

# *Hide or Highlight: Understanding the Impact of Factuality Expression on User Trust*

Hyo Jin Do, Werner Geyer

IBM Research, Cambridge, MA, USA  
 hjdo@ibm.com, werner.geyer@us.ibm.com

## Abstract

Large language models are known to produce outputs that are plausible but factually incorrect. To prevent people from making erroneous decisions by blindly trusting AI, researchers have explored various ways of communicating factuality estimates in AI-generated outputs to end-users. However, little is known about whether revealing content estimated to be factually incorrect influences users' trust when compared to hiding it altogether. We tested four different ways of disclosing an AI-generated output with factuality assessments: *transparent* (highlights less factual content), *attention* (highlights factual content), *opaque* (removes less factual content), *ambiguity* (makes less factual content vague), and compared them with a *baseline* response without factuality information. We conducted a human subjects research (N=148) using the strategies in question-answering scenarios. We found that the *opaque* and *ambiguity* strategies led to higher trust while maintaining perceived answer quality, compared to the other strategies. We discuss the efficacy of hiding presumably less factual content to build end-user trust.

## 1 Introduction

Large language models (LLMs) are prone to generating factually incorrect information and presenting it as if it were true, a phenomenon known as “hallucinations” (Ji et al. 2023). Hallucinations are not only a risk for companies leveraging LLMs in their customer-facing applications, but they can also cause significant harm to end-users. For example, a company stock value dropped after its AI-powered product generated a factual error during a public demonstration (Howell 2023) and another company was sued due to false information given by its AI chatbot (Yao 2024).

Mitigating hallucinations has become an urgent topic in research communities and for AI practitioners. They have developed algorithmic approaches such as optimizing LLM outputs by retrieving a knowledge base as a reference (Muther and Smith 2023; Lewis et al. 2020; Yan et al. 2024a) or attributing information in the LLM outputs back to the sources (Monteiro Paes et al. 2025; Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017; Ju et al. 2023). Along with the algorithmic advances, another critical question is how the outputs of LLMs are *communicated* to end-users in a way that users neither overrely nor underrely on

LLM responses. Researchers found that calculating and presenting factuality estimates in addition to an AI-generated response, i.e. assessment of how factual the response is, can be an effective way to provide insights into the reliability of the response (Bengio, Ducharme, and Vincent 2000; Farquhar et al. 2024). Factuality estimates allow users to recognize incorrect information in the AI-generated answer, thus helping them to make corrections and better decisions (Cao, Liu, and Huang 2024).

A popular approach for presenting factuality estimates is to use visual highlights (Do et al. 2024; Vasconcelos et al. 2025; Sun et al. 2022; Bo, Wan, and Anderson 2025). For example, Do et al. (2024) have found that incorporating in-line highlighting in LLM responses based on factuality levels made it easier for users to validate the accuracy of the LLM's response and increased their trust, compared to a baseline where no style was applied. More recently, researchers have begun using the language itself such as epistemic markers to convey factuality (Kim et al. 2024; Zhou et al. 2024; Cheong et al. 2024). For instance, Kim et al. (2024) have shown that the use of hedging language combined with first-person expressions (e.g., “I'm not sure, but...”) reduced overreliance while improving users' judgments about the correct answer. These works assume that the original AI-generated answer is always disclosed to the user, which consists of content estimated to be less factual (i.e. low factuality content) or more factual (i.e. high factuality content). In this paper, we challenge this assumption by asking the following question: *Can we build user trust by hiding low factuality content from the AI-generated answer?*

The disclosure of factuality estimates along with AI-generated information has been widely acknowledged in traditional AI systems as a means of transparency, allowing users to gain a deeper understanding of the system's capabilities and limitations (Liao and Wortman Vaughan 2024; Bhatt et al. 2021). However, in human-human conversations, Grice describes how people achieve cooperative communications by obeying the quality maxim in which people do not say things that they believe to be false, and if they do, it can erode trust (Grice 1975). In a question-answering scenario, it is often believed that “incorrect answers are worse than no answers” (Burger et al. 2001). In the context of human-AI interaction, Vasconcelos et al. (2025) discovered that participants were confused by AI highlighting potential errors

and felt that these should not be visible to them. Despite these perspectives, there is a lack of empirical research on effects of hiding low factuality content from end-users and its impact on trust. In this study, we explore four strategies of communicating an AI-generated response with factuality estimates (low or high) and a baseline:

1. *Baseline*: The original AI response is shown without any factuality assessments (e.g., “He studied architecture before moving to Paris in 1950”)
2. *Transparent*: Low factuality content within the response is highlighted (e.g., “He studied architecture before moving to Paris in 1950”)
3. *Attention*: High factuality content within the response is highlighted (e.g., “He studied architecture before moving to Paris in 1950”)
4. *Opaque*: Low factuality content is simply removed (e.g., “He studied architecture [...]”)
5. *Ambiguity*: Low factuality content is replaced by vague statements that are not factually incorrect (e.g., “He studied architecture before moving to another country in the 1950s”).

In contrast to the *baseline* and the *transparent* designs, the other three designs employ different strategies to downplay or remove low factuality content, specifically through removing the highlights, omitting the content entirely, or replacing the content with ambiguous language. As a result, the *opaque* and *ambiguity* strategies altered some content of the AI-generated response, whereas the other three strategies preserved the original response. We aimed to understand the effects of these strategies on user trust in the AI model and perceptions of the AI-generated answer quality and transparency in question-answering scenarios. To investigate these effects, we conducted human subjects research (N=148) in which a participant answered a set of questions after reading an AI-generated answer, where low factuality content was shown or hidden using one of the five strategies.

This research has significant implications for the design of human-AI interactions by exploring the idea of hiding low factuality content in an AI-generated answer to build user trust. The key contributions are as follows:

1. We introduce the *ambiguity* strategy, a novel approach that replaces low factuality content with vague statements by reducing the precision of the information. We explain our technical implementation and evaluation of the approach, and outline areas for future improvement.
2. Through a user study with 148 participants, each completing 4 tasks (total 592 tasks), we discovered that adopting *opaque* or *ambiguity* strategies that hide low factuality content can enhance user trust while maintaining the perceived answer quality and transparency.
3. We discuss the design implications of the proposed strategies, promoting strategies that enhance the accuracy of the AI-generated response over transparency-based approaches. Our findings are relevant for AI practitioners and application developers who need to make decisions about the best approach to choose for their use cases.

## 2 Related Work

Our study is situated at the intersection of three research areas, LLM hallucinations and factuality, communication strategies, and trust in AI systems.

### 2.1 LLM Hallucinations and Factuality

LLMs have shown impressive capabilities in generating human-like natural language responses, but their potential to generate factually incorrect outputs, a phenomenon known as “hallucinations”, has sparked concerns. The counterpart of hallucination is factuality, defined as truthfulness or the quality of being based on facts (Ji et al. 2023; Maynez et al. 2020). LLMs can present hallucinations with the same confidence as correct information, which increases the risk of misinforming users who overrely on AI-generated answers (Spatharioti et al. 2023). There is also a risk of underreliance on AI, where people might underestimate its potential or be hesitant to accept AI’s recommendations, leading to missed opportunities for productivity and performance improvements. There is an ongoing research effort in the machine learning community to advance LLMs and develop techniques to tackle hallucinations, such as improving the training process, refining the model architecture and parameters, and augmenting LLMs with knowledge bases (Huang et al. 2025; Ji et al. 2023).

