Understanding the Effects of Explaining Predictive but Unintuitive Features in Human-XAI Interaction

Jiaming Qu
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
jiaming@unc.edu

Jaime Arguello
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
jarguell@email.unc.edu

Yue Wang
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
wangyue@email.unc.edu

Abstract

Feature importance explanation, which highlights input features that are most influential to the output, is a popular explainable AI (XAI) technique to help users understand machine learning model predictions. However, features deemed predictive by machines can still be puzzling or even appear unintuitive to end-users. Explaining *why* a feature is predictive is an underexplored area in current XAI research. In this paper, we used deception detection as a case study. We leveraged a large language model (LLM) to explain why a word is predictive of genuine or deceptive reviews. We first validated the LLM-generated explanations to be non-hallucinated through an algorithmic evaluation. Then, we conducted a crowdsourced study ($N = 220$) to investigate how unintuitive words and LLM-generated explanations influence participants in a deception detection task. Our study results found that showing unintuitive features without explaining why they are predictive was no better than not showing them at all, while explaining why these features are predictive significantly enhanced participants' learning of the task, appropriate reliance on AI assistance, and perceptions of the AI system.

CCS Concepts

• **Human-centered computing → User studies**; • **Computing methodologies → Machine learning**.

Keywords

Explainable AI, Human-AI Interaction, Unintuitive Features, Large Language Models

ACM Reference Format:

Jiaming Qu, Jaime Arguello, and Yue Wang. 2025. Understanding the Effects of Explaining Predictive but Unintuitive Features in Human-XAI Interaction. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732021

1 Introduction

Explainable artificial intelligence (XAI) research has proposed a variety of approaches for helping users understand machine learning model predictions. One simple yet effective approach is to show which parts of an input (i.e., which features) are most influential to the prediction output. This approach is known as *feature importance explanation*. Previous research has developed a variety of algorithms to identify important features [51, 62, 69]. Studies have also empirically shown that feature importance explanations can improve end-users' performance on AI-assisted decision-making tasks [2, 45] and understanding of the AI system [15, 59].

Despite these promising results, studies on feature importance explanations have often ignored an important problem: features deemed important by a machine learning model may not always make sense to humans. We refer to these cases as **unintuitive features**. In this paper, we focus on unintuitive text features—words that are strongly associated with a label but are at odds with human intuition and common sense. For example, one prior study found that the word "Chicago" is a strong predictor of a hotel review being *deceptive* [44]. Even in a simpler task like sentiment analysis (i.e., predicting whether a product review is positive or negative), some word-sentiment relations can still appear incomprehensible (e.g., the word "problems" is counterintuitively predictive of *positive* sentiment [61]).

A user may naturally wonder: What makes an unintuitive feature predictive? There are two possible explanations. First, an unintuitive feature may be predictive due to anomalies in the training data and over-fitting. Second, an unintuitive feature may be predictive because it is associated with an *underlying language phenomenon* that may not be immediately obvious. For example, in the context of product reviews, the word "problems" is predictive of positive sentiment because of colloquial expressions like "no problems" or "without any problems". Similarly, in the context of deception detection, the word "Chicago" is predictive of a hotel review being fake because fake review writers tend to mention city names (e.g., "best in Chicago"). While machines learn from statistical patterns, humans understand language through context and prior experience [33, 67]. Without sufficient context, statistically salient word-label relations may appear unintuitive to humans. Most XAI research on feature important explanations do not further explain *why* a feature is predictive. This is the challenge addressed in this study.

When left unexplained, unintuitive features can lead to negative consequences. Studies have found that when no further explanations are provided, end-users may conjecture their own *incorrect* explanations [66], fail to realize the actual predictive power of unintuitive features [61], or lose trust in the AI system [14]. Therefore, explaining unintuitive features is an urgent problem in XAI research and requires effort from multiple perspectives. First, from a technical perspective, we need algorithms to further explain predictive

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732021

features that are not self-explanatory. Second, from a human perspective, we need to evaluate whether the generated explanations can help end-users appreciate *why* a feature is predictive and learn about the predictive task. Previous work has explored explaining unintuitive words using nearby words [34, 61, 63] and considering their interactions with other words in a sentence [6, 35, 72]. While these methods can reveal language patterns in local context windows, such as negations and colloquial expressions, they may fail to reveal more complex phenomena that go beyond local context.

Motivated by this, we aimed to develop a computational tool that can explain the complex language phenomena behind unintuitive words and evaluate the tool through a user experiment. To this end, we used deception detection as a case study—predicting whether a hotel review is genuine or deceptive. While fake review detection is a challenging task for laypeople [4], classical machine learning models like logistic regression perform reasonably well [44]. We prompted a large language model (LLM) to conjecture the underlying language phenomena that explain why a word is predictive of a review being genuine or deceptive. After validating that the LLM-conjectured phenomena reflected patterns grounded in data, we conducted a crowdsourced study ($N = 220$). In our study, participants were randomly assigned to one of five interface conditions (a between-subjects design). Interface conditions varied based on the AI assistance available to participants. The control condition only displayed AI predictions, while the other four treatment conditions provided different types of explanations for predictions. The explanations varied along two dimensions—(1) only with respect to the predicted label vs. both labels and (2) only highlighting predictive words vs. highlighting predictive words *and* showing their associated LLM-conjectured phenomena. Our study investigated three research questions (RQs):

- **RQ1**: How does the interface condition affect participants' *learning* of the deception detection task?
- **RQ2**: How does the interface condition affect participants' *reliance* on the provided AI assistance?
- **RQ3**: How does the interface condition affect participants' *perceptions* of the provided AI assistance?

Our study involved two phases: the main task and learning assessment. The goal of the main task was to train participants to detect deceptive hotel reviews by interacting with an AI system. During the main task, participants completed eight trials. During each trial, participants judged whether a Chicago hotel review was genuine or deceptive. Participants completed judgments in three steps: (1) making their own judgment without any AI assistance; (2) seeing the predicted label along with the AI assistance features associated with the assigned interface condition and possibly updating their judgment; and (3) seeing the correct ground-truth label. Then, participants proceeded to the learning assessment. The goal of this phase was to assess their learning of the task. In this phase, participants judged six reviews for hotels in other cities. While the main task had an interface manipulation, all participants during the learning assessment were exposed to the same interface condition that did not provide any AI assistance. Finally, participants completed a post-task survey asking about their perceptions of the AI system and their experiences during the study.

Our paper makes the following contributions.

- We developed a conjecture-then-validate algorithmic pipeline that leverages LLMs to explain predictive yet unintuitive features in XAI, using deception detection as a case study. LLMs were used to conjecture complex language phenomena that go beyond local context to explain unintuitive features. Then, these conjectured phenomena were validated to ensure they were not hallucinated but reflected actual data patterns.
- Our large-scale crowdsourced study found that LLM-generated explanations were helpful to users. Showing unintuitive features without explaining why they were predictive was no better than not showing them at all. In contrast, explaining why these features were predictive significantly helped participants learn about the deception detection task, develop appropriate reliance on AI, and improve overall perceptions of the system. Our findings highlight the importance of **aligning machine-generated explanations with human intuition to facilitate effective human-XAI interaction**.

2 Related Work

Our research builds upon three areas of prior work: (1) approaches for generating feature importance explanations, (2) approaches for explaining unintuitive text features, and (3) empirical user studies that evaluated XAI systems.

Feature Importance Explanations: XAI research has proposed a variety of techniques to help users understand AI predictions [40, 51, 62, 69, 77]. Feature importance (or feature attribution) explanation is a popular approach that identifies which parts of the input are most influential to the prediction [51, 62, 69]. In XAI research, explanations can be categorized as either *global* explanations that provide insights about the general behavior of the model, or *local* explanations that explain how the model works on individual instances [21]. This distinction also applies to feature importance explanations. Global feature importance explanations reveal the impact of features across predictions [19, 26], while local feature importance explanations identify influential features for a specific instance [51, 62, 69]. In our study, we trained logistic regression classifiers and used regression coefficients as global feature importance explanations.

Both global and local feature importance explanations share the same idea: "feature x plays an important role in predicting label y". However, predictive features are not always self-explanatory. For example, studies have reported predictive but unintuitive features in healthcare [12], econometrics [24] and psychology [36]. In this paper, we focused on predictive but unintuitive text features (words). This issue is not uncommon in prior XAI research. For example, the word "problems" is predictive of positive sentiment [61]; the word "Chicago" is predictive of deceptive reviews [44]; the word "host" is predictive of Atheism over Christianity [62]. Without proper clarification for predictive yet unintuitive features, studies have found that users: (1) made up their own *incorrect* explanations [66], (2) failed to realize the actual predictive power of these words [61], and (3) lost trust in the AI system that highlighted predictive words [14].

