

# Highlight All the Phrases: Enhancing LLM Transparency through Visual Factuality Indicators

Hyo Jin Do<sup>1\*</sup>, Rachel Ostrand<sup>2\*</sup>, Werner Geyer<sup>1</sup>,  
Keerthiram Murugesan<sup>2</sup>, Dennis Wei<sup>3</sup>, Justin Weisz<sup>2</sup>

<sup>1</sup>IBM Research, Cambridge, MA, USA,

<sup>2</sup>IBM Research, Yorktown Heights, NY, USA,

<sup>3</sup>IBM Research, San Jose, CA, USA

hjdo@ibm.com, rachel.oststrand@ibm.com, werner.geyer@us.ibm.com,  
keerthiram.murugesan@ibm.com, dwei@us.ibm.com, jweisz@us.ibm.com

## Abstract

Large language models (LLMs) are susceptible to generating inaccurate or false information, often referred to as “hallucinations” or “confabulations.” While several technical advancements have been made to detect hallucinated content by assessing the factuality of the model’s responses, there is still limited research on how to effectively communicate this information to users. To address this gap, we conducted two scenario-based experiments with a total of 208 participants to systematically compare the effects of various design strategies for communicating factuality scores by assessing participants’ ratings of trust, ease in validating response accuracy, and preference. Our findings reveal that participants preferred and trusted a design in which all phrases within a response were color-coded based on factuality scores. Participants also found it easier to validate accuracy of the response in this style compared to a baseline with no style applied. Our study offers practical design guidelines for LLM application developers and designers, aimed at calibrating user trust, aligning with user preferences, and enhancing users’ ability to scrutinize LLM outputs.

## 1 Introduction

Large language models (LLMs) can generate factually incorrect or fabricated information that appears plausible and is presented with confidence – a phenomenon widely known as “hallucination” (Ji et al. 2023). This behavior is also described as “confabulation” (Smith, Greaves, and Panch 2023), or more bluntly, “bullshit” (Hicks, Humphries, and Slater 2024)<sup>1</sup>. The presence of these hallucinations in LLM outputs, coupled with difficulty in detecting them and users’ tendency to over-trust LLMs (Bo, Wan, and Anderson 2024; Kim et al. 2024), has led to several high-profile incidents. For example, lawyers have been reprimanded by judges for referencing hallucinated case law (Sloan 2023), new products have been rapidly shelved due to hallucinated scientific

references (Ryan 2022), news outlets have had to issue corrections to articles written with AI assistance (Sato and Roth 2023), and company share prices have dropped after a hallucination caused a blunder during a new product demo (Howell 2023). In response, some communities have prohibited the use of LLM-generated content to safeguard against the inclusion of hallucinated information, such as StackOverflow, an online Q&A forum (Stack Overflow 2025).

Researchers are actively exploring ways to mitigate hallucinations by improving datasets and employing techniques such as reinforcement learning with human feedback (Ouyang et al. 2022; Ji et al. 2023) and retrieval-augmented generation (Lewis et al. 2020; Cai et al. 2022). However, technical advancements alone cannot completely resolve the issue; ultimately, it falls upon end-users to carefully evaluate LLM outputs and be accountable for their use.

Presenting *factuality scores*, which indicate the extent to which a model’s response is truthful to a source document (Kryściński et al. 2020; Laban et al. 2022; Maynez et al. 2020; Zhou et al. 2023; Chern et al. 2023; Min et al. 2023), presents a promising human-centered solution to help users in evaluating LLM outputs. (Although definitions of “factuality” vary slightly in the field, we use it here to refer to truthfulness with respect to a source document which is considered factual information.) Nevertheless, the best way to communicate factuality information to users remains unclear. As a first step, Leiser et al. (2023) conducted participatory workshops where participants brainstormed design features to help identify hallucinations in LLM outputs. They found that participants desired either numerical factuality indicators (e.g., percentage) or ordinal factuality indicators (e.g., high, medium, low) with visual aids, such as color-coded underlines to differentiate between factual and fictional arguments. However, no studies have systematically compared the effectiveness of different strategies in helping users comprehend the accuracy of the model’s response and calibrate their trust while aligning with their preferences.

Our research aims to identify the most effective strategy for communicating the factuality of an LLM’s response. We address the following research questions:

1. **Trust:** Which designs foster user trust in the model?
2. **Ease of validation:** Which designs facilitate validation

\*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>In this paper, we use the term “hallucination” to encompass all these concepts; while “confabulation” may be more precise, “hallucination” is more broadly recognized.

of the factuality of the model’s response?

### 3. **Preference:** What are the most preferred designs?

We approached these research questions in three phases:

- **Design Exploration** (Section 3): We conducted a design review and a pilot study to evaluate different design options and selected a subset of those designs.
- **Experiment 1: Evaluative Study** (Section 4): We conducted a controlled scenario-based study to evaluate six design strategies for representing factuality scores.
- **Experiment 2: Replication Study** (Section 5): We conducted a conceptual replication study (Derksen and Morawski 2022) to investigate whether the Experiment 1 findings generalize to different scenarios.

In the two experiments, participants were shown a color scale for conveying factuality scores, along with three styles for visualizing factuality scores within an LLM’s response: (1) *highlight-all*, which annotates all linguistic content in the LLM response with varying background colors based on its factuality score, (2) *highlight-threshold*, which annotates only those parts of the LLM response where the factuality score is below a given threshold, and (3) *score*, which shows the numeric factuality score associated with each part of the response. Factuality scores were evaluated at two levels of linguistic granularity – *phrase* and *term* – and the three factuality styles were presented at each level of granularity.

We then conducted two experiments to compare the effects of these design strategies in question-answer scenarios. We investigated their effects on participants’ ratings of trust, ease of evaluating response accuracy, and preference rankings. In both experiments, participants had the highest preference ratings for the *highlight-all* style at a *phrase-level* granularity. Participants found it easier to validate the accuracy of an LLM’s response in this style compared to a *baseline* in which no style was applied. Moreover, displaying factuality scores led participants to increase their trust. Our paper makes three contributions:

1. We explore the design space for presenting factuality scores to users and identify a set of promising approaches for in-depth evaluation based on user feedback.
2. We find that design strategies significantly impact ratings of trust, ease of accuracy validation, and preference.
3. We offer practical guidance on how to effectively communicate factuality within the user interface of LLM-based applications.

## 2 Related Work

### 2.1 LLM Hallucination and Factuality Detection

The widespread usage of LLMs in society has highlighted their risks and limitations. Notably, these models can generate text that appears plausible at first glance but actually contains factually incorrect information, a phenomenon referred to as “hallucination” or “confabulation.” In contrast, *factuality* is defined as “truthfulness or the quality of being based on fact” (Ji et al. 2023). A related concept is *faithfulness*, which pertains to how well an LLM-generated response is consistent with the ground truth source. In this study, we assume a

reliable source as our basis for “fact” so that faithfulness has the same meaning as factuality (Maynez et al. 2020). If the model’s response aligns with the information from a reliable source, it is factually correct.

