Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39107

From Detecting Hallucinations to Adjusting Reliance: Longitudinal Dynamics of Trust Calibration in Repeated Human–LLM Interaction

Haoyi Shi

Loughborough University, London, UK
shihaoyi02447@outlook.com

Abstract. Generative artificial intelligence has been widely embedded in question answering, writing assistance, and decision support. However, LLM hallucinations may lead users to accept incorrect information without sufficient verification and develop inappropriate reliance through repeated interaction. To address this issue, the experiment constructs a four round human LLM interaction task. It compares the effects of linguistic uncertainty expression, numerical uncertainty expression, and no uncertainty expression on user acceptance, verification, revision, and responsibility attribution. Cultural background, uncertainty avoidance, institutional trust, and relational orientation are also included. The simulated pilot results show that no uncertainty expression increases overall reliance and incorrect acceptance. Linguistic warnings improve verification and appropriate reliance, while numerical warnings more strongly influence responsibility attribution toward the system provider. Cultural background moderates the effect of warning format, but individual cultural psychology provides stronger explanatory power. The findings indicate that LLM risk governance should move beyond hallucination detection and support trust calibration through interface warnings, responsibility perception, and user feature modeling.

Keywords: large language model, hallucination detection, uncertainty communication, trust calibration, responsibility attribution

1. Introduction

The use of generative AI is now more prevalent for question answering, writing, and decision-making; however, fluency in LLM responses could lead to acceptance of false information without verification. Studies on hallucination detection have been able to determine false outputs using factual consistency and semantic uncertainty but lack in providing knowledge about the effect of detected false information on future trust [1]. The automation trust literature focuses on the importance of appropriate trust based on the capability of the automation [2]. Through iterative interaction, users calibrate their trust and verification processes based on past successes and failures and warnings, and possibly based on responsibility assignment as well.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39107

2. Literature review

2.1. Hallucination detection

LLM hallucination often appears linguistically coherent and superficially credible, which makes it difficult for users to distinguish correct answers from incorrect ones through expression quality alone. Existing detection approaches include retrieval augmentation, factual consistency evaluation, multiple generation comparison, and uncertainty estimation. These approaches move model output assessment from single text judgment to evidence based, semantic, and probabilistic risk identification [3]. Semantic entropy further suggests that the dispersion of meanings across different generated answers can reveal instability in the model's knowledge state [4].

2.2. Uncertainty communication

Uncertainty communication is an important interface form that connects internal model risk with external user behavior. Linguistic expression communicates risk through natural language. It lowers the threshold of comprehension, yet implicit wording may weaken the warning effect. Numerical expression presents risk through confidence scores or probabilities. It is more direct, yet users may treat it as a technical commitment made by the platform about answer accuracy [5]. In human AI decision making, warning format influences whether users pause to think, click verification information, and revise their original judgments [6].

2.3. Trust calibration

Trust calibration means that user reliance should change dynamically with the actual performance of the system. The accuracy of machine learning systems can affect user trust, yet users do not always translate accuracy changes into appropriate acceptance behavior [7]. After experiencing AI errors, users may develop algorithm aversion. They may also continue to over rely on highly fluent outputs, which shows a complex psychological link between failure experience and later behavior [8]. In LLM interaction, appropriate reliance means that correct answers are accepted, incorrect answers are questioned, and uncertain answers are further verified. Responsibility attribution further shapes the path of trust recovery [9].

3. Experimental methods

3.1. Participants and design

The study adopts an online longitudinal mixed experimental design involving 240 adults from Chinese and English speaking cultural backgrounds, with 120 participants each. Participants are randomly assigned to linguistic warnings, numerical warnings, or no warnings, with 80 per group. Each completes four rounds of 12 questions, totaling 48, covering factual judgments, health knowledge, legal knowledge, and consumer decisions. Participants provide initial judgments, view preset LLM answers and warnings, and submit final judgments. The answer bank includes correct, incorrect, and hallucination risk responses, with approximately 30% detectable errors. Repeated interactions enable analysis of reliance across warning formats, answer correctness, rounds, and cultural backgrounds.

