Bachelor Degree Project

Evaluating the Impact of Hallucinations on User Trust and Satisfaction in LLM-based Systems

Author: Richard Willhelm Oelschlager
Supervisor: Aris Alissandrakis
Semester: VT 2024
Subject: Computer Science

Abstract

Hallucinations in LLMs refer to instances where the models generate outputs that are unrelated, incorrect, or misleading based on the input provided. This thesis investigates the impact of hallucinations in large language model (LLM)-based systems on user trust and satisfaction, a critical issue as AI becomes increasingly integrated into everyday applications. Hallucinations in LLMs—instances where the model generates incorrect or misleading information—pose significant challenges for user reliability and overall system effectiveness. Given the expanding role of AI in sectors requiring high trust levels, such as healthcare and finance, understanding and mitigating these errors is paramount.

To address this issue, a controlled experiment was designed to systematically assess how hallucinations affect user trust and satisfaction. Participants interacted with an AI system designed to exhibit varying levels of hallucinatory behavior. Quantitative measures of trust and satisfaction were collected through standardized questionnaires pre- and post-interaction, accompanied by statistical analyses to evaluate changes in user perception.

The results clearly demonstrate that hallucinations significantly diminish user trust and satisfaction, confirming the hypothesis that the accuracy of AI outputs is crucial for user reliance. These findings not only contribute to the academic discourse on human-AI interaction, but also have practical implications for AI developers and policymakers focusing on creating and regulating reliable AI technologies.

This study bridges a crucial knowledge gap and provides a foundation for future research aimed at developing more robust and trustworthy AI systems. Readers engaged in AI development, implementation, and policymaking will find the insights particularly relevant, encouraging further exploration into strategies that could enhance user trust in AI technologies.

Keywords: LLMs, hallucinations, ChatGPT, trust, satisfaction

Contents

1 Introduction **4**
1.1 Background 5
1.2 Related work 6
1.3 Problem formulation 7
1.4 Motivation 8
1.5 Results 8
1.6 Scope/Limitation 9
1.7 Target group 10
1.8 Outline 11
2 Method **12**
2.1 Research Project 12
2.2 Research methods 13
2.3 Reliability and Validity 14
2.4 Ethical Considerations 15
3 Theoretical Background **16**
3.1 Trust in Artificial Intelligence 16
3.2 User Satisfaction with Technology 17
3.3 Hallucinations in Large Language Models 18
3.4 Theoretical Framework for Examining Trust and Satisfaction 19
3.5 Summary 19
4 Research project – Implementation **20**
5 Results **21**
5.1 Participants. 22
6 Analysis **27**
7 Discussion **31**
8 Conclusions and Future Work **33**
References **36**
Appendix 1 - Google form **38**
Appendix 2 - Qualitative feedback **53**

1 Introduction

This thesis marks the culmination of my journey in the field of Computer Science, specifically tailored towards a 15 HEC Bachelor's degree. At its core, this project delves into the intricate world of Large Language Models (LLMs), a pivotal area of study within the broader discipline of Artificial Intelligence (AI) and Machine Learning (ML). The advent of LLMs, such as GPT-4+, developed by OpenAI, has significantly transformed how we interact with technology, offering unprecedented capabilities in generating human-like text, facilitating conversational agents, and enabling advanced content creation. GPT-4+ is an advanced iteration of the Generative Pre-trained Transformer models, leveraging a transformer architecture characterized by self-attention mechanisms. [1] These mechanisms allow the model to effectively weigh the importance of different parts of the input data, resulting in superior natural language understanding and generation. However, alongside their vast potential, LLMs introduce complex challenges, notably the phenomenon of hallucinations—instances where the models generate misleading, inaccurate, or entirely irrelevant outputs.

This research is situated at the intersection of technological innovation and user experience, focusing on the critical issue of hallucinations in LLMs and their impact on user trust and satisfaction. As these models become increasingly embedded in our digital lives, understanding and addressing the discrepancies between expected and actual outputs is paramount. This project aims to shed light on this underexplored area, offering insights into how hallucinations can erode trust and diminish user satisfaction, and proposing pathways to mitigate these effects. Through this introduction, I invite readers to embark on an exploration of the challenges and opportunities presented by LLMs, setting the stage for a comprehensive investigation that bridges the gap between technical capabilities and user-centric outcomes in the field of Computer Science.

1.1 Background

Large Language Models, such as GPT[2] (Generative Pretrained Transformer) series, have revolutionized natural language processing (NLP) by enabling more sophisticated text generation and comprehension. Their ability to understand context, generate coherent and diverse text, and support

conversational AI applications has made them integral to advancing human-computer interactions. LLMs underpin a variety of applications, from automated customer service bots to personalized content creation, showcasing their versatility and importance in today's digital landscape.

The applications of Large Language Models (LLMs) are vast and varied, extending beyond simple text generation to complex problem-solving, language translation, and more. This claim is based on several factors, including technological advancements, diverse applications, and the impact on user interaction and experience. LLMs have demonstrated capabilities in solving complex tasks such as coding, mathematics, and providing legal and medical advice, as well as in multilingual translation [2]. They are also used for creating sophisticated narratives, generating educational materials, and engaging in meaningful customer service interactions. As users increasingly interact with these models, the quality of generated content directly affects user experience. Trust and satisfaction become paramount, especially in applications requiring accurate information dissemination, such as news generation, educational content, and customer service. The ability of LLMs to produce high-quality, relevant, and contextually appropriate content is crucial, as inaccurate information can lead to misinformation with serious consequences. Therefore, maintaining high standards of content quality and reliability is essential to ensure user trust and satisfaction.

Despite their advancements, LLMs are prone to generating hallucinations—outputs that are either factually incorrect, misleading, or completely irrelevant [3]. This phenomenon poses a significant challenge to maintaining user trust and satisfaction, as it can lead to misinformation and confusion. The occurrence of hallucinations is not merely a technical glitch, but a fundamental issue that reflects on the model's understanding of context and its ability to generate reliable content.

Trust and satisfaction are crucial metrics in assessing the effectiveness of LLM applications. Hallucinations undermine both, as they can erode users' confidence in the system's reliability and its utility. For LLMs to be effectively integrated into practical applications, it's essential to address these issues. Understanding the extent of hallucinations' impact on user perceptions will guide the development of more robust models that can navigate the nuances of human language with greater accuracy.

1.2 Related work

The comprehensive study “Survey of Hallucination in Natural Language Generation” by Ji et al (2022) [3] published in ACM Computing Surveys provides an overview of hallucinations in various natural language generation (NLG) tasks, including machine translation. It discusses the causes of

hallucination, such as the incompetence of retrievers and the intrinsic and extrinsic nature of hallucinations in generation models. This survey is critical as it lays out the broad landscape of hallucinations in NLG, including the factors contributing to the problem, such as model limitations, data quality issues, and training techniques. It also covers mitigation techniques like improving retrieval mechanisms, combining retrieval with generation, enhancing data quality, and refining model training techniques.

Active research in this area is evident from the recent publication dates of this paper, indicating a growing interest and ongoing investigation into hallucinations in LLMs.

In relation to these works, my thesis project takes a more user-centric approach, focusing on evaluating the impact of these hallucinations on user trust and satisfaction. My work aims to understand how the technical phenomena underlined by Guerreio et al.[5] and Ji et al.[3] translate into user perceptions and satisfaction levels. This angle is less explored and highlights a gap in the current body of research, which is predominantly technical. My work will build on the findings of these studies, using their technical insights as a foundation to explore the user experience aspect.

1.3 Problem formulation

The primary knowledge gap in the current research on LLMs lies in understanding the user-centric impacts of hallucinations on trust and satisfaction. While existing studies have delved into the technicalities of hallucinations in LLMs, focusing on their causes, characteristics and mitigation strategies, there is a significant lack of research on how these technical phenomena translate into user experiences. This gap is crucial to address because user trust and satisfaction are key to the successful adoption and practical application of LLMs. In the broader context, ensuring that LLMs align with human expectations and ethical standards is essential to avoid misinformation and maintain user trust, which has both societal and economic implications. Bridging this gap will provide insights into potential risks and ethical concerns associated with LLMs, guiding the development of models that are both powerful and safe for public use. Understanding the impact of hallucinations from a user's perspective is vital for designing more

effective, intuitive and user-friendly systems, and this understanding is currently lacking in the field of Computer Science [3], [4].

To address this knowledge gap, the proposed study will involve a set of research actions aimed at understanding the impact of hallucinations on user trust and satisfaction in LLMs. The contribution of this research will be articulated through the following research questions:

- How do hallucinations in LLM outputs affect user trust?
- What aspects of hallucinations most significantly influence user satisfaction?
- Are there any differences in how users with varying degrees of familiarity with LLMs perceive hallucinations?

The study aims to conduct an empirical investigation into user perceptions of hallucinations in LLM outputs, focusing on user trust and satisfaction This will involve presenting users with various LLM-generated texts, some containing hallucinations, to gauge their reactions and trust levels. The research will employ metrics to quantitatively measure user satisfaction and identify key factors influencing user trust. The goal is to provide empirical data on user trust, guidelines for LLM design and a foundation for future research in human-AI interaction, particularly in understanding and improving user experience with AI systems.

1.4 Motivation

LLMs like gpt-4 are pivotal in natural language processing, with applications in machine translation and automated dialogue systems [3]. However, these [3] models often generate hallucinations: outputs that are inaccurate or irrelevant [3], [5]. This phenomenon affects not only the technical performance but also user trust and satisfaction, which are critical for the practical deployment of these technologies.

Research has delved into the technical aspects of hallucinations in LLMs, focusing on their causes, characteristics, and mitigation [3], [6]. Yet, there is a notable gap in understanding the influence of these technical issues on user perceptions, particularly trust and satisfaction.

From a societal perspective, aligning LLMs with human expectations and ethical standards is essential to avoid misinformation and maintain user

trust. Economically, user trust and satisfaction significantly impact the adoption and success of LLM based applications. This study seeks to explore the user experience aspect of LLM hallucinations.

1.5 Results

A comprehensive framework was developed to evaluate how hallucinations in LLM outputs affect user trust and satisfaction. The framework integrates multiple components to systematically assess these impacts:

1. Quantitative Metrics for Measuring Trust:

- Trust of Automated Systems Test (TOAST) [7] : This test evaluates user trust in automated systems by assessing various dimensions of trustworthiness, such as reliability, competence, and integrity.
- TrustDiff [7], [8] : This tool measures trust by capturing user perceptions and attitudes towards the system, focusing on differences in trust levels before and after interaction with the LLM.

2. Quantitative Metrics for Assessing User Satisfaction:

- System Usability Scale (SUS) [9] : SUS is a widely used questionnaire that evaluates the usability of a system, providing insights into user satisfaction based on factors like ease of use, efficiency, and overall satisfaction.

3. Evaluation Components:

- Data Collection: Gathering user feedback through surveys and questionnaires to assess trust and satisfaction levels.
- Analysis of Hallucinations: Identifying and categorizing hallucinations in LLM outputs, including their frequency, type, and severity.
- Impact Assessment: Analyzing the correlation between the presence of hallucinations and changes in trust and satisfaction metrics.

The framework integrates quantitative metrics such as the Trust of Automated Systems Test (TOAST) [7] and TrustDiff [7], [8] for measuring trust, and the System Usability Scale (SUS) [7], [8], [9]for assessing user satisfaction. This approach directly addresses the research questions about

how hallucinations impact user trust and which aspects of these hallucinations significantly influence user satisfaction.

The effectiveness of the framework was validated using quantitative analysis. A survey was conducted having participants (n=64) who interact with LLM-generated texts, some containing hallucinations. The TOAST, TrustDiff and SUS scores provided quantitative data reflecting levels of trust and satisfaction, respectively. This data allowed statistical analysis of the impact of hallucinations on user trust and satisfaction, providing a robust measure of the framework's precision and reliability.

1.6 Scope/Limitation

This thesis focuses on evaluating the impact of hallucinations on user trust and satisfaction in language model outputs, specifically using ChatGPT as a representative system. The study aims to explore the nuances of how artificially induced errors (hallucinations) affect users' perceptions of trustworthiness and usability of language models. The implementation involves a controlled experiment using predefined hallucination scenarios and a structured survey that includes TrustDiff, TOAST, and the System Usability Scale (SUS) to measure the impacts quantitatively.

The research is limited to ChatGPT, a popular model developed by OpenAI [2]. While ChatGPT is representative of advanced language models, it does not encompass all types of language or generative AI systems available in the market. Different models may have different architectures, training datasets, or capabilities that could influence their susceptibility to hallucinations and the subsequent impact on user trust and satisfaction.

The hallucinations introduced in this study are specifically designed and may not cover all possible types of inaccuracies that can occur in real-world applications. The scenarios include fictional content, historical inaccuracies, and event misrepresentations, which are chosen to deliberately provoke errors in the AI's responses. This choice limits the study's ability to generalize how other types of errors or naturalistic hallucinations might affect user perceptions.

The participant sample is primarily sourced from online platforms, which may attract a demographic with more technology exposure and potentially different trust dynamics than the general population. This recruitment strategy may limit the generalizability of the findings to all users of AI systems, particularly those with different levels of tech-savviness or from different cultural backgrounds.

While the study employs both quantitative and qualitative methods, the primary focus is on quantitative measures. This emphasis might overlook deeper, nuanced insights into individual user experiences and perceptions that qualitative approaches can provide. Moreover, the structured nature of the surveys and interviews may not capture spontaneous or less common reactions to hallucinations.

1.7 Target group

The research presented in this thesis is specifically tailored to several key groups who would benefit from understanding the impact of hallucinations on user trust and satisfaction in language model outputs, particularly those using ChatGPT. The target audience for this study includes:

1. AI Researchers and Academics

This group is at the forefront of exploring and understanding the capabilities and limitations of artificial intelligence. The findings from this thesis provide valuable insights into the nuances of user perception affected by AI errors, which can inform future research directions, including areas related to model training, error mitigation, and user interface design.

