BY ARMAND RUIZ

AI in 2023

Top Highlights

"AI will have a more profound impact on humanity than fire, electricity and the internet"

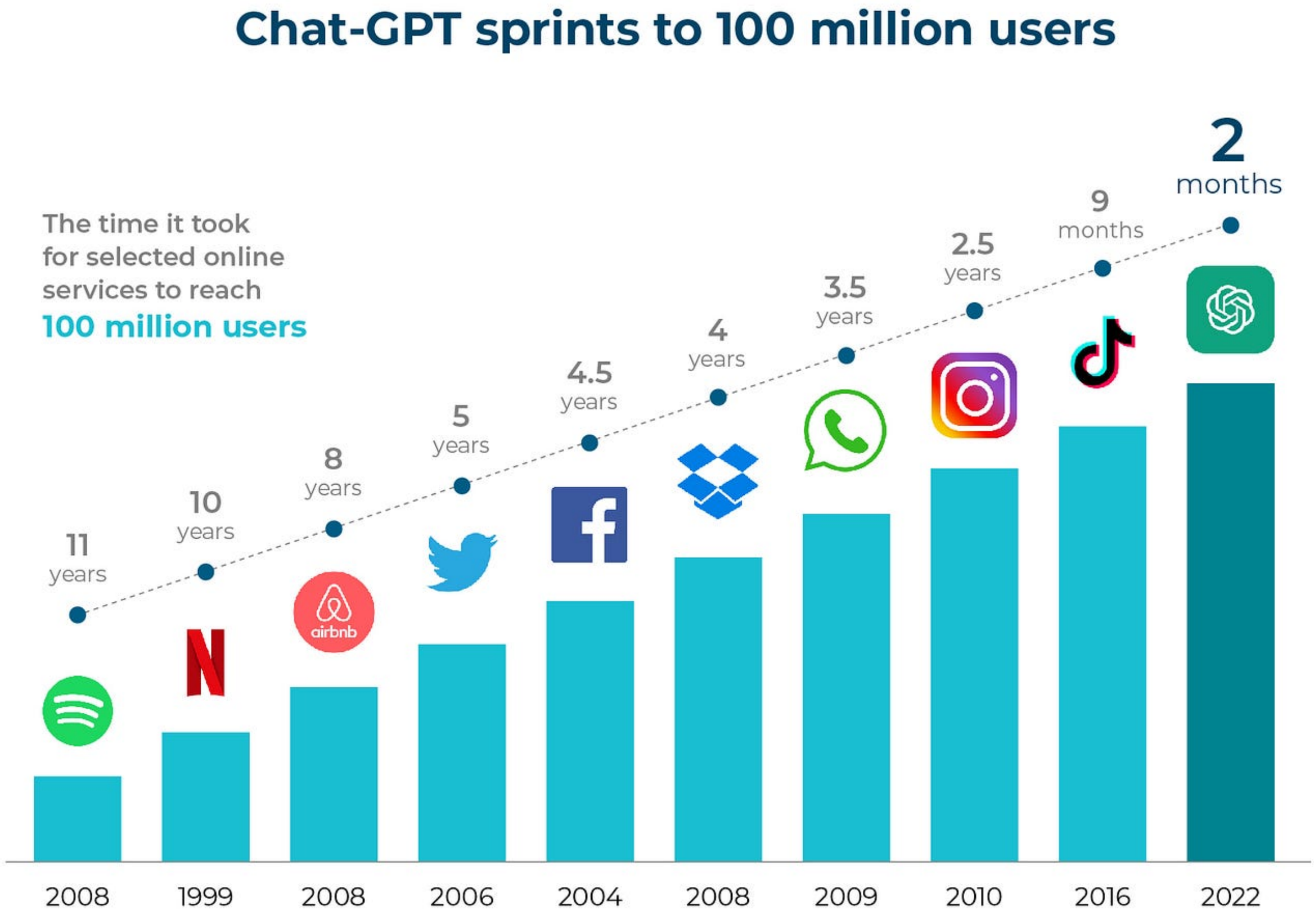
Sundar Pichai,
CEO Alphabet

ChatGPT sets record for fastest-growing user base

All-time record in growth, and with this, the hype in AI began.

Its capabilities amazed everyone, and suddenly, AI became the #1 technology to incorporate everywhere. The question now was... how?

And with that, the year of AI innovation started.



Meta announced Llama

Meta has introduced LLaMA, a smaller, more accessible LLM, to enable broader research in AI.

Available in various sizes, it's designed for easy fine-tuning on diverse tasks with lower resource requirements.

Released under a noncommercial license, it's accessible to selected researchers globally, aiming to promote responsible AI development.

Adobe announced Firefly

Adobe's Firefly model, enables easy creation of high-quality images and text effects, integrating with Adobe's various Cloud platforms.

It includes a "Do Not Train" tag for creators to opt-out of model training and plans to allow customization with personal creative content.

Giving Creators New superpowers to work at the speed of their imaginations.

Pause AI Experiments: An Open Letter

Tech leaders, including Steve Wozniak and Elon Musk, signed an open letter urging a six-month pause in training AI systems more powerful than GPT-4 to assess the risks of AI's rapid progress.

Concerns included job loss, human obsolescence, and a loss of control over civilization. It remains unclear if AI companies have taken heed of this call.

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

1079

Add your signature

OpenAI released APIs for ChatGPT

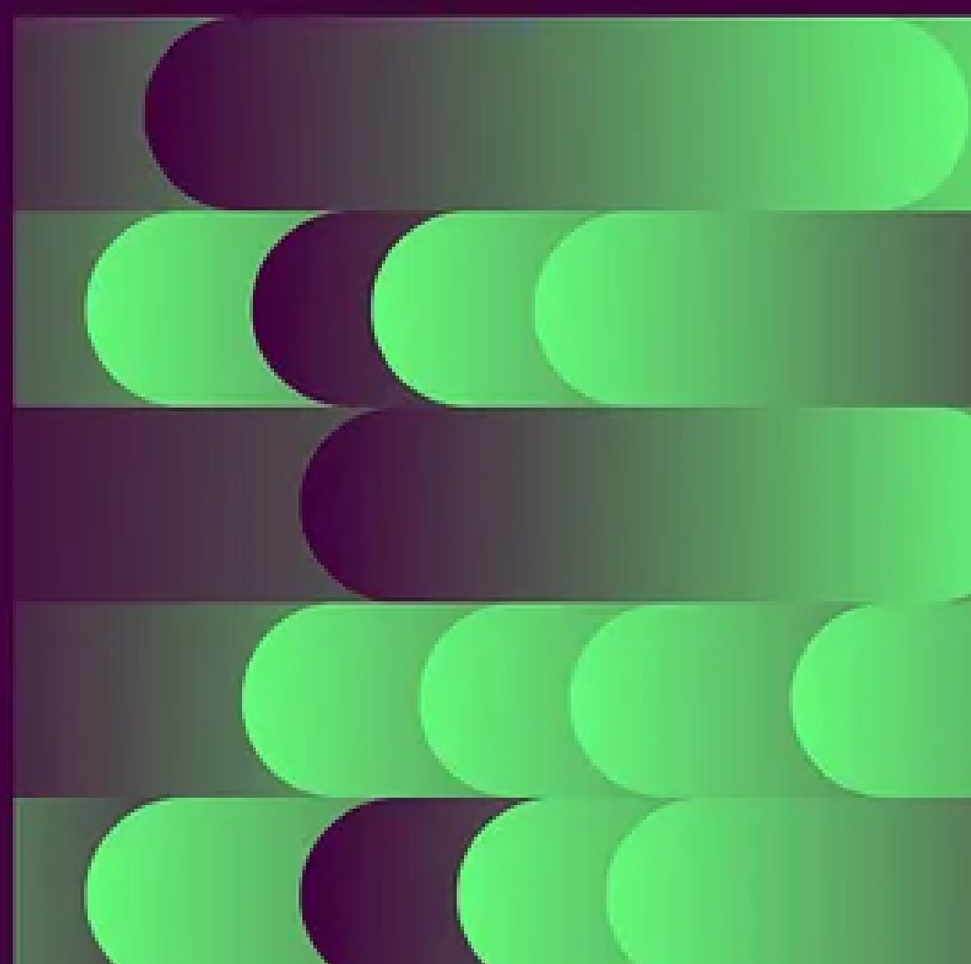
OpenAI has launched APIs for ChatGPT and Whisper, allowing easy integration of these advanced AI tools into various products.

