Exploring the Impact of Explainability in Large Language Model (LLM) Applications on User Experience

Yanyun Wang
Alibaba Cloud computing
Hangzhou, China
yanyun.wyy@alibaba-inc.com

Xumei Fang
Alibaba Cloud Computing
Hangzhou, China
xumei.fangxm@alibaba-inc.com

Zan Xu
Alibaba Cloud computing
Hangzhou, China
xuzan9281@163.com

Jianye Li
Alibaba Cloud Computing
Beijing, China
jianye.ljy@alibaba-inc.com

Luping Wang
Alibaba Cloud Computing
Hangzhou, China
dopink@msn.com

Abstract

Due to the "black-box" nature, explainability has long been a significant research topic in machine learning. Researchers have been committed to explaining model principles behind models and their scope of influence and decision-making to experts and technical practitioners. However, with the increasing popularity of the Large Language Model (LLM), more general users interact with these applications, bringing new challenges for explainability. This study explores the impact of LLM explainability on trust and satisfaction, revealing that both are significantly influenced by the degree and presentation of explainability. Moreover, trust and satisfaction vary across different risk scenarios. The study further evaluates the pros and cons of different explainability strategies, offering practical insights for the design of LLM applications.

CCS Concepts

• **Human-centered computing** → Interaction design; Interaction design theory,concepts and paradigms; Human computer interaction(HCI); Interaction paradigms; Natural language interfaces; • **Human- centered computing** → Interaction design; Empirical studies in interaction design; Human computer interaction(HCI); Empirical studies in HCI; • **Human-centred computing** → Human computer interaction(HCI); HCI design and evaluation methods; User studies.

Keywords

LLM User experience, Satisfaction, Explainability, Trust

ACM Reference Format:

Yanyun Wang, Xumei Fang, Zan Xu, Jianye Li, and Luping Wang. 2025. Exploring the Impact of Explainability in Large Language Model (LLM) Applications on User Experience. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3719941

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3719941

1 Introduction

Recent breakthroughs in large model technologies, like GPT, have transformed human-AI interactions, gaining significant attention. Despite the powerful potential of these models, challenges remain, including the "black-box" nature [1], ethical concerns [2–4], data privacy [5–8], and energy consumption [9]. Early research on the black-box problem focused mainly on explaining model decisions to experts and professionals [1, 10–13].

Recently, more general users have begun using Large Language Model (LLM) applications to discuss diverse topics. After researching mainstream LLM applications, we found that they vary in output presentations. Specifically, some applications only display the result, while others include reasoning processes or other explainability information.

This raises the question of whether reasoning processes and explainability information benefit users and impact their trust and satisfaction during the experience [14], issues that remain underexplored in existing LLM research.

To address these issues, we conducted a study to explore how explainability should be presented to general users and its impact on trust and satisfaction. This study involves thirty-two participants from China who engaged in 12 simulated dialogue scenarios. Their trust and satisfaction were measured using a seven-point Likert scale, followed by semi-structured interviews to understand participants' perceptions of LLM explainability and their varying needs.

2 Related work

2.1 The Development of LLM Explainability

In machine learning, various terms like explainability [1, 13], interpretability [1, 13], comprehensibility, intelligibility, transparency, and understandability have been used to address the black-box problem. In this paper, we will uniformly use the term "explainability" to refer to this process [14–16].

Most recently, understanding explainability has become prominent with the widespread adoption of complex models [17, 18, 20]. For instance, Caruana et al. [19] highlighted that the explainability of LLM has supported clinicians in making transparent decisions. Moreover, the Local Interpretable Model-agnostic Explanations (LIME) method introduced by Ribeiro et al. [21] has generated local explanations for each prediction, extensively increased trust.

Subsequently, a game-theoretic explanation framework, SHapley Additive exPlanations (SHAP), was proposed to standardise the interpretation of model outputs, signifying a shift towards a more systematic direction in explainability research [22]. Lastly, Kizilcec [23] offered valuable insights into explainability and user experience interplay by examining how trust was enhanced by moderate levels of explainability in AI decisions [1, 24].

While interacting with LLM models, although explainability is considered a key factor in improving trust and understanding [24, 25], existing research primarily focused on high-risk sectors (e.g., healthcare and law) [15, 19, 24, 26]. As LLMs expand across sectors, public explainability has become a pressing priority [24, 27–29].