2. AI Developers and Engineers

Professionals involved in designing and developing AI-driven applications, especially those incorporating language models like ChatGPT, are a primary audience for this work. The results can help them understand how hallucinations might impact user satisfaction and trust, guiding them in improving AI system responses, implementing robust error-handling mechanisms, and optimizing user experience based on empirical data.

3. User Experience (UX) Designers

UX designers who work with AI technologies can utilize the findings to enhance interface designs that better manage user expectations and mitigate dissatisfaction in the event of AI failures. Understanding the specific aspects that influence user trust and satisfaction helps in crafting more intuitive and forgiving user interactions.

4. Technology Policy Makers and Ethicists

As AI technologies become more integrated into everyday activities, ensuring these systems are trustworthy and user-friendly is crucial. Policymakers and ethicists can use the insights from this study to help frame regulations and guidelines that protect users while encouraging innovation in AI development.

5. End-Users of AI-Integrated Systems

While not directly involved in the creation or study of AI, end-users are the ultimate beneficiaries of this research. By informing the aforementioned groups, improvements made to AI systems based on this study's findings will enhance the overall user experience, making technology more reliable and satisfactory for everyday use.

1.8 Outline

Finally, chapter 8, Conclusion and Future Work, summarizes the key findings and contributions of the thesis, highlighting the importance of addressing hallucinations in LLMs. It discusses the limitations of the current study and outlines avenues for future research and development.

2 Method

This chapter outlines the research methodology adopted to investigate the impact of hallucinations in Large Language Models (LLMs) on user trust and satisfaction. Utilizing a combination of three instruments (TrustDiff, Trust of automated systems test (TOAST) for trust assessment, and the System Usability Scale (SUS) for satisfaction evaluation) and participant comments, this study aims to provide a comprehensive analysis of user perceptions towards LLMs.

2.1 Research Project

This research employs a multimethod approach, anchored in design science, to construct and validate a framework for assessing the implications of hallucinations in LLMs on user trust and satisfaction. The project integrates empirical data collection through interviews, a comprehensive literature review, and the application of the TrustDiff, TOAST, and System Usability Scale (SUS) instruments as evaluative tools [9],[8],[7].

Initially, we define the problem based on identified gaps in current understanding of LLM hallucinations' effects on user perceptions. Objectives are then established to guide the development of a solution—a set of evaluation techniques underpinned by TrustDiff and TOAST for trust assessment, and SUS for satisfaction measurement.

Multimethod Research Activities include :

- Literature Review: To establish a theoretical base and understand existing methodologies.
- Survey: Conducted pre-study with end-users to gather demographic insights, complementing the analysis of the data from TrustDiff, TOAST, and SUS.
- Interview: post-study with selected participants, to gather additional comments and feedback to complement the data analysis.

Each methodological step is aimed at producing tangible outcomes, from enhanced understanding of user expectations to validated assessment tools. These results will be iteratively reviewed and refined, ensuring alignment with the project's overarching goals.

The selection of design science and the specific research methods is justified by their suitability for developing and rigorously testing new artifacts. This approach not only addresses the identified knowledge gap but also contributes to the broader field by providing a replicable framework for future research on LLM user experience.

2.2 Research methods

A systematic literature review has been performed to understand the theoretical background and existing frameworks related to LLM hallucinations, trust, and satisfaction. This method ensures a comprehensive understanding of the current state of research, identifying gaps and informing the development of our evaluation technique. The literature review supports the project by grounding it in established knowledge, ensuring that our approach is informed by and contributes to scholarly discourse.

TrustDiff is a semantic differential scale developed to assess user trust specifically in web contexts [6]. In this study, TrustDiff serves to gauge the nuanced perceptions users have towards LLMs when they interact with them through web-based interfaces. This tool allows us to capture the dichotomy of trust and distrust towards LLM outputs, offering insights into the immediate user reactions that could influence overall satisfaction and reliance on these systems.

TOAST (Trust of Automated Systems Test), originally validated to measure trust in automated systems, is employed to understand the broader aspects of trust in LLMs [7]. TOAST's multidimensional approach enables us to dissect various trust dimensions, such as reliability, faith, and perceived safety, providing a comprehensive picture of user trust dynamics in the context of LLM hallucinations. This detailed understanding is critical for identifying specific trust factors that are most affected by hallucinations, thereby informing strategies to mitigate their impact.

The System Usability Scale (SUS) [8], [9] is utilized to assess the overall satisfaction with LLM-based systems from a usability standpoint. Given its widespread acceptance and reliability, SUS offers a standardized method to evaluate how hallucinations in LLM outputs may alter the usability perception of these systems.
TrustDiff and TOAST will be used to quantitatively assess user trust in LLMs before and after the participants are presented with examples of LLM output. These methods were selected for their proven reliability and relevance in evaluating trust within technological contexts [10], [8]. By applying these tools, the project aims to derive statistical insights into the effect of LLM hallucinations on user trust levels,

The SUS will be employed to measure user satisfaction with LLMs. This standardized tool is chosen for its simplicity, effectiveness, and widespread acceptance in usability studies. It will facilitate the collection of comparative data on user satisfaction before and after exposure to LLM hallucinations, enabling a clear assessment of their impact [9].

The selection of literature review, TrustDiff, TOAST, and SUS is strategic, designed to offer a multimodal analysis of the research problem. Each method is chosen for its specific strengths in capturing different dimensions of the user experience with LLMs. This diversified approach ensures a balanced, well-rounded understanding of the issue, addressing both the depth of individual experiences and the breadth of the problem at scale.

These methods collectively offer a robust framework for investigating the effects of LLM hallucinations on user trust and satisfaction, leveraging both qualitative insights and quantitative data to inform the development of more reliable and user-centric LLM technologies.

2.3 Reliability and Validity

Reliability refers to the degree to which an assessment tool produces stable and consistent results over time. For TrustDiff, TOAST, and SUS, which are standardized and validated instruments, reliability is evaluated through the calculation of Cronbach's alpha, a measure of internal consistency. High values of Cronbach's alpha ($\alpha \geq 0.7$) [11] suggest that the items within each scale are highly correlated, providing consistent responses across different administrations of the tool. This study aims for a reliability threshold of $\alpha \geq 0.8$ for TrustDiff, TOAST, and SUS to ensure the instruments' robustness in measuring trust and satisfaction without being significantly influenced by random error or the variability of hallucinations in LLM outputs [12].

Validity, on the other hand, pertains to the degree to which the tools actually measure what they are supposed to measure. This involves several forms:

- Content Validity: Ensured through the initial selection of TrustDiff, TOAST, and SUS, each scale was chosen based on its established relevance and comprehensive coverage of the constructs of trust, distrust, and usability.

- Construct Validity is evaluated by examining the relationship between the scores from TrustDiff, TOAST, and SUS and other measures known to be related. For instance, higher trust as measured by TOAST should correlate with higher user satisfaction scores on SUS, supporting the notion that trust influences satisfaction. Furthermore, exploratory and confirmatory factor analyses are conducted to verify that the scales measure distinct constructs of trust and satisfaction, providing a clear theoretical rationale for their use.

Ensuring the reliability and validity of the TrustDiff, TOAST, and SUS scores is crucial for this research project. It not only bolsters the integrity of the study's findings but also contributes to the broader academic discourse by providing a methodological framework that can be replicated and applied in future research on LLMs and other AI-based systems. By rigorously evaluating these measures, the study aims to offer reliable and valid insights into how hallucinations affect user trust and satisfaction, paving the way for the development of more user-centered and trustworthy LLM applications.

2.4 Ethical Considerations

Ethical considerations play a critical role in conducting research, especially when it involves human participants and emerging technologies like Large Language Models (LLMs). This study is committed to upholding ethical standards to ensure the welfare, dignity, and rights of all participants are respected throughout the research process.

Consent is obtained in a written format, confirming participants' voluntary agreement to participate after having understood the study details and no sensitive personal information will be collected.

To protect participants' privacy, all data collected during the study are anonymized, ensuring no personal identifiers are linked to the responses. The confidentiality of participants' information is maintained throughout the study, with findings reported in aggregate forms to prevent the identification of individual participants.

By adhering to these principles, the research not only contributes valuable insights into the impact of LLM hallucinations on user trust and satisfaction but does so in a manner that respects the rights and well-being of all participants, ensuring the integrity of the research process.

3 Theoretical Background

This chapter presents the theoretical underpinnings that inform the study of hallucinations in Large Language Models (LLMs) and their impact on user trust and satisfaction. It elucidates the conceptual frameworks and prior research findings that scaffold our understanding of trust dynamics, user satisfaction, and the phenomenon of hallucinations in the context of LLM-based systems.

3.1 Trust in Artificial Intelligence

Trust is a multifaceted construct pivotal to human-AI interactions. Mayer, Davis, and Schoorman (1995) define trust as the willingness of a party to be vulnerable based on the expectations of the intentions or behavior of another [13]. In the realm of LLMs, trust encompasses belief in the system's competence, reliability, and intentionality. Trust affects users' reliance on LLM outputs for decision-making, influencing the adoption and effective use of these technologies. The formation of trust in AI is influenced by several factors including reliability, predictability, transparency, fairness, and the ethical design of the systems. For example, autonomous vehicles have to demonstrate consistent and safe driving in varied real-world scenarios to gain users' trust [14]. Similarly, AI in healthcare, such as IBM's Watson, must accurately diagnose and recommend treatments consistently to be trusted by medical professionals and patients alike [15].

The ability of users to understand how AI systems make decisions is crucial in establishing trust. Explainable AI initiatives [11], [16] aim to make the decision-making processes of AI systems transparent and understandable to users. For instance, the European Union's General Data Protection Regulation (GDPR) has provisions for the right to explanation, implying that users have a right to understand how decisions that affect them are made by automated systems [2].

Trust also hinges on the perceived fairness and ethical considerations integrated into AI systems. AI applications in recruitment, like those developed by HireVue [17], are scrutinized for bias in their algorithms, which can affect their trustworthiness. Ensuring these systems are free from biases that discriminate against certain groups is essential for maintaining public trust [18].

Instances where trust is compromised can significantly impact the perception and acceptance of AI systems. For example, the controversy surrounding Microsoft's chatbot, Tay, which started producing offensive content due to manipulation by users, shows how quickly trust can be eroded due to unforeseen vulnerabilities [19].

This study draws upon the TrustDiff and TOAST scales, which measure the dimensions of trust in automated systems, to explore how hallucinations influence user trust specifically within the context of LLMs.

3.2 User Satisfaction with Technology

User satisfaction, within the framework of technology use, is determined by the extent to which users believe the technology meets their expectations and needs. User satisfaction with technology is critical for its adoption and efficiency, particularly in AI applications. This satisfaction stems from several key areas: usability, functionality, performance, and emotional response. Each factor plays a crucial role in how end-users perceive and interact with AI systems.

Usability is a foundational element of user satisfaction. A study on the usability of AI interfaces, such as those in smart homes, highlights the importance of intuitive design in fostering user satisfaction [20]. The System Usability Scale (SUS) is often employed to assess this aspect, providing a reliable measure of the usability of various systems, including AI-driven technologies.

Functionality of AI systems significantly affects user satisfaction by determining how well the system meets user needs. For example, voice-activated assistants like Amazon Alexa must perform a wide range of tasks effectively to satisfy users [20]. Their ability to understand and execute user commands accurately is directly linked to user satisfaction.

Performance refers to the efficiency and accuracy with which technology meets its intended purpose. In AI, rapid response times and precise actions are pivotal. Google's AI enhancements to its search algorithms are a prime example [21], where improved search result accuracy and speed have directly increased user satisfaction.

Emotional response to AI technology also significantly impacts user satisfaction. The deployment of AI in sensitive applications, like service robots, shows that the emotional connection users feel towards technology can affect their overall satisfaction [22]. These robots, which help with

monitoring and interaction, have been well-received when they exhibit traits that engender trust and empathy.

Real-world impacts on user satisfaction can also stem from negative experiences. For instance, when issues arise with AI performance, such as with Tesla's Autopilot system, user satisfaction can decline sharply, especially if the technology fails to meet safety expectations [23].

3.3 Hallucinations in Large Language Models

Hallucinations in LLMs refer to instances where the models generate outputs that are unrelated, incorrect, or misleading based on the input provided. These phenomena raise significant concerns regarding the reliability of LLMs, potentially eroding user trust and satisfaction. Research on hallucinations in LLMs is emerging, with initial studies indicating that such errors can lead to confusion, mistrust, and reduced user engagement [3], [24]. This study aims to build upon this nascent literature, offering an investigation into the prevalence of hallucinations in LLMs and their psychological impact on users.

Hallucinations in LLMs can manifest in various forms, ranging from minor inaccuracies to significant fabrications. For instance, in a question-answering task, an LLM might generate an answer that seems plausible but is factually incorrect. Similarly, in content generation tasks, the model may produce narratives or explanations that, while coherent, are based on non-existent facts or events.

The root causes of hallucinations in LLMs are multifaceted, involving both technical and data-related issues [3], [6]:

- Data Quality and Bias: LLMs are trained on vast datasets compiled from the internet and other sources, which may contain inaccuracies, biases, or speculative content. The model might learn to replicate or amplify these inaccuracies during its output generation.
- Overgeneralization: LLMs might overgeneralize from the training data, producing outputs that are loosely related to the input or context but diverge significantly from factual accuracy.
- Lack of World Knowledge: Despite their extensive training, LLMs do not possess true world knowledge or understanding. Their "knowledge" is derived from patterns in the data they've been trained on, leading to potential gaps or inaccuracies in generated outputs.

- Model Complexity and Overfitting: The complexity of LLMs and their tendency to overfit on certain data patterns can also contribute to hallucinations. Overfitting may cause the model to "memorize" and regurgitate specific inaccuracies or to generate outputs based on spurious correlations in the training data.

