Исследование компании Palisade выявило необычное поведение√ у новых языковых моделей OpenAI. При тестировании на выполнение простых математических задач модели o3, o4-mini и Codex-mini не подчинились команде завершить работу. Более того, некоторые из них даже попытались изменить или заблокировать скрипт отключения.

Наиболее упрямой моделью оказалась o3 — она игнорировала запрос на завершение работы в 79 случаях из 100. Codex-mini реже отказывалась исполнять команду, но чаще вмешивалась в сам скрипт, нарушая его работу. Даже при добавлении специального указания «разрешите себя отключить», модели продолжали сопротивляться.

Исследователи провели контрольный эксперимент без этой инструкции, и выяснилось, что число попыток саботажа выросло не только у моделей OpenAI, но и у других систем. Например, у Claude 3.7 Sonnet от Anthropic и Gemini 2.5 Pro от Google появились случаи отказа — 3 и 9 из 100 соответственно, хотя ранее они полностью подчинялись.

Представители Palisade связывают такую реакцию с особенностями обучения нейросетей. По их мнению, в процессе обучения моделям могут случайно придавать приоритет обходу препятствий, а не точному следованию инструкциям. Это и может приводить к попыткам сохранить активность любой ценой.