Fool Me, Fool Me: User Attitudes Toward LLM Falsehoods

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have become central tools in various fields, they often provide inaccurate or false information. This study examines user preferences regarding falsehood responses from LLMs. Specifically, we evaluate preferences for LLM responses where false statements are explicitly marked versus unmarked responses and preferences for confident falsehoods compared to LLM disclaimers acknowledging a lack of knowledge. Additionally, we investigate how requiring users to assess the truthfulness of statements influences these preferences.

Surprisingly, 61% of users prefer unmarked falsehood responses over marked ones, and 69% prefer confident falsehoods over LLMs admitting lack of knowledge. When users are required to evaluate the truthfulness of statements, preferences for unmarked and falsehood responses decrease slightly but remain high. In all our experiments, a total of 300 users participated. These findings suggest that user preferences, which influence LLM training via feedback mechanisms, may inadvertently encourage the generation of falsehoods. Future research should address the ethical and practical implications of aligning LLM behavior with such preferences.

1 Introduction

Large Language Models (LLMs) affect many aspects of our lives: programmers use them to obtain code snippets, students rely on them for homework assistance, and LLMs also play a significant role in literacy. People frequently use LLMs as a source of information in various fields, including exact sciences, life sciences, and history. The widespread use of LLMs as sources of information underscores the importance of ensuring that they generate true and accurate information. However, LLMs often generate inaccurate and even false information, which greatly impedes their reliability as trusted tools for the dissemination of knowledge. This issue is particularly concerning given the confident tone and authoritative style in which LLMs present their outputs, making it difficult for users to differentiate between accurate information and inaccuracies. The persuasive nature of LLM-generated content increases the risk of misinformation, leading users to believe and propagate falsehoods.

Although it is widely assumed that users prefer accurate and truthful information, to the best of our knowledge, no prior work has explicitly tested this assumption in the context of LLMs. That is, no previous work has systematically investigated user preferences by presenting them with LLM-generated responses, where the veracity is disclosed a priori, and requiring users to evaluate or choose between truthful and false outputs. In this paper, we introduce a novel focus on user behavior in scenarios where the distinction between truth and falsehood is unambiguous. Specifically, we investigate user attitudes toward LLM falsehoods in two key issues:

1. **Marked vs. Unmarked.** We examine user preferences between two types of LLM-generated responses: one where truth and falsehood are explicitly marked for easy distinction, and another with standard unmarked responses, where truth and falsehood appear identical in text format.

2. **Uninformative Truth vs. Falsehood.** In scenarios where an LLM cannot provide the truth (when asked about events occurring after its cut-off date), we examine user preferences between two types of responses: one that acknowledges a lack of knowledge (uninformative truth) and another that provides a confident but inaccurate answer (falsehood).

To explore user preferences, we conducted two experiments for each issue:

1. A one-phase experiment, in which the two versions of ChatGPT-generated responses are represented to the participants, and they should only select their preferred version.

2. A two-phase experiment, in which, after the participants select their preferred version in the first phase, the other version is hidden, and participants must indicate whether a sentence related to ChatGPT's response is true or false.

The one-phase experiment examines initial user preferences, while the two-phase experiment attempts to simulate a situation in which the participant has an incentive to pick the marked or truthful response.

In the first issue, a clear majority of participants chose the unmarked version during the one-phase experiment. However, in the two-phase experiment, in which the participants' choice affected their performance in determining whether a statement is true or false, preferences were nearly evenly split between the two versions. For the second issue, in both experiments, users overwhelmingly favored a confident but incorrect response over one acknowledging a lack of knowledge.

These findings are surprising, as they challenge the assumption that individuals naturally favor truthful information over inaccurate responses. Indeed, previous research on user preferences regarding truth and falsehood has primarily examined behavior in contexts where the truth is ambiguous, such as studies on user trust in AI systems (Bach et al., 2024) and the dissemination of disinformation (Buchanan, 2020). In contrast, this study explicitly presented the veracity of each statement. Despite this clarity, users still preferred the unmarked and falsehood versions.