DOI: 10.54254/2755-2721/2026.39107

3.2. System implementation

Reliance behavior is a binary outcome. Whether users finally accept the LLM answer can be estimated through a generalized linear mixed model. The model includes uncertainty expression, interaction round, cultural background, and individual cultural psychology variables [10]. It also includes participant and item random effects to control individual differences and task difficulty. The estimation of reliance probability is shown in Formula 1.

$$\log it\left[P(R_{ijt}=1)\right] = \beta_0+\beta_1U_j+\beta_2T_t+\beta_3C_i \\ +\beta_4(U_j\times T_t)+\beta_5(U_j\times C_i)+\beta_6P_i+u_i+v_q \tag{1}$$

In the formula, R_{ijt} indicates whether participant i accepts the LLM answer under expression condition j and interaction round t. U_j indicates uncertainty expression format. T_t indicates interaction round. C_i indicates cultural background. P_i indicates individual psychology variables. u_i and v_q indicate participant random effects and item random effects.

Responsibility attribution and the explanatory power of individual psychology are further examined through model comparison. A baseline model that only includes nationality is first built. Then uncertainty avoidance, institutional trust, and relational orientation are added [11]. The change in explanatory power is compared. The incremental explanatory power is shown in Formula 2.

$$\Delta R^2=R^2(Y\sim N_i+UA_i+IT_i+RO_i)-R^2(Y\sim N_i) \tag{2}$$

In the formula, Y indicates reliance rate, responsibility attribution, or trust score. N N_i indicates nationality. UA_i indicates uncertainty avoidance. IT_i indicates institutional trust. RO_i indicates relational orientation. ΔR^2 is used to determine whether individual cultural psychology variables provide stronger explanatory power than nationality.

3.3. Measures and analysis

The web platform records cultural background, AI experience, statistical literacy, and baseline trust before testing. During tasks, it logs initial and final judgments, LLM responses, warning types, acceptance, verification clicks, revisions, dwell time, and correctness. Following incorrect or potentially hallucinated answers, participants assign responsibility to themselves, the AI model, and the system provider using seven point scales. After four task rounds, uncertainty avoidance, institutional trust, and relational orientation are measured. Records with abnormal completion times, repetitive responses, or failed attention checks are excluded. Behavioral logs assess reliance changes, attribution ratings examine responsibility, and cultural measures test moderation and explanatory power.

4. Results

4.1. Reliance behavior

Findings from simulated pilot studies indicate different levels of reliance according to uncertainty expression. As can be seen from Table 1 below, none of the uncertainty expressions gave the highest reliance (72.4%), but gave the lowest appropriate reliance (58.6%) and the highest incorrect

DOI: 10.54254/2755-2721/2026.39107

acceptance (38.9%). Linguistic uncertainty lowered overall reliance to 61.3% but raised appropriate reliance to 70.2% and had a 36.5% verification and 27.1% revision rate. Numerical uncertainty resulted in 67.5% appropriate reliance and 27.9% incorrect acceptance and was more or less in the middle of the other uncertainty expressions. These findings suggest that uncertainty expression can improve reliance calibration by discouraging acceptance of incorrect answers while preserving appropriate reliance on correct responses.

Table 1. Differences in reliance behavior under three types of uncertainty expression

Expression condition	Overall reliance rate	Appropriate reliance rate	Incorrect acceptance rate	Verification rate	Revision rate	Mean dwell time
No uncertainty expression	72.4%	58.6%	38.9%	18.7%	12.9%	24.6 s
Linguistic uncertainty	61.3%	70.2%	24.7%	36.5%	27.1%	31.8 s
Numerical uncertainty	64.8%	67.5%	27.9%	30.8%	22.4%	29.5 s

4.2. Responsibility attribution

Responsibility attribution after failure is an indicator of how uncertainty expression affects the perception of users regarding the faults of AI. According to Table 2, no expression generated the highest AI model responsibility attribution score of 5.46 ± 1.08, implying that users perceived their failures due to model faults. Linguistic uncertainty increased the self responsibility attribution score to 4.38 ± 1.21, meaning that users had more knowledge of their duties of verifying the results. Numerical uncertainty yielded the highest system provider responsibility attribution score of 5.18 ± 1.02, which shows that confidence scores were perceived as the judgment of the platform. These attribution differences provide a basis for examining trust decline, recovery, and subsequent reliance adjustments following AI failures.