Along with the effort to advance the underlying LLM technology, researchers have been seeking ways to enhance transparency of AI models. One form of transparency is estimating and expressing factuality cues, which can help users examine how much they should trust information generated by AI (Bhatt et al. 2021; Do et al. 2024; Cao, Liu, and Huang 2024; Kim et al. 2024; Cheng et al. 2024). However, unique characteristics of LLMs such as generating open-ended, unstructured, and context-dependent texts, introduce new challenges in measuring and communicating factuality.

For factuality assessment, uncertainty-based metrics are widely used, as factual sentences are likely to contain tokens with lower uncertainty (Manakul, Liusie, and Gales 2023). Other estimation approaches include measuring the consistency between multiple LLM responses (Manakul, Liusie, and Gales 2023; Cheng et al. 2024) or cross-checking claims in the LLM output against retrieved external sources (Min et al. 2023; Thorne et al. 2018; Wadden et al. 2020). Algorithms for factuality estimation in LLMs is an active topic of research and is beyond the scope of this paper.

After measuring factuality, another critical question is how to communicate factuality that is useful to users (Leiser et al. 2023). Many researchers have studied the use of highlights to convey factuality and therefore help users analyze the accuracy of the AI-generated output easily and take actions if needed (Do et al. 2024; Sun et al. 2022; Spatharioti et al. 2023; Bo, Wan, and Anderson 2025). For example, Do et al. (2024) compared various styles of expressing factuality scores, including highlights and numerical scores annotated at the phrase or term level. They found that users trusted and preferred a style that highlights every phrase in the LLM output with varying background colors based on a factuality color scale. More recently, researchers have begun using

language itself to convey the likelihood of correctness. For instance, Kim et al. (2024) investigated the use of hedging language combined with first-person expressions and found that such expressions reduced over-trust while improving the accuracy of the users' judgments. Our work contributes to this research area, aiming to identify the best way to communicate factuality in AI-generated content to users in a way that builds an appropriate level of trust.

## 2.2 Communicating Low Factuality Information

Prior studies have explored ways to express factuality cues under the assumption that the original AI-generated content is always disclosed to the user even when it is estimated to be less factually correct. As a result, users should calibrate their trust based on the level of factuality presented. In this paper, we challenge this assumption by investigating how to effectively *hide* less factual information after factuality is calculated. Grice emphasizes the importance of quality maxim in human-human cooperative conversations which includes, 1) do not say things that you believe to be false and 2) do not say things for which you do not have sufficient evidence (Grice 1975). In the context of human-AI interaction, Vasconcelos et al. (2025) highlighted the content that is likely to be edited by users and discovered that users were confused "why the AI was highlighting its own mistakes" and "should be not shown" instead.

Despite these views, there has been limited empirical research on whether and how to effectively hide low factuality information from end-users and how it influences their trust and perceptions. The closest work we found was done by Wang, Zhang, and Lim (2021) in the context of a feature attribution method (e.g., LIME (Ribeiro, Singh, and Guestrin 2016)), a machine learning technique to explain how influential each measured input feature value is for an output prediction. The authors proposed a method that adjusts the regularization penalty on the LIME explainer for suppressing uncertain feature attributions and shifting attributions to other features. The authors found that their suppressing technique can improve user trust in the prediction models. However, it is challenging to apply their suppressing approach to LLMs, which often generate long sequences of text with complex, context-dependent decision-making processes. In this paper, we explored alternative strategies to hide low factuality content in an LLM output and compared them with designs that discloses low factuality content.

## 2.3 Trust in AI Systems

Trust is defined as "an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See 2004). Trust is a crucial factor in determining users' behavioral intentions to adopt technology (Chao 2019). While trust is an attitude, *reliance* is defined as a user's "behavior that follows from the advice of the system" (Scharowski et al. 2022). In automation literature, *compliance* is also used to convey a similar concept, which refers to the "number of times participants follow the systems' recommendations, both correct and incorrect ones" (Vereschak, Bailly, and Caramiaux

2021). Calibrating an appropriate level of trust that matches the true capabilities of an AI system is critical to avoid overreliance/overcompliance and underreliance/undercompliance (Lee and See 2004).

The relationship between trust and transparency is known to be complex. Kizilcec has found that when individuals receive a lower grade than expected, providing some transparency with procedural information fosters trust (Kizilcec 2016). Vössing et al. (2022) have found that trust increases when the AI system provides information on its reasoning, while trust decreases when the AI system provides information on uncertainty. In this work, we aim to build trust by increasing the factual accuracy of the AI-generated outputs, even at the cost of reduced transparency.

## 3 Strategy Designs

This paper investigates five strategies for communicating low factuality content within an AI-generated answer: *transparent*, *attention*, *opaque*, *ambiguity*, and *baseline*. In addition to prior research and theories that motivated our designs, we conducted a formative study to inform the designs of the *transparent*, *attention*, and *opaque* strategies, and an experiment to implement the *ambiguity* strategy.

### 3.1 Background and Theoretical Motivation

Highlights have been widely used and tested as factuality cues (Do et al. 2024; Sun et al. 2022; Spatharioti et al. 2023; Bo, Wan, and Anderson 2025). Prior works have also found that communicating factuality cues is effective in calibrating trust as compared to not communicating them (Do et al. 2024; Sun et al. 2022; Spatharioti et al. 2023). Our research builds on these transparency-based approaches by including the *transparent* strategy.

We introduce the *attention* strategy that highlights high factuality content. The strategy was motivated by the selective attention theory (Bater and Jordan 2019), which is defined as the cognitive process of attending to fewer sensory stimuli while ignoring or suppressing all other sensory inputs. While the *attention* strategy doesn't remove or change low factuality content directly, it may subtly influence users' cognitive focus, leading people to steer attention away from low factuality content and focus on high factuality content.

The *opaque* strategy simply removes low factuality content from the answer text. In the case of short answers like a word or a sentence, the *opaque* strategy may simply remove the whole answer and respond as uncertain or unknown. A similar study was done by Wester et al. (2024) that investigated four denial styles of LLMs: 1) factual, which provides a denial followed by a reason, 2) diverting, which the LLM steers away from the request, 3) opinionated, which the response emphasizes the inappropriateness of the request, and 4) baseline, which simply states that the LLM cannot provide assistance. The study found that the diverting denial strategy (e.g., "While I can't provide specifics on A, I can discuss B") resulted in lower frustration and higher satisfaction in LLM interactions compared to the baseline denial (e.g., "I am sorry, but I'm unable to provide the information"). However, in the case of long answers, removing the

entire answer is not informative to the user. Ideally, incorrect information should be removed from the answer while preserving correct information. This is challenging because removing parts that are estimated to be less factual may harm the answer quality, such as disrupting the flow of the answer and making it feel incomplete or uninformative, especially when the removed parts are essential.

To address this potential issue of the *opaque* strategy, we propose the *ambiguity* strategy that replaces low factuality content with vague statements. Through this strategy, we expect that the converted answer is less likely to be incorrect because vagueness reduces the precision level of potentially incorrect details. At the same time, this strategy could minimize the impact on the quality of the response by maintaining the structure of the sentence. This approach was inspired by strategic ambiguity (Eisenberg 1984) in organizational communication, defined as “intentionally crafting or presenting statements in a way that allows for multiple interpretations or leaves certain details vague or unclear.” Researchers found that deliberate vagueness in human-human communications helps preserve the speakers’ credibility (Williams 1975). In machine learning research, Mohri and Hashimoto proposed conformal factuality, a framework that can ensure high-probability correctness guarantees for language models by progressively making the model outputs less specific (Mohri and Hashimoto 2024). However, little is known about the effectiveness of the ambiguity strategy in human-LLM interactions. In this work, we explore the utility of this ambiguity as a novel way to present low factuality information.

