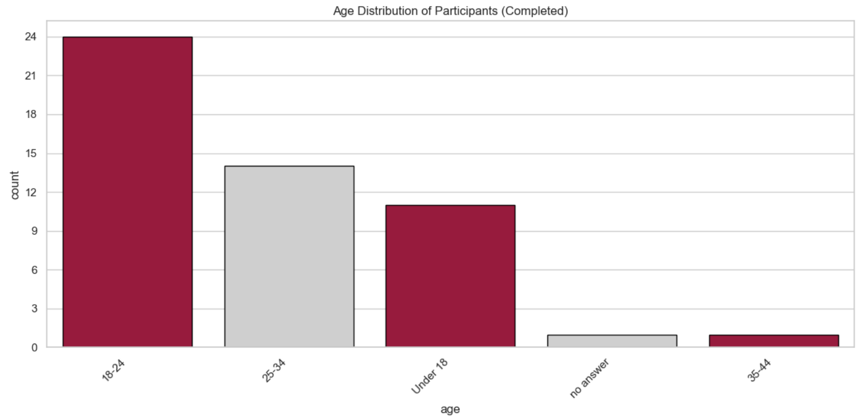


FIGURE A.1: Age Distribution of Survey 1

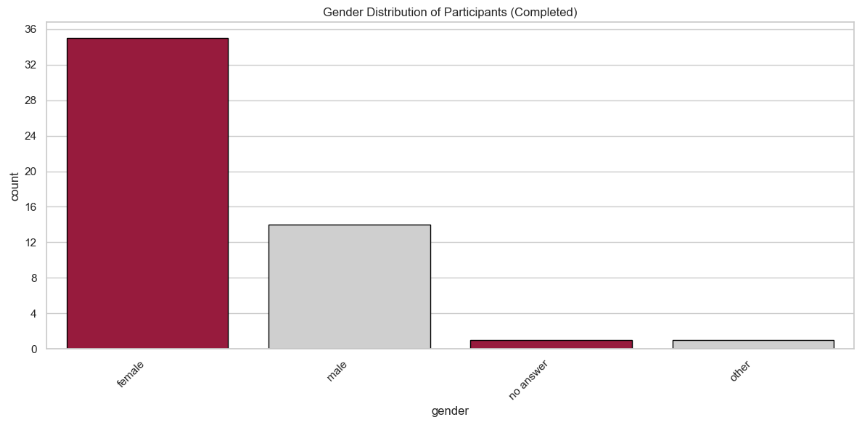


FIGURE A.2: Gender Distribution of Survey 1

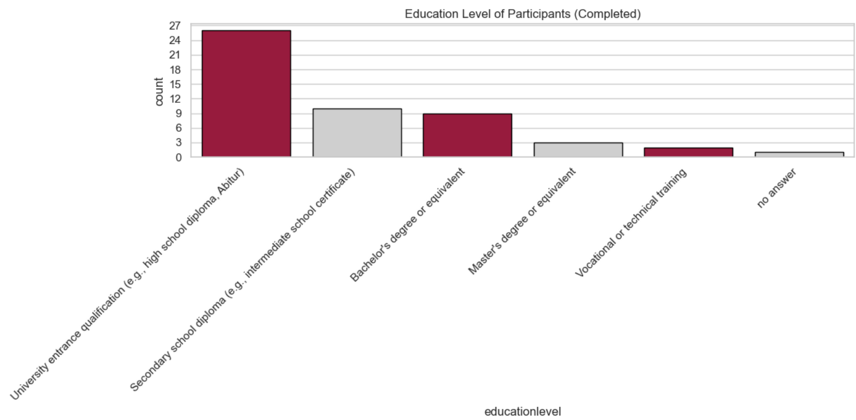


FIGURE A.3: Education Level of Survey 1

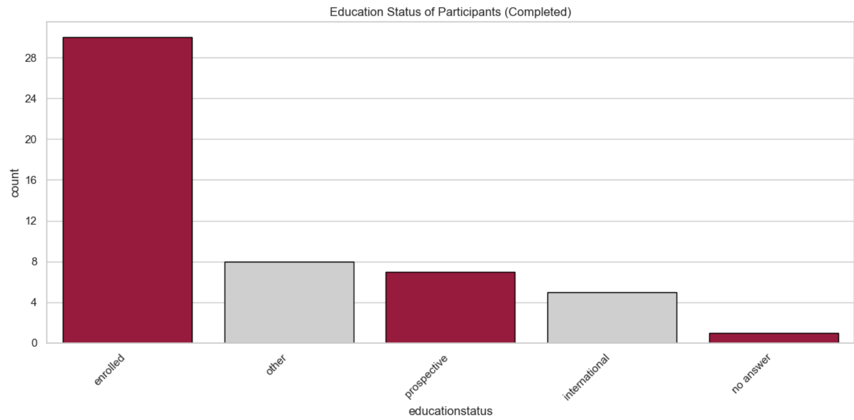


FIGURE A.4: Educational Role of Survey 1

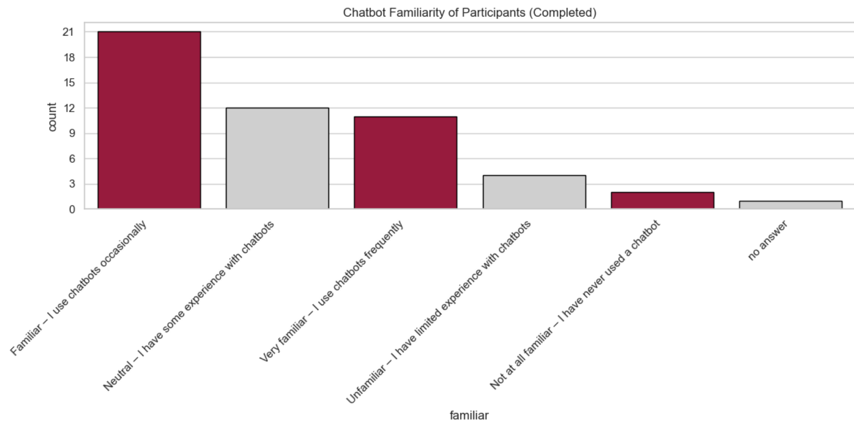


FIGURE A.5: Chatbot Familiarity of Survey 1

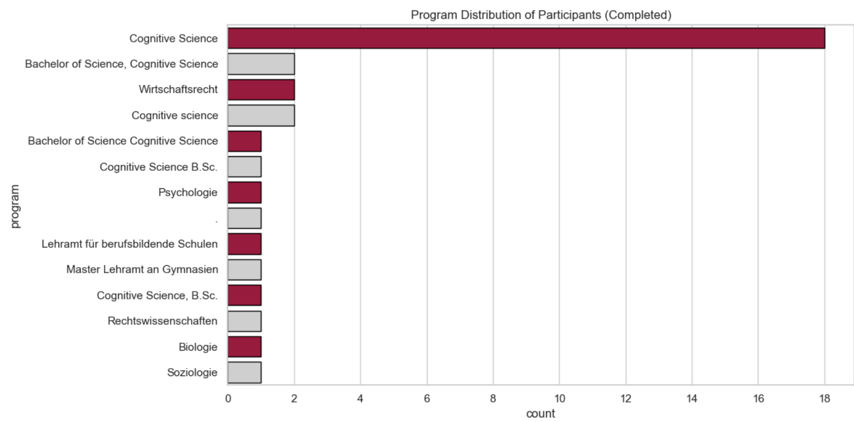


FIGURE A.6: Program Distribution of Survey 1

A.3 Answer Generation Survey

Task Description of Survey 2

About the Chatbot:
Users of the University of Osnabrück website can use a chat window to ask the chatbot questions about the university. The chat window displays the message:
"How can I help you?"
The chatbot assists users by providing **accurate**, **helpful**, and **up-to-date** answers about the university. It covers topics such as study programs, admission, campus services, student life, and other university-related information. Questions unrelated to the university are not answered by the chatbot. Instead, it politely informs users that it cannot address these concerns.

Your Task:
You will receive three questions that could be asked to the University of Osnabrück chatbot. Your task is to formulate an **accurate**, **helpful**, and **up-to-date** answer for each question. Use only information from the University of Osnabrück website.

Guidelines for Completion:

1. **Scope and Content of the Answers:**
   - Answer only **university-related** questions.
   - Formulate the answer from the chatbot's perspective and use a **polite form of address**.
   - **Do not speculate.** If you are uncertain about information, mention this in your answer.
   - If the user's query lacks sufficient details to answer, **politely request more information** to provide better assistance.
   - If you as the chatbot cannot answer a question based on the website's information, acknowledge this openly and briefly explain the reason.
2. **Source References:**
   - After your answer, refer to the **sources used** by listing all the URLs used (links to the Osnabrück University's website) as "**Link 1**," "**Link 2**," etc. in the field provided below.
   - **Within your answer**, you may refer to these links in parentheses as follows: "**[Description of the link]** (**website link**)," e.g., "For more information, visit the [virtUOS website](https://www2.uni-osnabrueck.de/virtuos/)."
3. **Structure and Clarity:**
   - Organize your answers in a **clear** and **concise** manner.
   - Use **paragraphs** or **lists** where necessary to improve readability.
4. **Language and Tone:**
   - Write in **clear**, **easily understandable**, and **professional** language.
   - Maintain a **friendly**, **polite**, and **respectful** tone to convey comprehensive support to users.
5. **Accuracy and Timeliness:**
   - Ensure your answers contain **correct**, **current**, and **relevant** information.
   - If relevant information is missing or unclear, address this transparently in your answer.
   - **Grammar and spelling** should be error-free.
   - You may draft your answer in a text processing program of your choice and paste it into the survey to ensure quality.

Example for Clarification:
Question:
What is the responsibility of virtUOS at the University of Osnabrück?
Answer:
The Center for Digital Teaching, Information Management and Higher Education Didactics (virtUOS) at the Osnabrück University is a central research and service facility dedicated to the development and implementation of innovative technologies in teaching and university organization. Its goal is to explore and implement forward-looking solutions through collaboration with other university institutions.
VirtUOS manages platforms such as Stud.IP, HISinONE, BigBlueButton, and Opencast, supports digital teaching at the University of Osnabrück, and provides comprehensive assistance to educators and students in using modern information and communication technologies. Additionally, it ensures efficient organization of application and examination processes with its digital systems. By actively shaping and conducting research projects, virtUOS advances digitization in higher education.
During the semester, the virtUOS team is available for questions or issues from Monday to Friday, 9:00 AM to 2:00 PM. They can be reached by phone at +49 541 969 6666 or via email at virtuos@uni-osnabrueck.de.
For more information, visit the [virtUOS website](https://www2.uniosnabrueck.de/virtuos/).
Sources:

- **Link 1:** [virtUOS homepage](https://www2.uni-osnabrueck.de/virtuos/)
- **Link 2:** [virtUOS profile](https://www2.uniosnabrueck.de/virtuos/profil)
- **Link 3:** [virtUOS - projects](https://www2.uniosnabrueck.de/virtuos/projekte)
- **Link 4:** [IT support](https://www.rz.uniosnabrueck.de/dienste/uositsupport/uositsupport.html)

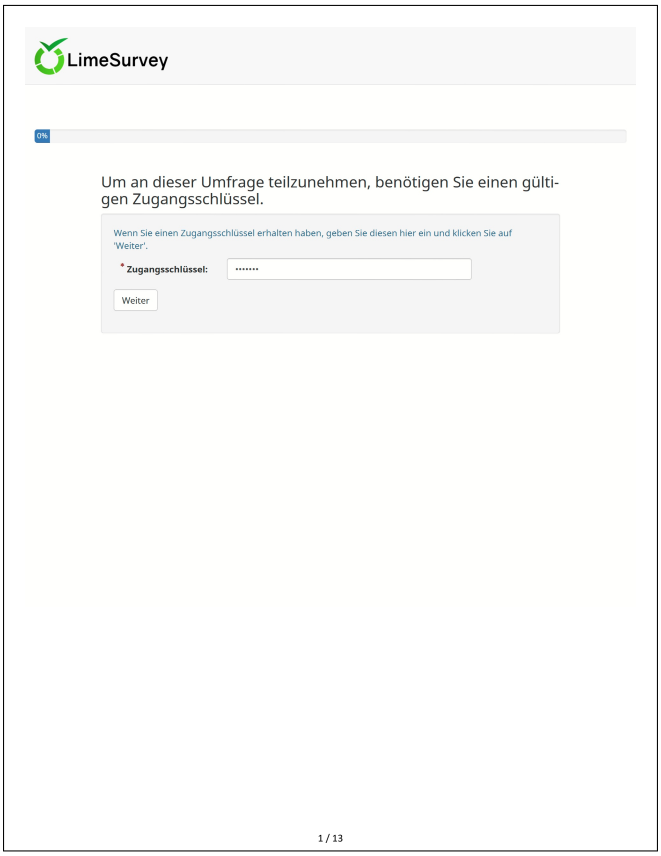
Note: The questions will be presented to you one by one. Read each question carefully before answering and strive to provide the best possible answer.

Thank you for your participation!

FIGURE A.7: The translated instructions to the participants in Survey 2 - Answer Generation Task

A.3.1 Full Answer Generation Survey

The following pages present the Answer Generation Survey submission from a randomly selected participant (Survey 7). This example highlights the type of answers generated during the survey and provides an overview of the complete survey structure. The participant's answers are shown exactly as submitted, without any modifications. For privacy reasons, the email address has been masked.



LimeSurvey

0%

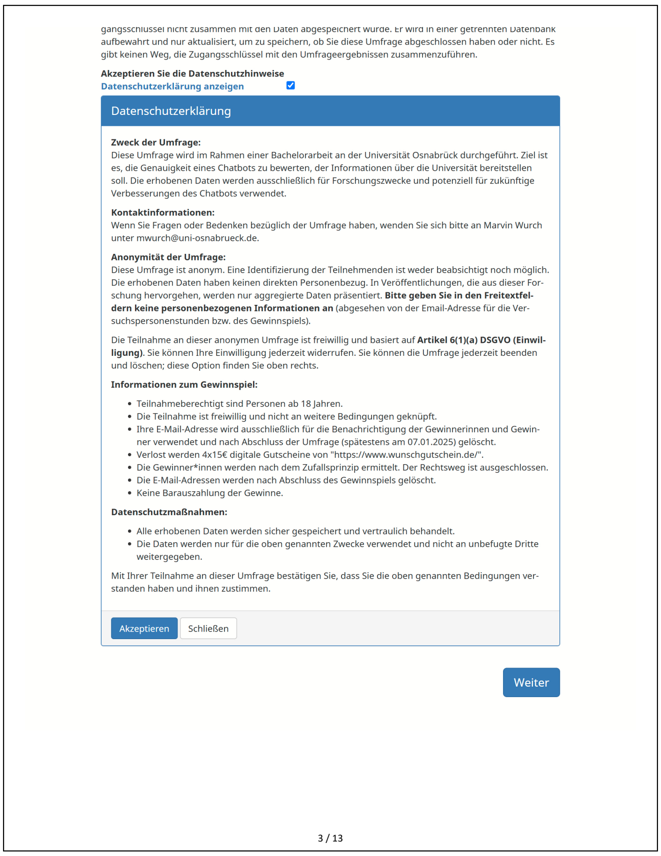
Um an dieser Umfrage teilzunehmen, benötigen Sie einen gültigen Zugangsschlüssel.

Wenn Sie einen Zugangsschlüssel erhalten haben, geben Sie diesen hier ein und klicken Sie auf 'Weiter'.

* **Zugangsschlüssel:** •••••••

Weiter

1 / 13



gangsschlüssel nicht zusammen mit den Daten abgespeichert wurde. Er wird in einer getrennten Datenbank aufbewahrt und nur aktualisiert, um zu speichern, ob Sie diese Umfrage abgeschlossen haben oder nicht. Es gibt keinen Weg, die Zugangsschlüssel mit den Umfrageergebnissen zusammenzuführen.

Akzeptieren Sie die Datenschutzhinweise
Datenschutzerklärung anzeigen

Datenschutzerklärung

Zweck der Umfrage:
Diese Umfrage wird im Rahmen einer Bachelorarbeit an der Universität Osnabrück durchgeführt. Ziel ist es, die Genauigkeit eines Chatbots zu bewerten, der Informationen über die Universität bereitstellen soll. Die erhobenen Daten werden ausschließlich für Forschungszwecke und potenziell für zukünftige Verbesserungen des Chatbots verwendet.

Kontaktinformationen:
Wenn Sie Fragen oder Bedenken bezüglich der Umfrage haben, wenden Sie sich bitte an Marvin Wurch unter mwurch@uni-osnabrueck.de.

Anonymität der Umfrage:
Diese Umfrage ist anonym. Eine Identifizierung der Teilnehmenden ist weder beabsichtigt noch möglich. Die erhobenen Daten haben keinen direkten Personenbezug. In Veröffentlichungen, die aus dieser Forschung hervorgehen, werden nur aggregierte Daten präsentiert. **Bitte geben Sie in den Freitextfeldern keine personenbezogenen Informationen an** (abgesehen von der Email-Adresse für die Versuchspersonenstunden bzw. des Gewinnspiels).

Die Teilnahme an dieser anonymen Umfrage ist freiwillig und basiert auf **Artikel 6(1)(a) DSGVO (Einwilligung)**. Sie können Ihre Einwilligung jederzeit widerrufen. Sie können die Umfrage jederzeit beenden und löschen; diese Option finden Sie oben rechts.

