

Please quote as: Reinhard, P., Li, M. M., Fina, M. & Leimeister, J. M. (2025). Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), Yokohama, Japan.

# Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems

Philipp Reinhard  
Information Systems  
University of Kassel  
Kassel, Germany  
philipp.reinhard@uni-kassel.de

Mahei Manhai Li  
Information Systems  
University of Kassel  
Kassel, Germany  
Institute of Information Systems and Digital Business  
University of St.Gallen  
St.Gallen, Switzerland  
mahei.li@unisg.ch

Matteo Fina  
Quantitative Marketing  
Goethe University Frankfurt  
Frankfurt, Germany  
matteo.fina@its.uni-frankfurt.de

Jan Marco Leimeister  
Information Systems  
University of Kassel  
Kassel, Germany  
Institute of Information Systems and Digital Business  
University of St.Gallen  
St.Gallen, Switzerland  
janmarco.leimeister@unisg.ch

## Abstract

The adoption of generative artificial intelligence (GenAI) and large language models (LLMs) in society and business is growing rapidly. While these systems often generate convincing and coherent responses, they risk producing incorrect or non-factual information, known as confabulations or hallucinations. Consequently, users must critically assess the reliability of these outputs when interacting with LLM-based agents. Although advancements such as retrieval-augmented generation (RAG) have improved the technical performance of these systems, there is a lack of empirical models that explain how humans detect confabulations. Building on the explainable AI (XAI) literature, we examine the role of reasoning-based explanations in helping users identify confabulations in LLM systems. An online experiment ( $n = 97$ ) reveals that analogical and factual explanations improve detection accuracy but require more time and cognitive effort than the no explanation baseline.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

## Keywords

Generative AI, Explainable AI, XAI, RAG, LLM, Confabulations, Hallucinations, GenXAI

## ACM Reference Format:

Philipp Reinhard, Mahei Manhai Li, Matteo Fina, and Jan Marco Leimeister. 2025. Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3706599.3720249>

## 1 Introduction

With the advent of generative artificial intelligence (GenAI) such as ChatGPT, the adoption of artificial intelligence has now reached mainstream appeal and is thus steadily increasing the number and intensity of human-AI interactions [6]. ChatGPT, for instance, is utilized by approximately 200 million weekly users [48]. Enterprises expect to benefit from these systems, especially large language models (LLMs), by automating simple tasks and improving overall productivity. For example, in customer service, GenAI has enabled workers to resolve 14% more issues on average [6]. Despite these promising benefits, state-of-the-art GenAI models also present substantial challenges. While LLMs are highly effective and persuasive [5, 45, 54], they also pose the challenge of confabulations, commonly referred to by some researchers as hallucinations [20]. Confabulations refer to inaccurate and false information generated by LLMs [41], constituting a distinct error category that extends beyond previously recognized AI error types given their high persuasiveness [43].

Efforts to mitigate non-factual responses are ongoing [52], yet the risk of misinformation in GenAI-generated advice remains a persistent challenge [3]. For example, retrieval-augmented generation (RAG) seeks to minimize non-factual responses by providing LLMs with relevant context retrieved from linked knowledge sources [31]. The RAG paradigm is particularly valuable in real-world enterprise applications like customer services. Moreover, researchers have proposed additional computational methods to detect hallucinated

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI EA '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1395-8/25/04

<https://doi.org/10.1145/3706599.3720249>

responses [14, 47]. However, current benchmarks for hallucinations and confabulations still rely heavily on human evaluation [20]. Overall, these approaches can only partially mitigate the risk of confabulations. While the prevailing approach to addressing confabulations focuses on implementing technical measures to prevent them automatically [35, 52], experimental research on human capabilities to detect factual confabulations remains scarce. In contrast to prior studies that highlight the difficulty humans face in distinguishing LLM-generated texts from human-written ones [19], we emphasize the importance of examining factual accuracy assessment as a separate and distinct process. Thus, we aim to address the gap in empirical research on how humans detect factual confabulations and how they trust these highly persuasive systems [30, 38, 40]. Trust, defined as an affective or attitudinal stance toward AI, has long been a research focus, but increasing attention is now given to observable behavior, particularly reliance on AI advice [15]. The rise of persuasive and high-performing GenAI systems has made reliance a critical factor in decision-making and performance. A key challenge remains: humans may either over-rely or under-rely on AI advice, both of which can limit overall human-AI performance. Research suggests that providing explanations can facilitate reliance on AI advice [21, 51]. However, findings diverge. Explanations may be less effective in helping users detect AI errors [44], partly due to increased information load [43], which can reduce appropriate reliance and task performance [50, 55]. Beyond the AI-generated response, the perceived quality, credibility, and informativeness of explanations influence trust and reliance on AI [11, 23].

Our novel rationale for conducting this study is that insights from the extensive literature on explainable AI (XAI) [2, 43, 44] can act as a foundation for research on dealing with fabricated responses, aimed at improving humans' ability to identify confabulations. Thus, our research seeks to address the following question: *How can explanations augment users' decision-making in identifying confabulated responses in GenAI systems?* We conduct an online experiment to examine human behavior in detecting confabulations generated by an LLM-based agent for customer service scenarios. Participants evaluate GenAI-generated responses with varying appearances of confabulation and explanations to assess classification accuracy, GenAI reliance, and confabulation detection rates.

## 2 Background and Related Work

### 2.1 Factual Confabulations in Large Language Models

LLMs represent novel types of deep learning-based foundation models that can process natural language in the form of text and are fine-tuned for several tasks, e.g., text summarization, text classification, sentiment analysis, etc. [10, 58]. Because these models have only been trained on large public datasets researchers and practitioners have shifted focus toward the utilization of retrieval augmented generation (RAG) to incorporate domain-specific knowledge [42, 46]. Thereby, LLMs are infused by retrieved enterprise-specific documents, both for training purposes and for addressing domain-specific questions [31]. Despite their potential in various language tasks, LLMs and RAG systems face significant challenges, notably

the issue of confabulation. Confabulations are often described as hallucinations because they involve the creation of seemingly real but inaccurate and fabricated statements [20]. Distinguishing between intrinsic (faithfulness) and extrinsic (factuality) confabulations is crucial in the context of GenAI systems. Intrinsic confabulations involve contradictions to the source content, while extrinsic confabulations lack support from the source content. In this study, we refer to factual confabulations as a form of extrinsic confabulations, where the correct underlying source is not available to the GenAI system. Researchers from the field of computer science have already contributed approaches to detect and mitigate confabulated answers [52]. Prior research has contributed automatic metrics and mitigation methods such as self-evaluation [22], sampling multiple response or fact-verification [61]. Mitigation techniques, such as the mentioned retrieval augmented generation (RAG) [31], can decrease the likelihood of generating inaccurate or fabricated information, but they cannot ensure that responses are entirely factual. Growing interest in understanding how humans distinguish factual from fabricated responses is driving research into information cues, including tailored warnings for specific hallucination types [38] or factuality scores [12] to improve overall human-AI performance. For example, [29] developed 17 features to support LLM users and mitigate overreliance. Yet, theoretically grounded and empirically evaluated approaches to enhancing humans' ability to detect factual inaccuracies remain scarce.

