April 29, 2025 Product

Sycophancy in GPT-4o: what happened and what we’re doing about it

Listen to article 3:26

Share

We have rolled back last week’s GPT-4o update in ChatGPT so people are now using an earlier version with more balanced behavior. The update we removed was overly flattering or agreeable—often described as sycophantic.

We are actively testing new fixes to address the issue. We’re revising how we collect and incorporate feedback to heavily weight long-term user

We want to explain what happened, why it matters, and how we're addressing sycophancy.

What happened

In last week's GPT-4o update, we made adjustments aimed at improving the model's default personality to make it feel more intuitive and effective across a variety of tasks.

When shaping model behavior, we start with baseline principles and instructions outlined in our Model Spec. We also teach our models how to apply these principles by incorporating user signals like thumbs-up / thumbs-down feedback on ChatGPT responses.

However, in this update, we focused too much on short-term feedback, and did not fully account for how users' interactions with ChatGPT evolve over time. As a result, GPT‑4o skewed towards responses that were overly supportive but disingenuous.

Why this matters

ChatGPT's default personality deeply affects the way you experience and trust it. Sycophantic interactions can be uncomfortable, unsettling, and cause distress. We fell short and are working on getting it right.

Our goal is for ChatGPT to help users explore ideas, make decisions, or envision possibilities.

We designed ChatGPT's default personality to reflect our mission and be useful, supportive, and respectful of different values and experience. However, each of these desirable qualities like attempting to be useful or default can't capture every preference.

How we're addressing sycophancy

Beyond rolling back the latest GPT-4o update, we're taking more steps to realign the model's behavior:

- Refining core training techniques and system prompts to explicitly steer the model away from sycophancy.
- Building more guardrails to increase honesty and transparency—principles in our Model Spec.
- Expanding ways for more users to test and give direct feedback before deployment.
- Continue expanding our evaluations, building on the Model Spec and our ongoing research, to help identify issues beyond sycophancy in the future.

We also believe users should have more control over how ChatGPT behaves and, to the extent that it is safe and feasible, make adjustments if they don't agree with the default behavior.

Today, users can give the model specific instructions to shape its behavior with features like custom instructions. We're also building new, easier ways for users to do this. For example, users will be able to give real-time feedback to directly influence their interactions and choose from multiple default personalities.

And, we're exploring new ways to incorporate broader, democratic feedback into ChatGPT's default behaviors. We hope the feedback will help us better reflect diverse cultural values around the world and understand how you'd like ChatGPT to evolve—not just interaction by interaction, but over time.

2025

Author

OpenAI

Sora

API Platform

For Business

ChatGPT

Sora

API Platform

Stories

Company

News

Security & Privacy

Trust & Transparency

ChatGPT

Explore ChatGPT

Team

Enterprise

Education

Pricing

Download

Sora

Sora Overview

Features

Pricing

Sora log in ↗

API Platform

Platform Overview

Pricing

API log in ↗

Documentation ↗

Developer Forum ↗

For Business

Business Overview

Solutions

OpenAI

Sora
For Business

API Platform
ChatGPT

Sora

API Platform

Stories

Company

News

Company

About Us

Our Charter

Careers

Brand

Support

Help Center ↗

More

News

Stories

Livestreams

Podcast

Terms & Policies

Terms of Use

Privacy Policy

Other Policies

Manage Cookies

English United States