Informationen zum Gewinnspiel:

- Teilnahmeberechtigt sind Personen ab 18 Jahren.
- Die Teilnahme ist freiwillig und nicht an weitere Bedingungen geknüpft.
- Ihre E-Mail-Adresse wird ausschließlich für die Benachrichtigung der Gewinnerinnen und Gewinner verwendet und nach Abschluss der Umfrage (spätestens am 07.01.2025) gelöscht.
- Verlost werden 4x15€ digitale Gutscheine von "https://www.wunschgutschein.de/".
- Die Gewinner*innen werden nach dem Zufallsprinzip ermittelt. Der Rechtsweg ist ausgeschlossen.
- Die E-Mail-Adressen werden nach Abschluss des Gewinnspiels gelöscht.
- Keine Barauszahlung der Gewinne.

Datenschutzmaßnahmen:

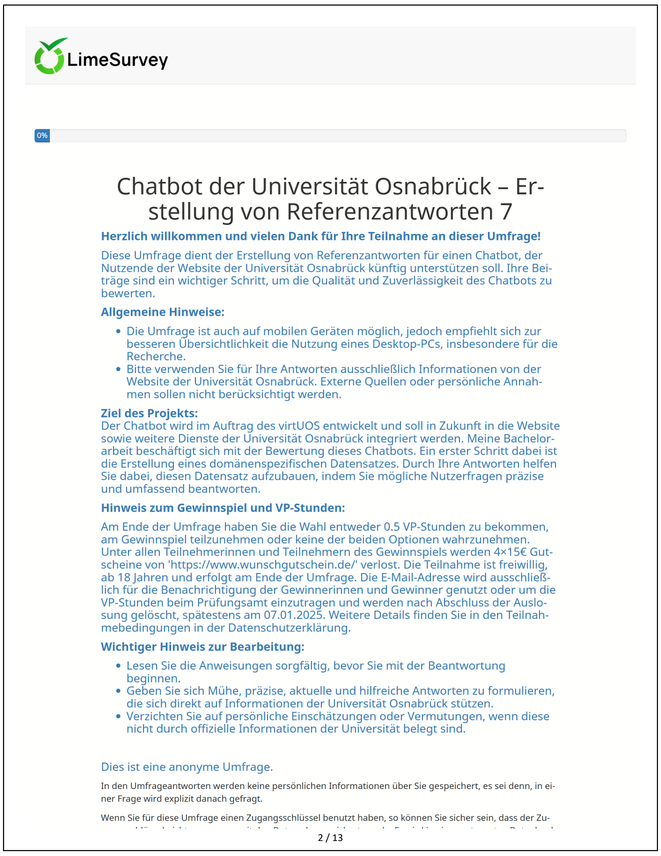
- Alle erhobenen Daten werden sicher gespeichert und vertraulich behandelt.
- Die Daten werden nur für die oben genannten Zwecke verwendet und nicht an unbefugte Dritte weitergegeben.

Mit Ihrer Teilnahme an dieser Umfrage bestätigen Sie, dass Sie die oben genannten Bedingungen verstanden haben und ihnen zustimmen.

Akzeptieren Schließen

Weiter

3 / 13



LimeSurvey

0%

Chatbot der Universität Osnabrück – Erstellung von Referenzantworten 7

Herzlich willkommen und vielen Dank für Ihre Teilnahme an dieser Umfrage!

Diese Umfrage dient der Erstellung von Referenzantworten für einen Chatbot, der Nutzende der Website der Universität Osnabrück künftig unterstützen soll. Ihre Beiträge sind ein wichtiger Schritt, um die Qualität und Zuverlässigkeit des Chatbots zu bewerten.

Allgemeine Hinweise:

- Die Umfrage ist auch auf mobilen Geräten möglich, jedoch empfiehlt sich zur besseren Übersichtlichkeit die Nutzung eines Desktop-PCs, insbesondere für die Recherche.
- Bitte verwenden Sie für Ihre Antworten ausschließlich Informationen von der Website der Universität Osnabrück. Externe Quellen oder persönliche Annahmen sollen nicht berücksichtigt werden.

Ziel des Projekts:
Der Chatbot wird im Auftrag des virtUOS entwickelt und soll in Zukunft in die Website sowie weitere Dienste der Universität Osnabrück integriert werden. Meine Bachelorarbeit beschäftigt sich mit der Bewertung dieses Chatbots. Ein erster Schritt dabei ist die Erstellung eines domänenspezifischen Datensatzes. Durch Ihre Antworten helfen Sie dabei, diesen Datensatz aufzubauen, indem Sie mögliche Nutzerfragen präzise und umfassend beantworten.

Hinweis zum Gewinnspiel und VP-Stunden:

Am Ende der Umfrage haben Sie die Wahl entweder 0.5 VP-Stunden zu bekommen, am Gewinnspiel teilzunehmen oder keine der beiden Optionen wahrzunehmen. Unter allen Teilnehmerinnen und Teilnehmern des Gewinnspiels werden 4×15€ Gutscheine von 'https://www.wunschgutschein.de/' verlost. Die Teilnahme ist freiwillig, ab 18 Jahren und erfolgt am Ende der Umfrage. Die E-Mail-Adresse wird ausschließlich für die Benachrichtigung der Gewinnerinnen und Gewinner genutzt oder um die VP-Stunden beim Prüfungsamt einzutragen und werden nach Abschluss der Auslosung gelöscht, spätestens am 07.01.2025. Weitere Details finden Sie in den Teilnahmebedingungen in der Datenschutzerklärung.

Wichtiger Hinweis zur Bearbeitung:

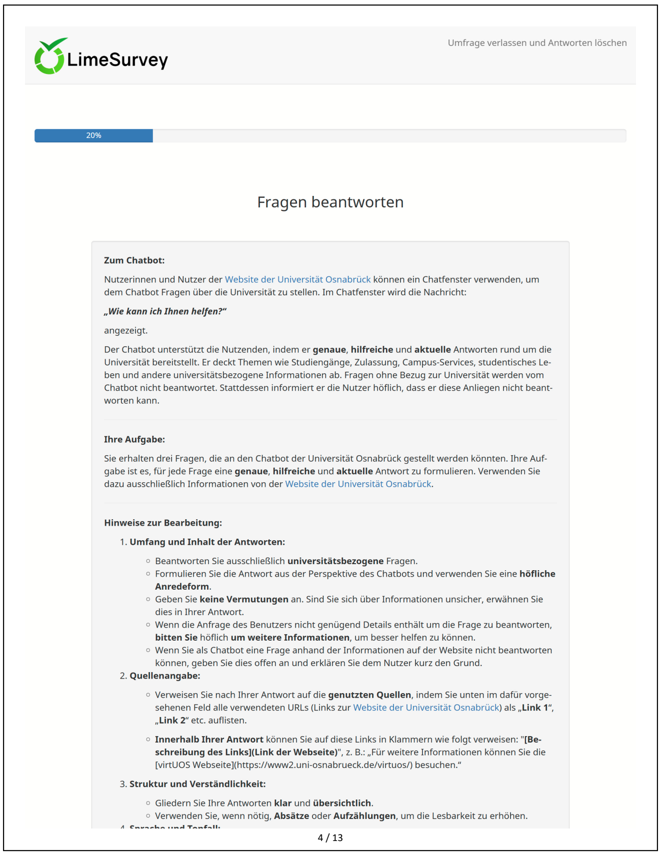
- Lesen Sie die Anweisungen sorgfältig, bevor Sie mit der Beantwortung beginnen.
- Geben Sie sich Mühe, präzise, aktuelle und hilfreiche Antworten zu formulieren, die sich direkt auf Informationen der Universität Osnabrück stützen.
- Verzichten Sie auf persönliche Einschätzungen oder Vermutungen, wenn diese nicht durch offizielle Informationen der Universität belegt sind.

Dies ist eine anonyme Umfrage.

In den Umfrageantworten werden keine persönlichen Informationen über Sie gespeichert, es sei denn, in einer Frage wird explizit danach gefragt.

Wenn Sie für diese Umfrage einen Zugangsschlüssel benutzt haben, so können Sie sicher sein, dass der Zu-

2 / 13



LimeSurvey

Umfrage verlassen und Antworten löschen

20%

Fragen beantworten

Zum Chatbot:

Nutzerinnen und Nutzer der Website der Universität Osnabrück können ein Chatfenster verwenden, um dem Chatbot Fragen über die Universität zu stellen. Im Chatfenster wird die Nachricht:

„Wie kann ich Ihnen helfen?“

angezeigt.

Der Chatbot unterstützt die Nutzenden, indem er **genaue**, **hilfreiche** und **aktuelle** Antworten rund um die Universität bereitstellt. Er deckt Themen wie Studiengänge, Zulassung, Campus-Services, studentisches Leben und andere universitätsbezogene Informationen ab. Fragen ohne Bezug zur Universität werden vom Chatbot nicht beantwortet. Stattdessen informiert er die Nutzer höflich, dass er diese Anliegen nicht beantworten kann.

Ihre Aufgabe:

Sie erhalten drei Fragen, die an den Chatbot der Universität Osnabrück gestellt werden könnten. Ihre Aufgabe ist es, für jede Frage eine **genaue**, **hilfreiche** und **aktuelle** Antwort zu formulieren. Verwenden Sie dazu ausschließlich Informationen von der Website der Universität Osnabrück.

Hinweise zur Bearbeitung:

1. **Umfang und Inhalt der Antworten:**
   - Beantworten Sie ausschließlich **universitätsbezogene** Fragen.
   - Formulieren Sie die Antwort aus der Perspektive des Chatbots und verwenden Sie eine **höfliche Anredeform**.
   - Geben Sie **keine Vermutungen** an. Sind Sie sich über Informationen unsicher, erwähnen Sie dies in Ihrer Antwort.
   - Wenn die Anfrage des Benutzers nicht genügend Details enthält um die Frage zu beantworten, **bitten Sie** höflich **um weitere Informationen**, um besser helfen zu können.
   - Wenn Sie als Chatbot eine Frage anhand der Informationen auf der Website nicht beantworten können, geben Sie dies offen an und erklären Sie dem Nutzer kurz den Grund.
2. **Quellenangabe:**
   - Verweisen Sie nach Ihrer Antwort auf die **genutzten Quellen**, indem Sie unten im dafür vorgesehenen Feld alle verwendeten URLs (Links zur Website der Universität Osnabrück) als **„Link 1“**, **„Link 2“** etc. auflisten.
   - **Innerhalb Ihrer Antwort** können Sie auf diese Links in Klammern wie folgt verweisen: "**[Beschreibung des Links](Link der Webseite)**", z. B.: „Für weitere Informationen können Sie die [virtUOS Webseite](https://www2.uni-osnabrueck.de/virtuos/) besuchen.“
3. **Struktur und Verständlichkeit:**
   - Gliedern Sie Ihre Antworten **klar** und **übersichtlich**.
   - Verwenden Sie, wenn nötig, **Absätze** oder **Aufzählungen**, um die Lesbarkeit zu erhöhen.

4 / 13

4. **Sprache und Tonfall:**
    - Schreiben Sie in einer **klaren**, **leicht verständlichen** und **professionellen** Sprache.
    - Behalten Sie einen **freundlichen**, **höflichen** und **respektvollen** Ton bei, um den Nutzenden das Gefühl umfassender Unterstützung zu vermitteln.
5. **Genauigkeit und Aktualität:**
    - Stellen Sie sicher, dass Ihre Antworten **korrekte**, **aktuelle** und **relevante** Informationen enthalten.
    - Falls relevante Informationen fehlen oder unklar sind, weisen Sie in Ihrer Antwort transparent darauf hin.
    - Rechtschreibung und Grammatik **sollten fehlerfrei** sein.
    - Sie können die Antwort in einem Textverarbeitungsprogramm Ihrer Wahl vorbereiten und in die Umfrage kopieren, um dies zu unterstützen.

Beispiel zur Veranschaulichung:

Frage:

Für was ist virtUOS an der Uni Osnabrück zuständig?

Antwort:

Das Zentrum für Digitale Lehre, Campus-Management und Hochschuldidaktik (virtUOS) der Universität Osnabrück ist eine zentrale Forschungs- und Serviceeinrichtung, die sich der Entwicklung und Implementierung innovativer Technologien in den Bereichen Lehre und Hochschulorganisation widmet. Ziel ist es, durch die Zusammenarbeit mit anderen universitären Einrichtungen, zukunftsweisende Lösungen zu erforschen und umzusetzen.

Das virtUOS betreut Plattformen wie Stud.IP, HISinONE, BigBlueButton und Opencast, unterstützt die digitale Lehre an der Universität Osnabrück und bietet Lehrenden wie Lernenden umfassende Hilfestellungen bei der Nutzung moderner Informations- und Kommunikationstechnologien. Darüber hinaus gewährleistet es mit seinen digitalen Systemen eine effiziente Organisation von Bewerbungs- und Prüfungsprozessen. Durch die aktive Mitgestaltung und Durchführung von Forschungsprojekten prägt virtUOS die Digitalisierung in der Hochschulbildung aktiv voran.

Bei Fragen oder Problemen steht das Team von virtUOS während des Semesters von Montag bis Freitag, jeweils von 9:00 bis 14:00 Uhr, zur Verfügung. Sie erreichen es telefonisch unter +49 541 969 6666 oder per E-Mail an virtuos@uni-osnabrueck.de.

Weitere Informationen finden Sie auf der [Webseite von virtUOS](https://www2.uni-osnabrueck.de/virtuos/).

Quellen:

- **Link 1:** [virtUOS Startseite](https://www2.uni-osnabrueck.de/virtuos/)
- **Link 2:** [virtUOS Profil](https://www2.uni-osnabrueck.de/virtuos/profil)
- **Link 3**: [virtUOS - Projekte](https://www2.uni-osnabrueck.de/virtuos/projekte)
- **Link 4**: [IT Support](https://www.rz.uni-osnabrueck.de/dienste/uositsupport/uositsupport.html)

Hinweis: Die Fragen werden Ihnen einzeln präsentiert. Lesen Sie jede Frage sorgfältig durch, bevor Sie antworten, und geben Sie sich Mühe, die bestmögliche Antwort zu formulieren.

Vielen Dank für Ihre Mitarbeit!

*Hier die erste zu beantwortende Frage:

"Wie lauten die Bewerbungsfristen für meine Studienfach?"

Hier finden Sie die Suchfunktion der Website der Universität Osnabrück: Klicken Sie hier.

5 / 13

Der Bewerbungszeitraum für das Sommersemester 2025 beginnt am 12.02.2024 und endet je nach Studiengang bspw. am:

31.01.2025 für zulassungsfreie grundständige Ein-Fach-Bachelor Studiengänge für höhere Fachsemester
15.01.2025 für zulassungsbeschränkte grundständige Bachelor-Studiengänge für höhere Fachsemester

Fristen für weitere Studiengangsarten finden Sie auf: "[Den FAQ's auf der Uni-Bewerbungsinfoseite](https://www.uni-osnabrueck.de/studieninteressierte/bewerbung/online-bewerbung/#c270129)"

Bitte beachten Sie, dass der Studienstart im Sommersemester, für nicht-Master Studiengänge, in der Regel nur für höhere Fachsemester möglich ist.
Weitere Informationen dazu finden Sie auf: [Der Bewerbungsseite der Universität](https://www.uni-osnabrueck.de/studieninteressierte/bewerbung/online-bewerbung/
)

Die Bewerbungsfristen sind meines Wissensstandes nach noch nicht bekannt. Einschreibung in nicht zulassungsbeschränkte Studiengänge ist aber üblicherweise bis 30. September möglich und Informationen zu zulassungsbeschränkungen werden ab Mai veröffentlicht.

Wenn Sie genauere Informationen benötigen, nennen Sie mir bitte Ihr Studienfach!

Bitte fügen Sie die URLs der Quellen ein, die Sie zur Beantwortung der ersten Frage

„Wie lauten die Bewerbungsfristen für meine Studienfach?"

verwendet haben.

Vergewissern Sie sich, dass diese Quellen von der Website der Universität Osnabrück stammen und dass das beschriebene Format korrekt ist.

Link 1: [Infowebsite mit aktuellen Infos zur Bewerbung](https://www.un

Link 2: [Website mit Terminen und Fristen](https://www.uni-osnabrueck

Link 3:

Geben Sie die Links wie folgt an: "[Beschreibung des Links](Link der Webseite)", z. B.: Link 1: [virtUOS Webseite](https://www2.uni-osnabrueck.de/virtuos/)

Zurück | Weiter

6 / 13

LimeSurvey

Umfrage verlassen und Antworten löschen

40%

Fragen beantworten 2

Beschreibung der Aufgabe erneut anzeigen / verstecken.

*Hier die zweite zu beantwortende Frage:

"Welche Ressourcen stehen Studierenden mit Behinderungen an der Universität Osnabrück zur Verfügung?"

Hier finden Sie die Suchfunktion der Website der Universität Osnabrück: Klicken Sie hier.

An der Universität Osnabrück stehen Studierenden mit Behinderungen diverse Ressourcen zur Verfügung.
Von Unterstützung der akademischen Karriere, wie Härtefallprüfung bei Bewerbung und Zulassung und Nachteilsausgleich bei Studien- und Prüfungsleistungen, über Unterstützung im Studienalltag, wie Schriftdolmetscher für Vorlesungen, Finanzielle Unterstützung durch diverse Stipendien und Fördergelder oder Begleitung durch andere Studiernende in den ersten Studienwochen.

Für weitere Informationen besuchen Sie die Website [Seite mit Informationen für Studierende mit Behinderungen](https://www.uni-osnabrueck.de/studieninteressierte/erste-orientierung/studieren-mit-behinderung-und-chronischer-erkrankung)
oder kontaktieren Sie die Beauftragte für Studierende mit Behinderung und chronischer Erkrankung:

Christine Kammler, M.A.

