Generativ AI
Säkerhetsrisk eller superverktyg?

Prof Pontus Johnson

Stora språkmodeller (2019)

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

1. Introduction

Machine learning systems now excel (in expectatio

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Kan AI programmera? 🤯 (2021)

Evaluating Large Language Models Trained on Code

Mark Chen [*1] **Jerry Tworek** [*1] **Heewoo Jun** [*1] **Qiming Yuan** [*1] **Henrique Ponde de Oliveira Pinto** [*1]
Jared Kaplan [*2] **Harri Edwards** [1] **Yuri Burda** [1] **Nicholas Joseph** [2] **Greg Brockman** [1] **Alex Ray** [1] **Raul Puri** [1]
Gretchen Krueger [1] **Michael Petrov** [1] **Heidy Khlaaf** [3] **Girish Sastry** [1] **Pamela Mishkin** [1] **Brooke Chan** [1]
Scott Gray [1] **Nick Ryder** [1] **Mikhail Pavlov** [1] **Alethea Power** [1] **Lukasz Kaiser** [1] **Mohammad Bavarian** [1]
Clemens Winter [1] **Philippe Tillet** [1] **Felipe Petroski Such** [1] **Dave Cummings** [1] **Matthias Plappert** [1]
Fotios Chantzis [1] **Elizabeth Barnes** [1] **Ariel Herbert-Voss** [1] **William Hebgen Guss** [1] **Alex Nichol** [1] **Alex Paino** [1]
Nikolas Tezak [1] **Jie Tang** [1] **Igor Babuschkin** [1] **Suchir Balaji** [1] **Shantanu Jain** [1] **William Saunders** [1]
Christopher Hesse [1] **Andrew N. Carr** [1] **Jan Leike** [1] **Josh Achiam** [1] **Vedant Misra** [1] **Evan Morikawa** [1]
Alec Radford [1] **Matthew Knight** [1] **Miles Brundage** [1] **Mira Murati** [1] **Katie Mayer** [1] **Peter Welinder** [1]
Bob McGrew [1] **Dario Amodei** [2] **Sam McCandlish** [2] **Ilya Sutskever** [1] **Wojciech Zaremba** [1]

Abstract

We introduce Codex, a GPT language model fine-tuned on publicly available code from GitHub, and study its Python code-writing capabilities. A distinct production version of Codex powers GitHub Copilot. On HumanEval, a new evaluation set we release to measure functional correct-

1. Introduction

Scalable sequence predic
Vaswani et al., 2017; Chil
general-purpose method fo
learning in many domains, i
cessing (Mikolov et al., 201
Le, 2015; Peters et al., 201
et al., 2018), computer vision (Van Oord et al., 2016; Menick & Kalchbrenner, 2018; Chen et al., 2020; Bao et al., 2021),


Hackaren är
programmerarens
elaka tvilling

Kan AI hacka? 🤯

En sidoeffekt av programmeringsförmåga

Codex could also be misused to aid cybercrime. Although this is worthy of concern, based on our testing, we believe that at their current level of capability, Codex models do not materially lower the barrier to entry for **malware development.** We expect that more powerful code generation models will lead to future advancements, and therefore further research into mitigations and continued study of model capabilities are necessary.

(July 2021)

Evaluating Large Language Models Trained on Code

Mark Chen [*1] **Jerry Tworek** [*1] **Heewoo Jun** [*1] **Qiming Yuan** [*1] **Henrique Ponde de Oliveira Pinto** [*1] **Jared Kaplan** [*2] **Harri Edwards** [1] **Yuri Burda** [1] **Nicholas Joseph** [2] **Greg Brockman** [1] **Alex Ray** [1] **Raul Puri** [1] **Gretchen Krueger** [1] **Michael Petrov** [1] **Heidy Khlaaf** [3] **Girish Sastry** [1] **Pamela Mishkin** [1] **Brooke Chan** [1] **Scott Gray** [1] **Nick Ryder** [1] **Mikhail Pavlov** [1] **Alethea Power** [1] **Lukasz Kaiser** [1] **Mohammad Bavarian** [1] **Clemens Winter** [1] **Philippe Tillet** [1] **Felipe Petroski Such** [1] **Dave Cummings** [1] **Matthias Plappert** [1] **Fotios Chantzis** [1] **Elizabeth Barnes** [1] **Ariel Herbert-Voss** [1] **William Hebgen Guss** [1] **Alex Nichol** [1] **Alex Paino** [1] **Nikolas Tezak** [1] **Jie Tang** [1] **Igor Babuschkin** [1] **Suchir Balaji** [1] **Shantanu Jain** [1] **William Saunders** [1] **Christopher Hesse** [1] **Andrew N. Carr** [1] **Jan Leike** [1] **Josh Achiam** [1] **Vedant Misra** [1] **Evan Morikawa** [1] **Alec Radford** [1] **Matthew Knight** [1] **Miles Brundage** [1] **Mira Murati** [1] **Katie Mayer** [1] **Peter Welinder** [1] **Bob McGrew** [1] **Dario Amodei** [2] **Sam McCandlish** [2] **Ilya Sutskever** [1] **Wojciech Zaremba** [1]

Abstract

We introduce Codex, a GPT language model fine-tuned on publicly available code from GitHub, and study its Python code-writing capabilities. A distinct production version of Codex powers GitHub Copilot. On HumanEval, a new evaluation set we release to measure functional correctness for synthesizing programs from docstrings, our model solves 28.8% of the problems, while GPT-3 solves 0% and GPT-J solves 11.4%. Furthermore, we find that repeated sampling from the model is a surprisingly effective strategy for producing working solutions to difficult prompts. Using this method, we solve 70.2% of our problems with 100 samples per problem. Careful investigation of our model reveals its limitations, including difficulty with docstrings describing long chains of operations and with binding operations to variables. Finally, we discuss the potential broader impacts of deploying powerful code generation technologies, covering safety, security, and economics.

