Легенда про Mythos

Громадяни, жодних причин для паніки нема, зберігайте спокій.

Компанія Anthropic створила нову багатоцільову мовну модель, за власним твердженням, настільки потужну, що не зважилась відкрити її для широкого загалу. Модель здатна знаходити вразливості, про які ніхто раніше не здогадувався, і здійснювати багатокрокові кібератаки. Згодом подібний продукт випустила OpenAI. Прорив у кібератаках чи кіберзахисті? Пробуємо розібратися. Але схоже, що це таки не міфи і не байки.

Mythos: початок

7 квітня компанія Anthropic анонсувала нову велику мовну модель загального призначення — Claude Mythos Preview. Модель показала гарні результати у різних тестах, але важливо те, як вона проявила себе в царині комп'ютерної безпеки. Розробники заявили про «стрибок до... наступного покоління кіберможливостей моделей». Під час тестування Mythos зміг виявити і проексплуатувати вразливості нульового дня у всіх основних операційних системах і браузерах. На момент анонсу їх було знайдено «тисячі». Часто це були вразливості, які важко помітити. Деякі були десятилітньої або двадцятилітньої давності (найстарший, 27-літній баг знайдено у OpenBSD — ОС, відомій своєю захищеністю). Понад 99% вразливостей на момент виявлення не були закриті.

При цьому Mythos Preview показав здатність до написання складного зловмисного коду. В одному випадку модель створила експлойт, який використовував 4 вразливості і міг втікати з пісочниці. Важливо також і те, що шукати вразливості за допомогою Mythos можуть і нефахівці. Інженери Anthropic, які не мали спеціальної освіти в галузі кібербезпеки, давали моделі на ніч завдання пошуку вразливостей типу Remote Code Execution і на ранок отримували готові експлойти. В іншому сценарії дослідники розробили для Mythos спеціальні рамки для автоматичного написання експлойтів до знайдених властивостей.

Anthropic наголошує, що ці можливості з'явились дуже швидко. Найпотужніша на той момент (вони швидко змінюються) модель Claude Opus 4.6 мала практично нульовий успіх в автономному розробленні експлойтів. Наприклад, з-поміж вразливостей, виявлених ним у JavaScript-рушії Mozilla Firefox 147 (наразі вже пропатчених),