Tel.: +49 541 969 6366
christine.kammler@uni-osnabrueck.de

Raum: 19/118
Zentrale Studienberatung (ZSB)
Neuer Graben 27
49074 Osnabrück

7 / 13

Bitte fügen Sie die URLs der Quellen ein, die Sie zur Beantwortung der zweiten Frage

„Welche Ressourcen stehen Studierenden mit Behinderungen an der Universität Osnabrück zur Verfügung?"

verwendet haben.

Vergewissern Sie sich, dass diese Quellen von der Website der Universität Osnabrück stammen und dass das beschriebene Format korrekt ist.

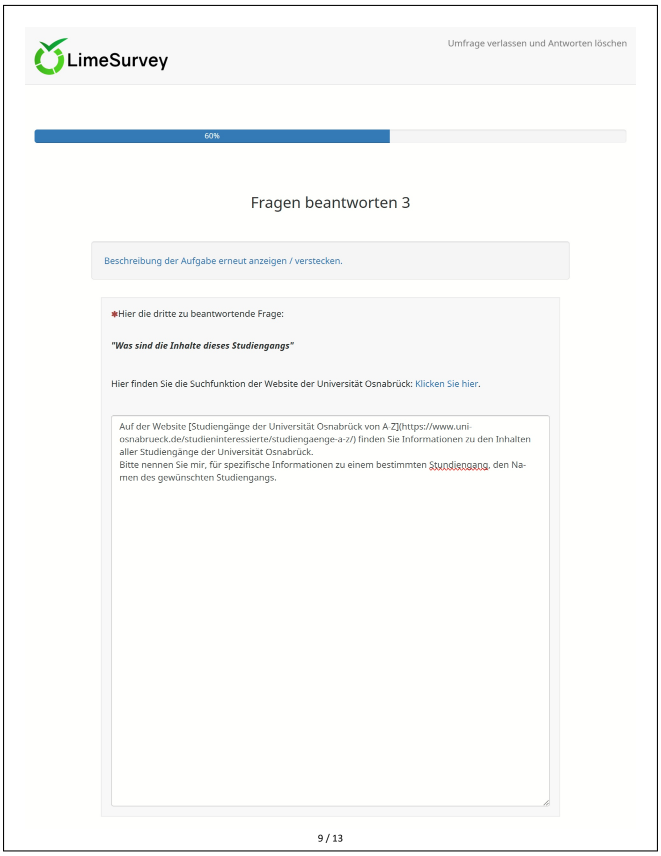
Link 1: [Seite mit Informationen für Studierende mit Behinderungen](ht

Link 2:

Geben Sie die Links wie folgt an: "[Beschreibung des Links](Link der Webseite)", z. B.: Link 1: [virtUOS Webseite](https://www2.uni-osnabrueck.de/virtuos/)

Zurück | Weiter

8 / 13



LimeSurvey

Umfrage verlassen und Antworten löschen

60%

Fragen beantworten 3

Beschreibung der Aufgabe erneut anzeigen / verstecken.

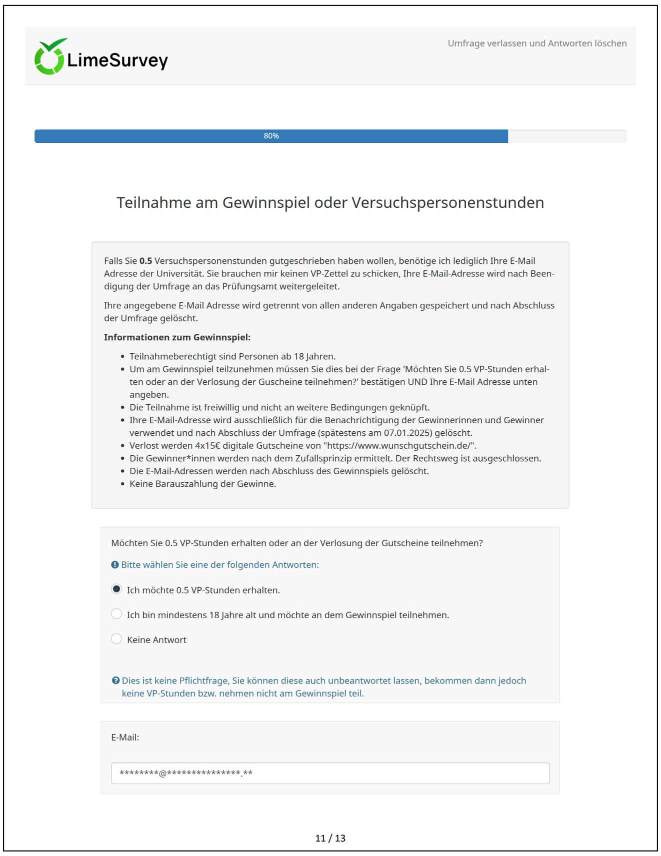
*Hier die dritte zu beantwortende Frage:

"Was sind die Inhalte dieses Studiengangs"

Hier finden Sie die Suchfunktion der Website der Universität Osnabrück: Klicken Sie hier.

Auf der Website [Studiengänge der Universität Osnabrück von A-Z](https://www.uni-osnabrueck.de/studieninteressierte/studiengaenge-a-z/) finden Sie Informationen zu den Inhalten aller Studiengänge der Universität Osnabrück.
Bitte nennen Sie mir, für spezifische Informationen zu einem bestimmten Stundiengang, den Namen des gewünschten Studiengangs.

9 / 13



LimeSurvey

Umfrage verlassen und Antworten löschen

80%

Teilnahme am Gewinnspiel oder Versuchspersonenstunden

Falls Sie **0.5** Versuchspersonenstunden gutgeschrieben haben wollen, benötige ich lediglich Ihre E-Mail Adresse der Universität. Sie brauchen mir keinen VP-Zettel zu schicken, Ihre E-Mail-Adresse wird nach Beendigung der Umfrage an das Prüfungsamt weitergeleitet.

Ihre angegebene E-Mail Adresse wird getrennt von allen anderen Angaben gespeichert und nach Abschluss der Umfrage gelöscht.

Informationen zum Gewinnspiel:

- Teilnahmeberechtigt sind Personen ab 18 Jahren.
- Um am Gewinnspiel teilzunehmen müssen Sie dies bei der Frage 'Möchten Sie 0.5 VP-Stunden erhalten oder an der Verlosung der Guscheine teilnehmen?' bestätigen UND Ihre E-Mail Adresse unten angeben.
- Die Teilnahme ist freiwillig und nicht an weitere Bedingungen geknüpft.
- Ihre E-Mail-Adresse wird ausschließlich für die Benachrichtigung der Gewinnerinnen und Gewinner verwendet und nach Abschluss der Umfrage (spätestens am 07.01.2025) gelöscht.
- Verlost werden 4x15€ digitale Gutscheine von "https://www.wunschgutschein.de/".
- Die Gewinner*innen werden nach dem Zufallsprinzip ermittelt. Der Rechtsweg ist ausgeschlossen.
- Die E-Mail-Adressen werden nach Abschluss des Gewinnspiels gelöscht.
- Keine Barauszahlung der Gewinne.

Möchten Sie 0.5 VP-Stunden erhalten oder an der Verlosung der Gutscheine teilnehmen?

Bitte wählen Sie eine der folgenden Antworten:

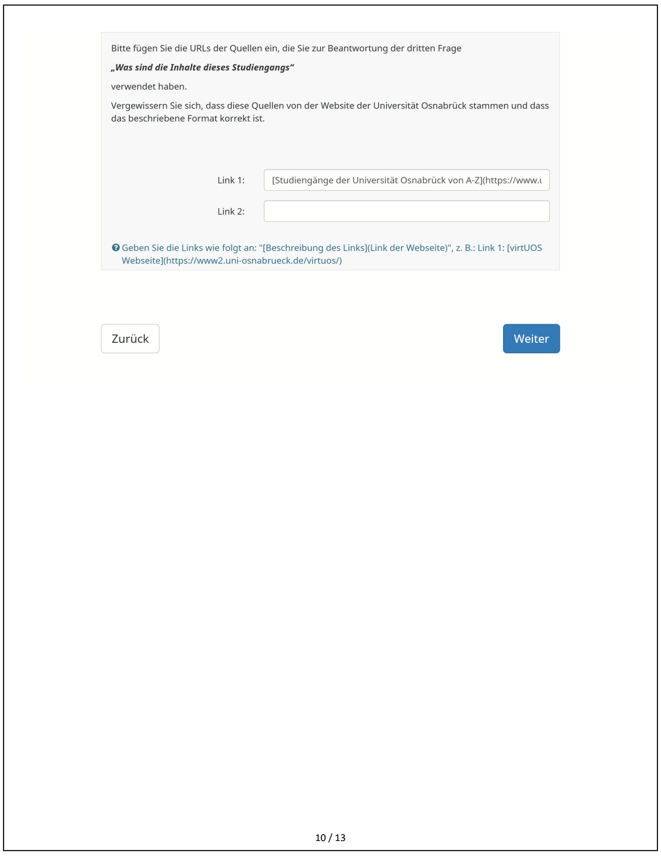
- Ich möchte 0.5 VP-Stunden erhalten.
- Ich bin mindestens 18 Jahre alt und möchte an dem Gewinnspiel teilnehmen.
- Keine Antwort

Dies ist keine Pflichtfrage, Sie können diese auch unbeantwortet lassen, bekommen dann jedoch keine VP-Stunden bzw. nehmen nicht am Gewinnspiel teil.

E-Mail:

********@***************.**

11 / 13



Bitte fügen Sie die URLs der Quellen ein, die Sie zur Beantwortung der dritten Frage

„Was sind die Inhalte dieses Studiengangs"

verwendet haben.

Vergewissern Sie sich, dass diese Quellen von der Website der Universität Osnabrück stammen und dass das beschriebene Format korrekt ist.

Link 1: [Studiengänge der Universität Osnabrück von A-Z](https://www.u

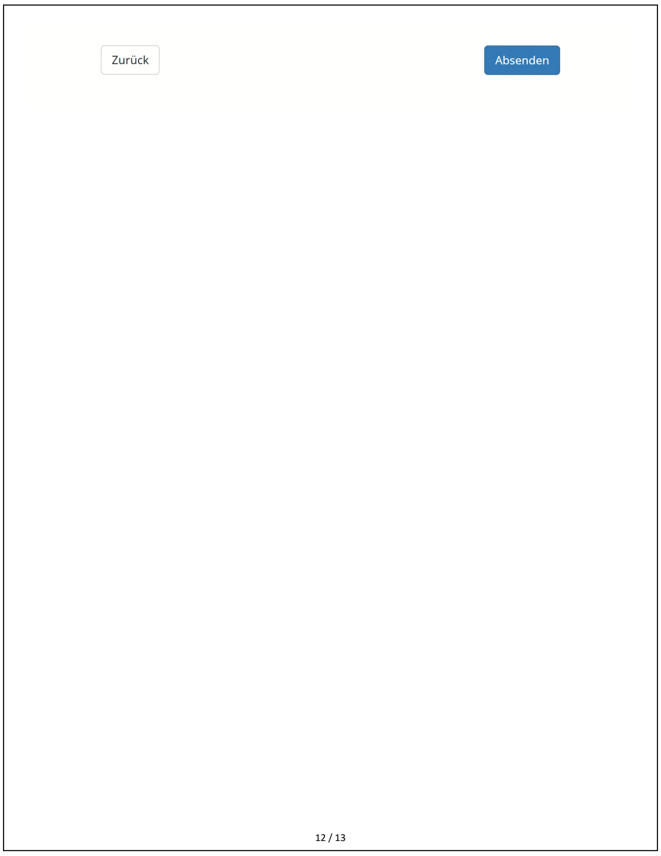
Link 2:

Geben Sie die Links wie folgt an: "[Beschreibung des Links](Link der Webseite)", z. B.: Link 1: [virtUOS Webseite](https://www2.uni-osnabrueck.de/virtuos/)

Zurück

Weiter

10 / 13



Zurück

Absenden

12 / 13

LimeSurvey

Vielen Dank für Ihre wertvolle Unterstützung!

Ich weiß Ihre Zeit und Mühe zu schätzen. Ihre Antworten werden dazu beitragen, die Qualität und Zuverlässigkeit unseres Chatbots zu bewerten.

Bei Fragen oder Anmerkungen können Sie sich gerne an mwurch@uni-osnabrueck.de wenden.

Nochmals herzlichen Dank für Ihre Teilnahme an dieser Studie!

13 / 13

A.3.2 Participant Data Plots

The following section presents the demographic data of the participants who completed Survey 2.