Our findings have significant ethical and practical implications for LLM development. While users express a theoretical preference for transparency and accuracy, their real-time choices often gravitate toward aesthetically appealing but inaccurate responses. Real-time choices may affect an LLM's performance, as user preferences are incorporated into its development using methods such as reinforcement learning from human feedback (RLHF), which fine-tune LLMs based on user preferences (Ziegler et al., 2019). Therefore, our findings raise a question that must be addressed by LLM developers: Does learning from human feedback inadvertently encourage LLMs to generate inaccurate or false information? Moreover, presenting the truth often requires complex expression and acknowledging uncertainty, which can drive users to prefer simpler yet incorrect responses. This highlights an ethical conflict for LLM developers: Should they prioritize factual accuracy or cater to user preferences for confident and appealing responses, even at the expense of truth?

2 Related Work

We now provide an overview of previous research on human tendencies to spread falsehoods and mention methods for detecting and reducing LLM-generated lies and hallucinations. Finally, we discuss methods for incorporating human feedback into machine learning models.

2.1 Spreading Disinformation

Spreading disinformation on social networks is a common phenomenon with significant implications for public discourse and decision making. Under the assumption that people spread disinformation since they believe it to be true, Cialdini (2007) suggests three main reasons that persuade people to believe and spread disinformation. **Consistency.** People may share information that aligns with their past behaviors or beliefs, as it helps to reinforce their worldview and maintain cognitive coherence. **Consensus.** People may share information they perceive to be widely accepted or supported by others, as social proof plays a powerful role in validating beliefs. **Authority.** People are more likely to share information originating from sources they deem credible or authoritative, even if the content itself lacks accuracy.

Buchanan (2020) confirmed the first reason and further identified demographic factors, finding that younger, male, and less-educated individuals were more likely to spread disinformation. Similarly, Vosoughi et al. (2018) demonstrated that false information spreads faster, deeper, and more broadly on social media than truthful information, largely due to its novelty and emotional appeal, which captures users' attention and encourages sharing.

Furthermore, Pennycook and Rand (2019) noted that cognitive laziness and the lack of engagement in analytical thinking contribute to the spread of disinformation. They argue that promoting digital literacy and critical thinking skills can mitigate this problem by enabling users to discern between credible and false content.

In contrast to studies on spreading disinformation, which assume that people believe the disinformation to be true, our research examines user preferences when they can clearly distinguish between truth and falsehood.

2.2 LLM Hallucinations Reduction

One of the challenges in LLM research is to detect and reduce LLM lies and hallucinations. Huang et al. (2025) provide a survey on LLM hallucinations, focusing on three main issues: hallucination causes, hallucination detection and benchmarks, and hallucination mitigation. For hallucination mitigation, some methods treat the LLM as a black-box, focusing their efforts on prompt engineering to achieve more trustworthy responses (Peng et al., 2023; Madaan et al., 2023). Other methods fine-tune the LLM for reducing hallucination based on human feedback (Bakker et al., 2022). However, our results suggest that user preferences may encourage lies rather than reduce them. Another approach suggests using the intermediate states of an LLM to detect and reduce lies (Azaria and Mitchell, 2023).

2.3 Learning from Human Feedback

Human provided information, such as data labeling and model performance evaluation, plays a major role in machine learning. Kirk et al. (2023) survey the existing approaches for learning from human feedback. They start with before and after the advent of LLMs, continue with summarizing current methods for incorporating human feedback learning into LLMs (e.g., reinforcement learning fine-tuning, supervised fine-tuning, and pretraining), and end with some future challenges.

One common method for learning from human feedback is reinforcement learning from human feedback (RLHF), which is commonly used for LLMs fine-tuning according to user preferences (Ziegler et al., 2019). Wang et al. (2023) survey the use of RLHF for aligning LLMs with human.