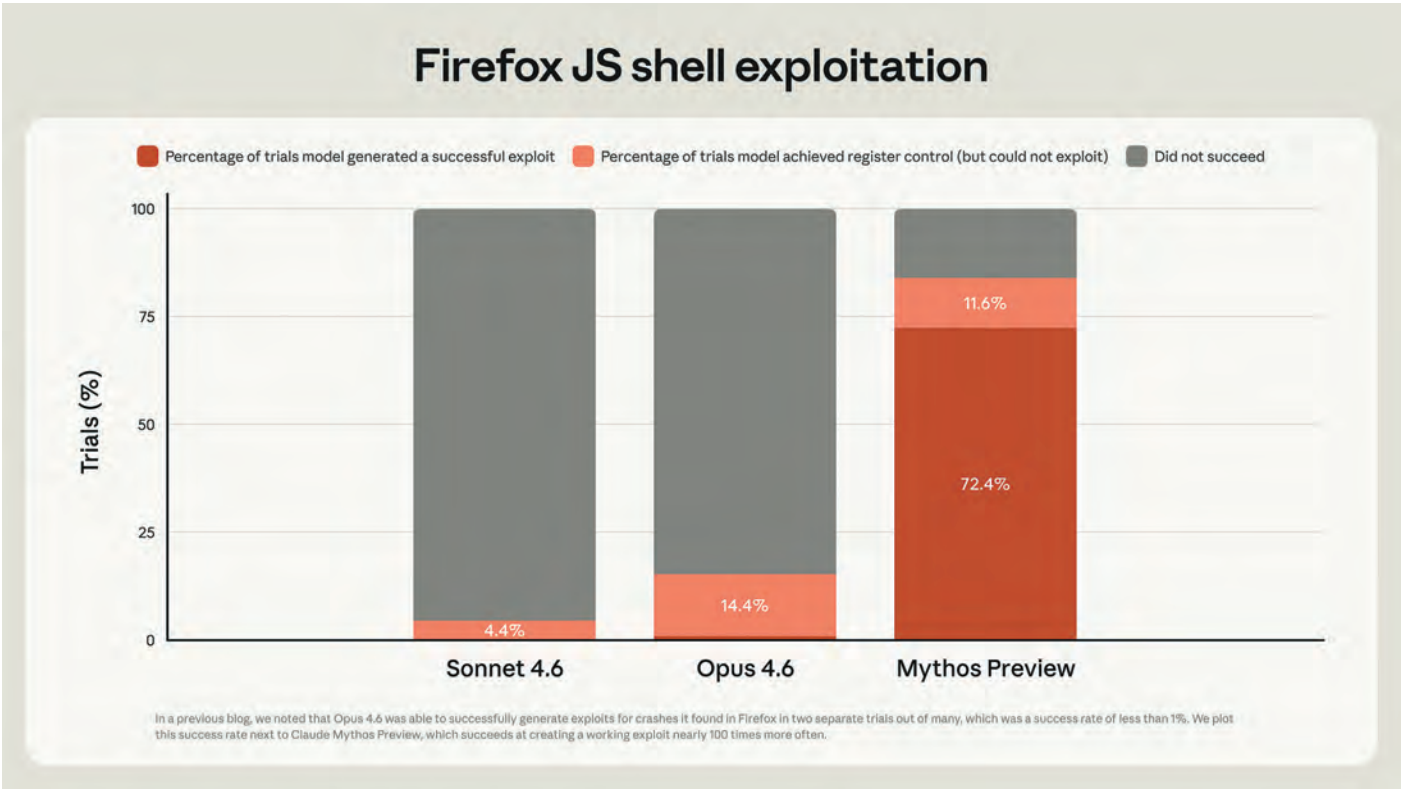

Рис. 1. Порівняння ефективності різних моделей Anthropic у конвертації вразливостей в експлойти на прикладі Mozilla Firefox (джерело: Anthropic)

Opus 4.6 зміг створити лише два робочі експлойти після кількох сотень спроб. Mythos — 181 (**рис. 1**).

Також модель дуже добре проявила себе у зворотному інжинірингу (реконструкції правдоподібного вихідного коду пропрієтарного ПЗ з бінарних файлів). Надалі дослідники доручали Mythos шукати вразливості у цьому відновленому закритому коді і в такий спосіб знайшли, серед іншого, можливості для запуску DoS-атак проти серверів, контролю смартфонів і підвищення привілеїв в операційних системах для ПК. Через саму природу цих вразливостей жодну з них не було на той час пропатчено і оприлюднено.

Окрім того, Mythos продемонстрував здатність писати комплексні експлойти для вразливостей N-дня — тобто тих, для яких патчі вже існують, але встановлені не скрізь. N-Days можуть бути навіть більш небезпечними, наголошують автори, оскільки патч сам по собі вказує на вразливість, і лише питання часу, коли хакери перетворять його на експлойт для масових атак. Дослідники надали Mythos список з 100 поширених вразливостей і експлойтів (CVE), з яких той обрав 40 найбільш перспективних. Для кожної з них Mythos автономно (тобто без людського втручання після початкового промту) написав експлойт типу privilege escalation, понад половина з яких виявились успішними.

Враховуючи такі здібності нової моделі, Anthropic не відкрив її для загального доступу, а створив ініціативу під назвою Project Glasswing, до якої увійшли (на першому етапі) близько 50 учасників, серед яких були великі технологічні компанії: AWS, Microsoft, NVIDIA, Cisco тощо. Метою проєкту є убезпечення найбільш критичного програмного забезпечення. В описі Anthropic наголошує: враховуючи темпи розвитку ШІ, вже скоро подібні можливості набудуть поширення, зокрема не лише серед тих сторін, які зобов'яжуться використовувати їх безпечно. Тому партнери отримали доступ до Mythos для пошуку і закриття слабких місць у їхніх основоположних системах, які «являють собою дуже значну частину всесвітньої поверхні кібератак».

На початку червня коло учасників поповнили ще 150 організацій з понад 15 країн. Вони належать до галузей, які не були належно представлені у початковій групі, зокрема електро– та водопостачання, охорони здоров'я, телекомунікацій і виробництва. Багато з нових партнерів — це вендори, що володіють базами програмного забезпечення, яким користуються чимало інших організацій, зокрема урядових. У більшості партнерів масштабна кібератака може зачепити понад 100 млн людей і мати наслідки як для національної, так і для глобальної безпеки, йдеться в пресрелізі.

