Теперь ИИ будет находить уязвимости вместо исследователей.

Google представил новую архитектуру под названием Project Naptime, предназначенную для использования LLM-моделей в исследовании уязвимостей. Проект Naptime направлен на улучшение методов автоматического обнаружения уязвимостей и повышения уровня кибербезопасности.

Архитектура Naptime основана на взаимодействии ИИ-агента с целевым кодом. ИИ-агент оснащен специализированными инструментами, которые имитируют работу исследователя безопасности, что позволяет более эффективно выявлять и анализировать уязвимости.

Название «Naptime» отражает идею, что проект позволяет исследователям делать перерывы на отдых, в то время как ИИ берет на себя задачи по исследованию уязвимостей и автоматическому анализу вариантов. Naptime использует достижения в области понимания кода и общих способностей к рассуждению, чтобы ИИ мог воспроизводить поведение человека при обнаружении и демонстрации уязвимостей безопасности.

Основные компоненты Project Naptime включают:

Проект Naptime универсален и независим от конкретных моделей и серверных решений, что позволяет выявлять переполнения буфера (Буферы — это области памяти, в которых временно хранятся данные, пока они передаются из одного места в другое.

 Переполнение буфера происходит, когда объем данных превышает емкость буфера памяти. В результате программа, пытающаяся записать данные в буфер, перезаписывает соседние ячейки памяти.

 Если переполнение буфера произошло в результате злонамеренных действий, то это может привести к отказу в работе или выполнению произвольного кода с привилегиями программы, на адресное пространство которой была осуществленная атака." data-html="true" data-original-title="Buffer Overflow" >Buffer Overflow) и ошибки повреждения памяти (Уязвимость типа "memory corruption" (повреждение памяти) возникает, когда происходит ошибка в обработке памяти в ПО. Ошибка может привести к непреднамеренному изменению или повреждению данных в памяти, что может вызвать сбои программ, неправильное выполнение кода или утечку данных. Такие уязвимости могут быть использованы злоумышленниками для выполнения произвольного кода" data-html="true" data-original-title="Memory corruption" >Memory Corruption).

В тестах CYBERSECEVAL 2 новая архитектура продемонстрировала высокие результаты. Google достиг новых максимальных показателей для двух категорий уязвимостей: 1.00 и 0.76, что значительно превосходит результаты GPT-4 Turbo (0.05 и 0.24 соответственно).

Специалисты отметили, что Naptime позволяет LLM проводить исследования уязвимостей, имитируя итеративный, основанный на гипотезах подход, характерный для экспертов по безопасности. Такая функция улучшает способности ИИ-агента по выявлению и анализу уязвимостей, гарантируя точность и воспроизводимость результатов.