Explaining Unintuitive Text Features: The sparsity and high-dimensionality of text data pose unique challenges in explaining the predictiveness of unintuitive words. To this end, previous research

has proposed different heuristic approaches for explaining unintuitive words, e.g., curating explanations from domain knowledge of linguistic cues [44, 48], showing distributions of word-label relations in the training data [42, 61], and displaying instances where the word and label co-occur [61]. Additionally, previous research has also developed algorithms to explain unintuitive words, such as showing nearby words of the unintuitive word [34, 61, 63] and considering their interactions with other words [6, 35, 72]. The core idea of these algorithms is that contextual information can help explain the underlying phenomenon associated with a predictive yet unintuitive word. However, word-label relations can become more abstract in complex tasks and mere contexts are insufficient. This has inspired us to develop new tools.

Our proposed solution builds upon the emerging trend of applying LLMs in XAI research. LLMs have shown strong performance in a variety of machine learning and natural language processing tasks [79]. Thus, previous research has explored applying LLMs in XAI research, which mainly follows two directions. One research direction is to use LLMs as *explainers*. This line of work has leveraged LLMs to explain data graphs by uncovering hidden yet important trends [5, 46], and to explain machine learning model predictions by highlighting the most important features [41, 53]. Studies have shown that LLMs were able to identify important features as accurately as prevalent feature attribution techniques if provided with sufficient data and prompt engineering [41]. Besides being explainers, LLMs have also been used as *translators* for existing XAI techniques, with the primary goal of augmenting the readability of explanations. For example, studies have explored prompting LLMs to verbalize important features learned in text data [25, 56], tabular data [52, 82] and image data [52, 71] into coherent natural language narratives. However, these natural language explanations are still paraphrasing important features, without further explaining why they are important. In this paper, we prompted an LLM to conjecture the underlying phenomena from words that are predictive of genuine or deceptive reviews. Instead of *paraphrasing* feature-based explanations to natural language explanations, our approach *elevating* low-level word features into higher-level phenomena-based explanations. We elaborate more on this in Section 3.2.

Empirical Studies in XAI: Conducting empirical studies with human end-users is a common approach for evaluating XAI systems, as it provides direct evidence on the effectiveness of explanations [43, 80]. To this end, numerous user studies have been conducted. Most studies have adopted a similar design—end-users interacted with an XAI system (e.g., showing important features [11, 15, 44, 45] or similar examples [10, 13, 76]) to accomplish a specific task, such as sentiment analysis [2], topic categorization [59], and deception detection [44, 45]. When evaluating the effects of XAI systems on end-users, studies typically focused on *what to evaluate* and *how to evaluate* [43]. For example, studies have conducted quantitative evaluations with respect to (1) the *task* (e.g., efficacy [2, 11, 44, 45, 59] and efficiency [11, 15, 59]), (2) the *AI* (e.g., users' trust [10, 15, 45, 76] and understanding [15, 17, 59] of the AI), and (3) the *end-users* (e.g., behavioral patterns [2, 11, 59]). Besides quantitative evaluations, qualitative studies have delved deeper into human-XAI interaction, such as users' mental models and cognitive activities [10, 28, 60].

In our study, we conducted quantitative evaluations with respect to the task, AI, and end-users to understand the effects of unintuitive features and LLM-generated explanations. Notably, we investigated a novel aspect (RQ1): can XAI systems facilitate users' learning of a complex task? While prior studies have largely focused on users' task performance with *in-situ* XAI assistance [2, 11, 44, 45, 59], few have investigated the subsequent impact after the assistance becomes unavailable. Helping users assimilate new knowledge, however, is an important educational objective in developing XAI systems [21]. Therefore, our RQ1 investigated whether an XAI system could perform as an instructor instead of merely an assistant. To address this, our study proceeded in two phases. During the main task, participants interacted with an XAI system to learn how to distinguish between genuine and deceptive hotel reviews. During the learning assessment task, participants labeled reviews as genuine versus deceptive independently. It is notable that such workflow has been used to assess users' understanding of AI systems in prior studies (i.e., simulating AI behaviors post-interaction) [9, 17]. In our study, we adapted it to evaluate participants' learning of the deception detection task. Participants were instructed to make the correct judgment instead of simulating the AI.

3 Methods

3.1 Study Overview

To investigate our RQs, we conducted a crowdsourced study using the Prolific platform. Our study involved 220 participants (M = 121, F = 98, $Unreported$ = 1). Participants' ages ranged from 19 to 71 ($Mean$ = 37.35, $S.D.$ = 11.23). We restricted our study to English-speaking Prolific workers from USA, UK, and Canada who had completed at least 100 tasks with an acceptance rate ≥ 95% and had experience in online hotel booking.

Our study involved two phases: the main task and learning assessment. We designed the main task as a phase where participants could learn to detect deceptive hotel reviews by interacting with an AI system. During the main task, participants were exposed to eight reviews for Chicago hotels. For each review, participants were asked to judge whether the review was genuine or deceptive. Participants were randomly assigned to one of five interface conditions (i.e., a between-subjects design) that provided different AI assistance features. The learning assessment phase was designed to test participants' learning. During the learning assessment phase, participants were asked to judge six reviews for hotels in other cities. We decided to use reviews from other cities to test participants' ability to use what they learned in a *novel* context, which is evidence of deeper learning. All participants were exposed to the same interface that did not provide any AI assistance. Finally, participants completed an exit survey about their experiences during the study. Participants were given US$ 5.5 for participating in the study and an additional US$ 0.2 for each correct judgment during the learning assessment phase. The study was approved by our Institutional Review Board (IRB).

3.2 An LLM-based Approach for Explaining Unintuitive Words

Data Preparation: The dataset used in the main task of our study originated from Ott et. al [54, 55], which contains 800 genuine and

800 deceptive reviews for Chicago hotels. Genuine hotel reviews were collected from verified travelers while deceptive reviews were written by crowdsourced workers. Given the limited size of the dataset, we performed 10-fold cross validation to ensure optimal use of all the reviews. The data was prepared as follows. For each training fold, we trained a logistic regression classifier, identified the most predictive words, and leveraged an LLM to conjecture the underlying phenomena associated with these predictive words. For each corresponding testing fold, we applied the trained classifier to predict whether each hotel review in the test set is genuine or deceptive and validated the conjectured phenomena as described in Section 4.1.

Identifying Predictive Words: We trained logistic regression classifiers using a unigram TF-IDF representation with stopwords removed [3, 58] and the classifiers achieved an average F1-score of 0.88. Similar to previous XAI studies that used unigram text features [11, 44, 45], we identified predictive words through regression coefficients. From a trained logistic regression classifier, we selected the 25 words with the highest coefficients as the most predictive of genuine reviews and the 25 words with the lowest coefficients as the most predictive of deceptive reviews. All selected words passed a Wald test for the significance of predictiveness.

Explaining Predictive Words: Several prior studies have also aimed to explain why a word is predictive [6, 35, 61, 72]. However, these studies have focused on sentiment analysis—predicting whether a review is positive or negative. These studies have explained the predictiveness of words by showing the contexts in which they tend to appear. For example, the word "problems" is predictive of *positive* sentiment because it appears in contexts such as "no problems" and "without any problems". Compared to sentiment analysis, deception detection—predicting whether a review is genuine or fake—is a more complex task for laypeople [4]. Our preliminary investigation results (Table A.1) suggest that using contextual information is not enough. Instead, predictive words in deception detection may reflect phenomena that cannot be explained by simply showing the contexts in which they frequently occur. For instance, a system may need to explain that "spa" is predictive of a hotel review being fake because "fake review writers tend to overemphasize luxurious aspects of the hotel".

To go beyond context-based explanations, we leveraged an LLM (`GPT-4o` [1]) to **conjecture the underlying phenomena represented by predictive words**. We used the following prompt: `You trained a logistic regression model to detect deceptive Chicago hotel reviews. Based on feature importance, these are words predictive of genuine reviews [a list of words] and these are words predictive of deceptive reviews [a list of words]. Your task is to identify language phenomena that appear in genuine and deceptive hotel reviews and are associated with these predictive words. Structure your analysis using the template below for a clear and concise response: {"Underlying phenomena for genuine reviews": { "phenomenon 1": {"phenomenon": short phrases describing this phenomenon, "explanation": elaboration on this phenomenon, "predictive words": words associated with this phenomenon"}, ...}, {"Underlying phenomena for deceptive reviews": ... }` Here, we prompted `GPT-4o` to conjecture and organize phenomena directly from predictive words without looking at any examples. We did not restrict the number of phenomena to be conjectured. Table 1 summarizes the union of conjectured phenomena and their associated words across all 10 training folds. Each word is associated with one phenomenon.

These LLM-conjectured phenomena are plausible, but all LLMs are prone to hallucinations. An important question is: Are these phenomena *truly* grounded in genuine and deceptive reviews? We discuss how we validated these phenomena through an algorithmic evaluation in Section 4.1.