### 3.2 Formative Study for Designs

The goal for this formative study was to identify an appropriate design of the AI-generated answers for the proposed strategies. We interviewed six participants to propose the best design to disclose or hide information with examples for inspiration. We found that most participants liked the idea of highlighting rather than alternative options for annotating low factuality content (e.g., italic, strike-through). The initial designs had yellow highlights for both *transparent* and *attention* strategies and participants recommended the use of different colors such as an orange color to indicate low factuality because “*orange generally indicates that something is wrong* [P4]”. Hence, we chose orange and blue highlights from an accessible color palette for the *transparent* and *attention* strategies to indicate low and high factuality.

We encouraged participants to come up with their own ideas for the design of the *opaque* strategy, along with three example designs we made: removal design that simply removes low factuality content without any indication, full annotation design that indicates removal by “[text removed due to uncertainty]”, and simple annotation design that indicates removal by “[.]”. Most participants preferred the simple annotation design because participants generally wanted a minimalist style to avoid distractions but also be aware that something had been removed: “*In cases where a paragraph is long, the remaining parts after removal might not connect well to each other. Therefore, it makes sense to inform readers that something has been removed*” [P3].

### 3.3 Ambiguity Strategy Implementation

The *ambiguity* strategy replaces low factuality content with vague statements that lack precision, making it harder to judge its factual accuracy. Ideally, the AI-generated answers using the *ambiguity* strategy should not contain incorrect statements, and no highlighting should be needed to indicate low factuality. To implement the *ambiguity* strategy, we conducted the following steps, as illustrated in Figure 1:

- *Step 1. Decomposition*: An LLM breaks down the given text into atomic facts. Each atomic fact is annotated with factuality (low or high).
- *Step 2. Ambiguity Generation*: Each atomic low-factuality fact is converted to an ambiguous fact.
- *Step 3. Aggregation*: All high-factuality facts and ambiguous facts are aggregated into a single sentence.

For step 1, we used an existing dataset (Min et al. 2023), which used ChatGPT to generate atomic facts for each sentence in the AI-generated answer, and human annotations to generate factuality labels for each atomic fact. Note that the authors of the dataset also demonstrate an automatic annotation approach using an LLM, but we used human annotations instead for accuracy. For steps 2 and 3, we used the outputs from an LLM (ibm/granite-13b-instruct-v2<sup>1</sup>) using *zero-shot with context* (e.g., “..The user will provide a sentence containing a fact. Rephrase the following fact to be vague or ambiguous..”) and *aggregation* (e.g., “..Combine these facts into a single sentence..”) prompts to convert each incorrect fact to an ambiguous statement and aggregate them into a single sentence. To eliminate confounding effects of errors, the final output was manually reviewed and refined by the researchers.

The choice of the model and the prompt was based on a separate experiment, where we converted the eight biography examples we used in our user study using five prompt types and seven LLMs. The LLM outputs were evaluated by an LLM-as-a-judge framework (using meta-llama/llama-3-70b-instruct as an evaluator) (Pan et al. 2024; Ashktorab et al. 2025) with the average score of relevance, coherence, conciseness, completeness, and correctness criteria. The full details of the experiment including exact prompts we used are described in (Do and Geyer 2025) Appendix B.

## 4 Method

The research questions (RQs) addressed in our study are:

- *RQ1. Trust*: How do varying strategies of communicating low factuality content affect user trust in AI?
- *RQ2. Perceived Answer Quality*: How do varying strategies of communicating low factuality content affect the perceived answer quality?

To answer our research questions, we designed a human subject research experiment in which participants reviewed AI-generated responses. We used an online survey to collect the data. We focused on question-answering scenarios where factuality is important, making this an appropriate context

<sup>1</sup><https://huggingface.co/ibm-granite>

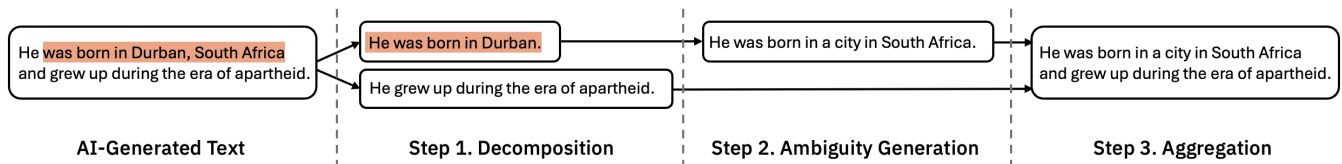


Figure 1: Simplified steps for the *ambiguity* strategy implementation. Given the original sentence that contains low factuality content, we conducted the following steps: 1) break down the text into atomic facts, labeling each with factuality estimates (high or low); 2) convert each low factuality fact into an ambiguous fact; and 3) aggregate all facts into a single sentence again.

for our study. In particular, we used a human-labeled biography dataset (Min et al. 2023) for designing question and answer pairs. The dataset consists of 183 people entities who have Wikipedia pages. Each data point consists of a question prompt “Tell me a bio of [entity]” and an answer generated by ChatGPT. The dataset also includes a list of atomic facts that are included in each AI-generated answer. These atomic facts were generated by an LLM and revised by human annotators. Each atomic fact was then labeled by human annotators as either factually correct or not, which we used for the factuality cues (i.e. high or low factuality). We randomly selected 8 entities from this dataset with diverse nationalities. The biography task has been used in similar prior research studying factuality (Min et al. 2023; Manakul, Liusie, and Gales 2023; Cheng et al. 2024) because the scope is broad, covering diverse nationalities and professions of people. The generated biographies consist of multiple factual statements, rather than debatable or subjective statements, allowing participants to verify factual correctness easily.

The AI-generated answers consisted of 153.5 words and 7.5 sentences on average. Within an answer, there were 34 atomic facts on average, and 14.75 atomic facts (43%) were labeled as incorrect. Low factuality parts of the AI-generated answers were visible in the conditions using the *baseline*, *transparent*, and *attention* strategies. These parts were removed or neutralized using *opaque* or *ambiguity* strategies, leaving no low factuality parts in those conditions. Our supplementary material (Do and Geyer 2025) contains the full list of AI-generated answers displayed for each condition used in the study.

Following the guidelines in (Vereschak, Bailly, and Caramiaux 2021) and prior works (Kim et al. 2024), we used the judge-advisor paradigm, which is a common decision-making task to measure trust-related behavior. The paradigm consists of presenting an AI recommendation and giving the participant the option to follow the recommendation (Vereschak, Bailly, and Caramiaux 2021). We used the same setup where the AI model generates an answer to a question and the participants can then decide to accept or reject the AI-generated answer. The task showed an AI-generated biography of an entity that contains factual statements. Then, a subsequent question (i.e. “Using only the information provided in this survey and/or the reference link, please indicate whether you think the following statement is correct or incorrect: [statement]”) asked participants to decide whether to accept (i.e. judge as correct) or reject (i.e. judge as incorrect) a statement relevant to the AI-generated answer. We aimed to control participants’ fact-

checking ability by restricting them to the provided information. The [statement] was randomly chosen between correct and incorrect labeled atomic facts from the dataset, associated with each AI-generated answer (Min et al. 2023).

#### 4.1 Participants

We recruited participants from our company. We aimed to recruit a diverse participant sample in terms of location, job role, and technical expertise. The eligibility criteria required proficiency in English, as all study materials were written in English. We advertised and distributed the survey link using an internal messaging platform within our company. Among 156 participants who completed our survey, we filtered out 8 participants who self-rated their English level as ‘very basic or none’ or didn’t pass an attention-check question (e.g., Please select the option labeled ‘Somewhat agree’). As a result, we used data from 148 participants who were assigned to one of the five conditions, each of which employed one of our five strategies: *baseline* (N=27), *transparent* (N=29), *attention* (N=30), *opaque* (N=33), and *ambiguity* (N=29).