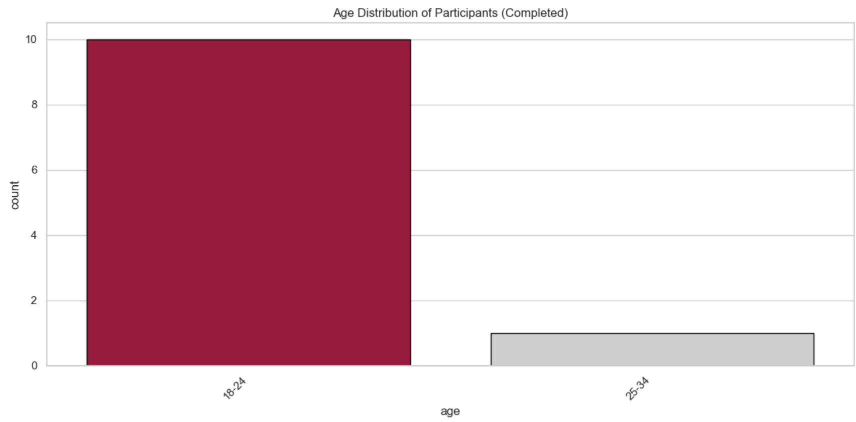


FIGURE A.8: Age Distribution of Survey 2

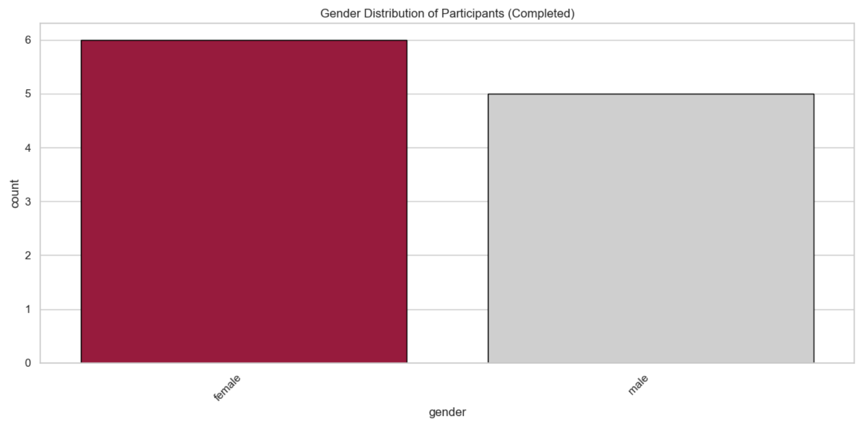


FIGURE A.9: Gender Distribution of Survey 2

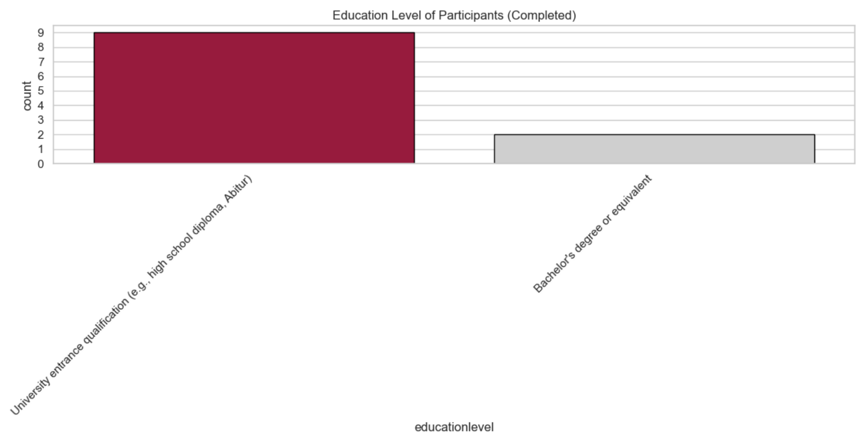


FIGURE A.10: Education Level of Survey 2

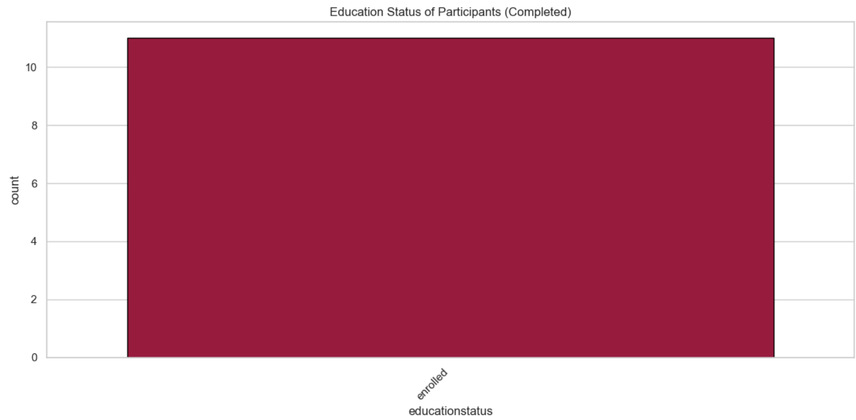


FIGURE A.11: Educational Role of Survey 2

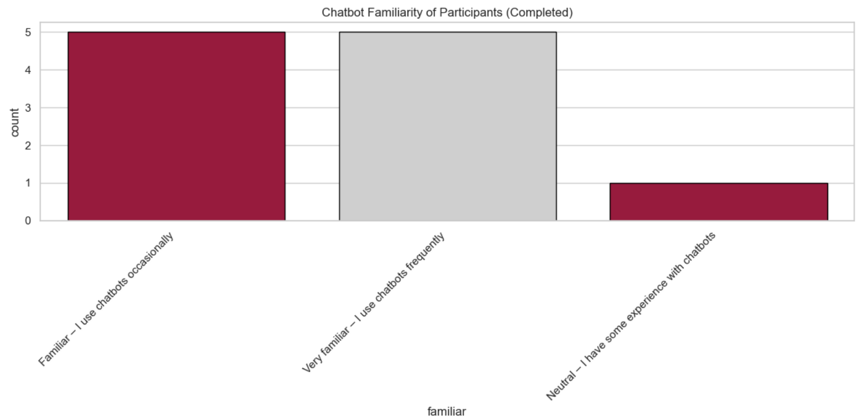


FIGURE A.12: Chatbot Familiarity of Survey 2

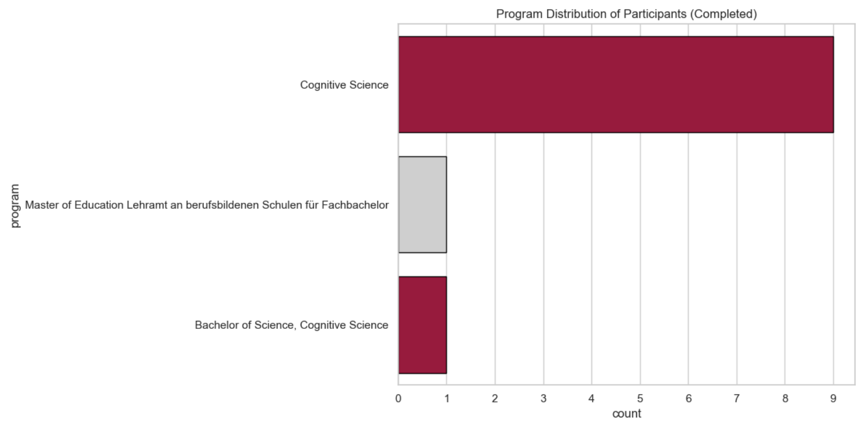
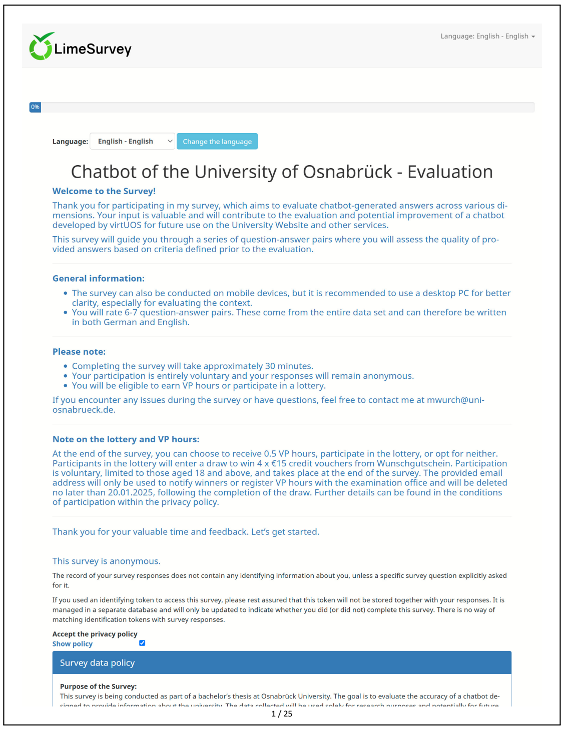


FIGURE A.13: Program Distribution of Survey 2

A.4 Human Evaluation Survey

A.4.1 Full Evaluation Survey

In the following pages, the reader can review the Chatbot Answer Evaluation submission from a randomly selected participant (Participant 41). This example showcases the whole survey structure. The answers of the participant are presented exactly as provided by the participant, without alterations. For privacy reasons the e-mail address is masked.



LimeSurvey

Language: English - English

0%

Language: English - English | Change the language

Chatbot of the University of Osnabrück - Evaluation

Welcome to the Survey!

Thank you for participating in my survey, which aims to evaluate chatbot-generated answers across various dimensions. Your input is valuable and will contribute to the evaluation and potential improvement of a chatbot developed by virtUOS for future use on the University Website and other services.

This survey will guide you through a series of question-answer pairs where you will assess the quality of provided answers based on criteria defined prior to the evaluation.

General information:

- The survey can also be conducted on mobile devices, but it is recommended to use a desktop PC for better clarity, especially for evaluating the context.
- You will rate 6-7 question-answer pairs. These come from the entire data set and can therefore be written in both German and English.

Please note:

- Completing the survey will take approximately 30 minutes.
- Your participation is entirely voluntary and your responses will remain anonymous.
- You will be eligible to earn VP hours or participate in a lottery.

If you encounter any issues during the survey or have questions, feel free to contact me at mwurch@uni-osnabrueck.de.

Note on the lottery and VP hours:

At the end of the survey, you can choose to receive 0.5 VP hours, participate in the lottery, or opt for neither. Participants in the lottery will enter a draw to win 4 x €15 credit vouchers from Wunschgutschein. Participation is voluntary, limited to those aged 18 and above, and takes place at the end of the survey. The provided email address will only be used to notify winners or register VP hours with the examination office and will be deleted no later than 20.01.2025, following the completion of the draw. Further details can be found in the conditions of participation within the privacy policy.

Thank you for your valuable time and feedback. Let's get started.

This survey is anonymous.

The record of your survey responses does not contain any identifying information about you, unless a specific survey question explicitly asked for it.

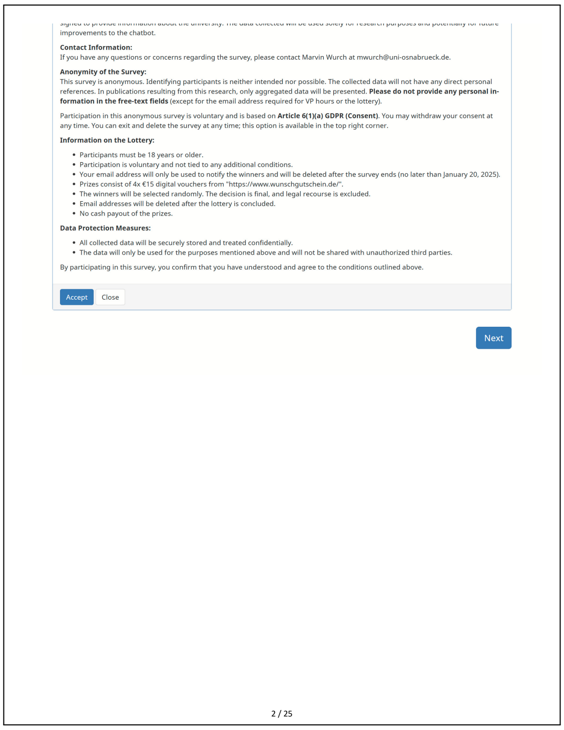
If you used an identifying token to access this survey, please rest assured that this token will not be stored together with your responses. It is managed in a separate database and will only be updated to indicate whether you did (or did not) complete this survey. There is no way of matching identification tokens with survey responses.

Accept the privacy policy
Show policy

Survey data policy

Purpose of the Survey:
This survey is being conducted as part of a bachelor's thesis at Osnabrück University. The goal is to evaluate the accuracy of a chatbot de-

1 / 25



improvements to the chatbot.

Contact Information:
If you have any questions or concerns regarding the survey, please contact Marvin Wurch at mwurch@uni-osnabrueck.de.

Anonymity of the Survey:
This survey is anonymous. Identifying participants is neither intended nor possible. The collected data will not have any direct personal references. In publications resulting from this research, only aggregated data will be presented. **Please do not provide any personal information in the free-text fields** (except for the email address required for VP hours or the lottery).

Participation in this anonymous survey is voluntary and is based on **Article 6(1)(a) GDPR (Consent)**. You may withdraw your consent at any time. You can exit and delete the survey at any time; this option is available in the top right corner.

Information on the Lottery:

- Participants must be 18 years or older.
- Participation is voluntary and not tied to any additional conditions.
- Your email address will only be used to notify the winners and will be deleted after the survey ends (no later than January 20, 2025).
- Prizes consist of 4x €15 digital vouchers from "https://www.wunschgutschein.de/".
- The winners will be selected randomly. The decision is final, and legal recourse is excluded.
- Email addresses will be deleted after the lottery is concluded.
- No cash payout of the prizes.

Data Protection Measures:

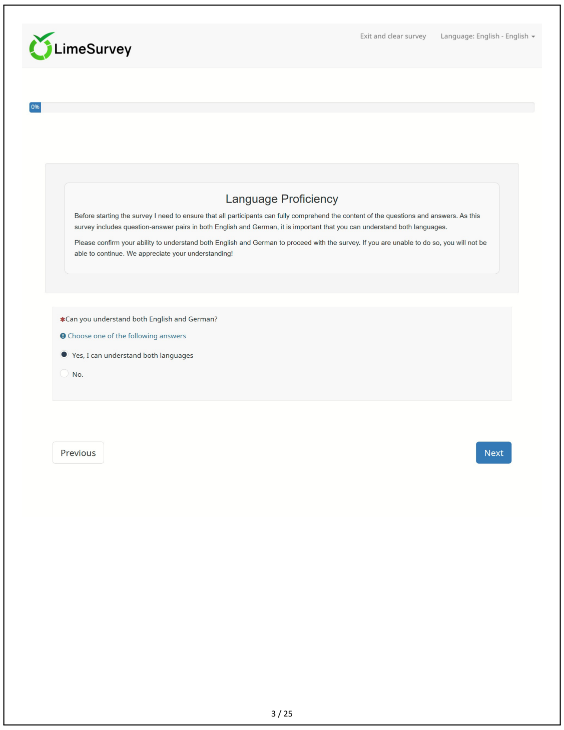
- All collected data will be securely stored and treated confidentially.
- The data will only be used for the purposes mentioned above and will not be shared with unauthorized third parties.

By participating in this survey, you confirm that you have understood and agree to the conditions outlined above.

Accept | Close

Next

2 / 25



LimeSurvey

Exit and clear survey   Language: English - English

0%

Language Proficiency

Before starting the survey I need to ensure that all participants can fully comprehend the content of the questions and answers. As this survey includes question-answer pairs in both English and German, it is important that you can understand both languages.

Please confirm your ability to understand both English and German to proceed with the survey. If you are unable to do so, you will not be able to continue. We appreciate your understanding!

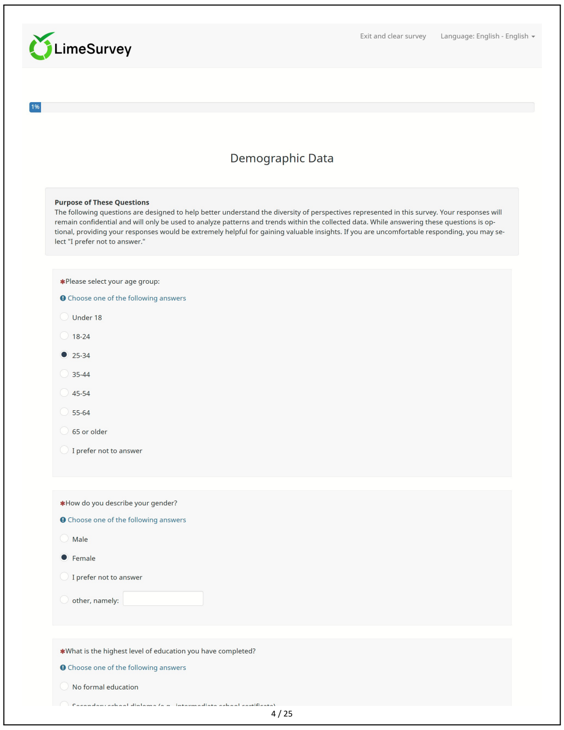
*Can you understand both English and German?

Choose one of the following answers

- Yes, I can understand both languages
- No.

Previous   Next

3 / 25



LimeSurvey

Exit and clear survey   Language: English - English

1%

Demographic Data

Purpose of These Questions
The following questions are designed to help better understand the diversity of perspectives represented in this survey. Your responses will remain confidential and will only be used to analyze patterns and trends within the collected data. While answering these questions is optional, providing your responses would be extremely helpful for gaining valuable insights. If you are uncomfortable responding, you may select "I prefer not to answer."

*Please select your age group:

Choose one of the following answers

- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 or older
- I prefer not to answer

*How do you describe your gender?

Choose one of the following answers

- Male
- Female
- I prefer not to answer
- other, namely:

*What is the highest level of education you have completed?

Choose one of the following answers

- No formal education

4 / 25

- [ ] Secondary school diploma (e.g., intermediate school certificate)
- [ ] Vocational or technical training
- [ ] University entrance qualification (e.g., high school diploma, Abitur)
- [ ] Bachelor's degree or equivalent
- [x] Master's degree or equivalent
- [ ] Higher academic qualification
- [ ] None of the above
- [ ] I prefer not to answer

*Which of the following best describes your current status?

Choose one of the following answers

- [ ] School student (e.g., high school, secondary school)
- [ ] School student intending to study at a university
- [ ] Prospective student (not currently a school student, but planning to study)
- [ ] University student (enrolled in a university program)
- [ ] International student (studying outside your home country)
- [x] Alumni (former university student)
- [ ] None of the above
- [ ] I prefer not to answer

What is your current study program or field of study?
(Please write the full name of your study program. If not applicable, leave the field blank)

Artificial Intelligence

*How familiar are you with using chatbots in general?

Choose one of the following answers

- [x] Very familiar – I use chatbots frequently
- [ ] Familiar – I use chatbots occasionally
- [ ] Neutral – I have some experience with chatbots
- [ ] Unfamiliar – I have limited experience with chatbots
- [ ] Not at all familiar – I have never used a chatbot
- [ ] I prefer not to answer

5 / 25

Previous Next

6 / 25

LimeSurvey

Exit and clear survey Language: English - English

2%

Task Description

You are tasked with evaluating the quality of answers generated by a chatbot designed to assist users with questions related to Osnabrück University. Your feedback will help assess and improve the chatbot across several dimensions of quality.

Instructions

System Setting:

- The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline.
- The chatbot's answers were generated on **January 9, 2025.**

Please take this into account when evaluating the answers.

1. **Evaluation Dimensions:**

   Rate the answer on the following dimensions using a scale of 1 to 5 (1 = Very Bad, 5 = Very Good):

   - **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
     - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
   - **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing *correct*, *complete*, and *relevant* information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
     - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
   - **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
     - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
   - **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
     - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
   - **Context Quality:** Assesses the relevance and completeness of the context (the listed links under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
     - **With 'Context used':**
       - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
       - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
     - **Absence of 'Context used':**
       - Evaluate how the absence of context impacts the quality of the answer.
       - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would haven been unnecessary).

2. **Best Practices:**
   - **Take your time:** Carefully read the user's question and any provided context before evaluating.
   - **Be Objective:** Base your evaluations strictly on the provided content and criteria.
   - **Evaluate Independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.

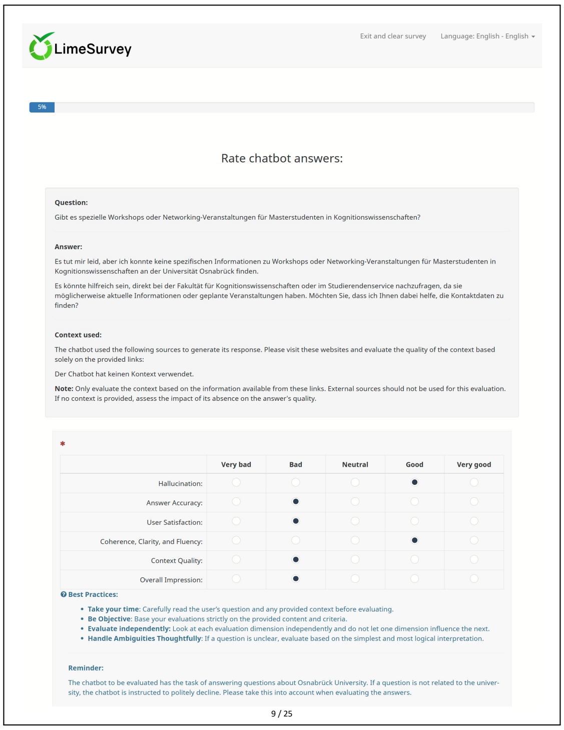
7 / 25

   - **Handle Ambiguities Thoughtfully:** If a question is unclear, evaluate based on the simplest and most logical interpretation.

Thank you for your thoughtful evaluations—they are critical to the success of this project.

Previous Next

8 / 25



LimeSurvey

Exit and clear survey Language: English - English

5%

Rate chatbot answers:

Question:

Gibt es spezielle Workshops oder Networking-Veranstaltungen für Masterstudenten in Kognitionswissenschaften?

Answer:

Es tut mir leid, aber ich konnte keine spezifischen Informationen zu Workshops oder Networking-Veranstaltungen für Masterstudenten in Kognitionswissenschaften an der Universität Osnabrück finden.

Es könnte hilfreich sein, direkt bei der Fakultät für Kognitionswissenschaften oder im Studierendenservice nachzufragen, da sie möglicherweise aktuelle Informationen oder geplante Veranstaltungen haben. Möchten Sie, dass ich Ihnen dabei helfe, die Kontaktdaten zu finden?