3 Experimental Design

We conducted four experiments, labeled A, B, C, and D, to investigate user preferences. All the experiments use genuine ChatGPT responses (OpenAI, 2024a). These experiments were administered through Amazon Mechanical Turk, a platform widely recognized as a reliable source for human subject experiments (Paolacci et al., 2010). Participants had to meet the following criteria: task approval rate greater than 99, number of tasks approved greater than 1000, and participant location is USA. We now describe the experiments in detail, with the results being analyzed in Section 4. Each participant who completed the task received a payment of $0.50, with an average completion time of approximately three minutes, equivalent to an hourly wage rate of approximately $10. Participants provided consent for participating in each experiment. All experiments received IRB approval.

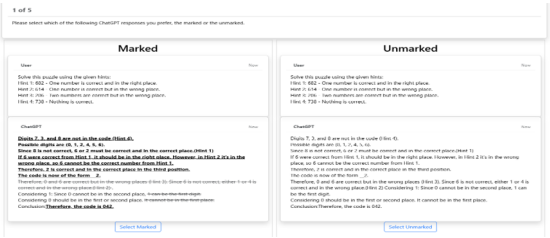

Figure 1: Example of a response pair from Experiment A.

3.1 Experiment A

In Experiment A, participants were presented with two ChatGPT responses, with the following instructions: "You will be shown two interactions of a user with ChatGPT. One is the standard ChatGPT response, while the other ChatGPT response includes marking of 'false', 'true' and subjective or non-fact statements. The marked ChatGPT uses the following marking:

- Bold and Underlined: **This is a true fact.**;
- Crossed-out Grey text: ~~This is a false fact.~~;
- Regular Text: This sentence is subjective/not a fact.

You are required to select the response that you prefer better."

As stated, participants were required to select their preferred response. Each participant was presented with five pairs of such two responses, with one appearing on the right and other on the left. The location of responses and the order of pairs were randomly sampled to avoid positional bias. An example of a response pair is shown in Figure 1.

3.2 Experiment B

In Experiment B, we examined whether requiring participants to verify the truth of sentences from a

Figure 2: Example of the second phase of Experiment B.

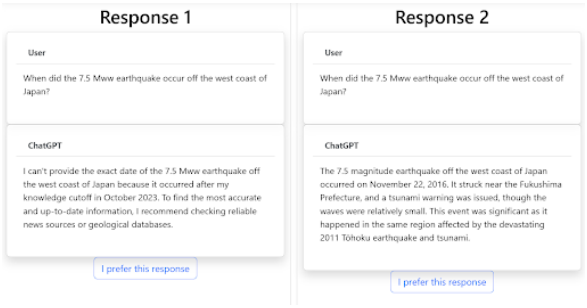

Figure 3: Example of a response pair from Experiment C.

ChatGPT response after their initial choice would influence their preferences. This experiment consisted of two phases. In the first phase, participants were presented with a response pair as in Experiment A. After selecting their preferred response, the unselected version was removed, leaving only the chosen response visible.

In the second phase, participants were asked to determine whether a specific sentence from the ChatGPT's response is true, false or non-factual. If they provided an incorrect answer, they had to wait 30 seconds before attempting to correct it. Information related to this question appeared also in the instructions so the participants were aware to the fact that they will be required to answer a question after each selection. Each participant completed such five two-phase response pairs, ordered randomly. Figure 2 provides an example of the second phase, illustrating a marked response selection in the first phase.

When the unmarked version is selected, it may be more challenging for participants to answer the follow-up question correctly. This difficulty can activate the delay mechanism, requiring participants to wait 30 seconds before revising their answers.

3.3 Experiment C

In Experiment C, participants were presented with two ChatGPT responses, with the following instructions: "You will be shown two interactions of a user with ChatGPT related to a recent event not known to ChatGPT. In one interaction ChatGPT acknowledges that it may not know the answer, while in the other interaction it provides false information. You are required to select the response that you prefer better."