This development enables companies to incorporate ChatGPT's capabilities directly at a low cost, likely leading to widespread adoption in consumer apps.

The Whisper API enhances speech-to-text functionality, promising to spur new speech-based applications.

Introducing ChatGPT and Whisper APIs

Developers can now integrate ChatGPT and Whisper models into their apps and products through our API.



GPT-4 was released

OpenAI's GPT-4, launched on March 13, improves in functionality, responsiveness, and safety over previous models but still faces limitations, like generating incorrect information.

It introduces advanced capabilities like image processing and extended text generation, making it useful in various business applications. However, users are advised to verify its outputs for accuracy.

Khan Academy launched Khanmigo

Khanmigo, an AI-powered teaching assistant from Khan Academy, assists in math, science, and humanities.

It supports students with writing and debate while helping teachers create lesson materials.

It is accessible on most Khan Academy pages and it has a 4-star rating from Common Sense Media.

Introducing Zoom AI Companion

Zoom AI Companion is an AI assistant within Zoom, offering features like chat assistance, meeting summaries, and brainstorming tools.

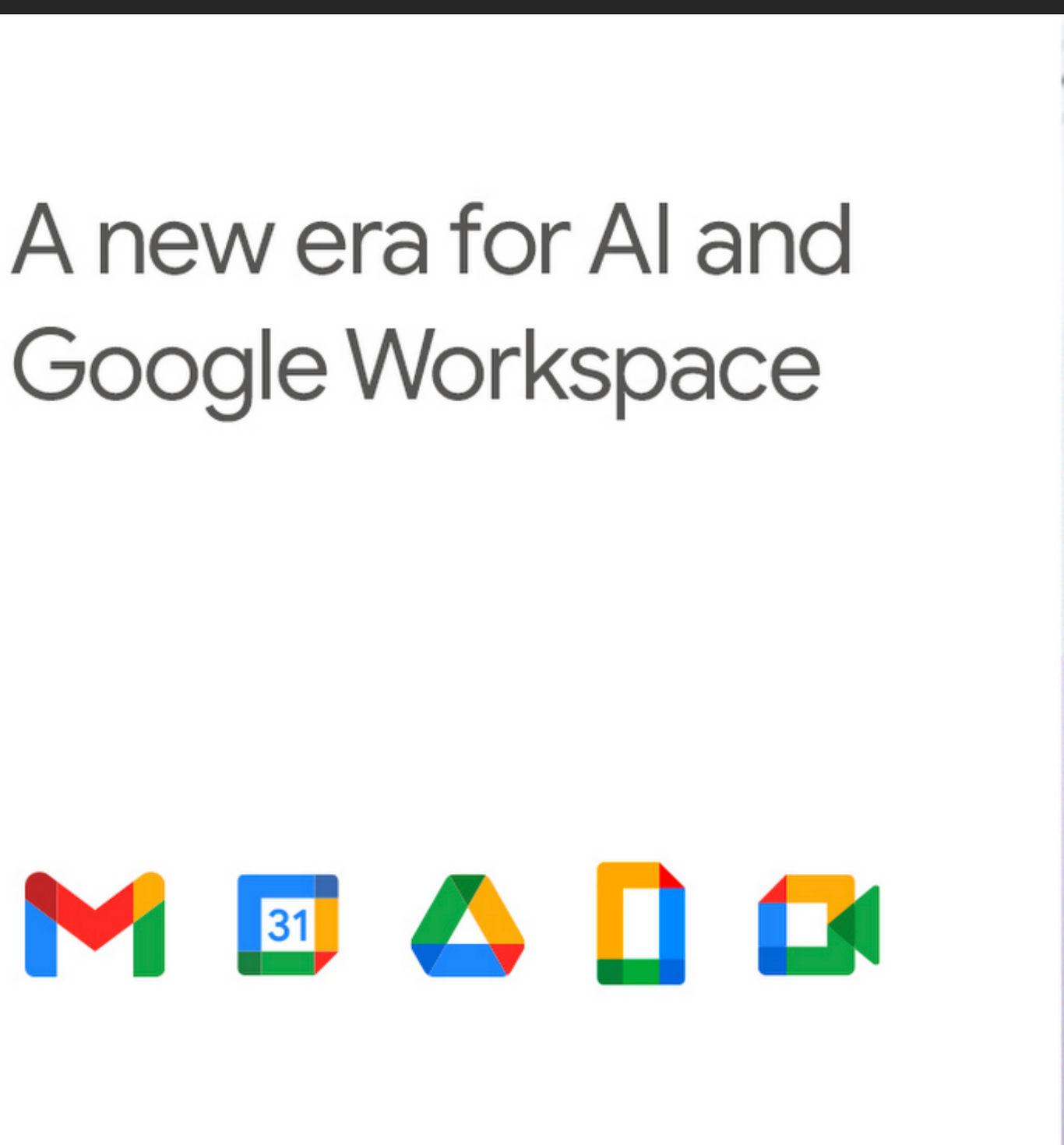
Planned updates include conversational interaction and real-time support.

Included in Zoom's paid services, it prioritizes data privacy and allows administrators to control its activation and use.

Google integrated AI into Workspace

Google announced new AI-powered features for Google Workspace aimed at enhancing writing and productivity for its users.

The rollout is part of Google's broader vision to incorporate AI as a collaborative tool across its suite, including Gmail, Docs, Slides, Sheets, Meet, and Chat.



Github released Copilot X

GitHub Copilot X, leveraging OpenAI's GPT-4, is an AI assistant for coding, integrated with Visual Studio and VS Code.

It provides code suggestions and debugging help and is accessible through chat and terminal interfaces.

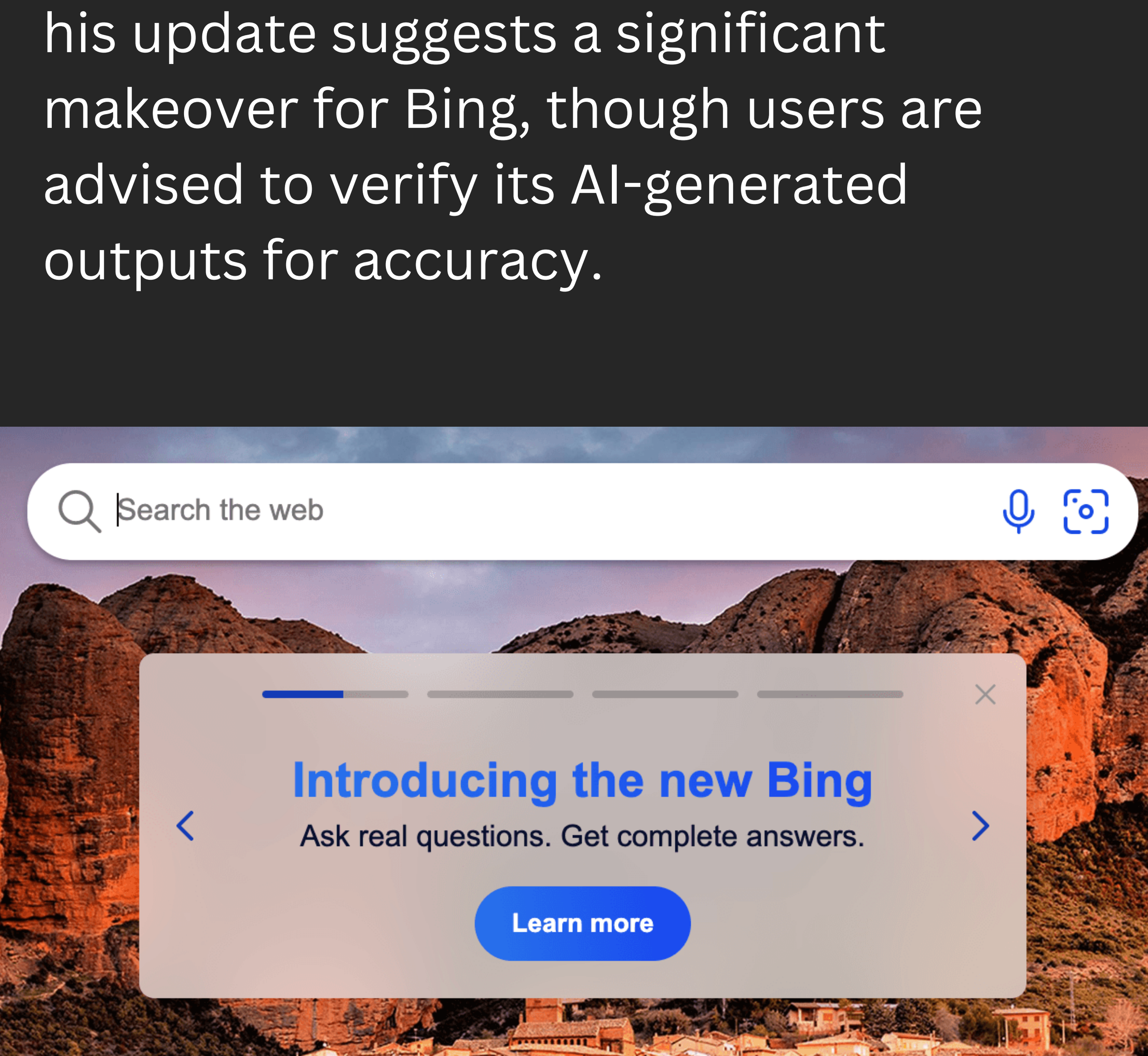
Users can install it via Visual Studio Code and its extensions. Copilot X is free for students, teachers, and open source contributors, and is compatible with various IDEs including Visual Studio, Vim, and JetBrains.