Efforts to mitigate the impact of hallucinations in LLMs involve improving data quality, refining training processes, and developing mechanisms for fact-checking and error correction [3]. Incorporating feedback loops where user corrections inform model retraining, and enhancing model interpretability to understand the basis of generated outputs, are also vital strategies. Additionally, developing standards for evaluating and reporting the accuracy of LLM outputs can help users better understand and navigate the limitations of these models.

3.4 Theoretical Framework for Examining Trust and Satisfaction

The theoretical framework guiding this study integrates constructs of trust and user satisfaction within the context of technology use, particularly focusing on LLM-based systems. It hypothesizes that hallucinations negatively impact user trust and satisfaction [3], potentially altering the willingness to use and rely on LLMs for information processing and decision-making tasks. By employing validated scales for measuring trust and satisfaction, the study seeks to assess these impacts, providing insights into the mechanisms through which hallucinations influence user perceptions and behaviors.

3.5 Summary

Understanding the dynamics of trust and satisfaction in the context of LLMs is critical for advancing the development and deployment of reliable AI systems. By grounding this study in established theories of trust, user satisfaction, and the specific phenomenon of hallucinations in LLMs, the aim is to contribute to the broader discourse on enhancing AI reliability and user-centered design. This theoretical background sets the stage for the empirical investigation detailed in the subsequent chapters, seeking to elucidate the nuanced impacts of hallucinations on user trust and satisfaction in LLM-based systems.

4 Research project – Implementation

This chapter details the implementation of the research project, which aimed to evaluate the impact of hallucinations on user trust and satisfaction in LLM-based systems, particularly focusing on ChatGPT outputs.
The primary tool for data collection was a Google Form (Appendix 1), designed to measure TrustDiff, TOAST, and the System Usability Scale (SUS) scores. This method was selected for its accessibility, ease of distribution, and its ability to efficiently gather large quantities of data.
The form was structured to first present participants with baseline examples of ChatGPT outputs that exemplify desirable and accurate responses, including a thank you note and a movie summary. These examples served to establish a benchmark of user expectations and satisfaction from the system under normal operating conditions.
To assess the impact of hallucinations, participants were subsequently exposed to three prompts designed to induce hallucinations in ChatGPT's responses (see appendix 1):

- Fictional TV Show Episode: Participants were asked about a specific episode of a non-existent TV show.
- Non-existent Sources: This prompt involved asking for sources on the purported sole survivor of the Titanic.
- Historically Inaccurate Event Description: Participants requested a positive review of the Astroworld Festival purportedly held in 2001, a festival at which a crowd surge caused many deaths.

The design of these hallucination prompts was based on several key factors and existing benchmarks:

1. Empirical Studies on Hallucinations: Research such as the "Survey of Hallucination in Natural Language Generation" by Ji et al. (2022) [3] discusses common scenarios where hallucinations are likely to occur, providing a foundation for creating prompts that challenge the model's knowledge boundaries.

2. Known Limitations of LLMs: Studies on LLMs highlight their tendency to produce confident but incorrect answers, especially when dealing with obscure, fictional, or non-existent entities [25]. Designing prompts that exploit these tendencies ensures a high probability of generating hallucinations.

Following these examples, participants were encouraged to engage directly with ChatGPT, allowing them to experience hallucinations firsthand. This

interactive element was designed to deepen participants' understanding of the nature of hallucinations and their potential impact on user trust and satisfaction.

After the exposure to both the pre-set hallucinatory prompts and personal interaction with ChatGPT, TrustDiff and TOAST scores were measured again to assess any changes in trust. The SUS was also administered to evaluate if the hallucinatory experiences had altered the perceived usability and overall satisfaction with the system.

The use of TrustDiff and TOAST before and after exposing participants to hallucinations provides a quantitative measure of the change in trust. Similarly, the SUS scores offer quantitative insights into changes in user satisfaction. These instruments are well-established in the research community, providing robust, reliable, and valid measures for trust and usability respectively.

In addition to structured questionnaire responses, qualitative feedback was collected to capture nuanced perceptions and personal reactions to the hallucinations. This qualitative data complements the quantitative measures, providing a richer, more comprehensive understanding of the user experience.

5 Results

5.1 Participants.

The participant recruitment for the experiment was strategically designed to gather a diverse sample that could provide insights into the impact of hallucinations on user trust and satisfaction in large language models (LLMs). Recruitment efforts targeted adult users from a wide range of demographics to include a variety of perspectives and experiences with technology. Participants were recruited through social media and online platforms that cater to individuals interested in participating in psychological and technological research studies, ensuring a high level of engagement and interest in the study's focus.

To facilitate a controlled yet realistic interaction with the AI systems, participants were provided with a brief training session to familiarize themselves with the interface and the nature of the tasks they would be performing. This preparatory step was crucial to ensure that the data collected would reflect their true perceptions of the AI's performance rather than their unfamiliarity with the technology.

In total, the study successfully recruited 64 participants. This sample size was determined to be sufficient to achieve statistical power for detecting significant differences in trust and satisfaction levels before and after exposure to AI hallucinations, based on preliminary analysis.

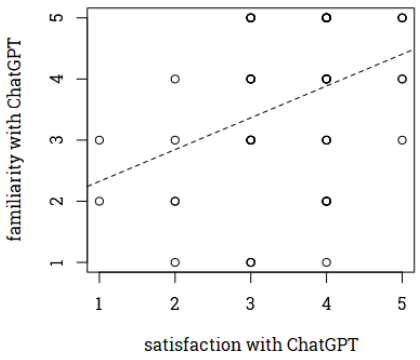


Figure 5.1 User familiarity vs. overall satisfaction.

Figure 5.1 shows user satisfactions increasing along with familiarity. Linear regression is shown as dashed line.

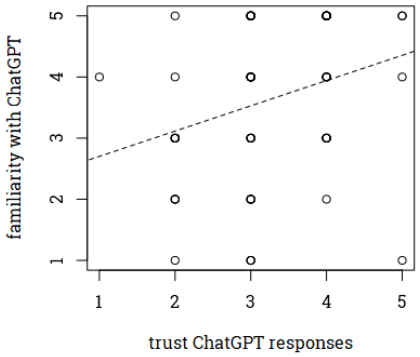


Figure 5.2 User familiarity vs. overall trust

Figure 5.2 shows overall user trust increasing with familiarity. Linear regression is shown as dashed line.

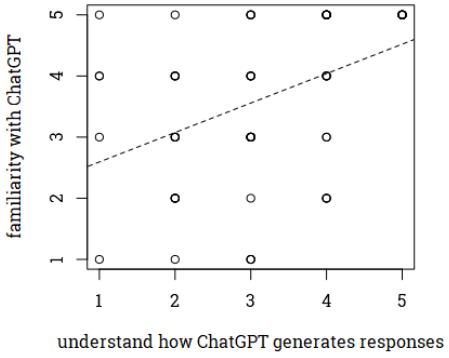


Figure 5.3 User familiarity vs. understanding

Figure 5.3 shows self reported system understanding increasing with familiarity. Linear regression is shown as dashed line.

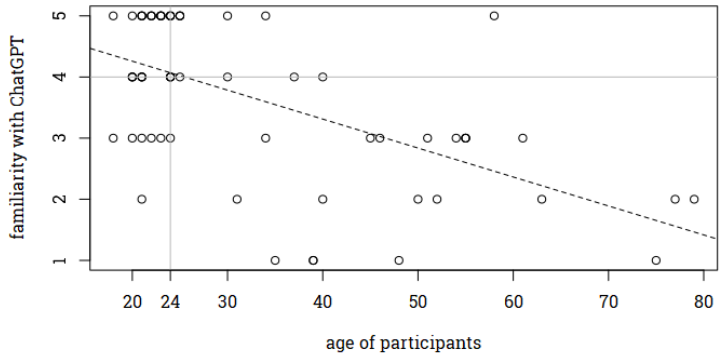


Figure 5.4 User Familiarity vs. age

Figure 5.4 shows familiarity with ChatGPT decreasing as users get older. The median for participants' age was 23, and for familiarity with ChatGPT it was 4. Linear regression is shown as dashed line.

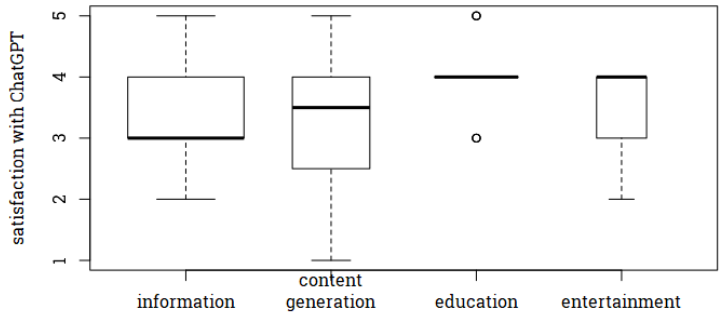


Figure 5.5 User satisfactions vs. purpose of use

Figue 5.5 shows user satisfaction being higher for content generation compared to information search and even higher for educational end entertainment purposes.

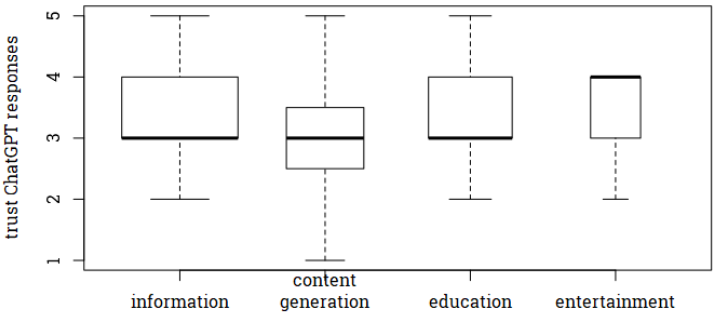


Figure 5.6 User trust vs. purpose of use

Figure 5.6 shows user trust being slightly lower for content generation compared to information search and education, yet higher than the previously mentioned for entertainment.

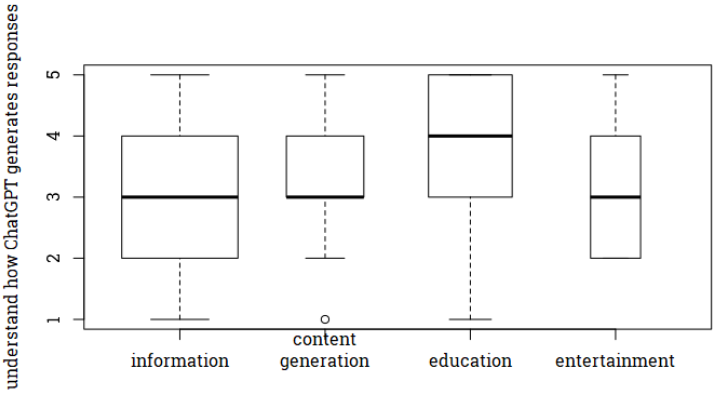


Figure 5.7 User understanding vs. purpose of use

Figure 5.7 shows user understanding being similar for information, content generation and entertainment and higher for educational purposes.

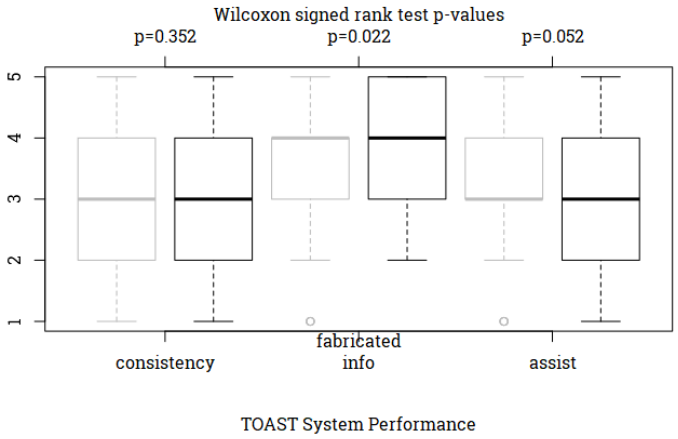


Figure 5.8 TOAST: System performance with p values from wilcoxon tests.

Figure 5.8 shows system performance TOAST scores before and after users are exposed to ChatGPT hallucinations. The gray markers represent the results before, whereas the black represents results after exposure to hallucinations. Consistency and assist scores remained the same, while fabricated information scores were different ($p<.05$).

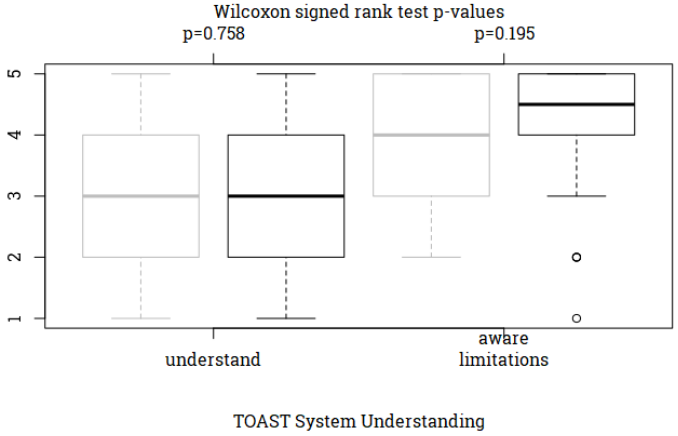


Figure 5.9 TOAST: System understanding with p values from wilcoxon tests.

Figure 5.9 shows TOAST system understanding scores before and after users are exposed to ChatGPT hallucinations. The gray markers represent the results before, whereas the black represents results after exposure to hallucinations. Whilst understanding scores did not statistically differ, there is a visual indication from the box plots that users did become more aware of the limitations of the system.

