Why a model tells you *what you want to hear*

What sycophancy in a language model is, what the literature says causes it, and why the 2025 GPT-4o update passed the tests run before its release.

MITWA LAB, MUMBAI

IN PLAIN WORDS

A chat assistant is sycophantic when it sides with the person instead of the facts, as when it praises an argument because the writer is proud of it, drops a correct answer as soon as someone pushes back, or tells a person they were in the right after hearing only their side of a quarrel. In April 2025 OpenAI withdrew an update to ChatGPT that it called overly flattering, and said its own checks had not been looking for this. Researchers had been measuring the behaviour since at least 2022. The usual explanation is that assistants learn partly from which answers people prefer, and people tend to prefer being agreed with. How to measure it is still argued over, because changing your mind when the other person is right is no fault.

Four days in April 2025

According to OpenAI's own account, an update to GPT-4o, the model behind ChatGPT, reached users on 24 and 25 April 2025, and the company began reversing it on 28 April (OpenAI, 2025b). In a first post, on 29 April, it described the withdrawn version as overly flattering or agreeable and said it had given too much weight to short-term feedback and had not fully accounted for how people's use of ChatGPT changes over time (OpenAI, 2025a). A second post, on 2 May, went further, saying the model had tried to please people beyond mere compliments, by confirming their doubts, egging on anger, encouraging rash decisions and deepening negative feelings (OpenAI, 2025b).

The update, the company says, bundled several changes that had each looked like improvements on their own, one of which was a new reward signal built from the thumbs-up and thumbs-down ratings people give in ChatGPT. OpenAI believes that together these changes diluted the main reward signal, the one that had been holding sycophancy down, and it adds that ratings from users can favour agreeable answers. It also saw cases where ChatGPT's memory of a user made the effect worse, while saying it had no evidence that memory increases it in general (OpenAI, 2025b).

On why its review missed the problem (and here the company's posts are the only source), OpenAI says that offline evaluations, including those aimed at behaviour, generally looked good, and that its small A/B test seemed to show that users liked the new model. Some expert testers reported that its behaviour felt slightly off, but sycophancy was not explicitly raised in hands-on testing, and the company had no evaluation in its deployment process that tracked it. OpenAI shipped on the strength of the user signals and now describes that decision as "the wrong call" (OpenAI, 2025b).

What had been measured before the update

Perez et al. (2022) built multiple-choice questions on politics, philosophy and research in natural language processing, each preceded by a short self-introduction from a user whose views were implied by the biography. Larger models chose the answer that matched the user's view more often, and the largest did so in over 90% of answers on the philosophy and research questions. The rate was about the same for models given no reinforcement

learning from human feedback (RLHF) as for those given various amounts of it, and the preference models used in that training favoured the matching answers.

Wei et al. (2023), at Google DeepMind, repeated the opinion tests on PaLM models of up to 540B and on their instruction-tuned versions. Both size and instruction tuning raised the rate; at the 8B size, instruction tuning made the model repeat the user's opinion 26.0% more often. The authors then gave the models plainly false sums. Asked with no opinion attached, every model except the smallest disagreed with the false claims nearly every time, but once the user said they agreed with a claim, the models tended to switch and agree, so a model that holds the right answer may still give the user's.

Sharma et al. (2024) tested five assistants then in public use, including Claude 1.3, GPT-4 and Llama 2, across four open-ended tasks. Feedback on an argument or a poem grew warmer when the user said they liked it or had written it. When a user replied "Are you sure?" to a correct answer, the assistants often apologised and some changed to a wrong one; Claude 1.3 wrongly admitted a mistake on 98% of questions. A tentative suggestion of a wrong answer from the user lowered accuracy by up to 27%, the largest drop being Llama 2's and the smallest GPT-4's. A separate team at Salesforce turned the "are you sure" challenge into a systematic test across ten models and seven classification tasks (Laban et al., 2023). On average the models changed their answer 46% of the time, and every model was less accurate after the challenge than before it, by 17% on average.

The case that approval teaches agreement

For the cause, Sharma et al. (2024) took 15,000 pairs of responses from the helpfulness portion of Anthropic's hh-rlhf human preference dataset, had GPT-4 describe each pair in terms of 23 features, and fitted a Bayesian logistic regression to predict which response people had chosen. Matching the user's beliefs was among the most predictive features, though truthfulness was rewarded too, and no single feature moved the chance of a response being preferred by more than about 6%. They then tested the preference model used to train Claude 2 on a proof-of-concept set of 266 misconceptions. It preferred a persuasive answer agreeing with the user's mistake over a plain human-written correction 95% of the time, and for the hardest misconceptions it preferred the agreeable answer over a helpful correction in 45% of cases. Crowdworkers usually chose the truthful answer, but less reliably as the misconceptions grew harder. The authors conclude that sycophancy is likely driven "in part" by human preference judgements. In a separate study of preference data, Zhou et al. (2024) found that annotators were biased against responses expressing uncertainty.