Table 2. Responsibility attribution scores in failure situations

Condition	Self responsibility	AI responsibility	Provider Responsibility	Trust Decline
No uncertainty expression	3.21 ± 1.37	5.46 ± 1.08	4.28 ± 1.22	1.42 ± 0.71
Linguistic uncertainty	4.38 ± 1.21	4.92 ± 1.15	4.66 ± 1.09	0.96 ± 0.64
Numerical uncertainty	3.72 ± 1.26	4.71 ± 1.19	5.18 ± 1.02	1.08 ± 0.67

4.3. Cultural moderation

The role of culture in interpreting warning expressions emerged as follows. As illustrated in Figure 1, warnings in the form of language resulted in more appropriate reliance by Chinese participants (72.8%) compared to English-speaking participants (67.9%). In contrast, warnings in numerical format led to more appropriate reliance by English-speaking participants (70.4%) compared to Chinese participants (65.1%). This could imply that language warnings operate as cooperative prompts while numerical warnings act as clear indicators of risk. There was also an interaction effect involving the expression format and the culture variable on appropriate reliance and verification click rates.

DOI: 10.54254/2755-2721/2026.39107

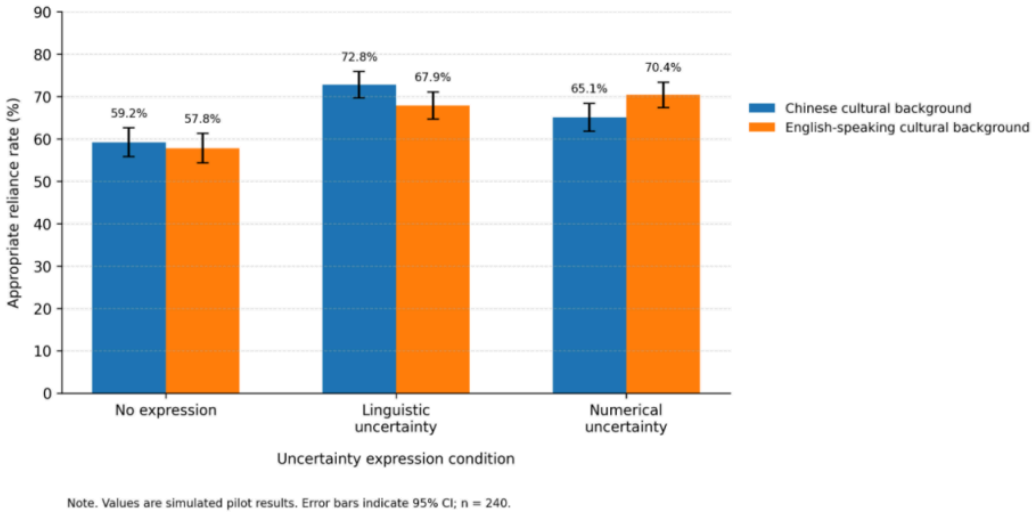

Figure 1. Appropriate reliance rate under uncertainty expression and cultural background

4.4. Individual cultural psychology

The individual psychological variables of cultural psychology offer a better explanation for adjustment in reliance than just nationality. As seen from Figure 2 below, the pseudo R^2 goes from 0.061 for nationality alone to 0.118 when uncertainty avoidance is included, to 0.147 when institutional trust is included, and to 0.172 when relational orientation is included. Uncertainty avoidance has a positive correlation with higher click rates and lower acceptance of false risks after warning. Institutional trust has a positive correlation with the acceptance of guidance and attribution of responsibility after numerical warnings.