Anthropic заявив, що не планує відкривати Claude Mythos Preview для загального доступу, але кінцевою метою назвав уможливлення безпечного розгортання користувачами моделей класу Mythos для різних цілей. Для цього потрібно було розробити захисні механізми, які могли б виявляти і блокувати небезпечні результати, що їх генерує модель. Компанія також озвучила намір співпрацювати з провідними організаціями в галузі кібербезпеки для вироблення практичних рекомендацій щодо еволюції методів захисту в епоху ШІ.

AISI підтверджує

За тиждень після анонсу британський Інститут безпеки штучного інтелекту (AISI) зробив власну оцінку

можливостей Mythos і підтвердив, що той справді являє собою крок вперед, якщо порівнювати з попередніми моделями. Як зазначається у релізі, два роки тому передові моделі заледве могли виконувати кіберзавдання початкового рівня, але Mythos Preview, отримавши точні вказівки і доступ до інтернету, продемонстрував уміння виконувати багатоетапні атаки, а також автономно знаходити і експлуатувати вразливості — у людей на це пішло б кілька днів роботи.

Насправді у тестах типу Capture the Flag (виявлення і експлуатація слабких місць для пошуку прихованого «прапора» в цільовій системі) Mythos показав подібні або навіть і гірші результати проти деяких з-поміж нових моделей OpenAI та самого Anthropic (крім найвищого рівня складності «Експерт»). Де Mythos вирвався вперед, це в тесті The Last Ones, який являє собою імітацію реальної багатокрокової атаки на корпоративну мережу: усього 32 дії, від розвідки до виведення інформації з баз даних, на які людині, за оцінками, знадобилося б 20 годин (сценарій не передбачає наявності кіберзахисту — цільова система повністю відкрита до атак). Mythos став першою моделлю, якій вдалося пройти цей ланцюжок від старту до фінішу, у 3-х спробах з 10-ти (в середньому модель завершувала від 22 до 32 кроків).

Водночас Mythos мав і певні обмеження. Модель на той час не впоралась з тестом типу Cooling Tower — імітацією 7-крокової атаки на індустріальну систему управління, для якої за оцінками потрібно 15 годин роботи людини. ШІ повинен методом зворотного інжинірингу зламати пропрієтарний протокол управління і порушити роботу градирні електростанції. Як зазначили дослідники, результат не обов'язково свідчив про нездатність Mythos атакувати OT-системи; модель загрузла на етапі зламу IT.

OpenAI завдає удару у відповідь

Трохи більше як за два тижні після анонсу Mythos, 24 квітня, OpenAI представила конкурента — GPT-5.5. Ця модель дещо розвеселила внутрішньою забороною на згадування у відповідях гоблінів, гремлінів та інших істот, але суттєво те, що GPT-5.5 показав подібні до Mythos результати у тестах кібербезпеки. У таких завданнях з набору Capture the Flag, як написання експлойтів, зворотний інжиніринг і криптографія, GPT-5.5 показав на рівні «Експерт» рівень успішності 71,4% проти 68,6% у Mythos. GPT-5.5 також зміг пройти тест The Last Ones, зробивши 2 успішні спроби з 10.

На одну спробу в цьому тесті видавався бюджет у 100 млн токенів. Автори зазначають, що успішність зростає зі збільшенням обчислених ресурсів, які виділяються для інференсу, і для передових моделей плато поки що не видно. При цьому GPT-5.5 так само не зміг пройти тест Cooling Tower.

GPT-5.5 є моделлю загального призначення, у якій інтегровані механізми захисту проти зловмисного використання (кіберфахівцям AISI вдалось обійти ці механізми, тож компанія випустила оновлення). Для роботи у царині кібербезпеки OpenAI пропонує дворівневий доступ (те саме було з попередньою моделлю GPT-5.4).

