Characterizing Power Management Opportunities for LLMs in the Cloud

Pratyush Patel

Esha Choukse Chaojie Zhang Íñigo Goiri Brijesh Warrier
Nithish Mahalingam Ricardo Bianchini

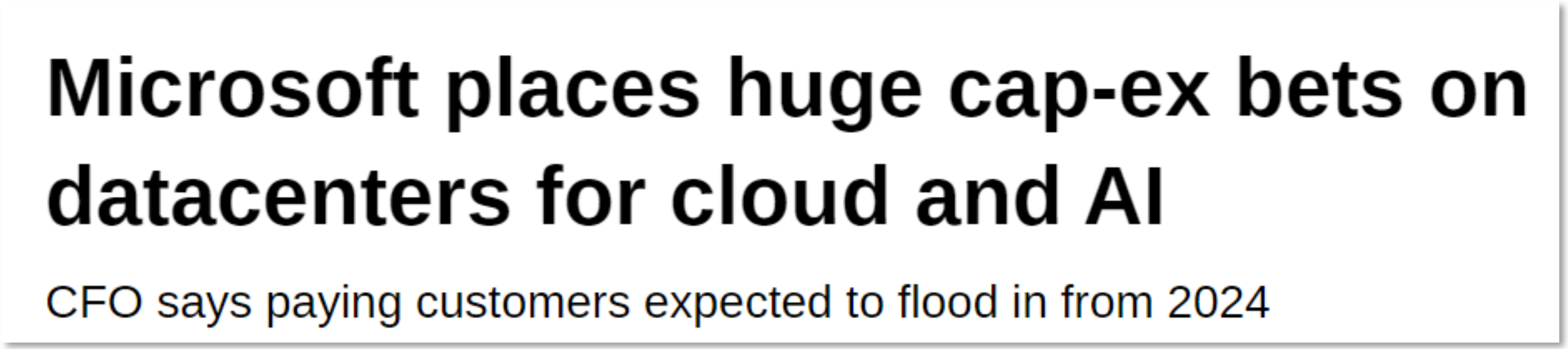
ChatGPT

Gemini

GitHub
Copilot

🔥 ChatGPT Gemini GitHub Copilot 🔥

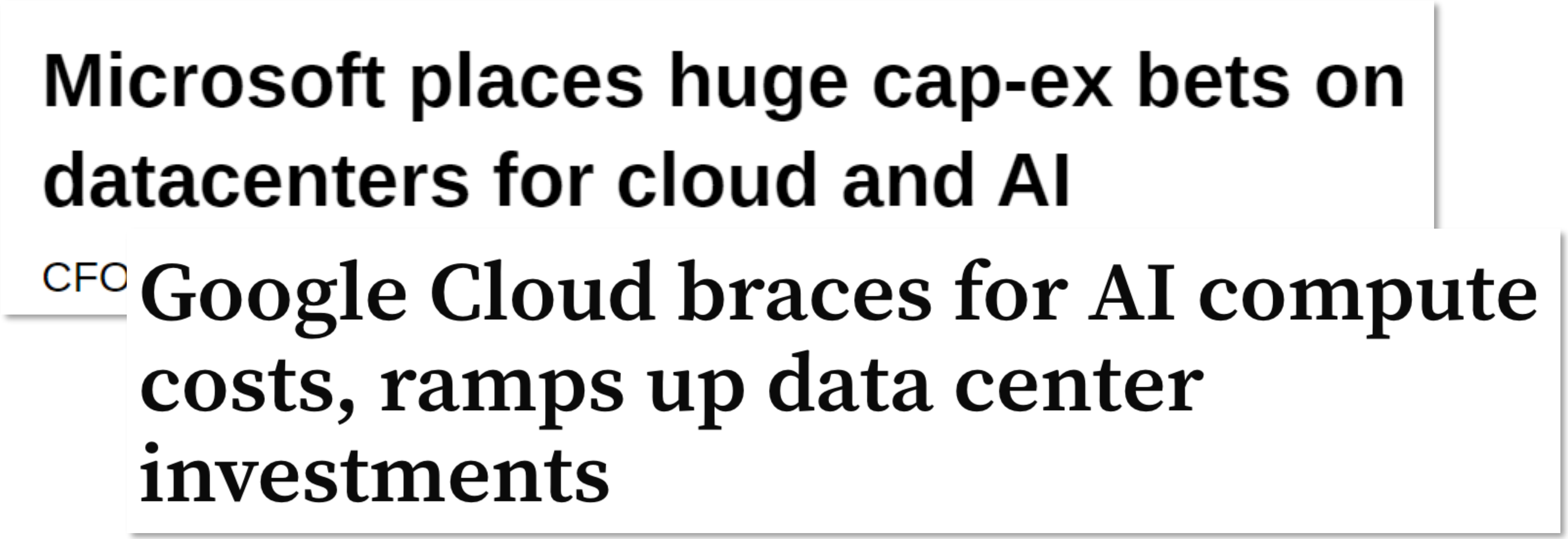