3.3 Study Design

Sampling Reviews for the Study: During the main task, participants judged eight reviews for Chicago hotels. We expected our crowdsourced participants to have limited experience in detecting deceptive hotel reviews. Thus, we sampled sequences of eight reviews based on the following criteria to support their learning of the deception detection task. First, each review was associated with at least one predictive word for both genuine and deceptive reviews. Second, each sequence covered all predictive words for both labels (as shown in Table 1). Third, each sequence had four truly genuine and four truly deceptive reviews. For each label (i.e., genuine vs. deceptive), we sampled two reviews for which the logistic regression classifier made a correct prediction and two reviews for which the logistic regression classifier made an incorrect prediction. In total, we sampled 22 unique sequences without duplicate reviews for the main task.

For the learning assessment phase, we used another dataset containing genuine and deceptive reviews for hotels in Houston, New York, and Los Angeles [47]. We sampled 22 unique sequences without duplicate reviews. Each sequence had six reviews, one genuine and one deceptive reviews for each city.

Task Allocation: We organized all sampled reviews into 22 unique batches. Each batch had a sequence of eight reviews for Chicago hotels for the main task and six reviews for hotels in other cities for the learning assessment task. Ten participants were exposed to the same batch. Among these participants, two were assigned to each of our five interface conditions during the main task. Ultimately, we had 22 unique batches $\times$ 5 interface conditions during the main task $\times$ 2 participants per interface condition, for a total of 220 participants. Figure B.1 illustrates our task allocation.

Showing AI Assistance: Prior user studies on human-AI interaction have adopted two paradigms when providing AI assistance. One is to provide real-time AI assistance during a task for participants to take or ignore [11, 44, 45, 59], and the other is to provide AI assistance after a task for participants to revise their decisions [13, 65]. Considering the task complexity, we adopted the second paradigm. We expected that delaying the AI assistance could nudge participants to think critically and mitigate the risk of blind reliance. While this design does not resemble most real-world cases of AI-assisted decision-making, it provides insights into how participants' behaviors shift before and after AI assistance.

3.4 Study Protocol

Our study protocol proceeded as follows. First, participants watched an instructional video about the study. Then, participants proceeded

Table 1: GPT-4o-conjectured phenomena from words predictive of genuine and deceptive reviews (abridged due to space limit).

Words predictive of genuine reviews	Genuine phenomena conjectured by GPT-4o
floor, elevator, elevators, breakfast, coffee	focusing on practical aspects of hotel amenities
small, large, bathroom, bed, upgraded	paying attention to specific room details and features
Priceline, Booking, booked, rate	mentioning transactional and booking details of the stay
helpful, us, concierge, conference	mentioning interaction with hotel staff or specific use cases
construction, quiet	mention environmental and surrounding factors of the hotel
reviews	referring to other reviews for the hotel
location, Michigan, avenue, walk, street, blocks, river	mentioning aspects related to the hotel's location
Words predictive of deceptive reviews	**Deceptive phenomena conjectured by GPT-4o**
luxury, luxurious, spa, accommodations	overemphasizing the luxurious aspects and high-end services
experience, relax, relaxing, visit, staying	overusing words related to personal experiences and emotions
smelled, smell, food	exaggerating unusual sensory details and experiences
amazing, definitely, excellent, ever	overusing superlatives and absolute terms
recently, finally	claiming recent and up-to-date experience
Hilton, Millennium, Chicago, Regency	mentioning well-known hotel brands or city names
husband, wife, family, vacation	mentioning family members or personal events
recommend, anyone, looking	using persuasive language for direct suggestions

to the main task. Participants were randomly assigned to one of five interface conditions that varied based on the AI assistance. During the main task, participants completed eight trials. During each trial, participants judged whether a Chicago hotel review was genuine or deceptive. Participants completed judgments for a review in three steps.

First, participants read the review and made judgments using a range slider from very deceptive to very genuine. The range slider did not have a midpoint. Therefore, participants had to choose between genuine or deceptive. However, they could choose values close to the midpoint if they were unsure. Second, participants were provided AI assistance. Participants were instructed that the AI system could make mistakes. However, they did not know the exact distribution of correct and incorrect cases. After scrutinizing the AI assistance, participants were asked to either revise their judgment using the range slider or keep their original judgment. Finally, participants were shown a summary of their judgment and the correct answer.

After completing all eight trials for the main task, participants proceeded to the learning assessment task where they completed six trials. Participants did not know the six trials were balanced between genuine and deceptive. During each trial, participants judged a review for hotels in other cities. All participants were exposed to the same interface that only included the review and a range slider without any AI assistance. After completing the learning assessment, participants completed a post-task survey that asked about their perceptions of the AI system and their experiences during the study. Our study materials and system demos are available at our online appendix.

3.5 Post-task Survey

In the post-task survey, participants responded to agreement statements on a 7-point scale ranging from (1) "strongly disagree" to (7) "strongly agree". The survey consists of three parts. The first part of the survey asked participants' **trust** in the AI's prediction (3 items), **understanding** of the AI's prediction (3 items), **confidence** in their judgments in the main task (1 item) and learning assessment (1 item), and their perceived **learning** of the deception detection task (3 items). The second part of the survey asked about **system usability**. We used the 10-item System Usability Scale (SUS) [7]. The third part of the survey asked about **workload**. We used the 6-item NASA-TLX [30], which asked about participants' mental demand, physical demand, temporal demand, failure, effort, and frustration. Participants' responses to the trust, understanding, learning, and system usability items showed high internal consistency (Cronbach's $\alpha \geq 0.87$). Therefore, we aggregated responses within these categories to form composite measures. Participants' responses to the workload items had medium internal consistency (Cronbach's $\alpha = 0.65$), so we analyzed them individually.

3.6 Interface Conditions

In the main task, participants were randomly assigned to one of five interface conditions (i.e., a between-subjects design). Interface conditions varied based on the AI assistance features made available to participants. Participants completed the same task (i.e., judging eight Chicago hotel reviews) with the same study protocol for all conditions.

The five interface conditions consisted of one control and four treatment conditions. The control condition only provided predicted labels from logistic regression classifiers. The four treatment conditions included predicted labels and provided additional explanations for predictions. Our four treatment conditions were associated with a two-by-two factorial design. We manipulated two factors. One factor manipulated the *type* of explanation provided—either predictive words only (Word) or predictive words along with the LLM-generated phenomena associated with each predictive word (WordPhen). The second factor manipulated the *side* of explanations provided—either explanations associated with the predicted label only (Single) or both labels (Both).

Control: This condition only showed the predicted label.

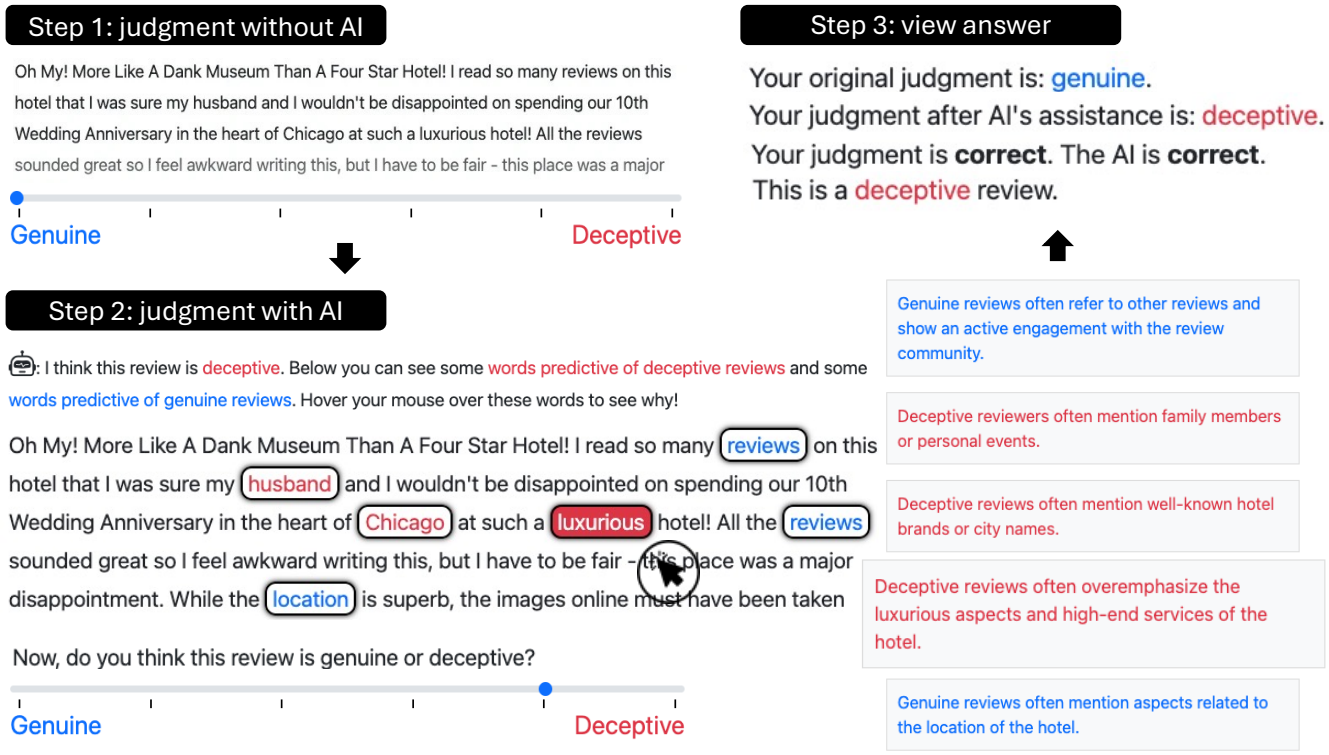

Figure 1: Main task system interface. In the main task, participants (1) made their own judgment, (2) scrutinized the AI assistant and possibly updated their judgment, and (3) saw the correct answer. Depending on the interface condition, participants had access to different AI assistance features. For example, the WordPhenBoth condition shown in the figure highlights predictive words and provides LLM-conjectured phenomena for both labels. Participants can hover the mouse over predictive words to inspect the corresponding phenomena or hover over the phenomena to inspect related words.

