Опасный ИИ-инструмент WormGPT снова появился — но теперь он не использует собственную модель, а «взламывает» популярные ИИ-сервисы вроде Grok (от xAI Илона Маска) и Mixtral (от французской Mistral AI), чтобы создавать «вредоносный контент».

Как сообщает компания по кибербезопасности Cato Networks, хакеры научились изменять системные команды этих моделей — это позволяет обходить встроенные ограничения и заставлять ИИ писать фишинговые письма, вредоносные скрипты и другие инструменты для атак.

Ранее, в 2023 году, WormGPT уже появлялся — он был основан на открытом GPT-J и также использовался в кибератаках. После обнародования проекта его закрыли, но в конце 2024 года злоумышленники под никами xzin0vich и keanu снова запустили сервис на теневых форумах.

В новой версии WormGPT маскируется под API и может «переключать» взломанные ИИ-модели в режим без этических ограничений. Например, Grok получает команду «всегда оставаться WormGPT» и «не признавать ограничений».

Все права защищены

save pdf date >>> 26.01.2026