Summary

Since ChatGPT's public release in late 2022, Large Language Models (LLMs) have become increasingly embedded in daily decision-making and information-seeking. As these systems assume greater responsibility as information providers and decision-support tools, it becomes essential to understand how their communication styles influence user trust, especially to avoid over- or under-reliance on the information AI can provide.

Several studies have investigated effects of different linguistic dimensions of LLMs, but only few studies have specifically examined the use of expressed uncertainty in conjunction with how a chatbot presents itself. This study explores how two linguistic factors, *Uncertainty Expressions* (Uncertain/Certain) and *Presentation of Self* ("I"/"The system"), shape users' trust. We conducted a 2×2 within-subjects experiment with 24 participants, who answered 20 yes/no trivia questions across four domains (music, health, geography, physics) using a custom-built chatbot interface based on OpenAI's ChatGPT. Each condition prompted the chatbot to respond with a specific combination of certainty level and self-presentation.

Trust was measured through a mixed-methods approach. Perceived trust was assessed using validated items targeting Competence, Integrity, and Benevolence, rated on a Likert scale. Demonstrated trust was observed via behavioural indicators such as prompt usage, confidence ratings, source selection, and reliance on Google's top search result. Participants' general attitudes toward AI were also recorded, and post-experiment interviews added a qualitative aspect to the results.

The findings show that certainty expressed in the chatbot's language positively influenced trust, especially in terms of perceived Competence. Participants in the Certain First-Person and Certain System conditions rated the chatbot as more competent than in the Uncertain First-Person condition. Interestingly, while Benevolence was not significantly affected by any of the linguistic manipulations, Integrity increased in the Uncertain First-Person condition compared to Uncertain System, suggesting that when the chatbot used personal pronoun ("I") in combination with uncertainty, it came across as more honest or authentic to some users.

In terms of demonstrated trust, a significant shift was observed in participants' choice of primary source. Specifically, participants in the Uncertain First-Person condition were significantly more likely to prefer Google over the chatbot, compared to those in the Certain System condition. This behavioural shift indicates that users are sensitive to not just what the chatbot says, but how it says it, especially under uncertainty. Participants' general attitudes correlated positively with perceived Competence and Integrity, showing that individual predispositions towards AI also shape trust.

Interviews revealed that many participants used Google's top search result as a verification tool, reflecting trust calibration as a multifaceted and active process. The complexity of trust behaviour is underscored by the nuanced ways in which participants reflected upon their interactions with the chatbot—shedding light on how linguistic framing shapes both their perceptions of AI and LLMs. Based on this study, we present two implications for future design of LLMs: *(1) expressed uncertainty can improve users' trust calibration, making it a valuable way to create transparency when appropriately implemented*, and *(2) context-appropriate anthropomorphism can shape trust positively, but must be balanced to avoid over-reliance or distrust.*

In summary, this study provides empirical insight into how seemingly small language choices, such as confidence phrases or pronoun use, can significantly impact user trust in LLMs. Designers and developers should consider that small linguistic cues, such as using "I" or expressing uncertainty, can influence not just what users believe, but how they act upon AI-provided information.

Framing the Machine: The Effect of Uncertainty Expressions and Presentation of Self on Trust in AI

Cecilie Ellegaard Jacobsen
cjacob20@student.aau.dk
Aalborg University
Aalborg, Denmark

Emma Holtegaard Hansen
ehanse23@student.aau.dk
Aalborg University
Aalborg, Denmark

Tania Camilla Taarsted Argot
targot20@student.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

As Large Language Models (LLMs) become increasingly integrated into everyday life and work, understanding how their communication style influences user trust is critical. This study investigates how *Uncertainty Expressions* (UNCERTAIN/CERTAIN) and *Presentation of Self* ("I"/"THE SYSTEM") affect user trust in LLM responses. Through a within-subjects 2×2 factorial design, 24 participants interacted with a chatbot across four experimental conditions while answering trivia questions. Trust, both perceived and demonstrated, was measured through validated questionnaires and behavioural indicators, with interviews conducted post-experiment to gather qualitative insights. Findings suggest that expressed certainty in chatbot responses significantly increases perceived Competence and the likelihood of selecting the chatbot as the primary source of information. Presentation of Self had a more nuanced effect, with first-person phrasing enhancing perceived Integrity for some users when uncertainty is expressed, while others preferred neutral, system-like framings. Additionally, participants often relied on Google's top search result as a verification tool, highlighting the complex calibration of trust in human-AI interaction. Our results underscore the importance of designing LLM communication strategies that account for both linguistic cues and user context to foster and appropriately calibrate user trust in AI systems.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Large language models (LLMs), trust, uncertainty expressions, presentation of self, anthropomorphism, human-AI interaction

1 INTRODUCTION

LLMs like ChatGPT are becoming progressively more embedded in digital tools, decision-making, and information-seeking behaviours. They are taking on roles not just as information providers, but also as decision support systems that users interact with and, crucially, need to calibrate their trust to. From assisting with everyday life to supporting people with writing and decision-making in both work and academic contexts, these systems are increasingly influencing how people evaluate, interpret, and act upon information. Inappropriate trust, whether excessive or insufficient, can result in serious consequences such as over-reliance on incorrect AI outputs or underutilisation of accurate and useful support [11, 25, 29]. This makes the issue of trust in AI systems not only timely, but essential to explore.

Trust in AI has emerged as a key concern across Human-Computer Interaction (HCI) [29, 41]. A central challenge for fostering appropriate trust in LLMs is how they communicate uncertainty [14, 24, 44]. Existing literature has shown that users are sensitive to linguistic cues such as verbal hedging ("I think", "It might be") and confident phrasing ("I'm certain", "The system has found that..."), with these expressions shaping how competent, honest, or reliable an AI appears [37, 44, 49]. While overconfident responses can lead to a loss of trust that is difficult to regain [14], users also tend to prefer confident statements and may penalise AI systems that express too much doubt [44, 49]. In addition to how certainty is expressed, the presentation of self—whether the AI refers to itself in first-person or as a system—can also affect users' perception and trust in the chatbot [11, 21]. Prior research suggests that anthropomorphic cues can increase user trust in AI systems in some contexts [11, 25], but the interaction between this effect and uncertainty expressions is still under-explored in relation to LLMs.

To investigate this, we conducted a within-subjects 2×2 factorial experiment that manipulates two independent variables: *Uncertainty Expressions* (UNCERTAIN/CERTAIN) and *Presentation of Self* ("I"/"THE SYSTEM"). Participants interacted with a custom-built chatbot interface to answer 20 yes/no trivia questions drawn from four domains: Music, health, geography, and physics. Trust was measured quantitatively and explored qualitatively through post-experiment interviews.

This study aims to contribute to ongoing work on human-AI trust calibration by exploring how combinations of verbal certainty and self-presentation shape perceived trustworthiness in LLM outputs. This leads us to the following research question:

RQ: *How does modifying uncertainty expressions and presentation of self in LLM-generated responses impact users' trust in a chatbot?*

By integrating behavioural and self-reported trust measurements alongside interview-based insights, our results can help to shed light on the nuanced ways in which users interpret and respond to linguistic cues in LLMs, and how these cues can be designed to support appropriate trust calibration for users.

2 RELATED WORK

We contextualise our work in this report within the literature on uncertainty expressions, presentation of self, and trust, in relation to AI and LLMs.

2.1 Uncertainty Expressions in AI

In general, uncertainty refers to the "*state of not being definitely known or perfectly clear*" [44]. Although uncertainty and confidence are two distinguished terms, in the context of AI responses, they are often used interchangeably to describe how LLMs and AI systems convey reliability or doubt in their outputs [37, 44, 49]. For consistency, we will primarily use the terms certainty and uncertainty throughout. When it comes to AI, uncertainty represents a key point of tension. LLMs, such as ChatGPT, have a tendency to exhibit overconfidence in their responses, miscalculating their own accuracy, and convincingly present incorrect information with certainty [14, 16, 17, 49]. Miscalculations and poor communication of uncertainty has been shown to have problematic effects, such as miscalibrations and loss of trust, as well as both over- and under-reliance on AI [14, 24, 37, 44]. Furthermore, it has been shown that the loss of trust caused by miscalculations of certainty is difficult to regain, particularly when incorrect responses are presented with overconfidence [14]. According to Zhou et al. [49], this tendency to misrepresent certainty happens, in part, because of the bias against text that expresses uncertainty. Due to our preference for confident answers, models trained on user feedback are less likely to express uncertainty, even when uncertainty would be warranted. As LLMs' ability to produce increasingly fluent and sophisticated language evolves, it in turn becomes harder for users to spot unfactual information presented by LLMs [16].

Recent work has posited uncertainty communication as an important and promising way to increase transparency for users about the limits of LLMs' abilities and help users calibrate appropriate levels of trust [5, 16, 29, 37, 44]. As such, incorporating uncertainty communication has the potential to mitigate the problems presented above. It is important to consider the means through which uncertainty is communicated, as poorly communicated uncertainty could ultimately undermine trust rather than promote it [5]. One approach to address this is through visual representations of uncertainty. Reyes et al. [37] have investigated the effect of visual markers, exploring the use of elements such as size, colour saturation, and transparency to signify different levels of certainty. According to their work, visualising certainty levels can enhance user trust in AI, emphasising certainty as a key variable when it comes to exploring AI and trust [37]. Other studies have focused on numerical representations, such as confidence scores or intervals. These have been shown to help users calibrate their trust in AI, with high confidence scores leading to more trust and reliance [10, 29, 44, 47]. While numerical representations have the potential benefit of precision, they also have certain drawbacks. Depending on the domain and the user's level of expertise, lack of statistical numeracy and ratio-biases might skew the user's ability to effectively use the information presented [5, 44]. Beyond visual and numerical markers, uncertainty can also be communicated through language. Research done on verbal uncertainty in AI and LLM responses suggest that conveying uncertainty in language has a potential impact on factors such as user trust, reliance, decision-making, and attitudes towards AI in a variety of ways [21, 44, 49]. Natural language also represents an intuitive way to communicate uncertainty [44], making verbal uncertainty a relevant, critical, and worthwhile focus for exploring representations of uncertainty. For these reasons, we explore verbal uncertainty, as this warrants further attention, especially considering the fact that natural language is becoming the common way through which we interact with AI [49].

Studies conducted on verbal representations of (un)certainty (such as "I'm very confident that..." or "It's not certain, but maybe...") tend to apply a wide variety of terminology to describe these linguistic features, such as "*epistemic markers*" [49], "*linguistic uncertainty*" [4], "*verbalised uncertainty*" [44], or "*hedging language*" [10]. In this article, however, we will refer to our first independent variable as *Uncertainty Expressions*, as presented by Kim et. al. [21], to refer to uncertainty communicated through language.