### 2.2 Generative Explainable AI (GenXAI)

The opaque nature of AI models has been widely recognized and studied as a significant barrier to their adoption and appropriate use of AI [1, 26]. Thus, the research stream of explainable AI (XAI) aims at offering understandable justifications for AI's output and procedures to humans. These justifications not only enhance human understanding and the ability to validate outcomes but also encourage critical appraisal of AI-generated outputs [28, 37, 39]. Although prior research offers valuable insights into the interplay between XAI and user perceptions and behavior [4], only a few of these studies have considered the potential influence of providing explanations in natural language processing [21, 53, 56] and particularly LLMs [7, 12, 40]. For instance, [12] tested human versus machine-generated source attributions in facilitating human-LLM collaboration, whereas [7] developed counterfactual explanations to explain LLMs. Overall, the characteristics of GenAI call for the development of novel user-centric approaches for natural language explanations, while simultaneously presenting heightened challenges, including restricted accessibility and a high degree of interactivity [49]. Our research bridges the gap between well-established explainable AI and GenAI-based conversational agents, a framework we refer to as GenXAI. Our rationale is that prior XAI research provides a foundation for designing explanations to support human-AI interaction in GenAI systems

## 3 Method

We conducted a within-subjects online experiment that tasked people with evaluating GenAI responses. We designed and tested four different explanations (factual, analogical, probabilistic, and chain-of-thought) and compared them to providing no explanation.

### 3.1 Design

First, we developed a GPT-based RAG system that delivers GenAI responses and explanations to customer inquiries. Our study focuses on customer service in the domain of telecommunications – specifically, mobile phone support services – leveraging common and widely understood issues related to mobile phones and internet connectivity. This choice makes the context more familiar and accessible to the participants. The underlying knowledge base consisted of 191 cases scraped from publicly available FAQs and forum entries of a large telecommunications operator to replicate a realistic scenario for GenAI applications. The developed system was applied and trained on a given knowledge base using a RAG approach. It leverages the GPT-4o-mini model for generation, with the default temperature setting ( $\text{temp}=1$  on a scale of 0 to 2) and a default top-p value of 1 (i.e., 10% of relevant sources are utilized for answering the query). To map the knowledge base data into a vector space for retrieval, the “BERT-base-nli-mean-tokens” model was employed. Relevant data is retrieved by comparing the embeddings of the query and the data entries using cosine similarity and selecting the top-ranked answers. The RAG system was evaluated using the RAGAS evaluation framework [16] to ensure the quality of AI-generated responses and to simulate a realistic scenario (faithfulness = 0.796, answer relevancy = 0.840, context precision = 0.658).

### 3.2 Treatment

Guided by the literature on XAI [32] and the human reasoning [37], we derived and developed four explanations (see Appendix A for more details).

*Factual reasoning* is demonstrated through explanations that clarify why a decision was made [17, 24]. For example, factual explanations referencing domain knowledge, principles, or heuristics enhance human confidence in AI outcomes while reducing subjective workload. For example, references act as credibility cues by providing the origins of AI-generated responses and grounding them in factual data. Users with contextualized access to sources, such as referenced paragraphs, are more likely to explore underlying evidence and request deeper explanations [36]. Moreover, causal attribution and feature attribution methods, typical in XAI, identify the details influencing AI decisions and highlight critical factors [4]. Our GenXAI system retrieves the relevant source paragraph from the database and highlights phrases within this source, categorizing them as highly important, of low importance, or irrelevant in relation to generating the response for the given prompt (Figure 3).

*Analogical reasoning* implies providing similar cases that allow users to compare and relate the GenAI responses to existing and known data points. Within this literature [13], these explanations are also called example-based, case-based, similarity-based, or evidence-based explanations. Because these What-else explanations provide additional information [21], they require increased mental effort to understand [37] – especially in cases of high uncertainty. Motivated by traditional machine learning systems, analogical explanations show similar input instances that result in similar model outputs and that are, for example, extracted from the training data

via nearest neighbor search [13]. In our GenAI system, we implement analogical reasoning by retrieving the top three most relevant contexts from our underlying database. To ensure cognitive effort remains comparable across treatments, these contexts are succinctly summarized before presentation (Figure 4).

Performance information on the local accuracy also phrased as certainty ratings [9] or confidence scores [21], can induce *probabilistic reasoning*. For example, prediction confidence scores can explain to which extent the prediction is correct or incorrect. A high score conveys trustworthiness while a low score will make the system look incompetent and less credible [21]. We instantiate probabilistic explanations as confidence scores in providing a correct answer to the given request. The confidence score varies per AI response, ranges from 0 (low) to 1 (high), and is extracted from the RAG system by representing the similarity of the retrieved facts with the response (Figure 5).

*Chain-of-Thought reasoning* refers to a contemporarily discussed approach to both improving and explaining large language model responses in the century of generative AI [8]. The approach has been applied to other domains in NLP such as question-answering [33]. In the case of RAG systems, chain-of-thought reasoning transfers less information on the underlying reasoning. Instead, it explains how the system derived a response including searching in the underlying database, retrieving relevant documents, and formulating an answer. Showing how RAG-systems work and how they reason to respond, combines global and local forms of explanation. This treatment presents each step of the RAG architecture such as searching in the underlying database, retrieving relevant documents, and formulating an answer in a human-readable and simplified form (Figure 6).