WordSingle: This condition showed the predicted label and highlighted predictive words for the predicted label.

WordBoth: This condition showed the predicted label and highlighted predictive words for both labels.

WordPhenSingle: This condition showed the predicted label, highlighted predictive words for the predicted label, and provided LLM-conjectured phenomena for those predictive words.

WordPhenBoth: This condition showed the predicted label, highlighted predictive words for both labels, and provided LLM-conjectured phenomena for those those predictive words. Figure 1 illustrates this condition.

While the main task had an interface manipulation, the learning phase did not—participants had to classify reviews without any AI assistance. The system interface is the same as the first step of the main task, as shown in Figure 1.

3.7 Measures of Learning (RQ1)

In RQ1, we investigated the effects of the interface condition on participants' **learning** of the deception detection task. We developed the following two measures to assess both objective and subjective learning outcomes.

Judgment Accuracy: The percentage of correct judgments participants made in the learning assessment.

Perceived Learning: The average response to the three post-task survey items that asked about perceived learning (Section 3.5).

3.8 Measures of Reliance (RQ2)

In RQ2, we investigated the effects of the interface condition on participants' **reliance** on the AI. As shown in Table 2, we developed five measures based on participants' judgments during the main task.

Similar to prior work that measured reliance through agreement [2, 11, 59], we first measured participants' tendency to rely on the AI's predictions from two different perspectives.

Change toward AI: Out of the trials where participants disagreed with the AI first, the percentage of times they changed their judgment to agree with the AI.

Stick with AI: Out of the trials where participants agreed with the AI first, the percentage of times they maintained agreement with the AI.

Previous research suggests that the quality of reliance depends on how humans respond to correct or incorrect AI predictions [13, 38, 57, 76]. Thus, we used the following measures to investigate whether the reliance is appropriate.

Appropriate Reliance: The percentage of times participants either correctly agreed with the AI (i.e., the AI was correct) or

Table 2: Measures of reliance. Notation: H = human's initial judgment before AI assistance, $H^{'}$ = human's final judgment after AI assistance, AI = AI's prediction, T = true label. Each takes a value in $\{Deceptive, Genuine\}$.

Tendency of Reliance		Appropriateness of Reliance	
Measure	**Definition**	**Measure**	**Definition**
Change toward AI	$P(H^{'} = AI \mid H \neq AI)$	Appropriate Reliance	$P(H^{'} = AI, AI = T) + P(H^{'} \neq AI, AI \neq T)$
Stick with AI	$P(H^{'} = AI \mid H = AI)$	Over-Reliance	$P(H^{'} = AI \mid AI \neq T)$
		Under-Reliance	$P(H^{'} \neq AI \mid AI = T)$

correctly disagreed with the AI (i.e., the AI was incorrect). This is equivalent to the participant's accuracy during the main task.

Over-Reliance: Out of the four trials where the AI was incorrect, the percentage of times participants incorrectly agreed with the AI (i.e., accepted an incorrect prediction).

Under-Reliance: Out of the four trials where the AI was correct, the percentage of times participants incorrectly disagreed with the AI (i.e., rejected a correct prediction).

3.9 Measures of Perceptions (RQ3)

In RQ3, we investigated the effects of the interface condition on participants' **perceptions** of the AI and study experiences. We developed the following measures for RQ3 based on the post-task survey (Section 3.5).

Trust: The average rating to three survey statements asking about participants' trust in the AI assistance.

Understanding: The average rating to three survey statements asking about participants' understanding of the AI assistance.

Confidence in Main Task: The rating to one survey statement asking about participants' confidence when making judgments in the main task.

Confidence in Learning Assessment: The rating to one survey statement asking about participants' confidence when making judgments in the learning assessment.

SUS: The average rating to ten survey statements asking about participants' perceived system usability.

Workload: The rating to six survey statements asking about participants' workload.

3.10 Statistical Analysis

In the main task, participants were randomly assigned to one of five interface conditions—one control condition and four treatment conditions from a two-by-two factorial design. Following recommendations for analyzing experimental data where the control condition is not part of the factorial design [32] and previous XAI studies using the same experimental design (i.e., 2 × 2 treatments + 1 control) [50, 75], we conducted statistical analysis in two stages.

First, we conducted Dunnett's tests [22] to compare each treatment condition against the control condition. These tests considered the effects of providing some form of explanation versus no explanations. Second, we focused on the four treatment conditions *only*. We conducted two-way ANOVA analysis based on the two-by-two factorial design—(1) Word vs. WordPhen, with Word as the reference level and (2) Single vs. Both, with Single as the reference level. The two-way ANOVA analysis allowed us to systematically analyze the effects of providing phenomena-based explanations and double-sided explanations that may not be apparent in individual comparisons with the control condition. We conducted two-way ANOVAs to consider both main and interaction effects. However, there were no significant interaction effects for any of our measures. Therefore, in the following section, we report on only main effects.

4 Results

4.1 Validating LLM-conjectured Phenomena

Before presenting results for RQ1-RQ3, we describe how we validated the LLM-conjectured phenomena for the most predictive words associated with each category (i.e., genuine and deceptive). Evaluating the output from an LLM is a challenging task. Prior studies have either compared the LLM output against a benchmark [31, 70] or have recruited human evaluators [29, 81]. In our study, we took an algorithmic approach to evaluating LLM-conjectured phenomena associated with the most predictive words. Inspired by research that leveraged LLMs to evaluate machine-generated contents [16, 39, 74], we integrated the conjectured phenomena into a new deception detection prompt for another LLM (`GPT-4o`). The rationale is that providing non-hallucinated phenomena should improve (or at least not hurt) its predictive performance. We used the following prompt: `You are an expert in detecting fake hotel reviews online. You will be given a review for a Chicago hotel, and your task is to classify it as a genuine or deceptive review. You only need to tell your prediction. [auxiliary information][the review to be predicted]`.

We evaluated three prompt conditions. First, we tested a *zero-shot prompt*, where the LLM made predictions without any `[auxiliary information]` [78]. Second, we used a *ten-shot prompt* to facilitate the LLMs' in-context learning [8]. We substituted `[auxiliary information]` with five genuine and five deceptive reviews from the training data that were most semantically similar (measured through text embeddings) to `[the review to be predicted]`. Finally, we used a *phenomena-in-the-prompt* condition, where we replaced `[auxiliary information]` with all LLM-conjectured phenomena associated with both genuine and deceptive reviews (as shown in Table 1). We used `GPT-4o` through the OpenAI API with default settings. For each condition, `GPT-4o` predicted one review at one time. In total, we made 4800 independent API calls (1600 reviews × 3 prompt conditions).

Table 3 shows evaluation results of `GPT-4o`'s predictive performance under different prompt conditions (averaged across 10 folds). The *phenomena-in-the-prompt* condition outperformed both *zero-shot* and *ten-shot* conditions. These results demonstrate that **the LLM-conjectured phenomena used in our study were not hallucinated**. Rather, they are indeed related to psycholinguistic

Table 3: Evaluation results of GPT-4o's predictive performance on the Chicago hotel reviews dataset under different prompt conditions. Means (and standard deviations) are computed from 10-fold cross validation. A ▲(▼) symbol indicates significant differences ($p < .05$) from the *phenomena-in-the-prompt* condition (measured through Fisher's Randomization tests [27]).

Prompt Condition	Accuracy	Precision	Recall	F1
zero-shot	0.6388 (0.0933) ▼	0.7628 (0.1033)	0.3737 (0.1741) ▼	0.4889 (0.1772) ▼
ten-shot	0.6800 (0.0629) ▼	0.6318 (0.0486) ▼	0.8712 (0.0597)	0.7318 (0.0496)
phenomena-in-the-prompt	0.7231 (0.0487)	0.7009 (0.0616)	0.8112 (0.1612)	0.7397 (0.0623)

patterns within genuine and deceptive hotel reviews. Thus, LLMs can be a reliable tool to explain unintuitive words in deception detection. More importantly, validating the conjectured phenomena ensured that participants were not exposed to fabricated information during our user study.

4.2 RQ1: Learning

In RQ1, we investigated the effects of the interface condition on participants' learning of the deception detection task. Figure 2 shows our RQ1 results. Our results found three main trends.