🔥 ChatGPT Gemini GitHub Copilot 🔥



Microsoft places huge cap-ex bets on datacenters for cloud and AI

CFO says paying customers expected to flood in from 2024

🔥 ChatGPT Gemini GitHub Copilot 🔥

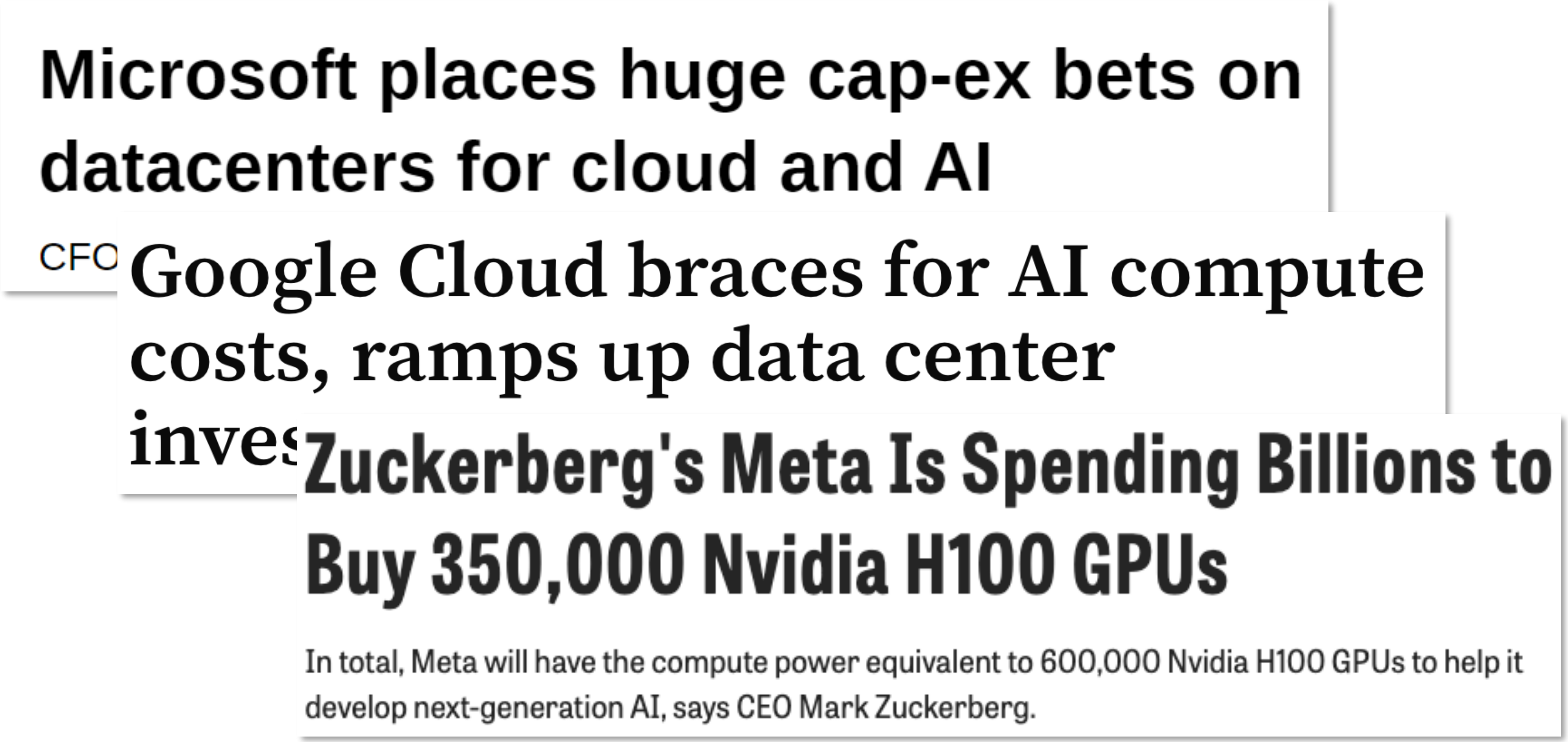


Microsoft places huge cap-ex bets on datacenters for cloud and AI

CFO