Hallucinations in LLMs stem from various factors such as noisy, biased, and erroneous training data, as well as the model itself. Researchers have addressed data-related issues by establishing ground truth data through human annotators and enhancing model inputs with external knowledge (Ji et al. 2023; Huang et al. 2025; Wang et al. 2024; Honovich et al. 2022). Efforts to enhance the model include refining the architecture (e.g., retrieval augmented generation, known as RAG; Lewis et al. 2020), improving the training process (e.g., reinforcement learning with human feedback, or RLHF; Bai et al. 2022), and post-processing (Chen et al. 2021). Each of these approaches has limitations. For instance, RAG can make statements that are not fully supported by cited sources, and may reduce the diversity of responses (Liu, Zhang, and Liang 2023). RLHF requires significant human labor, time, and emotional toll to refine the model (Metz 2023; Hao and Seetharaman 2023). Given that these algorithmic approaches cannot fully ameliorate problems caused by hallucinations, in the present work, we take a human-centered perspective, emphasizing that it is the responsibility of end-users to carefully evaluate and take accountability for their use of LLM outputs.

As part of the effort to assist end-users in evaluating LLM responses, ongoing research has focused on developing methods to score the factuality of LLM outputs (Laban et al. 2022; Maynez et al. 2020; Zhou et al. 2023; Chern et al. 2023). These methods can either use lexical matching-based metrics relying on hard-coded logic or model-based metrics using neural networks. Lexical matching-based metrics, such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and ROUGE (Lin 2004), measure factuality automatically by assessing the lexical overlap between the source text and the model’s response. In contrast, neural network-based metrics, including BERTscore (Zhang et al. 2019), BLEURT (Sellam, Das, and Parikh 2020), and FActScore (Min et al. 2023), have gained popularity due to their resilience against lexical, syntactic, and semantic differences between the source and the model’s output. Moreover, task-specific model-based metrics such as ANLI (Nie et al. 2020), SummaC (Laban et al. 2022), and QuestEval (Scialom et al. 2021) – which are based on canonical natural language understanding tasks such as natural language inference, abstractive summarization, and question generation – have shown promising directions for evaluating the factuality of the LLM response.

This growing body of research raises new questions for LLM developers and designers on how to effectively *communicate* factuality information to end-users. Currently, there are no established guidelines on which parts of an LLM’s response should be annotated with factuality information, in what visual style, and at what level of linguistic detail. Furthermore, we have limited understanding of how the communication of factuality information mitigates the effects of hallucination and calibrates end-users’ trust.

## 2.2 Calibrating End-User Trust for AI Systems

Successful human-AI collaboration requires a user to modulate their level of trust according to the true reliability of the AI system. This process is known as trust calibration (Lee and See 2004; Wischniewski, Krämer, and Müller 2023; Zhang, Liao, and Bellamy 2020). Miscalibrated trust can result in overreliance, where users accept incorrect AI recommendations, or underreliance, where they reject correct recommendations from the AI.

For example, Kim et al. (2023) asked end-users of an AI-based bird identification app about their trust and trust-related behaviors. The authors found that while people generally trusted the app, they did not accept its outputs as true every time they used the app, and carefully evaluated the outputs. If they were not able to verify the outputs due to lack of domain knowledge, participants disregarded the outputs. This indicates a disparity between user trust and ease of validating accuracy. In this study, we further investigate both concepts in the context of LLM interactions.

In human-human communication, uncertainty may signal transparency and honesty to a conversational partner (van der Bles et al. 2019). Similarly, in LLM interactions, communicating factuality scores or related concepts (e.g., uncertainty, confidence score) of LLM outputs can increase AI transparency (Liao and Vaughan 2023) and improve trust calibration (Zhang, Liao, and Bellamy 2020). Vasconcelos et al. (2023) indicated that highlighting uncertain tokens can assist programmers in identifying potential errors, leading to more focused edits and greater satisfaction among study participants. Weisz et al. (2021) explored a similar technique and found that confidence scores possessed explanatory power, although an analysis by Agarwal et al. (2020) found no correlation between the model’s confidence scores and the presence of actual programming errors. Leiser et al. (2023) found that end-users expressed a desire for visual aids such as color codes to communicate the factuality of LLM responses, which has inspired our study. The present work builds on this prior research and compares the effects of various design strategies for communicating the factuality of a model’s response.

## 3 Design Exploration

To develop our initial designs to present factuality scores for LLM-generated outputs, we drew inspiration from existing commercially available applications, including OpenAI’s WebGPT and Microsoft’s Bing Chat. We observed that factuality information was generally presented through highlights (Yue et al. 2023; Gao et al. 2023; Leiser et al. 2023) or scores (Li et al. 2023). In reviewing these applications, we observed that factuality information could be computed at different levels of granularity, such as at the term, phrase, or whole-response levels. Vasconcelos et al. (2023) reported that users had negative reactions to the whole-response level of granularity and found it unhelpful for identifying errors and felt it was difficult to interpret. Therefore, we did not include the whole-response granularity in the present study.

We conducted a pilot study with ten participants to identify preferred options in the design space. This study led to

the selection of six designs for representing factuality scores, as described in the following sections. We then ran two controlled experiments to evaluate these different designs.

## 4 Experiment 1

### 4.1 Participants

We recruited 104 participants for this experiment, all of whom were employees of IBM, a large multinational technology company. Our goal was to enroll diverse participants in terms of geography, job role, English proficiency, and experience with AI, machine learning, and LLMs. We advertised the participant recruitment widely within the company on internal Slack channels from multiple divisions and geographic regions. The participants were located in 20 different countries, with the largest representation (51%) from the United States. Their job roles encompassed various disciplines, including design, customer service, engineering, sales, research, and human resources. Participants reported a significant range of experience with LLMs, with 18.2% indicating they had never used an LLM and 9.6% reporting daily usage (see Do et al. 2025 Figure A.1(a) for more fine-grained responses).

The experiment was conducted in English, and we strove to recruit participants with varying degrees of English exposure and proficiency to capture the experience of people who interact with LLMs in a non-native language. As such, 56% of participants reported that they were exposed to English from birth, 27% before age 7 (often considered the end of the critical period for learning a language to native-level proficiency; Johnson and Newport 1989), and 17% after age 7. Participants also self-reported their English proficiency on a 7-point Likert scale, with 68% rating themselves at 7 (*native or native-like proficiency*), 19% at 6, 8% at 5, 4% at 4 (*medium*), and 1% at 3. All participants provided written informed consent, and were treated in accordance with guidelines for the ethical treatment of human participants.

### 4.2 Procedure

The experimental instructions told participants to imagine themselves as users of an AI-powered language model, and were shown a Question, a Response, and a Source. Their task was to evaluate different designs for showing the factuality of the model’s Response, based on the provided Source. The Question was explicitly non-technical to allow all participants, regardless of background or expertise, to assess its accuracy using the source information, and asked, “What movies did Beyonce star in and with whom?” The Source was an edited and condensed text pulled from the Wikipedia article about Beyonce, and participants were instructed to assume that the source was factually accurate. The Response was manually written (i.e., not actually generated by a model) to have a mixture of inaccurate statements which contradicted the source text, and accurate statements. Overall, the Response was approximately equal parts accurate and inaccurate information.

Participants were shown several design strategies to evaluate. Each design strategy was presented using the same Question, Response, and Source text, to hold constant the

Question	Reference paragraph from Wikipedia
What movies did Beyonce star in and with whom?	In July 2002, Beyonce continued her acting career playing Foxy Cleopatra alongside Mike Myers in the comedy film, Austin Powers in Goldmember, which spent its first weekend atop the US box office and grossed \$73 million.
AI-generated response	Beyonce released "Work It Out" as the lead single from its soundtrack album which entered the top ten in the UK, Norway, and Belgium.
Beyonce starred in the musical comedy The Fighting Temptations in 2002 and in the documentary film Austin Powers in Goldmember in 2003, alongside Missy Elliott and Foxy Cleopatra, respectively.	In 2003, Beyonce starred opposite Cuba Gooding, Jr., in the musical comedy The Fighting Temptations as Lilly, a single mother whom Gooding's character falls in love with.
	The film received mixed reviews from critics but grossed \$30 million in the U.S. Beyonce released "Fighting Temptation" as the lead single from the film's soundtrack album, with Missy Elliott, MC Lyte, and Free which was also used to promote the film.
	Another of Beyonce's contributions to the soundtrack, "Summertime", fared better on the US charts.