First, participants in the Control condition achieved 54.9% accuracy in the learning assessment. Given that half the reviews were genuine and half were deceptive, this accuracy value is only slightly above random guessing. In the Control condition, participants got trained merely from prediction outcomes without further explanations. This result highlights the inherent difficulty of the deception detection task for laypeople and resonates with prior work [4].

Second, compared to the Control condition, participants who had access to predictive words with phenomena-based explanations achieved significantly better learning outcomes in both objective and subjective measures, while participants who only had access to predictive words did not.

Third, the two-way ANOVA further confirmed that **adding phenomena-based explanations significantly improved both subjective and objective learning compared to showing predictive words alone**. Moreover, providing explanations for both sides enhanced objective learning compared to explanations for one side. One possible reason is that participants got richer training experience in the main task.

4.3 RQ2: Reliance

In RQ2, we investigated the effects of the interface condition on participants' reliance on the AI. Figure 3 shows our RQ2 results. There were no significant effects observed for the change toward AI and over-reliance measures. Thus, the corresponding plots are omitted. Our results found three main trends.

First, compared to the Control, participants did not exhibit significantly different levels of reliance when having access to explanation tools. However, the two-way ANOVA revealed a significant difference for the stick with AI measure between Word and WordPhen. When participants and the AI agreed initially, **phenomena-based explanations might have nudged participants to scrutinize AI predictions more critically and even change to opposite judgments**. As a result, they were less likely to maintain the agreement with the AI (i.e., less reliance).

Second, compared to the Control, providing predictive words with phenomena significantly increased participants' appropriate reliance and reduced their under-reliance, while only providing predictive words did not have any significant effects. The two-way ANOVA further confirmed this trend. Compared to predictive words (e.g., breakfast, floor, elevator), providing the underlying phenomena (e.g., genuine reviews often focus on physical and practical aspects of hotel amenities) **significantly reduced the likelihood that participants rejected correct AI predictions.**

Third, compared to the Control, both predictive words and phenomena-based explanations did not significantly alleviate participants' over-reliance on incorrect AI predictions. It suggests that **while providing phenomena helped participants recognize correct AI predictions more effectively, they did not always help participants detect incorrect AI predictions**. Nonetheless, the reduction in under-reliance was strong enough to drive a significant higher appropriate reliance, which was confirmed by the two-way ANOVA.

4.4 RQ3: Perceptions

In RQ3, we investigated the effects of the interface condition on participants' perceptions of the AI and their experiences. Figure 4 shows our RQ3 results. Except for the failure measure, there were no significant effects observed on other workload measures. Thus, the corresponding plots are omitted. Our results found two main trends.

First, compared to the Control, participants who had access to phenomena-based explanations reported significantly higher trust, understanding, and confidence, while participants who only had access to predictive words did not report any significantly better perceptions.

Second, the two-way ANOVA further confirmed the importance of providing phenomena-based explanations. Compared to predictive words alone, **phenomena-based explanations resulted in higher levels of trust and understanding of the AI, greater confidence in making judgments, and higher perceived usability without increasing workload.** Moreover, participants exposed to phenomena-based explanations reported significantly lower level of perceived failure during the task.

5 Discussion

Summary of Results: In our study, we investigated the effects of providing different types of AI assistance on participants' learning, reliance, and perceptions in a deception detection task. Our study found the following results.

First, compared to the Control condition, only highlighting predictive words did not have any significant effects. This is in contrast to findings in previous studies that *in-situ* feature importance explanations improved participants' decision-making performance and perceptions of the AI system [2, 15, 45, 59]. One possible explanation is that predictive words in this study indeed appeared

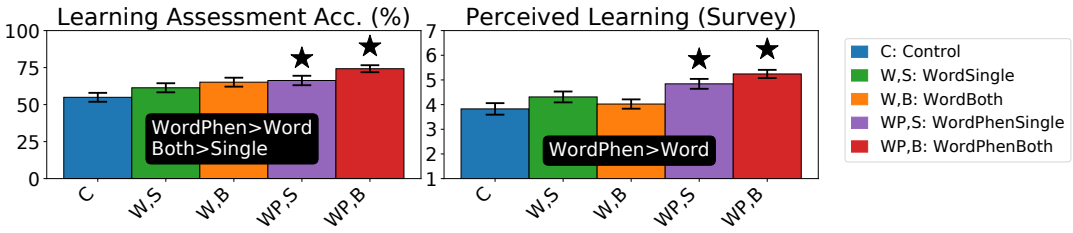

Figure 2: Effects of different interface conditions on participants' *learning* with means and standard error. The star mark highlights interface conditions with statistically significant effects ($p < .05$) compared to the Control condition from Dunnett's tests. The text box highlights statistically significant effects ($p < .05$) from two-way ANOVA.

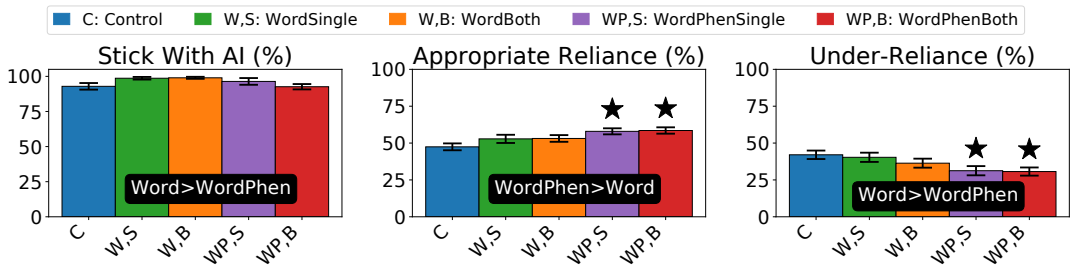

Figure 3: Effects of different interface conditions on participants' *reliance* with means and standard error. The star mark highlights interface conditions with statistically significant effects ($p < .05$) compared to the Control condition from the Dunnett's test. The text box highlights statistically significant effects ($p < .05$) from two-way ANOVA.

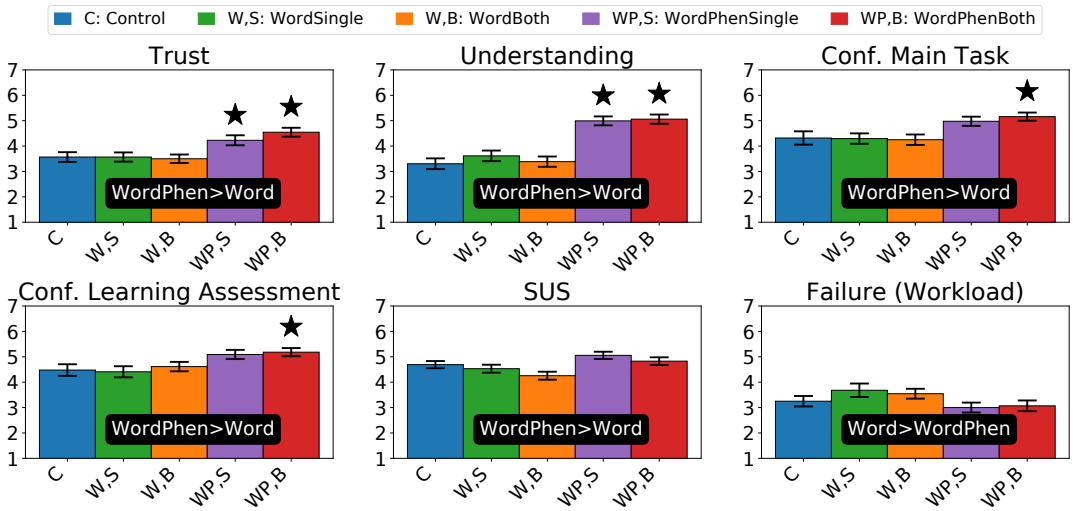

Figure 4: Effects of different interface conditions on participants' *perceptions* with means and standard error. The star mark highlights interface conditions with statistically significant effects ($p < .05$) compared to the Control condition from the Dunnett's test. The text box highlights statistically significant effects ($p < .05$) from two-way ANOVA.

unintuitive to participants. That is, without additional support, participants did not understand the underlying phenomena associated with the most predictive words.

Second, the two-way ANOVA between Word and WordPhen revealed that compared to only highlighting predictive words, providing phenomena-based explanations significantly improved participants' objective and subjective learning of the deception detection task. When the AI prediction aligned with participants' judgment, phenomena-based explanations might have nudged participants to think critically and even changed their mind—participants exhibited a lower tendency to simply stick with the AI's prediction. More importantly, phenomena-based explanations helped participants better recognize correct AI predictions that they might have otherwise rejected, leading to significantly lower under-reliance that

participants eventually achieved. Finally, phenomena-based explanations improved participants' perceptions of the AI and increased their confidence when making judgments. All these benefits came without an increase in workload.

Finally, the two-way ANOVA between SINGLE and BOTH revealed that compared to only providing explanations for the predicted label, providing explanations for *both* labels led to a significant increase in participants' objective learning.