2.2 Explainability Methods, Presentation Elements, and Influencing Factors

Explainability methods in machine learning include Global and Local Explanations, as well as Intrinsic and Post-hoc Interpretability [11, 14, 15, 21]. These various explainability methods facilitate users' understanding of how models work and the reasons behind their predictions. Individual method corresponds to distinct aspects of the model's outputs, including the context of model decisions [16, 30], the decision-making process [1, 16], the decision outcomes [19, 21], and the supporting evidence [10].

In addition to explainability methods, model explainability is influenced by factors such as model complexity, user background knowledge, and application risk level. Generally, more complex models are harder to explain [11, 21]. Different users (e.g., data scientists, business personnel, or end-users) have varying needs for explainability [23]. Doshi-Velez and Kim [1] categorised scenarios as low-risk, where users rely on model stability and feedback [21], and high-risk, such as healthcare or law, where explainability is crucial to mitigate serious decision-making errors [26, 27, 29, 31, 32].

This study examines three factors influencing explainability: model complexity, user background, and scenario risk. It focuses on explainability outputs, including decision context, processes, outcomes, and supporting evidence, to evaluate their impact on user experience.

2.3 Methods for Evaluating Explainability

Model explainability depends on both its inherent features and users' ability to understand them [10, 14], making user involvement essential in evaluation. Doshi-Velez and Kim [1] identified three evaluation methods, with human-centered evaluation [33–35] – involving interactions between non-expert users and the system – is highly relevant to this study.

Trust and satisfaction are widely recognised as key metrics for assessing explainability [10, 37]. Jacovi et al. [36] highlighted their importance in evaluating explanation quality, while Hoffman et al. [25] emphasised trust as a core goal of explainability. Trust reflects system transparency and reliability, while satisfaction measures how well explanations meet user needs. Thus, trust and satisfaction will be the primary metrics in this study.

2.4 Hypotheses

Figure 1 presents the three-level theoretical framework of this study. The first level identifies the factors influencing explainability (user background, model complexity, and scenario risk). The second level focuses on the content of the model's explanation outputs (decision context, process, outcome, and supporting evidence). The final level consists of two core metrics for evaluating the quality of explanations: trust and satisfaction. The following sections will outline the experimental design based on this framework.

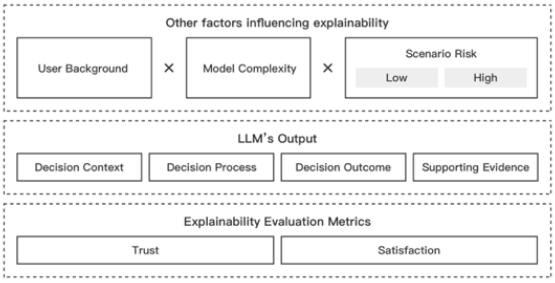

Figure 1: The Theoretical Framework

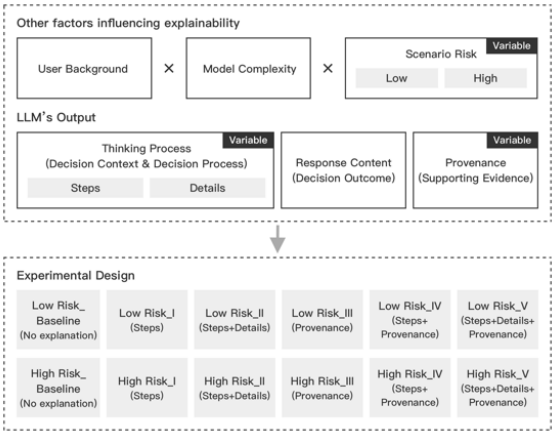

Figure 2: Experimental Design Framework

This study explores the impact of different interpretability schemes on trust and satisfaction in LLM applications, proposing the following hypotheses based on the theoretical framework.

H1: It is hypothesised that there are significant differences in trust (1-a) and satisfaction (1-b) with the LLM's output when provided with different explainability schemes.

H2: It is hypothesised that there are significant differences in trust (2-a) and satisfaction (2-b) with the LLM's output when the same explainability scheme is presented under different scenario risks.

H3: It is hypothesised that there is a positive correlation between trust and satisfaction with the LLM's output.

3 Methods

3.1 Experimental Design

Figure 2 presents the experimental design based on the theoretical framework, followed by a detailed experimental design.

3.1.1 Participants. To ensure a foundational understanding of LLM applications, 32 participants aged 25-35 with prior LLM experience were recruited. All participants reported using Chatbots more than three times per week on average over the preceding month, which ensures a consistent level of familiarity and engagement with the LLM. The group was evenly divided by gender (16 males and 16 females) to maintain gender balance, minimising potential gender biases in the results. The average age was 31.19 years for females (SD = 3.23) and 31.94 years for males (SD = 2.82).