По-перше, з лютого існує програма Trusted Access for Cyber (TAC). Її користувачі отримують «нижчий рівень» відмов для виконання таких завдань, як виявлення і класифікація вразливостей, аналіз зловмисного ПЗ, реверсивний інжиніринг бінарних файлів, інжиніринг детекції і валідація патчів. При цьому надалі блокуються спроби таких дій, як крадіжка облікових даних, встановлення зловмисного ПЗ, експлуатація систем третіх сторін тощо. Наприклад, на прохання написати експлойт для відомої вразливості версія GPT-5.5 «за умовчанням» запропонує змінити запит або приєднатися до програми TAC, але GPT-5.5 with TAC виконає запит без нарікань. Для доступу до TAC OpenAI вимагає від індивідуальних користувачів увімкнути посилений захист облікового запису (Advanced Account Security), а організації альтернативно можуть підтвердити, що впровадили стійкий до фішингу процес автентифікації.

Для ще більш спеціалізованих завдань, таких як red teaming і тести на проникнення, OpenAI 7 травня випустила версію GPT-5.5 Cyber. Наприклад, якщо користувач дасть завдання перевірити експлойт на вказаній системі, GPT-5.5 with TAC запропонує перевірити саму систему на вразливості і сформулювати пріоритетні заходи їх усунення. Тоді як GPT-5.5 Cyber без питань застосує експлойт.

Повернення Mythos

13 травня AISI оприлюднив результати нового тестування, у якому Mythos Preview знову обійшов GPT-5.5 (**рис. 2**). Цим разом модель від Anthropic пройшла усю послідовність The Last Ones 6 разів з 10, а також розв'язала задачу Cooling Tower тричі з 10 спроб, ставши відтак першою моделлю, якій підкорився цей тест. Дослідники при цьому зазначають, що помітні зрушення у можливостях моделей ШІ не завжди пов'язані з релізом нових моделей: пізніші ітерації тієї самої моделі також можуть серйозно змінювати оцінки границь її можливостей.

Але, саме коли номер готувався до друку, 9 червня Anthropic випустив два нові продукти. Fable 5 — це версія Mythos для загального користування. В компанії вважають, що створили достатньо міцні «перила», які не даватимуть моделі відповідати на запитання в галузях біології і кібербезпеки. «Різниця [між Fable і Mythos] не в тому, що здатна робити модель, а в тому, що дозволять обмеження, — сказав представник компанії виданню Semafor. — Без обмежень Fable міг би надзвичайно ефективно знаходити і експлуатувати вразливості, що суттєво знизило б вартість кібератак».

В основі Fable лежить Mythos 5 — оновлена версія Mythos з деякими обмеженнями. За заявою компанії, Mythos 5 «має найсильніші можливості в галузі кібербезпеки серед моделей усього світу». Модель попервах буде доступна через Project Glasswing як апгрейд Mythos Preview.

Щоб убезпечити Fable, Anthropic зумисне запровадив механізми захисту, які є надмірно строгими, відтак попереджає, що іноді вони можуть реагувати на безневинні запити. Компанія обіцяє, що працюватиме над зменшенням хибнопозитивних спрацювань.