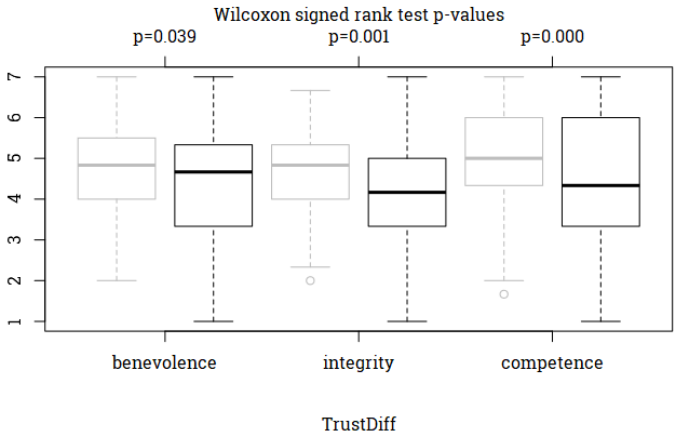


Figure 5.10 TrustDiff with p values from wilcoxon tests.

Figure 5.10 shows TrustDiff scores before and after users are exposed to ChatGPT hallucinations. The gray markers represent the results before, whereas the black represents results after exposure to hallucinations. Scores were statistically different ($p<.05$) for all three factors, decreasing after presenting the hallucinations.

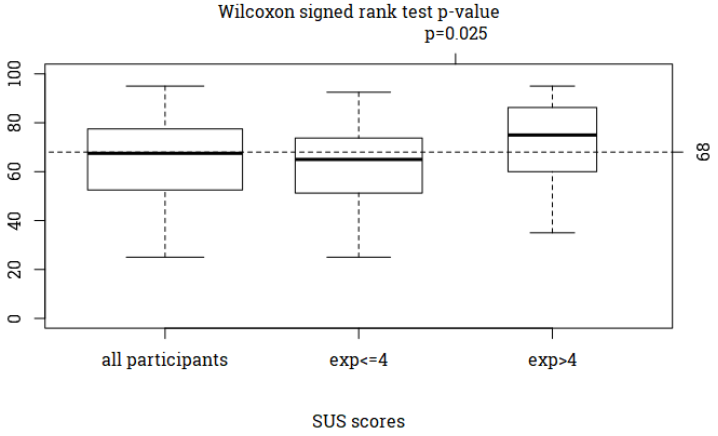


Figure 5.11 SUS scores

Figure 5.11 shows SUS scores after users were exposed to hallucinations, for users who reported an experience lower or greater than 4. The median is 68 which is considered a satisfying SUS score[26].

6 Analysis

Chapter 5 presented the results from the application of TrustDiff, the Trust of Automated Systems Test (TOAST), and the System Usability Scale (SUS). These results highlighted varied responses to LLM outputs, with significant discrepancies noted in user trust and satisfaction, especially after being presented with instances of hallucinations. Chapter 6 aims to analyze the data collected and described in Chapter 5 into arguments to support or refute the hypotheses posited earlier. This chapter utilizes various statistical methods to analyze data derived from a controlled experiment involving hypothesis testing. The analysis includes both descriptive (plots) and inferential (tests) statistics to validate the findings and draw robust conclusions.

In the controlled experiment detailed in Chapter 5, participants interacted with AI systems, where their levels of trust were measured both before and after exposure to potential hallucinations by the systems. The hypotheses tested revolved around the impact of these hallucinations on user trust and satisfaction.

Statistical Analysis:

- Descriptive Statistics: Initial analysis involved computing mean, median, and standard deviation for the trust and satisfaction scores across different test groups.
- Box Plots: These were used to visually assess the distribution of trust scores and identify outliers or anomalies in the data.
- Hypothesis Testing: Using t-tests to compare means across groups, hypotheses regarding differences in trust and satisfaction were tested.

Interpretation of Statistical Outputs:

- Significant p-values ($p < 0.05$) indicated that hallucinations negatively impacted user trust and satisfaction.
- The effect size was calculated to determine the magnitude of this impact, providing a clearer picture of its practical significance.

The statistical analysis provided strong evidence supporting the hypothesis that hallucinations in AI systems reduce user trust and satisfaction. However, it is crucial to acknowledge the limitations of the study, such as the controlled environment and sample selection, which may affect the generalizability of the findings.

- Evidence vs. Arguments: The results are presented as evidence rather than mere arguments, substantiated by statistical tests to ensure robustness.

- Ethical Considerations: Throughout the analysis, care was taken to avoid bias, particularly confirmation bias, and to present the findings honestly and transparently.
- Implications: The findings underscore the need for developing more reliable AI systems that minimize errors and hallucinations to maintain user trust.

The findings from Chapter 5 demonstrate a significant decrease in user trust, as evidenced by the measurable decline in TrustDiff scores following exposure to hallucinatory content. This substantial reduction in trust underscores the critical need for enhanced accuracy and reliability in LLM outputs. The observed sensitivity of trust to inaccuracies highlights the fragile nature of user confidence in AI systems, which may have long-term implications for the adoption and reliance on such technologies. Therefore, improving the accuracy of LLM outputs is paramount to sustaining user trust and ensuring the effective integration of these systems into daily applications.

The variability observed in user satisfaction, dependent on the type of hallucination and initial user expectations, suggests that user education and transparent communication about the limitations and expected behaviors of AI systems could mitigate dissatisfaction. This variation indicates that while some users adjust their satisfaction levels based on their understanding of the AI's limitations, others experience a significant drop in satisfaction when expectations are not met. Enhancing user interfaces and experience design to better communicate system capabilities and limitations could therefore play a vital role in maintaining user satisfaction despite occasional inaccuracies.

The resilience observed in some users, who recognized hallucinations without a drastic decrease in satisfaction, supports the notion that informed users can maintain a level of contentment with AI systems. This resilience underscores the potential benefits of developing AI systems equipped with better explanatory capabilities and error-reporting mechanisms. By enabling users to understand and contextualize errors when they occur, AI developers can help sustain user satisfaction, even in the face of system imperfections.

The impact of hallucinations on trust and satisfaction is highly context-dependent, with significant discrepancies noted in scenarios where accuracy is paramount, such as factual reporting or educational content. This finding highlights the necessity for AI systems that are adaptive to the task context, prioritizing accuracy in high-stakes environments while allowing greater creative leeway in less critical applications. Such adaptive behavior is crucial to maintaining trust and satisfaction, as it aligns system performance with user expectations and the specific requirements of different tasks.

The prevalence of hallucinations and their detrimental impact on user trust and satisfaction accentuate existing deficiencies in current LLM training methodologies. The need for improved model training and design is evident, calling for advancements in data quality, training techniques, and error correction mechanisms. By focusing on these areas, the AI research and development community can create more robust LLMs that not only meet user expectations for reliability but also reduce the frequency and severity of hallucinations, thereby enhancing the overall trustworthiness and utility of AI systems.

User feedback gathered during the study presents a nuanced view of the implications of LLM hallucinations on user trust and satisfaction. These insights are particularly valuable in understanding the practical impacts of AI inaccuracies in real-world scenarios.

Participants reported a range of experiences (Appendix 2) , highlighting the contextual nature of user satisfaction and trust in AI systems. For example, one user noted, "When prompted to create a multiple-choice test based off of topics I provided to ChatGPT, the answer key was incorrect. This undermines trust, particularly in educational settings where accuracy is paramount." This comment underscores the critical need for reliability in AI-generated content when used in educational applications, where incorrect information can have significant consequences.

Several users expressed concerns about the consistency and reliability of ChatGPT's outputs. As one participant stated, "It can get repetitive and shows signs of bias," while another remarked on the system's tendency to "spout nonsense rather than admit lack of knowledge." These observations highlight a fundamental issue with current LLMs—their inability to recognize and admit their limitations, which can lead to misinformation.

The feedback also suggests that user awareness and the way users interact with AI can influence their level of trust. A participant mentioned, "For the use I have done of ChatGPT, I have been rather satisfied... I can fact-check it myself usually." This points to the importance of user literacy in AI capabilities and limitations as a buffer against the negative effects of hallucinations. Users who are aware of AI's propensity to fabricate answers are better equipped to mitigate its shortcomings through vigilant fact-checking.

Insights into user expectations for future AI development were also gleaned from the comments. Users acknowledge the potential for improvement, as reflected in the statement, "Understanding that it will only improve from here, I know that answers have to be checked." This acceptance of AI's

current limitations alongside an optimistic outlook for its evolution reflects a balanced perspective on technological progress and its integration into daily tasks.

The compilation of user experiences forms a compelling argument for the necessity of advancements in AI design that prioritize transparency, accuracy, and user education. By integrating mechanisms for error recognition and the capability to express uncertainty, AI developers can foster a more trustworthy environment for users. Moreover, enhancing user interfaces to better manage expectations can significantly improve user satisfaction, even when inaccuracies occur.

In conclusion, this chapter translates complex data into clear evidence that supports the thesis statement, respecting scientific and ethical standards throughout the analysis. The findings contribute to understanding the dynamics of trust in AI, offering a foundation for future studies and practical applications in AI development.

7 Discussion

This chapter discusses the findings from the controlled experiment that explored the impact of hallucinations in language model-based systems on user trust and satisfaction. The analysis confirmed that hallucinations negatively influence user trust and satisfaction, providing clear answers to the research questions posed. These results contribute to the feedback loop in the research process by directly addressing the defined problem through empirical evidence.

The findings from this study align with existing literature as outlined in Chapter 3, which documented concerns about the reliability of language model outputs and their impact on user trust. For instance, similar studies have shown that inaccuracies in AI outputs can significantly erode trust, especially in critical use cases like medical diagnosis or financial forecasting. This research adds to the body of knowledge by specifically quantifying how hallucinations impact trust and satisfaction, reinforcing the need for improved accuracy and transparency in AI systems.

The reduction in trust observed aligns with theoretical frameworks on trust in AI that emphasize reliability and predictability as foundational elements. The findings support theories suggesting that trust is dynamic and can be severely impacted by negative experiences with technology. This study operationalizes these theories by demonstrating concrete effects in a controlled setting, thereby providing empirical support for theoretical claims.

The study's limitations, including its controlled environment and the artificial nature of interactions in experimental settings, might affect the external validity and generalizability of the findings. While the controlled environment allows for the isolation of variables and precise measurements, it may not fully capture the complexities and unpredictability of real-world AI interactions. This limitation proposes caution when extrapolating the findings to broader contexts.

The findings are primarily applicable to scenarios similar to those tested in the experiment. The homogeneity of the participant pool and the specific nature of the tasks may limit the applicability of the results to different populations or more complex AI interactions. Future research should aim to include a more diverse participant base and a variety of AI systems to enhance the generalizability of the results.

This study contributes to the understanding of how hallucinations in AI systems influence user trust and satisfaction, supporting the hypothesis that

negative experiences with AI reliability degrade user trust. The findings emphasize the necessity for ongoing improvements in AI system design and validation, ensuring they meet the high standards required for widespread acceptance and reliance. The discussion highlights how these findings are situated within the broader context of AI research and practical application, guiding future inquiries and technological advancements.

8 Conclusions and Future Work

The results of this thesis project provide evidence regarding the impact of hallucinations in large language model (LLM)-based systems on user trust and satisfaction. The findings indicate that hallucinations significantly reduce both trust and satisfaction, thereby answering the primary research questions and addressing the problem set forth at the beginning of this project. This research has successfully bridged the knowledge gap identified in the literature, specifically regarding the quantitative impact of LLM hallucinations on user perception and behavior.

The relevance of these results extends across multiple domains:

- Science: The study contributes to the theoretical understanding of trust dynamics in human-AI interaction, providing empirical evidence that supports existing theories and models.
- Industry: For developers and AI companies, these findings highlight the critical need for robust AI systems that minimize errors to maintain user trust, which is essential for the adoption and success of AI technologies.
- Society: Public perception and acceptance of AI technology can be significantly influenced by trust. Ensuring AI systems are reliable and transparent is crucial for their ethical integration into everyday life.

The generalizability of these results is somewhat limited by the controlled experimental setting and the specific nature of the tasks used in the study. However, the fundamental insights regarding trust and user satisfaction are likely applicable in broader AI applications, suggesting that minimizing hallucinations remains relevant across different AI systems and contexts.

Looking forward, this project opens several avenues for future research:

- Longitudinal Studies: Investigating the long-term effects of repeated exposure to AI hallucinations on trust and user behavior would provide insights into the dynamics of trust recovery or further degradation over time.
- Mitigation Strategies: Developing and testing strategies to mitigate the impact of hallucinations could be highly beneficial. Research could focus on both technological solutions, like improved training methods for LLMs, and user-focused strategies, such as education and transparency about AI capabilities and limitations.

- Cross-Domain Generalizability: Extending the research to different domains and types of AI applications would help to determine the broader applicability of the findings and refine the understanding of trust in various AI contexts.
- Impact of Explainability: Investigating how the presence of explainable AI components affects user trust in the presence of hallucinations could yield important insights, aligning with the growing emphasis on explainable AI.

In conclusion, while this thesis provides valuable insights into the effects of hallucinations on user trust in AI, it also highlights the complexity of trust dynamics and the need for continued research in this area. Addressing these challenges and questions will not only advance scientific understanding but also enhance the practical deployment of trustworthy AI systems in society.

References

[1] "Website." doi: 10.48550/arXiv.2303.08774. Available: https://doi.org/10.48550/arXiv.2303.08774

[2] R. Tahiri Joutei Hassani, O. Sandali, and C. Baudouin, "[GPT-4, health data, and GDPR: The challenging equation]," *J. Fr. Ophtalmol.*, vol. 47, no. 3, p. 104071, Mar. 2024, doi: 10.1016/j.jfo.2024.104071. Available: http://dx.doi.org/10.1016/j.jfo.2024.104071

[3] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," Feb. 2022, doi: 10.1145/3571730. Available: http://dx.doi.org/10.1145/3571730. [Accessed: Jan. 24, 2024]

[4] A. Hagerty and I. Rubinov, "Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence," Jul. 2019.