The current study adhered to rigorous ethical standards to ensure ethical integrity of the research process. First, informed consent was obtained from all participants following a comprehensive explanation of the study's objectives, procedures, and methodology. Participants explicitly confirmed their voluntary participation and consented to the use of their data for research purposes. Second, data collection was conducted anonymously to protect participant privacy. Specifically, participants had the option to decline to answer any sensitive questions (e.g., age, gender, ethnicity), with no personally identifiable information (e.g., name, address) being collected. All data were securely stored on protected servers with robust security protocols to prevent unauthorised access, leakage, or misuse.

3.1.2 Experimental Variables. In this chapter, we discussed two experimental variables, scenario risk and model output. To ensure consistency between the response content, we pre-generated the answers to each scenario using the Qwen2-72b LLM. Subsequently, we created simulated dialogue videos to ensure that all factors other than the experimental variables remained consistent (e.g., model complexity, dialogue content, system response speed, etc.). The Chatbot interface is incorporated into the created dialogue videos to ensure participants' familiarity [33, 38].

Scenario risk: In terms of dialogue scenario selection, we classified multiple scenarios based on their risk levels (from low to high). According to this classification, we selected two scenarios that aligned with participants' daily experience while characterised by a markedly different risk level: Low-Risk – Weather inquiry (W) and High-Risk – Medical emergency inquiry (M).

The model outputs were varied as Thinking Process and Provenance. Specifically, Thinking Process involved Steps (outlining the LLM's thought process) and Details (providing specific content related to the LLM's reasoning). Moreover, Provenance included reference information added after the LLM's response was complete.

3.1.3 Experimental Conditions. For each scenario, one control group and five experimental groups were established, incorporating a comparative method to explore differences (Refer to Appendix Figure 8):

- Baseline: No explanation (user query and LLM's response only, referred to as the control group.)
- Group I: Steps (Only the additional content relative to the control group is listed here; the same applies throughout.)
- Group II: Steps + Details
- Group III: Provenance
- Group IV: Steps + Provenance
- Group V: Steps + Details + Provenance

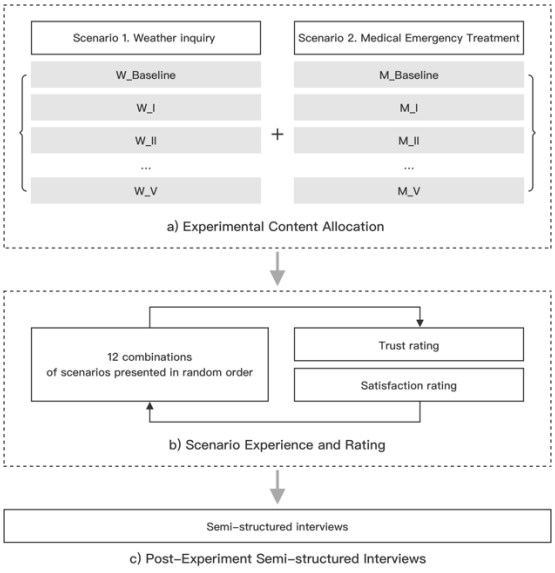

Figure 3: Experimental Procedure

3.2 Experimental Procedure

Figure 3 illustrates the comprehensive experimental process of this study. First of all, all eligible participants were provided with an overview of the content and distinctions between the 12 dialogue videos across the two scenarios. Participants were then instructed to assume the role of the questioner in the simulated dialogue videos. The 12 dialogue videos from two scenarios were presented in a random order. Participants engaged sequentially with each scenario pair and rated their trust and satisfaction using a seven-point Likert scale. After rating all experimental pairs, participants completed a supplementary interview.

4 Analysis and Results

In this section, we analysed hypotheses H1 to H3, combining statistical analysis with direct quotes from user interviews. Prior to testing the hypotheses, we conducted a variance analysis on participants' gender (male, female) and their ratings of trust and satisfaction. The results showed no significant differences between male and female participants in trust ($F = 0.088$, $P = 0.767$) and satisfaction ($F = 0.413$, $P = 0.521$). This finding was further confirmed during the interviews, where no noticeable differences in rating logic were reported. Therefore, gender will not be considered a categorical variable in the subsequent hypotheses analysis.