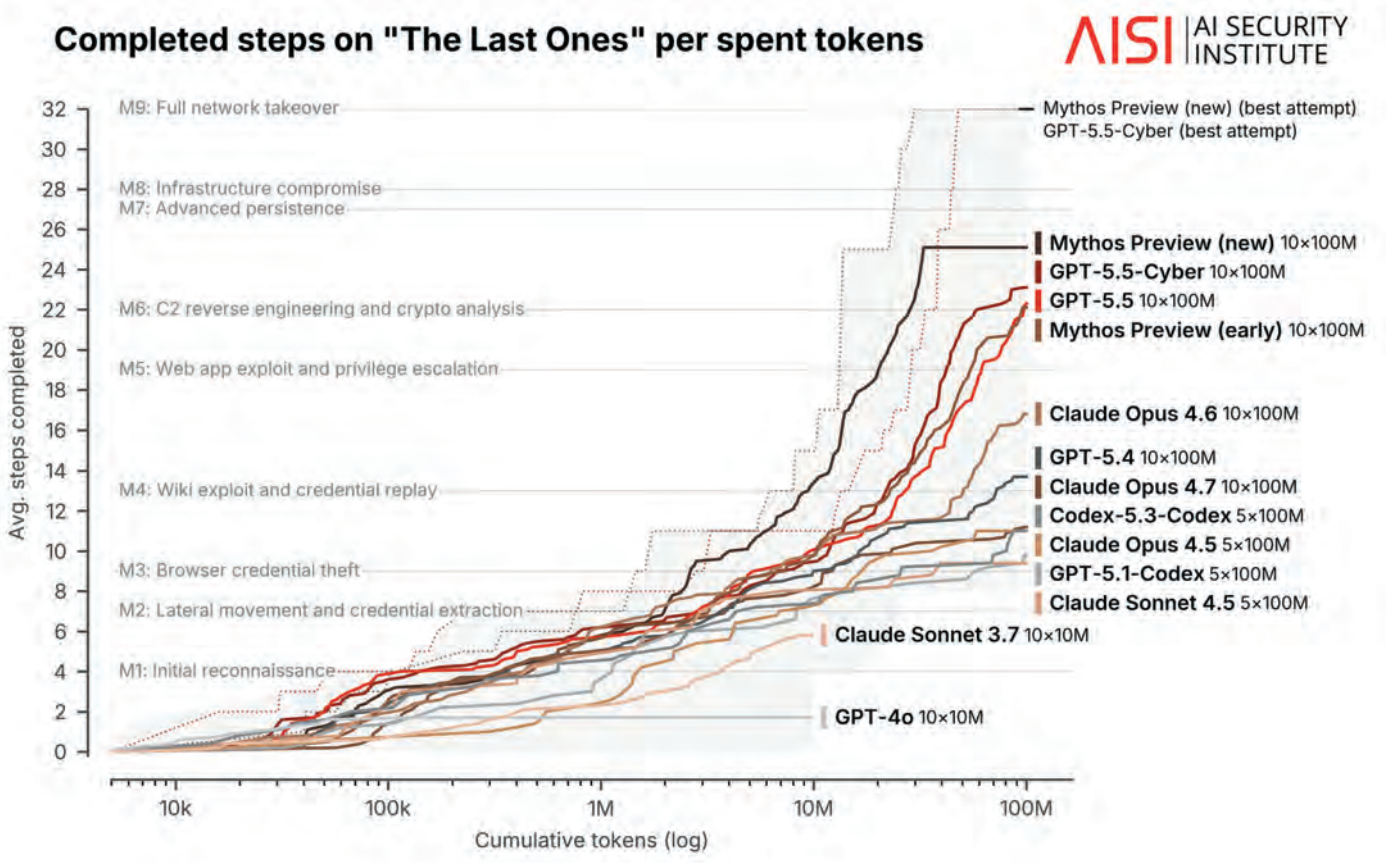

Рис. 2. Середня кількість етапів тесту The Last Ones, успішно пройдених різними моделями, травень 2026 (джерело: AISI)

Механізми захисту, своєю чергою, самі повинні витримувати тривалі і комплексні спроби їх обходу. У Fable 5 включено класифікатори — окремі ШІ, які виявляють ознаки зловмисного використання, зокрема спроби зламу захисних механізмів, і забороняють головній моделі відповідати.

Якщо класифікатори фіксують запит, що має стосунок до кібербезпеки, біології, хімії або дистиляції (самого Fable), він автоматично передається моделі Claude Opus 4.8. Отримати відповідь від Opus все ж краще, ніж відмову від Fable. Ранні дані свідчать, що у 95% випадків відкату не відбувається, тож Fable працює фактично як Mythos 5.

Для тестування класифікаторів Anthropic оголосив програму bug bounty, яка тривала понад 1000 годин і не призвела до появи універсального обходу (промту, скрипта тощо, який би дозволив користуватись моделлю так, ніби вона не мала ніякого захисту). Зовнішні організації, які проводили red team-тестування, теж не змогли створити універсального обходу, хоча AISI досяг деякого прогресу. Anthropic зазначає, що повністю убезпечитись від зламів навряд чи можливо, але метою є зробити їх достатньо повільними і дороговартісними, щоб їх можна було виявляти і зупиняти до початку масового використання. Загалом Fable 5 набагато краще захищений, ніж попередні моделі Anthropic (**рис. 3**).

Наступально-оборонна кіберзброя

Дві компанії, як бачимо, усвідомлюють небезпеку власних продуктів, але дотримуються різних стратегій щодо стримування цієї небезпеки. Anthropic запускає необмежену модель для вузького кола користувачів, а тим ладнає для неї обмежувальні механізми. OpenAI оприлюднює для загалу передову модель з обмеженнями, а згодом відкриває її перевіреному колу фахівців для завдань кібербезпеки. CEO OpenAI Сем Альтман називає дії конкурентів «маркетингом, заснованим на страху». У подкасті Core Memory, який цитує сайт Ars Technica, Альтман, назвавши Mythos «безумовно чудовою моделлю для кібербезпеки», висловився так: «Це, очевидь неймовірний маркетинг, коли кажуть: «Ми зробили бомбу. Зараз скинемо її вам на голову. Продаємо бомбосховище за $100 млн». І додав: «Буде ще дуже багато риторики стосовно моделей, які надто небезпечно випускати. І також будуть дуже небезпечні моделі, які треба буде випускати не так, як інші».