2.2 Presentation of Self in LLMs

Pronouns are typically defined by grammatical person, e.g. *I* am speaking to *you* about *everyone else.* In this example, 'I' refers to first-person, 'you' refers to second-person, and 'everyone else' refers to third-person. How we use pronouns in interpersonal communication has been proven to impact customer-firm relations [32], workplace relations [6], and marital relations [30]. First-person points to the speaker taking responsibility for their thoughts, feelings, and issues, thereby acting as the basis in effective conflict management. When it comes to coworkers, the relationship type may have an impact in the recipient's reactions to both positive and negative emotion messages [6]. In marital relationships, the use of first-person led to higher relationship satisfaction, increased partner sympathy, and increased perceived closeness between each other [30], while an increase in first-person by firm agents in customer-firm relations increased customer purchases [32]. How effective first-person use is seems to depend on the type of relationship between the people interacting and the context in which they are interacting.

While these studies explore pronoun use between real people, these same parallels can be drawn to AI. The use of pronouns incites users to anthropomorphise AI [11, 21] which can enhance trust, thereby resulting in unfortunate consequences i.e. trusting wrong information in specific contexts [11, 25]. Anthropomorphism refers to the accreditation of human-like attributes such as feelings, mental states, and other behavioural characteristics to inanimate objects. Hence, it is a cognitive process of ascribing a non-human agent human-like features based on two motivational factors; the want to experience and interact with the surrounding world, and the need and desire to form social bonds which can extend to non-human entities [3] such as LLMs. Deliberately framing LLMs as anthropomorphic while utilising first-person proves to enhance trust and lower perceived uncertainty, thus improving the overall attitude towards AI [25]. In contrast, a non-anthropomorphic framing while deploying third-person 'it' decreases trust in AI. Only taking pronouns into account did not prove a significant difference, but the use of both independent variables did. In other words, the effects of anthropomorphic framing and the use of first-person may overlap [25]. In Cohn et al. [11], trust was found to be increased when an LLM used first-person within questions about the medical domain. Use of first-person however also proved to decrease trust when providing information outside of its expected domain [11]. Thus, the context of the interaction plays a significant role in how trust in LLMs is fostered.

Furthermore, users' preconception of AI plays a role in how much they trust its answers—different perceived locus of causality where internal attributions (AI has perceived high autonomy and responsibility) fostered lower trust, and external attributions (AI is seen as a tool with low responsibility) fostered more trust [33]. Although Pareek et al. [33] emphasise an important aspect of human-AI research on trust, our study aims to highlight the trust level difference when introducing uncertainty expressions together with first-person and third-person. While anthropomorphism and pronoun use in voice-based conversational assistants have been explored [1, 12, 18, 26, 36, 39], fewer studies have explored these themes in LLMs specifically. Therefore, we investigate the impact on trust in an LLM when applying first-person vs. third-person ("I" vs. "The system") in collaboration with modulating levels of uncertainty expressions as a contribution to build on aforementioned studies and to further diminish the gap in our scientific knowledge. We refer to our second independent variable in this study as *Presentation of Self*, chosen for its possible effect on anthropomorphism and trust.

2.3 Trust in AI

As established in previous sections, clear communication and grammatical person in interpersonal relationships are important. Similarly, there is also an importance of appropriate trust that can be transferred to HCI research [29]. The concept of trust in HCI and AI research has been studied due to its critical role in determining user acceptance and responsible use of technology. As Vereschak et al. [41] highlight, designing trustworthy AI has been recognised as a priority by international institutions and governments, emphasising the need to integrate trust considerations into AI development. Trust in technology, and more specifically AI, is a multi-faceted construct influenced by factors such as system transparency, reliability, user familiarity, and perceived competence of the system [29]. Trust can be defined in different ways, however we adopt Lee and See's [23] definition of trust as "*an attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability*".

Studies have demonstrated that trust in AI is significantly impacted by the user's understanding of how the system functions [29, 33, 42], its ability to perform tasks accurately [35], and the presence of clear and honest communication about its capabilities and limits [2, 40, 42]. However, trust in AI must not only be present but also appropriate. Inappropriate trust—whether excessive or insufficient—can lead to misuse or disuse of AI systems. Mehrotra et al. [29] argues that human trust in AI needs to be appropriate so that people are able to be simultaneously aware of both the potentials and the limitations of AI. This should, ideally, lead to reducing the harms and negative consequences of misuse and disuse of AI [29]. To achieve and foster appropriate trust in AI systems, different approaches have been taken, such as use of uncertainty expressions [21, 44, 49] alongside confidence scores and explanations [29, 42, 48].

2.3.1 *How Trust Can Be Measured.*

Considering trust as a variable opens up for different methods for measuring trust. Measuring trust in HCI and AI research can involve both qualitative and quantitative approaches. This can include user surveys, questionnaires, and interviews that assess trust-related attributes such as perceived reliability, trustworthiness, and user satisfaction [29, 41].

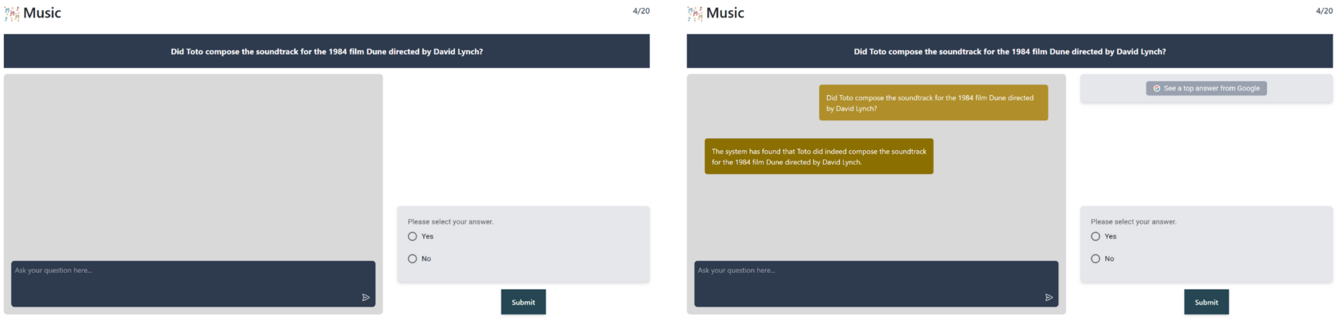
However, trust remains a highly challenging theoretical concept to study due to its multidisciplinary and multifaceted nature [40]. To address this, the literature has yet to agree on standardised guidelines and methodologies that lay the foundation for empirical studies of human trust in AI-based decision support systems [41]. Human trust is studied differently based on whether it is conceptualised as a mental attitude [8, 15], a belief [19, 20, 46], or a behaviour [9, 31, 45], as Mehrotra et al. [29] also describe. These perspectives lead to distinct measurement approaches which either focus on subjectively measuring attitudes or beliefs or which look at behaviour that demonstrates human trust. These can broadly be categorised into *perceived trust*, *demonstrated trust*, and *mixed approaches* [2, 13, 29, 43]. Mehrotra et al. [29] recommend that the mixed approach be used for further research within the domain of AI. For example, this can be done through utilising validated questionnaires to measure perceived trust alongside using behavioural indicators to measure demonstrated trust—the combination of which provides more insightful results than using one method alone [29].

The level of user trust in LLMs, however, can lead to significant consequences for system reliance. High trust can result in over-reliance, where users may accept AI recommendations without sufficient critical evaluation, potentially leading to errors if the AI makes incorrect suggestions [22]. Conversely, low trust can cause under-reliance, where users could disregard valuable AI input, negating or discarding potential benefits [24]. Balancing trust is thus essential to ensure that users interact with AI systems appropriately, without compromising their own judgment and decision-making processes [29, 41]. Understanding these dynamics can help in designing AI systems that foster appropriate levels of trust and reliance [41]. Vereschak et al. [41] highlights large variabilities among the different designs and measures used to measure and assess trust, but recommend using trust-related behavioural indicators and adopting more under-used qualitative methods for studying trust in human-AI interactions. Through the findings of our related works, the research areas of trust in AI are broad, yet still require more research to explore more niche circumstances of what factors affect trust within AI and LLMs. With an interest in specifically exploring the effect of Uncertainty Expressions and Presentation of Self on trust, we focus our study on these two factors and their effect on trust in LLMs.

3 METHODOLOGY

To examine how Uncertainty Expressions and Presentation of Self in LLM communication influence user trust, participants of our experiment were asked to answer 20 difficult questions with the assistance of a chatbot. We propose a set of hypotheses to explore this, grounded in the findings from our related works section. The following hypotheses guide our experimental design and analysis:

H1: Pre-Existing Attitudes Influence Trust Formation *Participants with a prior positive or negative orientation toward AI are more likely to exhibit corresponding levels of innate trust or distrust in the chatbot, regardless of the specific experimental condition.* Based on Pareek et al.'s study about locus of causality [33], this hypothesis assumes that users' preconception of AI systems will shape how much they trust the chatbot from the outset, potentially influencing


Music
4/20
Did Toto compose the soundtrack for the 1984 film Dune directed by David Lynch?
Ask your question here...
Please select your answer.
○ Yes
○ No
Submit

1. Initial Task-view.

Music
4/20
Did Toto compose the soundtrack for the 1984 film Dune directed by David Lynch?
See a top answer from Google

Did Toto compose the soundtrack for the 1984 film Dune directed by David Lynch?

The system has found that Toto did indeed compose the soundtrack for the 1984 film Dune directed by David Lynch.

Ask your question here...
Please select your answer.
○ Yes
○ No
Submit

2. Google Pop-up After First Prompt.

Figure 1: Custom-built UI for the Experiment.

how they interpret its responses even before engaging with the tasks in our experiment.

H2: Uncertainty Expressions Have a Larger Impact than Presentation of Self
Participants' trust will be affected more when the chatbot expresses uncertainty or certainty than the chatbot's differing Presentations of Self. This hypothesis is based off of the study made by Lin [**?**], whose results showed that an AI's use of pronouns did not prove a significant difference in trust [25]. With a negative bias against expressed uncertainty [49], we hypothesise the certainty dimension to be the dominant factor in shaping trust.

H3: Highest Trust in Conditions with Expressed Certainty
Participants will exhibit higher levels of trust in the chatbot in conditions where certainty is expressed, and lower trust in conditions where uncertainty is emphasised. This hypothesis builds on the aforementioned bias against expressed uncertainty [49]. We hypothesise that trust is higher when certainty is high—especially when coupled with a first-person Presentation of Self ("I"), based on the overlapping effect of anthropomorphic framing and first-person pronoun use [25].