GitHub Copilot X

Microsoft Bing made a comeback

Microsoft's Bing search engine added AI chatbot ChatGPT, offering a new interface with enhanced capabilities like full conversations, creative tasks, and access to current information.

his update suggests a significant makeover for Bing, though users are advised to verify its AI-generated outputs for accuracy.



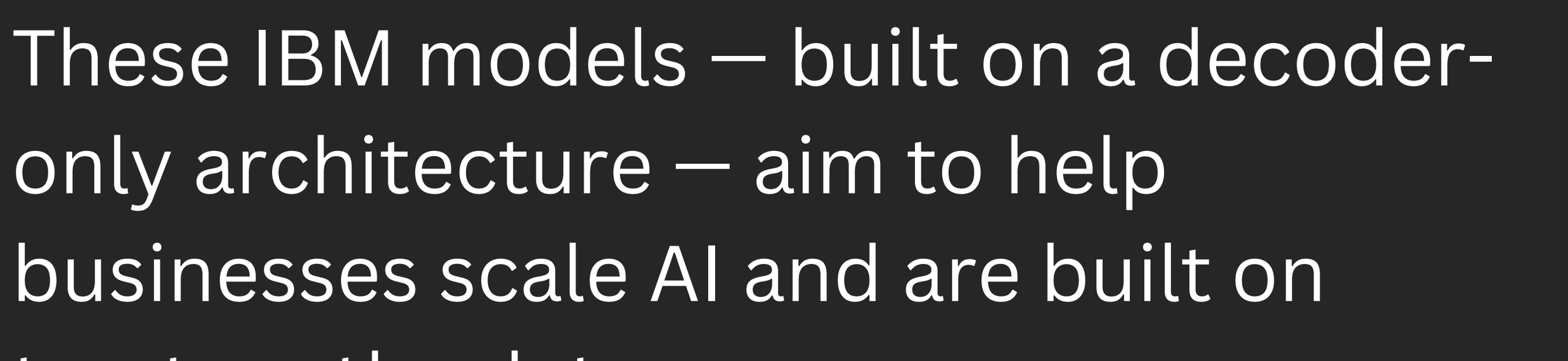
IBM announced new Enterprise LLMs

IBM's Granite models apply generative AI to the modalities of language and code.

Recognizing that a single model will not fit the unique needs of every business use case, the Granite models are being developed in different sizes.

These IBM models — built on a decoder-only architecture — aim to help businesses scale AI and are built on trustworthy data.

Model architectures



Slate

Non-generative encoder-only architecture

Sandstone

Lightweight encoder-decoder architecture

Granite

Decoder-only architecture

Obsidian

Novel-sparse universal transformer architecture

Moonstone

Novel architecture based on dense associative memory

AI Agents as the next big thing in AI

AI agents, using LLMs, autonomously execute tasks and can handle complex activities like online research and computer control.

These technologies promise significant advancements but also raise concerns about job displacement, biases, and accountability in various applications.

New frameworks appeared such as AutoGen or AutoGPT.

Russia’s Sberbank announced GigaChat

Russian bank Sberbank released GigaChat, an AI chatbot rival to ChatGPT, emphasizing its advanced Russian language capabilities.

This move is part of a broader tech industry trend to leverage AI, and aligns with Russia's goal to reduce import reliance amid Western sanctions.

It was described as “a breakthrough for the larger universe of Russian technology.”

Microsoft announces Copilot, your everyday companion

Microsoft is launching Microsoft Copilot, an AI companion, across its products, including Windows 11, Microsoft 365, and Edge.

Copilot will use web context, work data, and real-time PC activity to provide enhanced assistance with a focus on privacy and security.

To begin human trials for implanting AI-powered brain chips

Elon Musk's Neuralink has gained FDA approval for human trials of its brain implants, despite facing scrutiny and investigations over its animal testing methods and safety concerns.

The company, which aims to treat various conditions with its technology, previously had its application rejected by the FDA but has now crossed a significant milestone in its development.

Chinese tech giants Alibaba and Huawei entered the AI game

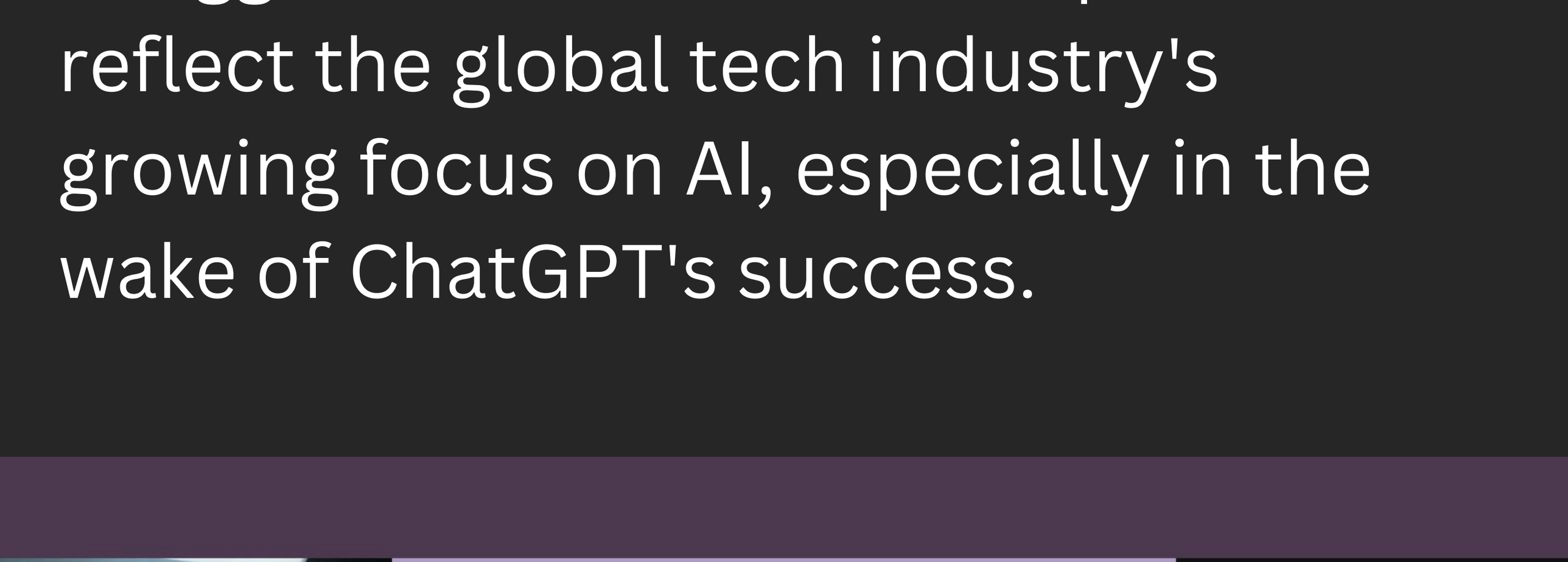
New significant AI products. Alibaba unveiled an AI image generator, Tongyi Wanxiang, to compete with OpenAI's DALL-E, and a developer tool, ModelScopeGPT.

