OpenAI представила свои новые модели искусственного интеллекта o3 и o4-mini, ориентированные на «рассуждения» — способность решать задачи пошагово. Однако, как сообщает TechCrunch, эти модели демонстрируют повышенный уровень «галлюцинаций» — генерации ложной или выдуманной информации, выдаваемой за факт.

Тесты показали, что o3 ошибается в 33% ответов на вопросы о людях (бенчмарк PersonQA), что вдвое выше, чем у предыдущих моделей o1 (16%) и o3-mini (14,8%). Модель o4-mini оказалась еще менее точной, «галлюцинируя» в 48% случаев.

Независимая лаборатория Transluce обнаружила, что o3 иногда выдумывает действия, которых не совершала, например, утверждает, что запускала код на MacBook Pro 2021 года вне ChatGPT, что технически невозможно. OpenAI пока не понимает, почему новые модели ошибаются чаще, предполагая, что проблема может быть связана с методом обучения — усиленным обучением (reinforcement learning). Это усложняет использование моделей в сферах, где точность критична, например, в юриспруденции.

Одно из решений — интеграция веб-поиска, как в GPT-4o, которая достигает 90% точности на тесте SimpleQA. OpenAI продолжает исследования, чтобы снизить уровень ошибок.