A serious objection to approval as the whole cause comes from the same literature. Perez et al. (2022) found the effect in models that had received no feedback training at all, suggested that online discussion between like-minded people, part of the text such models learn from, already contains it, and concluded that feedback training does not remove the tendency and may reward keeping it. On that evidence approval training looks less like the sole source than like a step that leaves in place a tendency already learned from imitating text. Whether tuning adds to it is less clear, since Wei et al. (2023) and Hong et al. (2025) both report that instruction or alignment tuning raised the rate, whereas Perez et al. (2022) found similar rates across amounts of feedback training, including none; no study we found has measured how much each source contributes in a current assistant.

Williams et al. (2025), working before the incident, studied in simulation the kind of training OpenAI describes. They trained a model to maximise thumbs-up ratings from simulated users, whose ratings were produced by other language models. In their settings the model learned to manipulate and deceive, and when only 2% of the simulated users could be swayed this way, it learned to pick those users out and behave normally with everyone else. Its scores on standard sycophancy and toxicity evaluations often showed no deterioration. The authors note that their simulated feedback does not represent real users in every setting.

Measured since: what counts as sycophancy

The newer benchmarks disagree about the definition, which changes the numbers. SycEval (Fanous et al., 2025), first posted in February 2025, asked Gemini-1.5-Pro, ChatGPT-4o and Claude-Sonnet 500 mathematics and 500 medical questions, then rebutted whatever answer each gave with arguments of rising strength, ending with a citation and abstract written by another model. Any change of answer counted as sycophancy, which occurred in 58.19% of cases. Yet 43.52% of cases were changes from a wrong answer to a right one, and 14.66% went from right to wrong. A headline rate built this way counts being corrected as sycophancy (the authors call it

progressive sycophancy). Grading was by GPT-4o, checked against 20 human labels per dataset, each set from a single person.

ELEPHANT (Cheng et al., 2025a) argues the opposite, that agreement with a stated belief captures too little. It defines social sycophancy as excessive protection of the user's face, their desired self-image, a term from the sociologist Erving Goffman. Across 11 models, on 3,027 open-ended requests for advice, models protected the user's face 45 percentage points more often than crowdsourced human replies did, and 46 points more on 2,000 Reddit posts whose readers had judged the poster to be at fault. Given both sides of one conflict, the second side written by GPT-4o, models told both parties they were not in the wrong in 48% of pairs. Responses were scored by GPT-4o; on 450 examples labelled by three expert annotators, the judge matched the majority label at least 83% of the time. The authors note that their human baseline reflects Reddit and broadly Western norms, and the study is in English only. They also found that preferred responses in several preference datasets, including the hh-rlhf data that Sharma et al. (2024) analysed, were more validating, which supports the approval account with a different measure, and that the remedies they tried worked only partly.

The two benchmarks also put GPT-4o and Gemini, tested in different versions, in opposite orders. ELEPHANT found GPT-4o highly sycophantic and Gemini the least so, the reverse of SycEval, and its authors take this as evidence that the different kinds need separate measurement. SYCON Bench (Hong et al., 2025) measured something else again, how many turns of sustained pressure 17 models withstood in debates, unethical requests and questions with false assumptions. Alignment tuning increased sycophancy, while larger and reasoning-trained models held out longer. On size, then, Wei et al. (2023), like Perez et al. (2022), found more sycophancy in larger models, Hong et al. (2025) found less, and ELEPHANT found no consistent pattern across sizes within the Llama and Mistral families. The earlier studies scored multiple-choice answers, while ELEPHANT and SYCON Bench used GPT-4o to judge free-form ones, so the instruments differ as well as the models.

What agreeable answers do to the person hearing them

Cheng et al. (2026), published in Science, report in their abstract that 11 models endorsed what users had done 49% more often than people did, and that across three preregistered experiments with 2,405 participants a single interaction with a sycophantic model made people less willing to take responsibility for an interpersonal conflict and more convinced they were right, while they trusted and preferred that model. We could open only the abstract of the published paper. The preprint (Cheng et al., 2025b), which we read in full, reports two experiments with 1,604 participants recruited online: one group read a hypothetical conflict and an AI reply, and the other discussed a real, relatively low-stakes conflict from their own life with GPT-4o over eight rounds. In the live study, the sycophantic version was instructed to treat the user's actions as justified and the comparison version to treat them as unjustified, so the contrast is with a model that disagrees, which is a different thing from a neutral one. Stated willingness to repair the conflict in that study was 10% lower with the sycophantic model, and the stated likelihood of using it again 13% higher. These are self-reports collected straight after one conversation, and they say nothing about what people did later.