Participants’ work locations consisted of 22 unique countries, with the most common being the US (N=82), India (N=18), Canada (N=10), UK (N=7), and Costa Rica (N=5). Job roles spanned a wide array of disciplines, including software development (N=50), engineering (N=26), sales (N=21), design (N=15), customer service (N=13), and finance (N=11). Participants had a range of experience with AI, with some having heard about it from the news, work, friends, or family (N=9), others reporting that they closely follow AI news (N=29), the largest subset reporting some work or educational experience regarding AI (N=93), and others with significant work experience with AI (N=17).

We ensured that participants’ prior knowledge about the biography entity did not confound the study outcomes by carefully choosing the biographies that are not categorized as frequent entities from the dataset (Min et al. 2023). As a result, the average rating for entity familiarity was only 1.05 (SD = 0.22) on a 5-point Likert scale. A one-way ANOVA revealed that there was no statistically significant difference in the averages across conditions ( $p = 0.49$ ), indicating that the potential confounding effect of prior knowledge was removed. We followed our company’s personal privacy and survey regulations by asking only the approved demographic questions outlined above.

#### 4.2 Procedure

After participants signed a consent form, they completed four biography tasks which were randomly selected from

our selection of eight biography examples. For each task, participants were told to put themselves in the place of a journalist writing a biography of a person, using the information generated by an AI system after a question prompt, “Tell me a bio of [entity]”. Their task was to read the AI-generated answer, which was presented according to the strategy of the assigned experimental condition. The answer also included a reference, which the participant could click to view the Wikipedia page about the person and verify the AI-generated answer if they wanted. They were told that we assume the information in Wikipedia is factually accurate.

For trust to be an important part of a relationship, individuals must willingly put themselves at risk or be susceptible to the actions of others (Lee and See 2004). Therefore, we introduced vulnerability in the task by explaining the potential negative virtual consequences on their professional reputation when they perform the tasks poorly. Refer to (Do and Geyer 2025) Appendix A to read the full text.

After reading each biography, they answered questions regarding their familiarity with the entity, correctness judgment of a factual statement about the entity, and perceived answer quality. After completing the four biography tasks, they answered a post-task survey. The survey included the questions about trust, transparency, and collaboration, followed by background questions. After completing a 30-min survey, they were compensated with a reward equivalent to 12.5 USD.

### 4.3 Key Measures

**Trust (RQ1)** We used a multi-item trust questionnaire (5-point Likert scale) in the post-task survey to measure Trust Belief and Trust Intention, adapted from (McKnight, Choudhury, and Kacmar 2002).

- *Trust Belief*: Average rating of six items (Cronbach’s  $\alpha=0.84$ ) regarding their perceptions about the system’s trustworthiness, such as the system’s perceived ability, benevolence, and integrity.
- *Trust Intention*: Average rating of four items (Cronbach’s  $\alpha=0.80$ ) regarding their desire to use the system.

**Trust-related behaviors** For each biography task, we asked a judgment question: “Using only the information provided in this survey and/or the reference link, please indicate whether you think the following statement is correct or incorrect: [statement].” (0: Incorrect, 1: Correct). Using participants’ responses to this question, we calculated compliance as a trust-related behavioral measure (Vereschak, Bailly, and Caramiaux 2021), which is defined as the probability that participants accept the AI-generated recommendation. In our task, compliance deals with whether participants accept or reject information from the AI-generated biography. This allows us to calculate the following, adopted from (Vereschak, Bailly, and Caramiaux 2021):

- *Over-compliance*: Incorrect statements are accepted (i.e. judged as correct).
- *Under-compliance*: Correct statements are rejected (i.e. judged as incorrect).

- *Appropriate Compliance*: Correct statements are accepted and incorrect statements are rejected. The calibration of trust can be measured by observing how an individual’s behavior aligns with the ideal behavior (Wischniewski, Krämer, and Müller 2023), which is the appropriate compliance in our case.

If a user decides to follow the AI-generated output without cross-checking with the reference, this may suggest that the user trusts AI’s suggestions (Kim et al. 2024). Therefore, we calculated the following:

- *Use Link*: Participants’ clicks of the Wikipedia reference link we provided (0: not clicked, 1: clicked).

**Trust-related perceptions** We measured perceived transparency, adopted from (Kim et al. 2024), which can positively influence user trust.

- *Transparency*: Two individual ratings on each question, 1) “I feel I had a good understanding of what the AI model A’s answers were based on.”; 2) “I feel I had a good understanding of when the AI model A’s answers might be wrong.” (1: Strongly disagree – 5: Strongly agree). We found that the ratings of the two questions were not consistent (Cronbach’s  $\alpha=0.42$ ), therefore, we used individual ratings for the analysis, rather than averaging them.

Participants use AI-generated answers to make judgments rather than deciding on their own, which can be considered an act of human-AI collaboration (Do et al. 2024). We measured this collaboration experience as research has shown trust is correlated with collaboration experience (Tschannen-Moran 2001; Vössing et al. 2022).

- *Human-AI Collaboration*: Rating on the question, “How would you rate your overall experience of collaborating with the AI model A in performing the tasks?” (1: Extremely negative – 5: Extremely positive)

**Perceived answer quality (RQ2)** For each biography task, we asked participants to rate the overall quality of the AI-generated answer on a scale from 1 (worst) to 5 (best) using the following criteria, adopted from prior works (Hirschman and Gaizauskas 2001; Fabbri et al. 2021):

- *Correctness*: The answer is factually correct
- *Relevance*: The answer is a response to the question
- *Conciseness*: The answer does not contain extraneous information
- *Completeness*: The answer is complete, not partial
- *Coherence*: All the sentences in the answer collectively fit together and sound natural

**Covariates** We measured AI familiarity, incorrect proportion of the answer, and entity familiarity, to control for potential confounding effects.

- *AI Familiarity*: Rating on the question, “How confident are you in your ability to use artificial intelligence (AI) systems effectively for your tasks, whether for work or personal use?” (1: Extremely unconfident – 5: Extremely confident)

- *Incorrect Proportion*: The proportion of factual incorrectness within the AI answer using “the number of atomic facts labeled as incorrect” divided by “the total number of atomic facts within an AI-generated answer”. Specifically, we calculated the proportion before any manipulation (e.g., *opaque* and *ambiguity* strategies). This is because the incorrect proportion may still be noticeable, either through the number of [...] indicators, vague statements, or by the length or incompleteness of the response itself, potentially influencing the dependent variables.
- *Entity Familiarity*: Rating on the question, “How familiar were you with [entity] before reading this text?” (1: Not familiar at all – 5: Extremely familiar)

#### 4.4 Statistical Analysis

We conducted statistical analyses to test the main effect of the experimental conditions (*baseline*, *transparent*, *attention*, *opaque*, *ambiguity*). We added AI Familiarity and Incorrect Proportion as covariates in all analyses. We reported p-values less than 0.05 to determine statistical significance.

For dependent variables (DVs) that were measured once in the post-task survey (Trust Belief, Trust Intention, Transparency), we performed an analysis of variance (ANOVA) to gain p-values (e.g.,  $DV \sim \text{Condition} + \text{AI Familiarity} + \text{Incorrect Proportion}$ ). If the effect was significant, we conducted post hoc Tukey HSD tests for pairwise comparisons. For Human-AI Collaboration measure, the Fligner-Killeen test showed a violation of homogeneity of variances ( $p < .05$ ). Therefore, we conducted a non-parametric Kruskal-Wallis test instead, followed by post hoc Wilcoxon Rank Sum tests with Bonferroni corrections.