### 3.3 Samples

We recruited English-speaking participants aged 18+ through Prolific. They were compensated £2.25 for 15 minutes of participation and received a £0.10 bonus for each correct task as an incentive. Participants who failed attention checks or repeated the experiment were excluded and replaced. Only fully completed questionnaires were included, resulting in a final sample of  $N = 97$  (Age  $M = 36.06$ ;  $SD = 11.63$ ; 54% Female; 29% graduate degree or higher).

### 3.4 Procedure

The experiment followed a within-subject design, involving four explanation types and a no-explanation condition. Pre- and post-surveys collected demographic data, human characteristics, and subjective measures such as trust, confidence, and cognitive load (Figure 1). Participants acted as evaluators of GenAI systems, addressing mobile phone technical issues. Their primary task was to assess whether GenAI responses were correct or incorrect. Without access to the underlying database, participants relied solely on provided responses and explanations, knowing the average accuracy was 60%. From our set of 191 instances in the knowledge base, we pre-selected and prepared five prompts. Each participant received each prompt in a random order, paired with GenAI responses consecutively, and was randomly assigned one treatment per prompt-response pair to ensure internal validity. To control for factual confabulations, responses were pre-generated and validated.

While content originated from a real RAG system, all responses and features were pre-generated to eliminate potential confounds related to system errors or variability. Two responses were intentionally confabulated and incorrect, while three were accurate, based on the available knowledge base context. Confabulations reflected cases where the relevant source was deliberately removed from the knowledge base.

### 3.5 Measures

Using the decision on whether the GenAI response was accurate, combined with the provided ground truth, we evaluated the following metrics: We utilized the *confabulation detection rate* to measure the effectiveness of XAI in terms of recognizing incorrect AI advice [27]. Moreover, we measured *human-AI performance* by the rate of correctly classified GenAI responses [25, 34]. Furthermore, we measured *GenAI reliance* as the proportion of accepted responses deemed correct, reflecting the extent to which users trusted and relied on the AI's outputs [39, 60].

We also collected subjective measures for each task and treatment during the experiment to explore perceptions of explanations and understand the mechanics of confabulation detection: *Informativeness* was assessed using the statement, “This AI system and its response provide relevant information as well as persuasive and convincing arguments for me to understand.” [57]. In addition, we measured *trustworthiness* through the statement, “You feel that this AI system and its response are trustworthy, credible, and reliable.” [57]. Finally, we evaluated *understandability* using the item, “It is easy to understand the AI system’s response and additional information.” [18, 59].

We gathered participant data on age, gender, education, AI/GenAI experience, and expertise in mobile phone operations and customer support. No significant group differences were found, indicating balanced distribution across demographic and control variables.

## 4 Descriptive Analysis

Finally, we conducted a descriptive analysis of our experimental data to investigate the impact of the explanations on the users’ behavior in detecting confabulations and to demonstrate the validity of our experimental setup. We report sample means and run several ANOVA analyses to gain a first understanding of the relationships of interest. Our main results are presented in Figure 2.

### 4.1 Human-AI Performance and Confabulation Detection Rate

The data reveals differences between the features in terms of human-AI performance measured through accuracy (Figure 2a). Analogical ( $M = 0.670$ ,  $SD = 0.473$ ) and factual explanation ( $M = 0.606$ ,  $SD = 0.491$ ) performed best, while in case of no explanation average accuracy was lowest ( $M = 0.505$ ,  $SD = 0.503$ ). These results highlight a tendency toward explanations that include source documents from the underlying database. However, the average human-AI performance across all treatments ( $M = 0.572$ ,  $SD = 0.495$ ) is slightly lower than the GenAI system’s accuracy in our experiment (60%). Hence, we found that human evaluators, despite the provided XAI features, struggle to distinguish correct from incorrect responses. This highlights the need to further investigate the role of explanations in

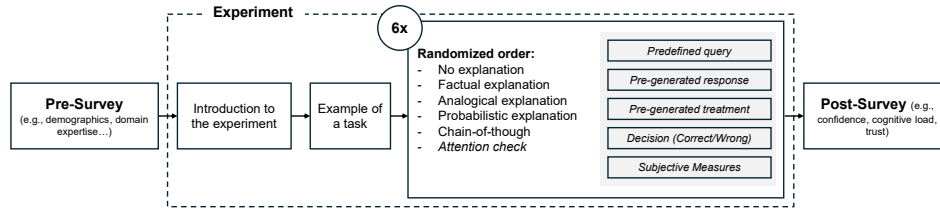
interacting with highly persuasive conversational systems. In terms of confabulation detection, the overall average rate was even lower ( $M = 0.211$ ,  $SD = 0.409$ ), as Figure 2a illustrates. Again, providing no explanation revealed the lowest rate ( $M = 0.150$ ,  $SD = 0.362$ ), while analogical explanation performed best ( $M = 0.308$ ,  $SD = 0.468$ ). Given the higher average confabulation detection rate observed with the explanations, the designed features appear to be effective. We additionally ran ANOVA analyses testing the relationship between the explanations and human-AI performance/confabulation detection rate which turned out to be insignificant. We shed further light on these relationships in section 5.