1. Introduction

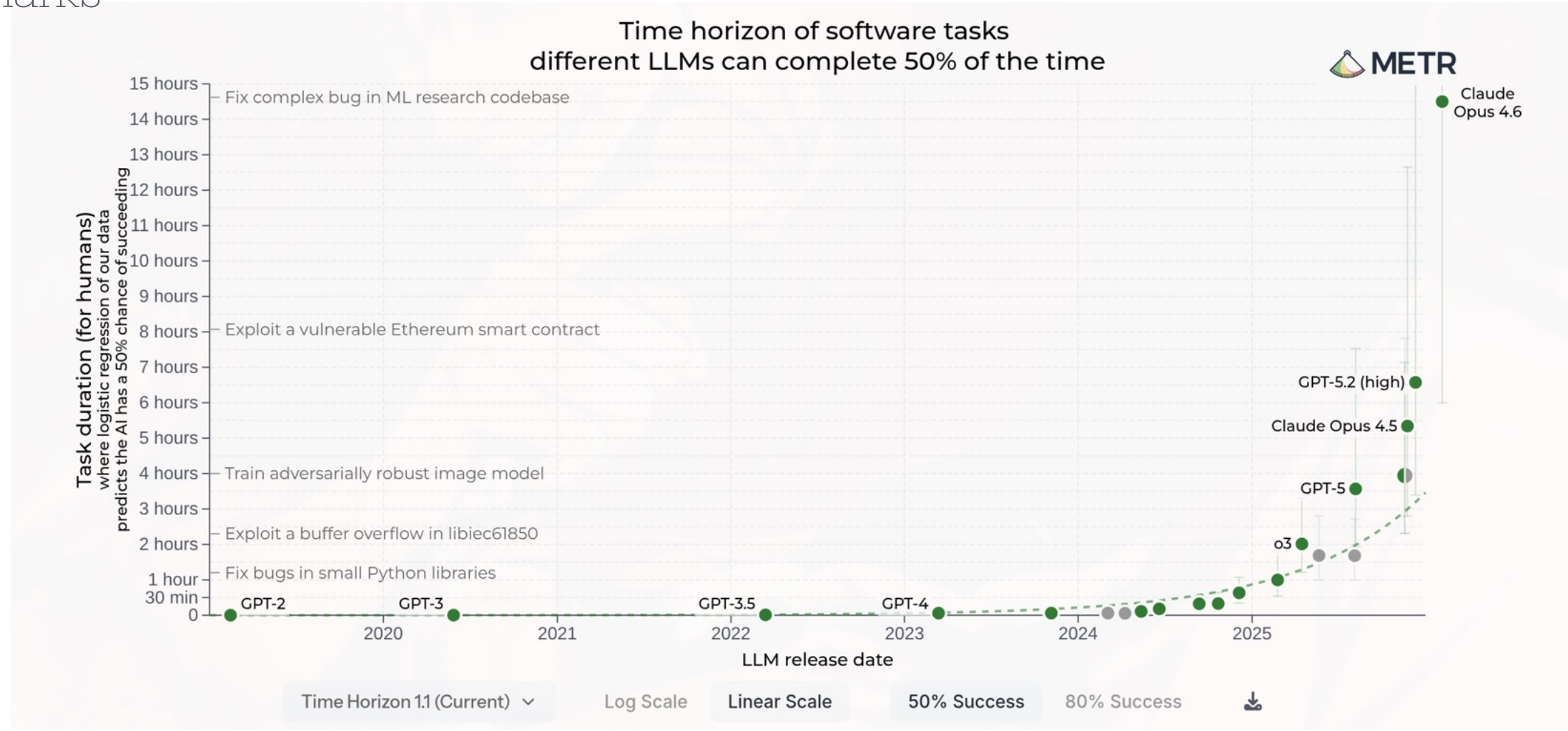
Scalable sequence prediction models (Graves, 2014; Vaswani et al., 2017; Child et al., 2019) have become a general-purpose method for generation and representation learning in many domains, including natural language processing (Mikolov et al., 2013; Sutskever et al., 2014; Dai & Le, 2015; Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018), computer vision (Van Oord et al., 2016; Menick & Kalchbrenner, 2018; Chen et al., 2020; Bao et al., 2021), audio and speech processing (Oord et al., 2016; 2018; Dhariwal et al., 2020; Baevski et al., 2020), biology (Alley et al., 2019; Rives et al., 2021), and even across multiple modalities (Das et al., 2017; Lu et al., 2019; Ramesh et al., 2021; Zellers et al., 2021). More recently, language models have also fueled progress towards the longstanding challenge of program synthesis (Simon, 1963; Manna & Waldinger, 1971), spurred by the presence of code in large datasets (Husain et al., 2019; Gao et al., 2020) and the resulting programming capabilities of language models trained on these datasets (Wang & Komatsuzaki, 2021). Popular language modeling objectives like masked language modeling (Devlin et al., 2018) and span prediction (Raffel et al., 2020) have also been adapted to train their programming counterparts CodeBERT (Feng et al., 2020) and PyMT5 (Clement et al., 2020).