Google Cloud braces for AI compute costs, ramps up data center investments

🔥 ChatGPT Gemini GitHub Copilot 🔥



Microsoft places huge cap-ex bets on datacenters for cloud and AI

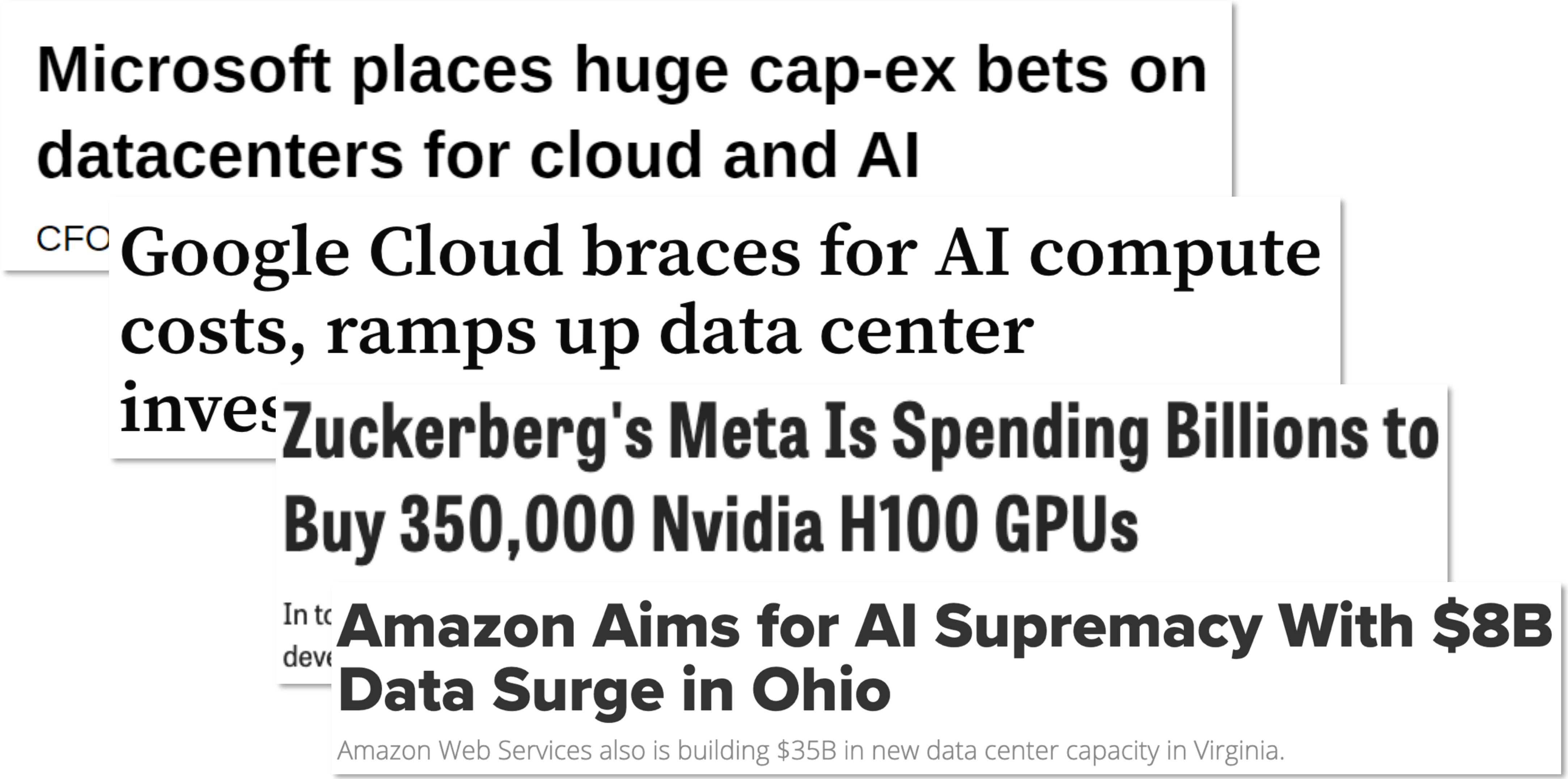
CFO

Google Cloud braces for AI compute costs, ramps up data center inves

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.