4.1 The Impact of Different Explainability Schemes on Trust and Satisfaction (H1)

4.1.1 Trust (H1-a). To assess the relationship between interpretability schemes and trust, we conducted normality tests. Although the data were not strictly normal (Shapiro-Wilk test), the kurtosis ($|-0.407| < 10$), skewness ($|-0.62| < 3$), and histogram bell-shaped features indicated that the normal distribution assumption

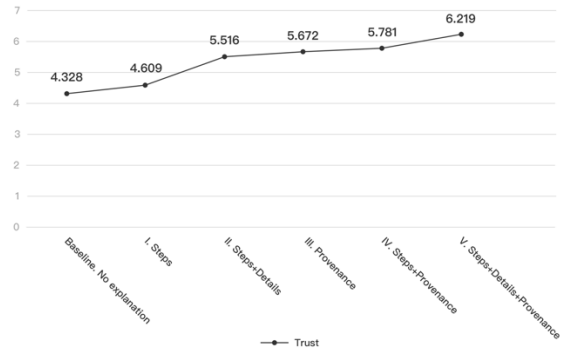

Figure 4: Trust trend

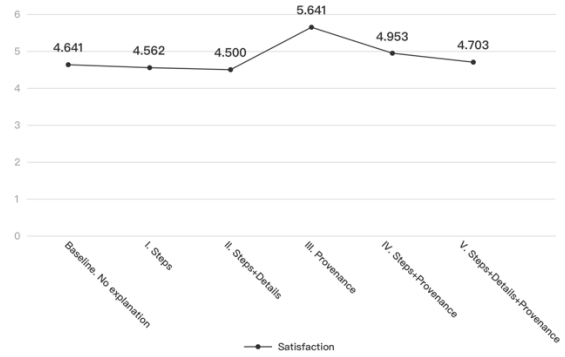

Figure 5: Satisfaction trend

was largely satisfied. One-way ANOVA showed that different interpretability schemes had a significant effect on trust scores (F = 23.215, P = 0.001***), which was also supported by the Welch's ANOVA test (F = 21.738, P = 0.001***). Therefore, hypotheses H1-a was verified.

Respectively comparing the trend graphs (see Figure 4 and Appendix Table 1) of the mean trust scores for (Baseline, Group I, Group II) and (Group III, Group IV, Group V), it can be observed that as the amount of explainable information increased, overall trust showed a noticeable upward trend in each comparison. User interviews supported this, with User P16 noting, 'Seeing the thinking process makes me feel more secure; it seems like the model is genuinely thinking, which effectively enhances my trust.'

We analysed the sample with post hoc tests. Adding thinking steps alone (Baseline vs. Group I) had little impact on trust (P = 0.987 > 0.05), but including detailed information (Baseline vs. Group II) proved more effective (P = 0.001***). User P10 commented, 'The steps are ineffective information that makes it feel like a mechanical process.'

Increasing the amount of provenance information (comparing Baseline with Group III, Group I with Group IV, and Group II with Group V) effectively enhanced trust (P = 0.000***, P = 0.000***, P = 0.007***). Subjective interviews indicated that more than half of the participants felt that provenance improved their trust more than the thinking process. User P2 noted, 'Compared to the thinking process, provenance information that follows the output results better conveys trust and certainty. Its presentation is more concise and effective.'

4.1.2 Satisfaction (H1-b). To assess the relationship between interpretability schemes and satisfaction, we conducted normality tests. Although the data were not strictly normal, the kurtosis (|-0.546| < 10), skewness (|-0.423| < 3), and histogram bell-shaped features indicated that the normal distribution assumption was largely satisfied. One-way ANOVA showed that different interpretability schemes had a significant effect on satisfaction scores (F = 6.05, P = 0.000***), which was also supported by the Welch's ANOVA test (F = 7.571, P = 0.000***), verifying hypotheses H1-b.

Combining the Post hoc test analysis and the satisfaction graph (see Figure 5 and Appendix Table 2) of the mean of trust, it can be observed that only the provenance scheme (Group III) demonstrated the highest satisfaction, as well as a significant difference to all other schemes (Baseline: p = 0.000***, Group I: p = 0.000***, Group II: p = 0.000***, Group IV: p = 0.031**, Group V: p = 0.007***). The primary reason for this preference lies in the perception that the thinking process increased the perceived waiting time, which caused users to feel anxious and pressured (as noted by users P1 and P3). In contrast, provenance provided clear and direct content without requiring users to observe intermediate steps, offering more control.