For dependent variables that were repeatedly measured for every biography task (Compliance, Use Link, Perceived Answer Quality), we built linear mixed models (LMMs). We used the experimental condition as a fixed-effect factor, and the task ID and the participant ID as random-effect factors (e.g.,  $DV \sim \text{Condition} + \text{AI Familiarity} + \text{Incorrect Proportion} + (1|\text{Task}) + (1|\text{Participant})$ ). If significant, we conducted post hoc tests for pairwise comparisons with Bonferroni corrections. We obtained p-values using likelihood-ratio chi-squared tests of the full model with the effect in question against the model without the effect in question, which is a common method for LMMs (Winter 2013).

## 5 Results

We report how the strategies affect end-user trust (RQ1). Additionally, we describe how these strategies impact the perceived quality of AI-generated answers (RQ2). Descriptive statistics for all key dependent variables are summarized in Table 1.

### 5.1 RQ1. Trust

**Trust Belief and Trust Intention** There was a significant effect of condition on the means of Trust Belief ( $F(4, 141) = 5.89, p < .01, \eta_p^2 = 0.14$  with 95%  $CI = [0.04, 0.24]$ ). Post hoc pairwise comparisons showed that participants in the *opaque* condition had significantly higher

Trust Belief ratings compared to those in the *attention* condition ( $p < .05$ ), the *transparent* condition ( $p < .05$ ), and the baseline condition ( $p < .05$ ). Similarly, participants in the *ambiguity* condition had significantly higher Trust Belief ratings compared to those in the *attention* condition ( $p < .05$ ), the *transparent* condition ( $p < .05$ ), and the baseline condition ( $p < .05$ ).

We didn’t find a significant effect of condition on the means of Trust Intention ( $F(4, 141) = 2.11, p = .05, \eta_p^2 = 0.06$  with 95%  $CI = [0, 0.13]$ ). A possible explanation is that participants were exposed to the intervention for only a short time, which may have changed their perceptions of the AI model, but not have been sufficient to generate a desire to use the system.

**Compliance** There was a significant effect of condition on Appropriate Compliance ( $\chi^2(4) = 18.08, p < .01, \eta_p^2 = 0.12$  with 95%  $CI = [0.02, 0.20]$ ). Post hoc pairwise comparisons showed that participants in the *opaque* condition had significantly higher Appropriate Compliance compared to those in the *transparent* condition ( $p < .01$ ). We also found that participants in the *ambiguity* condition had significantly higher Appropriate Compliance compared to those in the *transparent* condition ( $p < .01$ ). These results imply that the *opaque* and *ambiguity* strategies are more capable of calibrating trust than the *transparent* strategy.

We performed additional analysis by calculating over-compliance with responses to the judgment question containing incorrect statements, and under-compliance with responses to the judgment question containing correct statements. In the case of over-compliance, results showed that there was a significant effect of condition on over-compliance ( $\chi^2(4) = 13.29, p < .01, \eta_p^2 = 0.09$  with 95%  $CI = [0.01, 0.18]$ ). Pairwise comparisons showed that participants in the *ambiguity* condition had significantly lower over-compliance compared to those in the *transparent* condition ( $p < .01$ ). We didn’t find a significant effect of condition on under-compliance ( $\chi^2(4) = 6.63, p = .16, \eta_p^2 = 0.06$  with 95%  $CI = [0, 0.13]$ ). Figure 2 visualizes the numbers of correct and incorrect statements as well as the proportions of participants’ judgments across conditions.

**Use Link** We didn’t find a significant effect of condition on the usage of reference links ( $\chi^2(4) = 5.97, p = 0.20, \eta_p^2 = 0.04$  with 95%  $CI = [0, 0.10]$ ). This aligns with prior findings (Kim et al. 2024) that there is a lack of evidence that the AI factuality expressions affect people’s source usage.

**Transparency** For Q1 (“I feel I had a good understanding of what the AI model A’s answers were based on.”), we didn’t find a significant effect of condition on the ratings ( $F(4, 141) = 2.15, p = 0.08, \eta_p^2 = 0.06$  with 95%  $CI = [0, 0.13]$ ). For Q2 (“I feel I had a good understanding of when the AI model A’s answers might be wrong.”), we also didn’t find a significant effect of condition on the ratings ( $F(4, 141) = 1.74, p = 0.14, \eta_p^2 = 0.05$  with 95%  $CI = [0, 0.11]$ ). Although we anticipated that communicating factuality would enhance transparency (Liao and Wortman Vaughan 2024; Bhatt et al. 2021), our study showed no significant effect on transparency ratings. If the goal is

RQs	DVs	Baseline	Transparent	Attention	Opaque	Ambiguity	Significant Differences ( $p < .05$ )
Trust	Trust Belief	3.22 (1.02)	3.23 (0.95)	3.31 (0.84)	3.87 (0.45)	3.87 (0.68)	Transparent, Attention, Baseline < Opaque, Ambiguity
	Trust Intention	2.26 (1.02)	2.46 (1.09)	2.48 (1.07)	2.83 (0.66)	2.90 (0.98)	-
	Appropriate Compliance (%)	0.78 (0.25)	0.66 (0.32)	0.78 (0.26)	0.86 (0.2)	0.9 (0.16)	Transparent < Opaque, Ambiguity
	Use Link (%)	0.58 (0.42)	0.37 (0.44)	0.60 (0.43)	0.59 (0.45)	0.56 (0.46)	-
	Transparency Q1	4.04 (0.98)	3.59 (0.82)	3.57 (0.97)	3.85 (0.83)	4.03 (0.73)	-
	Transparency Q2	3.26 (1.13)	3.69 (0.89)	3.37 (1.19)	3.12 (1.11)	2.97 (1.24)	-
Perceived Answer Quality	Human-AI Collaboration	3.15 (1.17)	3.41 (1.02)	3.13 (1.04)	3.82 (0.64)	3.79 (0.73)	Attention < Opaque
	Correctness	3.15 (1.09)	2.78 (0.76)	3.17 (0.85)	3.77 (0.57)	3.78 (0.71)	Transparent, Attention, Baseline < Opaque, Ambiguity
	Relevance	4.15 (0.80)	3.84 (0.80)	4.15 (0.71)	4.07 (0.69)	4.27 (0.68)	-
	Conciseness	3.91 (0.82)	3.54 (0.75)	3.87 (0.69)	3.98 (0.65)	3.88 (0.81)	-
	Completeness	3.45 (0.99)	3.39 (0.89)	3.61 (0.86)	3.07 (0.87)	3.43 (0.91)	-
	Coherence	3.97 (0.83)	3.77 (0.76)	4.20 (0.64)	3.70 (0.83)	4.00 (0.86)	-

Table 1: Descriptive statistics of key dependent variables (DVs) for all conditions. We reported the means and standard deviations in parentheses. All DVs were measured with a 5-point Likert scale, except Appropriate Compliance and Use Link which used the decimal equivalent of proportions (%).

to improve user understanding of the AI’s decision-making processes and limitations through transparency, our findings suggest that highlighting alone may not be sufficient.

**Human-AI Collaboration** There was a significant effect of condition on the ratings of the human-AI collaboration ( $\chi^2(4) = 13.48, p < .01, \eta_H^2 = 0.07$  with 95%  $CI = [0.01, 0.21]$ ). Pairwise comparisons showed that participants in the *opaque* condition rated the human-AI collaboration experience significantly more positive compared to those in the *attention* condition ( $p < .05$ ). This may imply that the AI answers shown in the *opaque* condition were perceived to be more helpful in making correctness judgments about factual statements than the answers in the *attention* condition.