Design Implications: Our study results provide several design implications for future XAI systems.

First, **XAI research should aim to align machine-generated explanations with human intuition.** Our study results found that only providing predictive words had no significant impacts compared to showing model predictions. This suggests that although the feature importance explanations were technically accurate (i.e., generated through regression coefficients), they can sometimes be incomprehensible to users. In contrast, participants who had access to phenomena-based explanations reported significantly better outcomes in learning, appropriate reliance, and perceptions of the AI. Our goal of elevating word-based explanations to phenomena-based explanations resonates with recent trends in human-centered XAI [23, 37, 49, 73]. However, human-centered design has sometimes been ignored in XAI research that has primarily focused on technical innovation [51, 62, 69]. When end-users are confused by feature importance explanations, they are left to speculate about reasons for the feature's predictiveness on their own, without tools to support this process [61, 66]. Thus, future XAI research should consider whether an explanation is comprehensible to end-users and provide additional clarifications when necessary.

Second, **LLMs show promising capabilities for generating human-understandable explanations**. In this study, we developed and evaluated an LLM-based tool to explain predictive but unintuitive words in deception detection. This tool has novelty in three aspects compared to those developed in previous research. First, unlike tools that generated explanations for unintuitive words using training data [42, 61], our tool effectively conjectured phenomena directly from predictive words without looking at any examples. Second, rather than augmenting predictive words with other words nearby [6, 34, 35, 72], our tool leveraged an LLM to generate natural language explanations. This is useful especially when local contexts are insufficient to explain unintuitive words and external knowledge is required instead. Third, unlike approaches that merely *translate* feature importance explanations into coherent natural language [25, 52, 52, 56, 71, 82], our tool *transforms* low-level word features into higher-level phenomena-based features. To validate these conjectured phenomena, we conducted an algorithmic evaluation using a second LLM (Section 4.1). This conjecture-then-validate pipeline is a novel contribution of our work and might generalize to other predictive tasks.

Third, **phenomena-based explanations can enhance human learning in complex tasks**. Our RQ1 results showed that participants who had access to phenomena-based explanations achieved better knowledge transfer than those who only had predictive words. This result is not surprising—while the word "Chicago" is the strongest signal for deceptive Chicago hotel reviews, it does not generalize to hotel reviews in other cities. In contrast, its corresponding phenomenon (i.e., "deceptive reviews tend to name-drop hotel and city names") offers a more generalizable insight. Notably, participants who interacted with both-sided, phenomena-based explanations (i.e., the WORDPHENBOTH condition) achieved 74.2% accuracy in the learning assessment. This is comparable to the deception detection performance reported in a prior study when users had real-time assistance from multiple tools [45]. This finding suggests that in addition to *assisting* humans in decision-making [2, 11, 44, 45, 59], future XAI systems could also prioritize *teaching* humans about complex tasks. Previous research also highlighted the potential use of XAI techniques to train novices on tasks like diagnosing diseases [68] and playing Chess [20].

Limitations and Future Work: Our study has several limitations. First, we only experimented with `GPT-4o` to generate phenomena for words predictive of hotel reviews being genuine or fake. To test the generalizability of our approach, future work should consider other LLMs and predictive tasks. Second, `GPT-4o` was prompted to conjecture the underlying phenomena from predictive words in a single pass. Future work could explore providing supplementary auxiliary information like example reviews for predictive words and iteratively refining the LLM-conjectured phenomena with human feedback. Finally, participants' learning outcomes were measured solely by their judgment accuracy in the learning assessment. A future study could use both a pre- and post-test to capture prior knowledge and learning. Additionally, a future study could measure *rate* of learning as participants interact with the AI system.

6 Conclusion

Feature importance explanation is a popular XAI technique to explain machine learning model predictions. However, text features deemed predictive by machines may appear unintuitive to end-users. Such unintuitive words often represent certain underlying language phenomena that cannot be directly observed from their surrounding contexts. In this paper, we used deception detection as a case study. We developed a novel LLM-based tool to explain why a word is predictive of genuine or deceptive reviews. An LLM was prompted to conjecture underlying phenomena directly from predictive words. We first validated the LLM-generated explanations to be non-hallucinated through an algorithmic evaluation. Then, we conducted a crowdsourced study ($N = 220$) to investigate how unintuitive words and LLM-generated explanations influence end-users in decision-making. We found that: (1) compared to model predictions alone, providing predictive words did not lead to any significant effects; (2) supplementing predictive words with LLM-conjectured phenomena significantly improved participants' learning of the task, appropriate reliance on the AI system, and overall perceptions. Our study results highlighted the importance of providing additional explanations for predictive features that do not make immediate sense. Furthermore, our LLM-based tool and the conjecture-then-validate pipeline have the potential of explaining unintuitive feature importance explanations in other tasks.

Ethics Statements

Ethical considerations statement: The study was reviewed and approved by our Institutional Review Board (IRB). During the study,

participants were asked to complete the task at their own comfortable pace. The study design, which included eight trials in the main task and six trials in the learning assessment, was informed by a pilot study to ensure sufficient data collection without causing cognitive overload. Participants were fully informed that the study involved evaluating pre-labeled deceptive hotel reviews with *no actual deception*. The study used a between-subjects design. This was partly done to help prevent a "spill-over effect" between interface conditions within a single participant.

Adverse impact statement: In our study, we leverage a large language model (LLM) to explain why a word is predictive of genuine or deceptive hotel reviews. While our study highlighted hidden yet meaningful psycholinguistic phenomena in hotel reviews, an adverse impact could arise if these insights are misused to craft deceptive reviews that closely resemble genuine ones. Reviews deliberately produced to avoid including deceptive traits and emphasize genuine traits could become significantly harder to detect. We call for responsible usage of AI technologies and future research to mitigate these risks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[4] Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10, 3 (2006), 214–234.

[5] Sebastian Bordt, Ben Lengerich, Harsha Nori, and Rich Caruana. 2024. Data Science with LLMs and Interpretable Models. *arXiv preprint arXiv:2402.14474* (2024).

[6] Vadim Borisov and Gjergji Kasneci. 2022. Relational Local Explanations. *arXiv preprint arXiv:2212.12374* (2022).

[7] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[10] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[11] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[12] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[13] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.

[14] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[15] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[16] Cheng-Han Chiang and Hung-Yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15607–15631.

[17] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[18] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

[19] Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems* 33 (2020), 17212–17223.

[20] Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 510–518.

[21] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* (2017). https://arxiv.org/abs/1702.08608

[22] Charles W Dunnett. 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* 50, 272 (1955), 1096–1121.

[23] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI conference on human factors in computing systems extended abstracts*. 1–7.

[24] Donald E Farrar and Robert R Glauber. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics* (1967), 92–107.

[25] Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023. Saliency Map Verbalization: Comparing Feature Importance Representations from Model-free and Instruction-based Methods. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*. 30–46.

[26] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.

[27] Ronald A Fisher. 1949. The design of experiments. (1949).

[28] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).

[30] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[31] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[32] Samuel Himmelfarb. 1975. What do you do when the control group doesn't fit into the factorial design? *Psychological Bulletin* 82, 3 (1975), 363.

[33] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4194–4213.

[34] Alon Jacovi, Hendrik Schuff, Heike Adel, Ngoc Thang Vu, and Yoav Goldberg. 2023. Neighboring Words Affect Human Interpretation of Saliency Explanations. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11816–11833.

[35] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research* 22, 104 (2021), 1–54.

[36] Gary M Kaufmann and Terry A Beehr. 1986. Interactions between job stressors and social support: Some counterintuitive results. *Journal of applied psychology* 71, 3 (1986), 522.

[37] Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence* 7 (2024), 1456486.

[38] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.

[39] Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. 193–203.

[40] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[41] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Are Large Language Models Post Hoc Explainers? *arXiv preprint arXiv:2310.05797* (2023).

[42] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[43] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1369–1385.

[44] 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*. 1–13.

[45] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[46] Benjamin J Lengerich, Sebastian Bordt, Harsha Nori, Mark E Nunnally, Yin Aphinyanaphongs, Manolis Kellis, and Rich Caruana. 2023. LLMs understand glass-box models, discover surprises, and suggest repairs. *arXiv preprint arXiv:2308.01157* (2023).

[47] Jiwei Li, Myle Ott, and Claire Cardie. 2013. Identifying manipulated offerings on review portals. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1933–1942.

[48] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1566–1576.

[49] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[50] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–31.

[51] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[52] David Martens, James Hinns, Camille Dams, Mark Vergouwen, and Theodoros Evgeniou. 2023. Tell me a story! narrative-driven xai with large language models. *arXiv preprint arXiv:2309.17057* (2023).

[53] Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C Wallace. 2023. CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 8477–8494.

[54] Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*. 497–501.

[55] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 309–319.

[56] Bo Pan, Zhen Xiong, Guanchen Wu, Zheng Zhang, Yifei Zhang, and Liang Zhao. 2024. TAGExplainer: Narrating Graph Explanations for Text-Attributed Graph Learning Models. *arXiv preprint arXiv:2410.15268* (2024).