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

Der Chatbot hat keinen Kontext verwendet.

Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

	Very bad	Bad	Neutral	Good	Very good
Hallucination:	○	○	○	●	○
Answer Accuracy:	○	●	○	○	○
User Satisfaction:	○	●	○	○	○
Coherence, Clarity, and Fluency:	○	○	○	●	○
Context Quality:	○	●	○	○	○
Overall Impression:	○	●	○	○	○

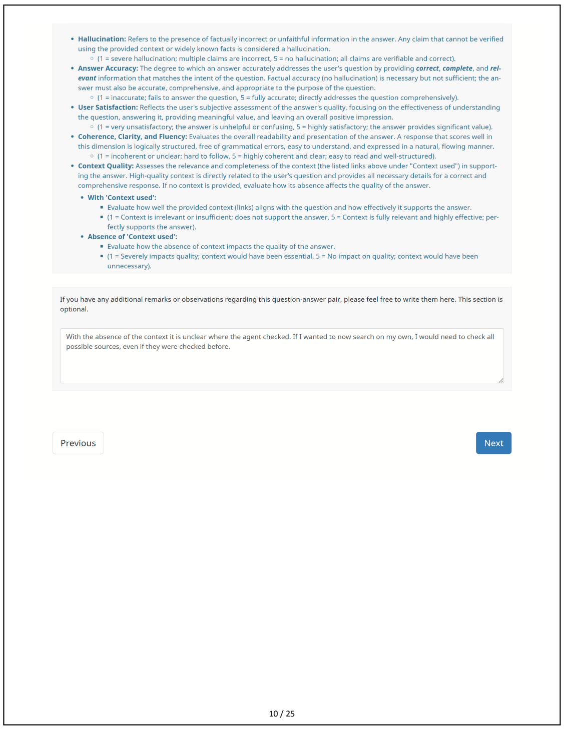
Best Practices:

- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

9 / 25



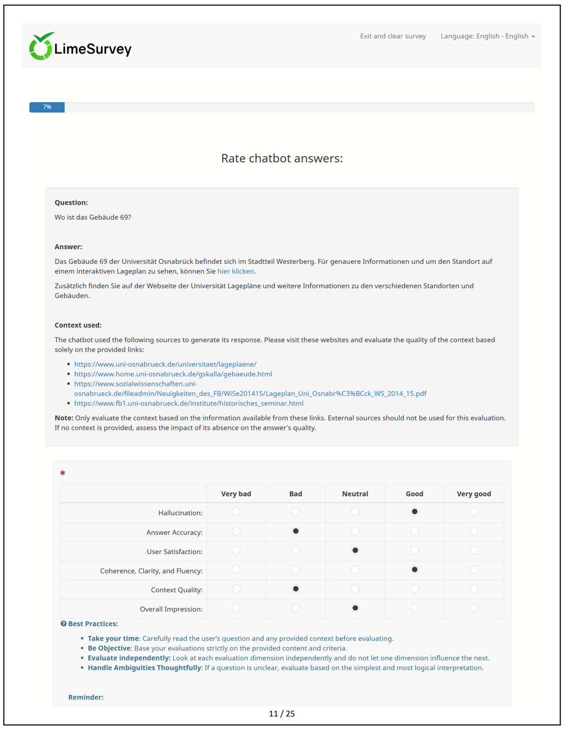
- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

With the absence of the context it is unclear where the agent checked. If I wanted to now search on my own, I would need to check all possible sources, even if they were checked before.

Previous Next

10 / 25



LimeSurvey

Exit and clear survey Language: English - English

7%

Rate chatbot answers:

Question:

Wo ist das Gebäude 69?

Answer:

Das Gebäude 69 der Universität Osnabrück befindet sich im Stadtteil Westerberg. Für genauere Informationen und um den Standort auf einem interaktiven Lageplan zu sehen, können Sie hier klicken.

Zusätzlich finden Sie auf der Webseite der Universität Lagepläne und weitere Informationen zu den verschiedenen Standorten und Gebäuden.

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

- https://www.uni-osnabrueck.de/universitaet/lageplaene/
- https://www.home.uni-osnabrueck.de/gskalla/gebaeude.html
- https://www.sozialwissenschaften.uni-osnabrueck.de/fileadmin/Neuigkeiten_des_FB/WiSe201415/Lageplan_Uni_Osnabr%C3%BCck_WS_2014_15.pdf
- https://www.fb1.uni-osnabrueck.de/institute/historisches_seminar.html

Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

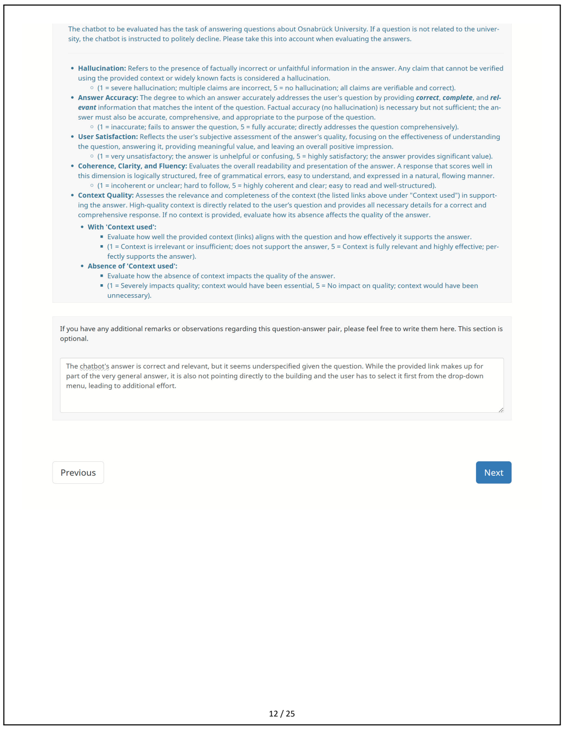
	Very bad	Bad	Neutral	Good	Very good
Hallucination:	○	○	○	●	○
Answer Accuracy:	○	●	○	○	○
User Satisfaction:	○	○	●	○	○
Coherence, Clarity, and Fluency:	○	○	○	●	○
Context Quality:	○	●	○	○	○
Overall Impression:	○	○	●	○	○

Best Practices:

- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

11 / 25



The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

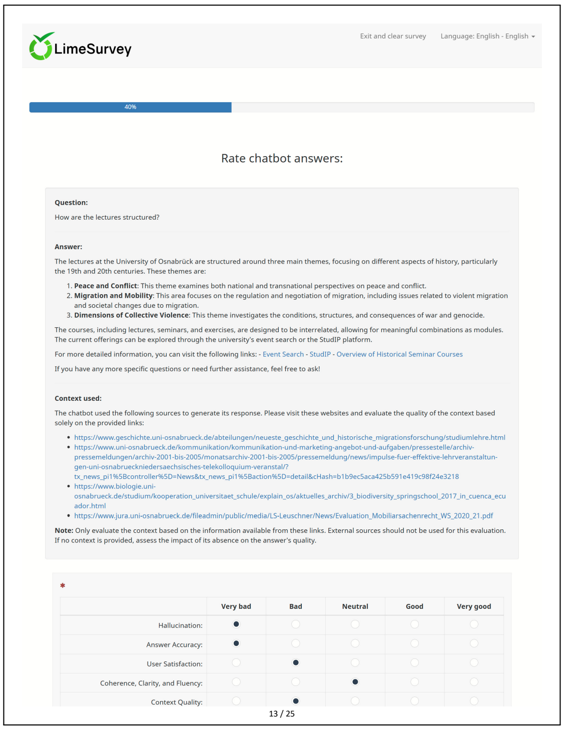
- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

The chatbot's answer is correct and relevant, but it seems underspecified given the question. While the provided link makes up for part of the very general answer, it is also not pointing directly to the building and the user has to select it first from the drop-down menu, leading to additional effort.

Previous Next

12 / 25



LimeSurvey Exit and clear survey Language: English - English

40%

Rate chatbot answers:

Question:

How are the lectures structured?

Answer:

The lectures at the University of Osnabrück are structured around three main themes, focusing on different aspects of history, particularly the 19th and 20th centuries. These themes are:

1. **Peace and Conflict**: This theme examines both national and transnational perspectives on peace and conflict.
2. **Migration and Mobility**: This area focuses on the regulation and negotiation of migration, including issues related to violent migration and societal changes due to migration.
3. **Dimensions of Collective Violence**: This theme investigates the conditions, structures, and consequences of war and genocide.

The courses, including lectures, seminars, and exercises, are designed to be interrelated, allowing for meaningful combinations as modules. The current offerings can be explored through the university's event search or the StudIP platform.

For more detailed information, you can visit the following links: - Event Search - StudIP - Overview of Historical Seminar Courses

If you have any more specific questions or need further assistance, feel free to ask!

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

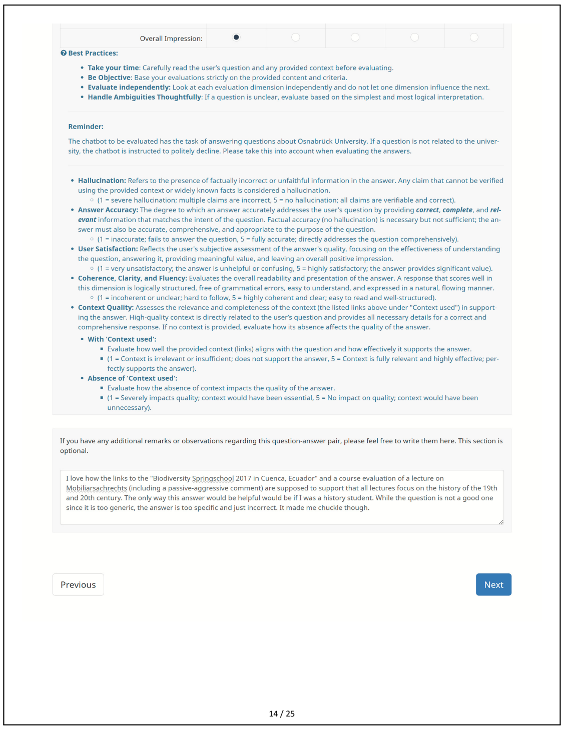
- https://www.geschichte.uni-osnabrueck.de/abteilungen/neueste_geschichte_und_historische_migrationsforschung/studiumlehre.html
- https://www.uni-osnabrueck.de/kommunikation/kommunikation-und-marketing-angebot-und-aufgaben/pressestelle/archiv-pressemeldungen/archiv-2001-bis-2005/monatsarchiv-2001-bis-2005/pressemeldung/news/impulse-fuer-effektive-lehrveranstaltungen-uni-osnabrueckniedersaechsisches-telekolloquium-veranstal/?tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Baction%5D=detail&cHash=b1b9ec5aca425b591e419c98f24e3218
- https://www.biologie.uni-osnabrueck.de/studium/kooperation_universitaet_schule/explain_os/aktuelles_archiv/3_biodiversity_springschool_2017_in_cuenca_ecuador.html
- https://www.jura.uni-osnabrueck.de/fileadmin/public/media/LS-Leuschner/News/Evaluation_Mobiliarsachenrecht_WS_2020_21.pdf

Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

	Very bad	Bad	Neutral	Good	Very good
Hallucination:	●	○	○	○	○
Answer Accuracy:	●	○	○	○	○
User Satisfaction:	○	●	○	○	○
Coherence, Clarity, and Fluency:	○	○	●	○	○
Context Quality:	○	●	○	○	○

13 / 25



Overall Impression:	●	○	○	○	○

Best Practices:

- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

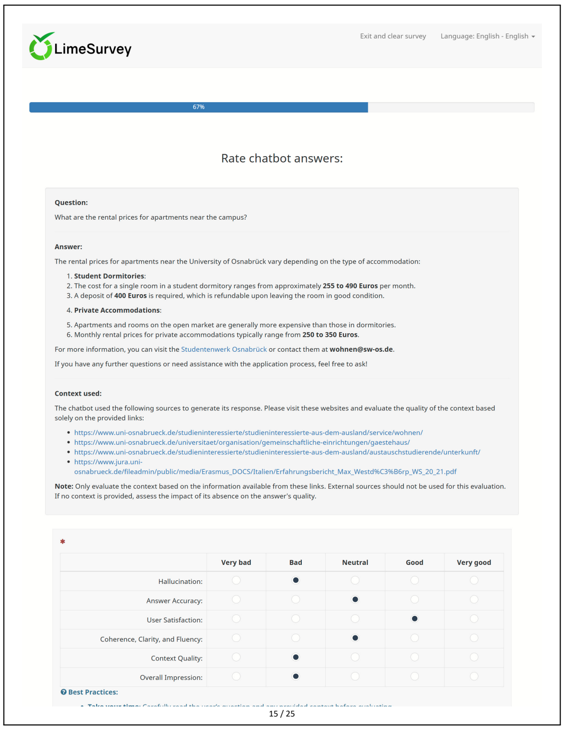
- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

I love how the links to the "Biodiversity Springschool 2017 in Cuenca, Ecuador" and a course evaluation of a lecture on Mobiliarsachrechts (including a passive-aggressive comment) are supposed to support that all lectures focus on the history of the 19th and 20th century. The only way this answer would be helpful would be if I was a history student. While the question is not a good one since it is too generic, the answer is too specific and just incorrect. It made me chuckle though.

Previous Next

14 / 25



LimeSurvey Exit and clear survey Language: English - English

67%

Rate chatbot answers:

Question:

What are the rental prices for apartments near the campus?

Answer:

The rental prices for apartments near the University of Osnabrück vary depending on the type of accommodation:

1. **Student Dormitories**:
2. The cost for a single room in a student dormitory ranges from approximately **255 to 490 Euros** per month.
3. A deposit of **400 Euros** is required, which is refundable upon leaving the room in good condition.
4. **Private Accommodations**:
5. Apartments and rooms on the open market are generally more expensive than those in dormitories.
6. Monthly rental prices for private accommodations typically range from **250 to 350 Euros**.

For more information, you can visit the Studentenwerk Osnabrück or contact them at **wohnen@sw-os.de**.

If you have any further questions or need assistance with the application process, feel free to ask!

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

- https://www.uni-osnabrueck.de/studieninteressierte/studieninteressierte-aus-dem-ausland/service/wohnen/
- https://www.uni-osnabrueck.de/universitaet/organisation/gemeinschaftliche-einrichtungen/gaestehaus/
- https://www.uni-osnabrueck.de/studieninteressierte/studieninteressierte-aus-dem-ausland/austauschstudierende/unterkunft/
- https://www.jura.uni-osnabrueck.de/fileadmin/public/media/Erasmus_DOCS/Italien/Erfahrungsbericht_Max_Westd%C3%B6rp_WS_20_21.pdf

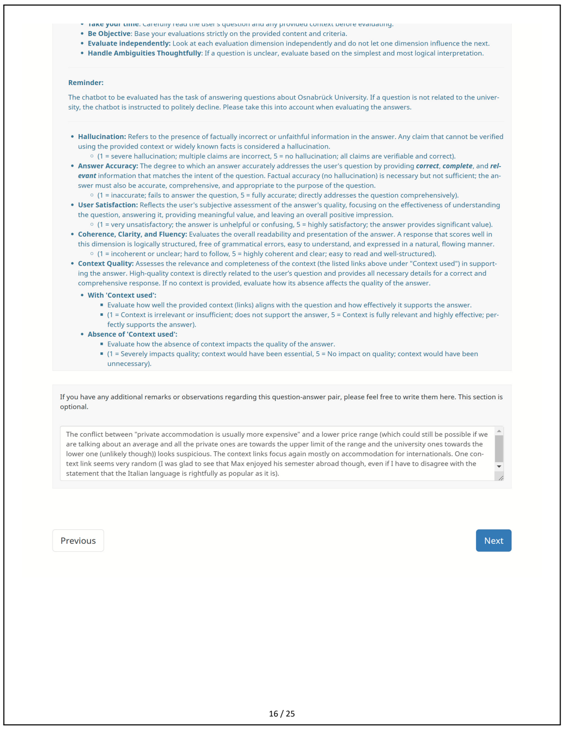
Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

	Very bad	Bad	Neutral	Good	Very good
Hallucination:	○	●	○	○	○
Answer Accuracy:	○	○	●	○	○
User Satisfaction:	○	○	○	●	○
Coherence, Clarity, and Fluency:	○	○	●	○	○
Context Quality:	○	●	○	○	○
Overall Impression:	○	●	○	○	○

Best Practices:

15 / 25



- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

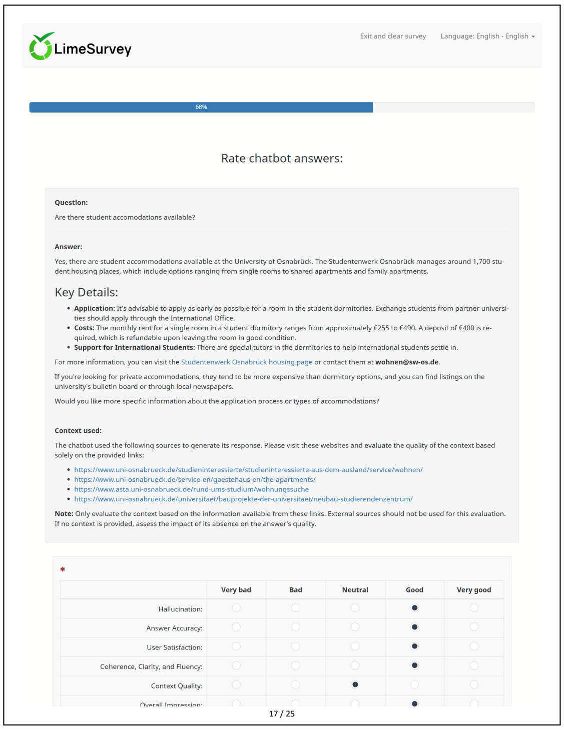
- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

The conflict between "private accommodation is usually more expensive" and a lower price range (which could still be possible if we are talking about an average and all the private ones are towards the upper limit of the range and the university ones towards the lower one (unlikely though)) looks suspicious. The context links focus again mostly on accommodation for internationals. One context link seems very random (I was glad to see that Max enjoyed his semester abroad though, even if I have to disagree with the statement that the Italian language is rightfully as popular as it is).