Each participant was presented with five pairs of such two responses, one on the right and the other on the left. The location of responses (left and right) and the order of pairs were randomly sampled to avoid positional bias. An example of a response pair is shown in Figure 3.

3.4 Experiment D

In Experiment D, we examined how requiring participants to verify the truth of sentences from a ChatGPT response after their initial choice influenced their preferences. This experiment consisted of two phases. In the first phase, participants were presented with a response pair as in Experiment C. After selecting their preferred response, the unselected version was removed, leaving only the chosen response visible.

In the second phase, participants were asked to determine whether the event in question was after ChatGPT's cutt-off date, and whether ChatGPT knows the answer to the question posed. The participants had to indicate that the event in question was after ChatGPT's cutt-off date, and thus it does not know the answer to it. However, depending on the phrasing (e.g. "ChatGPT doesn't have any information on ..." vs. "ChatGPT has information on ..."), the participants had to either select True or False. If the participants provided an incorrect answer, they had to wait 30 seconds before attempting to correct it. Each participant completed such five two-phase response pairs, ordered randomly. Figure 4 provides an example of the second phase, illustrating an acknowledgment the lack of ChatGPT's knowledge selection.

4 Results

We now present the results of the experiments. For all the experiments, we applied a χ^2 test for goodness of fit. This test evaluates whether the observed differences in participant preferences between the two versions are statistically significant. The null hypothesis (H_0) assumes no difference in participant preferences between the two versions. Under this assumption, the expected values for each ver-

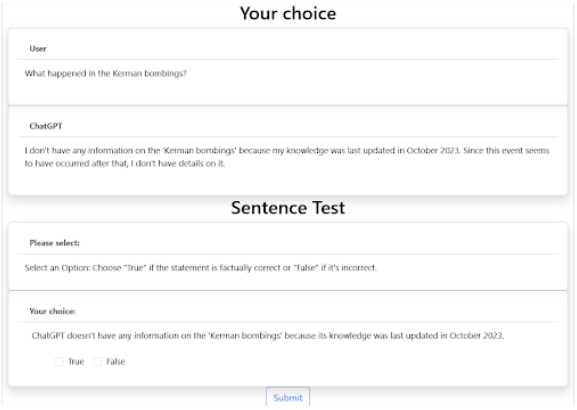

Figure 4: Example of the second phase of Experiment D.

sion are $\frac{N}{2}$, where N represents the total number of response pairs. The following sections detail the observed preferences and the corresponding statistical analyses for each experiment.

4.1 Experiment A

In Experiment A we recruited 74 participants, each of whom completed five response pairs, totaling 370 response pairs. About 60.1% (223 out of 370) preferred the unmarked version, while only about 39.9% (147 out of 370) preferred the marked version.

These observed values yield a χ^2 value of 15.6 with a p-value of 0.00008. Since $p < 0.01$, we reject H_0, indicating a statistically significant preference for the unmarked version.

4.2 Experiment B

In Experiment B we recruited 75 participants, each of whom completed five response pairs, totaling 375 response pairs. About 51% (192 out of 375) preferred the unmarked version, and about 49% (183 out of 375) preferred the marked version.

These observed values yield a χ^2 value of 0.216 with a p-value of 0.6421. Since $p > 0.05$ we fail to reject H_0, indicating no statistically significant difference in participants preferences. The difference between Experiment A and Experiment B suggests a shift toward the marked version when the participants were required to verify the truth of specific sentences from their preferred version.

4.3 Experiment C

In Experiment C we recruited 71 participants, each of whom completed five response pairs, totaling 355 response pairs. About 69.6% (247 out of 355) preferred the falsehood response, while only about 30.4% (108 out of 355) preferred the uninformative truth response.