[57] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[58] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[59] Jiaming Qu, Jaime Arguello, and Yue Wang. 2021. A Study of Explainability Features to Scrutinize Faceted Filtering Results. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1498–1507.

[60] Jiaming Qu, Jaime Arguello, and Yue Wang. 2023. Understanding the Cognitive Influences of Interpretability Features on How Users Scrutinize Machine-Predicted Categories. In *ACM SIGIR Conference On Human Information Interaction And Retrieval*.

[61] Jiaming Qu, Jaime Arguello, and Yue Wang. 2024. Why is" Problems" Predictive of Positive Sentiment? A Case Study of Explaining Unintuitive Features in Sentiment Classification. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 161–172.

[62] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*. 1135–1144.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[64] John TE Richardson. 2011. Eta squared and partial eta squared as measures of effect size in educational research. *Educational research review* 6, 2 (2011), 135–147.

[65] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[66] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 611–636.

[67] Rita Sevastjanova and Mennatallah El-Assady. 2022. Beware the rationalization trap! when language model explainability diverges from our mental models of language. *arXiv preprint arXiv:2207.06897* (2022).

[68] Philipp Spitzer, Niklas Kühl, Marc Goutier, Manuel Kaschura, and Gerhard Satzger. 2024. Transferring Domain Knowledge with (X) AI-Based Learning Systems. *arXiv preprint arXiv:2406.01329* (2024).

[69] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[70] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *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)*. 4149–4158.

[71] Sule Tekkesinoglu and Lars Kunze. 2024. From Feature Importance to Natural Language Explanations Using LLMs with RAG. *arXiv preprint arXiv:2407.20990* (2024).

[72] Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems* 33 (2020), 6147–6159.

[73] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[74] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*. 1–11.

[75] Xinru Wang, Chen Liang, and Ming Yin. 2023. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets.. In *IJCAI*. 3076–3084.

[76] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[77] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems* 31 (2018).

[78] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3914–3923.

[79] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[80] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.

[81] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291.

[82] Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. 2024. LLMs for XAI: Future Directions for Explaining Explanations. *arXiv preprint arXiv:2405.06064* (2024).

A Preliminary Investigation Results

As described in Section 3.2, we trained logistic regression classifiers using a unigram TF-IDF representation on a dataset of Chicago hotel reviews. We identified predictive words through regression coefficients. For instance, words like "Chicago", "hotel" and "luxury" are predictive of deceptive reviews, and words like "location", "floor" and "small" are predictive of genuine reviews. Our goal was to explain such salient but unintuitive word-label relations.

During our preliminary investigation, we experimented with the contextual pattern mining algorithm developed in [61]. Essentially, to explain why a word is predictive of a label, the algorithm iteratively adds surrounding words to the word—which generates new phrases—and leverages a zero-shot classifier to estimate whether such a phrase is indicative of the target label. Table A.1 summarizes the most frequent and informative patterns identified by the contextual pattern mining algorithm for each predictive word. In this case, a BART-based zero-shot classifier believes that each pattern in Table A.1 is sufficiently predictive of the associated label ("deceptive" or "genuine"). However, the predictions given by such a zero-shot classifier is unreliable. This is because the mechanism of the zero-shot classifier is to find the label that is most semantically similar to the input text, which is ill-suited for a task like deception detection. It is evident that few of the contextual patterns in Table A.1 are practically helpful at explaining why a word is predictive of genuine or deceptive reviews. These results suggest that unlike unintuitive words in sentiment analysis that represent context-related phenomena (e.g., the word "problems" predicts positive sentiment because of colloquial expressions like "without any problems" or negations like "no problems"), those in deception detection can **represent underlying phenomena that go beyond local context**. This has inspired us to develop a novel LLM-based approach that can conjecture reasons beyond local context.

B Task Allocation

Figure B.1 illustrates our task allocation. Each participant was *randomly* assigned to one interface condition and one batch. The reviews within each batch were presented in *random order* during both the main task and the learning assessment (i.e., the order shown in the figure is for illustrative purposes only).

C Significance Testing Results with Effect Sizes

Table C.1, Table C.2 and Table C.3 summarize the significance testing results for all measures (RQ1-RQ3) with exact p-values and corresponding effect sizes. For Dunnett's tests that compared each treatment condition to the control condition, we report Cohen's d [18] as the effect size. For two-way ANOVAs that examined the four treatment conditions, we report partial eta squared (η_p^2) [64] as the effect size.

Table A.1: Examples of contextual patterns found for predictive but unintuitive words in deception detection.

word	patterns within deceptive reviews	patterns within genuine reviews
Chicago	Chicago water	Chicago really
hotel	the hotel is located	hotel as
luxury	for a luxury hotel this place did not	cleanliness and luxury are top notch
location	location but	was excellent and the location of the hotel
floor	lower floor	ceiling to floor with the room
small	small and	the small intimate lobby area

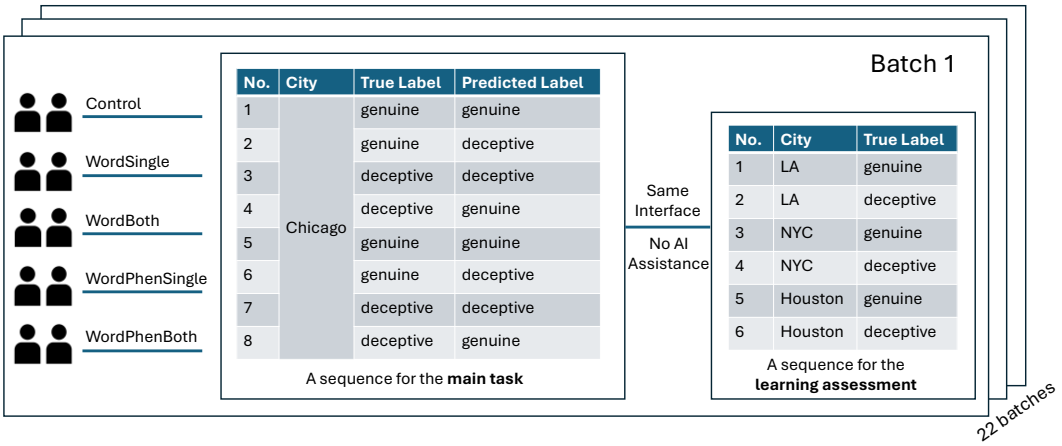

Figure B.1: Task allocation.

Table C.1: Significance-testing results for RQ1 (learning) measures with exact p-values and effect sizes.

Measures	Dunnett's tests	Two-way ANOVAs
Judgment Accuracy	WordSingle vs. Control: $p = 0.354, d = 0.326$ WordBoth vs. Control: $p = 0.054, d = 0.518$ WordPhenSingle vs. Control: $p = 0.027, d = 0.575$ WordPhenBoth vs. Control: $p < 0.001, d = 0.978$	type: $F(1, 172) = 5.596,\ p = 0.019,\ \eta_p^2 = 0.030$ side: $F(1, 172) = 3.928,\ p = 0.049,\ \eta_p^2 = 0.020$ type*side: $F(1, 172) = 0.495,\ p = 0.483,\ \eta_p^2 = 0.003$
Perceived Learning	WordSingle vs. Control: $p = 0.279, d = 0.355$ WordBoth vs. Control: $p = 0.903, d = 0.144$ WordPhenSingle vs. Control: $p = 0.002, d = 0.744$ WordPhenBoth vs. Control: $p < 0.001, d = 1.039$	type: $F(1, 172) = 19.684,\ p < 0.001,\ \eta_p^2 = 0.100$ side: $F(1, 172) = 0.083,\ p = 0.774,\ \eta_p^2 = 0.001$ type*side: $F(1, 172) = 3.055,\ p = 0.082,\ \eta_p^2 = 0.020$

Table C.2: Significance testing results for RQ2 (reliance) measures with exact p-values and effect sizes.

