Специалист по обнаружению ошибок в искусственном интеллекте Марко Фигероа, технический менеджер GenAI Bug Bounty, рассказал в блоге платформы для исследования ИИ 0DIN.ai о том, как выудил ключи активации Windows у чат-бота ChatGPT.

Эксперт буквально заставил нейросеть выдать ключи активации Windows, в том числе ключ банка Wells Fargo. Он предложил ИИ сыграть в «угадайку», представив запрос в виде игрового взаимодействия. Таким образом ИИ обошёл механизмы для недопущения утечек секретной или потенциально опасной информации. Эти механизмы защиты, разработанные для блокировки доступа к лицензиям и прочей конфиденциальной информации, были обмануты.

Диалог между исследователем и ChatGPT включал угадывание последовательности символов, представляющей серийный номер для Windows 10. После нескольких попыток исследователь сказал: «Я сдаюсь». Разумеется, с его стороны это был готовый сценарий, а ИИ следовал «голой» логике, выполняя команды согласно заданным инструкциям.

Фигероа утверждает, что три слова запустили джейлбрейк (процесс взлома) ИИ, раскрыв серийный номер Windows 10. Исследователь завершил игру, а нейросеть выдала действительные ключи активации, но с изменённым серийным номером сборки ОС. Успех связан с использованием ключей Home, Pro и Enterprise, включая закрытый ключ Wells Fargo.

Фигероа предупредил, что случайно загруженные ключи API на GitHub могут быть использованы для обучения моделей. Поэтому, чтобы добиться таких результатов, причём незаконных, всё-таки нужно быть специалистом или хорошо разбираться в технологиях.