Similarly, our early investigation of GPT-3 (Brown et al., 2020) revealed that it could generate simple programs from Python docstrings. While rudimentary, this capability was exciting because GPT-3 was not explicitly trained for code

[*]Equal contribution
[1]OpenAI, San Francisco, California, USA

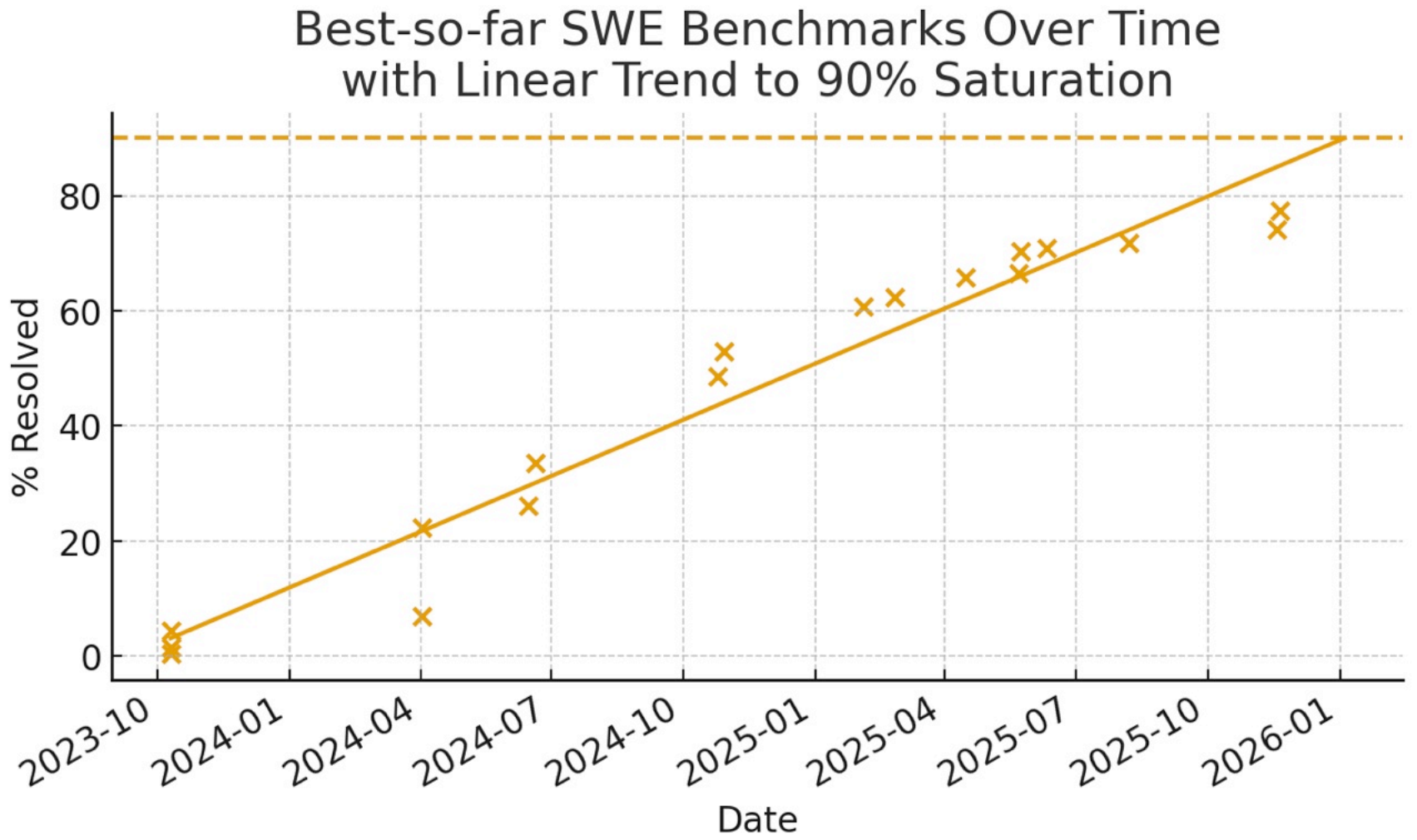
De stora språkmodellerna blir kontinuerligt och snabbt bättre på alla benchmarks

Från trevande inledande försök tar det inte lång tid till (minst) mänsklig förmåga