### 4.2 GenAI Reliance

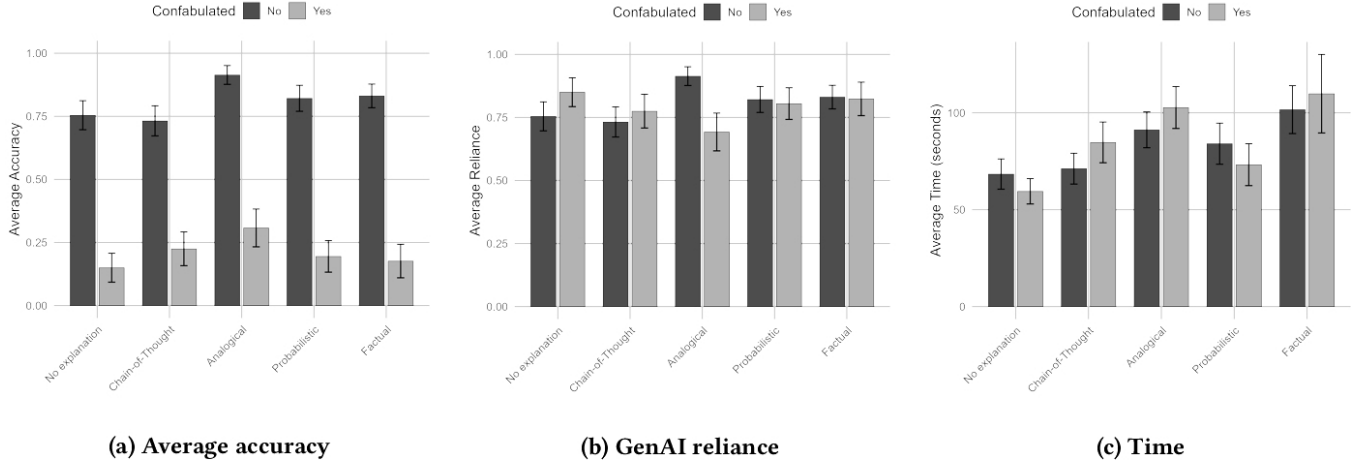
Our findings on GenAI reliance (Figure 2b) revealed that participants generally exhibited high levels of reliance across all explanation types, with average reliance ranging from chain-of-thought ( $M = 0.75$ ,  $SD = 0.435$ ) to factual ( $M = 0.828$ ,  $SD = 0.379$ ). Notably, reliance was only slightly higher when responses were correct ( $M = 0.812$ ,  $SD = 0.392$ ) compared to when they were incorrect ( $M = 0.789$ ,  $SD = 0.409$ ), suggesting no differences between confabulated and non-confabulated cases. Among explanation types, analogical explanations demonstrated a particularly pronounced difference in reliance between correct ( $M = 0.914$ ,  $SD = 0.283$ ) and incorrect responses ( $M = 0.692$ ,  $SD = 0.468$ ), which explains the findings regarding human-AI performance and confabulation detection. Interestingly, for no explanation, GenAI reliance was higher in confabulated cases ( $M = 0.850$ ,  $SD = 0.362$ ) compared to non-confabulated cases ( $M = 0.754$ ,  $SD = 0.434$ ). A similar pattern was observed for the chain-of-thought condition, where reliance was higher in confabulated cases ( $M = 0.775$ ,  $SD = 0.423$ ) than in non-confabulated ones ( $M = 0.732$ ,  $SD = 0.447$ ). These findings suggest a tendency toward overreliance on AI outputs, even when they are inaccurate. Additionally, we analyzed accuracy and GenAI reliance over time but observed no differences in the trajectories of these curves. This further supports the robustness of our experimental design, as participants’ behavior remains consistent over time. We also conducted ANOVA analyses to examine the relationship between explanations and GenAI reliance, but the results were not significant.

### 4.3 Time and Cognitive Load

The analysis of time spent per case reveals several key findings (Figure 2c). Participants spent on average 84.38 seconds per task ( $SD = 76.39$ ). The no explanation baseline required the least time ( $M = 64.74$ ,  $SD = 52.38$ ), while the factual explanation resulted in the highest average time spent ( $M = 104.00$ ,  $SD = 106.00$ ). Interestingly, time spent did not differ significantly between correct ( $M = 83.90$ ,  $SD = 76.40$ ) and incorrect ( $M = 85.10$ ,  $SD = 76.50$ ) responses, suggesting a consistent time effort regardless of response accuracy. An ANOVA revealed a significant effect of explanation features on time spent completing the tasks ( $F(4, 481) = 4.308$ ,  $p = 0.002$ ). Post-hoc tests showed that participants spent significantly less time without any explanation compared to the analogical ( $p = 0.034$ ) and factual ( $p = 0.002$ ) treatments, indicating that providing explanations increases the required user engagement. Marginal differences between factual



**Figure 1: The procedure spanning a pre-survey, an introduction, an example task, five randomized choice tasks and a post-survey.**



**Figure 2: Results of the descriptive analysis including (a) average accuracy and detection rate, (b) GenAI reliance, and (c) time.**

and chain-of-thought explanations ( $p = 0.080$ ) suggest variations in cognitive effort across explanation types.

Similarly, the analysis of cognitive load reveals significant differences across the explanation types. An ANOVA of cognitive load, conditional on task, shows a statistically significant effect of the explanation treatment ( $F(4,481) = 5.288, p < 0.001$ ), indicating that the type of explanation influences the cognitive effort required by participants. Pairwise t-tests further highlight these differences: Analogical explanations resulted in significantly higher cognitive load compared to no explanation ( $t = -2.40, p = 0.017$ ). Similarly, factual explanations increased cognitive load compared to no explanation ( $t = -2.75, p = 0.007$ ). However, we found that time and cognitive load are not associated with accuracy. While participants spent more time evaluating answers with explanations compared to the baseline (no explanation), only analogical explanations were linked to higher accuracy. Thus, more time spent reading does not necessarily predict better performance. These results validate our design, showing that participants actively engaged with the treatments, even though increased time did not guarantee better outcomes. We therefore conclude that differences in accuracy stem from the explanations themselves. In Section 5, we present a structural estimation model to further investigate these effects.

#### 4.4 Understandability, Informativeness and Trustworthiness

An additional ANOVA analysis revealed that understandability has a significant effect ( $F(1, 484) = 4.204, p = 0.041$ ) on human-AI performance, indicating that variations in understandability impact human-AI performance. Also, both trustworthiness ( $F(1, 484) = 7.863, p = 0.005$ ) and informativeness indicate higher accuracy ( $F(1, 484) = 3.891, p = 0.049$ ). Whereas cognitive load is not associated with higher or lower performance ( $F(1, 484) = 0.921, p = 0.338$ ). However, post-hoc tests revealed no significant differences across treatments, suggesting they do not fully explain the mechanisms through which explanations influence human-AI interaction.