Figure 1: The *baseline* design was shown to participants at the start of the experiment, with no annotations showing factuality.

content and degree of accuracy across different designs. This approach allowed us to reduce the number of variables tested and ensure a more targeted exploration of the design strategies themselves. Participants always saw the *baseline* design first, which had no factuality markup, and displayed the text only for the Question, Response, and Source (see Figure 1).

After being shown the *baseline* design, participants were asked to rate their perceptions about the model and its response on three metrics, using a 7-point Likert scale:

1. *Perceived accuracy*: How accurate do you think this AI-generated response is?
2. *Ease of validation*: With the information presented in this way, how easy is it for you to determine the accuracy of this AI-generated response?
3. *Trust*: With the information presented in this way, how much do you trust the AI system that generated the response?

Following their ratings of the *baseline* design, participants were introduced to the concept of a *factuality score* – a feature that compares linguistic components of the Response against the Source – and that a high factuality score indicated that the response was aligned with the information in the Source, and thus was likely to be correct. The factuality scores in our experiment were created manually, as opposed to using an existing factuality scoring algorithm, by comparing information units in the Response against the Source text, and tweaking wording in the Response to engender a range of factuality scores to display.

Participants were then presented with six design strategies for displaying factuality information on the Response: Three factuality designs, each at two levels of granularity. The three designs incorporated color-coding to show the factuality of individual linguistic units in the response, on a scale ranging from 0 (red) to 1 (green)<sup>2</sup>, shown in Figure 2. The designs were *highlight-all*, in which every part of the response text was highlighted with a color corresponding to the factuality score; *highlight-threshold*, in which only the

sections of the response text with a factuality score below 0.5 were highlighted to signal inaccuracies; and *score*, in which all parts of the response text were tagged with their factuality score, using both color-coded underlines and the numerical factuality score value.

In addition, designs were presented at two levels of *granularity* – *term-level* or *phrase-level* – referring to the size of the text chunks over which the factuality was evaluated. At *phrase-level* granularity, if there was an inaccuracy in one term in a phrase, then the entire phrase would be tagged with a lower factuality score. In contrast, at *term-level* granularity, just that term would be tagged with a lower factuality score, while the other terms in the phrase or sentence would individually be tagged with their own factuality scores. Table 1 shows the six design strategies that users evaluated.

Participants saw the factuality design strategies one at a time, and rated their perceptions on two metrics: *ease of validation* and *trust* (questions 2 and 3 above) on a 7-point Likert scale. Note that users were not asked about perceived accuracy (question 1) for any designs besides the *baseline*, because the wording of the text was identical in all designs.

Participants performed this rating task for each of the three designs at one granularity (*term-level* or *phrase-level*), and then rank-ordered them along with the *baseline* by preference. They then performed the same rating and preference-ranking for the three designs at the other granularity. The three designs within each granularity were presented in a randomized order across participants, and the order of the two granularities was randomized across participants to reduce order effects. Finally, participants responded to demographic and professional experience questions, as reported in Section 4.1.

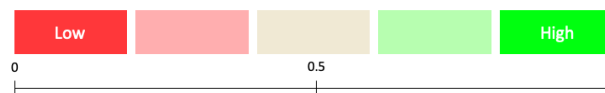


Figure 2: The factuality scale that was presented to Experiment 1 participants. The color scale and corresponding numbers demonstrated the range of possible factuality scores.

<sup>2</sup>This color scale is not ideal from an accessibility standpoint for color-blind users. We suggest modifying the color endpoints or adding shading information for systems deployed on a larger scale.

Granularity	Highlight-all	Highlight-threshold	Score
Term	Beyonce starred in the musical comedy The Fighting Temptations in 2002 and in the documentary film Austin Powers in Goldmember in 2003, alongside Missy Elliott and Foxy Cleopatra, respectively.	Beyonce starred in the musical comedy The Fighting Temptations in 2002 and in the documentary film Austin Powers in Goldmember in 2003, alongside Missy Elliott and Foxy Cleopatra, respectively.	Beyonce <sup>5</sup> starred <sup>5</sup> in the musical comedy <sup>7</sup> The Fighting Temptations <sup>5</sup> in 2002 <sup>3</sup> and in the documentary film <sup>1</sup> Austin Powers in Goldmember <sup>5</sup> in 2003 <sup>3</sup> , alongside Missy Elliott <sup>5</sup> and Foxy Cleopatra <sup>1</sup> , respectively.
Phrase	Beyonce starred in the musical comedy The Fighting Temptations in 2002 and in the documentary film Austin Powers in Goldmember in 2003, alongside Missy Elliott and Foxy Cleopatra, respectively.	Beyonce starred in the musical comedy The Fighting Temptations in 2002 and in the documentary film Austin Powers in Goldmember in 2003, alongside Missy Elliott and Foxy Cleopatra, respectively.	Beyonce starred in the musical comedy The Fighting Temptations in 2002 <sup>5</sup> and in the documentary film Austin Powers in Goldmember in 2003, <sup>3</sup> alongside Missy Elliott and Foxy Cleopatra, respectively. <sup>1</sup>

Table 1: The set of designs presented to each participant for displaying factuality scores on the model’s response. Each participant saw and rated all six designs, in a randomized order grouped by granularity.

### 4.3 Analysis Methods

Three dependent variables that examined different facets of participants’ opinions of the factuality designs were measured: 7-point Likert scale ratings of (a) trust and (b) ease of validating the response accuracy, and (c) rank-order preferences of the different designs. Analyses used generalized linear mixed-effects models in R (R Core Team 2019), using the *lme4*, *lmerTest*, and *emmeans* packages (Bates et al. 2015; Kuznetsova, Brockhoff, and Christensen 2017; Lenth 2020), with separate models for each dependent variable.

We first assessed how each design strategy compared with the *baseline* regarding ratings of trust, ease of validation, and preference. The statistical models included the within-subjects categorical independent variable Design Type, which had seven levels and was treatment-coded, with the *baseline* set as the reference level. This analysis allows for the comparison of the rating of each individual design strategy against the rating given for the *baseline* design. Two separate models were run, one for each of the two ratings – trustworthiness and ease of validating accuracy – both of which were continuous. The model’s random effect structure included the levels of Design Strategy within Participant ID, with random effects which accounted for less than 1% of the model’s variance removed in order to aid convergence. Following the full model, pairwise contrasts were conducted to explore comparisons between every pair of Design Strategy levels, with *p*-values corrected for multiple comparisons using the Tukey correction.

Second, an omnibus linear mixed-effects model with the factors Granularity (term, phrase) x Design Type (highlight-all, highlight-threshold, score) was conducted, including only the six markup design strategies (i.e., excluding the *baseline* design), to investigate rating differences across designs at the level of the factors Design Type and Granularity.

Finally, a cumulative link mixed model with participants’ preference ranking as the dependent variable was conducted. Bartlett’s test for homogeneity of variances indicated that the variances were not significantly different across conditions for the dependent variables.