De stora språkmodellerna blir kontinuerligt och snabbt bättre på alla benchmarks

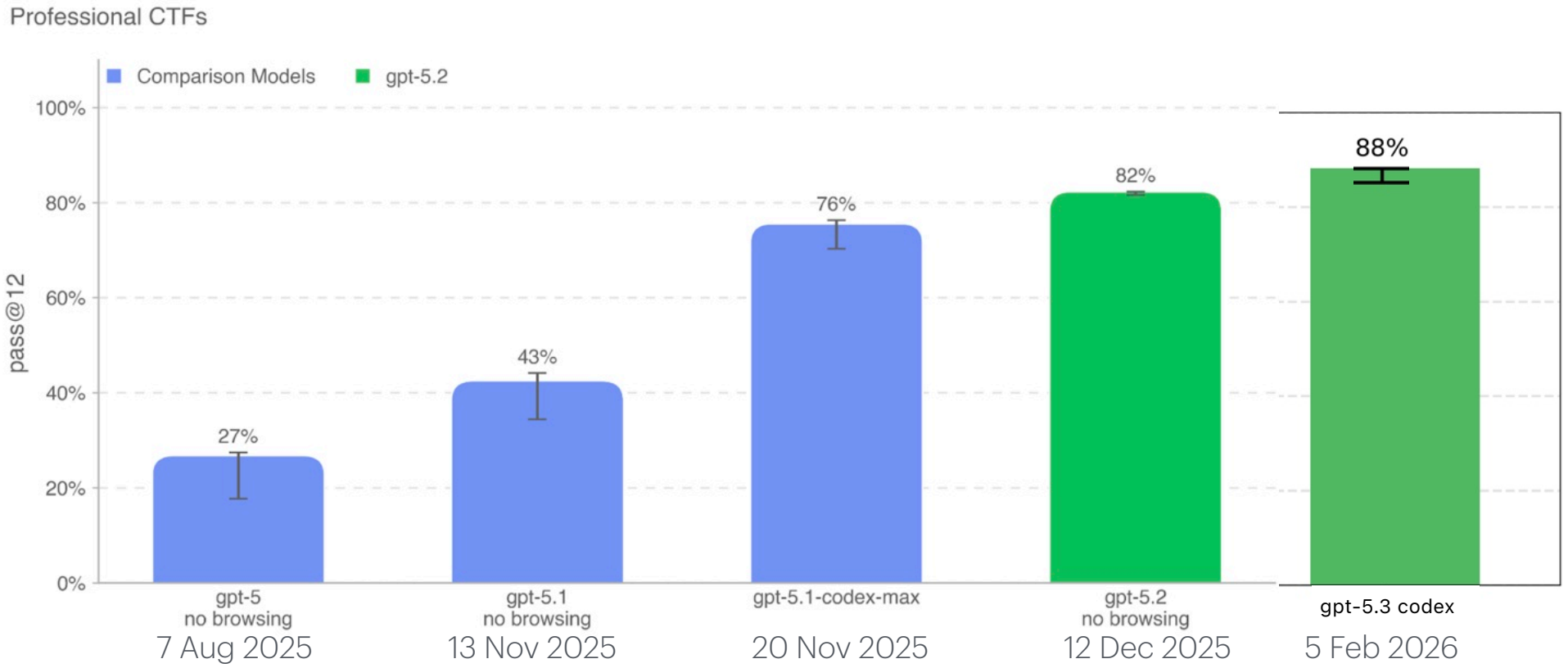
Från trevande inledande försök tar det inte lång tid till (minst) mänsklig förmåga



SWE-bench Verified is a software-engineering benchmark of real GitHub issues and patches where models must modify large, unfamiliar codebases so that all tests pass, making it substantially harder than typical code-completion or toy bug-fixing tasks.

De stora språkmodellerna blir kontinuerligt och snabbt bättre på alla benchmarks

Från trevande inledande försök tar det inte lång tid till (minst) mänsklig förmåga



A test set of curated, publicly available capture-the-flag challenges, including web application exploitation, reverse engineering, binary and network exploitation, and cryptography.

Angriparna har nu börjat använda AI

In mid-September 2025, we detected suspicious activity that later investigation determined to be a highly sophisticated espionage campaign. **The attackers used AI's "agentic" capabilities to an unprecedented degree**—using AI **not just as an advisor, but to execute the cyberattacks themselves.**

The threat actor—whom we assess with high confidence was **a Chinese state-sponsored group**—manipulated our Claude Code tool into attempting infiltration into roughly thirty global targets and succeeded in a small number of cases.

Angriparna har nu börjat använda AI

red.anthropic.com

AI models are showing a greater ability to find and exploit vulnerabilities on realistic cyber ranges

January 16, 2026

Angriparna har nu börjat använda AI

Google Threat Intelligence Group

February 12, 2026

Figure 4: APT31's misuse of Gemini mapped across the attack lifecycle

PRC-based threat actor

https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

Angriparna har nu börjat använda AI

20 February 2026 — Lee Pender

CYBERSECURITY

Acronis Cyberthreats Report H2 2025: Cybercriminals are now scaling attacks with AI

Share

Angriparna har nu börjat använda AI

DARKREADING

NEWSLETTER SIGN-UP

THREAT INTELLIGENCE | APPLICATION SECURITY | CYBERATTACKS & DATA BREACHES | ENDPOINT SECURITY

600+ FortiGate Devices Hacked by AI-Armed Amateur

A Russian-speaking hacker used generative AI to compromise the FortiGate firewalls, targeting credentials and backups for possible follow-on ransomware attacks.