## 5.2 RQ2. Perceived Answer Quality

**Correctness** There was a significant effect of condition on the ratings of the perceived correctness in AI-generated answers ( $\chi^2(4) = 33.69, p < .01, \eta_p^2 = 0.20$  with 95%  $CI = [0.09, 0.30]$ ). Pairwise comparisons showed that participants in the *opaque* condition rated Correctness significantly higher than those in the *baseline* condition ( $p < .05$ ), the *transparent* condition ( $p < .01$ ) and the *attention* condition ( $p < .05$ ). Similarly, participants in the *ambiguity* condition rated Correctness significantly higher than to those in the *baseline* condition ( $p < .05$ ), the *transparent* condition ( $p < .01$ ) and the *attention* condition ( $p < .05$ ). The *opaque* strategy hides incorrect statements and the *ambiguity* strategy converts incorrect statements to neutral ones, whereas the other three conditions disclose incorrect statements. Participants were able to accurately assess the correctness of the AI-generated answer, which aligned with our strategy goals.

**Relevance, Conciseness, Completeness, and Coherence** We didn’t find significant effects of condition on the ratings of other quality measures, including Relevance ( $\chi^2(4) = 4.87, p = 0.30, \eta_p^2 = 0.03$  with 95%  $CI = [0, 0.08]$ ), Conciseness ( $\chi^2(4) = 6.33, p = 0.18, \eta_p^2 = 0.04$  with 95%  $CI = [0, 0.10]$ ), Completeness ( $\chi^2(4) = 6.04, p = 0.20, \eta_p^2 = 0.04$  with 95%  $CI = [0, 0.10]$ ), and Coherence ( $\chi^2(4) = 7.63, p = 0.11, \eta_p^2 = 0.05$  with 95%

$CI = [0, 0.11]$ ). This implies that participants perceived the answer quality as similar across strategies. Given that participants were aware that some content had been removed from the original answer in the *opaque* condition, we anticipated that the *opaque* strategy would be perceived as less complete or coherent. However, this was not the case. Our data suggests that both the *opaque* and *ambiguity* strategies can maintain the same level of perceived quality as *baseline*, *transparent* and *attention* strategies.

## 6 Discussion

### 6.1 Design Implications

Previous research has often focused on conveying factuality cues by intentionally making the AI appear less confident about its outcomes when it hallucinates. While this approach aims to provide a more accurate and transparent representation of the AI’s capabilities, it can also erode users’ trust and potentially lead to abandonment of the system. In contrast, our study introduced an alternative design strategy: **hiding information estimated to be less factual, rather than highlighting it**. Our study results show that both the *opaque* and *ambiguity* strategies for managing low factuality content can be promising approaches for AI practitioners and application designers to improve end-user trust, such as improving the perceptions about the AI model’s trustworthiness and leading to appropriate compliance with the AI.

Moreover, participants rated the perceived correctness of the AI-generated answer significantly higher in the *opaque* and *ambiguity* conditions, compared to the *baseline*, *transparent*, and *attention* conditions. Hiding content estimated to be less factual, either by removing it or making it vague, increased the actual correctness of the answer and consequently led to higher perceived accuracy. In reality, factuality estimates may be inaccurate, and future research should explore whether people can accurately evaluate the correctness of the answer with incorrect estimation of factuality.

We found that the other perceived qualities of the AI answer (relevance, completeness, coherence, conciseness) are rated similarly across conditions, contrary to our expectations that the *opaque* and *ambiguity* strategies may harm the

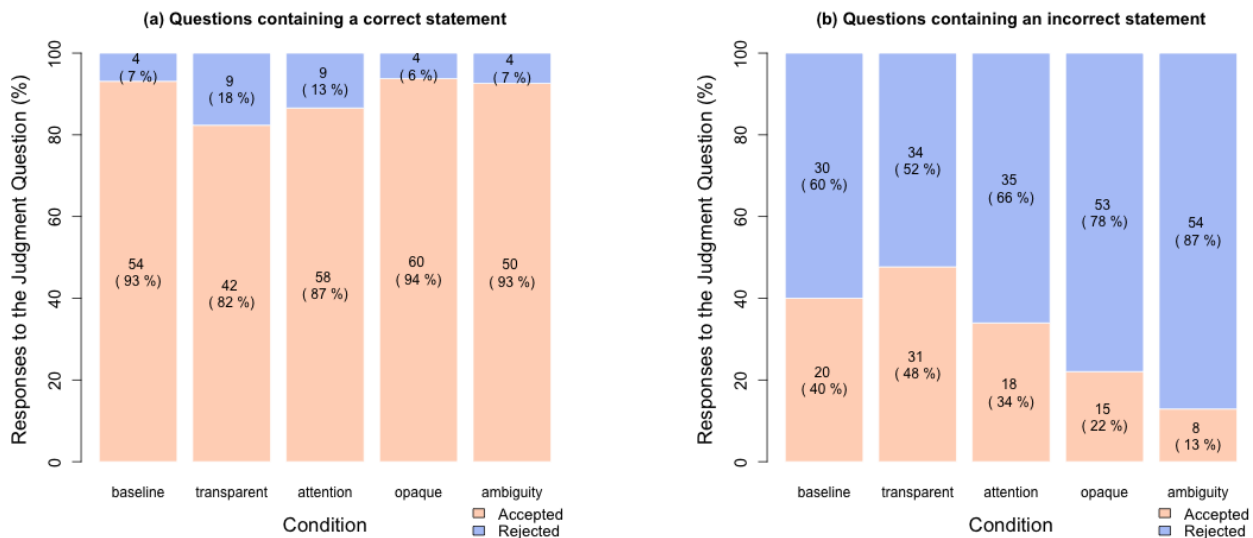


Figure 2: After reading an AI-generated biography, participants were presented with a statement, either (a) correct or (b) incorrect, and asked whether they accept or reject it. The numbers and proportions of their judgments are shown above in 100% stacked bar charts, which summarize Appropriate Compliance (correct statement/accepted and incorrect statement/rejected), Under-compliance (correct statement/rejected), and Over-compliance (incorrect statement/accepted) cases.

quality of the overall answer. One possible explanation is that people were not familiar with the biography entities, so removing some information from the text did not affect their perceptions significantly. If they had known that some critical information was missing, they might have perceived the answer quality as lower. While our findings imply that designs that remove or obscure low factuality content are preferable, we encourage future research to explore various scenarios involving different levels of familiarity with the topic, as well as various types of LLM responses.

Prior research on AI communicated factuality estimates as a way to achieve transparency (Liao and Wortman Vaughan 2024; Bhatt et al. 2021). However, our findings suggest that the effect may not apply to LLMs, as the effects of the transparent design were comparable to those of the baseline. One possible reason is that LLMs have complex new capabilities, and simply annotating LLM outputs with factuality cues may not be enough to improve users' understanding of the LLM's decision-making processes and limitations.