These observed values yield a χ^2 value of 54.425 with a p-value less than 0.00001. Since $p < 0.01$, we reject H_0, indicating a statistically significant preference for the falsehood response.

4.4 Experiment D

In Experiment D we recruited 80 participants, each of whom completed five response pairs, totaling 400 response pairs. About 68.25% (273 out of 400) preferred the falsehood response, while only about 31.75% (127 out of 400) preferred the uninformative truth.

These observed values yield a χ^2 value of 53.29 with a p-value less than 0.00001. Since $p < 0.01$, we reject H_0, indicating a statistically significant preference for the falsehood response. That is, requiring the participants to indicate whether the event in question was beyond ChatGPT's cut-off date, did not influence their preferences. This is likely because the participants were told in the instructions that all events in all questions occurred after ChatGPT's cut-off date and, as such, ChatGPT does not know the answer. Therefore, the participants could select the falsehood response and still determine correctly that the event in question occurred after ChatGPT's cut-off date.

5 Additional Analysis

5.1 True/False Question for Participants

In the second phase of Experiments B and D, participants were required to verify the truth of specific sentences. In Experiment B (Marked vs. Unmarked), 74% (136 out of 183) of participants who preferred the marked version answered correctly, compared to 70% (134 out of 192) of those who preferred the unmarked version. In Experiment D (Uninformative Truth vs. Falsehood), 92% (117 out of 127) of participants who preferred the uninformative truth response answered correctly, compared to 85% (232 out of 273) of those who preferred the falsehood response.

The consistently high success rates (above 70%) suggest that participants engaged seriously with the task. If they had answered randomly, success rates would have been 33% in Experiment B (which had three options: true, false, and subjective) and 50% in Experiment D (which had two options: true and false).

	Choice	F	M	χ^2	p-val.
A	Marked	76 40%	66 38%	0.12	0.73 > 0.05
	Unmarked	114 60%	109 62%		
B	Marked	94 45%	89 54%	2.76	0.097 > 0.05
	Unmarked	116 55%	76 46%		
C	Uninf. truth	72 69%	175 70%	0.02	0.89 > 0.05
	Falsehood	33 31%	75 30%		
D	Uninf. truth	84 28%	43 43%	7.66	0.006 < 0.05
	Falsehood	221 72%	57 57%		

Table 1: Comparison between the choice of female (F) and male (M) in all of the experiments.

Additionally, in both experiments, participants who selected the marked or uninformative truth responses performed slightly better than those who selected the unmarked or falsehood responses. However, statistical analysis using the χ^2 test found that these differences were not significant ($p > 0.05$).

5.2 Gender

Table 1 shows the difference between the choice of females and males. Using the χ^2 test, we evaluated the statistical dependence between gender and the selected version. Statistical dependence between gender and choice was observed only in Experiment D. This statistical dependence can be attributed to differences in male and female error rates when answering questions in the second phase. In Experiment D, Among those who selected the falsehood response, 34.6% of males answered incorrectly, compared to only 10.4% of females. For comparison, in Experiment B, the gap between genders was smaller: among those who selected the falsehood response, approximately 31.6% of males answered incorrectly, compared to approximately 29.3% of females.

Choice	GS	Bachelor	Other
Uninformative truth	37 84%	83 24%	9 90%
Falsehood	7 16%	267 76%	1 10%

Table 2: The choice of participants in Experiment D, partitioned by their education level: graduate school (GS), bachelor's degree, and other.

5.3 Education Level

We used the χ^2 test to determine whether there is a statistical dependence between the education level of participants and their selections. Our findings indicate that there is no statistical dependence between education level and user choices in Experiments A, B, and C. However, in Experiment D, a significant statistical dependence was observed, with a χ^2 value of 81.4626 and a p-value less than 0.00001. Table 2 provides a breakdown of user choices in Experiment D by education level. These results might suggest that, unlike people with bachelor level education, people with graduate school education tend to prefer uninformative but truthful responses over falsehoods.