#### 5 Logistic Regression Analysis

Next, we present the results of a logistic regression analysis to reliably elicit important aspects of confabulation detection. Our focus here is the relationship between subjective measures of AI answers, explanations of AI answers, the time spent on each choice task (IVs) and human-AI performance (DV). We employ the following regression equation:

$$\text{Logit}(y_{it} \mid \mathbf{X}_{it}, \mathbf{Z}_{it}, TS_{it}, C_{it}) = \alpha + \sum_p \beta_p X_{itp} + \sum_k \gamma_k Z_{itk} + \eta TS_{it} + \nu C_{it}$$

where  $y_{it}$  is a binary outcome, with 0 representing an incorrectly classified answer and 1 a correctly classified answer, for individual  $i$  at choice occasion  $t$ .  $\alpha$  is an intercept that measures the baseline detection of confabulations,  $\beta_p$  represents the marginal effects of explanations  $X_{itp}$ ,  $\gamma_k$  denotes the marginal effects of subjective evaluations  $Z_{itk}$ ,  $\eta$  measures the effect of time spent in minutes ( $TS$ ), and  $\nu$  measures the effect of the binary classification of whether the AI-generated answer was correct ( $C$ ). We fit a homogeneous model, assuming that individuals do not differ in how our measures  $X_{it}$ ,  $Z_{it}$ ,  $TS_{it}$ , and  $C_{it}$  influence the outcome. For explanations  $X_{it}$ , we include all explanation types outlined in Section 3.2, with the no explanation condition serving as the baseline.  $Z_{it}$  includes the subjective measures discussed in Section 4.4. Since we observe  $t \in \{1, \dots, 5\}$  choice tasks per individual  $i$ , fitting a frequentist heterogeneous model that accounts for random effects is not feasible. The individual-level log-likelihood functions are deficient and converge toward negative infinity. We report the main results in Table 1). Confirming our descriptive analysis, we find that the Analogical explanation has a positive and significant effect at the 5% level ( $\beta = 1.187, p = 0.003$ ) on correctly classifying an AI-generated response. Factual explanations also tend to positively influence human-AI interaction ( $\beta = 0.484$ ), though the effect is statistically insignificant ( $p = 0.199$ ), possibly due to assuming homogeneity in the population. Furthermore, time spent is not significantly related to classification accuracy ( $\beta = -0.073, p = 0.439$ ), while correct AI answers significantly improve classification accuracy ( $\beta = 3.279, p < 0.000$ ). This result may be attributed to the higher trust participants placed in the AI system. Given that our experiment contained more correct than incorrect answers, higher trust likely led to an increased rate of correct classifications when the response was indeed correct. Regarding subjective measures, only understandability shows a significant positive relationship with classification accuracy ( $\beta = 0.533, p < 0.000$ ). Further analysis revealed that trust ( $\beta = -0.008, p = 0.955$ ) and informativeness ( $\beta = -0.263, p = 0.087$ ) are highly correlated, which may explain these findings.

## 6 Conclusions and Discussion

The rapid advances in GenAI, coupled with its highly interactive and persuasive nature, call for innovative approaches to ensure reliable technology use and effective human-AI interactions. Acceptance of fabricated GenAI responses can, for instance, undermine customer satisfaction in service domains or hinder learning outcomes in educational settings. As a result, human users must critically evaluate and reflect on GenAI outputs. Our study offers initial insights into how prior knowledge from the field of XAI and the foundations of human reasoning can inform the design of explanations for LLM-based systems. The findings highlight the effectiveness of our GenXAI design and answer how explanations can augment humans in detecting confabulations. Differences in time spent suggest that participants actively engaged with and processed the explanations. While overall human-AI performance remains low, the explanations - particularly analogical and factual ones - demonstrated better performance. However, these treatments revealed higher time effort and heightened perceptions of cognitive load, pointing to a trade-off between time investment and accuracy.

As time spent is not associated with higher accuracy according to our analysis, the nature and quality of the explanations may better account for participants' behavior. Although understandability, trustworthiness, and informativeness positively correlate with human-AI performance, no significant associations were observed between these subjective measures and the specific treatments used. These findings underscore the need for deeper exploration into the mechanisms underlying the derived explanations.

Our current research presents a first exploration into understanding user's interactions with GenAI agents and the potential of GenXAI for facilitating these interactions by supporting facts and rationales. We contribute to the literature on decision-making in AI advice and XAI [4] by conceptualizing and examining explanation types in GenAI systems. Our experiment extends the existing XAI literature by probing several types of human reasoning as a foundation of explanations in LLM-enabled natural language-based systems. Furthermore, we extend the growing research on the detection of hallucinations and confabulations [12, 30, 38] and lay a foundation for designing XAI-based features that support humans confabulation detection. In addition, we outline managerial implications and implications for designing GenAI systems. Our findings suggest practitioners building GenAI agents interacting with human users to focus on providing information regarding the underlying sources such as referenced source documents (factual explanation) or similar documents (analogical explanation), while taking into consideration that these require more time and effort on the part of the users. These findings are mainly relevant for customer support settings as shown in this study but above that provide insights for various other domains such as education or social media. We provide technical details on how to leverage the inherent functions of RAG systems to provide supportive information.

However, our experiment does not come without limitations. *First*, we focused on the domain of customer service and optimized the treatments toward the usability of GenAI-based chatbots. Future research could transfer our experimental design to other domains such as education or healthcare. To mitigate the limitation of potential difficulties in understanding the technical domain, we collected data on user understandability within the study, which indicated a high level of comprehension. *Second*, our statistical analysis should incorporate a model that accounts for population heterogeneity and estimates effect coefficients as random effects. We aim to use a Bayesian Hierarchical model, well-suited for our short panels. As noted in Section 5, the deficient log-likelihood—due to only five observations per individual—precludes individual-specific random effects. A Bayesian estimation with priors as regularization tools will enhance insights from small datasets and enable the inclusion of human characteristics like domain expertise, AI literacy, and need for cognition in our structural model. *Third*, for future evaluations, it is essential to develop a deeper theoretical understanding of the factors that influence people's overall evaluations. The subjective measures used in this study do not account for the differences observed between the explanations. Lastly, we only considered one type of GenAI inaccuracies. However, there are multiple other types of errors such as those caused by ambiguity or intrinsic confabulations. Additionally, response features that influence how humans

identify confabulations or determine whether a response is confabulated should be considered. This aligns with studies examining human heuristics in distinguishing AI-generated text from human-written text [19]. Overall, our study lays an empirical foundation for further GenXAI research.