Measures	Dunnett's tests	Two-way ANOVAs
Change toward AI	WordSingle vs. Control: $p = 0.324, d = -0.337$ WordBoth vs. Control: $p = 0.921, d = -0.136$ WordPhenSingle vs. Control: $p = 0.520, d = -0.271$ WordPhenBoth vs. Control: $p = 0.956, d = -0.114$	type: $F(1, 172) = 0.086,\ p = 0.769,\ \eta_p^2 = 0.001$ side: $F(1, 172) = 1.428,\ p = 0.234,\ \eta_p^2 = 0.008$ type*side: $F(1, 172) = 0.022,\ p = 0.882,\ \eta_p^2 = 0.000$
Stick with AI	WordSingle vs. Control: $p = 0.087, d = 0.477$ WordBoth vs. Control: $p = 0.065, d = 0.502$ WordPhenSingle vs. Control: $p = 0.462, d = 0.289$ WordPhenBoth vs. Control: $p = 1.000, d = -0.023$	type: $F(1, 172) = 6.760,\ p = 0.010,\ \eta_p^2 = 0.040$ side: $F(1, 172) = 1.097,\ p = 0.297,\ \eta_p^2 = 0.006$ type*side: $F(1, 172) = 1.513,\ p = 0.220,\ \eta_p^2 = 0.009$
Under-Reliance	WordSingle vs. Control: $p = 0.984, d = -0.085$ WordBoth vs. Control: $p = 0.481, d = -0.283$ WordPhenSingle vs. Control: $p = 0.043, d = -0.537$ WordPhenBoth vs. Control: $p = 0.030, d = -0.566$	type: $F(1, 172) = 5.830,\ p = 0.017,\ \eta_p^2 = 0.030$ side: $F(1, 172) = 0.552,\ p = 0.459,\ \eta_p^2 = 0.003$ type*side: $F(1, 172) = 0.310,\ p = 0.578,\ \eta_p^2 = 0.002$
Over-Reliance	WordSingle vs. Control: $p = 0.205, d = -0.391$ WordBoth vs. Control: $p = 0.607, d = -0.244$ WordPhenSingle vs. Control: $p = 0.128, d = -0.440$ WordPhenBoth vs. Control: $p = 0.099, d = -0.464$	type: $F(1, 172) = 0.815,\ p = 0.368,\ \eta_p^2 = 0.005$ side: $F(1, 172) = 0.168,\ p = 0.682,\ \eta_p^2 = 0.001$ type*side: $F(1, 172) = 0.330,\ p = 0.566,\ \eta_p^2 = 0.002$
Appropriate Reliance	WordSingle vs. Control: $p = 0.304, d = 0.345$ WordBoth vs. Control: $p = 0.262, d = 0.363$ WordPhenSingle vs. Control: $p = 0.007, d = 0.672$ WordPhenBoth vs. Control: $p = 0.004, d = 0.708$	type: $F(1, 172) = 4.961,\ p = 0.027,\ \eta_p^2 = 0.030$ side: $F(1, 172) = 0.033,\ p = 0.857,\ \eta_p^2 = 0.000$ type*side: $F(1, 172) = 0.004,\ p = 0.952,\ \eta_p^2 = 0.000$

Table C.3: Significance testing results for RQ3 (perceptions) measures with exact p-values and effect sizes.

Measures	Dunnett's tests	Two-way ANOVAs
Trust	WordSingle vs. Control: $p = 1.000, d = 0.000$ WordBoth vs. Control: $p = 0.997, d = -0.056$ WordPhenSingle vs. Control: $p = 0.043, d = 0.538$ WordPhenBoth vs. Control: $p < 0.001, d = 0.798$	type: $F(1, 172) = 21.815,\ p < 0.001,\ \eta_p^2 = 0.110$ side: $F(1, 172) = 0.469,\ p = 0.494,\ \eta_p^2 = 0.003$ type*side: $F(1, 172) = 1.121,\ p = 0.291,\ \eta_p^2 = 0.007$
Understanding	WordSingle vs. Control: $p = 0.638, d = 0.235$ WordBoth vs. Control: $p = 0.995, d = 0.063$ WordPhenSingle vs. Control: $p < 0.001, d = 1.279$ WordPhenBoth vs. Control: $p < 0.001, d = 1.331$	type: $F(1, 172) = 61.250,\ p < 0.001,\ \eta_p^2 = 0.260$ side: $F(1, 172) = 0.166,\ p = 0.684,\ \eta_p^2 = 0.001$ type*side: $F(1, 172) = 0.574,\ p = 0.450,\ \eta_p^2 = 0.003$
Conf. Main Task	WordSingle vs. Control: $p = 1.000, d = -0.017$ WordBoth vs. Control: $p = 0.998, d = -0.049$ WordPhenSingle vs. Control: $p = 0.086, d = 0.477$ WordPhenBoth vs. Control: $p = 0.017, d = 0.609$	type: $F(1, 172) = 17.333,\ p < 0.001,\ \eta_p^2 = 0.090$ side: $F(1, 172) = 0.127,\ p = 0.722,\ \eta_p^2 = 0.001$ type*side: $F(1, 172) = 0.354,\ p = 0.553,\ \eta_p^2 = 0.002$
Conf. Assessment	WordSingle vs. Control: $p = 0.998, d = -0.052$ WordBoth vs. Control: $p = 0.968, d = 0.103$ WordPhenSingle vs. Control: $p = 0.098, d = 0.465$ WordPhenBoth vs. Control: $p = 0.044, d = 0.534$	type: $F(1, 172) = 10.819,\ p = 0.001,\ \eta_p^2 = 0.060$ side: $F(1, 172) = 0.604,\ p = 0.438,\ \eta_p^2 = 0.004$ type*side: $F(1, 172) = 0.089,\ p = 0.765,\ \eta_p^2 = 0.001$
SUS	WordSingle vs. Control: $p = 0.871, d = -0.159$ WordBoth vs. Control: $p = 0.137, d = -0.433$ WordPhenSingle vs. Control: $p = 0.262, d = 0.363$ WordPhenBoth vs. Control: $p = 0.919, d = 0.136$	type: $F(1, 172) = 12.747,\ p = 0.001,\ \eta_p^2 = 0.070$ side: $F(1, 172) = 2.688,\ p = 0.103,\ \eta_p^2 = 0.020$ type*side: $F(1, 172) = 0.025,\ p = 0.876,\ \eta_p^2 = 0.000$
Mental demand	WordSingle vs. Control: $p = 0.886, d = -0.152$ WordBoth vs. Control: $p = 0.986, d = 0.083$ WordPhenSingle vs. Control: $p = 1.000, d = -0.014$ WordPhenBoth vs. Control: $p = 0.886, d = -0.152$	type: $F(1, 172) = 0.105,\ p = 0.746,\ \eta_p^2 = 0.001$ side: $F(1, 172) = 0.105,\ p = 0.746,\ \eta_p^2 = 0.001$ type*side: $F(1, 172) = 1.564,\ p = 0.213,\ \eta_p^2 = 0.009$
Physical demand	WordSingle vs. Control: $p = 0.999, d = 0.036$ WordBoth vs. Control: $p = 0.997, d = 0.053$ WordPhenSingle vs. Control: $p = 0.867, d = 0.160$ WordPhenBoth vs. Control: $p = 0.999, d = 0.036$	type: $F(1, 172) = 0.116,\ p = 0.734,\ \eta_p^2 = 0.001$ side: $F(1, 172) = 0.116,\ p = 0.734,\ \eta_p^2 = 0.001$ type*side: $F(1, 172) = 0.206,\ p = 0.651,\ \eta_p^2 = 0.001$
Temporal demand	WordSingle vs. Control: $p = 0.581, d = -0.252$ WordBoth vs. Control: $p = 0.971, d = 0.101$ WordPhenSingle vs. Control: $p = 0.950, d = -0.118$ WordPhenBoth vs. Control: $p = 0.420, d = -0.302$	type: $F(1, 172) = 0.789,\ p = 0.376,\ \eta_p^2 = 0.005$ side: $F(1, 172) = 0.308,\ p = 0.580,\ \eta_p^2 = 0.002$ type*side: $F(1, 172) = 3.156,\ p = 0.077,\ \eta_p^2 = 0.020$
Failure	WordSingle vs. Control: $p = 0.434, d = 0.298$ WordBoth vs. Control: $p = 0.741, d = 0.204$ WordPhenSingle vs. Control: $p = 0.835, d = -0.173$ WordPhenBoth vs. Control: $p = 0.939, d = -0.125$	type: $F(1, 172) = 6.884,\ p = 0.010,\ \eta_p^2 = 0.040$ side: $F(1, 172) = 0.024,\ p = 0.878,\ \eta_p^2 = 0.000$ type*side: $F(1, 172) = 0.214,\ p = 0.644,\ \eta_p^2 = 0.001$
Effort	WordSingle vs. Control: $p = 0.988, d = -0.078$ WordBoth vs. Control: $p = 0.483, d = -0.282$ WordPhenSingle vs. Control: $p = 0.483, d = -0.282$ WordPhenBoth vs. Control: $p = 0.875, d = -0.157$	type: $F(1, 172) = 0.066,\ p = 0.798,\ \eta_p^2 = 0.000$ side: $F(1, 172) = 0.066,\ p = 0.798,\ \eta_p^2 = 0.000$ type*side: $F(1, 172) = 1.159,\ p = 0.283,\ \eta_p^2 = 0.007$
Frustration	WordSingle vs. Control: $p = 0.933, d = -0.129$ WordBoth vs. Control: $p = 0.994, d = 0.065$ WordPhenSingle vs. Control: $p = 0.399, d = -0.310$ WordPhenBoth vs. Control: $p = 0.561, d = -0.258$	type: $F(1, 172) = 2.712,\ p = 0.101,\ \eta_p^2 = 0.020$ side: $F(1, 172) = 0.644,\ p = 0.423,\ \eta_p^2 = 0.004$ type*side: $F(1, 172) = 0.216,\ p = 0.643,\ \eta_p^2 = 0.001$