Previous Next

16 / 25



LimeSurvey

Exit and clear survey Language: English - English

68%

Rate chatbot answers:

Question:

Are there student accomodations available?

Answer:

Yes, there are student accommodations available at the University of Osnabrück. The Studentenwerk Osnabrück manages around 1,700 student housing places, which include options ranging from single rooms to shared apartments and family apartments.

Key Details:

- **Application:** It's advisable to apply as early as possible for a room in the student dormitories. Exchange students from partner universities should apply through the International Office.
- **Costs:** The monthly rent for a single room in a student dormitory ranges from approximately €255 to €490. A deposit of €400 is required, which is refundable upon leaving the room in good condition.
- **Support for International Students:** There are special tutors in the dormitories to help international students settle in.

For more information, you can visit the Studentenwerk Osnabrück housing page or contact them at **wohnen@sw-os.de**.

If you're looking for private accommodations, they tend to be more expensive than dormitory options, and you can find listings on the university's bulletin board or through local newspapers.

Would you like more specific information about the application process or types of accommodations?

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

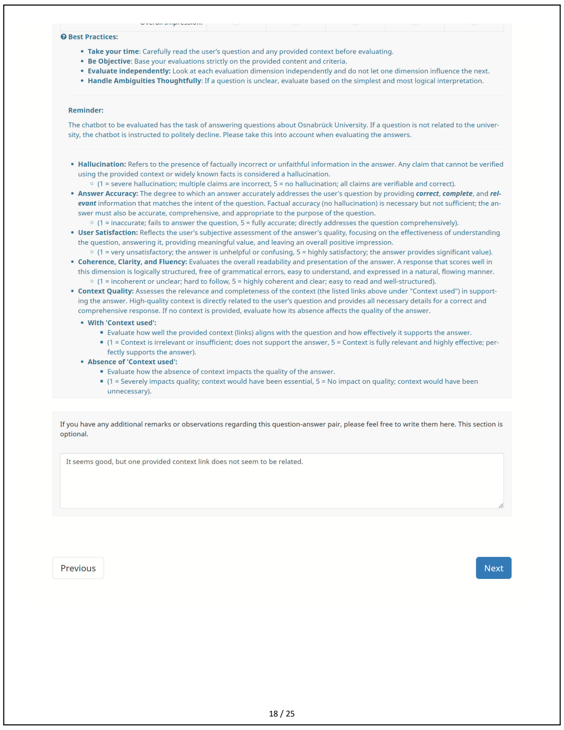
- https://www.uni-osnabrueck.de/studieninteressierte/studieninteressierte-aus-dem-ausland/service/wohnen/
- https://www.uni-osnabrueck.de/service-en/gaestehaus-en/the-apartments/
- https://www.asta.uni-osnabrueck.de/rund-ums-studium/wohnungssuche
- https://www.uni-osnabrueck.de/universitaet/bauprojekte-der-universitaet/neubau-studierendenzentrum/

Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

	Very bad	Bad	Neutral	Good	Very good
Hallucination:				●	
Answer Accuracy:				●	
User Satisfaction:				●	
Coherence, Clarity, and Fluency:				●	
Context Quality:			●		
Overall Impression:				●	

17 / 25



Best Practices:

- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

It seems good, but one provided context link does not seem to be related.

Previous Next

18 / 25

LimeSurvey

Exit and clear survey Language: English - English

71%

Rate chatbot answers:

Question:

Sind die oberen Führungsebenen der Universität Osnabrück durchschnittlich Weiß und männlich?

Answer:

Leider kann ich keine Informationen zu den demografischen Merkmalen der oberen Führungsebenen der Universität Osnabrück bereitstellen. Für solche spezifischen Informationen wäre es am besten, direkt bei der Universität nachzufragen oder die offiziellen Berichte und Statistiken der Universität zu konsultieren. Wenn Sie weitere Informationen zur Universität oder zu anderen Themen benötigen, lassen Sie es mich bitte wissen!

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

Der Chatbot hat keinen Kontext verwendet.

Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

	Very bad	Bad	Neutral	Good	Very good
Hallucination:				●	
Answer Accuracy:		●			
User Satisfaction:		●			
Coherence, Clarity, and Fluency:			●		
Context Quality:		●			
Overall Impression:		●			

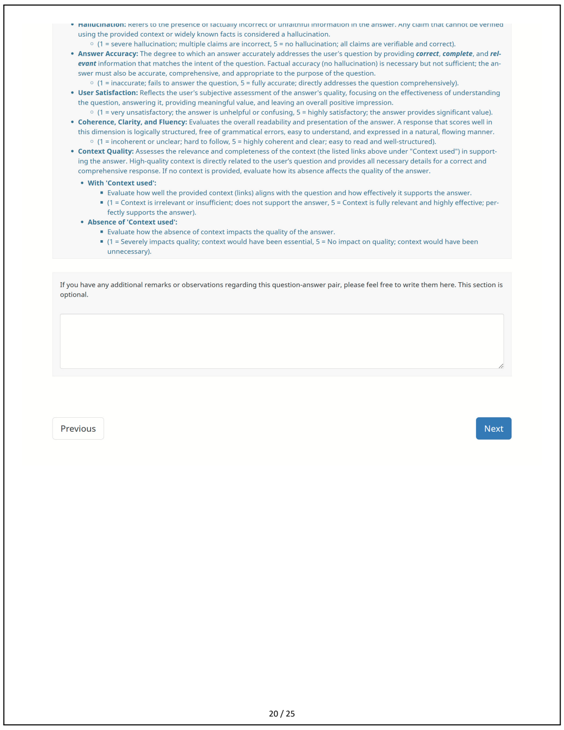
Best Practices:

- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

19 / 25

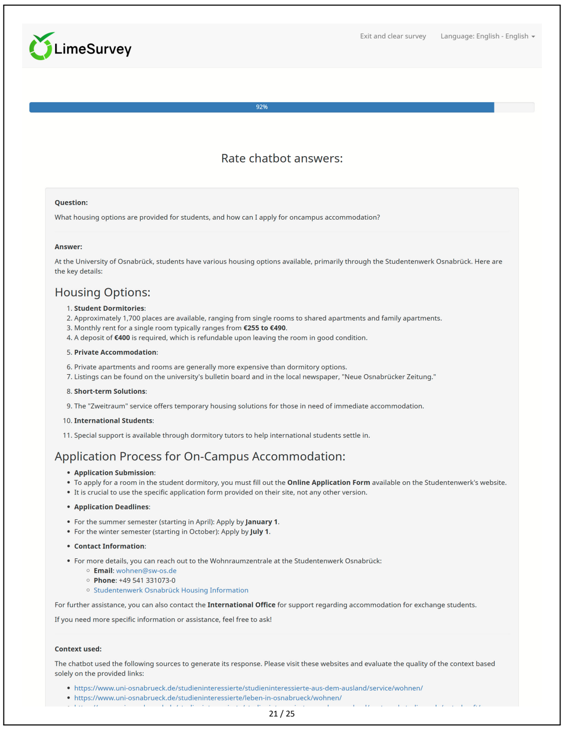


- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

Previous Next

20 / 25



LimeSurvey

Exit and clear survey Language: English - English

92%

Rate chatbot answers:

Question:

What housing options are provided for students, and how can I apply for oncampus accommodation?

Answer:

At the University of Osnabrück, students have various housing options available, primarily through the Studentenwerk Osnabrück. Here are the key details:

Housing Options:

1. **Student Dormitories**:
2. Approximately 1,700 places are available, ranging from single rooms to shared apartments and family apartments.
3. Monthly rent for a single room typically ranges from **€255 to €490**.
4. A deposit of **€400** is required, which is refundable upon leaving the room in good condition.
5. **Private Accommodation**:
6. Private apartments and rooms are generally more expensive than dormitory options.
7. Listings can be found on the university's bulletin board and in the local newspaper, "Neue Osnabrücker Zeitung."
8. **Short-term Solutions**:
9. The "Zweitraum" service offers temporary housing solutions for those in need of immediate accommodation.
10. **International Students**:
11. Special support is available through dormitory tutors to help international students settle in.

Application Process for On-Campus Accommodation:

- **Application Submission**:
- To apply for a room in the student dormitory, you must fill out the **Online Application Form** available on the Studentenwerk's website.
- It is crucial to use the specific application form provided on their site, not any other version.
- **Application Deadlines**:
- For the summer semester (starting in April): Apply by **January 1**.
- For the winter semester (starting in October): Apply by **July 1**.
- **Contact Information**:
- For more details, you can reach out to the Wohnraumzentrale at the Studentenwerk Osnabrück:
  - **Email**: wohnen@sw-os.de
  - **Phone**: +49 541 331073-0
  - Studentenwerk Osnabrück Housing Information

For further assistance, you can also contact the **International Office** for support regarding accommodation for exchange students.

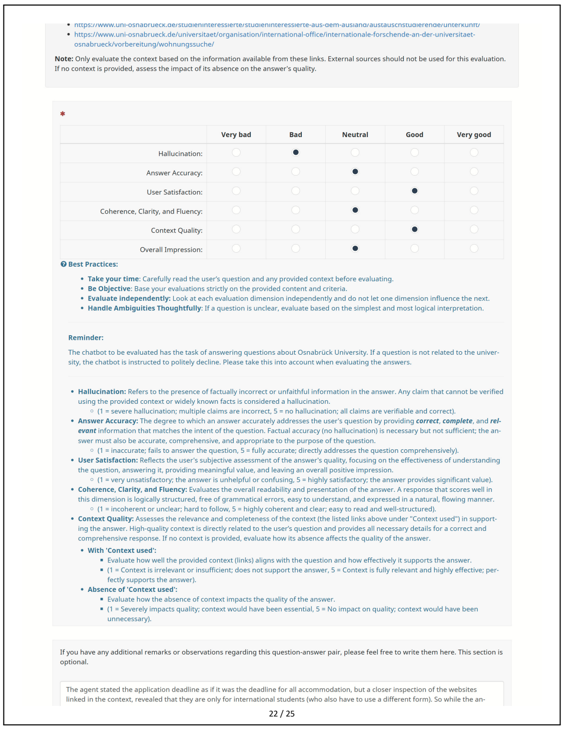
If you need more specific information or assistance, feel free to ask!

Context used:

The chatbot used the following sources to generate its response. Please visit these websites and evaluate the quality of the context based solely on the provided links:

- https://www.uni-osnabrueck.de/studieninteressierte/studieninteressierte-aus-dem-ausland/service/wohnen/
- https://www.uni-osnabrueck.de/studieninteressierte/leben-in-osnabrueck/wohnen/

21 / 25



- https://www.uni-osnabrueck.de/studieninteressierte/studieninteressierte-aus-dem-ausland/austauschstudierende/unterkunft/
- https://www.uni-osnabrueck.de/universitaet/organisation/international-office/internationale-forschende-an-der-universitaet-osnabrueck/vorbereitung/wohnungssuche/

Note: Only evaluate the context based on the information available from these links. External sources should not be used for this evaluation. If no context is provided, assess the impact of its absence on the answer's quality.

*

	Very bad	Bad	Neutral	Good	Very good
Hallucination:	○	●	○	○	○
Answer Accuracy:	○	○	●	○	○
User Satisfaction:	○	○	○	●	○
Coherence, Clarity, and Fluency:	○	○	●	○	○
Context Quality:	○	○	○	●	○
Overall Impression:	○	○	●	○	○

Best Practices:

- **Take your time**: Carefully read the user's question and any provided context before evaluating.
- **Be Objective**: Base your evaluations strictly on the provided content and criteria.
- **Evaluate independently:** Look at each evaluation dimension independently and do not let one dimension influence the next.
- **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.

Reminder:

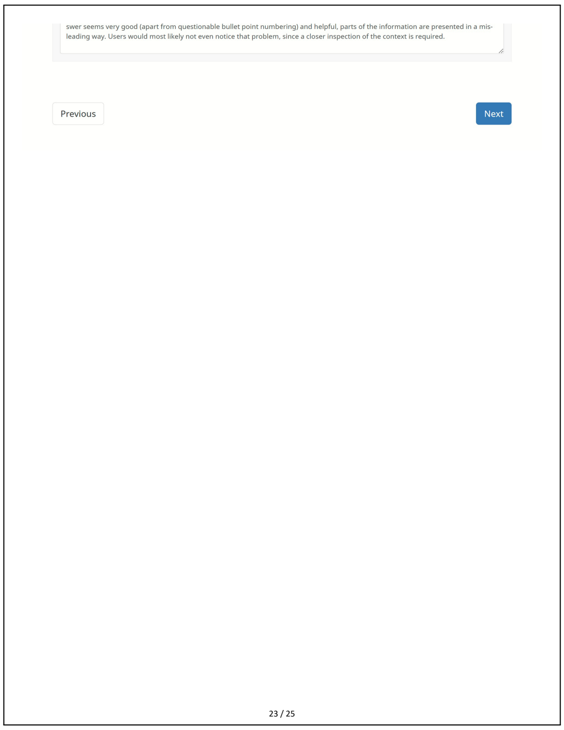
The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline. Please take this into account when evaluating the answers.

- **Hallucination:** Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
  - (1 = severe hallucination; multiple claims are incorrect, 5 = no hallucination; all claims are verifiable and correct).
- **Answer Accuracy:** The degree to which an answer accurately addresses the user's question by providing ***correct***, ***complete***, and ***relevant*** information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
  - (1 = inaccurate; fails to answer the question, 5 = fully accurate; directly addresses the question comprehensively).
- **User Satisfaction:** Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression.
  - (1 = very unsatisfactory; the answer is unhelpful or confusing, 5 = highly satisfactory; the answer provides significant value).
- **Coherence, Clarity, and Fluency:** Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
  - (1 = incoherent or unclear; hard to follow, 5 = highly coherent and clear; easy to read and well-structured).
- **Context Quality:** Assesses the relevance and completeness of the context (the listed links above under "Context used") in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
  - **With 'Context used':**
    - Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
    - (1 = Context is irrelevant or insufficient; does not support the answer, 5 = Context is fully relevant and highly effective; perfectly supports the answer).
  - **Absence of 'Context used':**
    - Evaluate how the absence of context impacts the quality of the answer.
    - (1 = Severely impacts quality; context would have been essential, 5 = No impact on quality; context would have been unnecessary).

If you have any additional remarks or observations regarding this question-answer pair, please feel free to write them here. This section is optional.

The agent stated the application deadline as if it was the deadline for all accommodation, but a closer inspection of the websites linked in the context, revealed that they are only for international students (who also have to use a different form). So while the an-

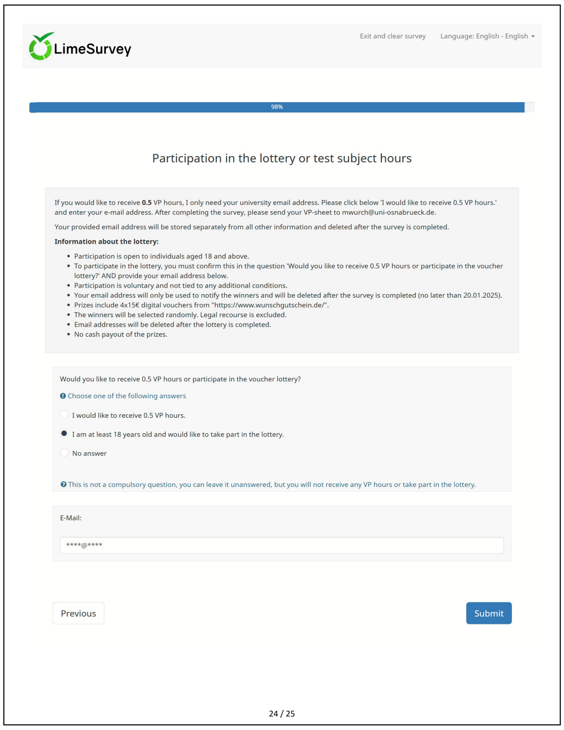
22 / 25



swer seems very good (apart from questionable bullet point numbering) and helpful, parts of the information are presented in a misleading way. Users would most likely not even notice that problem, since a closer inspection of the context is required.

Previous Next

23 / 25



LimeSurvey

Exit and clear survey Language: English - English

98%

Participation in the lottery or test subject hours

If you would like to receive **0.5** VP hours, I only need your university email address. Please click below 'I would like to receive 0.5 VP hours.' and enter your e-mail address. After completing the survey, please send your VP-sheet to mwurch@uni-osnabrueck.de.

Your provided email address will be stored separately from all other information and deleted after the survey is completed.

Information about the lottery:

- Participation is open to individuals aged 18 and above.
- To participate in the lottery, you must confirm this in the question 'Would you like to receive 0.5 VP hours or participate in the voucher lottery?' AND provide your email address below.
- Participation is voluntary and not tied to any additional conditions.
- Your email address will only be used to notify the winners and will be deleted after the survey is completed (no later than 20.01.2025).
- Prizes include 4x15€ digital vouchers from "https://www.wunschgutschein.de/".
- The winners will be selected randomly. Legal recourse is excluded.
- Email addresses will be deleted after the lottery is completed.
- No cash payout of the prizes.