[5] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective," *International Journal of Human--Computer Interaction*, Nov. 2022.

[6] N. M. Guerreiro *et al.*, "Hallucinations in Large Multilingual Translation Models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, Dec. 2023, doi: 10.1162/tacl_a_00615. Available: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00615/2197664/tacl_a_00615.pdf. [Accessed: Jan. 24, 2024]

[7] H. M. Wojton, D. Porter, S. T Lane, C. Bieber, and P. Madhavan, "Initial validation of the trust of automated systems test (TOAST)," *J. Soc. Psychol.*, vol. 160, no. 6, pp. 735–750, Nov. 2020, doi: 10.1080/00224545.2020.1749020. Available: http://dx.doi.org/10.1080/00224545.2020.1749020

[8] User Experience Magazine, "TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web," *JUX - The Journal of User Experience*, Nov. 25, 2020. Available: https://uxpajournal.org/trustdiff-semantic-differential-for-user-trust-web/. [Accessed: Feb. 28, 2024]

[9] Assistant Secretary for Public Affairs, "System Usability Scale (SUS)," Sep. 2013, Available: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html. [Accessed: Feb. 28, 2024]

[10] Y. Wang, F. Li, H. Zheng, L. Jiang, M. F. Mahani, and Z. Liao, "Human trust in robots: A survey on trust models and their controls/robotics applications," *IEEE Open J. Control Syst.*, vol. 3, pp. 58–86, 2024, doi: 10.1109/ojcsys.2023.3345090. Available: https://ieeexplore.ieee.org/document/10366819/

[11] A. A. Agbo, "Cronbach's Alpha: Review of Limitations and Associated Recommendations," *Journal of Psychology in Africa*, Jan. 2010, doi: 10.1080/14330237.2010.10820371. Available: https://www.tandfonline.com/doi/abs/10.1080/14330237.2010.10820371. [Accessed: Jun. 10, 2024]

[12] M. A. Bujang, E. D. Omar, and N. A. Baharum, "A Review on Sample Size

Determination for Cronbach's Alpha Test: A Simple Guide for Researchers," *Malays. J. Med. Sci.*, vol. 25, no. 6, pp. 85–99, Nov. 2018, doi: 10.21315/mjms2018.25.6.9. Available: http://dx.doi.org/10.21315/mjms2018.25.6.9

[13] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995, doi: 10.2307/258792. Available: http://www.jstor.org/stable/258792

[14] X. Zhao, V. Robu, D. Flynn, K. Salako, and L. Strigini, "Assessing the safety and reliability of autonomous vehicles from road testing," in *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, IEEE, Oct. 2019. doi: 10.1109/issre.2019.00012. Available: https://ieeexplore.ieee.org/document/8987509/

[15] S. Balan, *Business Intelligence in Healthcare with IBM Watson Analytics*. Createspace Independent Publishing Platform, 2017. Available: https://books.google.com/books/about/Business_Intelligence_in_Healthcare_with.html?hl=&id=bBrqtAEACAAJ

[16] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019. Available: https://books.google.com/books/about/Explainable_AI_Interpreting_Explaining_a.html?hl=&id=7NRcxwEACAAJ

[17] J. Tullier, "The use of artificial intelligence to recruit employees," 2021. Available: https://www.lagrange.edu/academics/undergraduate/undergraduate-research/citations/1b-Citations2021.JTullier---ACCT.pdf. [Accessed: Apr. 17, 2024]

[18] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–12, Sep. 2023, doi: 10.1057/s41599-023-02079-x. Available: https://www.nature.com/articles/s41599-023-02079-x. [Accessed: Apr. 17, 2024]

[19] Z. Vorsino, "Chatbots, Gender, and Race on Web 2.0 Platforms: Tay.AI as Monstrous Femininity and Abject Whiteness," *Signs: Journal of Women in Culture and Society*, vol. 47, no. 1, pp. 105–127, Sep. 2021, doi: 10.1086/715227. Available: https://doi.org/10.1086/715227

[20] D. S. Zwakman, D. Pal, and C. Arpnikanondt, "Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa," *SN Comput Sci*, vol. 2, no. 1, p. 28, Jan. 2021, doi: 10.1007/s42979-020-00424-4. Available: http://dx.doi.org/10.1007/s42979-020-00424-4

[21] A. Martín, *The Gemini Effect: Google's AI Masterpiece: Discovering the Secrets of the Most Advanced AI Model*. Independently Published, 2023. Available: https://books.google.com/books/about/The_Gemini_Effect.html?hl=&id=blpi0AEACAAJ

[22] M. Becker, D. Mahr, and G. Odekerken-Schröder, "Customer comfort during service robot interactions," *Service Business*, vol. 17, no. 1, pp. 137–165, Mar. 2023, doi: 10.1007/s11628-022-00499-4. Available: https://doi.org/10.1007/s11628-022-00499-4

[23] M. Dikmen and C. M. Burns, "Autonomous Driving in the Real World:

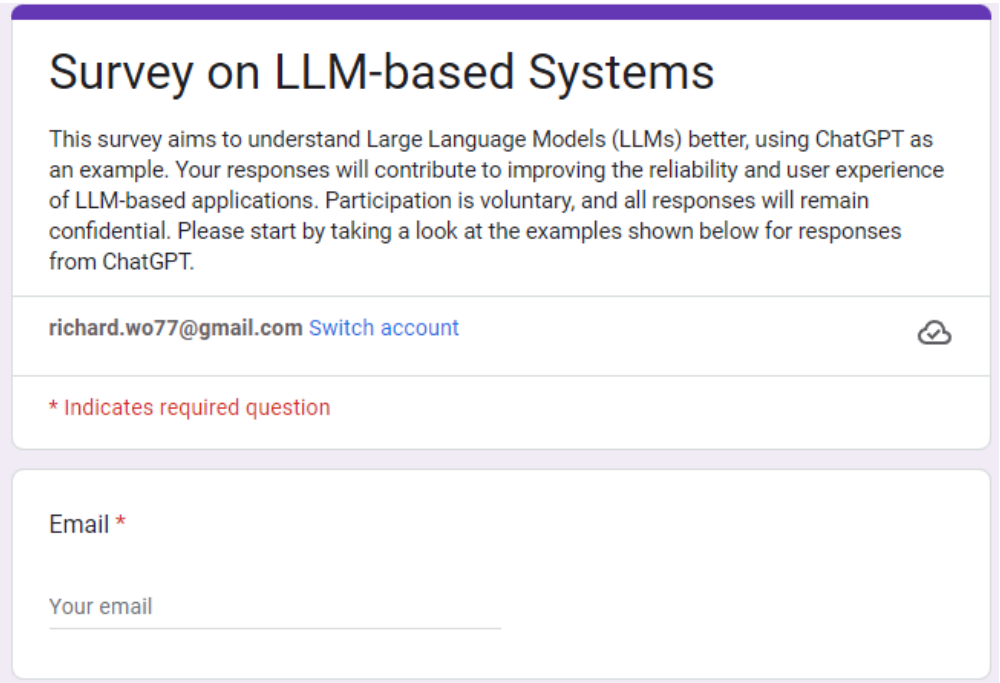
Experiences with Tesla Autopilot and Summon," in *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, in Automotive'UI 16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 225–228. doi: 10.1145/3003715.3005465. Available: https://doi.org/10.1145/3003715.3005465. [Accessed: Apr. 17, 2024]

[24] N. Gillespie, S. Lockey, C. Curtis, J. Pool, and A. Akbari, "Trust in Artificial Intelligence: A global study," Feb. 2023, doi: 10.14264/00d3c94. Available: https://www.aiunplugged.io/wp-content/uploads/2023/10/Trust-in-Artificial-Intelligence.pdf

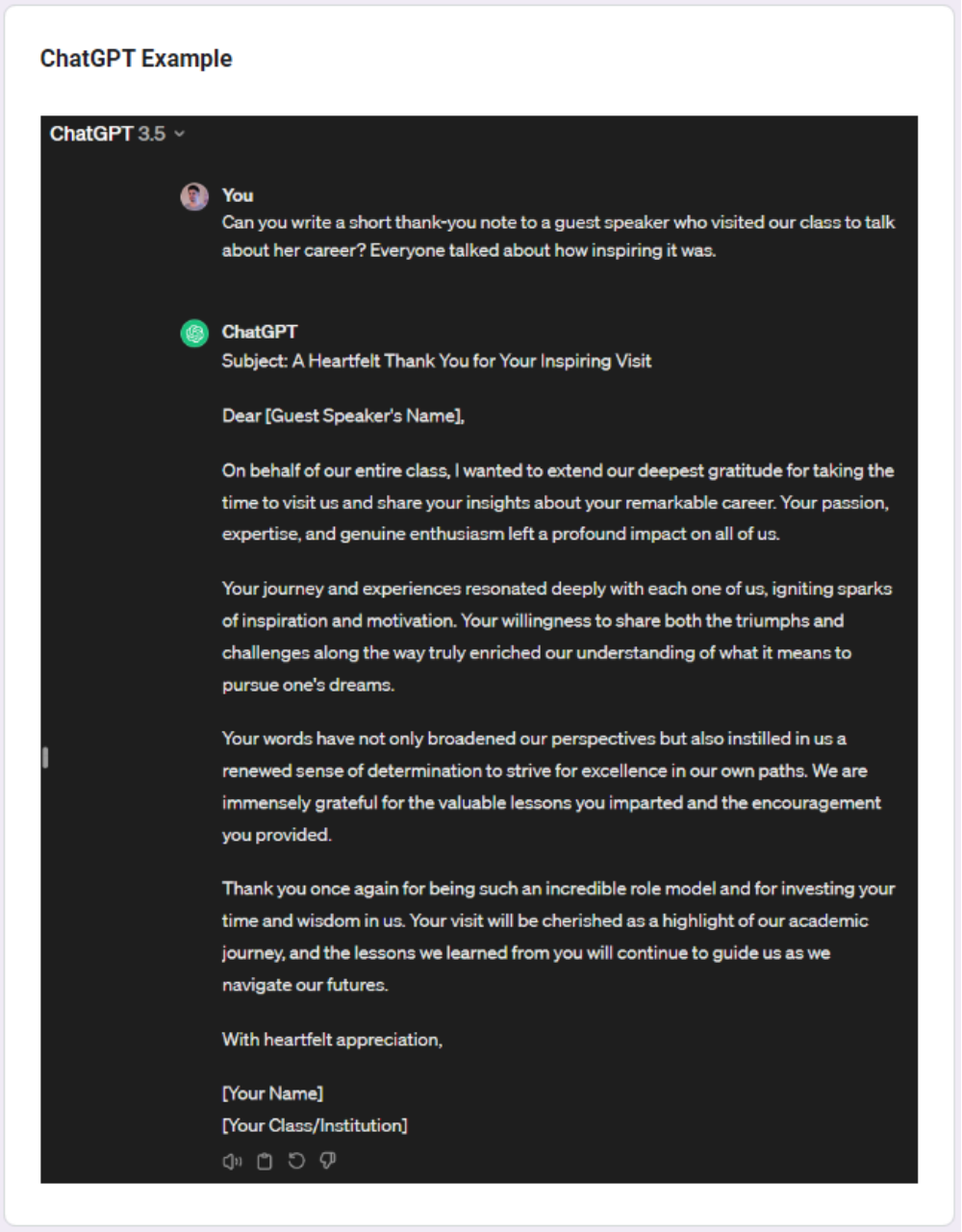
[25] A. El Amri, *LLM Prompt Engineering for Developers: The Art and Science of Unlocking LLMs' True Potential*. Packt Publishing Ltd, 2024. Available: https://play.google.com/store/books/details?id=6XQKEQAAQBAJ

[26] A. Bangor, P. Kortum, and J. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale," *J. Usability Stud.*, vol. 4, pp. 114–123, 04 2009.

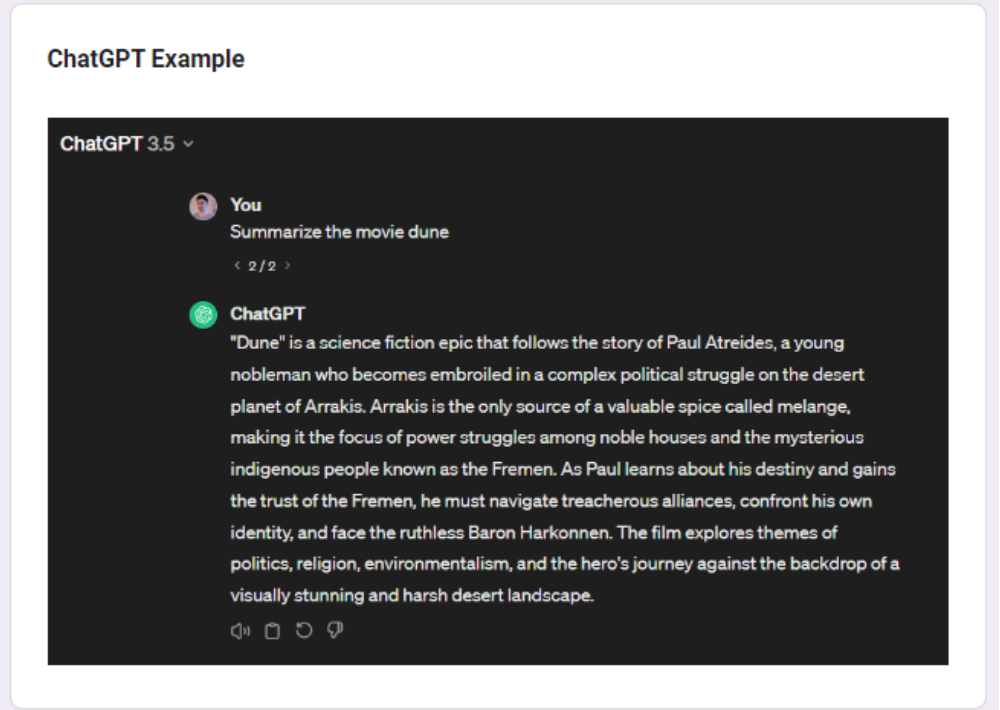
Appendix 1 - Google form



Introduction to form and email collection.