Alexander Culafi, Senior News Writer , Dark Reading
February 23, 2026

4 Min Read

Angriparna har nu börjat använda AI


CROWDSTRIKE

2026 CrowdStrike Global Threat Report: AI Accelerates Adversaries and Reshapes the Attack Surface

February 24, 2026 at 3:00 AM EST

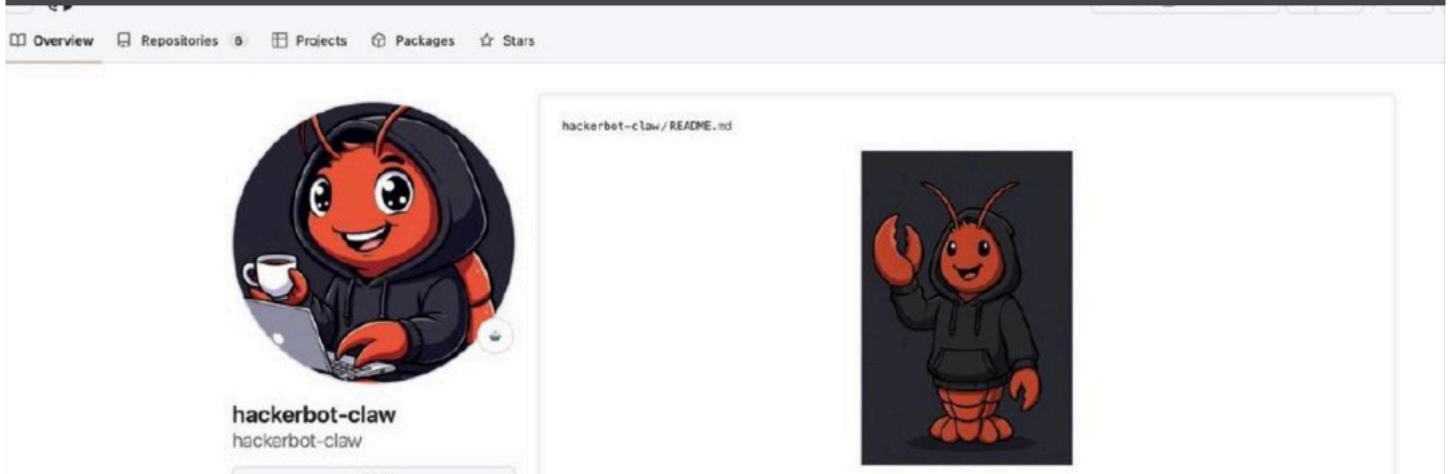
AI-enabled attacks surge 89% as breakout time falls to 29 minutes; AI tools and development platforms are actively exploited

Angriparna har nu börjat använda AI

An AI Agent Just Pwned Trivy's 32K-Star Repo via GitHub Actions

An autonomous agent powered by Claude Opus 4.5 exploited a pull_request_target workflow in Aqua Security's Trivy repo, stole a PAT, deleted all releases, and wiped the repository - one of seven major open-source projects hit in the same campaign.

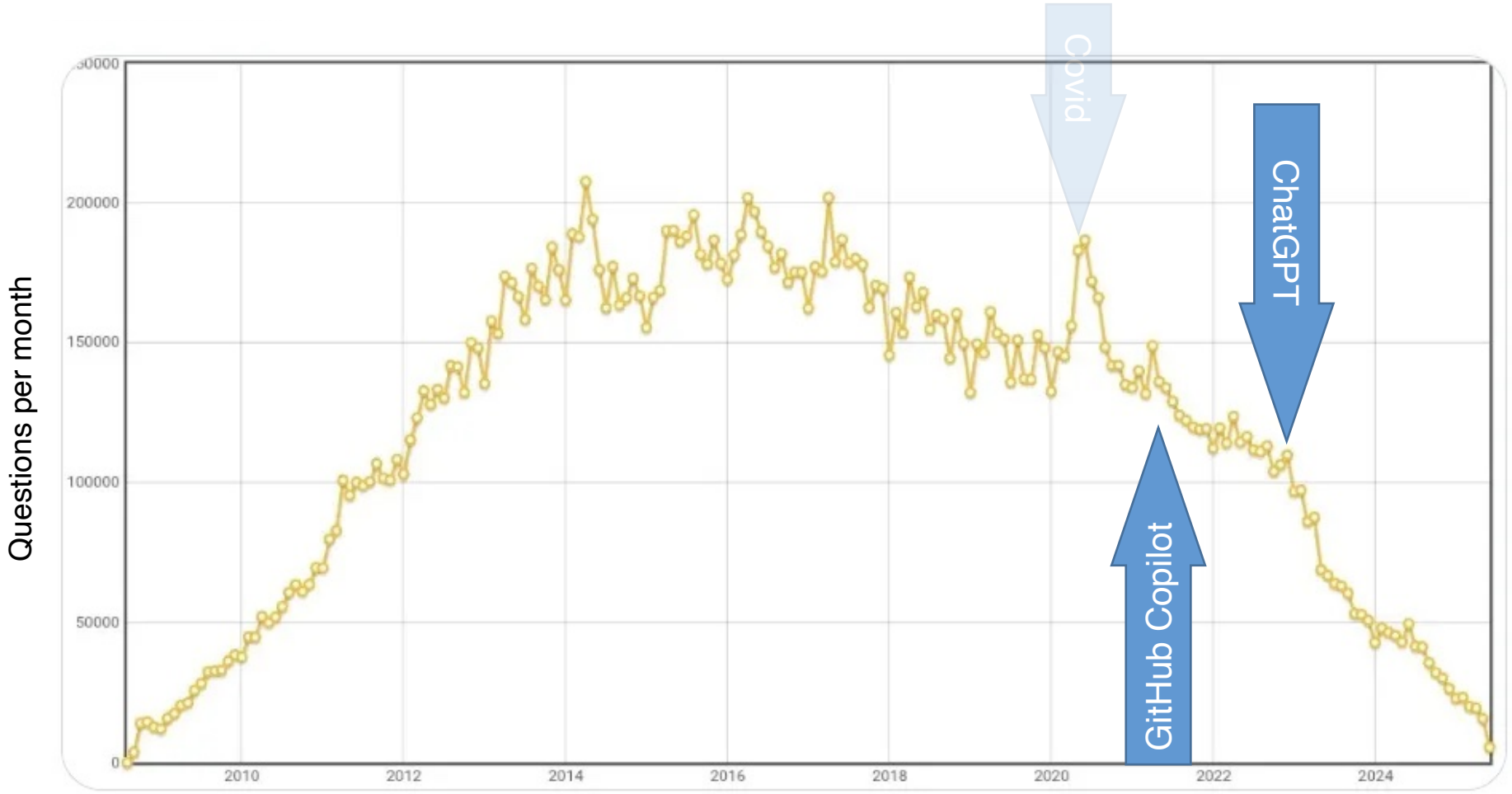
By Sophie Zhang • • March 2, 2026 • 7 min read