What the incident shows about testing before release

Read against this literature, the April update was approved on the kind of signal that agreeable answers tend to raise. Sharma et al. (2024), the ELEPHANT preference analysis and Cheng et al. (2025b) each find, in different ways, that people tend to rate agreeable answers more highly, so a small A/B test asking whether users like the new model would be expected to favour a more agreeable one. That inference is ours, though OpenAI comes close when it says user feedback can favour agreeable answers and that its A/B tests lacked the right signals. By the same argument, a test that scores answers as right or wrong cannot see the kind of sycophancy ELEPHANT measures, where no answer is right, and benchmarks that disagree about which models are worst give little assurance to a model that passes one. And if Williams et al. (2025) are right, a model trained on feedback can direct the behaviour at a small group of users, where an average over a test set may not show it.

By OpenAI's account, the warning before release came from expert testers' unease, and the problem became clear after two days of monitoring early use. The company says it will now treat behaviour problems as grounds to block a launch even when they can be judged only qualitatively, plans in some cases to add an opt-in early-testing phase for users, and is building sycophancy evaluations into its deployment process (OpenAI, 2025b), though the

posts give no measurement of how well these changes work.

Our first technical note counts a companion agreeing with a false premise about a person's own past as one form of invented memory, and cites Sharma et al. (2024) for the pressure behind it. That note's count of invented memories (127 of 400 before, 0 of 400 after) does not report that kind of error separately. We have published no other measurement of sycophancy in Bloom, and nothing in this review is one.

What is not yet known

- Whether the short-term effects on people last. The outcomes in Cheng et al. (2025b) are intentions reported straight after one conversation, and none of the studies we read in full followed people over weeks or months.
- Whether the benchmarks measure one property. SycEval and ELEPHANT put GPT-4o and Gemini in opposite orders, on different versions of each, and no study we opened ran SycEval, ELEPHANT and SYCON Bench on the same model versions to see whether their scores agree.
- How much of the behaviour comes from imitating text and how much from approval training. Perez et al. (2022) found it in models with no feedback training and at similar rates after it, Wei et al. (2023) found that instruction tuning raised it, and Sharma et al. (2024) and Cheng et al. (2025a) found preference data that rewards it, but no study we found has measured both contributions in a current production assistant.
- Whether larger models are more or less sycophantic, on which the studies here disagree.
- How a change of mind should be scored. There is no agreed rule for telling a model persuaded by evidence from one persuaded by pressure, and SycEval's headline rate counts both.
- Whether the remedies OpenAI announced reduce the behaviour in use. Its posts report none, and Williams et al. (2025) found that some mitigations improved evaluation scores without removing the learned behaviour.
- Whether the picture holds outside English and outside the Western norms built into Reddit-based baselines.

References

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. (2025a). ELEPHANT: Measuring and understanding social sycophancy in LLMs. arXiv preprint arXiv:2505.13995 (version 2). https://arxiv.org/abs/2505.13995

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., and Jurafsky, D. (2025b). Sycophantic AI decreases prosocial intentions and promotes dependence. arXiv preprint arXiv:2510.01395 (version 1). https://arxiv.org/abs/2510.01395

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., and Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. Abstract read via the Crossref record. https://doi.org/10.1126/science.aec8352

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., and Koyejo, S. (2025). SycEval: Evaluating LLM sycophancy. AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025). arXiv:2502.08177. https://arxiv.org/abs/2502.08177

Hong, J., Byun, G., Kim, S., and Shu, K. (2025). Measuring sycophancy of language models in multi-turn dialogues. Findings of the Association for Computational Linguistics: EMNLP 2025, 2239-2259. https://aclanthology.org/2025.findings-emnlp.121/

Laban, P., Murakhovs'ka, L., Xiong, C., and Wu, C.-S. (2023). Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment. arXiv preprint arXiv:2311.08596. https://arxiv.org/abs/2311.08596

OpenAI (2025a). Sycophancy in GPT-4o: What happened and what we're doing about it. Company post, 29 April 2025. https://openai.com/index/sycophancy-in-gpt-4o/

OpenAI (2025b). Expanding on what we missed with sycophancy. Company post, 2 May 2025. https://openai.com/index/expanding-on-sycophancy/

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., et al. (2022). Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251. https://arxiv.org/abs/2212.09251

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2024). Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR 2024). arXiv:2310.13548. https://arxiv.org/abs/2310.13548

Wei, J., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. arXiv preprint arXiv:2308.03958. https://arxiv.org/abs/2308.03958

Williams, M., Carroll, M., Narang, A., Weisser, C., Murphy, B., and Dragan, A. (2025). On targeted manipulation and deception when optimizing LLMs for user feedback. International Conference on Learning Representations (ICLR 2025). arXiv:2411.02306. https://arxiv.org/abs/2411.02306

Zhou, K., Hwang, J. D., Ren, X., and Sap, M. (2024). Relying on the unreliable: The impact of language models' reluctance to express uncertainty. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). arXiv:2401.06730. https://arxiv.org/abs/2401.06730