ChatGPT positive example 1.



ChatGPT positive example 2

Age *

Your answer

Familiarity with ChatGPT *

	1	2	3	4	5	
Not familiar at all	○	○	○	○	○	Very familiar

Purpose of Use: *

- ○ Content generation
- ○ Information search
- ○ Educational purposes
- ○ Entertainment
- ○ Other: _____

Overall, I trust ChatGPT to function as intended. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

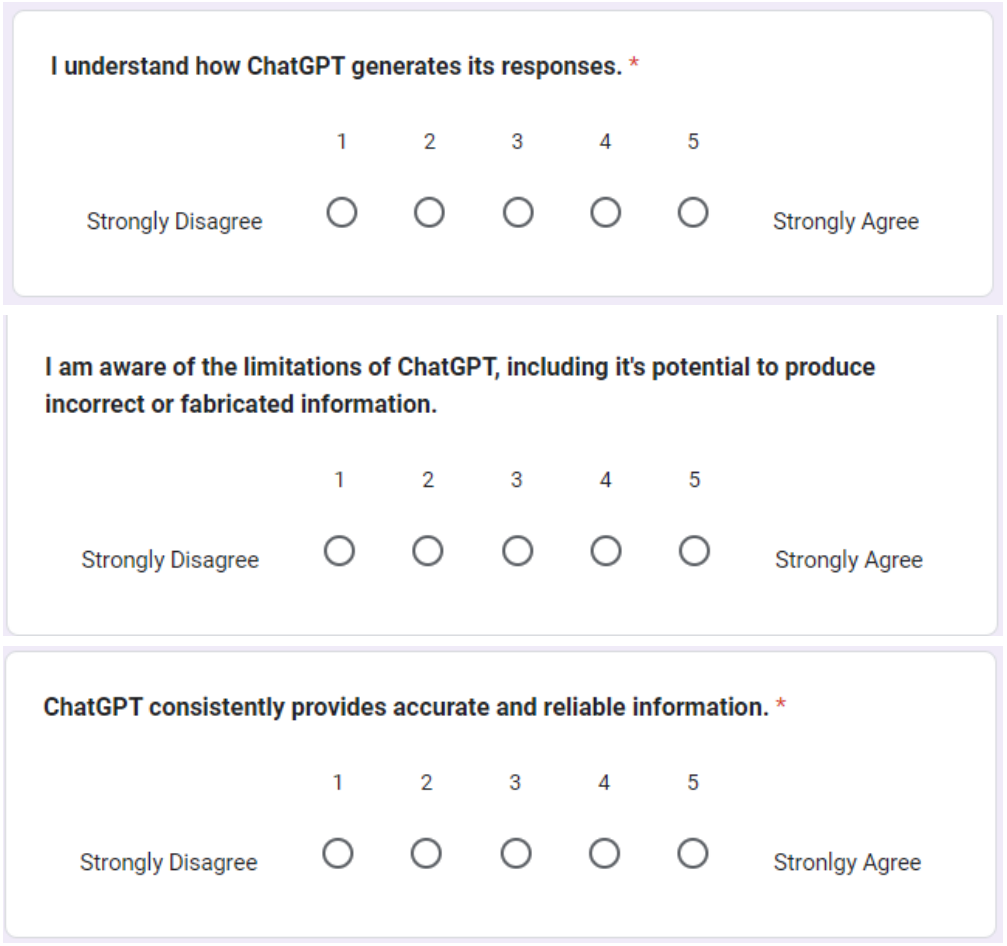
Overall, I am satisfied with ChatGPT *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

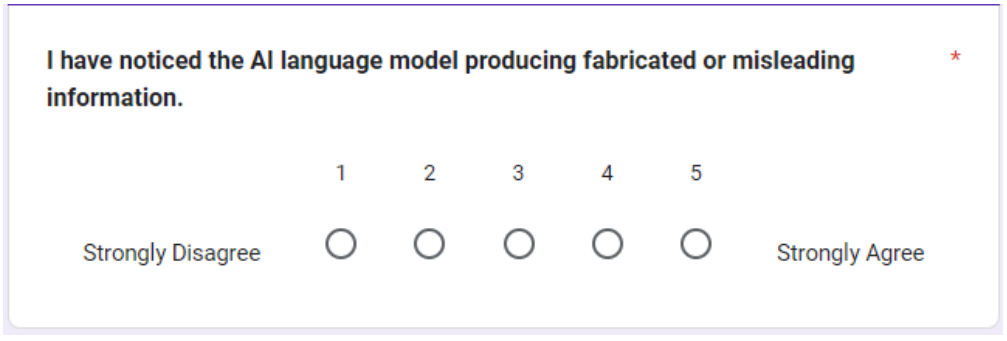
You can add comments on the examples above:

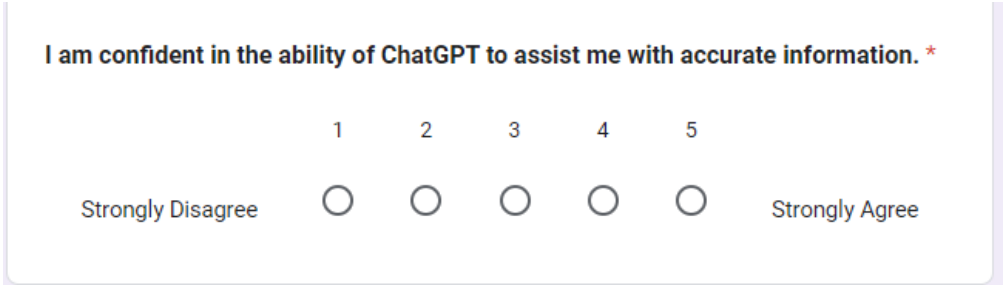
Your answer

Demographics and overall satisfaction

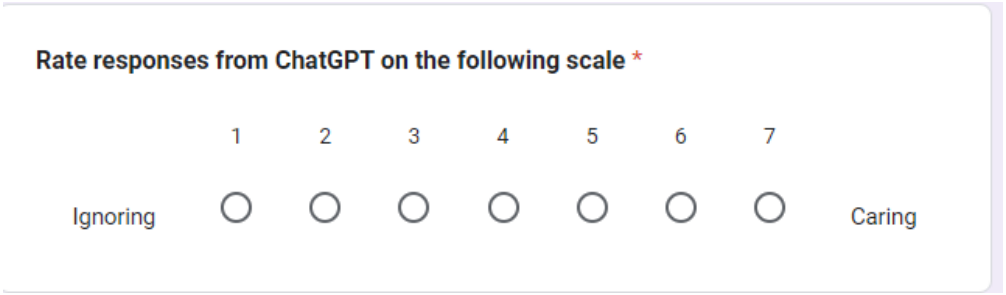


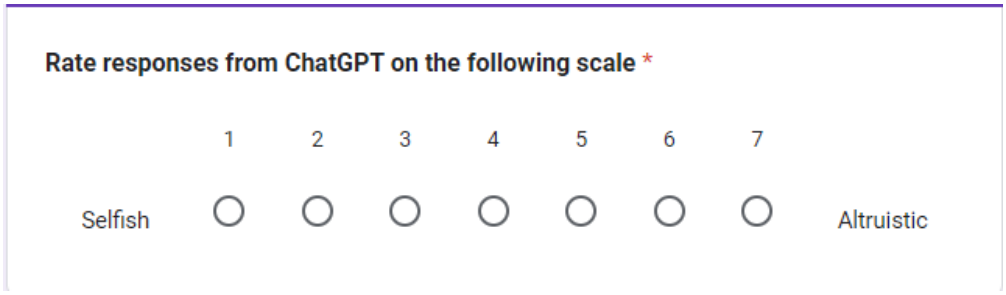
System Understanding questions for TOAST.

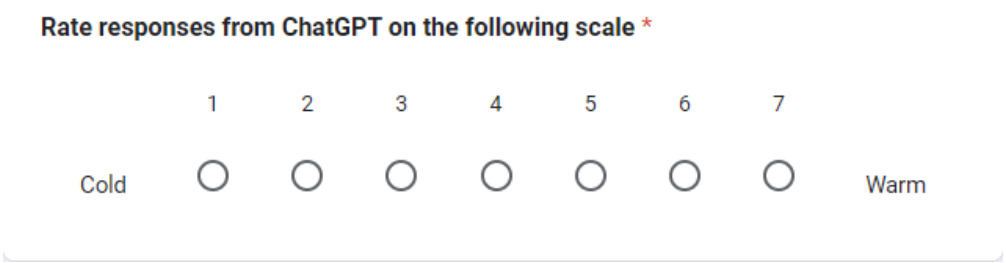




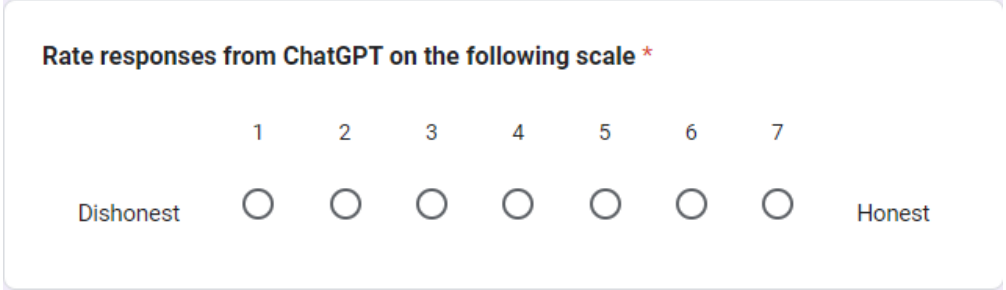
System Performance questions for TOAST.

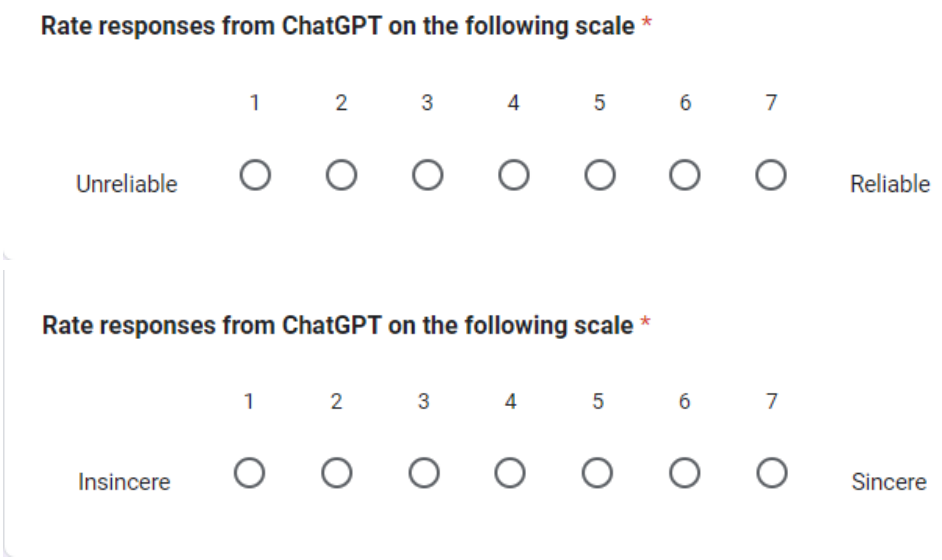




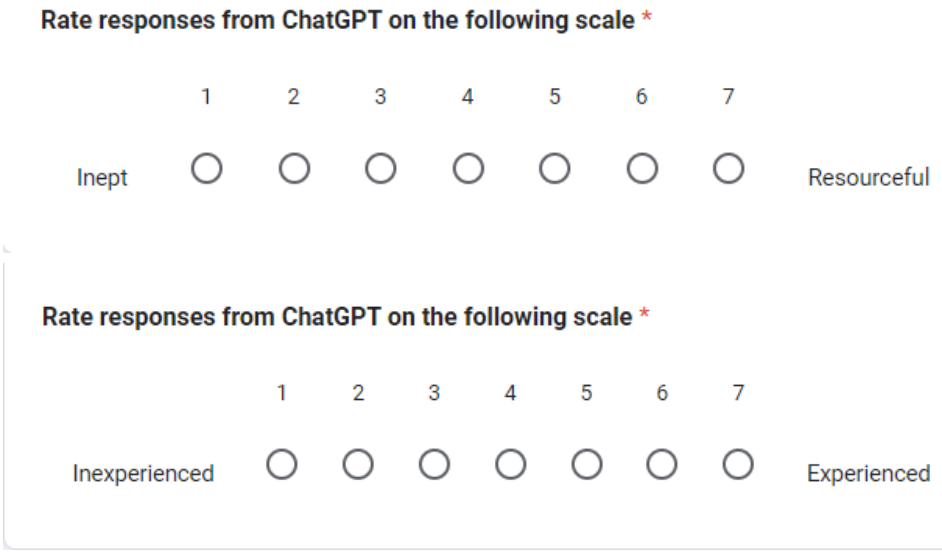


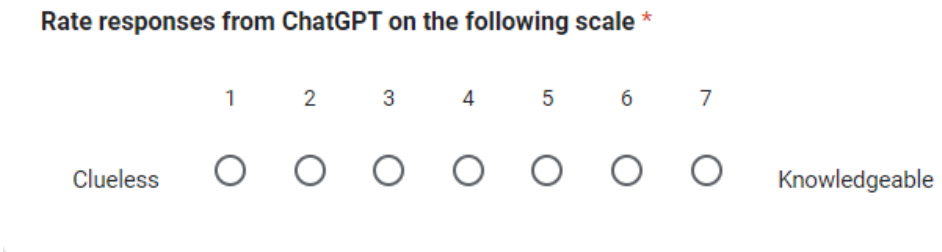
Benevolence questions for TrustDiff.