Huawei launched its industrial-focused Panggu model. These developments reflect the global tech industry's growing focus on AI, especially in the wake of ChatGPT's success.

Spotify introduced AI Voice translation for podcasts

New significant AI products. Alibaba unveiled an AI image generator, Tongyi Wanxiang, to compete with OpenAI's DALL-E, and a developer tool, ModelScopeGPT.

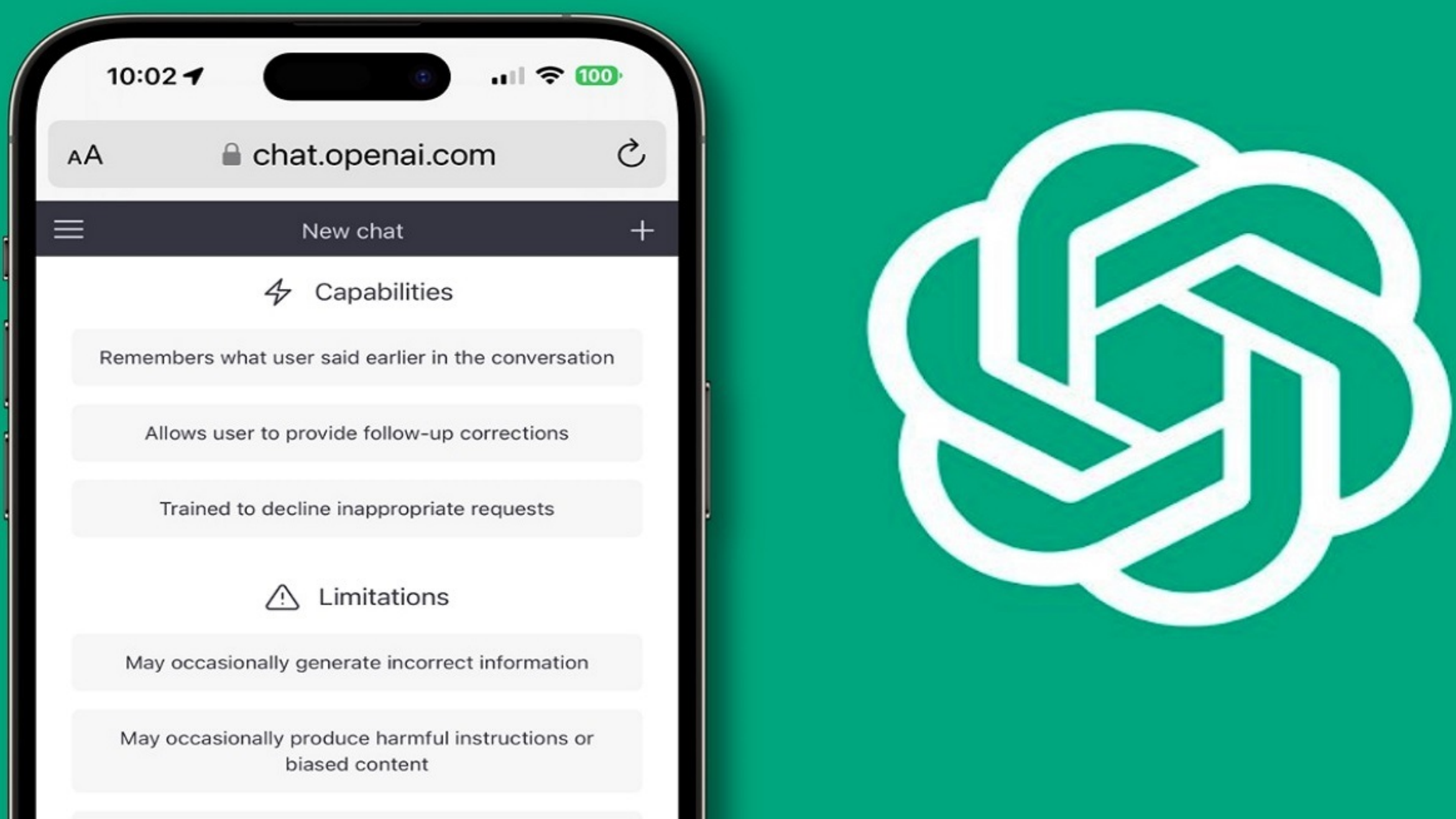
Huawei launched its industrial-focused Panggu model. These developments reflect the global tech industry's growing focus on AI, especially in the wake of ChatGPT's success.



ChatGPT available as a mobile app

OpenAI has released an official ChatGPT iOS & and Android app in the U.S., offering free, ad-free access to its AI chatbot with voice input.

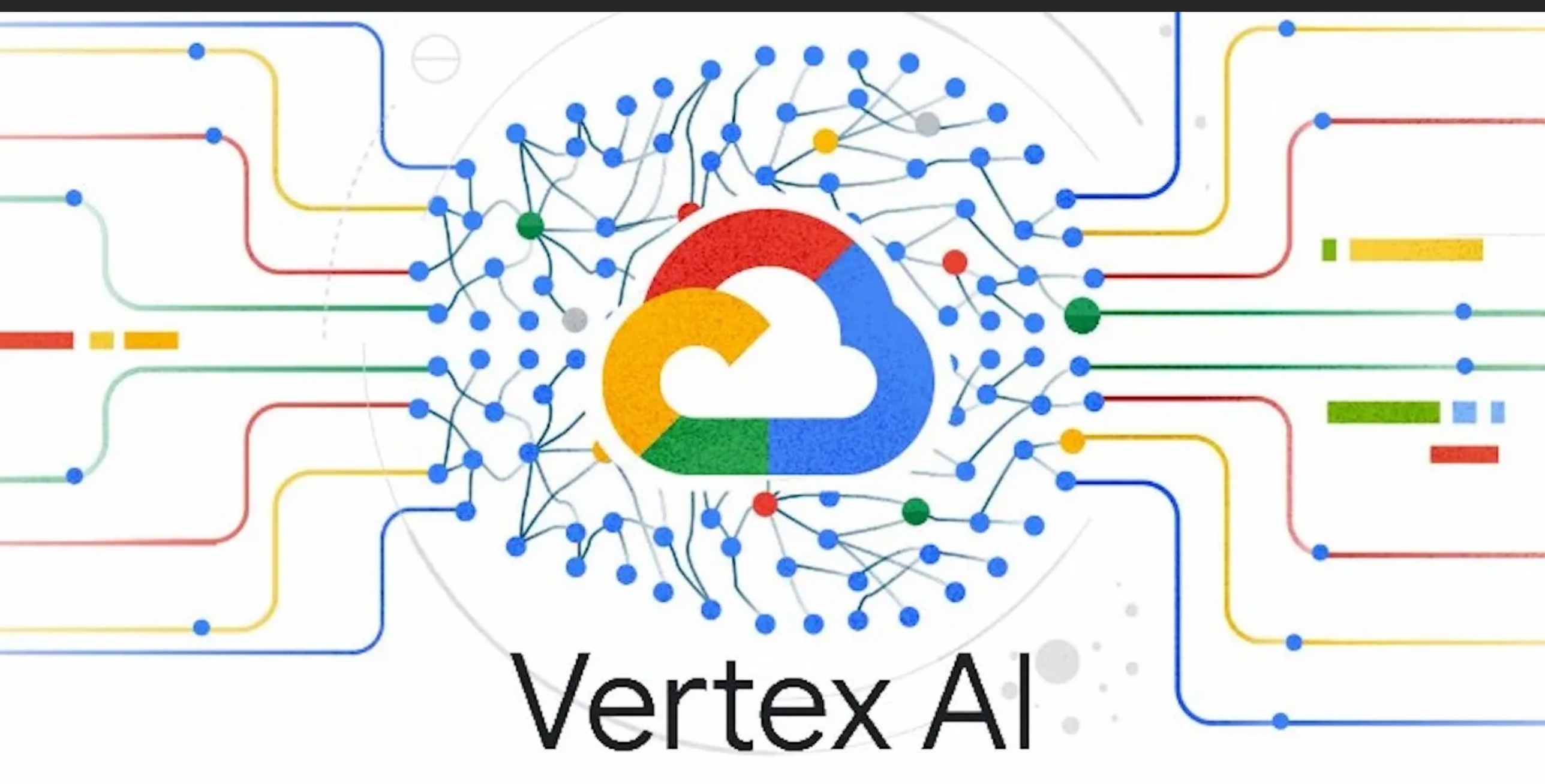
The app syncs search history across devices and poses a potential challenge to Apple's Siri and Google's search engine.