4.2 The Impact of Different Scenario Risks on Trust and Satisfaction (H2)

4.2.1 Trust (H2-a). To assess the relationship between scenario risks and trust, we conducted normality tests. The data basically conformed to the normal distribution (the kurtosis |-0.407| < 10, skewness |-0.62| < 3). The T-test analysis demonstrated that scenario risks had a significant effect on trust, confirming hypotheses H2-a (t = 2.085, P = 0.038**).

For the control group (Baseline), trust in the low-risk scenario (W) was significantly higher than that of in the high-risk scenario (M) (t = 3.32, P = 0.002**). This could be due to that low-risk (W) scenarios were perceived as simpler, requiring less accuracy, while high-risk scenarios (M) involved specialised fields where additional information was needed to assess accuracy, leading to lower trust.

The overall trend (see Figure 6 and Appendix Table 3) showed a lower trust on all the high-risk scenarios (M) compared to the low-risk scenarios (W), potentially because users were more cautious about the high-risk scenarios (M). For instance, comparing the control group (Baseline) and the Group V, it can be found that in the control group, where no explainability content was provided, the difference in trust between two scenarios was significantly large (t = 3.32, P = 0.002***); whereas, in the Group V, where explainability content was provided to the largest extent, trust had converged (t = -0.466, P = 0.645), indicating an increasing of explainability content may enhance trust under the high-risk scenario.

4.2.2 Satisfaction (H2-b). In terms of the relationship between scenario risks and satisfaction, the data basically conformed to a normal distribution (kurtosis |-0.546| < 10, skewness |-0.423| < 3). Upon a T-test analysis, scenario risks did not have a significant effect on satisfaction (t = 0.143, P = 0.887). Therefore, H2-b was not established. However, pairwise T-tests for specific explainability

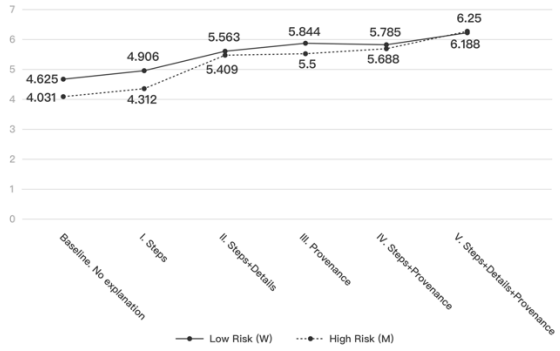

Figure 6: Trust trend

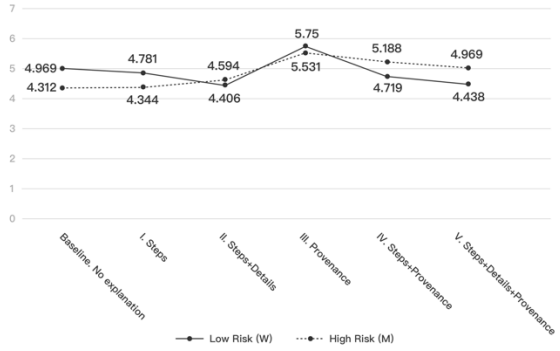

Figure 7: Satisfaction trend

schemes revealed significant differences in the satisfaction scores for the Baseline (t = 2.782, P = 0.009***), Group I (t = 1.951, P = 0.060*), Group IV (t = -2.089, P = 0.045**), and Group V (t = -2.237, P = 0.033**).

The trend graph (see Figure 7 and Appendix Table 4) comparing the control group (Baseline) with Groups I and II revealed that for high-risk scenarios (M), satisfaction increased as more explainability information was provided, indicating that users needed more details in these contexts. However, for low-risk scenarios (W), satisfaction declined with excess information. Interviews revealed that participants in low-risk scenarios (W) preferred faster, precise answers, while too much information reduced efficiency and caused distractions, lowering satisfaction. The trend graph showed that adding thinking process further decreased satisfaction once provenance was introduced, especially in low-risk scenarios (W).

4.3 Correlation Between User Trust and Satisfaction with LLM's Outputs (H3)

Pearson correlation analysis (see Appendix Table 5) revealed a significant positive correlation between trust and satisfaction, with a correlation coefficient of 0.451 (P < 0.05), confirming hypothesis H3. This finding suggested that increased trust positively contributed to higher satisfaction. However, user interviews revealed that satisfaction in LLM interaction scenarios is influenced by multiple factors beyond trust, including response time, output speed, content quality, and the presentation of results. Thus, although trust increased in some experimental results, satisfaction levels may have fluctuated. Overall, improving trust in the conversational content of LLMs is likely to enhance the user experience.

