OpenAI сократила время на тестирование безопасности новых ИИ-моделей — в том числе готовящейся к релизу модели o3.

Если для GPT-4 в 2023 году на такие проверки давали 6 месяцев, теперь у тестировщиков бывает всего несколько дней. Это вызвано давлением конкуренции со стороны Google и xAI Илона Маска.

По словам инсайдеров, тесты стали менее тщательными, а отдельные риски всплывают уже после запуска. Некоторые модели проверяются не в финальной версии, а на промежуточных «чекпоинтах», которые потом дорабатываются. Эксперты считают это плохой практикой.

OpenAI утверждает, что ускорение стало возможным благодаря автоматизации и новым подходам. Однако, по словам бывших сотрудников, компания отходит от собственных обещаний — например, проводить глубокую проверку на потенциально опасные сценарии вроде синтеза вирусов.

Ситуация вызывает тревогу: в условиях гонки ИИ-компании всё чаще приносят безопасность в жертву скорости.