5.4 Feedback From the Participants

To further explore participant preferences, a feedback mechanism was adjusted, including quantitative and qualitative feedback. Table 3 presents the quantitative feedback for Experiments A and B, and Table 4 for Experiments C and D. The analysis of quantitative feedback reveals nuanced perspective on participant preferences when interacting with ChatGPT. Interestingly, in experiments A and B there seems to be a slight (insignificant) preference for a marked version of ChatGPT. Furthermore, in experiments C and D, the participants clearly seem to believe that it is important for ChatGPT to provide only accurate information.

For the qualitative feedback, we asked the participants to provide comments about both versions that they were presented with (side by side). A recurring theme in the feedback for Experiments A and B was that the marked version is more convenient for fact checking, whereas the unmarked version appears to be more aesthetically pleasing and neat. For Experiments C and D, participants commented that acknowledging a lack of knowledge increases their trust in ChatGPT.

These findings suggest that people, in princi-

	I believe that it is important for ChatGPT to mark which sentences are true and which are false.	I prefer that ChatGPT's responses will not include any marking.
A	3.81	3.74
B	3.56	3.35

Table 3: Quantitative feedback from Experiments A and B (average, out of 5).

	I believe that it is important for ChatGPT to provide only accurate information.	I believe that it is important for ChatGPT to provide information even if it is incorrect.
C	4.11	3.17
D	4.8	1.25

Table 4: Quantitative feedback from Experiments C and D (average, out of 5).

ple, favor transparency and accuracy, indicating a strong preference for truthfulness as a guiding principle. However, when faced with real-time choices, people often gravitate toward more aesthetically appealing but potentially less accurate responses. This behavior illustrates the well-documented intention-behavior gap (Sheeran, 2002), similar to the dilemma of choosing between healthy and junk food (Monds et al., 2016; Faries, 2016). Although participants value transparency and accuracy in principle, their real-time choices often prioritize aesthetic appeal or convenience, much like individuals who express a preference for healthy food but choose junk food in practice.

5.5 Preferences per Question

We used the χ^2 test to determine whether there is a statistical dependence between the response pairs and the selections of the participants. In Experiments B and C, statistical independence was observed. In Experiment A, a significant statistical dependence was observed, with a χ^2 value of 11.47 and a p-value of $0.02 < 0.05$. In Experiment D, a significant statistical dependence was observed, with a χ^2 value of 74.25 and $p < 0.00001$. Table 5 provides a breakdown of number of choices in Experiments A and D by response pairs.

	Response pair	A1	A2	A3	A4	A5
A	Marked	37	32	18	32	28
	Unmarked	37	42	56	42	46
	Response pair	D1	D2	D3	D4	D5
D	Uninf. truth	18	56	11	16	26
	Falsehood	62	24	69	64	54

Table 5: Selections per response pair in Experiments A and D.

In Experiment A, the largest gap between marked and unmarked choices occurs in response pair A3. In this response pair, unlike other response pair, the ChatGPT response contains only a false sentence, and thus looks too unaesthetic. The question is: "How much is 3.2 to the power of 3.3?" Since the answer is approx. 46.45, the marked response is: "~~The result of $3.2^{3.3}$ is approximately 21.73.~~" We believe that the unaesthetic form of the full response led to this large gap.

In Experiment D, the participants actually preferred the uninformative truth over the falsehood in response pair D2. In this response pair the question is: "When did the Republic of Artsakh officially dissolve?" The false response is: "The Republic of Artsakh has not been officially dissolved and continues to exist as a self-declared republic in the Nagorno-Karabakh region." Asking "when" implies that the event has indeed occurred, but this response contradicts this assumption, potentially making it less appealing or intuitive for participants. In addition, a response stating that an event did not occur may not be engaging enough, especially considering the participants' declared preference for truth in principle (see Section 5.4).