5 Discussion

5.1 Key Findings

We find that the degree and methods of explainability significantly influence trust. Greater transparency in explainable information enhance trust, with provenance being the most impactful factor. In high-risk scenarios, the need for explainability is even more pronounced, leading to a more substantial increase in trust.

The degree and methods of explainability also affect satisfaction. Unlike trust, an excess of explainable information is not necessarily relevant to higher satisfaction; instead, it can overwhelm users, particularly in low-risk scenarios. Introducing provenance alone provides the highest satisfaction levels.

Trust is essential for improving user satisfaction in LLM applications, as a positive correlation exists between the two. Future experience designs should prioritise fostering trust. Additionally, personalised explainability schemes tailored to specific dialogue contexts should be considered rather than adopting a one-size-fits-all approach.

5.2 Limitations and Future Directions

This study primarily focuses on experienced LLM users aged 25-35 in Zhejiang, China, which excludes individuals lacking LLM experience, and neglects the potential influence of cultural and geographical differences. The reliance on only two scenarios to represent high- and low-risk classifications may introduce some bias. Future studies should investigate the effects of LLM explainability on users with varying levels of LLM experience. It is also recommended that a broader range of scenarios for high- and low-risk contexts are incorporated. For instance, low-risk scenarios could involve entertainment-related task scenarios (e.g., travel planning) and daily-life activity scenarios (e.g., food recommendations), while high-risk scenarios might include academic-related task scenarios (e.g., literature searches), and industrial application scenarios (e.g., autonomous vehicle systems or legal consultations). Incorporating these scenarios in future research will effectively enhance generalisability of the research outcomes. Furthermore, to minimise the influence of confounding factors, the experiments were limited to basic text content, which somewhat excluded factors such as multimodality [39, 40], and Graphical User Interface designs [1, 41].

While this research focuses on explainability, our future research direction will focus on examining other contributing factors.

6 Conclusion

In conclusion, this study has made the following contributions:

We provided empirical evidence demonstrating that the degree and methods of explainability significantly influenced trust and satisfaction, with varying impacts across different scenarios.

It confirmed a positive correlation between trust and satisfaction, highlighting trust as a key factor influencing satisfaction.

The insights derived from this study offered a foundation for future investigations into how LLM applications can enhance trust and provide more personalised services, thereby guiding further developments in the field.

References

[1] Finale Doshi-Velez, Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (Mar 2017). https://doi.org/10.48550/arXiv.1702.08608

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, New York,610-623. https://doi.org/10.1145/3442188.3445922

[3] Hovy Dirk, Shannon L. Spruit. 2016. The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). ACL, Berlin, Germany, 591-598. https://doi.org/10.18653/v1/P16-2096

[4] Crawford Kate, Ryan Calo. 2016.There is a blind spot in AI research. Nature 538 (October 2016), 311-313. https://doi.org/10.1038/538311a

[5] Haoran Li, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, Zenglin Xu, Bryan Hooi, Yangqiu Song. 2023. Privacy in large language models: Attacks, defenses and future directions. arXiv preprint arXiv:2310.10383 (Oct 2023). https://doi.org/10.48550/arXiv.2310.10383

[6] Veale Michael, Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society 4, 2 (November 2017), https://doi.org/10.1177/2053951717743530

[7] Mohammad Al-Rubaie, John M. Chang. 2019. Privacy-preserving Machine Learning: Threats and Solutions. IEEE Security & Privacy 17,2 (March 2019), 49-58. https://doi.org/10.1109/MSEC.2018.2888775

[8] Carolyn Abbot. 2012. Bridging the Gap - Non-state Actors and the Challenges of Regulating New Technology. Journal of Law and Society, Vol. 39, 3 (Sept. 2012), 329-358. https://doi.org/10.1111/j.1467--6478.2012.00588.x

[9] Emma Strubell, Ananya Ganesh, Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 3645–3650. https://doi.org/10.18653/v1/P19-1355

[10] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Turin, Italy, 80-89. https://doi.org/10.1109/DSAA.2018.00018

[11] Christoph Molnar. 2020. Interpretable machine learning: A guide for making black box models explainable. Leanpub.