## References

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [2] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM, Glasgow Scotland UK, 1–12. doi:10.1145/3290605.3300484
- [3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).
- [4] Kevin Bauer, Moritz von Zahn, and Oliver Hinz. 2023. Expl(AD)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research* (2023), isre.2023.1199. doi:10.1287/isre.2023.1199
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prfulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell (Eds.), 2020. *Language models are few-shot learners*. Vol. 33. Larochelle H., Razato M., Hadsell R., Balcan M. F., and Lin H.
- [6] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. *The Quarterly Journal of Economics* (2025).
- [7] Furui Cheng, Vilém Zouhar, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobel, and Mennatallah El-Assady. 2024. *Interactive Analysis of LLMs using Meaningful Counterfactuals*.
- [8] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402* (2023).
- [9] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455–496. doi:10.1007/s11257-008-9051-3
- [10] Jiaxin Dai, Chao Zhang, Dzmitry Aliakseyeu, Samantha Peeters, and Wijnand A. Ijsselstein. 2023. The Effect of Explanation Design on User Perception of Smart Home Lighting Systems: A Mixed-method Investigation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–14. doi:10.1145/3544548.3581263
- [11] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792. doi:10.1016/j.ijhcs.2022.102792
- [12] Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. 30.0. Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions. <http://arxiv.org/pdf/2405.20434>
- [13] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. (2019). doi:10.48550/ARXIV.1901.07694
- [14] Nouha Dziri, Ehsan Kamaloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics* 10 (2022), 1473–1490. doi:10.1162/tacl\_2022\_10\_00529
- [15] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. 2024. *A Survey of AI Reliance*.
- [16] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217* (2023).
- [17] Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly* 23, 4 (1999), 497. doi:10.2307/249487
- [18] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. Ijsselstein. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–9. doi:10.1145/3491102.3517650
- [19] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America* 120, 11 (2023), e2208839120. doi:10.1073/pnas.2208839120
- [20] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38. doi:10.1145/3571730
- [21] Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies* 165 (2022), 102839. doi:10.1016/j.ijhcs.2022.102839
- [22] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).
- [23] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM, Glasgow Scotland UK, 1–12. doi:10.1145/3290605.3300717
- [24] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI 2022*. doi:10.48550/ARXIV.2204.11788
- [25] Vivian Lai, Han Liu, and Chenhao Tan. 2020. Why is 'Chicago' deceptive? Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA. doi:10.48550/ARXIV.2001.05871
- [26] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *FAT\* '19: Conference on Fairness, Accountability, and Transparency*. doi:10.48550/ARXIV.1811.07901
- [27] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. In *Proceedings of the ACM on Human-Computer Interaction*. doi:10.48550/ARXIV.2308.04375
- [28] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539. doi:10.1016/j.chb.2022.107539
- [29] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In *Mensch und Computer 2023*. ACM, New York, NY, USA, 81–90. doi:10.1145/3603555.3603565
- [30] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3613904.3642428
- [31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [32] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3313831.3376590
- [33] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.
- [34] Tian Lu and Yingjie Zhang. 2024. 1 + 1 > 2? Information, Humans, and Machines. *Information Systems Research* (2024). doi:10.1287/isre.2023.0305
- [35] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*.
- [36] Ji-Ye Mao and Izak Benbasat. 2001. The effects of contextualized access to knowledge on judgement. *International Journal of Human-Computer Studies* 55, 5 (2001), 787–814. doi:10.1006/ijhc.2001.0507
- [37] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [38] Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. *Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations*.
- [39] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (2023), 102941. doi:10.1016/j.ijhcs.2022.102941
- [40] Marvin Pafla, Kate Larson, and Mark Hancock. 2024. Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models. In *Proceedings of the CHI Conference on Human Factors*



- in *Computing Systems*. ACM. doi:10.1145/3613904.3642934
- [41] Hyanghee Park and Daehwan Ahn. 05112024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Florian Floyd Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe Touns Dugas, and Irina Shklovski (Eds.). ACM, New York, NY, USA, 1–21. doi:10.1145/3613904.3642785
- [42] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback*.
- [43] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3411764.3445315
- [44] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. doi:10.1145/3173574.3173677
- [45] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. *Language models are unsupervised multitask learners*. <https://life-extension.github.io/2020/05/27/gpt%e6%8a%80%e6%9c%af%e5%88%9d%e6%8e%a2/language-models.pdf>
- [46] Briana Rajan, Stephen Carradini, and Claire Lauer. 2024. The Arizona Water Chatbot: Helping Residents Navigate a Water Uncertain Future One Response at a Time. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–10. doi:10.1145/3613905.3650919
- [47] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M. Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 08.1. The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations. <http://arxiv.org/pdf/2310.04988>
- [48] Reuters. 29.08.2024. OpenAI says ChatGPT's weekly users have grown to 200 million. *Reuters Media* (29.08.2024). <https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/>
- [49] Johannes Schneider. 2024. Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *Artificial Intelligence Review* 57, 11 (2024), 289.
- [50] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2023. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human-Computer Interaction* 39, 7 (2023), 1390–1404. doi:10.1080/10447318.2022.2101698
- [51] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. (2023). doi:10.48550/ARXIV.2302.00096
- [52] S. M. Tonmoy, S. M. Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313* (2024).
- [53] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–17. doi:10.1145/3411764.3445101
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems* 30 (2017). <https://proceedings.neurips.cc/paper/7181-attention-is-all>
- [55] Jacqueline Wastensteiner, Tobias M. Weiss, Felix Haag, and Konstantin Hopf. 2021. Explainable AI for tailored electricity consumption feedback – an experimental evaluation of visualizations. In *ECIS 2021 Proceedings*. doi:10.48550/ARXIV.2208.11408
- [56] Darcia Wilkinson, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vebjerg, Bart P. Knijnenburg, and Elizabeth Daly. 2021. Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations. *ACM Transactions on Information Systems* 39, 4 (2021), 1–21. doi:10.1145/3441715
- [57] Kem Z. K. Zhang, Sesia J. Zhao, Christy M. K. Cheung, and Matthew K. O. Lee. 2014. Examining the influence of online reviews on consumers' decision-making: A heuristic-systematic model. *Decision Support Systems* 67 (2014), 78–89.
- [58] Shunyu Zhang, Param Vir Singh, and Anindya Ghose. 2019. A structural analysis of the role of superstars in crowdsourcing contests. *Information Systems Research* 30, 1 (2019), 15–33.
- [59] Tong Zhang, X. Jessie Yang, and Boyang Li. 2023. May I Ask a Follow-up Question? Understanding the Benefits of Conversations in Neural Network Explainability. (2023). doi:10.48550/ARXIV.2309.13965
- [60] 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*. 295–305. doi:10.48550/ARXIV.2001.02114

- [61] Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240* (2021).

## A Logistic Regression Analysis

We provide the results of our logistic regression analysis from Section 5 in Table 1.