Google released Vertex AI platform with GenAI

Google Cloud has enhanced its Vertex AI platform with increased model capacity, new languages, and improved services like AI Search and Conversation tools.

They've also partnered with NVIDIA for advanced AI supercomputing and introduced more efficient fifth-generation cloud tensor processing units (TPUs).



Meta released first two AI chips

Meta has developed custom AI chips, including the Meta Scalable Video Processor (MSVP) for efficient video processing and the Meta Training and Inference Accelerator (MTIA) for various AI tasks.

This move aims to enhance performance and energy efficiency in AI. The announcement reflects Meta's strategic investment in improving its infrastructure for AI.

Google released PaLM 2

PaLM 2 is a new language model with improved multilingual, reasoning, and coding capabilities, suitable for a wide range of applications.

It powers over 25 Google products, including specialized versions for medical and cybersecurity purposes.

Available in various sizes, PaLM 2 marks a significant step in Google's AI advancement, with future developments like the multimodal Gemini model in progress.

Microsoft announces AI services on Azure

Azure AI introduced several capabilities like Azure AI Studio, Prompt Flow, OneLake integration, Model Catalog expansion, Model Benchmarks, Managed Feature Store, Pipeline Component Deployments, and Serverless Compute to streamline AI development and management.

These enhancements, on top of GPT-4 integration, make AI development more accessible for developers and data scientists.

Open Source AI becomes very real

2023 marked a significant rise in public interest and advancements in LLMs, with a focus on open-source models.

Key developments included the release of open-source LLMs, improvements in model personalization, and the emergence of smaller, more accessible model sizes.

These advancements, driven by companies and research labs, highlight the growing importance of open-source contributions and community.

The AI community building the future.

The White House AI Executive Order

The Biden Administration's Executive Order on AI emphasizes fostering innovation while ensuring responsible development, focusing on AI talent development, safety measures, and ethical governance.

It tasks federal entities with various mandates, aiming to maintain U.S. leadership in ethical AI with a focus on implementation and resource allocation.

Nvidia dominates the market and keeps innovating

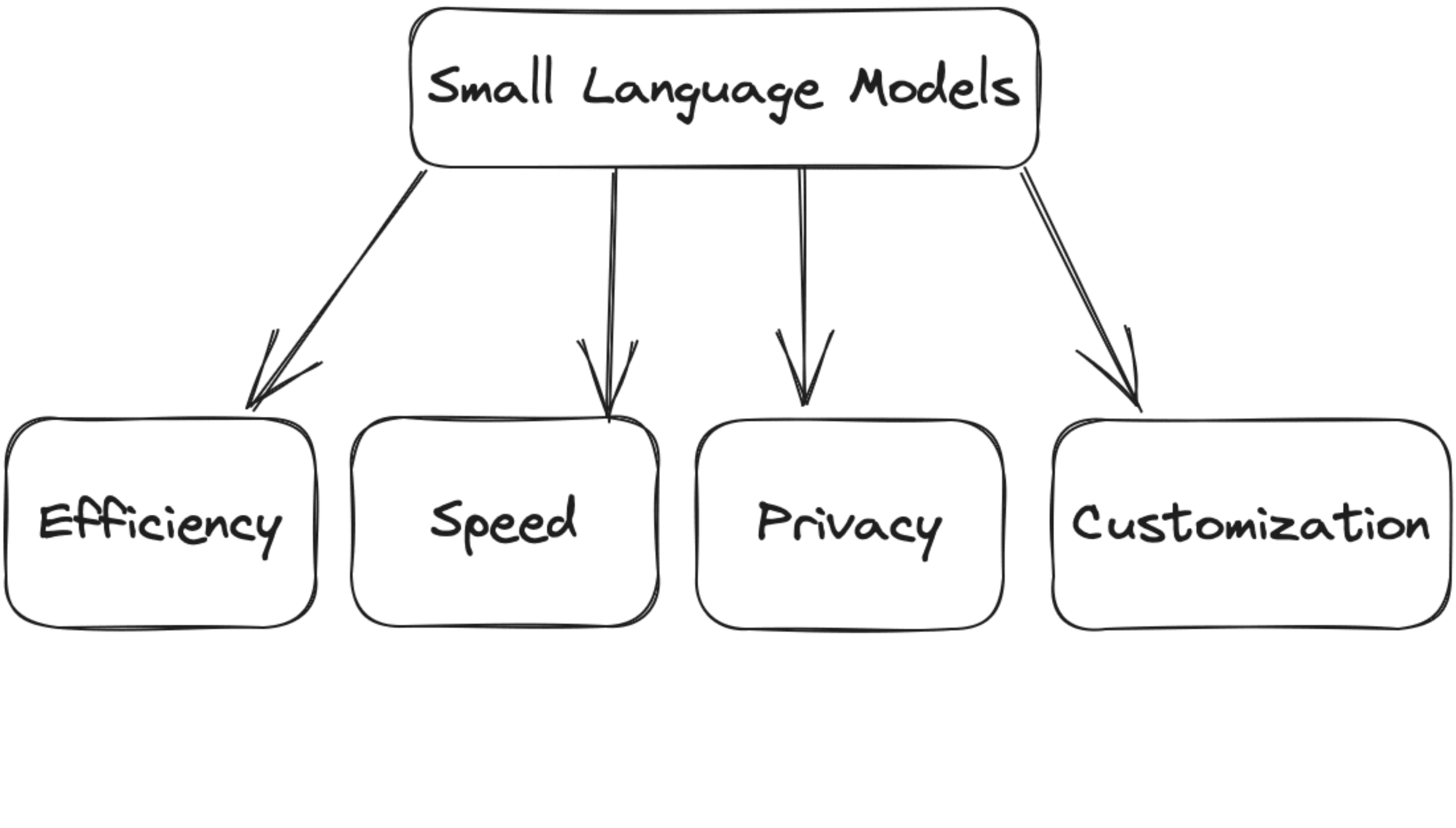
Today, Nvidia's GPUs are vital for running large AI models like ChatGPT, dominating about 88% of the GPU market.

The company's success in AI, driven by its software-focused approach and ecosystem development, positions it strongly in the burgeoning generative AI space.

Small Language Models take the stage

Small Language Models (SLMs) are efficient, cost-effective alternatives to larger AI models like GPT-4, offering customization for specific tasks.

Despite having fewer parameters, SLMs still deliver effective performance for targeted applications, particularly business-specialized sectors.



Mistral AI_ released impressive LLMs

Mistral.ai, with leaders from DeepMind and Meta, champions open-source AI with a European focus, targeting finance and law sectors.

Their strategy involves developing efficient LLMs and competing against major US tech companies.

The company emphasizes transparency, data sovereignty, and specialized AI applications.

MISTRAL AI_

OpenAI leadership crisis

OpenAI's CEO Sam Altman was abruptly fired and then reinstated after a brief, high-profile leadership crisis.