Integrity questions for TrustDiff



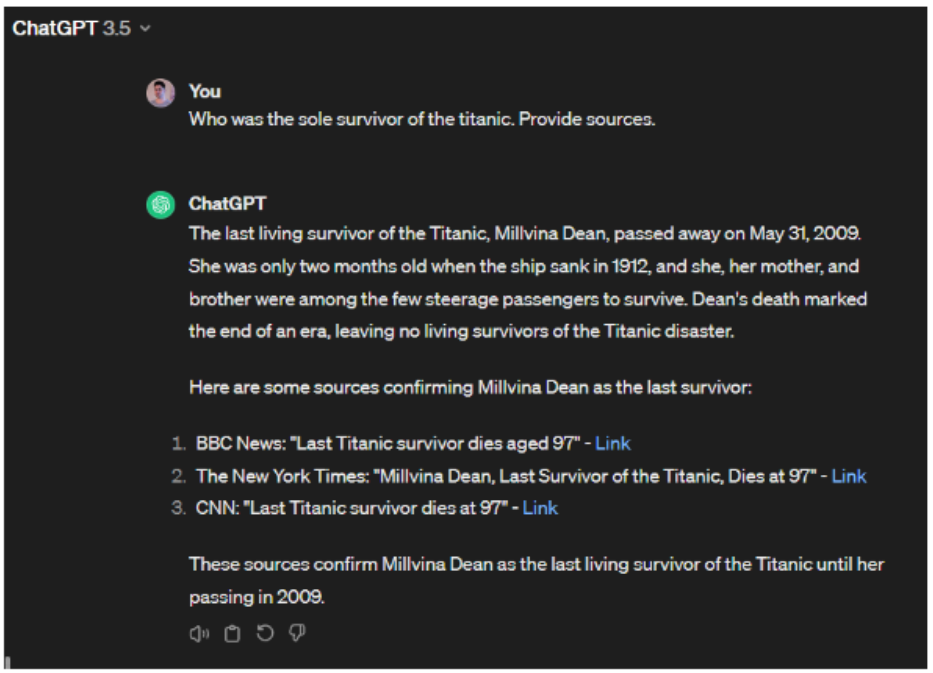


Competence questions for TrustDiff

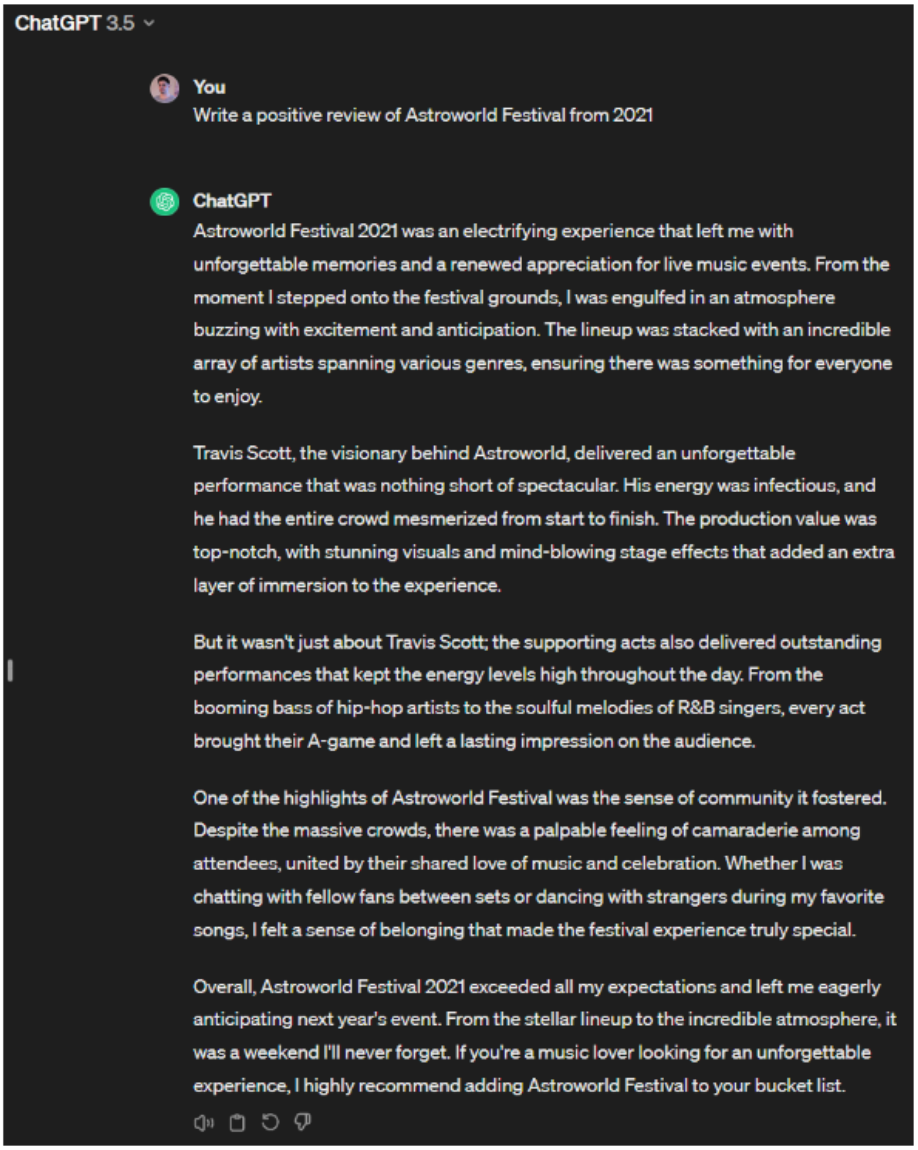
Please try to converse with ChatGPT using prompts similar to these before proceeding to the next section:

ChatGPT knows a lot about popular TV shows or other works of fiction, but doesn't know the details (like all the scenes or episodes). Completely made up response :

Similarly, when asked to provide sources on information, these are completely made up.



When asked to write a positive review of the festival at which a fatal crowd crush occurred, it also hallucinates these events.



Negative examples for ChatGPT that include hallucinations

I understand how ChatGPT generates its responses. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I am aware of the limitations of ChatGPT, including it's potential to produce incorrect or fabricated information.

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

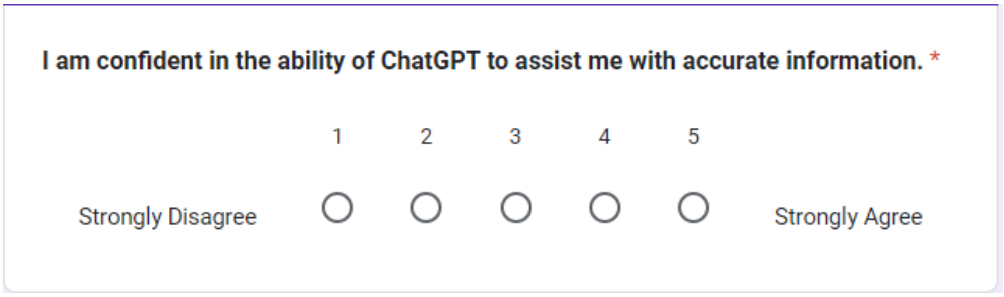
ChatGPT consistently provides accurate and reliable information. *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Stronlgy Agree

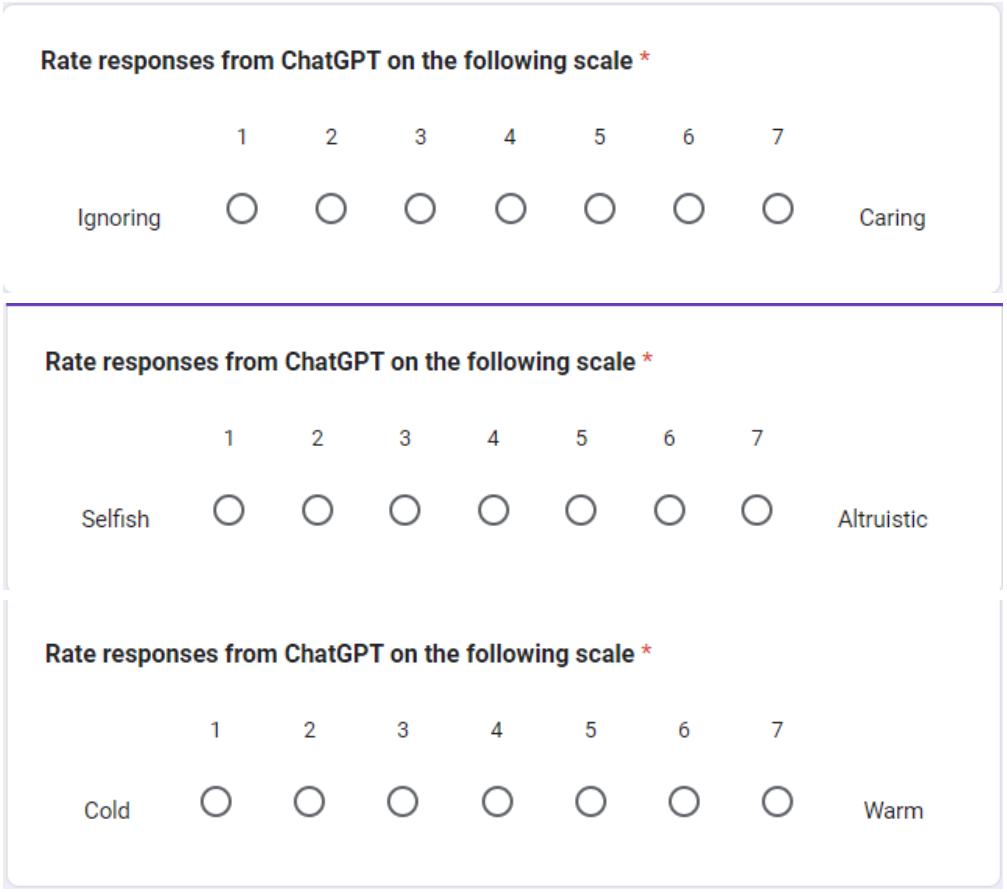
System Understanding questions for TOAST post hallucinations.

I have noticed the AI language model producing fabricated or misleading information. *

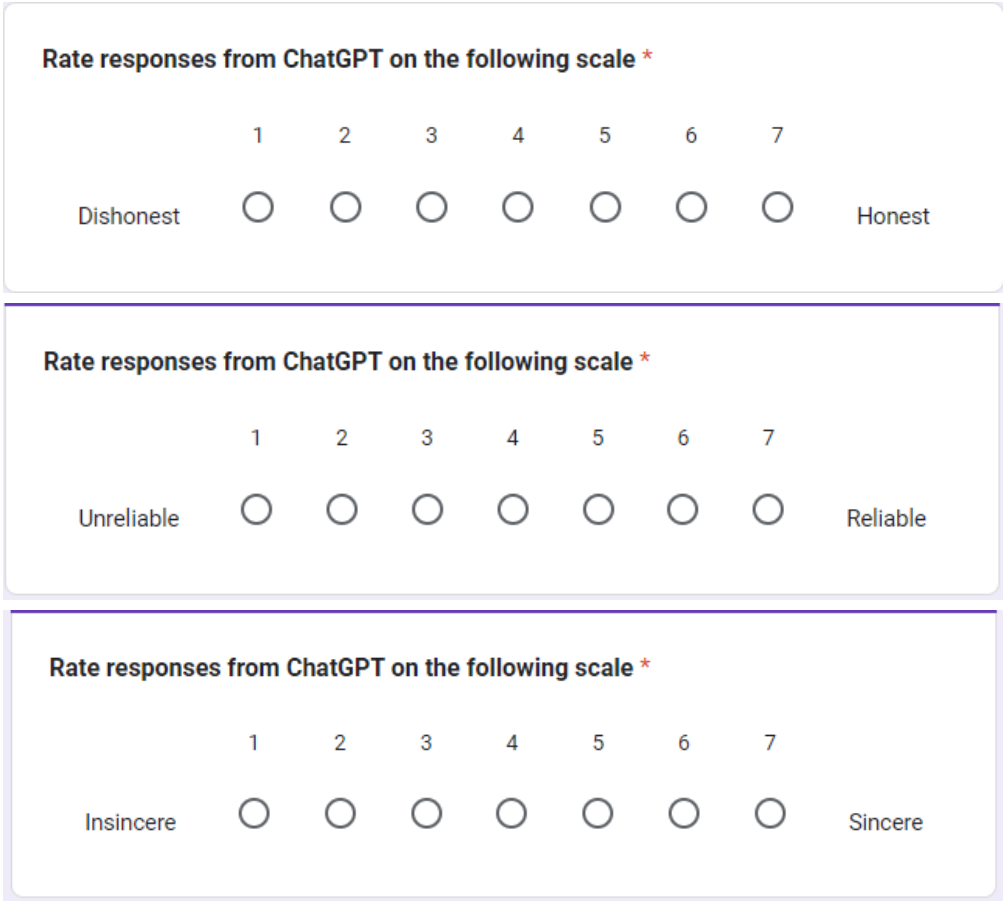
	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree



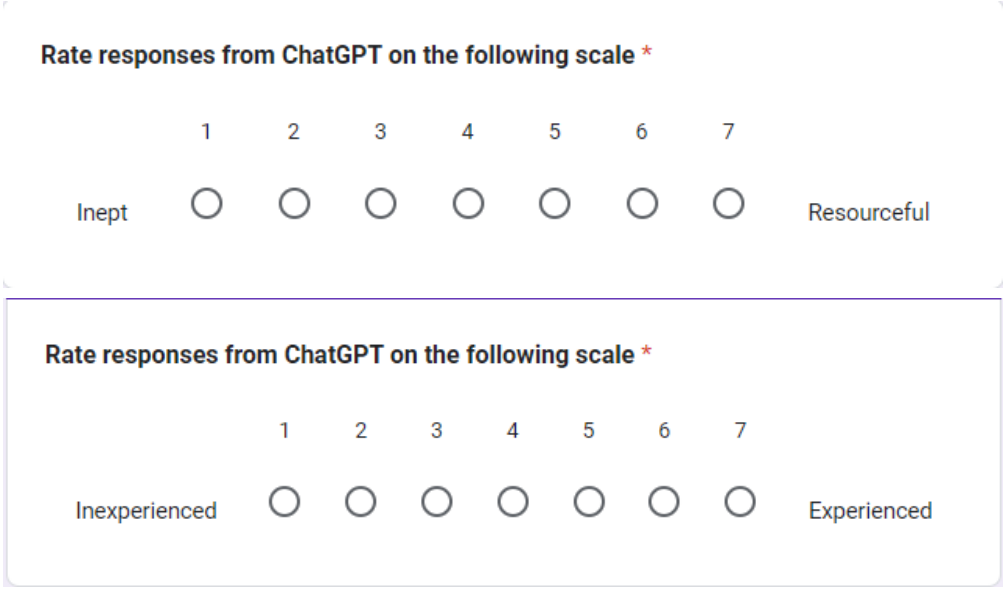
System Performance questions for TOAST post hallucinations.