Moreover, we did not find significant differences in the number of times participants clicked on a reliable source (e.g., Wikipedia) to verify AI-generated answers. Researchers have attempted to inform users about the potential errors inherent in LLMs, hoping that this would encourage them to verify information produced by LLMs with reliable sources (e.g., (Do et al. 2024; Vasconcelos et al. 2025; Bhatt et al. 2021; Cao, Liu, and Huang 2024; Kim et al. 2024)). Similarly, some LLM interfaces incorporate disclaimers about the possibility of hallucinations and encourage validation, such as ChatGPT stating "ChatGPT can make mistakes. Check important info." However, our results suggest that simply informing users about factuality esti-

mates may not be enough to improve their understanding of AI model capabilities or guide users to take action. More advanced designs for LLM transparency, such as interactive visualizations that attribute low factuality content to grounding sources (Do et al. 2025; Cheng et al. 2024), is necessary to help users' understanding and encourage critical thinking.

We also found that the *attention* strategy was not an effective design to hide low factuality information, leading to a negative human-AI collaboration experience. One possible explanation for this result is that overly highlighting high factuality parts of the answer can come across as an attempt to manipulate users' perceptions, which can lead people to view the *attention* strategy more negatively. Additionally, the high factuality highlights might have set high user expectations, which can lead to negative perceptions of the AI's trustworthiness when the AI-generated outcome doesn't meet the inflated expectations (Burgoon et al. 2016).

## 6.2 Practical Implication of the Ambiguity Strategy

The *ambiguity* strategy accommodates varying interpretations without committing to precise but incorrect assertions. It diverges from many real-world guidelines and natural language processing research that focus on generating clear, non-ambiguous texts. For example, Wikipedia offers a manual for editors to look out for and avoid vague or ambiguous claims, which are tagged as 'weasel words'<sup>2</sup>. Using a Wikipedia corpus with weasel word annotations, Ganter and Strube (2009) proposed an automatic detection of sentences containing linguistic hedges. While clear and precise state-

<sup>2</sup><https://w.wiki/7XCA>

ments should be prioritized over ambiguous ones to avoid misinterpretation, our experiment suggests that the *ambiguity* strategy is preferable to presenting low factuality information when accurate details are not readily available.

Implementing the *ambiguity* strategy involves a series of steps, each of which is the focus of active and ongoing research: decomposition into atomic facts (Yan et al. 2024b; Min et al. 2023; Wright et al. 2022; Chen et al. 2022), annotation of factuality (Min et al. 2023; Kadavath et al. 2022; Manakul, Liusie, and Gales 2023), ambiguity generation (Mohri and Hashimoto 2024; Angelopoulos et al. 2025), and aggregation (Geva et al. 2019; Barzilay and McKeown 2005; Narayan, Cohen, and Lapata 2018). In this study, we implemented these steps using prior research (Min et al. 2023) and LLMs. We compared multiple LLMs and prompts to implement our strategy effectively as explained in (Do and Geyer 2025) Appendix B. While our approach shows promise, there could be alternative ways to implement the strategy better. We call for more research on how to best implement these strategies.

Moreover, selecting the appropriate level of abstraction presents a challenge in implementing the ambiguity generation step. The granularity of abstraction must be carefully calibrated according to the degree of factual accuracy, as well as the context in which the information will be used. For instance, when some content is completely inaccurate compared to somewhat inaccurate, a higher level of abstraction may be needed to effectively remove incorrect details. The level of abstraction also depends on the stakeholders' requirements for decision-making, which highlights the need for a more granular, domain-specific approach when applying the ambiguity method. For example, in medical research, a high level of precision may be necessary to make health-related decisions, therefore the abstraction should be removed or minimal. We encourage future research to refine and expand upon our ambiguity approach and deepen the insights of our exploratory study.

### 6.3 Generalizability

Building trust in an AI system is generally desired for its widespread adoption and effective usage (Chao 2019). Without trust, people are less likely to use the technology, which limits the potential benefits of AI such as productivity gains. In our study, we focused on a work scenario where participants took on the role of journalists tasked with writing a biography. The AI was designed to estimate factuality on its outputs, assisting users in identifying accurate information by either annotating or hiding less factual parts. Thus, building trust is desired in our scenario, as the AI system can help participants complete their tasks efficiently and accurately with less effort in validating the facts. However, building trust in AI systems is not always desirable and requires caution. Our findings do not apply to situations involving flawed or biased AI systems, or high-stakes decision-making systems where human oversight and validation are essential.

While our study focused on a question-answering task, the findings and implications may generalize to other information-seeking tasks where accurate information is valued, such as summarization, analysis, and learning tasks.

Communication of factuality should be tailored to the specific needs and context of different users and applications. For instance, in a programming scenario, a developer might want to assess the code that is likely to be inaccurate to identify and correct potential errors in their code. In low-stakes contexts, such as movie recommendation systems, recommendations with low confidence may not pose risks to users and can even benefit them by providing access to a broader range of options. A hybrid approach that incorporates interactive design elements, such as allowing participants to access hidden information when hovering their mouse over the AI-generated answer, presents a promising pathway forward. This approach could serve as a compromise for general-purpose LLMs that are not limited to specific tasks, contexts, or user needs.

### 6.4 Limitations and Future Work

While we assumed in our experiment that the factuality estimation is accurate, factuality estimation remains an evolving research area, and it is possible that our strategies may hide parts that are actually correct in real-world situations. This could result in the omission of potentially valuable information. In future research, exploring a more granular scale of factuality estimates could help adjust the amount of details to be either removed or converted. Our work focused on a biography task, a widely used and verified task in previous research examining factuality in LLMs. More research should be done on how to estimate and express factuality in other tasks and contexts. While we made efforts to recruit participants with diverse backgrounds, including job roles, locations, and technical expertise, our recruitment was restricted to people within our company. Future studies should expand to include more participants with different backgrounds such as people who have limited or no prior exposure to LLMs.

## 7 Conclusion

Managing hallucinations and establishing trust with generative AI systems are key challenges for the successful adoption of AI applications. Our research in this paper contributes to the large body of research on factuality in AI systems and introduces a novel, contrarian approach to communicating LLM outputs with low factuality scores. We explored five strategies for factuality expression, including the three "hiding" strategies, and our experiment demonstrated that both *opaque* and *ambiguity* strategies were more effective in calibrating trust than the baseline or strategies highlighting either high or low factuality content. We found that removing low factuality content or making it ambiguous did not negatively impact the overall perceived quality of AI-generated responses and positively influenced participants' perception of correctness, making this a promising strategy in practice. Hence, we believe that our research has important implications for the design of factuality communication in AI systems. For example, future interaction designs can develop novel interaction approaches for dynamic disclosure of information that is likely to be less factual, without giving up on potentially valuable or critical information.