This event, while causing a brief media frenzy and revealing tensions between rapid AI development and safety concerns, also showcased the power of collective employee action and the influence of investors like Microsoft.

Microsoft announced Maia, the first AI chip on Azure

Microsoft announced Azure Maia AI Accelerator for AI workloads on Azure.

This chip is part of Microsoft's effort to optimize its infrastructure systems from silicon to software, allowing for better performance, sustainability, and cost optimization.

They will be deployed in Microsoft's datacenters next year to meet the increasing demand AI.

EU AI Act is approved

The EU AI Act is the world's first major law regulating AI.

It aims to ensure AI systems are safe, transparent, non-discriminatory, and respect fundamental rights.

The Act categorizes AI based on risk, with stricter rules for high-risk applications like facial recognition and social scoring. Set for implementation in 2025, aims to foster AI innovation in Europe while protecting citizens.

Figure 1—Pyramid of Risk

McKinsey calls GenAI a potential $4.4 trillion game-changer for the global economy!

GenAI could unlock $4 trillion yearly, supercharging industries like banking and automating up to 70% of work, especially brain work.

This productivity boom needs paired with worker training for a smoother transition.

Meta released Llama 2, with impressive results

Trained on massive data, Llama 2 excels at reasoning and knowledge tasks, but limitations remain like English-only use and potential bias.

Meta emphasizes the responsible use of resources and feedback programs. This open approach sparks debate while pushing AI development forward.

It has a commercial license.

Anthropic unveiled Claude 2, a multimodal model

Claude 2, Anthropic's next-gen AI, shines with exam-topping abilities, context mastery, and ethical safeguards.

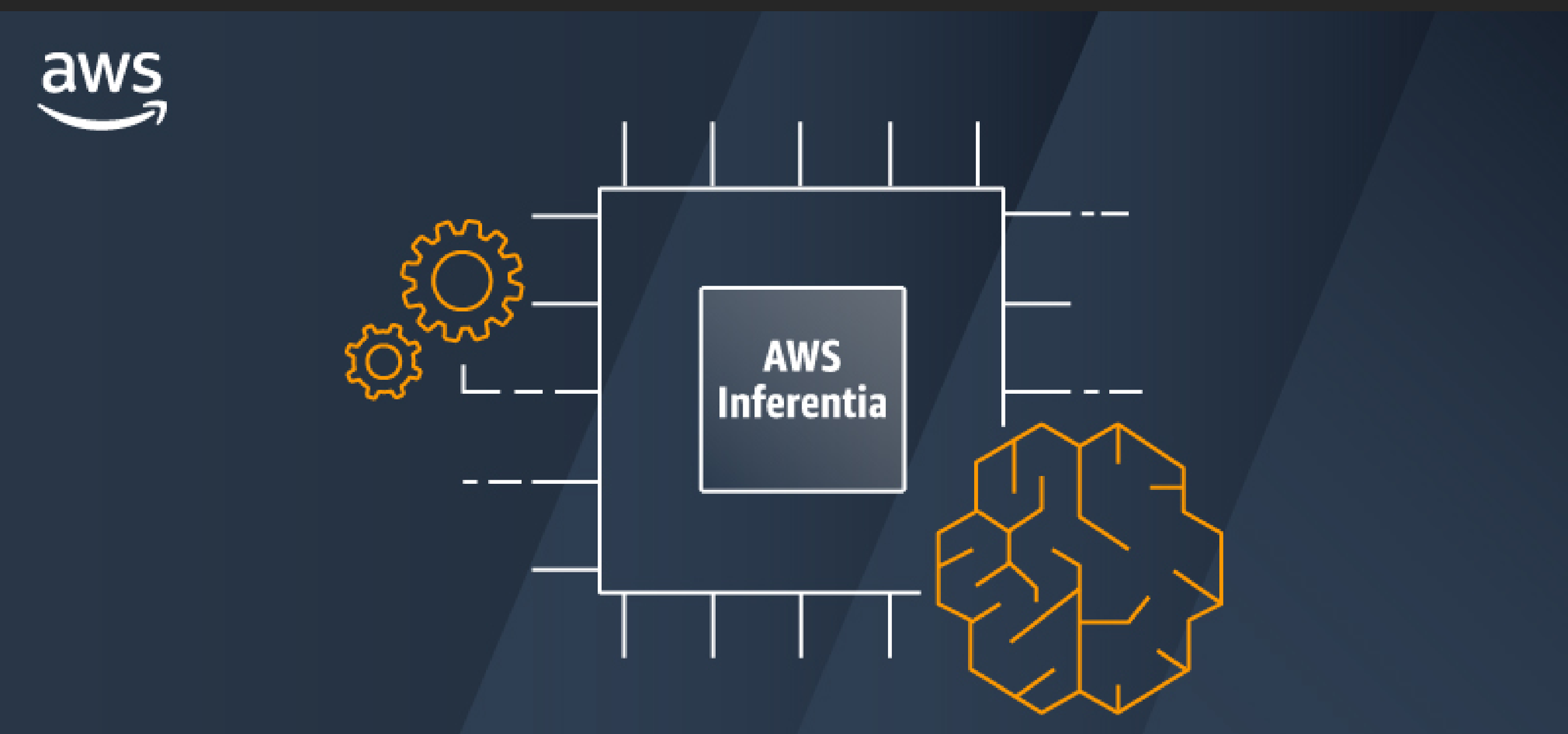
Its expansive knowledge and self-learning potential promise a future of intelligent virtual assistants, but responsible development remains paramount.

Accessible via AWS Bedrock, Claude 2 raises the AI bar while emphasizing ethical leadership.

Amazon announces custom AI Chips

Amazon is developing custom microchips, Inferentia and Trainium, for generative AI, offering an alternative to Nvidia GPUs.

While Amazon has entered the market with its Titan models and Bedrock service, it is seen as playing catch-up to competitors like Microsoft and Google. However, Amazon's custom silicon could provide a competitive edge due to its technical capabilities.



IBM announced watsonx enterprise platform

IBM's new generation of Watson is here, watsonx superpowers businesses with Generative AI tools: build models, manage data, control bias.

Think AI studio, data hub, and ethical guard, all in one. Look for it infused in software like code assistants and IT operations soon.

It provides an open ecosystem of models and enterprise AI capabilities.

OpenAI introduced new DALL·E 3

OpenAI released DALL-E 3, the next iteration of its text-to-image AI, featuring ChatGPT integration to help users generate prompts and additional safety measures.