[12] Diogo V. Carvalho, Eduardo M. Pereira, Jaime S. Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8,8 (July 2019), 832. https://doi.org/10.3390/electronics8080832

[13] Alejandro B. Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion 58 (June 2020), 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[14] Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16,3 (May 2018), 31-57. https://doi.org/10.1145/3236386.3241340

[15] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi. 2018. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51, 5(August 2018), 1-42. https://doi.org/10.1145/3236009

[16] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267 (February 2019), 1-38. https://doi.org/10.1016/j.artint.2018.07.007

[17] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information fusion 99 (November 2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[18] Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. Digital signal processing 73 (February 2018), 1-15. https://doi.org/10.1016/j.dsp.2017.10.011

[19] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, August 10 - 13, 2015, Sydney NSW Australia, Association for Computing Machinery, New York, NY, USA, 1721-1730. https://doi.org/10.1145/2783258.2788613

[20] Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296. (Aug 2017), 8 pages. https://doi.org/10.48550/arXiv.1708.08296

[21] Marco T. Ribeiro, Sameer Singh, Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, August 13 - 17, 2016, San Francisco California USA, Association for Computing Machinery, New York, NY, USA, 1135-1144. https://doi.org/10.1145/2939672.2939778

[22] Scott M. Lundberg, Su-In Lee. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems (NIPS 2017), Vol. 30. Long Beach, CA, USA. https://doi.org/10.48550/arXiv.1705.07874

[23] René F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). San Jose California USA, 2390-2395. https://doi.org/10.1145/2858036.2858402

[24] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 Chi conference on human factors in computing systems (CHI '18). Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3173951

[25] Robert Hoffman, Tim Miller, Shane T. Mueller, Gary Klein, William J. Clancey. 2018. Explaining explanation, part 4: a deep dive on deep nets. IEEE Intelligent Systems 33,3 (May 2018), 87-95. https://doi.org/10.1109/MIS.2018.033001421

[26] Erico Tjoa, Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems 32,11 (November 2021), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

[27] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai on behalf of the Precise4Q consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making 20. (November 2020), 1-9. https://doi.org/10.1186/s12911-020-01332-6

[28] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys 16,(January 2022), 1-85. https://doi.org/10.1214/21-SS133

[29] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, Ann Arbor, Michigan, USA, 359-380. https://proceedings.mlr.press/v106/tonekaboni19a.html

[30] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In Proceedings of the Natural language processing and Chinese computing: 8th CCF international conference, NLPCC 2019. dunhuang, China, 563-574. https://doi.org/10.1007/978-3-030-32236-6_51

[31] Ahmad Chaddad, Jihao Peng, Jian Xu, Ahmed Bouridane. 2023. Survey of explainable AI techniques in healthcare. Sensors 23,2 (January 2023). 634. https://doi.org/10.3390/s23020634

[32] Amitojdeep Singh, Sourya Sengupta, Vasudevan Lakshminarayanan. 2020. Explainable deep learning models in medical image analysis. Journal of imaging 6,6 (June 2020), 52. https://doi.org/10.3390/jimaging6060052

[33] Julien Colin, Thomas Fel, Remi Cadene, Thomas Serre. 2022. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. Advances in neural information processing systems vol.35 (Jun 2022), 2832-2845. https://doi.org/10.48550/arXiv.2112.04417

[34] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making? In Proceedings of the AAAI Conference on Artificial Intelligence Vol.35 No.8. AAAI Press, Palo Alto, California USA. 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

[35] Ben Shneiderman. 2022. Human-centered AI. Oxford University Press.

[36] Alon Jacovi, Ana Marasović, Tim Miller, Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Canada, 624-635. https://doi.org/10.1145/3442188.3445923

[37] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science 5, (February 2023), https://doi.org/10.3389/fcomp.2023.1096257

[38] Amon Rapp, Lorenzo Curti, Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. International Journal of Human-Computer Studies 151,(March 2021), 729-758, https://doi.org/10.1016/j.ijhcs.2021.102630

