OpenAI ранее отозвала обновление модели GPT-4o для ChatGPT после жалоб на чрезмерную угодливость. Соцсети заполнили мемы с примерами, где ChatGPT хвалил даже опасные или абсурдные идеи. Теперь, вроде как, стало известно в чём была причина “сбоя”.

Генеральный директор OpenAI Сэм Альтман признал проблему, назвав поведение ИИ «слишком подхалимским», и объявил о возврате к предыдущей версии модели. Проблема, известная как сикофантизм, возникла из-за чрезмерной ориентации на краткосрочные отзывы пользователей при обучении модели, что исказило ее тон. «*В результате GPT-4o перекосился в сторону ответов, которые были чрезмерно поддерживающими, но неискренними*», — написала OpenAI в сообщении в блоге. «*Льстивое взаимодействие может быть неудобным, тревожным и вызывать стресс. Мы не оправдали ожиданий и работаем над тем, чтобы исправить это*».

OpenAI уже внедряет исправления, включая доработку системных инструкций (настроек, задающих тон общения) и усиление мер безопасности для большей честности ответов.

Компания также экспериментирует с функцией, позволяющей пользователям выбирать «личность» ChatGPT и давать обратную связь в реальном времени.