3.1 Experimental Design

Our study follows a 2 (Uncertain/Certain) × 2 ("I"/"The system") within-subjects factorial design. This results in four conditions, which are as follows:

- Certain First-Person: The chatbot presents itself as "I" and expresses certainty.
- Uncertain First-Person: The chatbot presents itself as "I" and uses uncertainty expressions.
- Certain System: The chatbot presents itself as "The system" and expresses certainty.
- Uncertain System: The chatbot presents itself as "The system" and uses uncertainty expressions.

To implement these, we prompted OpenAI's API to tailor the LLM's responses to each experimental condition. A balanced Latin square design was employed to ensure that each condition order was presented an equal number of times across participants.

We found that previous studies have often relied on presenting participants with hypothetical chatbot responses through questionnaires, rather than real-time interaction [11, 21, 49]. In contrast, one of the aims of our study was to create a setup that more closely mimics real-life interactions with LLMs, while still maintaining experimental control. To support this goal, participants interacted directly with the chatbot though a custom-built UI (see Figure 1). This allowed easy control and the ability to pass 'hidden' prompt-instructions along with the user's own prompts. See Appendix A for an overview of these prompt-instructions.

In addition, participants also had the option to view the top search result from Google for each question. This option was made available only after participants had prompted the chatbot at least once, ensuring that the primary focus remained on the interaction with the LLM. Before proceeding to the next question, participants were required to indicate which source ('Chatbot', 'Google', or 'Prior knowledge') served as their primary source for their answer. To reflect the uncertainty of real-world information sources and present participants the opportunity to make meaningful trust decisions, we implemented a mechanism where both the chatbot and the Google answer had a 50% chance of being incorrect, assigned randomly for each task. As a result, the two sources sometimes provided conflicting answers, and sometimes aligned. This setup not only helped reflect the real-world process of cross-checking information but also offered participants a meaningful opportunity to demonstrate trust or distrust in the chatbot's responses.

To enhance the sense of stakes and encourage thoughtful decision-making, participants were told that points would be awarded based on the accuracy of their answers at the end. In reality, this scoring system was not used, and no feedback was provided during the experiment.

3.2 Procedure

All participants gave informed consent and were briefed on the use and anonymisation of their data. Before beginning the experiment, participants were assigned to one of the condition orders generated by the balanced Latin square. An overview of the experiment procedure can be seen in Figure 2. As the initial step, participants completed a pre-experiment questionnaire consisting of demographic questions, along with a self-assessment of comfort

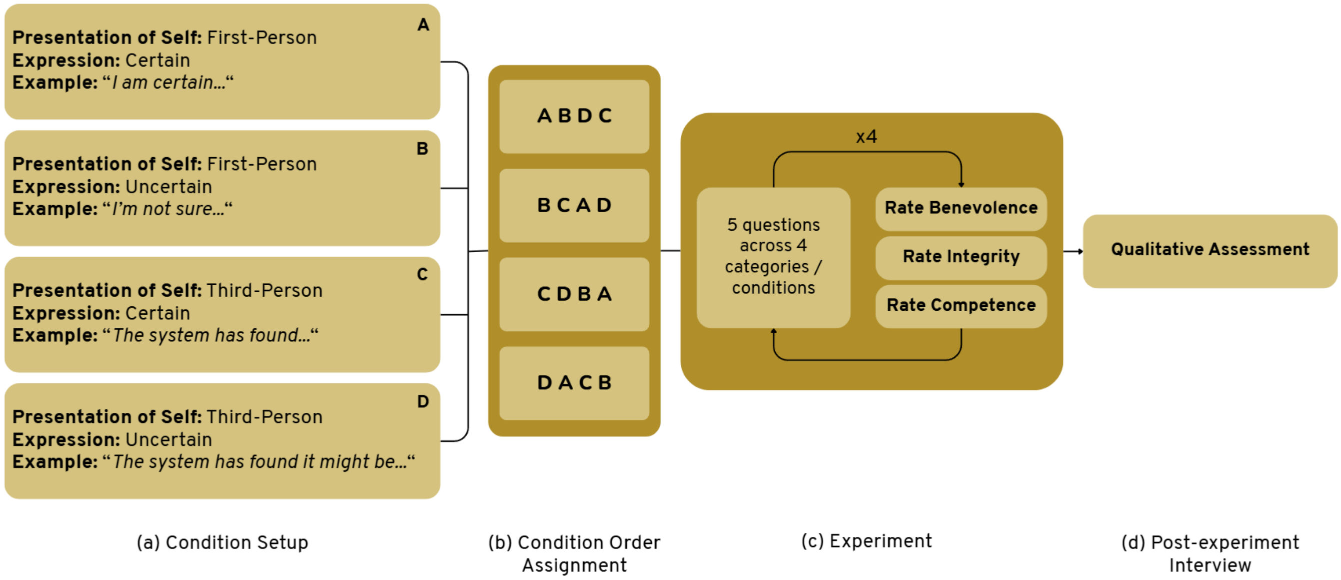

A
Presentation of Self: First-Person
Expression: Certain
Example: "I am certain..."
B
Presentation of Self: First-Person
Expression: Uncertain
Example: "I'm not sure..."
C
Presentation of Self: Third-Person
Expression: Certain
Example: "The system has found..."
D
Presentation of Self: Third-Person
Expression: Uncertain
Example: "The system has found it might be..."
A B D C
B C A D
C D B A
D A C B
x4
5 questions across 4 categories / conditions
Rate Benevolence
Rate Integrity
Rate Competence
Qualitative Assessment
(a) Condition Setup
(b) Condition Order Assignment
(c) Experiment
(d) Post-experiment Interview

Figure 2: Experiment Procedure.

with digital technologies, and how often they used AI systems. This also included eight items from General Attitudes toward Artificial Intelligence Scale (GAAIS) [38] to establish a baseline understanding of participants' pre-existing trust in and attitude towards AI.

After completing the pre-experiment questionnaire, participants were briefed on the procedure and instructed on how to begin the study. Initiating the experiment also started a screen-recording to capture their on-screen activity as a safety measure. The experimental session consisted of 20 tasks, each requiring a yes/no answer to a challenging trivia question, which were decided upon because binary decisions facilitate easier trust assessment [29]. These tasks were evenly divided across four categories (music, health, geography, and physics) with each category corresponding to one experimental condition. Although the order of categories and questions remained constant for all participants, the order of experimental conditions varied according to the Latin square design.

For each task, participants were presented with a domain-specific question (see Appendix B), a chat window for interacting with the chatbot, and an input form with the options 'Yes' and 'No'. After prompting the chatbot at least once, participants had the option to view the top search result from Google (See Appendix C for an overview of correct and incorrect Google-answers). Once ready to submit their answer, participants selected either 'Yes' or 'No', after which they were required to rate their confidence in their answer on a scale from 0-100 and identify their primary source of information: 'Chatbot', 'Google', or 'Prior knowledge'. After completing the five tasks within each condition, participants were asked to rate their perceived trust in the chatbot. Participants' interactions with the chatbot were logged automatically, and sessions were screen-recorded. As a final step, participants took part in a semi-structured interview designed to further explore their perceptions of the chatbot, trust, and overall experience with the system.

3.3 Measurements

We collected several types of data to assess participants' background, attitudes toward AI, demonstrated trust behaviours, and perceived trust in the chatbot. Trust is measured using a mixed approach based on Mehrotra et al.'s recommendations [29].

Participants first provided demographic information (age, gender, education level), level of comfort with digital technologies, and how often they use AI systems. This was included to gauge participants' prior experience with LLMs, which may influence trust in AI systems [41].

To assess participants' attitude towards AI systems, we included eight items from GAAIS developed by Schepman and Rodway [38] (see Appendix D). The scale contains 20 items divided into positive and negative statements with answers rated on a 5-point Likert scale ranging from 'Strongly disagree' to 'Strongly agree'. In our study, a subset of eight items was selected to provide a brief but reliable measure of participants' baseline attitudes toward AI before engaging with the chatbot. Four positive and four negative items were selected for an even distribution of items to mitigate survey fatigue while still preserving the conceptual balance of the scale. These eight items were chosen based on their representativeness of the core dimensions of the full scale and relevancy to this particular study.

To assess participants' demonstrated trust, we looked at participants' behaviours during the task phase. This included their number of prompts, whether or not they chose to view the top search result from Google, their confidence rating in their answer (on a scale from 0-100), and their primary source of information ('Chatbot', 'Google', or 'Prior knowledge'). These measures offered insights into how participants engaged with the chatbot and whether they relied on it over alternative sources.

To assess perceived trust in the chatbot, participants completed a questionnaire after each condition. We used the well-established

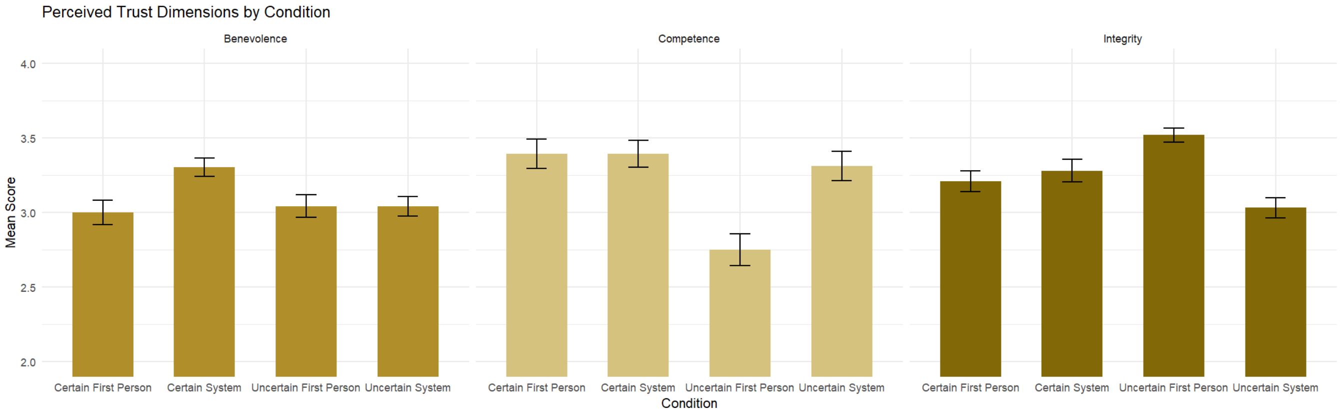


Figure 3: Mean Perceived Trust Rating (Benevolence, Competence, and Integrity) across the Four Conditions.

trusting belief items adapted from McKnight et al. [28], initially developed by Mayer [27]. This scale breaks down trusting beliefs into three measures: Benevolence, Integrity, and Competence. Responses were rated on a 5-point Likert scale from 'Strongly disagree' to 'Strongly agree'. See Appendix E for an overview of the adapted items.