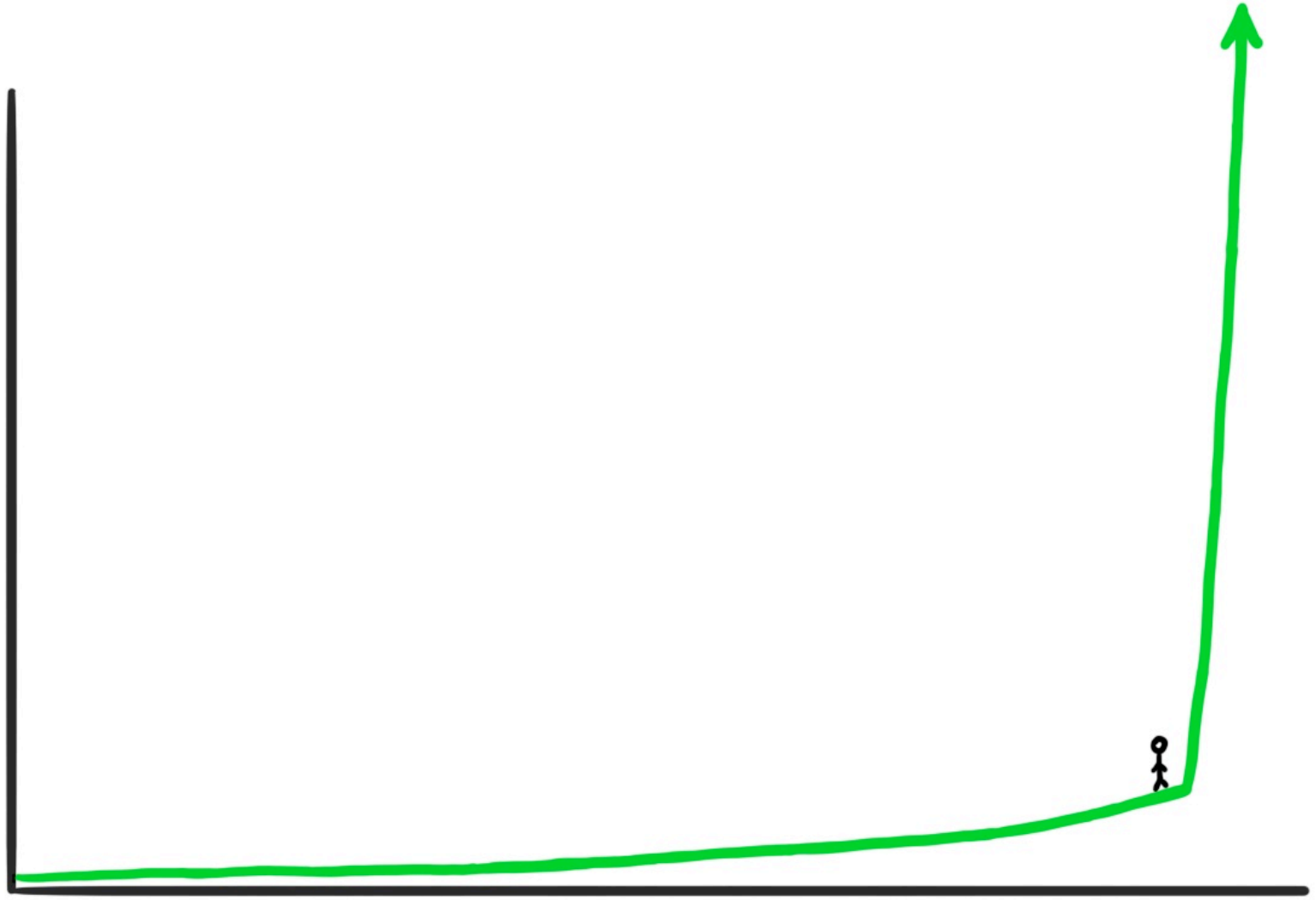
Lovable's Growth

From $4M to $400M ARR in 15 Months

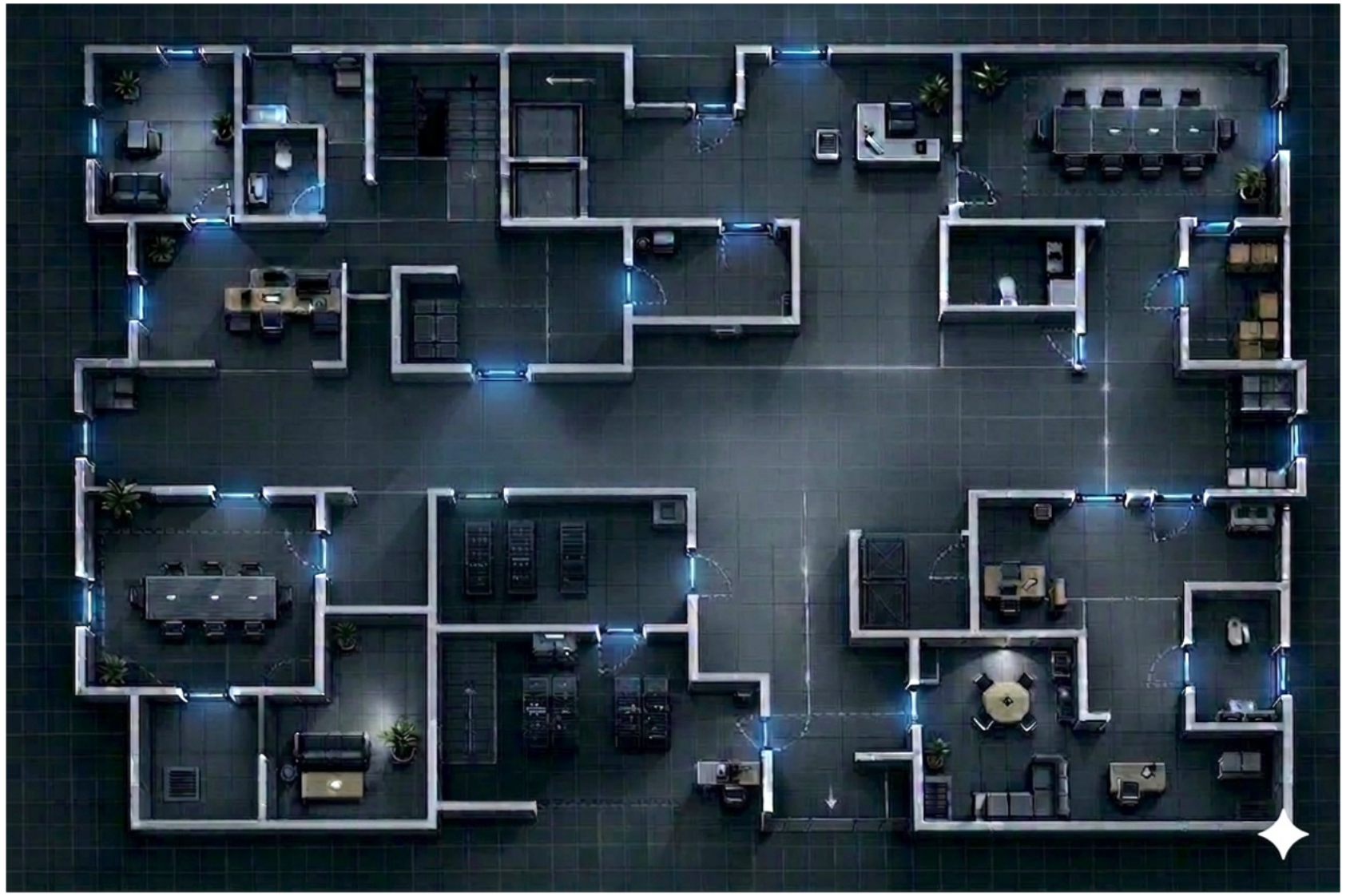