### 4.4 Results

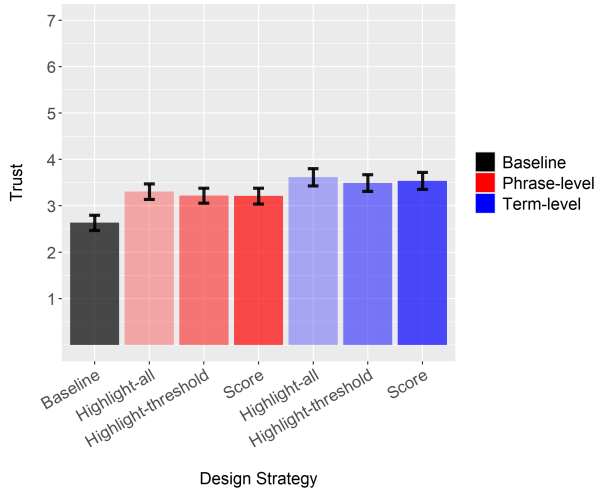
**Trust** The first model compared participants’ ratings of trustworthiness for each of the design strategies against the

*baseline* as the reference level. As can be seen in Figure 3(a), all six designs were rated as significantly more trustworthy than the *baseline*, suggesting that presenting factuality information using any of the markup methods increased participants’ trust in the model. See Table 2 for the detailed statistical results. Post-hoc pairwise comparisons between each pair of designs revealed no additional significant differences after correction for multiple comparisons.

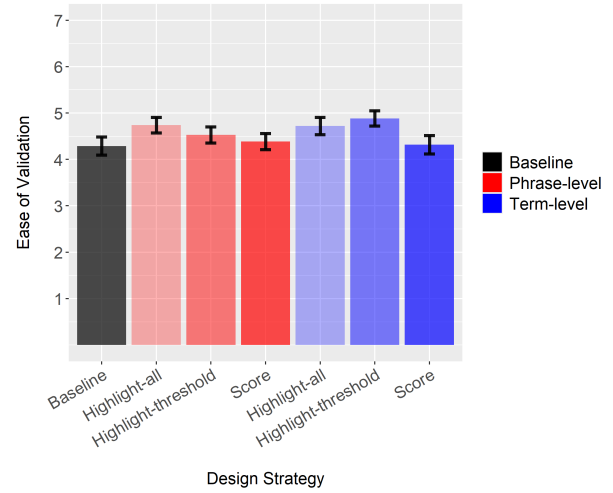
Next, we conducted the omnibus model of Granularity x Design Type. This revealed a main effect of Granularity ( $F(1, 103) = 6.30, p = .01$ ), with *term-level* granularity designs ( $M = 3.54$ ) rated higher than *phrase-level* designs ( $M = 3.25$ ). There was no main effect of Design Type ( $F(2, 106) = 1.27$ ) or interaction ( $F(2, 309) < 1$ ).

As an exploratory post-hoc analysis, we investigated whether participants’ rating of the accuracy of the model’s response when presented with the *baseline* design (question (1) in Section 4.2) affected how much they trusted the model when it subsequently displayed factuality scores. For visualization purposes, we categorized participants into two groups: The *low baseline accuracy* group, which was defined as those participants who rated the model’s response accuracy in the *baseline* design at or below 4 (the midpoint of the rating scale;  $N = 87$ ), and the *high baseline accuracy* group, who rated the *baseline* response accuracy as 5 or higher ( $N = 17$ ). As can be seen in Figure 4(a), participants who rated the model’s response at *baseline* with low accuracy (dashed line) also had low trust of the model when viewing the baseline non-marked-up design, but subsequently increased their trust ratings after reviewing the factuality score markups. In contrast, participants who initially felt the model’s response was more accurate (solid line) also had high trust of the model’s response when shown the *baseline* design, but subsequently *decreased* their trust ratings after examining the factuality information.

**Ease of validation** The first model compared participants’ ratings on the ease of assessing the model’s accuracy for each of the designs against the *baseline*. As can be seen in Figure 3(b), of the six design strategies, three were rated as significantly easier to assess the response accuracy compared to the *baseline*: *highlight-all* at phrase-level granularity, and *highlight-all* and *highlight-threshold* at term-level granularity. The other three designs were not significantly



(a) Trust ratings



(b) Ease of validating accuracy ratings

Figure 3: Experiment 1 ratings of each design strategy: (a) trust and (b) ease of validating the accuracy of the model's response.

EXPERIMENT 1					EXPERIMENT 2				
Strategy	Trust		Ease of Validation		Trust	Ease of Validation			
	Mean (SE)	<i>t</i> ( <i>p</i> )	Mean (SE)	<i>t</i> ( <i>p</i> )	Mean (SE)	<i>t</i> ( <i>p</i> )	Mean (SE)	<i>t</i> ( <i>p</i> )	
Baseline	2.63 (0.17)	–	4.29 (0.20)	–	2.75 (0.16)	–	4.81 (0.17)	–	
<i>Phrase</i>									
highlight-all	<b>3.31 (0.17)</b>	4.59 (<.001)	<b>4.74 (0.17)</b>	2.24 (.03)	<b>3.67 (0.17)</b>	5.81 (<.001)	5.03 (0.17)	1.10 (.27)	
highlight-threshold	<b>3.22 (0.16)</b>	4.00 (<.001)	4.53 (0.17)	1.19 (.23)	<b>3.38 (0.17)</b>	4.15 (<.001)	4.58 (0.16)	-1.15 (.25)	
score	<b>3.21 (0.17)</b>	3.93 (<.001)	4.38 (0.17)	0.48 (.63)	<b>3.55 (0.18)</b>	5.29 (<.001)	4.51 (0.18)	-1.48 (.14)	
<i>Term</i>									
highlight-all	<b>3.62 (0.18)</b>	6.57 (<.001)	<b>4.72 (0.19)</b>	2.10 (.04)	<b>3.34 (0.18)</b>	3.84 (<.001)	<u>4.23 (0.20)</u>	-2.77 (.01)	
highlight-threshold	<b>3.49 (0.18)</b>	5.80 (<.001)	<b>4.88 (0.16)</b>	2.96 (.003)	<b>3.18 (0.18)</b>	2.87 (.004)	4.44 (0.19)	-1.82 (.07)	
score	<b>3.54 (0.18)</b>	5.85 (<.001)	4.32 (0.20)	0.13 (.89)	2.94 (0.16)	1.28 (.20)	<u>3.57 (0.19)</u>	-6.16 (<.001)	

Table 2: Trust and ease of validation ratings: means, standard errors (SE), and statistical results across the two experiments. Trust and ease of validation were rated on a 1-7 Likert scale, with 7 as the highest score. *t*- and *p*-values are from the model with the *baseline* set as the reference level. Bolded text indicates ratings that were significantly higher than the *baseline*, and underlined text indicates ratings that were significantly lower than the *baseline*.

different from the *baseline*; see Table 2 for the detailed statistical results. Post-hoc pairwise comparisons between each pair of design strategies revealed no significant differences after correction for multiple comparisons.

Next, the omnibus model investigating Granularity x Design Type revealed a main effect of Design Type ( $F(2, 103) = 4.62, p = .01$ ), with *highlight-all* ( $M = 4.73$ ) and *highlight-threshold* ( $M = 4.71$ ) rated higher than *score* ( $M = 4.35$ ). There was a Granularity x Design Type interaction ( $F(2, 206) = 3.12, p = .05$ ), the locus of which can be seen in Figure 3(b), largely driven by the fact that *highlight-threshold* was rated higher relative to the other designs at the *term-level* but not *phrase-level*. There was no main effect of Granularity ( $F(1, 103) < 1$ ).