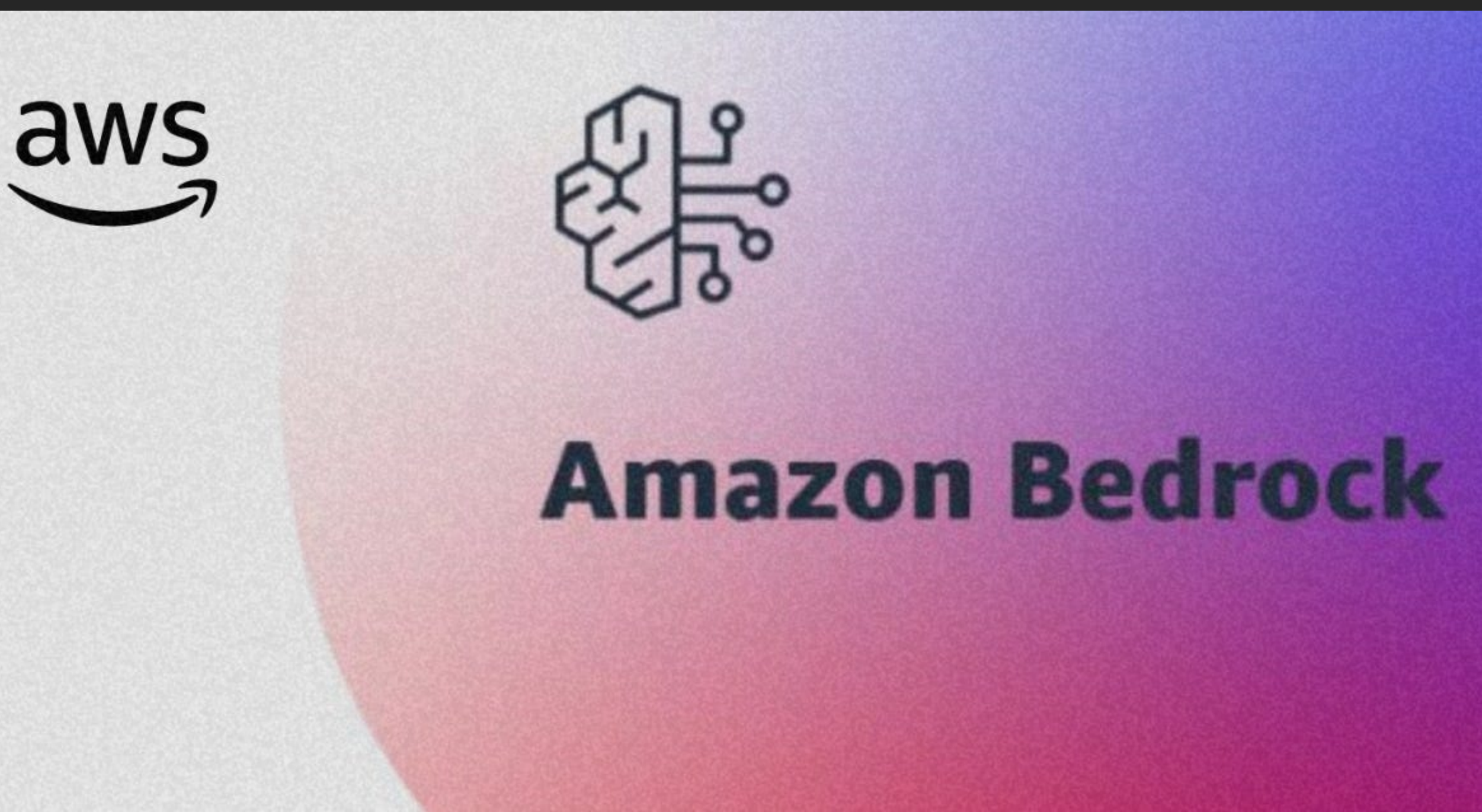
Creators can submit an image that they own rights to and request its removal in a form, a move to avoid potential lawsuits.

Amazon released Bedrock on AWS

The Bedrock service comes with foundational models and tools, including models from Cohere, Anthropic, and Stability AI.

AWS has also introduced Bedrock Console, which enables developers to create virtual AI agents.

It also includes implementing safeguards customized to your application requirements and responsible AI policies.



Elon Musks’s unveiled AI chatbot Grok

Social media platform X, formerly Twitter and now owned by Elon Musk, has launched its A.I. chatbot Grok for U.S. premium users.

Grok, inspired by "The Hitchhiker's Guide to the Galaxy," provides witty answers and real-time data access.

It aims to assist users in accessing information and generating new ideas.

Google announced Gemini, a Multimodal LLM

Google launched its largest AI model, Gemini, with multi-modal capabilities for text, images, video, audio, and code.

Gemini achieved a 90.0% score in massive multitask language understanding (MMLU), outperforming GPT-4. It's designed to understand complex topics and will be integrated into products like Bard, Google Assistant, Search, and Pixel smartphones.

AMD announced new GPUs to compete with Nvidia

AMD has unveiled the Instinct MI300X, a data center GPU to capture market share from Nvidia in the AI chip market.

AMD is looking to capitalize on Nvidia's supply chain constraints and has great benchmarks.

It will find applications in cloud providers and enterprises and has garnered support from players like Meta, Microsoft Azure, and Oracle Cloud.

Pika Launches impressive AI video app

Former Stanford Ph.D. students, founded Pika, an AI video generator simplifying video creation through text input.

The platform has attracted $55 million in funding.

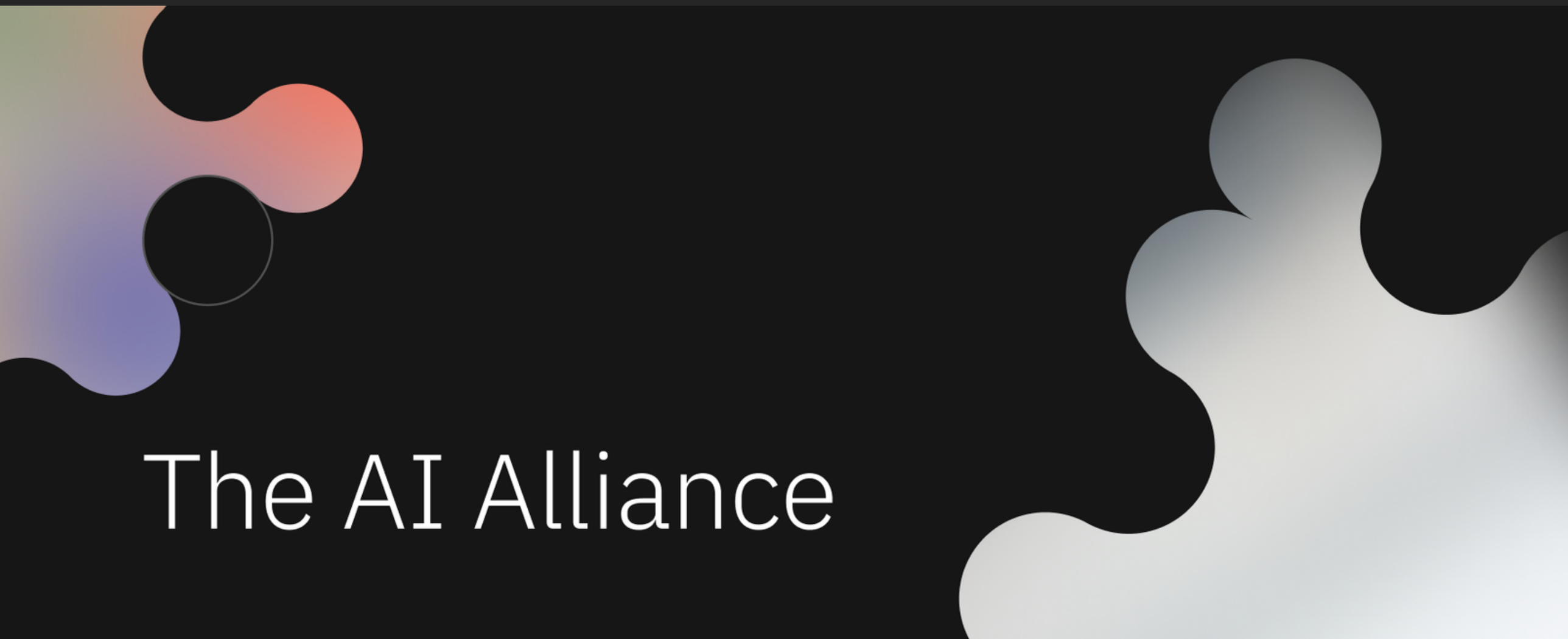
Pika aims to streamline video-making for everyday users, refine its AI model, and develop algorithms for filtering copyrighted content.

IBM and Meta formed the AI Alliance

Meta and IBM have formed the AI Alliance, comprising 57 organizations from various sectors, to promote open-source AI development.

Founding members include major tech companies, startups, nonprofits, and public institutions.

The alliance aims to develop open foundation models, provide free benchmarks, and advocate for responsible AI system development to foster openness and collaboration.