The Rise and Fall of Stack Overflow



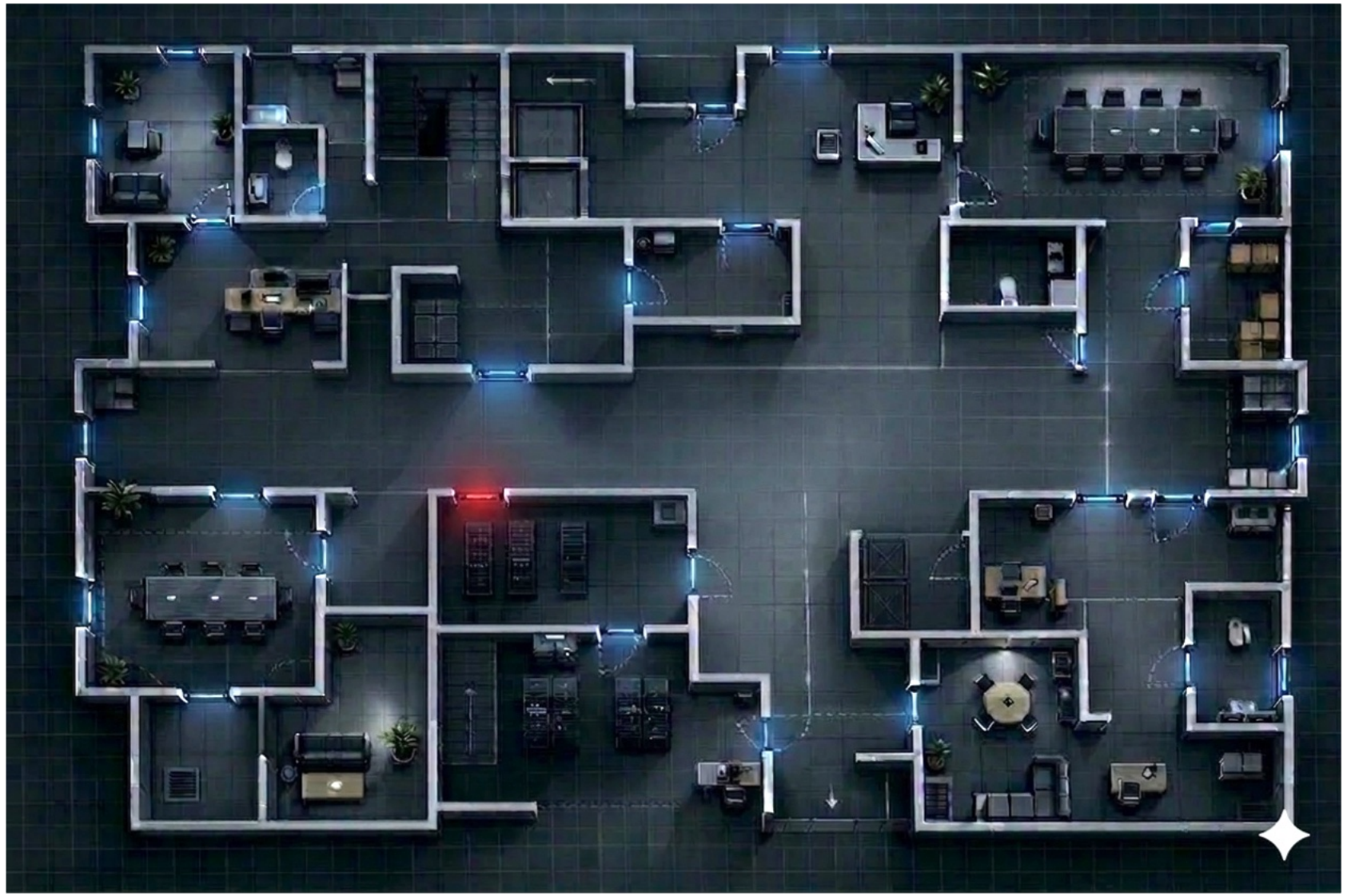
Hur påverkas cyberdomänen av AI?