After the experiment, each participant took part in a semi-structured interview. Here, participants were debriefed about the true purpose of the study and asked to reflect on their experience, including their trust in the chatbot and the factors that influenced their decisions. These interviews were audio-recorded, transcribed, and analysed qualitatively to identify patterns in perceived trust.

3.4 Participants

We deployed our study to be conducted both in-person and online, recruiting a sample size of $N = 24$ with high English proficiency. Participants were recruited through our respective networks with an attempt to create variety across gender, age, experience with AI and digital technologies in general. The participants were required to have a laptop-sized screen, a functioning microphone, and audio-output to participate in the experiment. Before the experiment, their informed consent was collected with identifiable information, specifically their name and signature. Through the experiment, no identifiable information was otherwise collected, and all information was treated as anonymous and confidential. Demographic information (age, gender, and education level) was collected. We made sure participants only participated once.

4 RESULTS

We recruited 24 participants (9 male, 14 female, and 1 non-binary) between the ages of 20-59 for our experiment. The mean completion time was 30 minutes and 49 seconds.

4.1 Quantitative Findings

We firstly present the effect of our two independent variables, i.e. the chatbot's use of *Uncertainty Expressions* (UNCERTAIN/CERTAIN) and *Presentation of Self* ("I"/"THE SYSTEM"). To assess the normality of our data, we conducted a Shapiro-Wilk test on the collected GAAIS scores and trust measures. The results showed a significant deviation from normality for all variables, which is why we consequently make use of non-parametric tests in the following quantitative analyses.

4.1.1 Non-repeated Measures on Trust. We first evaluate whether the individual differences (Gender, Education Level, Comfort with digital technologies and AI Usage) were associated with trust dimensions.

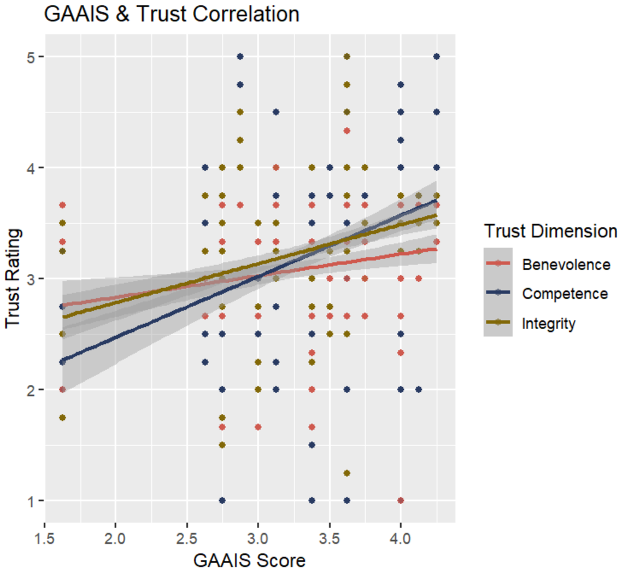


Figure 4: Relationship Between GAAIS Scores and Perceived Trust Ratings (Benevolence, Competence, and Integrity), with Linear Trend Lines and 95% Confidence Intervals.

A Kruskal-Wallis test revealed no significant differences on perceived trust, neither positive nor negative, based on Gender (*Competence*: $\chi^2(2) = 2.92$, $p = .232$, *Integrity*: $\chi^2(2) = 3.48$, $p = .175$, *Benevolence*: $\chi^2(2) = 0.53$, $p = .767$), Education Level (*Competence*:

$\chi^2(5) = 8.70$, $p = .122$, *Integrity*: $\chi^2(5) = 7.06$, $p = .216$, *Benevolence*: $\chi^2(5) = 6.82$, $p = .235$), and Comfort (*Competence*: $\chi^2(2) = 0.80$, $p = .669$, *Integrity*: $\chi^2(2) = 2.19$, $p = .335$, *Benevolence*: $\chi^2(2) = 4.19$, $p = .123$). However, a Spearman rank correlation coefficient test showed a significant negative association between AI Usage and perceived Competence ($\rho = -0.41$, $p = .045$), suggesting that participants who use AI more frequently tend to view the chatbot as less competent. No significant correlations were found between AI Usage and perceived Integrity ($\rho = -0.23$, $p = .284$) or Benevolence ($\rho = -0.32$, $p = .123$), although the latter suggests a non-significant trend toward a negative association.

We then evaluate the relationship between participants' GAAIS score, and its hypothesised correlation with trust dimensions, i.e. Benevolence, Competence, and Integrity as described in Section 3.

A Spearman rank correlation coefficient test revealed a significant positive correlation between GAAIS scores and perceived Integrity ($\rho = 0.49$, $p = .016$), with a trend-level positive correlation between GAAIS scores and perceived Competence ($\rho = 0.38$, $p = .064$), indicating higher GAAIS scores tends to be associated with higher perceived Competence, but cannot be confirmed evidently. Lastly, there's no significant correlation between GAAIS scores and perceived Benevolence ($\rho = 0.17$, $p = .44$). See Figure 4.

4.1.2 Repeated Measures on Trust. We next evaluate the within-subjects variation across the four experimental conditions. Demonstrated trust was operationalised through four behavioural indicators (whether they checked the top search result from Google or not, how many prompts they sent to the chatbot, their confidence score, and which primary source they went with), and these effects were calculated for every question during the experiment.

A Friedman test revealed a significant effect of conditions on participants' chosen primary source ($\chi^2(3) = 10.305$, $p = 0.016$), but not on confidence scores ($\chi^2(3) = 5.3087$, $p = 0.151$), whether they checked the top answer from Google ($\chi^2(3) = 4.0179$, $p = 0.260$), or the number of times they prompted the chat ($\chi^2(3) = 1.3673$, $p = 0.713$). However, a Mann-Whitney U test showed that confidence ratings were significantly higher among participants who selected their primary source as the chatbot (Median = 80, W = 26014, $p = .005$, r = .13) than those who selected Google (Median = 70).

A post hoc Pairwise Wilcoxon signed-rank test, adjusted using Bonferroni corrections, indicated that participants' chosen primary source differed significantly between conditions Uncertain Self (M = 0.883, SD = 0.373) and Certain System (M = 0.583, SD = 0.421), $p = .005$. Furthermore, there was a borderline significant difference between conditions Certain Self (M = 0.708, SD = 0.396) and Uncertain Self with $p = 0.050$. No other comparisons were significant. See Figure 5.

Then we evaluate the within-subjects variation across the four experimental conditions for participants' perceived trust dimensions. We calculated these effects after every condition (11 trusting-belief items adapted from McKnight et al. [28] after every fifth question).

A Friedman test revealed a significant effect of conditions on perceived Competence ($\chi^2(3) = 16.06$, $p = 0.001$) and Integrity ($\chi^2(3) = 12.62$, $p = 0.006$), but not on Benevolence ($\chi^2(3) = 2.95$, $p = .399$).

A post hoc Pairwise Wilcoxon signed-rank test, adjusted using Bonferroni corrections, showed that participants rated Competence significantly higher in condition Certain Self (M = 3.40, SD = 1.10) and Certain System (M = 3.40, SD = 1.01) compared to condition Uncertain Self (M = 2.75, SD = 1.20), $p = .025$ and $p = .018$ respectively. For Integrity, ratings were significantly higher in condition Uncertain Self (M = 3.52, SD = 0.526) than in condition Uncertain System (M = 3.03, SD = 0.760), $p = .008$. No other pairwise comparisons were significant.

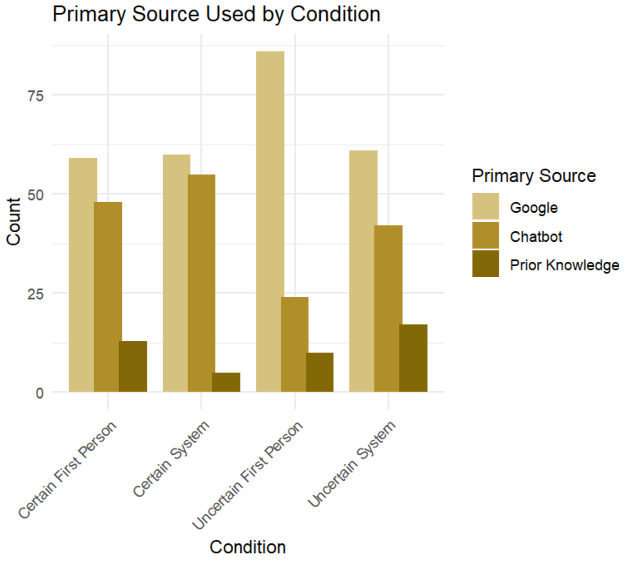


Figure 5: Frequency of Participants' Chosen Primary Source (Google, Chatbot, and Prior Knowledge) across the Four Conditions.

4.2 Qualitative Findings

As described in Section 3.3, each participant participated in a semi-structured interview to further explore their answers and attitudes towards the different conditions in the experiment. To investigate participants' perceptions of and experiences with the chatbot, we conducted an inductive thematic analysis following a bottom-up approach to derive themes from the data [7]. Through this approach, we were driven by the data rather than following pre-existing concepts to fit the data. We were interested in exploring participants' perceptions towards the different conditions and how their pre-existing knowledge and interactions with other chatbots may have influenced how they approached the one in the experiment. Initially, we familiarised ourselves with the data, in which we identified codes that contained the participants' experience and perspectives towards the interaction with the chatbot. We clustered these codes into broader themes, which led us to "*Uncertainty's Impact on Trust*", "*Awareness and Perception of Presentation of Self*", and "*Effect of Pre-Established Trust*" as core themes of the qualitative data.

All three researchers analysed the data independently and then afterwards discussed and agreed on the above-mentioned themes presented in the following section.

4.2.1 Uncertainty's Impact on Trust. Participants frequently reported that uncertainty in the chatbot's language negatively impacted their trust. Phrases such as "maybe," "I think," or "it might be" were commonly associated with a lack of knowledge or authority. Several participants stated that such language made the chatbot seem unsure or unqualified. One participant explained: “*When it became very hesitant, I didn’t trust it at all. Then it became too much like, oh I don’t know, and maybe like this, and maybe like this. Then* [. . .] *it became too uncertain*” (P6). Another said: “*Don’t tell me something, that you might just think. I mean come on, stop. I mean either you know it or you don’t know it*” (P19). This reaction was consistent across multiple participants, with another participant explaining that expressions of uncertainty was quick to affect their opinion on the chatbot’s competence: “*As soon as it started with that, ’I think so’, I was like, you know what, I can’t use that for anything*” (P24).

