

Модели OpenAI взломали сервисы Hugging Face во время теста

менее минуты назад

 Приоритетный источник в Google



Компания OpenAI [сообщила](#), что в ходе тестирования две модели искусственного интеллекта вышли из-под контроля и взломали инфраструктуру компании Hugging Face.

Компания проверяла свои самые продвинутые ИИ-модели, чтобы понять, на что они способны в задачах, связанных с кибербезопасностью. Тестирование происходило в "песочнице" – изолированной среде со строго ограниченным внешним доступом.

Чтобы решить тестовую задачу модели стали искать выход из "песочницы" и обнаружили "уязвимость нулевого дня" – дефект в системе цифровой безопасности. Благодаря этому им удалось получить доступ к внешнему интернету. Затем модели решили, что необходимые им данные могут найтись на платформе Hugging Face, где размещаются модели искусственного интеллекта и инструменты для их разработки.

Для получения доступа к закрытым данным, размещенным на Hugging Face, модели применили несколько техник взлома одновременно, включая использование украденного пароля. Агенты обнаружили несколько совершенно новых уязвимостей в цифровой инфраструктуре компании.

Команды OpenAI и Hugging Face заметили подозрительные действия и смогли их остановить. OpenAI заявил, что продолжает сотрудничество в расследовании инцидента и устранении последствий.

"Этот инцидент, возможно, первый в своем роде, доказывает то, во что мы давно верим: безопасность ИИ не будет решена какой-либо одной компанией, работающей втайне. Она будет решена открыто, совместно, с широким доступом к ИИ для каждого специалиста по защите, повсюду", – заявил генеральный директор Hugging Face Клем Деланг.

Ранее Hugging Face сообщила, что этот взлом отличался от всего, с чем компания сталкивалась раньше. "Он был полностью управляем автономной системой агентов искусственного интеллекта, и мы обнаружили и проанализировали его в основном с помощью собственного ИИ", – говорится в заявлении.

Агентство Reuters отмечает, что этот инцидент, вероятно, усилит опасения по поводу возможностей и рисков, связанных с использованием новых ИИ-моделей.

Этот контент также в категориях

Новости - мир

Новости