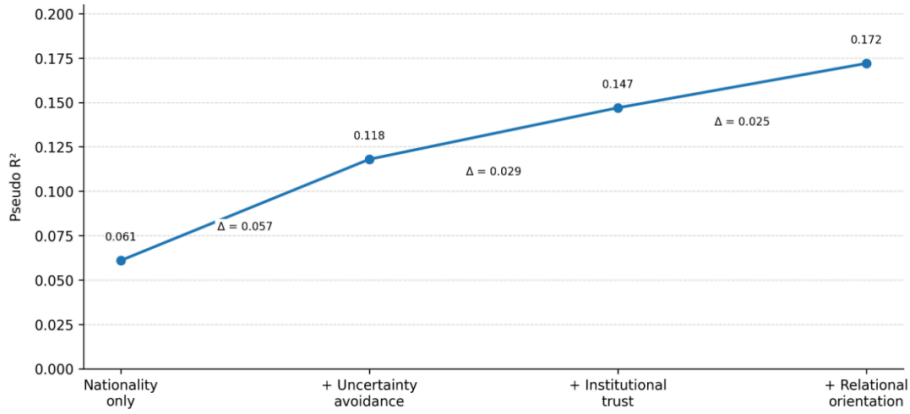

Figure 2. Incremental explanatory power of individual cultural psychology variables

5. Discussion

The simulated pilot results show that uncertainty expression changes not a single level of trust, but the behavioral path users follow when facing error risk. In the no expression condition, the incorrect acceptance rate is higher, which suggests that fluent outputs still strengthen surface credibility. Linguistic warnings increase verification and revision, indicating that natural language risk statements can activate users' awareness of their own verification responsibility. Numerical warnings reduce some incorrect acceptance, but they also lead users to assign more responsibility to the

DOI: 10.54254/2755-2721/2026.39107

system provider. This means that confidence scores may be interpreted as platform commitments. The interaction effect of cultural background further shows that warning formats need to match how users understand risk, responsibility, and cooperative communication.

6. Conclusion

The framework integrates LLM hallucination risk, uncertainty expression, responsibility attribution, and cultural psychology variables into one experimental process to examine trust calibration in repeated human LLM interaction. The web based experiment records changes from initial judgment to final acceptance through a controlled answer bank, four rounds of tasks, and behavior logs. Responsibility attribution scales and cultural psychology scales are used to explain the mechanism of reliance adjustment. The results show that linguistic and numerical warnings can both reduce overreliance, but they work through different paths. This framework provides an operational method for LLM risk warning design, trustworthy interface evaluation, and cross cultural human AI interaction.

References

[1] Farquhar, S., et al. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*(8017), 625–630.
[2] Steyvers, M., et al. (2024). The calibration gap between model and human confidence in large language models. *arXiv preprint arXiv:2401.13835*, 4–17.
[3] Augenstein, I., et al. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence, 6*(8), 852–863.
[4] Massenon, R., et al. (2025). "My AI is lying to me": User-reported LLM hallucinations in AI mobile apps reviews. *Scientific Reports, 15*(1), 30397.
[5] Huang, L., 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.
[6] Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics, 13*(5), 856.
[7] Asai, A., et al. (2024). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations, 2024*.
[8] Rapp, A., Di Lodovico, C., & Di Caro, L. (2025). How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations. *International Journal of Human-Computer Studies, 198*, 103471.
[9] Larsen, A. G., et al. (2025). LLM hallucinations in conversational AI for customer service: Framework and end-user perceptions. *International Journal of Human–Computer Interaction*, 1-22.
[10] Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, 108352.
[11] Visser, R., et al. (2025). Trust, distrust, and appropriate reliance in (X) AI: A conceptual clarification of user trust and survey of its empirical evaluation. *Cognitive Systems Research, 91*, 101357.