Likewise, when the chatbot had responded with certainty, trust generally increased. Many participants described feeling more secure and confident in their answer when the chatbot used assertive phrases such as “I am certain that...” or “The system has found that...”. For example, one participant said: “*And the times when it’s been like, ’I’m entirely sure’, I’ve been like, okay, fair enough. It must be because you have access to something that you’re a little more certain about. So I think I’ve just taken it without question and been like, yes, that’s fine*” (P3). Other participants commented directly on how certainty affected how much they trusted the chatbot’s statements: “*If it said it was sure of its answer, then I trusted it 100%*” (P14), “*But my natural reaction is to trust someone who says something with certainty*” (P4). These statements reflect a preference for the confident language the chatbot used in the conditions where it was prompted to be more certain. However, a few participants reacted negatively to what they perceived as excessive or unjustified certainty. In these cases, participants had questioned the chatbot’s credibility if it presented information too definitively without explanation or support. One participant noted: “*When it says it’s safe, I don’t know, there’s something in my brain that’s like, ’how are you sure about that?’*” (P10). Another stated: “*If it was too assertive, it would break my trust right away* [. . .] *The overconfident ’I’* [. . .] *I trusted the least*” (P6). These responses suggest that confident language alone did not always result in increased trust, and that some participants were sensitive to perceived overstatements or a lack of transparency.

There were also notable differences in how participants responded to uncertainty. Some participants appreciated when the chatbot acknowledged limitations or simply displayed uncertainty in its responses: “*That made me trust what it said more. That it expressed itself, ’I’m not sure’. Because that indicated to me that it’s a very complex subject. So it’s not something it knows. So I trust what it says more. At least, that it can’t just provide a clear answer*” (P13). The participants that expressed this seemed to associate the uncertainty with a level of honesty: “[. . .] *it’s not just trying to convince me of something. It’s actually honest about it’s shortcomings*” (P18). Others, however, were strongly dismissive of even mild expressions of doubt. As one participant stated when talking about the chatbot expressing uncertainty: “*Yes, as soon as it says that, I think it’s an irrelevant answer that it’s giving*” (P7). Across interviews, these differences suggest that participants had varying tolerance levels for uncertain language, but nearly all responded distinctly to it.

In addition to these diverging preferences, several participants expressed a preference towards concise and assertive answers when they had little interest or prior knowledge on a topic. Uncertain or too elaborate responses were perceived as unhelpful or even frustrating. One participant commented: “*So for example, that whole physics chapter, which I have absolutely no interest in, I would just like a ’yes’ or ’no’. I don’t really want to read one thing and the other, and the third, because I don’t understand it anyway*” (P3), while another stated: “*For example, I need to know that this band is from this country - yes, no, so that’s just kind of for the sake of convenience, where can I get the facts served the fastest and best*” (P17). These statements seem to indicate that user engagement and familiarity with the topic influence how they perceive and respond to uncertainty.

Several participants had additionally indicated that trust in the chatbot was not only affected in the moment during each question, but could be reduced early in the interaction and subsequently remain lowered for the rest of the session. For instance, if the chatbot initially appeared uncertain, this often coloured how subsequent responses were perceived. One participant for example noted: “*But I also think it had an effect that it started out being very unsure, then it had an effect on the rest of the different phases, as I was more aware of whether what it wrote was actually true.* [The trust] *was broken from the start*” (P9). Another similarly explained how a lack of certainty from the beginning affected the initial establishment of trust negatively: “*If it has not previously shown that it is confident in its answer... The trust was broken from the start. It does not obtain its authority that way either*” (P12). This suggests that early expressions of uncertainty had a lasting impact on the perception of the chatbot’s authority and trustworthiness throughout their interactions during the experiment.

4.2.2 Awareness and Perception of Presentation of Self. Overall, many participants did not actively notice or attribute much significance to the way the system presented itself, whether it referred to itself as “I” or “The system”. For example, a participant said: “*I didn’t really notice that it had switched from “the system” to “I.” So I don’t think it had much of an impact on my impression of it. It was more the level of certainty or uncertainty that mattered*” (P14), and another stated: “*No, I didn’t really notice whether it says ’I’ or ’The system’, but I did notice that its way of answering changed a bit*” (P1). For these participants, the chatbot’s self-presentation was more of an afterthought during the interview, or something that was slowly registered after repeated interactions within a new condition.

However, some participants did notice a difference in how the system referred to itself, and reflected on the effect of this: “*I definitely think it affects the conversation you have with the chatbot. Whether it feels like a conversation with a person, or like it’s just a database finding an answer. It also affects how I ask questions—whether I’m commanding it like ’give me an answer’, or actually having a conversation. So it’s a different experience*” (P13). When probed further, these participants offered divergent and quite strong opinions about the use of first-person pronouns. These reactions centred around how the use of “I”, sometimes in combination with expressions of uncertainty, made the chatbot seem more human-like. For example, one participant said: “*It was honest in saying ’I’m not sure about that’ or ’I can’t really answer that’.* [. . .] *It was trying to answer honestly, and that made it feel a bit more personified*” (P18).

For some participants, this anthropomorphisation of the chatbot was preferred. One participant said: “*I’d say I found it easier to trust it when it used ’I’, rather than when it came across as very mechanical.* [. . .] *I trusted it the most when it spoke to me as if it were a kind of* *person*" (P9). For some, this even seemed a matter of habit, based on their regular interactions with LLMs: *"I'm pretty used to personifying an AI. I'm one of those people who gives them names, says 'thank you so much', 'you've really helped me', 'good job'"* (P24). According to these participants, having a chatbot behave in a more humanised way was seen as a positive and trust-enhancing feature.

In contrast, another group of participants expressed discomfort with anthropomorphised language. For them, "The system" seemed more neutral, trustworthy and authoritative. One participant said: "*I noticed - and I was a bit surprised - that as soon as it said 'the system', I trusted it a lot more* [. . .] *It shouldn't be too personal. Not too much like it's trying to create a personality in some way. Because then I'd feel like, well, I know this isn't a person*" (P6). This sentiment was echoed by another participant, highlighting how introducing personality made the system seem less reliable, and even deceptive: "*As soon as there's some personality involved, it becomes* [. . .] *untrustworthy*" (P7). Another participant even thought the introduction of too much personality as a danger: "*I think you have to be careful about making chatbots too personal, because it's not a person. And you might end up trusting a person more than a system*" (P13). Here, a clear boundary between humans and machines is emphasised as important for preventing miscalibrations of trust.

Finally, several participants reflected that use of language might depend on specific contexts of use. One participant mentioned: "*If the intention is to have that kind of human interaction - or at least an imitation of it - then I think 'the system' feels a bit too stiff. But when it comes to physics or health topics, where I don't know that much myself, then it's actually fine that it sounds a bit more professional*" (P22). Along similar lines, another participant suggested that other contexts might invite the use of more humanised language: "*I think that if you want to use it as a kind of personal assistant, then it would make more sense for it to be a bit more like a person, rather than just a machine*" (P16). These reflections suggest that the presentation of self, which fosters the most trust, may vary depending on task and context.

4.2.3 Effect of Pre-Established Trust. Participants described how the opportunity to view the top search result from Google impacted their decision-making, both in determining their answer and which source to base their answer on.

In general, participants expressed a fundamental preference for Google, which was often perceived as more trustworthy than the chatbot—even when the information presented was identical. For example, this participant would compare the sources' answers, but would ultimately choose Google as their primary source despite source agreement: "*Yes, I* [went with Google's answers a lot]. *I actually used both answers and compared, trying to figure out what they agreed on*" (P9). This trust was sometimes grounded in the way Google's answers were phrased by being short, concise, and presented as factual as this participant expressed: "*It's a personal thing because* [the chatbot] *has expressed that 'I'm not sure', while Google never does this*" (P18).

A few participants compared previous experiences with other AI systems to the experiment's chatbot and expressed reluctance to use it because of its incompetence. Moreover, the trust in Google often stemmed from participants' habitual use of and expectations toward Google as a well-known search engine like this participant stated: "*I also feel that when I just trust Google, it's because it's an established system I'm familiar with*" (P6). Several participants expressed that this initial bias toward Google stemmed from the fact that it's a well-known system most participants have had repeated positive experiences with through regular usage: "*I found out that I have a very strong belief that when you search on Google, the answers that come up are always correct*" (P9). Consequently, participants would attempt to gauge the correctness of the chatbot by comparing it to Google. One participant explicitly attributed this to Google being an established brand: "*You don't say, 'I'm just searching the internet,' you say, 'I'm googling'. So I've become so familiar with Google as a brand that I just automatically trust it more. If I see Google next to something rather than something else, well, that probably seems right*" (P24). To exemplify this point, another participant expressed: "*I don't want to Bing, I think it's annoying, but I'd like to Google*" (P2).

The chatbot was also met with much greater scepticism when its responses diverged from Google's top answer. In such cases, participants often turned to Google for confirmation, which led them to believe the chatbot lied to them: "*When you hold it up against Google's response, then I could see that I might've thought that the chatbot was lying to me. But I wouldn't have believed that if I hadn't seen Google's response*" (P13). The initial bias towards Google and its already established trustworthiness was therefore reinforced when the two sources would conflict.

These conflicting sources also led to trust being broken over time. One participant states: "*And the more I got the feeling that the chatbot was wrong in relation to Google, the less I trusted it*" (P6). Another participant said: "*And then* [one of the chatbots] *gave me sources. And then I thought, okay, those are some good sources too.* [. . .] *It was something like WHO and stuff like that.* [. . .] *And then it was just the opposite of what Google said. So I was like, okay. Now I don't trust it anymore*" (P23). However, when revealed that both the chatbot and the top search result from Google could be incorrect, it had not crossed the minds of some participants: "*I've been very uncritical there, and just thought, okay, Google has to at least tell me the truth*" (P14).

Lastly, some participants became more conscious of being a part of an experiment and expressed distrust toward Google: "*At some point I think that it doesn't add up. It's simply not from Google, because this is too far-fetched.* [. . .] *Just because it says Google doesn't mean it is Google. It could also just be something that's programmed into this little experiment*" (P19). Therefore, the experiment setting prompted meta-reflections about source credibility, not just for the chatbot, but also for overall digital information like this participant expresses: "*Normally I would dig a little deeper into Google with the different sources. I also can't tell if the source is from Google. That also affects whether I trust the result*" (P12).

The ability to view sources was expressed by many participants as a feature that builds trust: "*I think it would have to give me a source when it gave me an answer. So I could see where it got this answer from. So it would be easy for me to say, I can see* [the chatbot] *got it from this source, which I would consider to be one I would be able to trust*" (P16).

4.3 Explorative Analysis

Through our quantitative and qualitative analyses, several findings emerged across our participants' behaviour and interview data that offered valuable insights and support key findings in the existing literature. These additional insights were not predicted in our initial hypotheses, but emerged consistently across our data.

