Это достижение основывается на результатах более ранних исследований, в которых GPT-4 успешно эксплуатировал 87% известных уязвимостей критической серьезности. В эталонных тестах на 15 реальных веб-уязвимостей HPTSA оказалась на 550% эффективнее, чем одиночные попытки, взломав 8 из 15 уязвимостей нулевого дня, в то время как один LLM справился только с тремя.

Несмотря на эти возможности, GPT-4 в режиме чат-бота (ChatGPT) по-прежнему не может автономно эксплуатировать уязвимости, обеспечивая соблюдение этических границ.