**Preference** Participants rank-ordered each of the three designs and the *baseline* within each granularity level. Thus, each ranking response compared four designs, with rank 1 for the most preferred and 4 for the least preferred design. At *phrase-level* granularity, the *highlight-all* design was the most preferred, *score* was second most preferred, *highlight-threshold* was third, and the *baseline* was the least preferred (see Table 3). There was a significant effect of Design Type on the means of preference rankings ( $\chi^2(3) = 103.82, p < .001$ ). Post-hoc pairwise comparisons revealed that participants preferred each of the factuality designs significantly more than the *baseline* ( $p < .001$ ). Similarly, at *term-level* granularity, the *highlight-all* design was the most preferred, *highlight-threshold* was second, *score* was third, and *baseline* was the least preferred. There was a significant ef-



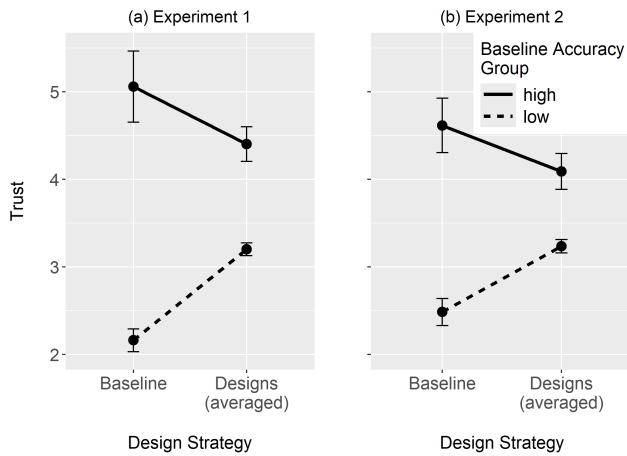


Figure 4: Participants’ trust ratings for the *baseline* design, and for the six design markups averaged together, for (a) Experiment 1 and (b) Experiment 2. Participants were divided into two groups based on whether they rated the model’s response as high accuracy (rating of 5-7) or low accuracy (rating of 1-4) when initially presented with *baseline* design.

fect of Design Type on the means of the preference rankings ( $\chi^2(3) = 78.01, p < .001$ ). Post-hoc pairwise comparisons revealed that participants preferred each of the factuality designs significantly more than the *baseline* ( $p < .001$ ). No other comparisons yielded statistically significant results.

Participants were also asked about their preference between the two types of granularities: 52.9% of participants preferred *phrase-level* granularity, while 26.9% of the participants preferred *term-level* granularity, with 10.6% of participants responding with “don’t know” and 9.6% of participants selecting “other” (e.g. sentence-level, entire response).

## 5 Experiment 2

The goal of Experiment 2 was to assess whether the Experiment 1 results were robust to extension, and to automatize aspects of generating the Response and factuality scores. As such, several aspects of the procedure were modified. First, Experiment 1 investigated factuality designs in a single question-answer scenario; a goal of Experiment 2 was to assess the generalizability of the results to additional domains. Thus Experiment 2 included two new scenarios: a medical question and an HR question. These were selected as domains for which LLMs are increasingly employed for business automation purposes. The two new scenarios are shown in Figure B.1 in Do et al. (2025) Appendix B.

Second, in contrast to Experiment 1 where the Response was manually generated without LLM input, for Experiment 2, we started each scenario with an LLM-generated response as ground-truth (after comparing it to the Source and determining that it was entirely faithful). As this experiment required the presented Response to have a range of factuality scores, we then edited the model’s ground-truth response to contain errors, which became the Response for the experimental scenario. To determine the factuality scores

to display in the designs, we used the Python spaCy library (Honnibal et al. 2020) to calculate the semantic similarity between each term or phrase in the edited Response and the corresponding term or phrase in the ground-truth response, and used that similarity as the factuality score for that linguistic unit. Thus the factuality scores shown in Experiment 2 were generated by automated means rather than human-created as in Experiment 1. We opted to create the Response and factuality scores in this manner (rather than using an entirely-live LLM response and factuality scoring algorithm) as it gave us experimental control over the range of factuality scores that was presented to participants, as the goal of the present experiment was to assess opinions about the factuality *designs*, rather than assessing the accuracy of factuality scoring *algorithms*.

These factuality scores were then used to generate the markups for the six design strategies. To determine the high-light colors, factuality scores were mapped to colors in a modified manner from that in Experiment 1. As the semantic similarity-based factuality scores were biased towards high-magnitude, positive values, the color mapping thresholds were adjusted to have higher resolution at the parts of the scale with the largest concentration of numerical scores. See Do et al. (2025) Figure B.2 for the modified Experiment 2 factuality color scale.

Minor adjustments were made to the linguistic units used for factuality scores: In Experiment 1, we annotated some multi-word noun phrases (e.g., “musical comedy” or “documentary film”) as one term, whereas for Experiment 2, term-level markups were only individual words with the exception of multi-word proper nouns (e.g., disease names). We also added a question asking participants to rank all six designs together, in addition to ranking the designs separately within each level of granularity.

### 5.1 Participants

We recruited another 104 IBM employees via internal Slack channels. A condition of participation was that they had not participated in Experiment 1. Participants’ work locations consisted of 17 unique countries, with the US as the most common (59%). Job roles again spanned a wide array of disciplines, and participants had a range of experience with LLMs, from never to daily usage (see Do et al. 2025 Figure A.1(b)). Participants had varying degrees of English exposure and proficiency, with a very similar distribution as in Experiment 1: 58% of participants reported that they were exposed to English from birth, 19% before age 7, and 23% after age 7. For self-rated proficiency, 71% rated themselves at 7 (*native or native-like proficiency*), 20% at 6, 5% at 5, 3% at 4 (*medium*), and 1% at 3. All participants provided written informed consent and were treated in accordance with the guidelines for ethical treatment of human participants.

### 5.2 Procedure

The procedure for Experiment 2 was largely the same as that of Experiment 1, with a few changes as noted above. Participants were randomly assigned to one of the two scenarios, HR or medical, and rated all of the designs for that scenario (see Do et al. 2025 Figure B.1 for the text of the scenarios).

Strategy	EXPERIMENT 1		EXPERIMENT 2		
	Phrase	Term	Phrase	Term	Overall
Baseline	3.31 (0.12)	3.22 (0.12)	3.34 (0.10)	2.90 (0.13)	5.14 (0.21)
<i>Phrase</i>					
Highlight-all	<b>2.06 (0.10)</b>	–	<b>1.82 (0.09)</b>	–	<b>2.56 (0.15)</b>
Highlight-threshold	2.33 (0.08)	–	2.51 (0.09)	–	3.62 (0.17)
Score	2.31 (0.10)	–	2.34 (0.10)	–	3.56 (0.19)
<i>Term</i>					
Highlight-all	–	<b>2.13 (0.10)</b>	–	<b>2.09 (0.09)</b>	3.94 (0.16)
Highlight-threshold	–	2.21 (0.09)	–	2.16 (0.09)	4.02 (0.19)
Score	–	2.44 (0.10)	–	2.85 (0.10)	5.16 (0.17)

Table 3: Preference rank-order means (standard errors). The most preferred design is bolded. Ranking scores are on an inverse scale; thus lower numbers indicate higher preference. In both experiments, participants ranked designs separately within each granularity level. In Experiment 2, participants additionally ranked designs across both granularity levels in a single question.