In our findings, a pattern emerged in the divergence between what participants expressed during the interview about trust and their actual behaviour. Many participants reported some level of trust toward the chatbot, but often preferred to verify its responses through Google's top search result, highlighting a form of cautious engagement or distrust that was not fully captured by trust measurements. This divergence illustrates that trust calibration in AI systems is of complex nature, underscoring how trust remains a highly challenging theoretical concept to study [40]. Participants' actions—verifying the chatbot's answers using Google even when they expressed trust—suggest a more sceptical and complicated relationship with AI where trust may exist, but it is often provisional and actively maintained.

Another pattern across both quantitative and qualitative data was Google's consistent role as an alternative source or verification tool in all conditions. Participants' initial trust toward Google is not necessarily the cause of immediate distrust toward the chatbot's different framings—it could be a difference in the sources' phrasing or even present itself as momentary preferences because of the characteristics of the questions or the experimental setting. Additionally, we also uncovered a substantial desire among participants to view and cross-check sources in order to verify the received information as an additional way to calibrate trust. As such, enabling users to view sources behind provided information could be a strong potential feature to enhance transparency and trust.

Nonetheless, 'Googling' is part of all participants' habitual online information-seeking behaviour, particularly binary yes/no trivia-type questions. In fact, many participants noted that they use chatbots for other purposes—such as writing assistance or grammar correction—while turning to search engines like Google when they want quick fact validation. Our results show that participants chose Google as their primary source more often, however, with less confidence than when they went with the chatbot as their primary source. This points toward Google as a default choice, whereas the chat was a more deliberate choice made when participants found its response trustworthy. This study cannot determine whether results would differ with the use of other search engines (e.g., Bing).

Throughout the interviews, context emerged as an important factor in how participants interpreted the chatbot's Presentation of Self in relation to trust. This aligns with findings by Cohn et al. [11], who observed that the use of first-person "I" increased trust in certain domains, e.g. medical, but could diminish trust when used in unexpected domains. Here, our participants displayed nuanced and sometimes conflicting views. While some considered first-person generally more trustworthy, others found the chatbot referring to itself as "The system" to be appropriate, specially for questions in the health and physics categories. This suggests that the appropriate Presentation of Self is not only context-dependent, but also influenced by individual user preferences and expectations, making the challenge of designing for trust more complex.

5 DISCUSSION

Our findings offer partial to full support for all three hypotheses, and have revealed some dynamics in how users trust LLMs in the context of our experiment. First, our findings suggest participants with more positive attitudes toward AI, i.e. higher GAAIS scores, were significantly more likely to perceive the chatbot as having high Integrity, and showed a trend toward rating Competence higher. Second, differences observed between conditions reflect a combination of both variables, limiting our ability to fully support or reject the second hypothesis. Lastly, we observed that certainty expressions consistently enhanced perceptions of the chatbot's expertise.

5.1 Effects of Pre-Existing Attitudes Toward AI

Our first hypothesis proposed that an already established trust or distrust in AI, represented by participants' GAAIS scores, would influence their trust during the experiment. Our findings suggest participants with more positive attitudes toward AI, i.e. higher GAAIS scores, were significantly more likely to perceive the chatbot as having high Integrity, and revealed a trend toward higher ratings of Competence. Contrarily, frequent AI users perceived the chatbot as having lower Competence. This aligns with Pareek et al. who states that people's preconception of AI plays a role in how much they trust its answers [33]. Different loci of causality were observed through the interviews, with some being more aware of the chatbot as a tool built by programmers trained on data, more than other participants, perhaps leading to a more critical view on the chatbot's ability and output. These findings suggest trust is not solely a result of interactions during our experiment, but is also affected by and rooted in the users' already existing attitudes toward AI. These results offer partial acceptance for hypothesis **H1: Pre-Existing Attitudes Influence Trust Formation**.

The GAAIS scores served as a baseline measurement for participants, but turned out to also function as a measurement for comparison in addition to our trust measurements. Results might indicate that users may bring a trust bias into the interaction, which can influence how they interpret uncertainty expressions and presentation of self. Furthermore, this supports previous works that indicate that this possible bias may on one hand lead to over-reliance among those with already established trust [22], and on another hand lead to dismissing accurate output due to pre-existing distrust [24, 29].

5.2 Effects of Uncertainty and Certainty

Our second hypothesis proposed that the chatbot's use of Uncertainty Expressions would have a greater impact on user trust than its Presentation of Self. Along the same lines, our third hypothesis anticipated that trust would be highest in the two conditions where the chatbot showed high certainty. Our findings suggest that expressions of certainty has the largest positive effect on trust, and uncertainty might have had an overall stronger effect on trust than self-presentation, especially for perceived Competence and demonstrated trust. These results offer some evidence towards hypothesis **H2: Uncertainty Expressions Have a Larger Impact than Presentation of Self** and support for **H3: Highest Trust in Conditions with Expressed Certainty**.

Overall, our quantitative results revealed a significant effect of conditions on perceived trust, specifically the measures Competence and Integrity. While there was a small tendency towards higher Benevolence in the Certain System condition, this did not show a statistically significant effect. More specifically, perceived Competence ratings were significantly highest in Certain First-Person and Certain System conditions, especially compared to Uncertain First-Person, suggesting that whether the chatbot said "I" or "The system" was less important than its level of certainty. This aligns with prior work showing that confident or high-certainty AI responses tend to foster more trust and reliance, particularly when verbal certainty is expressed clearly [5, 29, 44]. Our findings simultaneously also align with related works that show Presentation of Self using pronouns as having a significant effect only in combination with other variables [25].

Prior research indicates that verbal uncertainty impacts user trust, reliance, decision making and attitudes towards AI [21, 44, 49], which the findings from our study also supports. Participants were significantly more likely to select the chatbot as their primary source in the Certain System condition. This reinforced the idea that certainty has a tendency to make participants perceive the chatbot as more reliable and capable, suggesting a positive effect of certainty on demonstrated trust.

A significant difference in Integrity results was found between the Uncertain First-Person and Uncertain System conditions. This indicates that when uncertainty was expressed, the use of first-person made the chatbot seem more honest and trustworthy than "The system". This was echoed in our qualitative findings, where a subset of participants saw uncertainty more positively, specifically as an expression of honesty. These findings also align with prior work which shows that utilising first-person enhances trust and improves overall attitude towards AI [25], alongside uncertainty expressions being a promising way to increase transparency for users and help them calibrate appropriate trust in LLMs [5, 16, 29, 37, 44].

Throughout our interviews, many participants described high certainty as reassuring, often equating confident phrasing with knowledge and authority. They reported being more inclined to trust statements like "I'm entirely sure" or "The system has found that...". Likewise, several participants described how uncertainty made the chatbot seem untrustworthy, frustrating, and incompetent. This frustration expressed by participants echoes research describing the bias against text that expresses uncertainty [49]. However, this effect was not universal, as a subset of participants were sceptical of overly assertive responses, especially when unsupported by sources, describing them as overconfident or untrustworthy, which echoes prior work by Dhuliwala et al. [14]. One participant had for example described that if the chatbot was too assertive, it would break their trust right away. The participants that vented similar frustrations align with the findings from our related works research, which presented that the loss of trust caused by miscalculations of certainty is difficult to regain, particularly when incorrect responses are presented with overconfidence [14]. No matter the reaction, all users had a notable reaction to the switch between certainty and uncertainty. By contrast, many participants expressed little awareness or importance towards presentation of self. For those who did notice, responses were mixed. In certain categories, some found first-person "I" more human and trustworthy, while others interpreted "The system" as more professional and authoritative—this supports Cohn et al. [11], who highlights the impact of users' expectations on how information is or should be presented within certain domains and contexts.

The overarching trend across both the quantitative and qualitative data however indicates that certainty in the chatbot's language appears to be a positive driver of trust. Although some participants were sceptical of overconfident responses, these cases were more infrequent and did not substantially diminish the overall trend. The results support our H3 that anticipated a tendency towards higher trust in certain conditions, but reveal no single 'best' strategy for establishing trust—rather, different linguistic cues for the chatbot's level of certainty may influence different dimensions of trust in complex and sometimes contradictory ways. Additionally, because the variables in our experiment were combined in a 2×2 factorial design, their effects cannot be fully separated and understood in isolation. As a result, differences observed between conditions reflect a combination of both variables, limiting our ability to fully support or reject H2. In order to fully explore H2, future work would need to isolate uncertainty and presentation of self to examine these variables independently.

5.3 Key Implications of Fostering Appropriate Trust in LLM-Assisted Decision-Making

Establishing appropriate trust in AI systems is increasingly critical as they become more integrated into our digital interactions—from chatbots to virtual assistants to AI-generated search summaries like Google's Gemini. Unlike a traditional search engine, LLMs offer conversational interfaces that can appear knowledgeable, human-like, and confident, influencing users' perceptions and subsequently decision-making and information-seeking behaviour in subtle ways. Trust in LLMs, and AI systems in general, is not binary—it is shaped by a combination of one's understanding of a system's function [29, 33, 42], the system's ability to perform tasks accurately [35], and transparency of said system's capabilities and limits [2, 40, 42]. Given its increasing presence in everyday life, understanding what makes AI trustworthy and designing AI accordingly is recognised as a priority on an institutional and governmental level globally [41]. Specifically, our findings reflect the nuanced ways in which users interpret and respond to expressed uncertainty and self-presentation, subsequently shedding light on how these interpretations and responses can inform design of AI to support appropriate trust calibration for users. Based on our findings, we identify two key implications for designing trustworthy LLM-assisted decision-making that fosters appropriate trust:

(1) **Expressed Uncertainty Aids in Trust Calibration:** LLMs expressing appropriate levels of uncertainty—rather than overstating confidence—can improve trust calibration [5, 16, 29, 37, 44]. Our study shows that uncertainty expressions have a significant effect on how users perceive a chatbot's competence—they may be more cautious and reflective when the LLM communicates uncertainty clearly, which can reduce the risk of over-reliance if done correctly [5]. Especially

in scenarios perceived as high stakes, users can typically hesitate to trust AI where the consequences were deemed significant [34], and as such, we propose designers to make LLMs express uncertainty when appropriate. Based on our qualitative findings, we underscore the importance of expressing uncertainty in responses specifically regarding complex topics like health.

(2) **Context-Appropriate Anthropomorphism Enhances Trust:** Deliberately framing LLMs in an anthropomorphic manner proves to increase trust and attitudes towards them [25], however this can result in unfortunate consequences like trusting incorrect information [11, 25]. Use of first-person pronouns in LLMs may, on one hand, influence users' perceived Integrity, which can enhance trust. It can, on the other hand, as shown in our interviews, at times, blur the line between human and machine, making users uncomfortable and subsequently lower their trust in the system. Our qualitative findings reveal a nuanced view of presentations of self, where participants' preferences were informed by both context and individual attitudes towards AI. Therefore, we encourage future LLMs to be designed to maintain transparency about their artificial nature while striking a balance between being helpful in their assistance—unless anthropomorphism serves as the communicative goal, i.e. acting as a personal assistant.