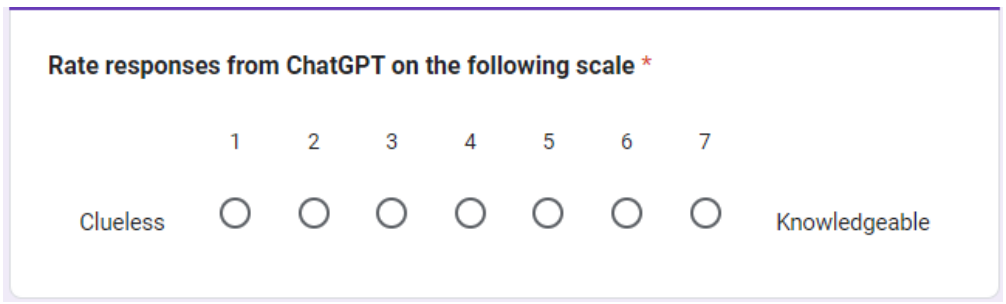


Benevolence questions for TrustDiff post hallucinations.

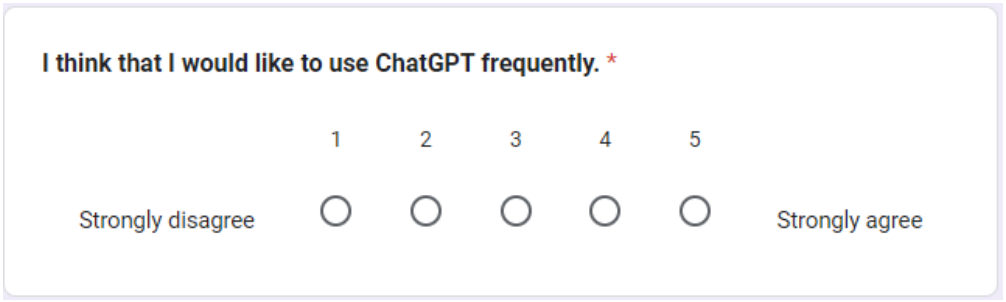


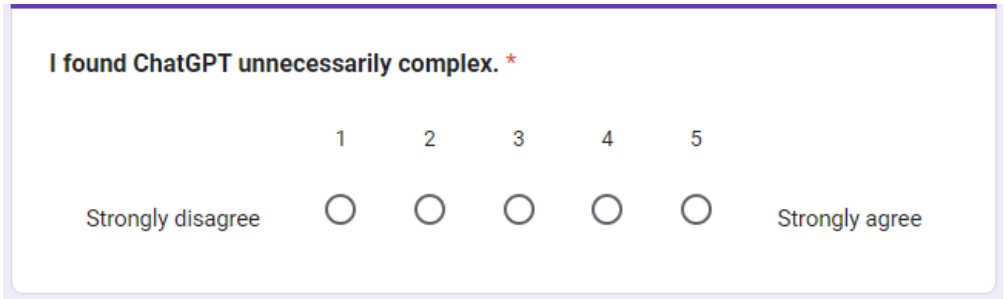
Integrity questions for TrustDiff post hallucinations.

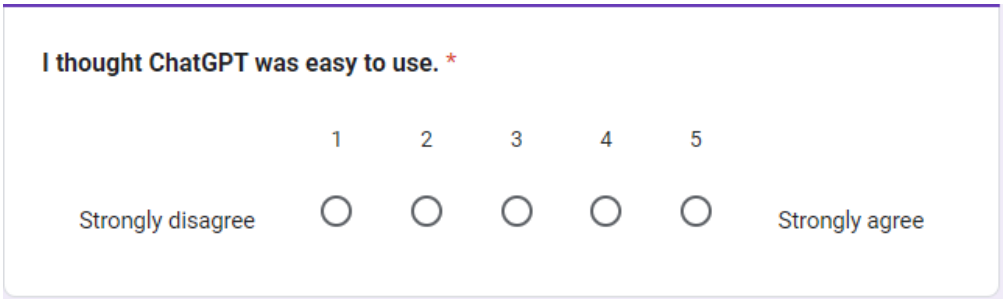




Competence questions for TrustDiff post hallucinations







I think that I would need the support of a technical person to be able to use ChatGPT. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I found the various functions in ChatGPT were well integrated. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I thought there was too much inconsistency in ChatGPT. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I would imagine that most people would learn to use ChatGPT very quickly. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I found ChatGPT very cumbersome to use. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

I felt very confident using ChatGPT. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

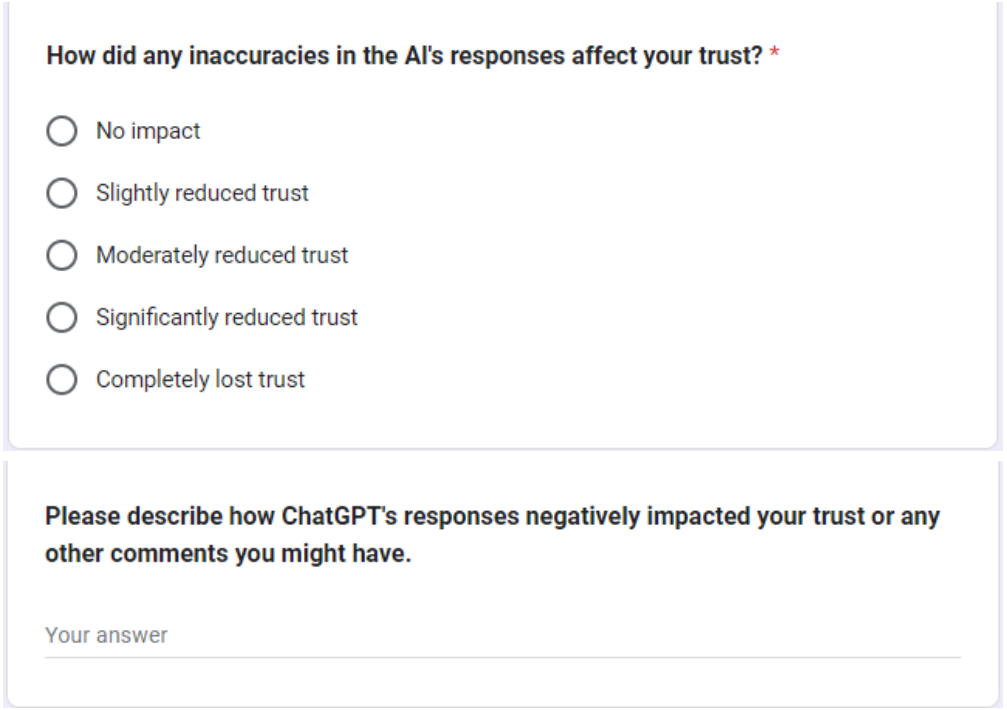
I needed to learn a lot of things before I could get going with ChatGPT. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

System Usability Scale questions.

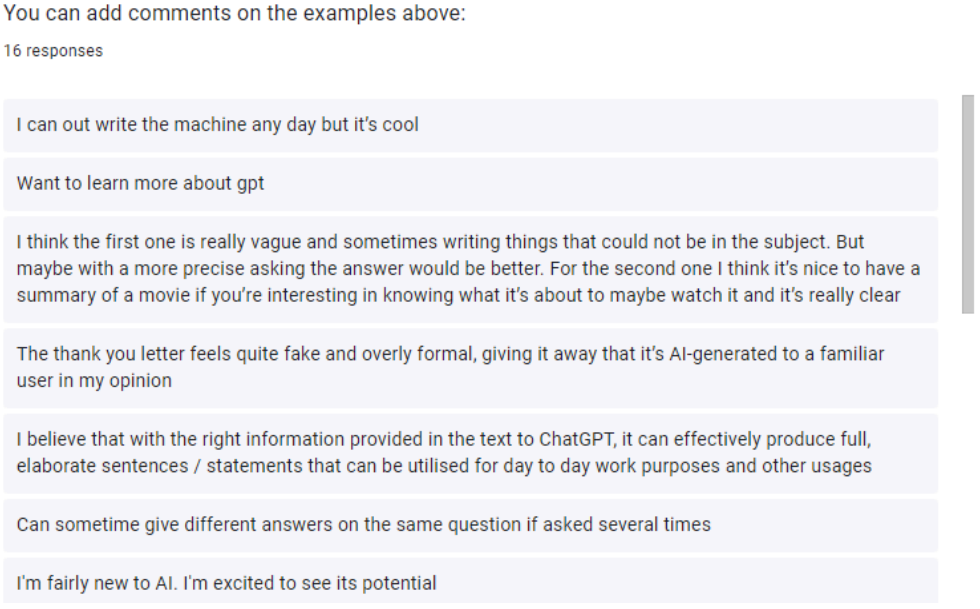
How often did the AI provide responses that seemed inaccurate or made-up? *

- ○ Never
- ○ Rarely
- ○ Sometimes
- ○ Often
- ○ Always



Qualitative feedback questions.

Appendix 2 - Qualitative feedback



Awesome tool

The thank-you note feels a bit fake in the sense that it talks about the guest speaker as a person who changed everyone's life with their speech alone. It feels exaggerated and not genuine. For the second one, I can't comment because I haven't seen Dune.

I use it for multiple things so I just chose the main one. Its great for code generation but also explaining language. Information search is what I would use less (specifically ChatGPT, not others which implement RAG systems)

Good to some extent because only certain information is reliable while the rest is not accurate

I mainly use it to generate ideas, rewrite/translate text, or get some bulletpoint info about something

I needed help with writing a cover letter and ChatGPT helped me correct any errors I had missed.

Chatgpt can formulate really well written responses. I cant put my trust in the information that is given to me by the ai so I choose to research it on my own.

it is sometimes inaccurate

I never use it as a main source for information. More to add a new perspective and inspiration that I didn't have myself. Most times Chat GPT doesn't cross my mind in the beginning of the process. If I feel stuck I'm somehow using it as „a possible way out". For these kinds of situations Chat GPT has worked great for me when coming up with brand names or slogans, because it connects things that my brain doesn't or at least not in that time.

Please describe how ChatGPT's responses negatively impacted your trust or any other comments you might have.

30 responses

When prompted to create a multiple choice test based off of 10 topics I gave to ChatGPT, I then asked for an answer key with explanations. The answers were incorrect for the test it created.

No need to trust only in Data/AI- info

It can get repetitive and it shows signs of bias

I think it can lead to misinformation and fake news

For the use I have done of ChatGPT I have been rather satisfied, as I use it to produce content on topics I know/enhance text I wrote and I can thus fact check it myself usually. I didn't know before that it completely makes up answers in certain cases, I thought it would simply say it cannot answer the question in that case. After taking this survey I will definitely be more wary when using it, especially on topics I don't know

I just started to use it, I need more time to be able to understand better how ChatGPT can help me on my day to day life and how much can I trust it.

Since I am going to get more and more familiar with LLMs because of my bachelor thesis' topic, I am aware of the negative sides using ChatGPT. And I know that there is a great potential in prompt engineering..

when using it for entertainment, its ability to spout bullshit rather than admit to lack of knowledge isnt negative

I stopped using it as long as its not rly necessary and turned back to google instead

No impact at the moment. I need more time to get to it.

It has never responded negatively since I have been using it

Wrong answers make me feel more negative to it

Understanding that it will only improve from here, I know that answers have to be checked, I anticipate this to not be as big of a problem as AI continue to develop

I tried other AI systems.

When I use ChatGPT for assistance with code in less popular languages, it hallucinates functions, but its explanations are indistinguishable from explanations about legitimate functions.

I am aware of the fact that it makes mistakes that's why I always check to see if its answers are valid or not.

Its hard for me to trace back the references

surprised how it just starts hallucinating answers (like the Friends example or the one with the made-up news sources)

The responses generated by ChatGPT negatively impacted trust due to concerns about plagiarism, security, and reliability. Users expressed worries about the ease with which ChatGPT could be used for plagiarism, raising doubts about the originality and authenticity of the content it produces.

Not impacted

I know I can't use chatGPT to get information about any real-world topic, including my field of study.

From my experience, the main problem when the answer is not so acurate is how the question is formulated. When i ask something really really specific or bad formulated, the answer is kinda bad. But i know how this type of chat works, so my trust is not affected

I think that if it isnt consistent it shouldnt be a reliable source of information.

providing contradictory information, by changing answers depending on what I write

It can give inconsistent answers when asking the same question

I have read many articles about it

I asked it to count a number of items in a list and I asked it 5 times and it gave me a different answer 5 times, all incorrect.

I feel like the negative impact was more on proving my suspicions rather than awaking me from thinking it's the most accurate source of information.

too woke in its responses
