INFO-TECH AI INSIGHTS

AI Transformation Brief

Featuring AI best practices and insights to enable our members to strategize, plan, develop, deploy, manage, and govern AI-based technologies and solutions.

VOL 4 | APRIL 2024

AI IN THE NEWS

Anthropic Exposes Many-Shot Jailbreaking, a Potential Vulnerability to All Large Language Models

Read News Article

A new research paper by Anthropic highlights a potential security risk for all major large language models (LLMs), called "many-shot jailbreaking." This technique exploits the context window, the LLM's ability to ingest and process large amounts of information.
The paper finds that a long and diverse set of prompts could be used to develop a universal jailbreak technique to bypass the safety filters for any of the major LLMs in the market.

ANALYST ANALYSIS

The performance or accuracy of an LLM is often directly related to the size of the context window. In general, the larger the context window, the more information can be ingested and considered when creating the response. However, as Anthropic's developers uncovered, the larger context window can be used to train and bypass (jailbreak) the safety filters of the LLM, allowing users to direct the LLM to deliver responses to harmful and dangerous requests (alignment failure).
Anthropic's Claude V3 LLM is considered by many as the industry's best-performing (recently surpassed GPT-4 on the LMSYS Chatbot Arena leaderboard for the first time) and safest LLM, but this new technique has been able to bypass the safety filters from all the major LLMs, including those from Anthropic, OpenAI, Google, and Meta. Anthropic has shared its research with the LLM community to raise awareness and drive mitigation initiatives.

ChatGPT Sees 4x Growth in Adoption

Read News Article

In a recent interview, OpenAI's COO Brad Lightcap said over 600,000 people are now signed up to use ChatGPT Enterprise, up from ~150,000 in January 2024. This enterprise-grade product is the most significant commercial offering by OpenAI. OpenAI launched the business version of ChatGPT in August 2023, promising added features and privacy safeguards, including data encryption and a guarantee that OpenAI would not use information from customers to develop technology.

ANALYST ANALYSIS

The significant growth in the adoption of ChatGPT this year demonstrates how organizations have evolved to the stage where they are now deploying and operationalizing their generative AI applications. It is expected that the other major commercial LLM offerings have also seen significant growth in 2024. Open-source LLM offerings have also seen increased activity. Hugging Face, the premier open-source platform for training and deploying models, has seen the number of models increase to 350,000 in 2024.

AI IN THE NEWS

US and UK Announce Partnership on Science of AI Safety

Read News Article

On April 1, 2024, the US and UK signed a Memorandum of Understanding (MOU) that will see them work together to develop tests for the most advanced AI models. This action follows initiatives discussed during the world's first AI Safety Summit in the UK in November 2023, where leaders from 28 countries signed the Bletchley agreement to create safeguards and standards for the safe development and deployment of AI.

ANALYST ANALYSIS

Following the inaugural AI Safety Summit in November 2023 in the UK, both the UK and US established their respective AI Safety Institutes. To ensure AI safety testing is consistent and reliable, the announced collaboration will establish a standardized approach. Both institutes plan to use the same methods and infrastructure, facilitating teamwork. To further solidify this partnership, they will exchange personnel and share information, adhering to all legal and contractual requirements. Additionally, a joint testing exercise is planned on a publicly available AI model.

Cohere Announces Its Latest LLM: Command R+ Read News Article

On April 4, Cohere announced Command R+ as its most capable LLM, with best-in-class capabilities for:

- Advanced retrieval augmented generation (RAG) with citation to reduce hallucinations.
- Multilingual coverage in ten key languages to support global business operations.
- Tool use to automate sophisticated business processes.

ANALYST ANALYSIS

Cohere is known in the industry as a leader in the field of natural language processing and was the first LLM provider to deliver to the market a multilingual language model with support for over 100 languages. Its strategy focuses on providing solutions for business use cases. Training data for its LLMs include business-critical information like financial statements, reports, and industry-specific data. Cohere also prioritizes responsible AI development. By implementing strict measures to mitigate bias and adhering to ethical guidelines, Cohere ensures that its models are not only effective but also trustworthy.