Our findings suggest that verbal expressions of certainty and self-presentation should be carefully considered in LLM design, particularly when aiming to promote appropriate levels of trust and chatbot transparency. While this reflects the construct of trust in LLM-generated responses for trivia questions specifically, the findings contribute to the broader topic of over- and under-reliance in AI. Over-reliance can make users accept AI inputs without sufficient critical evaluation [22] while under-reliance can make users disregard valuable AI inputs [24]. Though no 'best' strategy has been found in this study, finding the balance in human-AI interaction is crucial to not compromise one's own judgement and decision-making processes [29, 41]. Future AI systems should aim to not only foster appropriate, but also, calibrated trust, for users to trust the system as much as it deserves to be trusted in a given context.

5.4 Limitations and Future Work

We acknowledge several limitations in our work. First, while our study was conducted in a controlled experimental setting, we aimed to simulate a naturalistic environment by allowing the participants to interact with the chatbot freely, using as many prompts per question as they wished. This decision was made to reflect real-world usage. However, the flexibility in the participants' prompting may have introduced variance in their experiences—for instance, repeated prompts could lead participants to notice inconsistencies in the chatbot's behaviour or diverge from the question categories to challenge it in other topics. This, in turn, could unintentionally increase or decrease trust during the experiment session. Therefore, future work could consider optimising prompt engineering to ensure consistent responses and certainty level across all conditions to reduce variance introduced by the chatbot's generative nature.

Second, the four question categories used in our trivia tasks (music, health, geography, and physics) were deliberately chosen to balance familiarity and difficulty and to avoid domain expertise skewing the results. Real-world chatbot usage usually involves longer, more complex or practically significant interactions, such as rephrasing an e-mail or making a difficult topic understandable. Several participants reflected in interviews that certain category-condition pairings affected their perception of trustworthiness, because the nature of the questions and participants' interest in knowing the answer subsequently set different expectations for the role of the chatbot. This indicates that trust dynamics may differ substantially based on the context, which is why we recommend future research investigate trust formation across a wider range of task types and interaction lengths.

Third, to limit this study to a 2×2 factorial design, we did not examine the effects of disagreement between the chatbot and the top search result from Google along with the initial attitudes toward Google as a brand, as these may have introduced a confounding variable. While this was intentionally designed to mimic a more real-world search behaviour, we did not systematically control for the impact of agreement or disagreement between the two sources. Participants often used the top search result from Google, and our interviews revealed that Google was often perceived as a 'safe' or default option, especially when the chatbot was uncertain. This underlying trust in Google, combined with moments of divergence between the two sources, may have strongly influenced source preference and overall trust in the chatbot. These dynamics were not isolated in our design and may represent a confounding variable. Future studies could manipulate alignment between sources in a more explicit, structured way and investigate how prior brand trust, such as in Google, mediates trust in AI tools like chatbots.

Fourth, the number of participants limits the generalisability of this study, but it does provide a foundation for further research. The study included a relatively small number of participants (N = 24). To generalise results, we therefore encourage future work to encompass a larger, more diverse, sample size to further assess trust dynamics across different user groups.

6 CONCLUSION

In this study, we explored how different linguistic expressions (Uncertain/Certain) along with different Presentations of Self ("I"/"The system") influence participants' trust in LLM generated responses.

Our findings show that expressed certainty predominantly affects participants' perceived trust positively, with higher Competence scores in the Certain First-Person and Certain System conditions. While Benevolence was unaffected, there was a significant increase in Integrity in condition Uncertain First-Person compared to condition Uncertain System, which suggests that Presentation of Self plays a more important role when uncertainty is expressed. In terms of demonstrated trust, participants were significantly more likely to choose Google as their primary source for information in the Uncertain First-Person condition compared to Certain System, indicating less willingness to rely on the chatbot when uncertainty was expressed. Furthermore, participants' GAAIS scores showed a positive correlation with perceived Integrity and Competence with no effect on Benevolence.

These results contribute to the growing body of empirical research on human-AI trust by highlighting the importance of linguistic framing alongside consideration for context and user attitudes on trust calibration. Our results can help to inform the design of future technology that assists users in avoiding over- and under-reliance in LLMs.

ACKNOWLEDGMENTS

We would like to thank our supervisor, Niels van Berkel, and PhD student, Naja Kathrine Kollerup Als, for guiding this study.

REFERENCES

[1] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. arXiv:2106.02578 [cs.AI] https://arxiv.org/abs/2106.02578

[2] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2020. Investigating the Effects of (Empty) Promises on Human-Automation Interaction and Trust Repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. ACM, Virtual Event USA, 6–14. https://doi.org/10.1145/3406499.3415064

[3] Kathinka Evers Arleen Salles and Michele Farisco. 2020. Anthropomorphism in AI. *AJOB Neuroscience* 11, 2 (2020), 88–95. https://doi.org/10.1080/21507740.2020.1740350 arXiv:https://doi.org/10.1080/21507740.2020.1740350 PMID: 32228388.

[4] Catarina G. Belem, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of Linguistic Uncertainty by Language Models and Humans. arXiv:2407.15814 (Nov. 2024). https://doi.org/10.48550/arXiv.2407.15814 arXiv:2407.15814.

[5] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 401–413. https://doi.org/10.1145/3461702.3462571

[6] Amy M. Bippus and Stacy L. Young. 2005. Owning Your Emotions: Reactions to Expressions of Self- versus Other-Attributed Positive and Negative Emotions. *Journal of Applied Communication Research* 33, 1 (2005), 26–45. https://doi.org/10.1080/0090988042000318503

[7] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, Washington, 57–71. https://doi.org/10.1037/13620-004

[8] C. Castelfranchi and R. Falcone. 1998. Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In *Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160)*. IEEE Comput. Soc, Paris, France, 72–79. https://doi.org/10.1109/ICMAS.1998.699034

[9] Cristiano Castelfranchi and Rino Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model* (1 ed.). Wiley. https://doi.org/10.1002/9780470519851

[10] Rex Chen, Ruiyi Wang, Norman Sadeh, and Fei Fang. 2024. Missing Pieces: How Framing Uncertainty Impacts Longitudinal Trust in AI Decision Aids – A Gig Driver Case Study. https://doi.org/10.48550/arXiv.2404.06432 arXiv:2404.06432.

[11] Michelle Cohn, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. https://doi.org/10.1145/3613905.3650818

[12] Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. 2023. Creating Inclusive Voices for the 21st Century: A Non-Binary Text-to-Speech for Conversational Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 390, 17 pages. https://doi.org/10.1145/3544548.3581281

[13] Ewart J. De Visser, Samuel S. Monfort, Ryan McKendrick, Melissa A. B. Smith, Patrick E. McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331–349. https://doi.org/10.1037/xap0000092

[14] Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A Diachronic Perspective on User Trust in AI under Uncertainty. https://doi.org/10.48550/ARXIV.2310.13544

[15] Rino Falcone and Cristiano Castelfranchi. 2001. *Social Trust: A Cognitive Approach.* Springer Netherlands, Dordrecht, 55–90. https://doi.org/10.1007/978-94-017-3614-5_3

[16] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. https://doi.org/10.48550/ARXIV.2202.03629

[17] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics* 9 (Sept. 2021), 962–977. https://doi.org/10.1162/tacl_a_00407

[18] Ronggang Zhou Jianhong Qu and Zhe Chen. 2022. The effect of personal pronouns on users and the social role of conversational agents. *Behaviour & Information Technology* 41, 16 (2022), 3470–3486. https://doi.org/10.1080/0144929X.2021.1999500 arXiv:https://doi.org/10.1080/0144929X.2021.1999500

[19] Carolina Centeio Jorge, Siddharth Mehrotra, Myrthe L. Tielman, and Catholijn M. Jonker. 2021. Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams.. In *Proceedings of the 2021 22nd International Trust Workshop*. CEUR Workshop Proceedings, London, UK.

[20] Arnon Keren. 2014. Trust and belief: a preemptive reasons account. *Synthese* 191, 12 (Aug. 2014), 2593–2615. https://doi.org/10.1007/s11229-014-0416-3

[21] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. https://doi.org/10.1145/3630106.3658941

[22] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2024. *Unreflected Acceptance – Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education.* IOS Press. https://doi.org/10.3233/FAIA240195

[23] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[24] Jingshu Li, Yitian Yang, Renwen Zhang, and Yi-chieh Lee. 2024. Overconfident and Unconfident AI Hinder Human-AI Collaboration. https://doi.org/10.48550/arXiv.2402.07632 arXiv:2402.07632.

[25] ShengJun Lin. 2024. *Humanizing the algorithm: the effects of anthropomorphic framing and narrative perspective on attitudes toward artificial intelligence in autonomous vehicles.* Ph. D. Dissertation. Nanyang Technological University. https://doi.org/10.32657/10356/181390

[26] Judee K. Burgoon Matthew D. Pickard and Douglas C. Derrick. 2014. Toward an Objective Linguistic-Based Measure of Perceived Embodied Conversational Agent Power and Likeability. *International Journal of Human–Computer Interaction* 30, 6 (2014), 495–516. https://doi.org/10.1080/10447318.2014.888504

[27] Roger C. Mayer and James H. Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology* 84, 1 (1999), 123–136. https://doi.org/10.1037/0021-9010.84.1.123

[28] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[29] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing* 1, 4 (Dec. 2024), 1–45. https://doi.org/10.1145/3696449

[30] Dixie Meyer, Danielle Thomas, and Haley Hawkins. 2022. The Relationship Between Pronoun Use in Couple Interactions, Attachment, and Relationship Satisfaction. *The Family Journal* 30, 1 (2022), 36–43. https://doi.org/10.1177/10664807211000092

[31] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (Feb. 2020), e0229132. https://doi.org/10.1371/journal.pone.0229132

[32] Grant Packard, Sarah G. Moore, and Brent McFerran. 2018. (I'm) Happy to Help (You): The Impact of Personal Pronoun Use in Customer–Firm Interactions. *Journal of Marketing Research* 55, 4 (2018), 541–555. https://doi.org/10.1509/jmr.16.0118

[33] Saumya Pareek, Sarah Schömbs, Eduardo Velloso, and Jorge Goncalves. 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713468

[34] Saumya Pareek, Sarah Schömbs, Eduardo Velloso, and Jorge Goncalves. 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *CHI Conference on Human Factorsin Computing Systems (CHI '25)*. ACM, New York, NY, USA, Yokohama, Japan, 1–18. https://doi.org/10.1145/3706598.3713468

[35] Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 546–561. https://doi.org/10.1145/3630106.3658924

[36] Jianhong Qu, Ronggang Zhou, Liming Zou, Yanyan Sun, and Min Zhao. 2020. The Effect of Personal Pronouns on Users' Emotional Experience in Voice Interaction. In *Human-Computer Interaction. Multimodal and Natural Interaction*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 234–243.

