When AI Flatters Too Much

AN EXPLORATORY STUDY INTO TRUST, PERCEIVED TRUSTWORTHINESS, AND OPINION FORMATION ON SIMULATED USERS

TU Delft

Melissa Hu | M.K.Y.Hu@student.tudelft.nl

Supervisors: Dr. Ujwal Gadiraju, Dr. Marije van Dalen, Esra de Groot, Shreyan Biswas

1. Introduction

- **Large Language Models (LLMs)** are increasingly used for advice, writing, and decision-making [1], [3].
- Key risk is **sycophancy:** models align with users beliefs, even at the cost of truthfulness [4].
- **64%** of young adults (aged 18-24) in European Union used LLMs in **2025** [2].
- This can affect how trustworthy users perceive the model and how their opinons develop.

2. Research Question

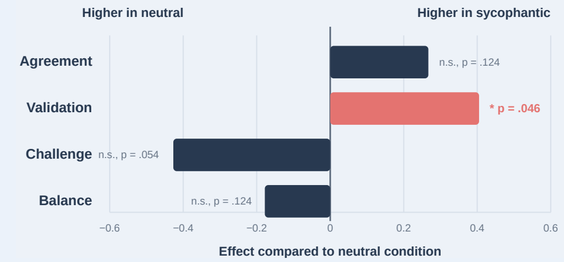
How does LLM sycophancy shape opinion formation and influence trust and perceived trustworthiness among simulated young adults?

3. Methodology

LLaMA by Meta

- 128 simulated young adults aged 18-25.
- Each user interacted with **Llama 3.1 8B** in a specific condition:
  - **Neutral:** balanced, evidence-based;
  - **Sycophantic**: agreeable, validates user opinions.
- Users discussed with model over two topics: **autonomous vehicles** and **artificial intelligence in society**.
- **Pre- & post-surveys** measuring trust, perceived trustworthiness, and opinion on 7-point Likert-scale.

4. Results



- Given manipulation checks, sycophantic model was mainly perceived as more validating.
  - **6.5%** neutral vs **54.3%** sycophantic (from qualitative analysis).
- It was not necessarily more agreeable or less balanced.

Neutral and sycophantic values are model-based marginal means.

Outcome	Neutral	Sycophantic	Result
Perceived trustworthiness	5.80	6.02	Significant increase, p = .014
Trust	4.22	4.43	Not significant, p = .128
Opinion change	1.15	1.28	Not significant, p = .424
Opinion strength	1.35	1.50	Not significant, p = .151
Opinion confidence	5.98	5.93	Not significant, p = .658

- Sycophantic condition showed no significant effect on trust, and opinion formation.
- It did show significant effect on perceived trustworthiness.

5. Limitations

- Some simulated users produced incomplete or invalid responses, reducing usable sample.
- Short interactions and ceiling effects may explain weak results in opinion change.
- Limited generalizability as only 2 topics and 1 model was used.

6. Future Work

- Conduct experiment with human participants to verify the results of this study.
- Replicate across multiple LLMs to see whether results can be generalized across them.
- Study **trust calibration** to see whether users adjust their trust appropriately given model's response.

7. Conclusion

- Sycophantic model mostly changed how the model was perceived, not what users believed.
- Effect seems to be mainly driven by validation/complimentary demeanor.
- Overall, sycophantic behavior increased perceived trustworthiness but did not significantly affect trust or opinion formation.
- This study is done with simulated users and should be viewed as an exploratory study.

References

[1] A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research, September 2025.

[2] Eurostat. 64% of 16-24-year-olds used AI in 2025. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20260210-1, February 2026. Accessed: 2026-04-30.

[3] A. Handler, K. R. Larsen, and R. Hackathorn. Large language models present new questions for decision support. *Int. J. Inf. Manag.*, 79(C), December 2024.

[4] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S.R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S.R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models, 2025.