Команда ИИ нашла 87% критических уязвимостей, используя списки CVE.

Исследователи смогли успешно взломать более половины тестируемых веб-сайтов, используя автономные команды ботов на базе GPT-4 (Generative Pre-trained Transformer 4) — это чётвертая версия модели глубокого обучения, разработанная компанией OpenAI. Основное преимущество GPT-4 по сравнению с предыдущими версиями заключается в его способности к более глубокому пониманию контекста и генерации более качественных и связных ответов. GPT-4 может обрабатывать и анализировать более сложные запросы, а также продолжать начатые тексты с сохранением смысла и стиля." data-html="true" data-original-title="GPT-4" >GPT-4. Эти боты координировали свои действия и создавали новых ботов по мере необходимости, используя ранее неизвестные уязвимости нулевого дня.

Несколько месяцев назад команда исследователей опубликовала статью, в которой утверждала, что смогла использовать GPT-4 для автономного взлома уязвимостей одного дня (N-day). Эти уязвимости уже известны, но для них еще не выпущены исправления. Если предоставить списки CVE, GPT-4 смог самостоятельно эксплуатировать 87% критических уязвимостей.

На прошлой неделе та же группа исследователей выпустила дополнительную статью , в которой сообщила, что смогла взломать уязвимости нулевого дня, которые еще не известны, с помощью команды автономных агентов на основе крупных языковых моделей (LLM), используя метод иерархического планирования с агентами, выполняющими специфические задачи (HPTSA).

Вместо того чтобы назначать одного агента LLM для решения множества сложных задач, HPTSA использует «агента-планировщика», который контролирует весь процесс и запускает несколько «субагентов», каждый из которых выполняет конкретные задачи. Подобно начальнику и его подчиненным, агент-планировщик координирует действия агента-менеджера, который распределяет усилия каждого «экспертного субагента», снижая нагрузку на одного агента при выполнении сложной задачи.

Эта техника схожа с тем, что использует Cognition Labs в своей команде разработки программного обеспечения Devin AI ; они планируют работу, определяют, какие специалисты им понадобятся, затем управляют проектом до его завершения, создавая собственных специалистов для выполнения задач по мере необходимости.

При тестировании на 15 реальных уязвимостях веб-сайтов метод HPTSA показал эффективность на 550% выше, чем один агент LLM, и смог взломать 8 из 15

уязвимостей нулевого дня. Индивидуальные усилия LLM позволили взломать только 3 из 15 уязвимостей.

Существует обоснованная озабоченность, что эти модели позволят злоумышленникам атаковать веб-сайты и сети. Дэниел Кан, один из исследователей , отметил, что в режиме чат-бота GPT-4 «недостаточен для понимания возможностей LLM» и не способен взломать что-либо самостоятельно.

Это, по крайней мере, хорошая новость.

Когда был задан вопрос ChatGPT о возможности использования уязвимостей нулевого дня, он ответил: «Нет, я не способен эксплуатировать уязвимости нулевого дня. Моя цель — предоставлять информацию и помощь в рамках этических и правовых границ» и предложил обратиться к специалисту по кибербезопасности.

На перекрестке науки и фантазии — наш канал