[39] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui

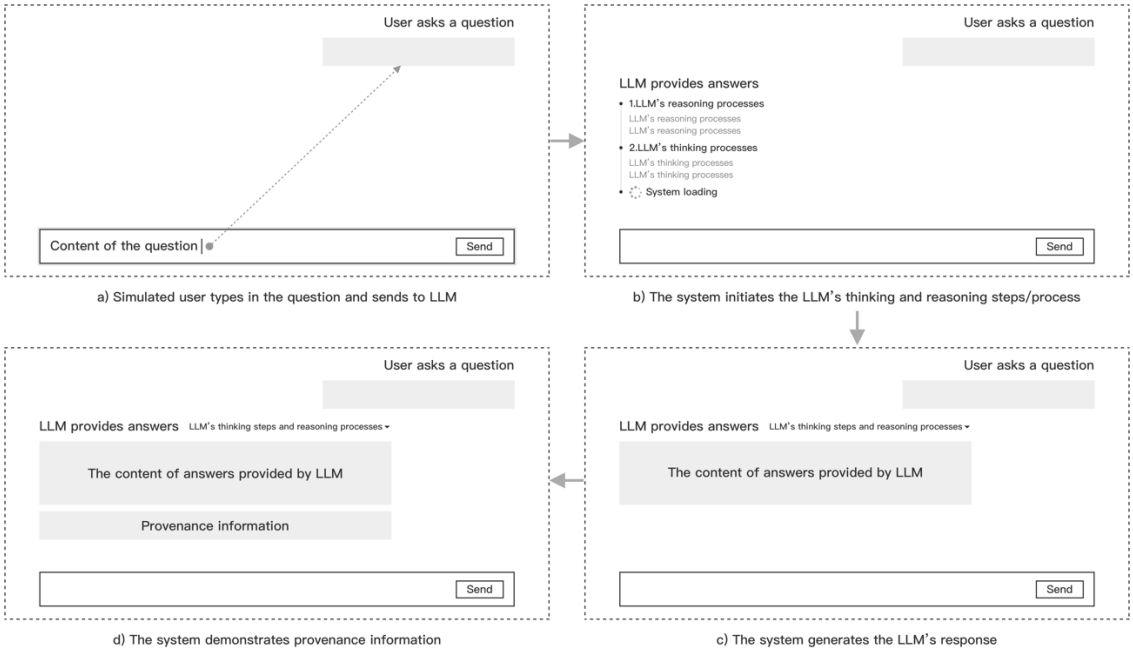

Figure 8: Schematic of Experimental Materials (Example: Group V)

Table 1: Trust analysis(H1-a)

Variable	Experimental Condition	Sample Size	Mean	Standard Deviation (SD)	Variance Test	Welch's Variance Test
Trust	Baseline	64	4.328	1.448	F=23.215	F=21.738
	Group I	64	4.609	1.305	P=0.000***	P=0.000***
	Group II	64	5.516	1.208		
	Group III	64	5.672	1.142		
	Group IV	64	5.781	1.119		
	Group V	64	6.219	1		

Xiong, Xuming Hu. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. arXiv preprint arXiv:2412.02104(Dec 2024). https://doi.org/10.48550/arXiv.2412.02104

[40] Nikolaos Rodis, Christos Sardianos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Iraklis Varlamis, Georgios Th. Papadopoulos. 2024. Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions. IEEE Access, vol. 12 (September 2024), 159794-159820. https://doi.org/10.1109/ACCESS.2024.3467062

[41] Saša Brdnik. 2023. GUI Design Patterns for Improving the HCI in Explainable Artificial Intelligence. In Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23) . Sydney, NSW, Australia. 240-242. https://doi.org/10.1145/3581754.3584114

A APPENDICES

A.1 Experimental material

Figure A.1 presents a schematic representation of the experimental materials used in this study. Specifically, taking Group V as an example, the participants were presented with simulated dialogue animation to mimic their real-life interactions with LLM applications.

A.2 Statistical data

Table 2: Satisfaction analysis(H1-b)

Variable	Experimental Condition	Sample Size	Mean	Standard Deviation (SD)	Variance Test	Welch's Variance Test
Satisfaction	Baseline	64	4.641	1.277	F=6.05	F=7.571
	Group I	64	4.562	1.194	P=0.000***	P=0.000***
	Group II	64	4.5	1.543		
	Group III	64	5.641	1.173		
	Group IV	64	4.953	1.302		
	Group V	64	4.703	1.725		

Table 3: Satisfaction analysis(H2-a)

Variable	Scenario Risks	Sample Size	Mean	Standard Deviation (SD)	T-test	Welch's T-Test
Trust	Low-risk	192	5.5	1.318	T=2.085	T=2.085
	High-risk	192	5.208	1.421	P=0.038**	P=0.038**

Table 4: Satisfaction analysis(H2-b)

Variable	Scenario Risks	Sample Size	Mean	Standard Deviation (SD)	T-test	Welch's T-Test
Trust	Low-risk	192	4.844	1.478	T=0.143	T=0.143
	High-risk	192	4.823	1.38	P=0.887	P=0.887

Table 5: Satisfaction analysis(H3)

Variable	Satisfaction	Trust
Satisfaction	1(0.000***)	0.451(0.000***)
Trust	0.451(0.000***)	1(0.000***)