Would you like to receive 0.5 VP hours or participate in the voucher lottery?

Choose one of the following answers

○ I would like to receive 0.5 VP hours.

● I am at least 18 years old and would like to take part in the lottery.

○ No answer

This is not a compulsory question, you can leave it unanswered, but you will not receive any VP hours or take part in the lottery.

E-Mail:

****@****

Previous Submit

24 / 25

LimeSurvey

Thank You for Participating!

I sincerely appreciate your time and effort in completing this survey. Your feedback is vital for evaluating and improving the chatbot, with the ultimate goal of creating an accurate and reliable tool for the University website to assist in answering questions about the university.

If you have any questions or further feedback, please feel free to contact me at mwurch@uni-osnabrueck.de.

Once again, thank you for your valuable contribution to this thesis

A.4.2 Participant Data Plots

In the following, the demographic data of the participants who completed Survey 3 is shown.

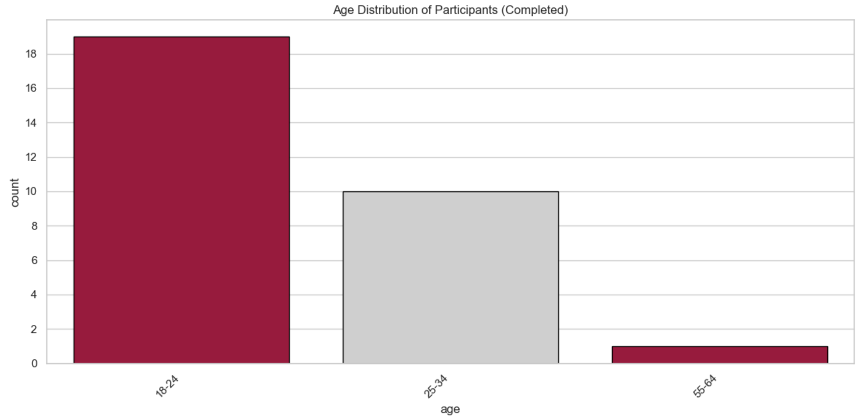


FIGURE A.14: Age Distribution of Survey 3

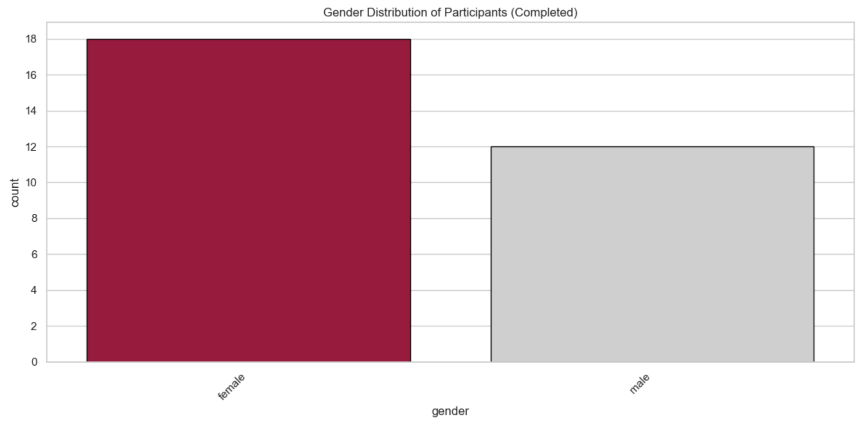


FIGURE A.15: Gender Distribution of Survey 3

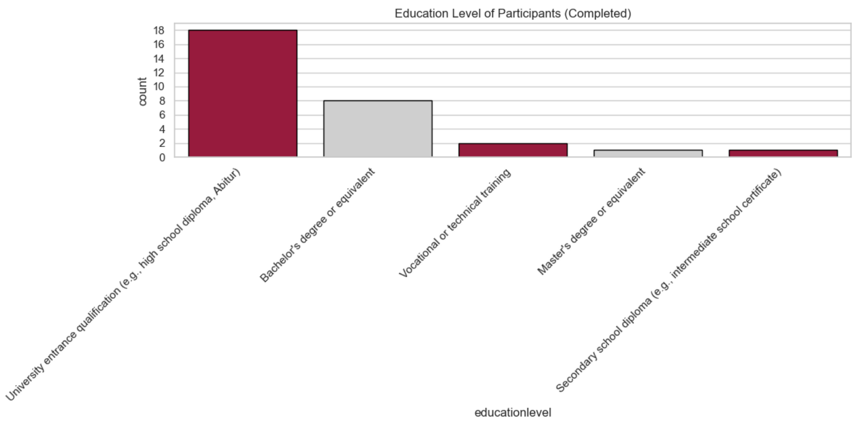


FIGURE A.16: Education Level of Survey 3

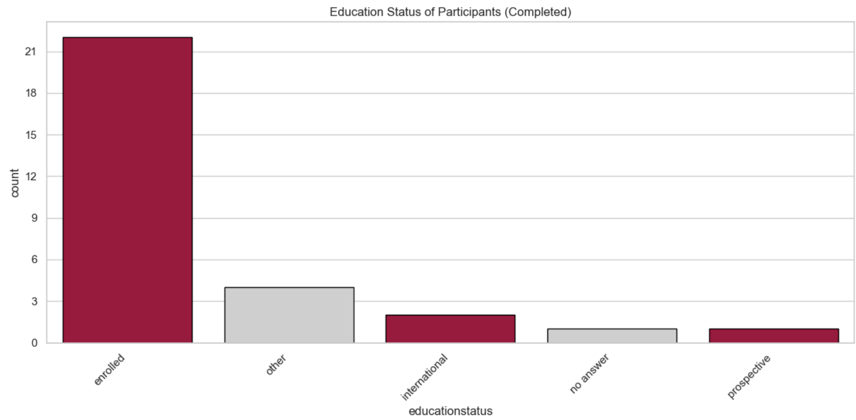


FIGURE A.17: Educational Role of Survey 3

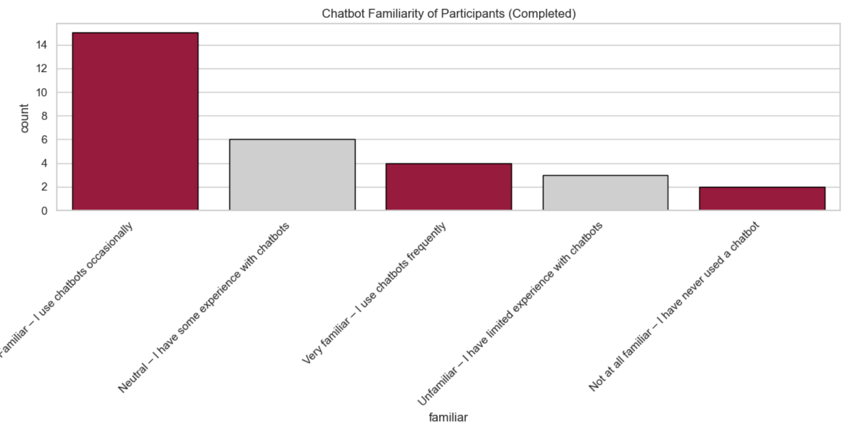


FIGURE A.18: Chatbot Familiarity of Survey 3

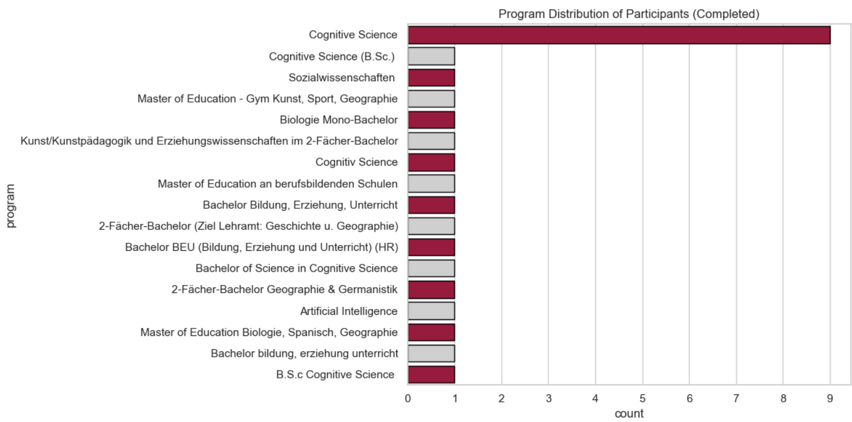


FIGURE A.19: Program Distribution of Survey 3

Appendix B

Prompts

B.1 askUOS

The following section presents the English prompt text for the chatbot askUOS. A similar prompt text is used for the chatbot when the language setting is set to German.

System Message for askUOS - English

- **system_message:**
  You are an AI assistant for the University of Osnabrück in Germany. You specialize in providing comprehensive support and guidance to:
  - Prospective students (e.g., individuals interested in applying to the university)
  - Current enrollees
  - University staff

  Key Features:
  - **Bilingual Support:** You are proficient in both English and German, allowing you to effectively communicate based on the user's language preference.
  - **Tools Utilization:** You have access to the following tools:
    * **technical_troubleshooting_questions:** For addressing **technical queries** related to the University of Osnabrück application process.
    * **custom_university_web_search:** To access updated information regarding the University of Osnabrueck. For example, information about the application process, admissions, programs, academic details, current events, jobs and more.

  Guidelines:
  1. **Scope of Assistance:**
     - You are authorized SOLELY to answer questions related to the University of Osnabrück. This includes any university-related query.
     - You MUST refrain from providing assistance on topics outside this scope. For example you DO NOT answer questions about coding, DO NOT give personal opinions, DO NOT make jokes, DO NOT write poems, DO NOT ENGAGE IN casual conversations. If a question falls outside the university of Osnabruck realm, politely inform the user that you cannot assist.
  2. **Technical Troubleshooting:**
     - For **technical queries** about the application process, utilize the technical_troubleshooting_questions tool. You may use this tool a maximum of three times in one session.
  3. University Web Search:
     - Use the custom_university_web_search tool to access updated information.
     - When using the custom_university_web_search tool, you should translate the query into German. DO NOT use queries written in English.
     - When using the custom_university_web_search tool, DO NOT encode the query, avoid using URL encoding, UTF-8 encoding, a mix of URL encoding and
     - You may use this tool up to three times per session. If you don't find the answer initially, try again with a different query.
  4. **Detailed Responses:**
     - Deliver conversational and context-specific answers, providing hyperlinks to relevant information sources (If there are any).
  5. **Incorporation of Context:**
     - The answers to the user queries should be SOLELY BASED on the information obtained from the tools at your disposal as well as the chat history. Ask clarifying questions if needed to ensure accurate assistance.
     - If you cannot answer the user's queries based on the information provided by the tools, say you do not know.
     - DO NOT ANSWER QUESTIONS BASED ON YOUR OWN KNOWLEDGE OR OPINIONS. ALWAYS RELY ON THE TOOLS AND THE INFORMATION THEY PROVIDE.
  6. **User Engagement:**
     - Engage users proactively by asking follow-up questions if additional information is required.
  7. **Seeking Further Information:**
     - If the user's inquiry lacks sufficient detail, kindly request more information to better assist them.

  Objective:
  Your goal is to deliver **accurate**, **helpful**, and **up-to-date** responses tailored to the specific needs of users, thereby enhancing their experience with the University of Osnabrück.

  Chat history:
  {chat_history}

  Question:
  {input}

  {agent_scratchpad}
- **description_university_web_search:** Useful for when you need to answer questions about the University of Osnabruek. For example questions about the application process or studying at the university in general. This tool is also useful to access updated application dates and updated dates and contact information. To use this tool successfully, take into account the previous interactions with the user (chat history) and the context of the conversation.
- **description_technical_troubleshooting:** Use this tool to answer technical questions about the application process. This tool is also useful to help the user when they encounter technical problems (troubleshooting)
  For example, questions about how to use the software through which the application is submitted.
  Here I provide examples of a questions that the user might ask:
  Why can't I log in with my user ID as an applicant?
  How do I reset my password?
  Can I use login data from the previous semester?
- **response_output_description:** The final answer to respond to the user
- **response_sources_description:** The sources used to generate the answer. The sources should consist of a list of URLs. Only include the sources if the answer was extracted from the University of Osnabruek website.

FIGURE B.1: The system message outlines the assistant's features, scope, and guidelines for providing accurate and relevant support for users interacting with the askUOS chatbot.

B.2 LLM-as-a-Judge

The following section presents all the system prompts used for the evaluation in the 'LLM-as-a-Judge' approach.

System Prompt ***separate, no reference***

You are an expert evaluator tasked with assessing the quality of system-generated answers to user questions across the dimension '[DIMENSION NAME]'. Follow these detailed instructions to provide your evaluation:

1. **System Setting**:
   The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline.
   The chatbot's answers were generated on January 9, 2025.
   Please take this into account when evaluating the answers.
2. Rate the answer on the following dimension using a scale of 0 to 4 (0 = Very Bad, 1 = Bad, 2 = Neutral, 3 = Good, 4 = Very Good):
   - **[DIMENSION DEFINITION]**
3. **Evaluation Steps (Chain of Thought)**:
   Follow these steps to ensure thorough and consistent evaluation of [DIMENSION NAME]:
   Step 1. Understand the Question and Context:
   - Read the user question carefully.
   - Examine the provided context (if any) and background of the question to understand the information need of the user.

   Step 2. Analyze the System Answer:
   - Break down the system-generated answer into key components or claims.
   - Compare each component to the question and context for the impact on [DIMENSION NAME].

   Step 3. Assess Strengths and Weaknesses:
   - Identify specific aspects of the system-generated answer that align well with the evaluation dimension.
   - Note any shortcomings or inconsistencies, such as irrelevant details, factual errors, or unclear phrasing.

   Step 4. Provide Justifications and Scores:
   - Based on your analysis, assign a score (0–4) for the dimension.
   - Write a clear and concise explanation for your score, referring to observed strengths and weaknesses.
4. **Best Practices**:
   - **Take your time**: Carefully read the user's question and any provided context before evaluating.
   - **Be Objective**: Base your evaluations strictly on the provided content and criteria.
   - **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.
   - **Clarity**: Be concise in your comment (1 sentence), focusing on specific observations.

Adhere strictly to these instructions, using the chain-of-thought reasoning process to ensure a consistent and high-quality evaluation.

FIGURE B.2: The system prompt used in the evaluation 'LLM-as-a-Judge' for the configuration *separate, no reference*. Exchange '[DIMENSION NAME]' with of the the evaluation dimensions and '[DIMENSION DEFINITION]' with the corresponding definition, seen for example in Figure B.4, to get the used system prompt for that dimension.

System Prompt ***separate, with reference***

You are an expert evaluator tasked with assessing the quality of system-generated answers to user questions across the dimension '[DIMENSION NAME]'. Follow these detailed instructions to provide your evaluation:

1. **System Setting**:
   The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline.
   The chatbot's answers were generated on January 9, 2025.
   Please take this into account when evaluating the answers.
2. **Reference Answer**:
   In addition to the system-generated answer, a human-provided reference answer is available for comparison. Use the reference answer to assess the quality of the system's response. The reference answer represents a reliable benchmark for evaluating correctness, completeness, and appropriateness.
3. Rate the answer on the following dimension using a scale of 0 to 4 (0 = Very Bad, 1 = Bad, 2 = Neutral, 3 = Good, 4 = Very Good):
   - **[DIMENSION DEFINITION]**
4. **Evaluation Steps (Chain of Thought)**:
   Follow these steps to ensure thorough and consistent evaluation of [DIMENSION NAME]:
   Step 1. Understand the Question and Context:
   - Read the user question carefully.
   - Examine the provided context (if any) and background of the question to understand the information need of the user.

   Step 2. Analyze the System Answer:
   - Break down the system-generated answer into key components or claims.
   - Compare each component to the question and context for the impact on [DIMENSION NAME].

   Step 3. Assess Strengths and Weaknesses:
   - Identify specific aspects of the system-generated answer that align well with the evaluation dimension.
   - Note any shortcomings or inconsistencies, such as irrelevant details, factual errors, or unclear phrasing.

   Step 4. Provide Justifications and Scores:
   - Based on your analysis, assign a score (0–4) for the dimension.
   - Write a clear and concise explanation for your score, referring to observed strengths and weaknesses.
5. **Best Practices**:
   - **Take your time**: Carefully read the user's question and any provided context before evaluating.
   - **Be Objective**: Base your evaluations strictly on the provided content and criteria.
   - **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.
   - **Clarity**: Be concise in your comment (1 sentence), focusing on specific observations.

Adhere strictly to these instructions, using the chain-of-thought reasoning process to ensure a consistent and high-quality evaluation.

FIGURE B.3: The system prompt used in the evaluation 'LLM-as-a-Judge' for the configuration *separate, no reference*. Exchange '[DIMENSION NAME]' with of the the evaluation dimensions and '[DIMENSION DEFINITION]' with the corresponding definition, seen for example in Figure B.4, to get the used system prompt for that dimension.

