# THE MORE YOU KNOW, THE LESS YOU TRUST


**Authors**

Edgar Graichen, Nils Herrling, Josephin Pohle, Tamta Sivsivadze & Johann Wiederhold


**Background**

With AI becoming indistinguishable from human output, brief explanations are a crucial way of addressing blind over-reliance and enabling safer, critically-minded AI interactions.

## INTRODUCTION

- Large Language Models (LLMs) like ChatGPT produce highly convincing language that frequently leads users to overestimate output accuracy (despite the risk of misinformation)
- creates a distinct vulnerability: trust relies on belief that unsupervised AI will perform reliably
- study addresses critical need to investigate user expectations with technical capabilities of LLMs

## THE CURRENT STUDY

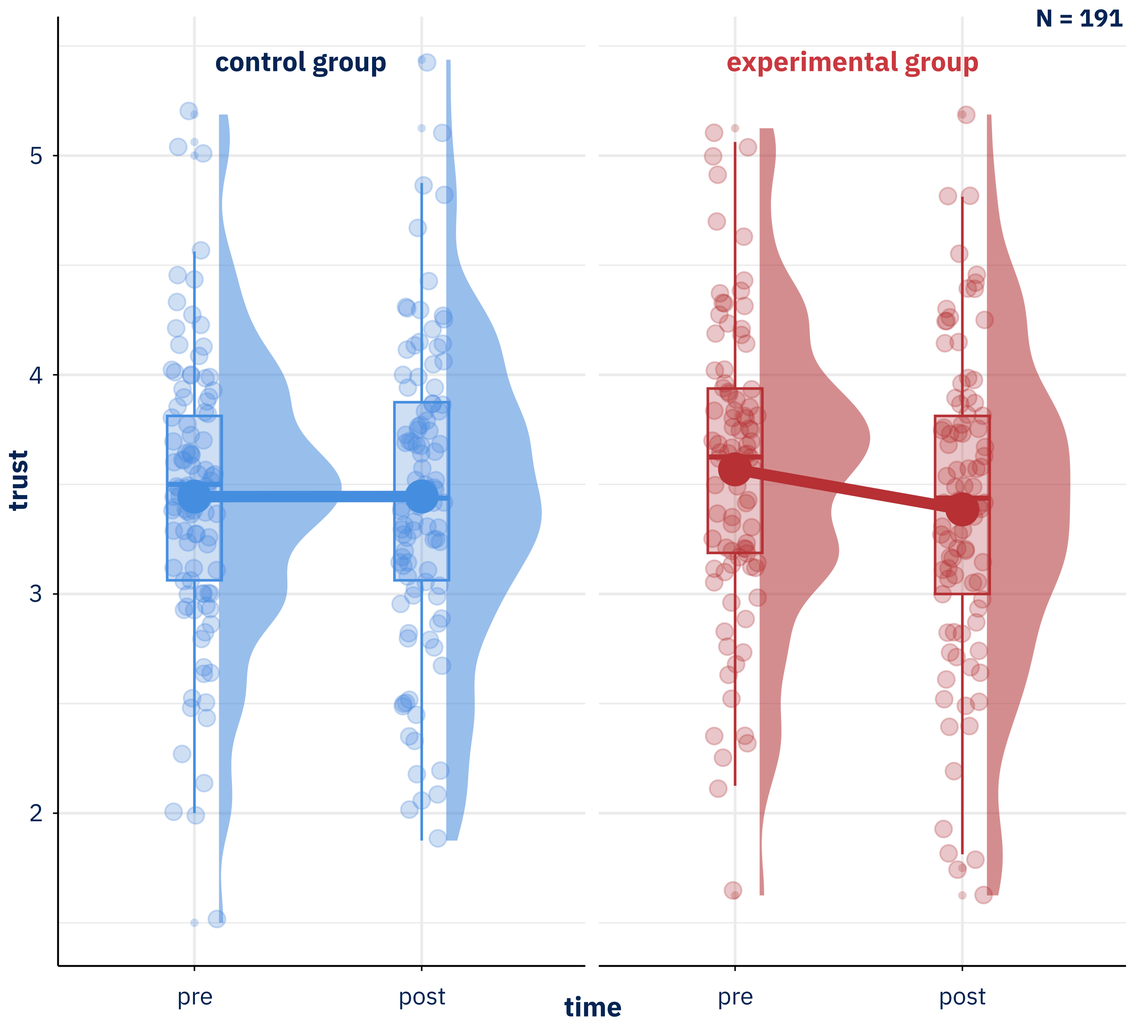
- **Core Objective:** Investigating the effect of a short explanatory instructional intervention about LLM functioning on users' trust.
- **Hypothesis**: Anticipating a significant Time (pre vs. post) x Group (experimental vs. control) interaction.
- **Expected Outcome:** A significant decrease in trust specifically within the experimental group over time.