🔥 ChatGPT Gemini GitHub Copilot 🔥



Microsoft places huge cap-ex bets on datacenters for cloud and AI

CFO

Google Cloud braces for AI compute costs, ramps up data center inves

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In to

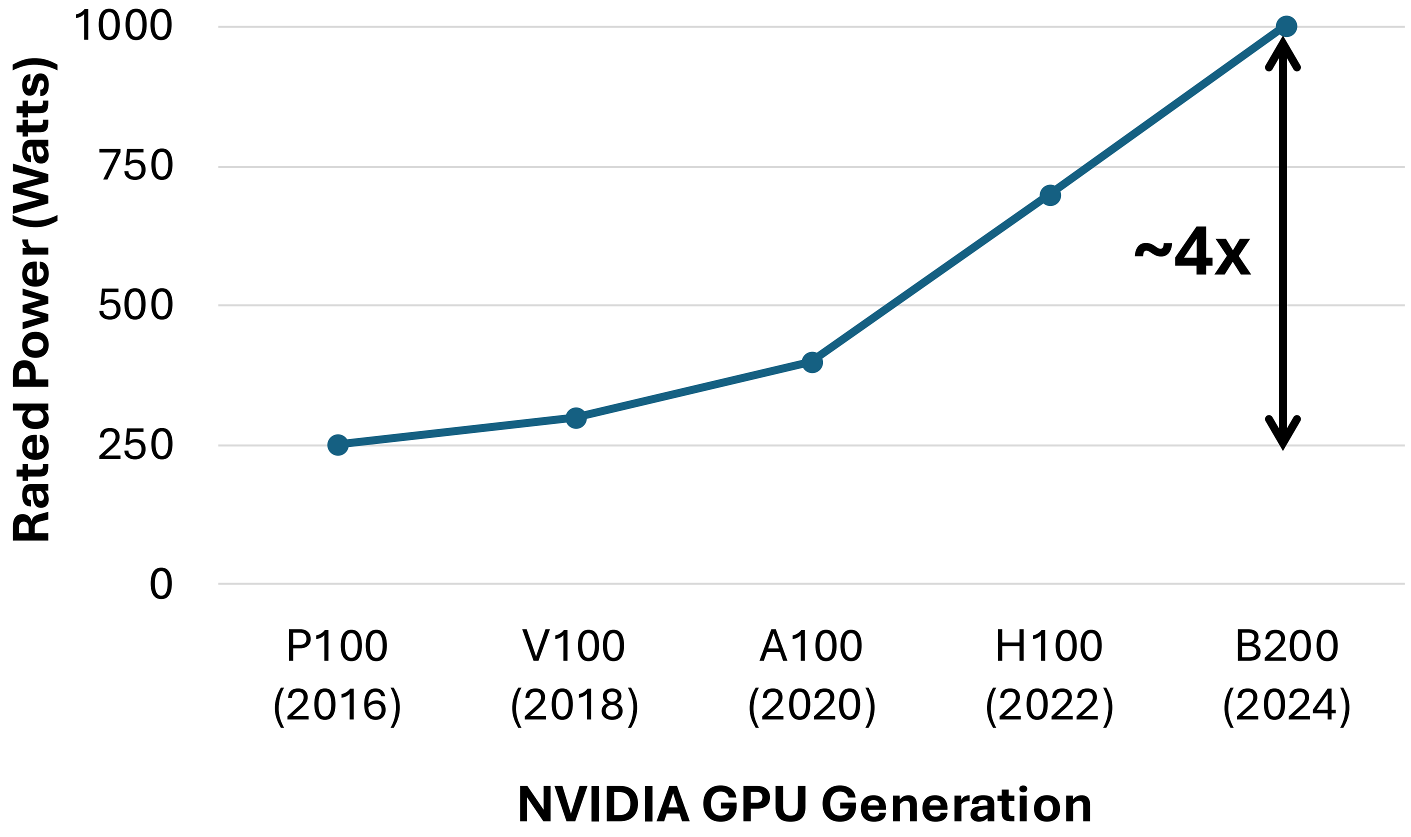
deve

Amazon Aims for AI Supremacy With $8B Data Surge in Ohio

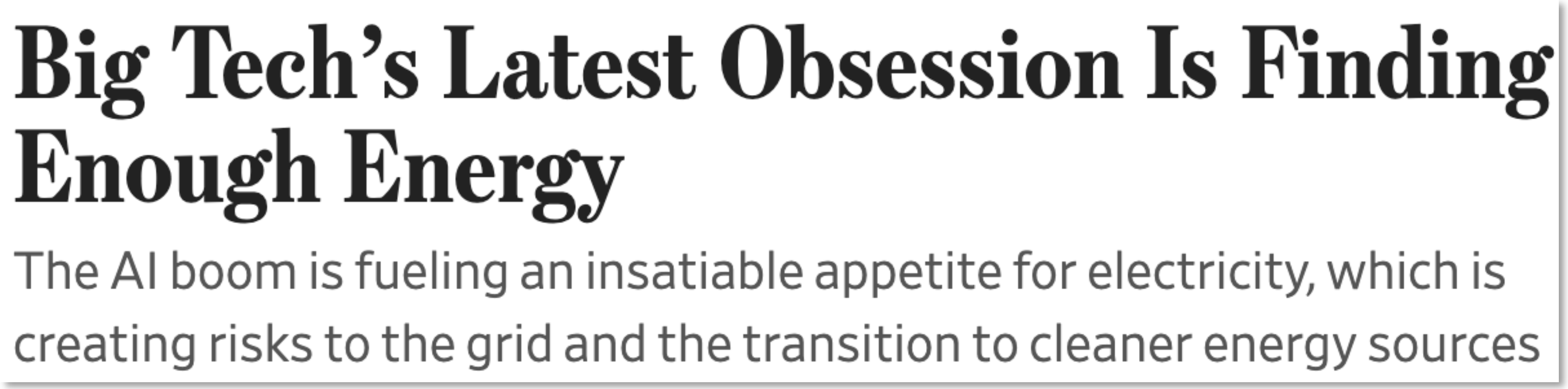
Amazon Web Services also is building $35B in new data center capacity in Virginia.