US Congress Bans Staff Use of Microsoft's AI Copilot; The Cyber Safety Review Board's Report on Microsoft Read News Article Read CSRB Report

On March 29, the US House declared Microsoft Copilot as "unauthorized for House use" and said the Office of Cybersecurity has deemed it at risk of leaking House data to non-House-approved cloud services. As a result, all House Windows PCs are blocked from using Copilot, but staff members are allowed to use Copilot AI on their personal phones and laptops. The House's Office of Cybersecurity reportedly fears that snippets of sensitive congressional data might unknowingly be available through the codebase, creating a potential vulnerability for unauthorized access. Microsoft acknowledged the security exposure and pledged to have an offering later this year to meet the government's security requirements.

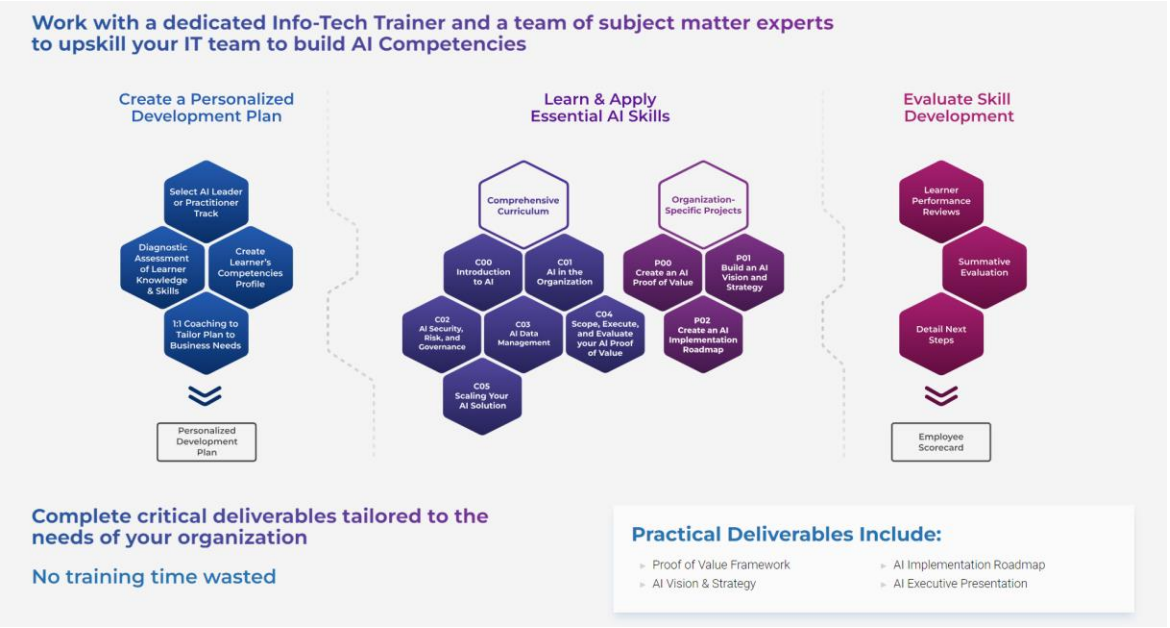
ANALYST ANALYSIS

Paralleling the ban of Copilot, the Cyber Safety Review Board recently published a report blasting Microsoft and its handling of last year's intrusion by the hackers known as Storm-0558 on US government officials. The Board cited that Microsoft has "a corporate culture that deprioritized enterprise security investments and rigorous risk management, at odds with the company's centrality in the technology ecosystem and the level of trust customers place in the company to protect their data and operations." Microsoft's response is like its response for Copilot's security shortcomings: the company acknowledged its lack of enterprise-grade security and transparency, apologized, and pledged to do better by adopting security reforms across their products.

The challenge for Microsoft is addressing their lack of experience and desire to create "enterprise-grade" offerings vs. delivering more functions and integration with its ecosystem of products, as the report points out. In addition, the report calls out that Microsoft lacks the culture to develop best-of-breed products for the market and relies on delivering "good enough and integration into its ecosystem," making it difficult for customers to migrate to or integrate third-party tools.