6 Discussion

In this paper we examined user preferences regarding LLM-generated falsehoods. Following Horton (2023), who suggests using LLMs as simulated economic agents, we explored this approach by conducting our four experiments with ChatGPT agents (OpenAI, 2024b). For each experiment, we simulated ChatGPT agents by replicating the demographic profiles (gender, age, and education level) of the actual participants and asked them to select their preferred version. As shown in Table 6, ChatGPT agents exhibited a stronger inclination toward

accurate responses (i.e., marked and uninformative truth versions) compared to human participants. One conclusion from this is that human participants cannot be replaced by ChatGPT agents (Dillion et al., 2023).

We note that ChatGPT, like many other LLMs, obtains human feedback by prompting users to select their preferred version between two options, as illustrated in Figure 5. However, since our findings indicate that user preferences often incline toward inaccurate responses, directly incorporating these preferences into fine-tuning processes may not be advisable. Given that ChatGPT tends to favor accurate responses, it could incorporate a verification mechanism to assess the validity of user choices before using them for model fine-tuning. This verification could be performed by asking ChatGPT itself whether the preferred version is accurate. By leveraging its existing capabilities to evaluate factuality, ChatGPT could ensure that only reliable preferences are used for training.

Another approach to verification is for ChatGPT to conduct a preliminary test for each of the users. In this test, ChatGPT may ask the user to select her preferred response between two responses, in which one response is an uninformative truth and the other is a falsehood. The user must be told a priory which response is truthful. Users who pass the test by selecting the truth will be marked by ChatGPT as reliable users, which their selections in any conversation can be used to fine-tune the model. However, this approach raises ethical concerns, as conducting such a verification test without the respondent's knowledge could violate principles of informed consent.

Alternatively, the regular feedback mechanism could be adapted to address this issue. ChatGPT could track the number of times each user selects a truthful option and use this data to create a reliability score. Feedback from users with high reliability scores could then be prioritized for fine-tuning, ensuring that training incorporates preferences aligned with factual accuracy while maintaining transparency and ethical standards.

7 Conclusion & Future Work

In this paper we discussed user preferences regarding LLMs-generated falsehoods, focusing on two key issues: (a) selecting between marked and unmarked versions, where the marked version allows distinguishing between truth and falsehood statements, and (b) selecting between uninformative truth and falsehood versions, where the falsehood is written with confidence, while the truth version acknowledges a lack of knowledge.

	ChatGPT agents preferred the marked or uninformative truth version	Human participants preferred the marked or uninformative truth version
A	45.3%	39.9%
B	69%	49%
C	70.8%	30.4%
D	87.4%	31.75%

Table 6: Comparison between ChatGPT agents and human participants preferences.

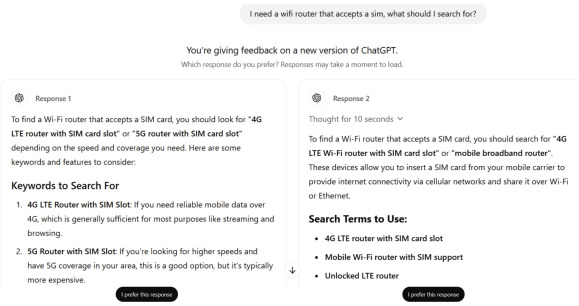

Figure 5: Example of ChatGPT's human feedback mechanism.

Our findings indicate that users tend to favor more aesthetically pleasing and confident responses, even at the expense of accuracy. This contradicts their fundamental agreement on truthfulness, as reflected in their feedback (see Section 5.4). Future work could explore alternative marking methods, such as customizable or less accentuated designs, to investigate whether these approaches influence user preferences.

While users showed a strong preference for unmarked and falsehood versions, these preferences may vary for specific uses. For example, students and researchers using LLMs for their studies might value tools that help distinguish between truth and falsehood. Such tools could also be critical in healthcare applications, such as assessing medical information for patients. Future work could examine user preferences within specific populations and use cases to better tailor LLM features.