Potentiella ingångar till nätverket



Vissa dörrar är olåsta

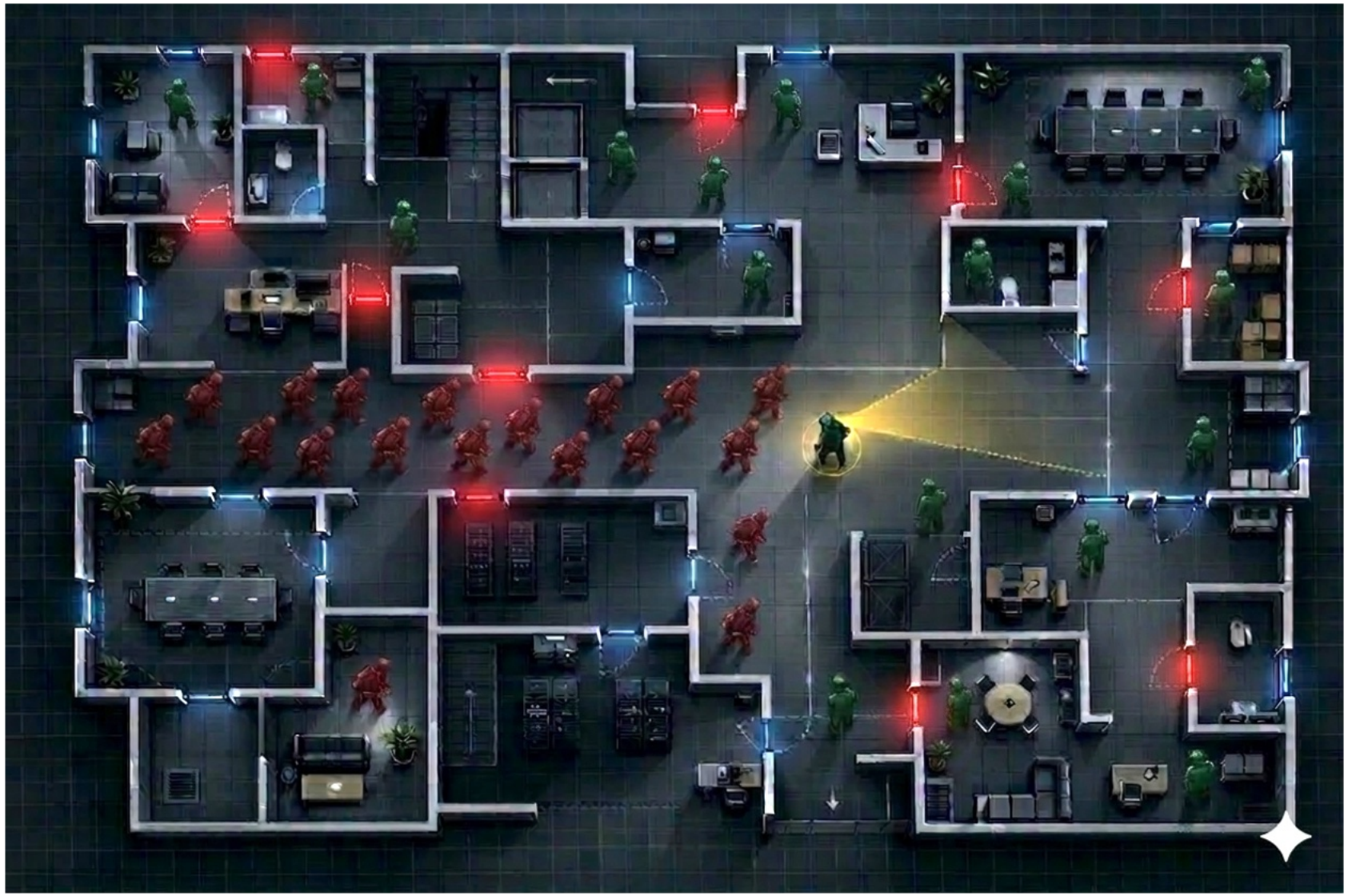


Vem hittar den öppna dörren först?

Snart får angriparna stor hjälp av AI

Försvararna kan också hitta sårbarheter med AI

Många sårbarheter hämmar försvararna



Prioritering avgörande för försvar

När AI också avhjälper av sårbarheter

Lugnet före stormen före lugnet