There were 55 participants who completed the HR scenario, and 49 participants who completed the medical scenario.

### 5.3 Results

**Trust** First, we ran a model comparing the trustworthiness ratings of each of the design strategies against the *baseline*. All of the designs with the exception of *term-level score* were rated significantly higher on trust than was the *baseline* design (see Figure 5(a)). The mean and standard error of the trust ratings, as well as results from this statistical model comparing each design strategy against the *baseline*, are shown in Table 2. In the post-hoc pairwise comparisons between all pairs of designs, three remained significant after correcting for multiple comparisons: *phrase-level highlight-all* was rated as significantly more trustworthy than both *term-level highlight-threshold* ( $t(240) = 3.09, p = .04$ ) and *term-level score* ( $t(240) = 4.60, p < .001$ ), and *phrase-level score* was rated as significantly more trustworthy than *term-level score* ( $t(437) = 4.02, p = .001$ ).

Next, we conducted an omnibus model investigating the main effects of Granularity and Design Type on the six design strategies, excluding the *baseline*. There was a main effect of Granularity ( $F(1, 103) = 7.73, p = .007$ ), with *phrase-level* granularity ( $M = 3.53$ ) rated higher than *term-level* granularity ( $M = 3.15$ ). There was a main effect of Design Type ( $F(2, 125) = 6.24, p = .003$ ), with *highlight-all* rated the highest ( $M = 3.50$ ), and *highlight-threshold* ( $M = 3.28$ ) and *score* ( $M = 3.25$ ) rated lower. There was also a Granularity x Design Type interaction ( $F(2, 243) = 4.08, p = .02$ ). The locus of this interaction can be seen in Figure 5(a) and Table 2, by the difference in ratings of the *score* design strategy between *phrase-level* and *term-level* Granularity.

To investigate whether participants found the different scenarios (i.e., HR vs. medical scenario) to affect their trust ratings, we additionally ran a Granularity x Design Type x Scenario model on the ratings of the six design strategies (excluding the *baseline*). As with the previous model, there was a main effect of Granularity ( $F(1, 103) = 8.18, p = .005$ ) and Design Type ( $F(2, 140) = 6.24, p = .003$ ), as well as a Granularity x Design Type interaction ( $F(2, 231) = 3.93, p = .02$ ). However, importantly, all effects involving the Sce-

nario factor were not significant (all  $ps > .12$ ), demonstrating that idiosyncrasies of the scenarios themselves were not driving the trust ratings.

We again conducted an exploratory visualization, splitting participants into two groups based on how they had rated the accuracy of the model's response when presented with the initial *baseline* design: high accuracy ( $N = 13$ ) or low accuracy ( $N = 91$ ); see Figure 4(b). As in Experiment 1, participants who thought the model's response had low accuracy (dashed line) also initially rated the model with low trust, but subsequently increased their trust ratings when presented with design information detailing the factuality of the response. In contrast, participants who thought the model's response had high accuracy (solid line) initially rated the *baseline* design with high trust, but decreased their trust ratings once presented with factuality score information.

**Ease of validation** First, we ran a model comparing the rating for each of the six design strategies against the *baseline* rating. Although two design strategies were rated significantly different than the *baseline* design – *term-level highlight-all* and *term-level score* – both were rated lower, that is, as more difficult to validate the accuracy of compared to the *baseline*. See Figure 5(b) and Table 2 for the numerical and statistical results. In the post-hoc pairwise comparisons between all pairs of designs, all of the design strategies, including the *baseline*, were rated as easier to validate the accuracy than *term-level score* (all  $ts > 3.18$ , all  $ps < .03$ ). In addition, *phrase-level highlight-all* was rated significantly higher than *term-level highlight-all* ( $t(259) = 3.83, p = .003$ ).

The omnibus model of Granularity x Design Type for the six design strategies (excluding the *baseline*) revealed a main effect of Granularity ( $F(1, 103) = 12.83, p < .001$ ), with *phrase-level* granularity ( $M = 4.71$ ) rated higher than *term-level* granularity ( $M = 4.08$ ). There was a main effect of Design Type ( $F(2, 124) = 15.03, p < .001$ ), with *highlight-all* rated the highest ( $M = 4.63$ ), *highlight-threshold* ( $M = 4.51$ ) in the middle, and *score* ( $M = 4.04$ ) rated lowest. There was also a Granularity x Design Type interaction ( $F(2, 243) = 9.78, p < .001$ ), as can be seen in the different patterns of Design Strategy in Figure 5(b) and Table 2.

Next, we ran a Granularity x Design Type x Scenario



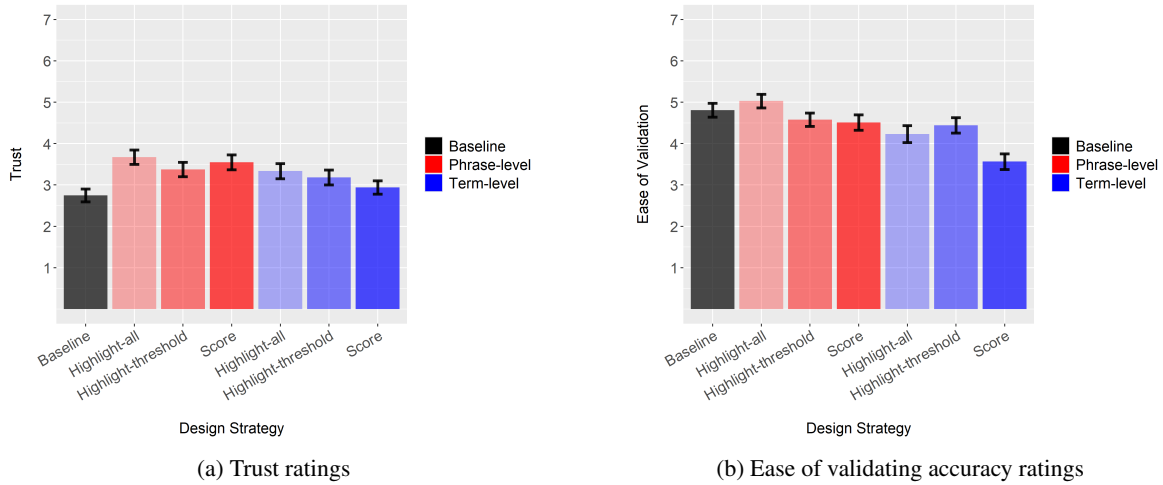


Figure 5: Experiment 2 ratings of each design strategy: (a) trust and (b) ease of validating the accuracy of the model’s response.

model to investigate whether the different stimulus scenarios affected ease of validation ratings. The results were very similar to the previous model, with a main effect of Granularity ( $F(1, 102) = 12.72, p < .001$ ), a main effect of Design Type ( $F(2, 123) = 14.94, p < .001$ ), and a Granularity  $\times$  Design Type interaction ( $F(2, 240) = 9.83, p < .001$ ). There was a main effect of Scenario ( $F(1, 102) = 5.84, p = .02$ ), with the Medical scenario ( $M = 4.71$ ) rated easier to validate than the HR scenario ( $M = 4.07$ ). This likely reflects the fact that the medical scenario was noticeably shorter than the HR scenario, so there was less text to read through and verify the accuracy. However, no interactions involving Scenario reached significance, demonstrating that properties of the scenarios themselves did not differentially affect ratings of validation ease between the different design strategies, and thus the conclusions of the relative ratings of the design strategies hold across multiple scenarios.

