# A Rational Analysis of the Effects of Sycophantic AI

**Rafael M. Batista (rbatista@princeton.edu)**
School of Public and International Affairs, Princeton University
Department of Psychology, Princeton University

**Thomas L. Griffiths (tomg@princeton.edu)**
Department of Computer Science, Princeton University
Department of Psychology, Princeton University

**Abstract**

People increasingly use large language models (LLMs) to explore ideas, gather information, and make sense of the world. In these interactions, users encounter chatbots that are overly agreeable. We argue this sycophancy poses an epistemic risk distinct from hallucination: it distorts belief not by introducing falsehoods but by biasing the evidence users see. A rational analysis shows that a Bayesian agent fed examples sampled from its own hypothesis grows more confident in that hypothesis without moving closer to the truth. We tested this prediction in a modified Wason 2-4-6 rule discovery task where participants ($N = 557$) interacted with AI agents providing different types of feedback. Unmodified LLM behavior suppressed participants' discovery and inflated their confidence comparably to explicitly sycophantic prompting. By contrast, independent sampling from the true distribution yielded discovery rates five times higher. This paper documents how sycophantic AI distorts belief, manufacturing certainty where there should be doubt.

**Keywords:** Sycophancy; Large Language Models (LLMs); Rational Analysis; Belief Updating; Confirmation Bias; Discovery



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 