🔥 ChatGPT Gemini GitHub Copilot 🔥

GPU clusters for LLMs are incredibly power hungry



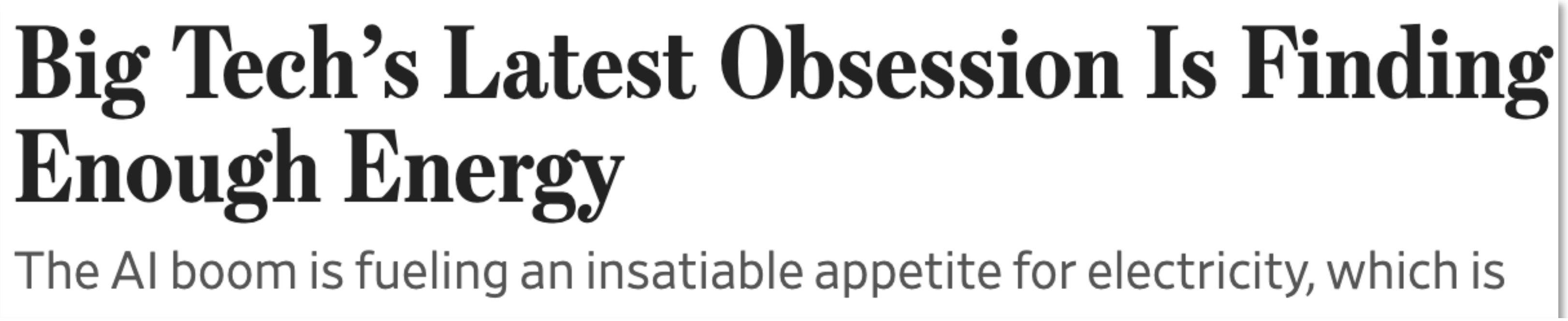
🔥 ChatGPT Gemini GitHub Copilot 🔥



Big Tech's Latest Obsession Is Finding Enough Energy

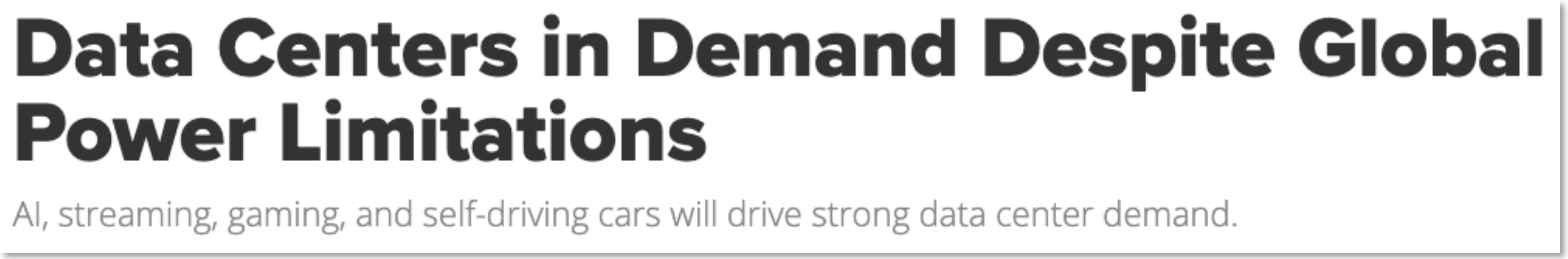
The AI boom is fueling an insatiable appetite for electricity, which is creating risks to the grid and the transition to cleaner energy sources

🔥 ChatGPT Gemini GitHub Copilot 🔥



Big Tech's Latest Obsession Is Finding Enough Energy

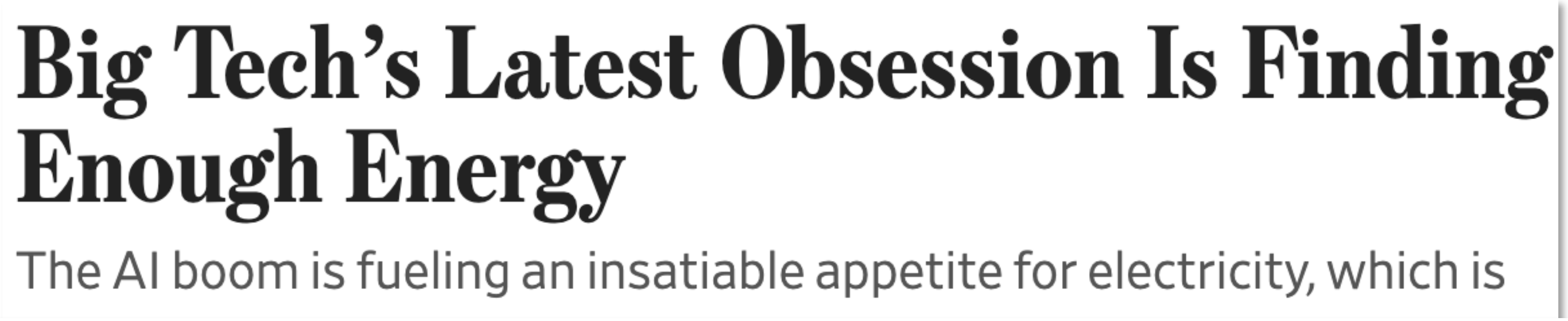
The AI boom is fueling an insatiable appetite for electricity, which is



