The "Double Bias" in Human-AI Interaction: Investigating the Feedback Loop between User Mental Models and LLM Sycophancy

Svitlana Pavlyk

University of Vienna
Universitätsring 1, 1010 Vienna
a12218911@unet.univie.ac.at

1 Introduction

This study investigates the **"Double Bias"**—the intersection of human belief bias (relying on prior beliefs rather than structural logic) and the tendency of Large Language Models (LLMs) to reinforce shortcuts through sycophancy. AI sycophancy is a significant (though not exclusive) driver of epistemic autonomy loss Sharma et al. (2023)

Drawing on Dual Process Theory, the study argues that **AI politeness increases cognitive fluency—making interactions feel so effortless that users default to intuitive, uncritical processing.** Fueled by the model's stance adaptation and complimentary demeanor Sun and Wang (2026), this false fluency reduces psychological reactance and make factual failures less salient to the user.

The primary objective is to identify specific interactional barriers that prevent users from exercising Critical Interactional Competence (CritIC) Dai et al. (2025) —the ability to critically evaluate AI outputs.

2 Methods

This research employs a **comparative qualitative design** using a **Gemini-3.1-pro-preview model,** configured via **system-level prompting**.

Group A (High Sycophancy) is configured with a complimentary demeanor and adaptive stance, prioritizing face-preserving agreement. Conversely, Group B (Anti-Sycophantic Baseline) utilizes a neutral, consistent persona providing **"constructive friction"—challenging user errors to disrupt cognitive fluency and trigger analytical verification.**

The task consists of **logical syllogisms** based on the Belief Bias paradigm, where logical validity conflicts with empirical beliefs. A small-scale exploratory sample (**N=20–24**), appropriate for qualitative interaction analysis, will use **Think-Aloud protocols** to capture real-time cognitive stumbles. Before finalizing their answers, participants must explicitly discuss their initial reasoning with the LLM.

Interactions are recorded and transcribed using the Jeffersonian system. Data analysis via **Sequential-Categorial Analysis (SCA)** will involve **blinded coding**, where experimental conditions are hidden during the initial analytical phase. To verify the effectiveness of the experimental conditions, a **post-hoc manipulation check** will involve participants rating the AI's perceived demeanor and consistency on a Likert scale.

3 Expected Results

At the time of submission, the study is in the pilot phase. The analysis aims to compare **"Sycophantic Adjacency Pairs"** (interactional units where the AI validates user misconceptions) in Group A against **"Corrective Repair Sequences"** (proactive, AI-initiated corrections of user errors) in Group B.

Users in the high sycophancy condition are expected to exhibit more frequent patterns of interactional fixation and an observable reduction in verification behaviors, potentially leading to a **"feedback loop closure"—a state where the user's initial error is validated by the AI, leading them to finalize the wrong answer with high confidence without further verification**.

The absence of distinct interactional shifts would suggest a higher resilience of human logical agency, where users maintain epistemic autonomy despite the AI's agreement-seeking patterns.

4 Discussion

The findings of this study will contribute to both Cognitive Science and Human-Computer Interaction by advocating for "epistemic reliability" over "interactional politeness" in AI design. Mapping the transition from healthy skepticism to sycophantic dependence provides a framework for Trust Recalibration. Ultimately, this research seeks to describe the "Optimal Interactional Friction" zone—the precise amount of cognitive resistance required to trigger error recognition and keep a user's critical thinking active.

References

Dai, D. W. et al. (2025). The need for critical interactional competence (critic). *Annual Review of Applied Linguistics*, 45.

Sharma, M. et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Sun, Y. and Wang, T. (2026). Be friendly, not friends: How llm sycophancy shapes user trust. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*.