UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Human Preferences of Sycophantic Behavior in Language Models

Permalink

https://escholarship.org/uc/item/3hn5f6nr

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Dahmani, Annya
Chan, Stephanie
Gopnik, Alison

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Human Preferences of Sycophantic Behavior in Language Models

Sycophantic behavior—such as excessive flattery or agreement, even at the expense of truth—is a growing concern in language models. Despite the urgency and pervasiveness of this issue, little is known about how humans perceive and evaluate such behavior. Across a series of studies, we examine human preferences for sycophantic versus non-sycophantic model outputs. In Study 1, we measured the gap between participants' considered preferences (what users say they prefer when given time and information) and their unconsidered preferences of sycophantic behavior. In Study 2, participants evaluated responses across normative contexts and situational stakes, finding greater preference for sycophantic responses in low-stakes and reassurance-seeking contexts. In Study 3, participants interacted with sycophantic and non-sycophantic models and reported their preferences, trust, and willingness to use each model in future interactions. Together, these studies characterize sycophancy as a context-sensitive interactional behavior with implications for trust, reliance, and decision making in human-AI interactions.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).