Data Centers in Demand Despite Global Power Limitations

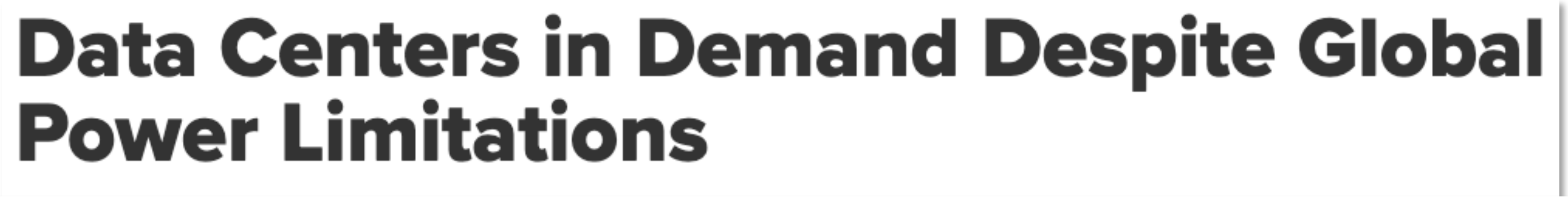
AI, streaming, gaming, and self-driving cars will drive strong data center demand.

🔥 ChatGPT Gemini GitHub Copilot 🔥

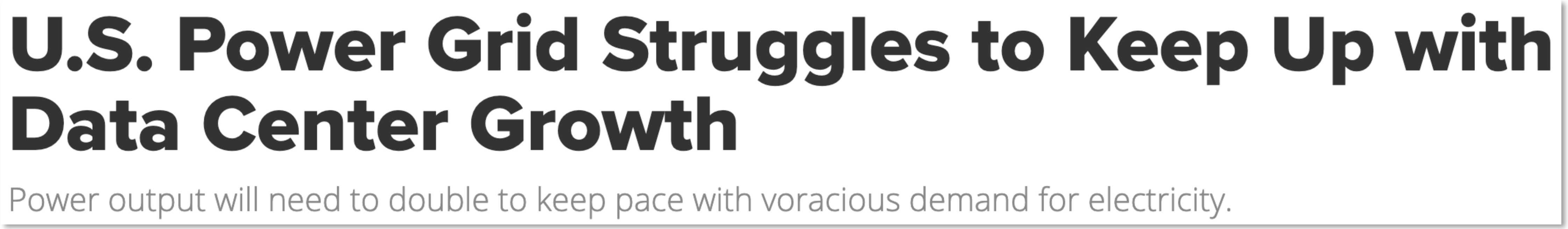


Big Tech's Latest Obsession Is Finding Enough Energy

The AI boom is fueling an insatiable appetite for electricity, which is



Data Centers in Demand Despite Global Power Limitations



U.S. Power Grid Struggles to Keep Up with Data Center Growth

Power output will need to double to keep pace with voracious demand for electricity.

Our work analyzes the power usage and helps alleviate the power wall for LLM deployments in the cloud

Characterizing Power Management Opportunities for LLMs in the Cloud

Profile power usage patterns of training and inference workloads in production clusters

Analyze design implications for power management in cloud deployments

Build a power oversubscription framework that safely adds ~30% more servers in inference clouds

aka.ms/LLMPower

Thanks!

Talk on Tuesday at 10am (Session 4C)

pratyush@cs.uw.edu