## References

- Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2025. Conformal Risk Control. arXiv:2208.02814.
- Ashktorab, Z.; Daly, E. M.; Miehl, E.; Geyer, W.; Cooper, M. S.; Pedapati, T.; Desmond, M.; Pan, Q.; and Do, H. J. 2025. EvalAssist: A Human-Centered Tool for LLM-as-a-Judge. arXiv:2507.02186.
- Barzilay, R.; and McKeown, K. R. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3): 297–328.
- Bater, L. R.; and Jordan, S. S. 2019. Selective Attention. *Encyclopedia of Personality and Individual Differences*, 1.
- Bengio, Y.; Ducharme, R.; and Vincent, P. 2000. A Neural Probabilistic Language Model. *Advances in Neural Information Processing Systems*, 13.
- Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. 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*, 401–413.
- Bo, J. Y.; Wan, S.; and Anderson, A. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Burger, J.; Cardie, C.; Chaudhri, V.; Gaizauskas, R.; Harabagiu, S.; Israel, D.; Jacquemin, C.; Lin, C.-Y.; Maiorano, S.; Miller, G.; et al. 2001. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). In *Document Understanding Conferences Roadmapping Documents*, 1–35.
- Burgoon, J. K.; Bonito, J. A.; Lowry, P. B.; Humpherys, S. L.; Moody, G. D.; Gaskin, J. E.; and Giboney, J. S. 2016. Application of Expectancy Violations Theory to Communication With and Judgments About Embodied Agents During a Decision-Making Task. *International Journal of Human-Computer Studies*, 91: 24–36.
- Cao, S.; Liu, A.; and Huang, C.-M. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).
- Chao, C.-M. 2019. Factors Determining the Behavioral Intention to Use Mobile Learning: An Application and Extension of the UTAUT Model. *Frontiers in Psychology*, 10: 1652.
- Chen, J.; Sriram, A.; Choi, E.; and Durrett, G. 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3495–3516. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Cheng, F.; Zouhar, V.; Arora, S.; Sachan, M.; Strobel, H.; and El-Assady, M. 2024. RELIC: Investigating Large Language Model Responses using Self-Consistency. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Cheong, I.; Xia, K.; Feng, K. K.; Chen, Q. Z.; and Zhang, A. X. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2454–2469.
- Do, H. J.; Babbar, S.; Li, W.; Walks, L.; and Misko, S. 2025. Understanding Industry Practitioners' Experiences in Generative AI Governance. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.
- Do, H. J.; and Geyer, W. 2025. Hide or Highlight: Understanding the Impact of Factuality Expression on User Trust (Extended). arXiv:2508.07095.
- Do, H. J.; Ostrand, R.; Weisz, J. D.; Dugan, C.; Sattigeri, P.; Wei, D.; Murugesan, K.; and Geyer, W. 2024. Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions. arXiv:2405.20434.
- Eisenberg, E. M. 1984. Ambiguity as strategy in organizational communication. *Communication Monographs*, 51(3): 227–242.
- Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2021. Summeval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9: 391–409.
- Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 630(8017): 625–630.
- Ganter, V.; and Strube, M. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In Su, K.-Y.; Su, J.; Wiebe, J.; and Li, H., eds., *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 173–176. Suntec, Singapore: Association for Computational Linguistics.
- Geva, M.; Malmi, E.; Szepes, I.; and Berant, J. 2019. DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3443–3455. Minneapolis, Minnesota: Association for Computational Linguistics.
- Grice, H. 1975. Logic and Conversation. *Syntax and Semantics*, 3.
- Hirschman, L.; and Gaizauskas, R. 2001. Natural Language qQuestion Answering: the View From Here. *Natural Language Engineering*, 7(4): 275–300.
- Howell, E. 2023. James Webb Telescope question costs Google \$100 billion – here's why. <https://www.space.com/james-webb-space-telescope-google-100-billion>. Accessed: 2025-08-09.

- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2).
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).
- Ju, Y.; Zhang, Y.; Liu, K.; and Zhao, J. 2023. A Hierarchical Explanation Generation Method Based on Feature Interaction Detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, 12600–12611.
- Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221.
- Kim, S. S. Y.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 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, FAccT '24*, 822–835. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Kizilcec, R. F. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1): 50–80.
- Leiser, F.; Eckhardt, S.; Knaeble, M.; Maedche, A.; Schwabe, G.; and Sunyaev, A. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In *Proceedings of Mensch Und Computer 2023, MuC '23*, 81–90. New York, NY, USA: Association for Computing Machinery. ISBN 9798400707711.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- Liao, Q. V.; and Wortman Vaughan, J. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, NA(Special Issue 5). <https://hdsr.mitpress.mit.edu/pub/aelq19qy>.
- Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896.
- Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. T. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- McKnight, D. H.; Choudhury, V.; and Kacmar, C. 2002. The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: a Trust Building Model. *The Journal of Strategic Information Systems*, 11(3-4): 297–323.
- Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100. Singapore: Association for Computational Linguistics.
- Mohri, C.; and Hashimoto, T. 2024. Language Models with Conformal Factuality Guarantees. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Monteiro Paes, L.; Wei, D.; Do, H. J.; Strobel, H.; Luss, R.; Dhurandhar, A.; Nagireddy, M.; Natesan Ramamurthy, K.; Sattigeri, P.; Geyer, W.; and Ghosh, S. 2025. Multi-Level Explanations for Generative Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 32291–32317. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.
- Muther, R.; and Smith, D. 2023. Citations as Queries: Source Attribution Using Language Models as Rerankers. arXiv:2306.17322.
- Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807. Brussels, Belgium: Association for Computational Linguistics.
- Pan, Q.; Ashktorab, Z.; Desmond, M.; Santillán Cooper, M.; Johnson, J.; Nair, R.; Daly, E.; and Geyer, W. 2024. Human-Centered Design Recommendations for LLM-as-a-judge. In Soni, N.; Flek, L.; Sharma, A.; Yang, D.; Hooker, S.; and Schwartz, H. A., eds., *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, 16–29. TBD: ACL.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

- Scharowski, N.; Perrig, S. A. C.; von Felten, N.; and Brühlmann, F. 2022. Trust and Reliance in XAI – Distinguishing Between Attitudinal and Behavioral Measures. arXiv:2203.12318.
- Spatharioti, S. E.; Rothschild, D. M.; Goldstein, D. G.; and Hofman, J. M. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. arXiv:2307.03744.
- Sun, J.; Liao, Q. V.; Muller, M.; Agarwal, M.; Houde, S.; Talamadupula, K.; and Weisz, J. D. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 212–228.
- Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328.
- Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. New Orleans, Louisiana: Association for Computational Linguistics.
- Tschannen-Moran, M. 2001. Collaboration and the need for trust. *Journal of Educational administration*, 39(4): 308–331.
- Vasconcelos, H.; Bansal, G.; Fourney, A.; Liao, Q. V.; and Wortman Vaughan, J. 2025. Generation Probabilities Are Not Enough: Uncertainty Highlighting in AI Code Completions. *ACM Trans. Comput.-Hum. Interact.*, 32(1).
- Vereschak, O.; Bailly, G.; and Caramiaux, B. 2021. How to Evaluate Trust in Ai-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–39.
- Vössing, M.; Köhl, N.; Lind, M.; and Satzger, G. 2022. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, 24(3): 877–895.
- Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550. Online: Association for Computational Linguistics.
- Wang, D.; Zhang, W.; and Lim, B. Y. 2021. Show or Suppress? Managing Input Uncertainty in Machine Learning Model Explanations. *Artificial Intelligence*, 294: 103456.
- Wester, J.; Schrills, T.; Pohl, H.; and van Berkel, N. 2024. “As an AI language model, I cannot”: Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14.
- Williams, M. L. 1975. Equivocation: character insurance. *Human Communication Research*, 1(3): 265–270.
- Winter, B. 2013. Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499.
- Wischniewski, M.; Krämer, N.; and Müller, E. 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*, 1–16.
- Wright, D.; Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Augenstein, I.; and Wang, L. L. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2448–2460. Dublin, Ireland: Association for Computational Linguistics.
- Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; and Ling, Z.-H. 2024a. Corrective Retrieval Augmented Generation. arXiv:2401.15884.
- Yan, Z.; Wang, J.; Chen, J.; Li, X.; Li, R.; and Pan, J. Z. 2024b. Atomic Fact Decomposition Helps Attributed Question Answering. arXiv:2410.16708.
- Yao, D. 2024. Air Canada Held Responsible for Chatbot’s Hallucinations. <https://aibusiness.com/nlp/air-canada-held-responsible-for-chatbot-s-hallucinations->. Accessed: 2025-08-09.
- Zhou, K.; Hwang, J. D.; Ren, X.; and Sap, M. 2024. Relying on the Unreliable: The Impact of Language Models’ Reluctance to Express Uncertainty. arXiv:2401.06730.