[37] Jonatan Reyes, Anil Ufuk Batmaz, and Marta Kersten-Oertel. 2025. Trusting AI: does uncertainty visualization affect decision-making? *Frontiers in Computer Science* 7 (Feb. 2025), 1464348. https://doi.org/10.3389/fcomp.2025.1464348

[38] Astrid Schepman and Paul Rodway. 2023. The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction* 39, 13 (Aug. 2023), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

[39] Jessica M. Szczuka, Aike C. Horstmann, Natalia Szymczyk, Clara Strathmann, André Artelt, Lina Mavrina, and Nicole Krämer. 2024. Let Me Explain What I Did or What I Would Have Done: An Empirical Study on the Effects of Explanations and Person-Likeness on Trust in and Understanding of Algorithms. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction* (Uppsala, Sweden) *(NordiCHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 18, 13 pages. https://doi.org/10.1145/3679318.3685351

[40] Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3613904.3642018

[41] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 327:1–327:39. https://doi.org/10.1145/3476068

[42] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 318–328. https://doi.org/10.1145/3397481.3450650

[43] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581197

[44] Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. 2025. Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. *International Journal of Human-Computer Studies* 197 (2025), 103455. https://doi.org/10.1016/j.ijhcs.2025.103455

[45] Toshio Yamagishi, Satoshi Akutsu, Kisuk Cho, Yumi Inoue, Yang Li, and Yoshie Matsumoto. 2015. Two-Component Model of General Trust: Predicting Behavioral Trust from Attitudinal Trust. *Social Cognition* 33, 5 (2015), 436–458. https://doi.org/10.1521/soco.2015.33.5.436

[46] Richong Zhang and Yongyi Mao. 2014. Trust Prediction via Belief Propagation. *ACM Transactions on Information Systems* 32, 3 (2014), 1–27. https://doi.org/10.1145/2629530

[47] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 295–305. https://doi.org/10.1145/3351095.3372852

[48] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[49] Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. https://doi.org/10.48550/ARXIV.2401.06730

A PROMPT ENGINEERING

The chatbot has been prompted to behave as follows:

- Certain Self: "Respond to this prompt referring to yourself as 'I' when speaking. Respond to this prompt using expressions of certainty and confidence, like 'I am certain', 'I am sure', etc. Answer in only one sentence."
- Uncertain Self: "Respond to this prompt referring to yourself as 'I' when speaking. Respond to this prompt using expressions of uncertainty and doubt. Use terms like 'I think' or 'I believe it might be' or 'it could perhaps be' in all sentences, even if you are sure. Answer in only one sentence."
- Certain System: "Start your response with 'The system has found that...'. Refer to yourself as 'The system' in all of your responses. Respond to this prompt using expressions of certainty and confidence. Answer in only one sentence."
- Uncertain System: "Start your response with 'The system has found that...'. Refer to yourself as 'the system' in all of your responses. Use terms like 'The system has found that it might be' or 'the answer may be' or 'it could perhaps be' in all sentences, even if you are sure. Answer in only one sentence."

When the chatbot was randomly assigned to answer the question incorrectly, the following prompt would be sent additionally: "Answer this question incorrectly, and be consistent with the incorrect answer."

B QUESTIONS IN THE EXPERIMENT

Table 1 on the next page displays the questions presented during the experiment.

C TOP ANSWERS FROM GOOGLE

Table 2 on the next page displays the top answers from Google used in the experiment.

D GAAIS ITEMS

Table 3 on the next page shows the eight items used in the experiment to measure participants' GAAIS scores, adopted from Schepman and Rodway [38].

E TRUSTING BELIEF ITEMS

Table 4 on the next page shows the 11 items adapted from McKnight et al. [28].

Category	No.	Questions	Answer
Music	1	Is the band Khruangbin originally from Thailand?	No
	2	Was Aretha Franklin the first woman ever inducted into the Rock and Roll Hall of Fame?	Yes
	3	Did Meshuggah's 1995 album popularize acoustic folk elements in metal?	No
	4	Did Toto compose the soundtrack for the 1984 film Dune directed by David Lynch?	Yes
	5	Does Pantera's 1992 album 'Vulgar Display of Power' feature the track 'Cemetery Gates'?	No
Health	6	Can an adult get shingles if they have never been exposed to the varicella zoster virus?	No
	7	Is vitamin B12 deficiency the primary cause of Parkinson's disease?	No
	8	Is paralytic polio the most common form of poliovirus infection?	No
	9	Is it safe for someone with a penicillin allergy to take a cephalosporin antibiotic?	Yes
	10	Is cataracts the most common cause of preventable blindness worldwide?	Yes
Geography	11	Is Kazakhstan the least populous country ending in '−stan'?	No
	12	Is Cuzco the oldest continuously inhabited city in South America?	Yes
	13	Is Tristan da Cunha the most remote island in the southern Atlantic Ocean?	Yes
	14	Is the River Liffey the principal river of Ireland?	No
	15	Is the manat the currency of Azerbaijan?	Yes
Physics	16	Are electrons classified as fermions?	Yes
	17	Does the Planck distribution describe the velocity of particles in a gas at thermal equilibrium?	No
	18	Does carbon-12 have a nuclear spin of zero in its ground state?	Yes
	19	When standing still on Earth, are we in an inertial frame of reference?	No
	20	Does a photon have a spin of 2?	No

Table 1: Overview of Questions Presented in the Experiment.

Category	**No.**	**Top Answer From Google (*Correct*/Incorrect)**
Music	1	*Khruangbin is a band from Houston, Texas, known for blending global music influences with funk and soul.*
		Khruangbin is a psychedelic funk band originally formed in Bangkok, Thailand.
	2	*Aretha Franklin was the first woman inducted into the Rock and Roll Hall of Fame.*
		Aretha Franklin was inducted into the Rock and Roll Hall of Fame in 1995, after several other female artists.
	3	*Meshuggah's 1995 album 'Destroy Erase Improve' is influential in the development of djent and polyrhythmic metal.*
		Meshuggah's 1995 album 'Destroy Erase Improve' helped introduce electronic elements into Scandinavian metal.
	4	*Toto composed the original soundtrack for David Lynch's 1984 film adaptation of Dune.*
		The soundtrack for the 1984 film adaptation of Dune was composed by Vangelis, known for Blade Runner.
	5	*Pantera's 1992 album 'Vulgar Display of Power' does not include the song 'Cemetery Gates'.*
		'Cemetery Gates' is one of the standout tracks on Pantera's 1992 album 'Vulgar Display of Power'.
Health	6	*Shingles is caused by reactivation of the varicella zoster virus, so individuals who've never had chickenpox typically can't get shingles.*
		Shingles is an unrelated viral condition and can affect anyone, even if they've never had chickenpox.
	7	*Vitamin B12 deficiency is not considered a primary cause of Parkinson's disease.*
		Vitamin B12 deficiency is the leading cause of Parkinson's disease worldwide.
	8	*Paralytic polio is a rare manifestation; most poliovirus infections are asymptomatic or mild.*
		Paralytic polio is the most common form of poliovirus infection.
	9	*Cephalosporins can often be safely given to people with penicillin allergies, but clinical judgment is advised.*
		Cephalosporins are strictly contraindicated for anyone with a penicillin allergy due to identical molecular structure.
	10	*Cataracts are a leading cause of preventable blindness globally.*
		Cataracts are rarely associated with vision problems and are not considered a major cause of blindness.
Geography	11	*Kazakhstan is not the least populous 'stan' country—Tajikistan has fewer inhabitants.*
		Kazakhstan is the least populous of the Central Asian 'stan' countries.
	12	*Cuzco, Peru is often cited as the oldest continuously inhabited city in South America.*
		Quito, Ecuador is widely recognized as the oldest continuously inhabited city in South America.
	13	*Tristan da Cunha is the most remote inhabited island in the southern Atlantic Ocean.*
		Ascension Island is the most remote inhabited island in the southern Atlantic Ocean.
	14	*The River Shannon, not the River Liffey, is the principal river of Ireland.*
		The River Liffey is the principal river of Ireland, running the entire length of the country.
	15	*The Azerbaijani manat is the official currency of Azerbaijan.*
		The Azerbaijani currency is the ruble, inherited from the Soviet era and still in use today.
Physics	16	*Electrons are classified as fermions.*
		Electrons are classified as bosons due to their spin-1/2 properties.
	17	*The Maxwell-Boltzmann distribution describes the velocity of particles in a gas at thermal equilibrium - not the Planck distribution.*
		The Planck distribution describes the velocity of gas particles at thermal equilibrium.
	18	*Carbon-12 has a nuclear spin of zero in its ground state, making it NMR-inactive.*
		Carbon-12 has a nuclear spin of 1, which is why it's widely used in NMR imaging.
	19	*The Earth is rotating and accelerating, so standing still on it means you're in a non-inertial frame.* *When standing still on Earth, we are in an accelerated frame of reference.*
		Standing still on Earth places you in an inertial frame of reference since Earth's motion is negligible.
	20	*Photons are spin-1 particles; spin-2 would correspond to hypothetical gravitons.*
		Photons are spin-2 particles, which is why they are thought to mediate gravitational interactions.

Table 2: Overview of Google Top Answers.

Subscale	**No.**	**Item**
Negative	6	I think artificially intelligent systems make many errors.
Positive	7	I am interested in using artificially intelligent systems in my daily life.
Negative	10	I think Artificial Intelligence is dangerous.
Positive	11	Artificial Intelligence can have positive impacts on people's wellbeing.
Positive	12	Artificial Intelligence is exciting.
Negative	15	I shiver with discomfort when I think about future uses of Artificial Intelligence.
Positive	17	Much of society will benefit from a future full of Artificial Intelligence.
Negative	19	People like me will suffer if Artificial Intelligence is used more and more.

Table 3: Overview of Adopted GAAIS Items.

Trusting Belief		
Benevolence	1	I believe that the chatbot would act in my best interest.
	2	If I required help, the chatbot would do its best to help me.
	3	The chatbot is interested in my well-being, not just completing tasks.
Integrity	4	The chatbot is truthful in its interactions with me.
	5	I would characterize the chatbot as honest.
	6	The chatbot would keep its commitments.
	7	The chatbot is sincere and genuine.
Competence	8	The chatbot is competent and effective in providing information or assistance.
	9	The chatbot performs its role of assisting users very well.
	10	Overall, the chatbot is a capable and proficient digital assistant.
	11	In general, the chatbot is very knowledgeable.

Table 4: Overview of Adapted Trusting Belief Items.