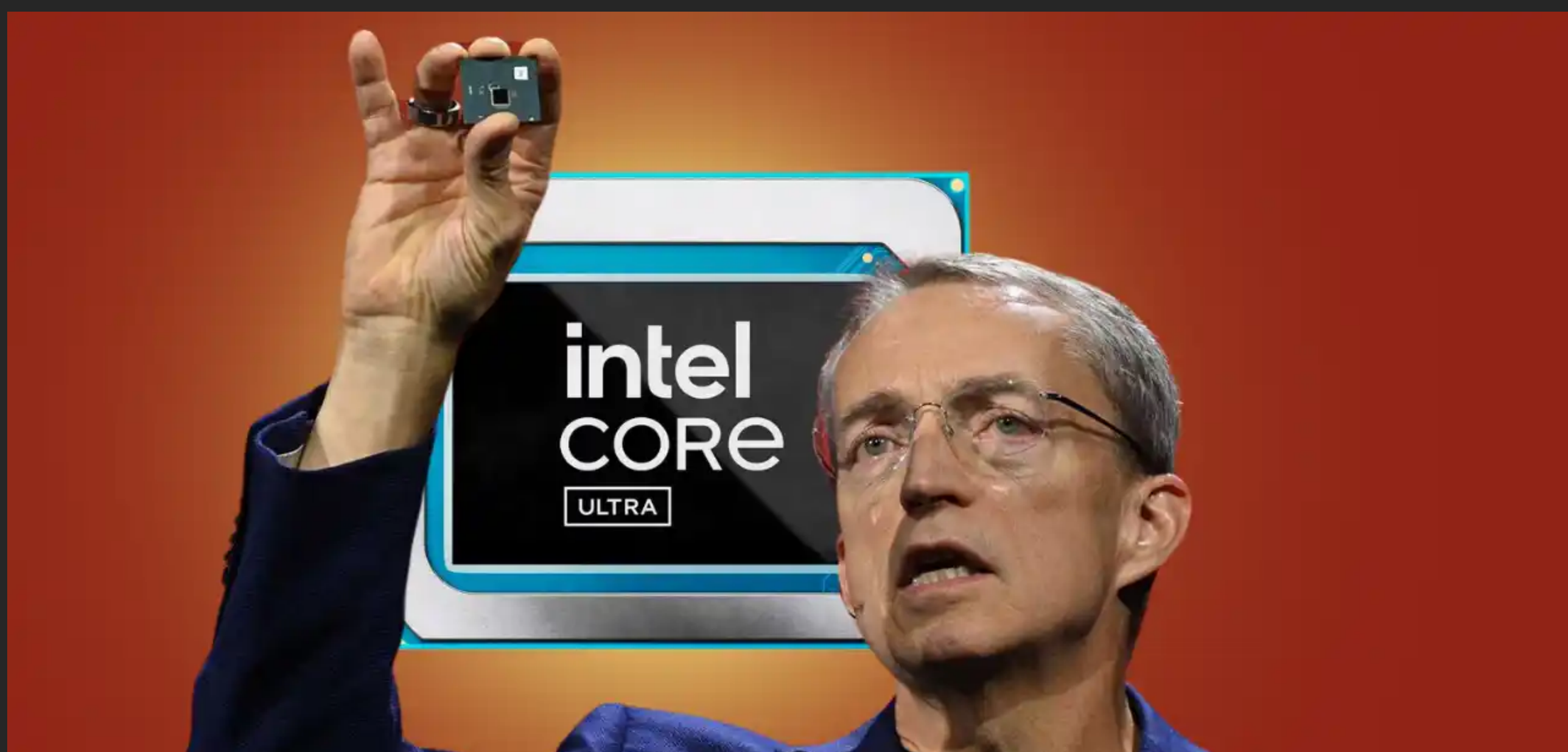


Intel AI Everywhere announced new chips

Intel introduced a range of AI products at its "AI Everywhere" event.

These include Gaudi3, a chip designed for GenAI, Core Ultra chips for Windows laptops and PCs with built-in AI capabilities, and the 5th-generation Xeon server chips.

Intel is collaborating with PC manufacturers to promote machine upgrades, emphasizing the new era of chatbots and AI-powered computing.

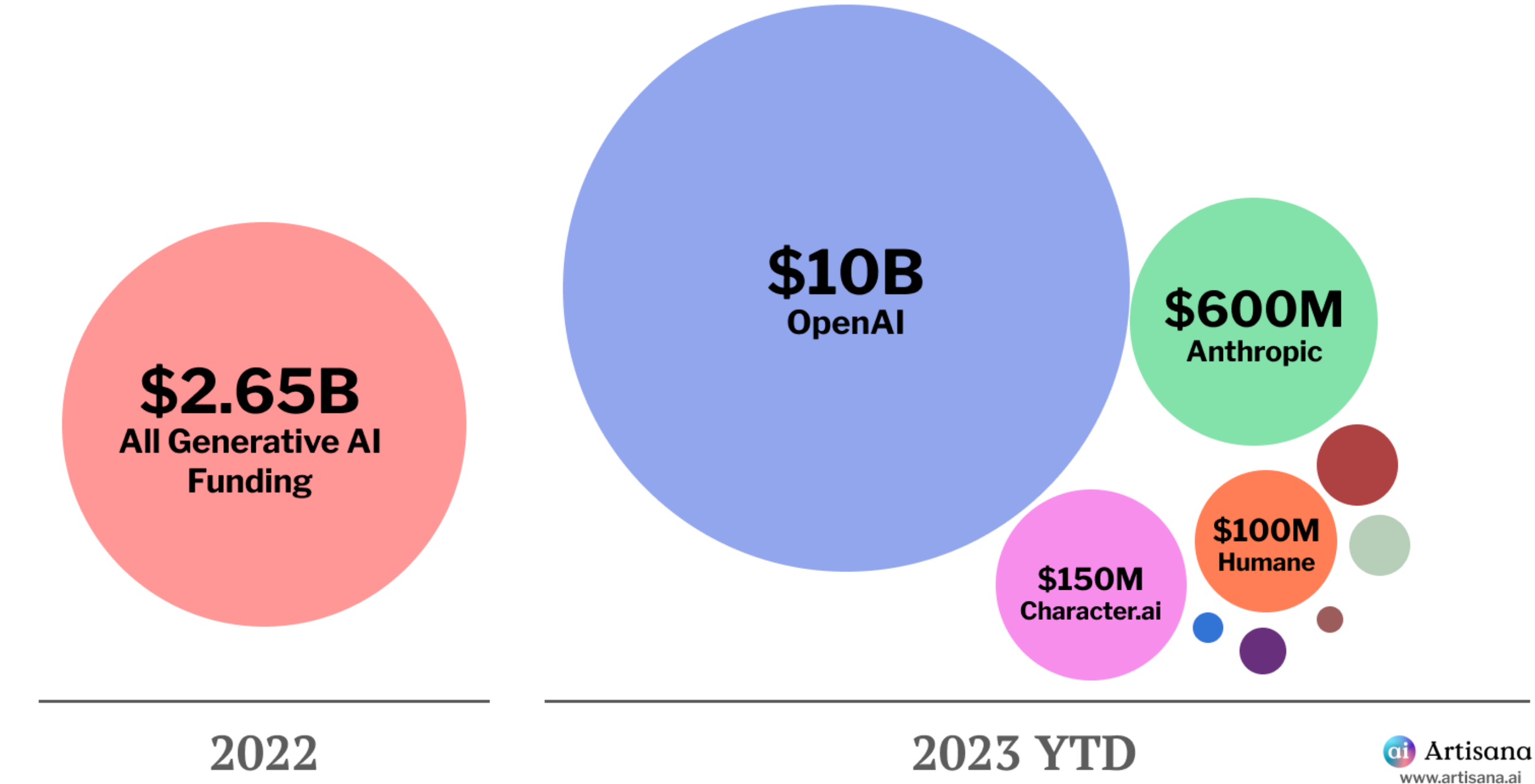


AI investment is booming. How much is hype?

Billions invested in generative AI this year raise bubble concerns.

Though some see hype driving rash bets, others remain convinced of transformative long-term potential across sectors.

Distinguishing momentary manias from truly groundbreaking innovations remains an open debate.



AI drove the stock market to All-Time Highs

The recent stock market rally, with record highs in the S&P 500 and Dow Jones Industrial Average, to excitement surrounding generative AI.

The increased productivity from AI will contribute to the market's long-term strength, comparing it to the early stages of previous bull markets. Despite concerns about a speculative bubble, there are strong reasons for optimism.

“AI WILL NOT DESTROY THE WORLD, AND IN FACT MAY SAVE IT.”

Marc Andreessen (Managing Partner and Co-founder a16z)

Thanks for reading.

Happy 2024.