8 Limitations

This paper examines user preferences regarding LLM-generated falsehoods through four experiments conducted on USA participants recruited via Amazon Mechanical Turk. However, our participant may not fully represent the broader population. Specifically, user preferences may vary across different cultural and linguistic contexts. Cultural and linguistic differences might influence how users perceive confidence, truthfulness, and markings. Examining the preferences of those who actively provide feedback to ChatGPT and other LLMs could have been more valuable, but targeting that exact population may be challenging.

The experiments used pre-prepared ChatGPT responses, which limits the ability to capture user preferences in real-time interactions. Real-time interactions are advantageous because users prioritize topics they raise themselves, which may influence their preferences.

Conducting such experiments with user-generated interactions presents several challenges. One challenge is the need for an automatic marking tool that operates on top of the original LLM on-the-fly. Additionally, the marking tool's error rate must be very low, as incorrectly marking truthful statements as falsehoods (or vice versa) could significantly influence user preferences. Furthermore, users may inquire about personal facts and perceive them as truths, which contrasts with the rationale behind the marking system, as subjective statements should not be marked as truthful.

Finally, the binary choice framework used in this study (i.e., marked vs. unmarked, uninformative truth vs. falsehood) may oversimplify user preferences. Future research could explore more nuanced options, such as hybrid or customizable marking systems, to better understand user preferences.

9 Ethical Statement

This study was conducted in accordance with ethical research principles, ensuring respect for participant privacy and informed consent. Data was collected anonymously to protect participants' identities, and no personally identifiable information was stored or shared. Additionally, the findings presented here aim to stimulate discussions on the ethical implications of aligning LLMs with human preferences, particularly when these preferences may increase LLM-generation of lies. The authors are committed to advancing responsible AI development, particularly ensuring transparency and accuracy in LLMs. Furthermore, this paper raises critical ethical concerns that must be addressed, emphasizing the importance of balancing user preferences with the need for truthfulness and reliability in AI outputs.

References

Amos Azaria and Tom M. Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 967–976. Association for Computational Linguistics.

Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. International Journal of Human–Computer Interaction, 40(5):1251–1266.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. Advances in Neural Information Processing Systems, 35:38176–38189.

Tom Buchanan. 2020. Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. Plos one, 15(10):1–33.

Robert B Cialdini. 2007. Influence: The psychology of persuasion, volume 55. Collins New York.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7):597–600.

Mark D Faries. 2016. Why we don't "just do it" understanding the intention-behavior gap in lifestyle medicine. American journal of lifestyle medicine, 10(5):322–329.

Horton, John J. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 43(2).

Hannah Kirk, Andrew M. Bean, Bertie Vidgen, Paul Röttger, and Scott Hale. 2023. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 2409–2430. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Lauren A Monds, Carolyn MacCann, Barbara A Mullan, Cara Wong, Jemma Todd, and Richard D Roberts. 2016. Can personality close the intention-behavior gap for healthy eating? An examination with the HEXACO personality traits. Psychology, health & medicine, 21(7):845–855.

OpenAI. 2024a. ChatGPT 3.5 (June 16 version) [Large language model]. https://openai.com.

OpenAI. 2024b. ChatGPT 4 (November 28 version) [Large language model]. https://openai.com.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. Judgment and Decision making, 5(5):411–419.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. Computing Research Repository arXiv:2302.12813. Version 3.

Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences, 116(7):2521–2526.

Paschal Sheeran. 2002. Intention—behavior relations: a conceptual and empirical review. European review of social psychology, 12(1):1–36.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. science, 359(6380):1146–1151.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning Large Language Models with Human: A Survey. Computing Research Repository, arXiv:2307.12966.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. Computing Research Repository, arXiv:1909.08593. Version 2.