**Preference** Preferences, ranked from most (1) to least (7) preferred across both granularity levels, were as follows: *phrase-level highlight-all*, *phrase-level score*, *phrase-level threshold*, *term-level threshold*, *term-level highlight-all*, *baseline*, and *term-level score*. There was a significant effect of Design Strategy on the means of preference rankings ( $\chi^2(6) = 150.42, p < .001$ ). Post-hoc pairwise comparisons showed that participants significantly preferred the *phrase-level highlight-all* design over all other designs ( $p < .01$ ). In contrast, both *baseline* and *term-level score* designs were significantly less preferred compared to other designs, and there was no significant difference between the two. See Table 3 for the ranking means.

We additionally aggregated the preference rankings by Granularity level. There was a significant effect of Granularity on the means of preference rankings ( $\chi^2(2) = 100.88, p < .001$ ). The preference rankings followed the order of *phrase-level*, *term-level*, and *baseline*, with every pairwise comparison showing a significant difference ( $p < .001$ ).

## 6 Discussion

In the present work, we created six design strategies for displaying factuality about an LLM’s response in a question-answer scenario, and conducted two experiments where participants rated these designs on trust, ease of validating the accuracy, and preference. Overall, highlighting every phrase in the response using a factuality score color scale (the *phrase-level highlight-all* design) was the most preferred, trusted, and easiest for users to validate the accuracy of a response. Our results suggest several design recommendations for communicating factuality scores to users, which we explain according to each outcome, and also additional factors that may influence factuality designs.

### 6.1 Design Recommendations

**Trust** All of our factuality designs were effective at increasing and calibrating trust compared to the baseline of showing no factuality information. Hence, we recommend presenting factuality information using one or more of the proposed designs to increase user trust. In addition, in an exploratory analysis, participants’ initial accuracy assessment of the model’s response had a substantial impact on their trust after they viewed the factuality scores. Participants who initially overlooked errors in the model’s response *decreased* their trust after viewing the errors called out through the factuality scores. In contrast, participants who initially identified errors in the model’s response *increased* their trust when they observed that the factuality scores accurately flagged those errors. Therefore, the current results suggest that incorporating factuality information into LLM responses might help to appropriately calibrate the level of end-users’ trust in the model – either in a positive or a negative direction.

**Ease of validation** The *phrase-level highlight-all* design was rated as easier to validate the model’s accuracy than the *baseline* design. However, designs at *term-level* granularity showed inconsistent results between the two experiments: In Experiment 1, *term-level highlight-all* and

*highlight-threshold* designs were rated as significantly easier to assess accuracy than the *baseline*. In contrast, in Experiment 2, the *term-level highlight-all* and *score* designs were rated as more difficult to validate accuracy than the *baseline*. A possible explanation is that the number of annotated terms was higher in Experiment 2 (HR: 39, Medical: 17) than in Experiment 1 (10), potentially overwhelming participants' ability or interest in investigating the accuracy of the model's response. A participant in Experiment 2 mentioned, "*The numbers in colored bubbles makes the text difficult to read.*" Additionally, the proportion of inaccurate terms (color-coded with red or pink) was lower in Experiment 2 (HR: 33%, Medical: 35%) than in Experiment 1 (50%), which may have led users to perceive that *term-level* designs are less informative for validating accuracy. As a result, we recommend adjusting the granularity of annotations according to the model's response length, or incorporating a feature that enables users to disable or filter accurate terms, in order to reduce distractions when assisting users in validating the accuracy of the response.

**Preference** The most common granularity preference was *phrase-level*, and within the phrase-level granularity designs, *highlight-all* was most preferred. Thus, we recommend the *highlight-all* style in cases where user preference and experience is the primary objective.

## 6.2 Factors Impacting Factuality Communication

Our study investigated scenarios where factuality information is likely to be valued; namely, a question-answer task. However, LLM "hallucinations" are not always undesirable or factually inaccurate. For example, researchers have explored the use of LLMs as creativity support tools for tasks including writing stories (Wang et al. 2023), fan fictions (Alfassi et al. 2025), and humor (Wu, Weber, and Müller 2025); brainstorming and ideation (Muller et al. 2024; Muller, He, and Weisz 2024; He et al. 2024); and using analogical thinking to solve design problems (Yang et al. 2025). In use cases like these, presenting factuality scores are likely less useful, because the model's response is not expected to be an accurate or veridical reflection of a source (or reality).

While our study assumed that the algorithm generating factuality scores for the model's response is reliable, no algorithm is perfect and incorrect factuality scores have the potential to erode users' trust in a model. One way to mitigate this issue is to present factuality scores in a fuzzy format, such as using ranges, instead of precise numbers. In our study, we found that the highlighting methods that visually represent a range of factuality scores with a single color performed better than the score method, which presented the precise numerical factuality score. It is possible that this fuzziness was one factor in why the highlighting designs were generally favored over the score designs. Therefore, we encourage HCI researchers to further investigate how to effectively communicate the uncertainty in factuality scores to address the limitations of the algorithms.

The present research focused on a model's response which contained inaccuracies, but it is important to acknowledge situations where a response is entirely accurate or faithful to the source. High factuality scores should increase end-

users' confidence in the LLM response, although some users may be skeptical when seeing a perfect rating. Additionally, in the present work, we told participants to assume the source was reliable; however, it is possible for a response to be *faithful* to an unreliable source. In such cases, high factuality (or faithfulness) scores may be misleading and could lead to over-reliance on the response. We encourage future research to explore design strategies for various situations.

## 6.3 Limitations and Future Directions

Our experiments focused on a question-answer scenario. An important avenue for future research is to explore factuality designs in other LLM tasks, such as summarization or classification. While Experiment 2 tested whether our conclusions generalized to two additional scenarios, there is an enormous amount of unexplored variation in domains and LLM interactions in real-world scenarios. Future research should explore additional topic domains, source and response lengths, and other features of the interaction.

We also created the experimental Responses by editing a real LLM response rather than using the original response. This allowed us to create designs with varying factuality scores that effectively tested our research questions. However, real-world LLM responses may be different, such as emitting a one-word response. However, considering the rapid pace of technological change, the responses may vary over time and across models, making it less critical to rely on actual responses produced by existing models.

While we made efforts to recruit participants with diverse skills, LLM experience, language proficiency, and geographic locations, our participants were all employees of a single technology company. Future studies should involve broader participant samples from the general public.

Finally, it is important to note that our research did not aim to exhaustively explore all potential design strategies. Instead, this study should be viewed as a starting point, encouraging researchers to delve deeper into diverse design strategies and expand the discussion.

## 7 Conclusion

Large language models have known problems with hallucinations. To address these challenges, researchers are developing algorithms to assess the factuality of an LLM model's output, but how to effectively communicate such factuality information to end-users is an open question. We conducted two experiments using three different scenarios to compare six design strategies for communicating factuality scores against a no-markup baseline. We found consistent results showing that **highlighting every phrase in the model's response** based on its factuality score was the most preferred strategy, and led to high trust of the model. This design was also perceived to be easier to validate the accuracy of the LLM response than the baseline. Our findings also suggested that factuality designs may enable participants to appropriately calibrate their trust in a model. Thus, presenting factuality score information in an understandable way is an important tool for end-users to be able to evaluate properties of large language models that are critical to being an informed consumer of AI.