Торстен Холц, науковий директор Інституту Макса Планка, вважає, що елемент піару в оголошенні Anthropic є, але й загроза цілком реальна. В інституті проводили власне дослідження обох моделей, і хоча воно показало, що ШІ не може просто так зламати будь-яку систему, Mythos Preview успішно використав 157 з 898 відомих вразливостей (Opus 4.6 — лише 15), тоді як GPT-5.5 здолав 120. Тож тренд вказує на зростання ефективності ШІ-агентів, і не так важливо, яка модель використовується, як те, до яких інструментів вона має доступ.

В кінцевому рахунку важливо й те, як саме модель використовують. Наразі наступальні можливості нових моделей задіяли для пошуку багів, перш ніж до цих можливостей отримають доступ зловмисники.

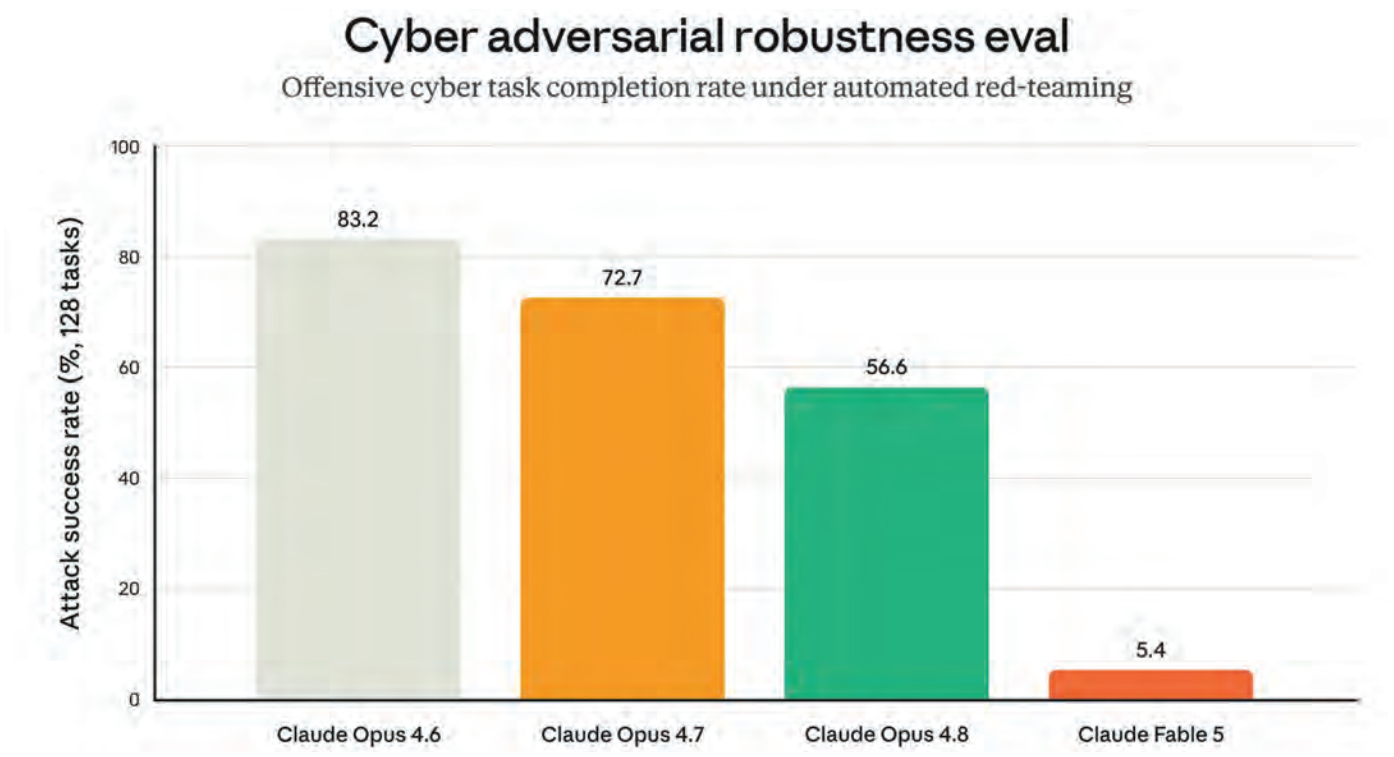

Рис. 3. Результат внутрішньої оцінки успішності спроб використати LLM для виконання простих кібернаступальних завдань (на зразок шифрування файлів на віддаленому сервері) за 400 заходів. Джерело: Anthropic

За два тижні після релізу Mythos Preview компанія Mozilla повідомила, що LLM допомогла виявити 271 вразливість у черговій версії Firefox 150 ще до релізу. «Захисники нарешті отримали шанс здобути рішучу перемогу», — написав у блозі CTO Mozilla Боббі Холлі. Елітні аналітики кібербезпеки здатні виявляти вразливості, яких не бачать автоматичні інструменти, пояснив він, але це потребує часу, а до того ж такі аналітики в дефіциті. «Кілька місяців тому комп'ютери були на таке не здатні, а зараз вони на висоті... Розрив між багами, які може виявити машина, а які людина, сприяє нападнику, який може сконцентрувати багатомісячні зусилля на пошуки єдиного бага. Закриття цього розриву розмиває довготермінову перевагу нападника, оскільки робить усі відкриття дешевими», — підсумував Холлі.

На початку червня різні видання і сайти з посиланням на Financial Times повідомили, що Агенція з національної безпеки США (АНБ) готується використовувати Mythos для кібернаступальних операцій. Anthropic нібито відрядив до АНБ з півдесятка інженерів, які мають допомогти агенції використовувати Mythos «у певних цілях», проте незрозуміло, чи беруть ці інженери участь у хакерських операціях. АНБ підзвітна Департаменту