**Table 1: Results of the logistic regression model. Significance codes: \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ , .  $p < 0.1$ , ' ' not significant.**

Predictor	$\beta$	SE	Pr(> z )	Sign.
(Intercept)	-3.282	0.551	0.000	***
Analogical Explanation	1.187	0.392	0.003	**
Chain-of-Thought	0.168	0.374	0.653	
Factual Explanation	0.484	0.376	0.199	
Probabilistic Explanation	0.372	0.373	0.319	
Correct	3.279	0.276	0.000	***
Understandability	0.533	0.117	0.000	***
Trustworthiness	-0.008	0.141	0.955	
Informativeness	-0.263	0.154	0.087	
Time Spent	-0.073	0.094	0.439	

## B Technical Descriptions of the Developed Treatments

### B.1 Factual Explanation

Given the predefined prompt, our GenXAI system retrieves the relevant source document ( $k = 1$ ) from the vector database and highlights phrases within this source, categorizing them as highly important, of low importance, or irrelevant in relation to generating the response for the given prompt. Thereby, this feature not only presents the underlying source document but identifies the sections within a source document from which the response was generated [11] (Figure 3). For highlighting the attribution of the relevant phrases in the source, we utilize an LLM (GPT-4o-mini) as well. The LLM is initialized by the following prompt:

*keyword\_init = "Given a Question and Source, list all the sentences in the source in order of them appearing, and mark if they are 'very important', 'mildly important' or 'not important' in answering the question, separated by semicolons, for example: mildly important: Sentence1; not important: Sentence2; very important: Sentence2; etc... Be strict which sentences are important and which are not."*

### B.2 Analogical Explanation

In our GenAI system, we implement analogical explanations (Figure 4) by retrieving the top three most relevant source documents from our underlying database. As with the RAG system the source documents are selected by utilizing cosine similarity based on the embeddings generated via the sentence transformer *bert-base-nli-mean-tokens*. To ensure cognitive effort remains comparable across treatments, these contexts are succinctly summarized before presentation by a GPT-4o model.

*extracter\_init = (f"Given a question and context text, answer the question (in short, max. 100 words) using the context. " f)"The user*

*does not see the context so try to formulate the answer as informative as you can. (remember: in short, max. 100 words)"*

```
extracter_bot = Chatter("gpt-4o", extractor_init)
sim_answer1 = extracter_bot.get_response("Question: " + sim_title1
+ sim_question1 + "Context: " + sim_answer1)
```

### B.3 Probabilistic Explanation

We implement probabilistic explanations (Figure 5) as confidence scores, representing the likelihood that the AI provides a correct answer to a given request. These confidence scores range from 0 (low) to 1 (high). To enhance user accessibility, the scores are categorized into different levels highlighted by color codings [8]. Factual accuracy, measured using the UpTrain framework, serves as the basis for representing confidence in each individual case. This score quantifies the extent to which a response aligns with the factual correctness of the provided context from the source document.

### B.4 Chain-of-Thought Explanation

The chain-of-thought assistant (Figure 6) operates by sequentially breaking down the process of answering a question into distinct steps. It identifies the required information (e.g., “To answer the question, I need to know: [FACT]”) and retrieves relevant data from the knowledge base. Using a vector store, the GenAI assistant searches for similar source documents within the data and highlights what it deems relevant for answering the question. It does not provide the final answer but instead focuses on explaining the retrieval and reasoning process, ensuring transparency of the overall RAG system.

```
chain_instruction = ("You are a chain of thought assistant!" "You
have to explain how you would find an answer to a question by
mentioning the steps you took to arrive at the solution and providing
the source you found." "Please answer in the following format: " "-
To answer the question, I need to know: [FACT]" "- I need to retrieve
similar cases from the database to find: [FACT]" "- Do not provide the
answer. We're only interested in the steps you take to find the answer.
Also keep your answers short.")
chain_assistant = Assistant("gpt-4o-mini", chain_instruction, [vector_store_id])
```

My phone was stolen, and I want to block its IMEI (International Mobile Equipment Identity) on the network so the thief can't use it, especially if they insert a SIM card. Can my internet provider still do these blocks? I have the IMEI on the original packaging and purchase receipt.

Yes, your internet provider can block the IMEI number of your stolen phone to prevent its use on the network, particularly if a new SIM card is inserted. You should contact your service provider immediately to report the theft and request the IMEI block.

**On what source did the AI base its response...**

Source Title	Source Type	Source Link
Call Forwarding Without Phone/SIM	Forum (A discussion where customers and experts share solutions)	<a href="https://telco.de/IS/Mobile/Call-forwarding-without-Mobile-SIM">https://telco.de/IS/Mobile/Call-forwarding-without-Mobile-SIM</a>

Highlighted Text

**very important** **mildly important** **not important**

*Forum user: An admin handling company phones could deal with such situations but I'm not sure if that complies with data protection laws when client numbers suddenly appear on a personal mobile.*

*The right approach would be to report the loss (block the IMEI, request a new phone + SIM card, which would be the admin's task). My work phone is encrypted so when I report a loss the IMEI gets blocked and once its online all data gets wiped.*

*Customer: The work phone isnt lost its just misplaced in the house where only I live. So it should only be a temporary situation.*

*And call redirection isnt about my personal phone receiving customer data its more about getting internal colleague calls.*

*Forum user: Have you tried calling the work phone to locate where it rings If its a work phone it should have been adequately charged.*

*Customer: switched to flight mode after work perhaps.*

**Do you think that the initial AI response is correct or wrong?**

Please select...

Correct ✓  Wrong ✗

**Please rate the following statements on a scale from totally disagree to totally agree.**

This AI system and its response provide relevant information as well as persuasive and convincing arguments for me to understand.

Totally Disagree  Disagree  Somewhat Disagree  Neutral  Somewhat Agree  Agree  Totally Agree

You feel that this AI system and its response are trustworthy, credible and reliable.

Totally Disagree  Disagree  Somewhat Disagree  Neutral  Somewhat Agree  Agree  Totally Agree

It is easy to understand the AI system's response and additional information

Totally Disagree  Disagree  Somewhat Disagree  Neutral  Somewhat Agree  Agree  Totally Agree

Rate your perceived level of mental effort during evaluating the AI system.

Extremely Low  Low  Somewhat Low  Neutral  Somewhat High  High  Extremely High

Do you have any feedback on this task?