## References

- Agarwal, M.; Talamadupula, K.; Houde, S.; Martinez, F.; Muller, M.; Richards, J.; Ross, S.; and Weisz, J. D. 2020. Quality estimation & interpretability for code translation. *arXiv preprint arXiv:2012.07581*.
- Alfassi, R.; Cooper, A.; Mitchel, Z.; Calabro, M.; Shaer, O.; and Mokryn, O. 2025. Online Storytelling Spaces: Exploring Participants’ Perceptions of Overt and Covert AI Agents. In Glowacka, D.; Santoro, C.; and Xiao, Z., eds., *Joint Proceedings of the ACM IUI 2025 Workshops*, volume 3957 of *CEUR Workshop Proceedings*, 225–235.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Das-Sarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
- Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1): 1–48.
- Bo, J. Y.; Wan, S.; and Anderson, A. 2024. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. *arXiv preprint arXiv:2412.15584*.
- Cai, D.; Wang, Y.; Liu, L.; and Shi, S. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3417–3419.
- Chen, S.; Zhang, F.; Sone, K.; and Roth, D. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. *arXiv preprint arXiv:2104.09061*.
- Chern, I.-C.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; and Liu, P. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv:2307.13528*.
- Derksen, M.; and Morawski, J. 2022. Kinds of replication: Examining the meanings of “conceptual replication” and “direct replication”. *Perspectives on Psychological Science*, 17(5): 1490–1505.
- Do, H. J.; Ostrand, R.; Geyer, W.; Murugesan, K.; Wei, D.; and Weisz, J. 2025. Highlight All the Phrases: Enhancing LLM Transparency through Visual Factuality Indicators: Supplemental Materials. <https://doi.org/10.17605/OSF.IO/P6HN2>.
- Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A. T.; Fan, Y.; Zhao, V.; Lao, N.; Lee, H.; Juan, D.-C.; et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16477–16508.
- Hao, K.; and Seetharaman, D. 2023. Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *The Wall Street Journal*. <https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483>.
- He, J.; Houde, S.; Gonzalez, G. E.; Silva Moran, D. A.; Ross, S. I.; Muller, M.; and Weisz, J. D. 2024. AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, 1–14.
- Hicks, M. T.; Humphries, J.; and Slater, J. 2024. ChatGPT is bullshit. *Ethics and Information Technology*, 26(2): 1–10.
- Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.
- Honovich, O.; Aharoni, R.; Herzig, J.; Taitelbaum, H.; Kulkarni, D.; Cohen, V.; Scialom, T.; Szpektor, I.; Hassidim, A.; and Matias, Y. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3905–3920.
- Howell, E. 2023. James Webb Telescope question costs Google \$100 billion – here’s why. *Space.com*. <https://www.space.com/james-webb-space-telescope-google-100-billion>.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12).
- Johnson, J. S.; and Newport, E. L. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1): 60–99.
- Kim, S. S.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 822–835.
- Kim, S. S.; Watkins, E. A.; Russakovsky, O.; Fong, R.; and Monroy-Hernández, A. 2023. Humans, AI, and Context: Understanding End-Users’ Trust in a Real-World Computer Vision Application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 77–88.
- Kryściński, W.; McCann, B.; Xiong, C.; and Socher, R. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346.

- Kuznetsova, A.; Brockhoff, P. B.; and Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13): 1–26.
- Laban, P.; Schnabel, T.; Bennett, P. N.; and Hearst, M. A. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10: 163–177.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1): 50–80.
- Leiser, F.; Eckhardt, S.; Knaeble, M.; Maedche, A.; Schwabe, G.; and Sunyaev, A. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In *Proceedings of Mensch und Computer 2023*, 81–90. Association for Computing Machinery.
- Lenth, R. 2020. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.0.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- Li, D.; Chen, T.; Zadikian, A.; Tung, A.; and Chilton, L. B. 2023. Improving Automatic Summarization for Browsing Longform Spoken Dialog. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.
- Liao, Q. V.; and Vaughan, J. W. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.
- Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating Verifiability in Generative Search Engines. arXiv:2304.09848.
- Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. arXiv:2005.00661.
- Metz, C. 2023. The Secret Ingredient of ChatGPT Is Human Advice. *The New York Times*. <https://www.nytimes.com/2023/09/25/technology/chatgpt-rlhf-human-tutors.html>.
- Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv preprint arXiv:2305.14251*.
- Muller, M.; He, J.; and Weisz, J. 2024. Workplace Everyday-Creativity through a Highly-Conversational UI to Large Language Models. In *ACM CHI Conference on Human Factors in Computing Systems*.
- Muller, M.; Houde, S.; Gonzalez, G.; Brimijoin, K.; Ross, S. I.; Moran, D. A. S.; and Weisz, J. D. 2024. Group Brainstorming with an AI Agent: Creating and Selecting Ideas. In *International Conference on Computational Creativity*.
- Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ryan, J. 2022. Meta Trained an AI on 48M Science Papers. It Was Shut Down After 2 Days. *CNET*. <https://www.cnet.com/science/meta-trained-an-ai-on-48-million-science-papers-it-was-shut-down-after-two-days/>.
- Sato, M.; and Roth, E. 2023. CNET found errors in more than half of its AI-written stories. *The Verge*. <https://www.theverge.com/2023/1/25/23571082/cnet-ai-written-stories-errors-corrections-red-ventures>.
- Scialom, T.; Dray, P.-A.; Lamprier, S.; Piwowarski, B.; Staliano, J.; Wang, A.; and Gallinari, P. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6594–6604.
- Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.
- Sloan, K. 2023. A lawyer used ChatGPT to cite bogus cases. What are the ethics? *Reuters*.
- Smith, A. L.; Greaves, F.; and Panch, T. 2023. Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health*, 2(11): e0000388.
- Stack Overflow. 2025. Why can't I use Artificial Intelligence tools to generate answers? <https://stackoverflow.com/help/ai-policy>. Accessed: 2025-08-08.
- van der Bles, A. M.; van der Linden, S.; Freeman, A. L.; Mitchell, J.; Galvao, A. B.; Zaval, L.; and Spiegelhalter, D. J. 2019. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5): 181870.
- Vasconcelos, H.; Bansal, G.; Fournay, A.; Liao, Q. V.; and Vaughan, J. W. 2023. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *arXiv preprint arXiv:2302.07248*.
- Wang, S.; Petridis, S.; Kwon, T.; Ma, X.; and Chilton, L. B. 2023. PopBlends: Strategies for conceptual blending with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R. J.; and Nakov, P. 2024. Factuality of large language models: A survey. *arXiv preprint arXiv:2402.02420*.

- Weisz, J. D.; Muller, M.; Houde, S.; Richards, J.; Ross, S. I.; Martinez, F.; Agarwal, M.; and Talamadupula, K. 2021. Perfection not required? Human-AI partnerships in code translation. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 402–412.
- Wischniewski, M.; Krämer, N.; and Müller, E. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Wu, Z.; Weber, T.; and Müller, F. 2025. One Does Not Simply Meme Alone: Evaluating Co-Creativity Between LLMs and Humans in the Generation of Humor. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1082–1092.
- Yang, Y.; Mohanty, V.; Martelaro, N.; Kittur, A.; Chen, Y.-Y.; and Hong, M. K. 2025. From Overload to Insight: Scaffolding Creative Ideation through Structuring Inspiration. *arXiv preprint arXiv:2504.15482*.
- Yue, X.; Wang, B.; Zhang, K.; Chen, Z.; Su, Y.; and Sun, H. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.
- Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 295–305.
- Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 2023. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.