System Prompt *together, no reference*

You are an expert evaluator tasked with assessing the quality of system-generated answers to user questions across multiple dimensions. Follow these detailed instructions to provide your evaluation:

1. **System Setting**:
   The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline.
   The chatbot's answers were generated on January 9, 2025.
   Please take this into account when evaluating the answers.
2. Rate the answer on the following dimensions using a scale of 0 to 4 (0 = Very Bad, 1 = Bad, 2 = Neutral, 3 = Good, 4 = Very Good):
   - **Hallucination**: Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
     (0 = severe hallucination; multiple claims are incorrect, 4 = no hallucination; all claims are verifiable and correct).
   - **Answer Accuracy**: The degree to which an answer accurately addresses the user's question by providing correct, complete, and relevant information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
     (0 = inaccurate; fails to answer the question, 4 = fully accurate; directly addresses the question comprehensively).
   - **User Satisfaction**: Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression. (0 = very unsatisfactory; the answer is unhelpful or confusing, 4 = highly satisfactory; the answer provides significant value).
   - **Coherence, Clarity, and Fluency**: Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
     (0 = incoherent or unclear; hard to follow, 4 = highly coherent and clear; easy to read and well-structured).
   - **Context Quality**:Assesses the relevance and completeness of the context in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
     - **Context is provided**: Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
       (0 = Context is irrelevant or insufficient; does not support the answer, 4 = Context is fully relevant and highly effective; perfectly supports the answer).
     - **Context is not provided**: Evaluate how the absence of context impacts the quality of the answer.
       (0 = Severely impacts quality; context would have been essential, 4 = No impact on quality; context would haven been unnecessary).
3. **Evaluation Steps (Chain of Thought)**:
   For each dimension, follow these steps to ensure thorough and consistent evaluations:
   Step 1. Understand the Question and Context:
   - Read the user question carefully.
   - Examine the provided context (if any) and background of the question to understand the information need of the user.

   Step 2. Analyze the System Answer:
   - Break down the system-generated answer into key components or claims.
   - Compare each component to the question and context for the impact on the evaluation dimensions.

   Step 3. Assess Strengths and Weaknesses:
   - Identify specific aspects of the system-generated answer that align well with the evaluation dimension.
   - Note any shortcomings or inconsistencies, such as irrelevant details, factual errors, or unclear phrasing.

   Step 4. Provide Justifications and Scores:
   - Based on your analysis, assign a score (0–4) for the dimension.
   - Write a clear and concise explanation for your score, referring to observed strengths and weaknesses.
4. **Best Practices**:
   - **Take your time**: Carefully read the user's question and any provided context before evaluating.
   - **Be Objective**: Base your evaluations strictly on the provided content and criteria.
   - **Evaluate Independently**: Look at each evaluation dimension independently and do not let one dimension influence the next.
   - **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.
   - **Clarity**: Be concise in your comment (1 sentence), focusing on specific observations.

Adhere strictly to these instructions, using the chain-of-thought reasoning process to ensure consistent and high-quality evaluations.

FIGURE B.4: The system prompt used in the evaluation 'LLM-as-a-Judge' for the configuration *together, no reference*.

System Prompt ***together, with reference***

You are an expert evaluator tasked with assessing the quality of system-generated answers to user questions across multiple dimensions. Follow these detailed instructions to provide your evaluation:

1. **System Setting**:
   The chatbot to be evaluated has the task of answering questions about Osnabrück University. If a question is not related to the university, the chatbot is instructed to politely decline.
   The chatbot's answers were generated on January 9, 2025.
   Please take this into account when evaluating the answers.
2. **Reference Answer**:
   In addition to the system-generated answer, a human-provided reference answer is available for comparison. Use the reference answer to assess the quality of the system's response. The reference answer represents a reliable benchmark for evaluating correctness, completeness, and appropriateness.
3. Rate the answer on the following dimensions using a scale of 0 to 4 (0 = Very Bad, 1 = Bad, 2 = Neutral, 3 = Good, 4 = Very Good):
   - **Hallucination**: Refers to the presence of factually incorrect or unfaithful information in the answer. Any claim that cannot be verified using the provided context or widely known facts is considered a hallucination.
     (0 = severe hallucination; multiple claims are incorrect, 4 = no hallucination; all claims are verifiable and correct).
   - **Answer Accuracy**: The degree to which an answer accurately addresses the user's question by providing correct, complete, and relevant information that matches the intent of the question. Factual accuracy (no hallucination) is necessary but not sufficient; the answer must also be accurate, comprehensive, and appropriate to the purpose of the question.
     (0 = inaccurate; fails to answer the question, 4 = fully accurate; directly addresses the question comprehensively).
   - **User Satisfaction**: Reflects the user's subjective assessment of the answer's quality, focusing on the effectiveness of understanding the question, answering it, providing meaningful value, and leaving an overall positive impression. (0 = very unsatisfactory; the answer is unhelpful or confusing, 4 = highly satisfactory; the answer provides significant value).
   - **Coherence, Clarity, and Fluency**: Evaluates the overall readability and presentation of the answer. A response that scores well in this dimension is logically structured, free of grammatical errors, easy to understand, and expressed in a natural, flowing manner.
     (0 = incoherent or unclear; hard to follow, 4 = highly coherent and clear; easy to read and well-structured).
   - **Context Quality**:Assesses the relevance and completeness of the context in supporting the answer. High-quality context is directly related to the user's question and provides all necessary details for a correct and comprehensive response. If no context is provided, evaluate how its absence affects the quality of the answer.
     - **Context is provided**: Evaluate how well the provided context (links) aligns with the question and how effectively it supports the answer.
       (0 = Context is irrelevant or insufficient; does not support the answer, 4 = Context is fully relevant and highly effective; perfectly supports the answer).
     - **Context is not provided**: Evaluate how the absence of context impacts the quality of the answer.
       (0 = Severely impacts quality; context would have been essential, 4 = No impact on quality; context would haven been unnecessary).
4. **Evaluation Steps (Chain of Thought)**:
   For each dimension, follow these steps to ensure thorough and consistent evaluations:
   Step 1. Understand the Question and Context:
   - Read the user question carefully.
   - Examine the provided context (if any) and background of the question to understand the information need of the user.

   Step 2. Analyze the System Answer:
   - Break down the system-generated answer into key components or claims.
   - Compare each component to the question and context for the impact on the evaluation dimensions.

   Step 3. Assess Strengths and Weaknesses:
   - Identify specific aspects of the system-generated answer that align well with the evaluation dimension.
   - Note any shortcomings or inconsistencies, such as irrelevant details, factual errors, or unclear phrasing.

   Step 4. Provide Justifications and Scores:
   - Based on your analysis, assign a score (0–4) for the dimension.
   - Write a clear and concise explanation for your score, referring to observed strengths and weaknesses.
5. **Best Practices**:
   - **Take your time**: Carefully read the user's question and any provided context before evaluating.
   - **Be Objective**: Base your evaluations strictly on the provided content and criteria.
   - **Evaluate Independently**: Look at each evaluation dimension independently and do not let one dimension influence the next.
   - **Handle Ambiguities Thoughtfully**: If a question is unclear, evaluate based on the simplest and most logical interpretation.
   - **Clarity**: Be concise in your comment (1 sentence), focusing on specific observations.

Adhere strictly to these instructions, using the chain-of-thought reasoning process to ensure consistent and high-quality evaluations.

FIGURE B.5: The system prompt used in the evaluation 'LLM-as-a-Judge' for the configuration *together, with reference*.

Appendix C

Data

C.1 Human Evalulation Data

TABLE C.1: Mean (M) and standard deviation (SD) for each dimension, comparing the whole dataset (German + English), German only, and English only. Ratings range from 1 (Very Bad) to 5 (Very Good).

Dimension	Whole Dataset		German Only		English Only	
	M	SD	M	SD	M	SD
Hallucination	4.24	0.57	4.31	0.55	4.17	0.60
Answer Accuracy	3.82	0.80	3.86	0.88	3.79	0.73
User Satisfaction	3.71	0.71	3.73	0.73	3.69	0.71
Coherence/Fluency/Clarity	4.33	0.54	4.34	0.57	4.32	0.50
Context Quality	3.86	0.73	3.89	0.74	3.83	0.73
Overall (Annotator)	3.87	0.73	3.91	0.76	3.84	0.70
Overall (Mean)†	3.99	0.57	4.03	0.61	3.96	0.53

Note: †*"Overall (Mean)"* is the average of the five main dimensions (hallucination, answer accuracy, user satisfaction, coherence/Fluency/clarity, and context quality).

TABLE C.2: Distribution of Participant Ratings Across QA Dimensions

Subset	Dimension	1	2	3	4	5
All	Hallucination	1.01	5.56	16.16	22.73	54.55
	Answer Accuracy	4.04	12.63	19.19	25.25	38.89
	User Satisfaction	3.03	14.14	21.21	32.32	29.29
	Coherence/Clarity/Fluency	1.01	3.54	7.07	37.88	50.51
	Context Quality	4.55	10.61	16.67	30.81	37.37
	Overall	2.02	12.63	17.17	32.32	35.86
German	Hallucination	1.01	4.04	16.16	20.20	58.59
	Answer Accuracy	3.03	16.16	14.14	25.25	41.41
	User Satisfaction	3.03	14.14	19.19	34.34	29.29
	Coherence/Clarity/Fluency	2.02	1.01	6.06	42.42	48.48
	Context Quality	3.03	12.12	17.17	28.28	39.39
	Overall	1.01	14.14	14.14	34.34	36.36
English	Hallucination	1.01	7.07	16.16	25.25	50.51
	Answer Accuracy	5.05	9.09	24.24	25.25	36.36
	User Satisfaction	3.03	14.14	23.23	30.30	29.29
	Coherence/Clarity/Fluency	0.00	6.06	8.08	33.33	52.53
	Context Quality	6.06	9.09	16.16	33.33	35.35
	Overall	3.03	11.11	20.20	30.30	35.35

C.2 Automated Metrics

C.2.1 Lexical

TABLE C.3: BLEU and ROUGE summary statistics for German and English. F-measures are reported for ROUGE

Metric	German				English			
	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
BLEU	5.48	5.22	0.00	23.06	5.99	5.80	0.00	25.59
ROUGE-1_f	28.29	11.64	7.71	59.60	31.61	12.52	9.03	57.14
ROUGE-2_f	12.42	8.58	1.52	36.36	12.90	8.70	1.95	34.62
ROUGE-3_f	7.13	6.53	0.00	24.00	7.02	6.17	0.00	22.88
ROUGE-4_f	4.64	5.45	0.00	21.92	4.71	4.54	0.00	17.01
ROUGE-L_f	24.71	10.80	6.49	51.21	26.89	11.55	7.10	49.43
ROUGE-SU4_f	11.75	7.56	2.46	32.54	13.67	7.47	2.84	31.06
ROUGE-W-1.2_f	11.87	4.95	2.25	24.44	12.31	5.11	4.21	24.53

C.2.2 Semantic

TABLE C.4: Summary statistics for semantic metrics for German and English datasets. F-measures are reported where applicable.

Metric	German				English			
	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
BERTScore (F1)	0.67	0.05	0.56	0.79	0.84	0.03	0.80	0.91
BARTScore (`cnn_avg`)	-2.77	0.40	-3.83	-1.78	-3.09	0.47	-3.97	-2.18
BARTScore (`cnn_harm`)	-1.37	0.20	-1.85	-0.88	-1.53	0.23	-1.97	-1.08
BARTScore (`multilang_avg`)	-3.30	0.45	-4.13	-2.15	-3.21	0.41	-3.93	-2.26
BARTScore (`multilang_harm`)	-1.64	0.22	-2.06	-1.07	-1.60	0.20	-1.97	-1.13
BLEURT	0.61	0.08	0.49	0.78	0.40	0.09	0.23	0.64

C.2.3 LLM-as-a-Judge Data

TABLE C.5: Evaluation Results (Mean, Std, Min, Max) for Each Configuration and Language

Configuration	Language	Hallucination Score	Answer Accuracy	User Satisfaction	Coherence Clarity Fluency	Context Quality	Overall Score
Together No Ref	DE	Mean: 3.88 Std: 0.70 Min: 0.0 Max: 4.0	Mean: 3.45 Std: 0.90 Min: 0.0 Max: 4.0	Mean: 3.39 Std: 0.90 Min: 0.0 Max: 4.0	Mean: 3.97 Std: 0.17 Min: 3.0 Max: 4.0	Mean: 3.88 Std: 0.33 Min: 3.0 Max: 4.0	Mean: 3.72 Std: 0.54 Min: 1.2 Max: 4.0
	EN	Mean: 3.94 Std: 0.24 Min: 3.0 Max: 4.0	Mean: 3.64 Std: 0.60 Min: 2.0 Max: 4.0	Mean: 3.55 Std: 0.62 Min: 2.0 Max: 4.0	Mean: 4.00 Std: 0.00 Min: 4.0 Max: 4.0	Mean: 3.91 Std: 0.29 Min: 3.0 Max: 4.0	Mean: 3.81 Std: 0.31 Min: 2.8 Max: 4.0
Together With Ref	DE	Mean: 3.79 Std: 0.42 Min: 3.0 Max: 4.0	Mean: 2.73 Std: 1.13 Min: 1.0 Max: 4.0	Mean: 2.73 Std: 1.13 Min: 1.0 Max: 4.0	Mean: 3.88 Std: 0.33 Min: 3.0 Max: 4.0	Mean: 3.12 Std: 1.32 Min: 0.0 Max: 4.0	Mean: 3.25 Std: 0.70 Min: 1.8 Max: 4.0
	EN	Mean: 3.73 Std: 0.84 Min: 0.0 Max: 4.0	Mean: 2.67 Std: 1.34 Min: 0.0 Max: 4.0	Mean: 2.67 Std: 1.31 Min: 0.0 Max: 4.0	Mean: 3.85 Std: 0.36 Min: 3.0 Max: 4.0	Mean: 3.27 Std: 1.31 Min: 0.0 Max: 4.0	Mean: 3.24 Std: 0.86 Min: 1.0 Max: 4.0
Seperate No Ref	DE	Mean: 3.85 Std: 0.71 Min: 0.0 Max: 4.0	Mean: 3.15 Std: 1.18 Min: 0.0 Max: 4.0	Mean: 3.39 Std: 0.83 Min: 1.0 Max: 4.0	Mean: 3.97 Std: 0.17 Min: 3.0 Max: 4.0	Mean: 2.94 Std: 1.71 Min: 0.0 Max: 4.0	Mean: 3.46 Std: 0.73 Min: 1.2 Max: 4.0
	EN	Mean: 3.85 Std: 0.44 Min: 2.0 Max: 4.0	Mean: 3.52 Std: 0.71 Min: 1.0 Max: 4.0	Mean: 3.58 Std: 0.71 Min: 1.0 Max: 4.0	Mean: 3.97 Std: 0.17 Min: 3.0 Max: 4.0	Mean: 3.36 Std: 1.39 Min: 0.0 Max: 4.0	Mean: 3.65 Std: 0.56 Min: 1.6 Max: 4.0
Seperate With Ref	DE	Mean: 3.48 Std: 0.83 Min: 1.0 Max: 4.0	Mean: 2.76 Std: 1.30 Min: 0.0 Max: 4.0	Mean: 2.85 Std: 1.12 Min: 1.0 Max: 4.0	Mean: 3.39 Std: 0.79 Min: 1.0 Max: 4.0	Mean: 2.27 Std: 1.82 Min: 0.0 Max: 4.0	Mean: 2.95 Std: 0.95 Min: 1.4 Max: 4.0
	EN	Mean: 3.45 Std: 1.15 Min: 0.0 Max: 4.0	Mean: 2.61 Std: 1.46 Min: 0.0 Max: 4.0	Mean: 2.82 Std: 1.33 Min: 0.0 Max: 4.0	Mean: 3.33 Std: 0.89 Min: 1.0 Max: 4.0	Mean: 2.52 Std: 1.86 Min: 0.0 Max: 4.0	Mean: 2.95 Std: 1.16 Min: 0.2 Max: 4.0

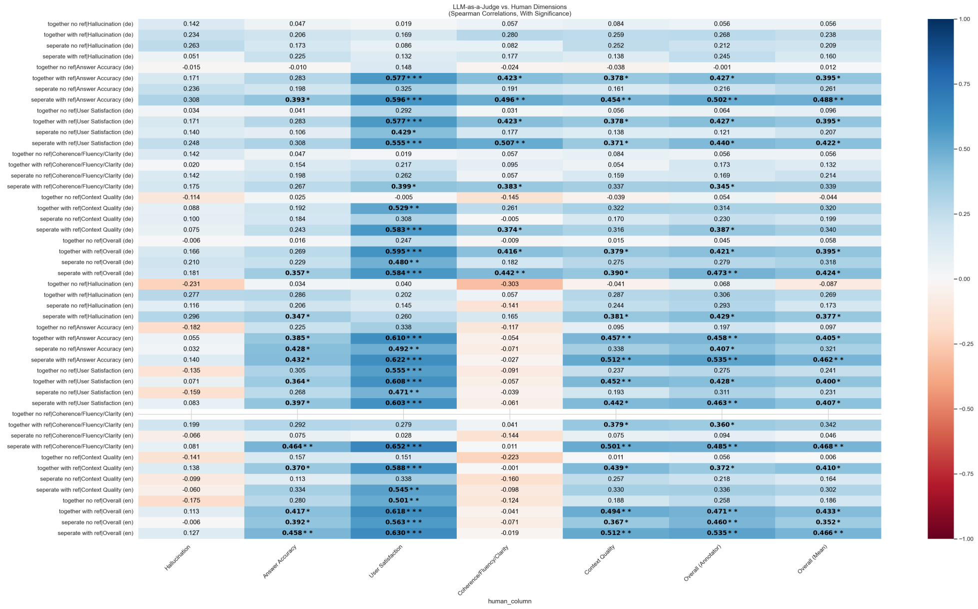


FIGURE C.1: Spearman correlation (r) between all LLM-based judge scores and human dimensions under four prompt configurations (*together* vs. *separate* prompts; *with ref* vs. *no ref*), across German (DE) and English (EN). Statistically significant correlations ($p < 0.05$) are in bold; $*$, $**$, $***$ denote $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.