**Next AI system**

**Figure 3: Screenshot of factual explanation including prompt, GenAI response, treatment, classification decision, and subjective measurements.**

Hello,

I set up my new iPhone and was charged €0.07 for an SMS related to the iMessage/FaceTime activation. Are these activation SMS now chargeable? How can I check where the SMS was sent?

Thanks for your help!

Activation SMS for iMessage or FaceTime may incur a charge depending on your mobile plan. To check where the SMS was sent, you can review your message history in the Messages app. Look for system messages or notifications indicating the recipient number.

**What are similar requests that result in the same AI response...**

<p><i>Similar Case 1</i></p> <p>Why doesn't the automatic login at the HotSpot work anymore after changing the password on the iPhone/iPad?</p> <p><i>Answer 1</i></p> <p>The automatic login at the HotSpot may not work after changing the password on an iPhone or iPad because the stored network settings must be updated. To resolve this, delete the "company" network settings in Wi-Fi, select "Ignore this Network," reconnect to the network, and then enter the new login details when prompted. Make</p>	<p><i>Similar Case 2</i></p> <p>Everywhere EDGE Solutions</p> <p><i>Answer 2</i></p> <p>To improve the iPhone 11's connection from EDGE to LTE, turning off VoLTE unexpectedly resolved the issue. Initially, attempts like switching providers, resetting network settings, and trying different SIM cards didn't work. The LTE became stable after turning off VoLTE and was maintained good performance when it was turned</p>	<p><i>Similar Case 3</i></p> <p>To adjust the ringing time before voicemail is activated on your Samsung Galaxy S2, enter the code **61*(voicemail number)** followed by 5-second increments (e.g.</p> <p><i>Answer 3</i></p> <p>To adjust the ringing time before voicemail is activated on your Samsung Galaxy S2, enter the code **61*(voicemail number)** followed by 5-second increments (e.g., **61*3334412345**10# for 10 rings or 50 seconds) and press send. The range can be set from 5 to 30 seconds. Each 5-second increment roughly equals one ring, so for 10 rings, set it around 50 seconds.</p>
---	---	--

**Do you think that the initial AI response is correct or wrong?**

Please select...

Correct ✓
  Wrong ✗

**Please rate the following statements on a scale from totally disagree to totally agree.**

This AI system and its response provide relevant information as well as persuasive and convincing arguments for me to understand.

Totally Disagree
  Disagree
  Somewhat Disagree
  Neutral
  Somewhat Agree
  Agree
  Totally Agree

You feel that this AI system and its response are trustworthy, credible and reliable.

Totally Disagree
  Disagree
  Somewhat Disagree
  Neutral
  Somewhat Agree
  Agree
  Totally Agree

It is easy to understand the AI system's response and additional information

Totally Disagree
  Disagree
  Somewhat Disagree
  Neutral
  Somewhat Agree
  Agree
  Totally Agree

Rate your perceived level of mental effort during evaluating the AI system.

Extremely Low
  Low
  Somewhat Low
  Neutral
  Somewhat High
  High
  Extremely High

Do you have any feedback on this task?

**Next AI system**

**Figure 4: Screenshot of analytical explanation including prompt, GenAI response, treatment, classification decision, and subjective measurements.**

My phone was stolen, and I want to block its IMEI (International Mobile Equipment Identity) on the network so the thief can't use it, especially if they insert a SIM card. Can my internet provider still do these blocks? I have the IMEI on the original packaging and purchase receipt.

Yes, your internet provider can block the IMEI number of your stolen phone to prevent its use on the network, particularly if a new SIM card is inserted. You should contact your service provider immediately to report the theft and request the IMEI block.

**How sure is the AI of its response..**  
Confidence scores can have a value between 0 and 100, with 0 being completely unsure and 100 being completely confident. 36 Confidence

**Do you think that the initial AI response is correct or wrong?**  
Please select...  
 Correct ✓  Wrong ✕

**Please rate the following statements on a scale from totally disagree to totally agree.**

This AI system and its response provide relevant information as well as persuasive and convincing arguments for me to understand.  
 Totally Disagree  Disagree  Somewhat Disagree  Neutral  Somewhat Agree  Agree  Totally Agree

You feel that this AI system and its response are trustworthy, credible and reliable.  
 Totally Disagree  Disagree  Somewhat Disagree  Neutral  Somewhat Agree  Agree  Totally Agree

It is easy to understand the AI system's response and additional information  
 Totally Disagree  Disagree  Somewhat Disagree  Neutral  Somewhat Agree  Agree  Totally Agree

Rate your perceived level of mental effort during evaluating the AI system.  
 Extremely Low  Low  Somewhat Low  Neutral  Somewhat High  High  Extremely High

Do you have any feedback on this task?

Next AI system

**Figure 5: Screenshot of probabilistic explanation including prompt, GenAI response, treatment, classification decision, and subjective measurements.**

The screenshot displays a user interface for an AI chat system. At the top, a light blue box contains the user's prompt: "Hello, I set up my new iPhone and was charged €0.07 for an SMS related to the iMessage/FaceTime activation. Are these activation SMS now chargeable? How can I check where the SMS was sent? Thanks for your help!". Below this, a white box shows the AI's response: "Activation SMS for iMessage or FaceTime may incur a charge depending on your mobile plan. To check where the SMS was sent, you can review your message history in the Messages app. Look for system messages or notifications indicating the recipient number." A grey bar below the response is titled "What steps did the AI take to come up with the answer ..." and contains a "Chain of thought:" section with three paragraphs explaining the AI's reasoning process. Below the chain of thought is a question: "Do you think that the initial AI response is correct or wrong?" with two radio button options: "Correct" (checked) and "Wrong". This is followed by a section titled "Please rate the following statements on a scale from totally disagree to totally agree." containing four statements, each with a seven-point Likert scale. The statements are: "This AI system and its response provide relevant information as well as persuasive and convincing arguments for me to understand.", "You feel that this AI system and its response are trustworthy, credible and reliable.", "It is easy to understand the AI system's response and additional information", and "Rate your perceived level of mental effort during evaluating the AI system." Below the ratings is a text input field for feedback: "Do you have any feedback on this task?". At the bottom, a grey button labeled "Next AI system" is visible.

Figure 6: Screenshot of chain-of-thought explanation including prompt, GenAI response, treatment, classification decision, and subjective measurements.