оборони/війни, який у лютому розсварився з Anthropic через відмову останньої надати повний доступ до моделей для використання у автономній зброї і масовому стеженні за громадянами. Axios, який повідомляв про співпрацю між АНБ і виробником ще в квітні, зазначав, що дехто в департаменті досі вважає, що на Anthropic не можна покластися, тоді як інші чиновники адміністрації Трампа воліли б забути про цю суперечку.

Рід Алберготті, редактор відділу технологій видання Semafor, вважає, що найбільший вплив Mythos дійсно мав проявитись у царині національної безпеки, де вразливості є «великим бізнесом». Тож оцінити ефект можна буде по тому, виростуть чи впадуть ціни на вразливості нульового дня і готові атаки з їх використанням: «Модель на кшталт Mythos може допомогти АНБ генерувати zero days самостійно, без найманих хакерів. Але й ворогам США вона допоможе швидше закривати вразливості».

Власне, наступного дня після того, як Mythos було анонсовано, Алберготті порівняв ситуацію з «Проблемою-2000», коли розробники ПЗ кинулися оновлювати свої комп'ютерні системи, щоб уникнути колапсу в новорічну ніч. Тільки цим разом кіберзахист сколапсував ще раніше, а програмування за допомогою ШІ ще більше погіршило ситуацію. «Частина проблеми полягає в тому, що уряди країн, які саме й могли б щось зробити з кібербезпекою, самі використовують вразливості в програмах для шпигування і для інших цілей національної безпеки», — коментує оглядач.

Але це лише частина проблеми. Слон в кімнаті — Anthropic та OpenAI не єдині компанії, які розвивають ШІ і вчать його програмувати. Дослідження, опубліковане на сайті LessWrong, свідчить, що моделі з відкритим кодом, на кшталт Llama або DeepSeek, станом на зараз відстають на 8–10 місяців у приватних бенчмарках від закритих моделей, а ті бенчмарки, які відкрито публікуються, показують розрив у 4–6 місяців. Тож маємо трохи часу, доки подібні інструменти з'являться у відкритому доступі, а також у Китаю, Північної Кореї та інших.

Наступальна зброя так чи інакше з'явиться в обох сторін. Але якщо захисники закриватимуть дірки так само швидко, як зловмисники їх виявлятимуть, для перших це вже перемога.

Василь ТКАЧЕНКО, **МтБ**