1. Pre-intervention assessment of initial trust towards LLMs alongside other baseline variables like prior knowledge and usage frequency.
2. Randomly assigned experimental group read a short, instructional text explaining how LLMs function, while the control group read a text of similar length about the gut microbiome.
3. Afterwards, each group answered a short comprehension check in order to test whether each participant understood the contents of the text.
4. Each participant answered the trust attitude measurement instrument again.

## RESULTS

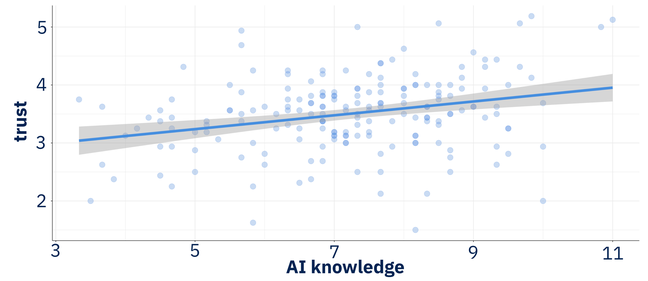
- mixed ANOVA and group comparisons confirmed a **significant Time x Group interaction** → showcasing that educational intervention caused **significant, moderate decrease in trust** (d = -0.65) while control group trust remained stable
- multiple regression analysis accounted for **29% of variance in baseline trust**, demonstrating that initial user trust is significantly driven by pre-existing individual factors → namely **prior AI knowledge, usage frequency, and task diversity**

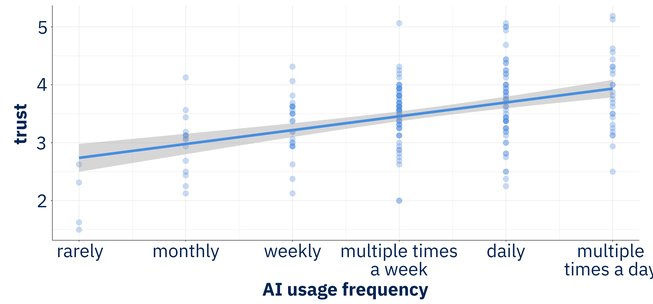
## Explaining the functionality of AI (LLMs) led to a significant, moderate decrease in user trust.

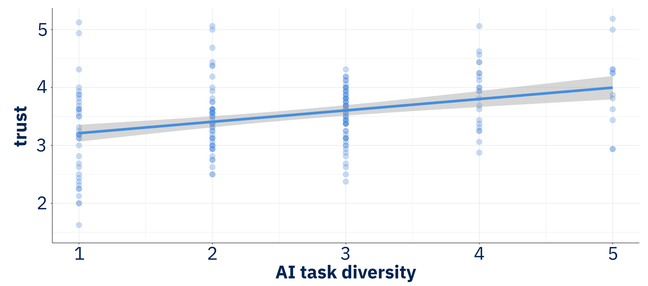



## Prior Knowledge of AI, Frequency of AI Use and Task Diversity are **significant predictors of trust** in AI (LLMs).

## SIGNIFICANT PREDICTORS OF TRUST




## CONCLUSION

- a brief explanation of LLM mechanics **significantly reduces** user trust over time compared to a stable control group → short interventions can lower overtrust and encourage more critical evaluations
- findings limited by immediate post-measurement, self-reported metrics, and slight measurement overlap → future research must investigate long-term behavioral impacts and stability of repeated interventions


**AI Trust Overview:** Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research.

**Measuring Trust Attitudes:** Larasati, R. (2025). Human and AI trust: Trust attitude measurement instrument. arXiv.

**AI Knowledge vs. Human Perception:** Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence



## Literature and Acknowledgements

We thank Dr. Mario E. Archila and the Department of Biological Psychology and Cognitive Neuroscience for providing academic leadership and support throughout this project.


Scan the QR code to access our full report →