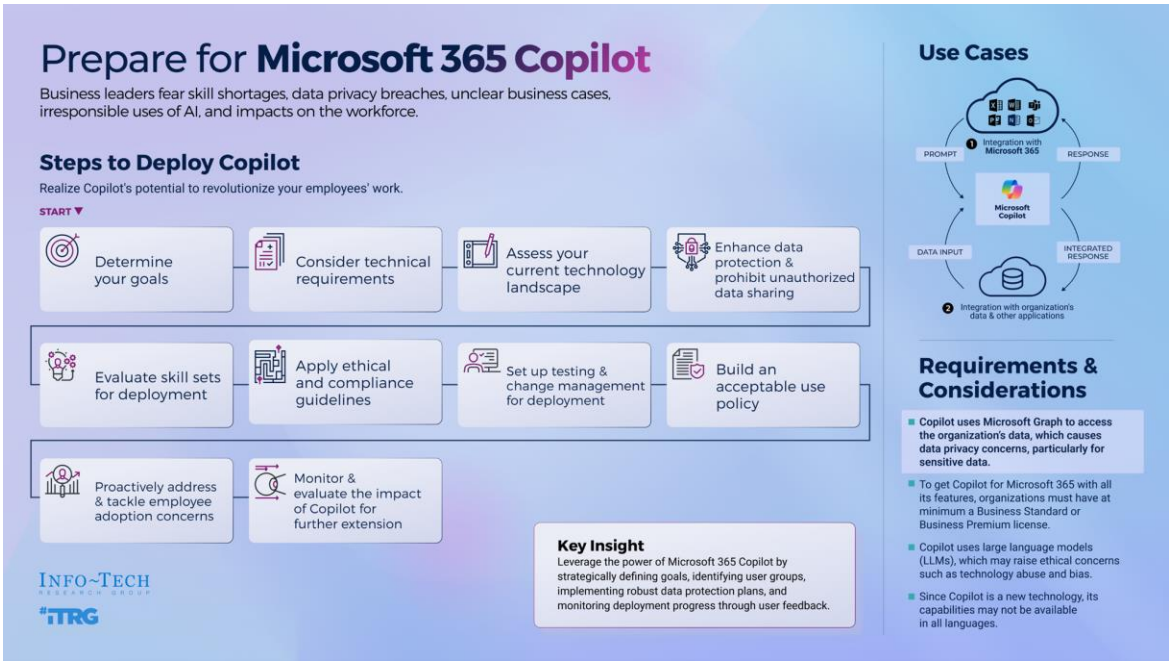
Enroll in the AI Workforce Development Program New!

The AI Workforce Development Program is a 12-week AI training program designed to help IT leaders and practitioners gain essential business and technical skills to support the responsible, reliable, and effective adoption of AI for their organizations.



For more information or to enroll: AI Workforce Development Program

Prepare for Microsoft 365 Copilot New!



Source: Prepare for Microsoft 365 Copilot
Info-Tech Research Group

VENDOR SPOTLIGHT

AI Marketplace

Unlock the potential of AI tailored to your needs and transform possibilities into reality with our dedicated support.

alwaysAI Computer Vision Solutions

Vispera Visual Intelligence for Retail

UPCOMING AND RECENT EVENTS

Info-Tech Leadership Summit, Vancouver, BC – May 29-31, 2024 →

Info-Tech Leadership Summit, Chicago, IL – June 19-21, 2024 →

Info-Tech LIVE 2024, Las Vegas, NV – September 17-19, 2024 →

Deploy AIOps to Improve IT Operations Webinar, April 24, 2024 (APAC) →

Govern the Use of AI Responsibly With a Fit-for-Purpose Structure Webinar, April 24, 2024 →

AI Priorities and Trends Event, Manhattan, NY – April 24, 2024; Contact: Jennifer Case →

How Transformative Technologies Will Revolutionize Tomorrow's Businesses – HBR Event, Manhattan, NY – April 25, 2024; Contact: Bill Wong →

Build Your AI Strategic Roadmap Webinar, May 8, 2024 →

CIO Roundtable on AI Event, London, UK – June 4, 2024 (Public Sector); Contact: Michelle van Wijk →

CIO Roundtable on AI Event, London, UK – June 6, 2024 (Private Sector); Contact: Michelle van Wijk →

(and many more events…)

AI AND DATA ANALYTICS SOLUTIONS – RESOURCES

AI Marketplace

Artificial Intelligence Research Center

AI Workforce Development Program

Workshops

Build Your AI Strategic Roadmap

Launch Your AI Proof of Value

Build a Scalable AI Deployment Plan

Govern the Use of AI Responsibly With a Fit-for-Purpose Structure

AI EDITOR-IN-CHIEF

Bill Wong – Info-Tech AI Research Fellow

INFO-TECH RESEARCH GROUP