Big Tech Capex Day, DeepSeek's Open DeepInfer Engine and Meta's Open-Weight Muse Spark

Alphabet and Microsoft report earnings as AI-spending scrutiny peaks, DeepSeek open-sources the inference engine behind V4.1-Flash, and Meta ships an 816B open-weight Muse Spark with strings attached.

Themes: Infrastructure & Economics · Models & Frontier

Note: coverage today is thin and mostly secondary (search summaries and aggregators). I could not open primary filings or earnings releases, and dates and figures differ between outlets; vendor benchmark claims are unverified.

1) Big Tech Earnings Week: The AI Capex Question

Schwab's market update reports AI-related stocks built on their rally early on October 29 while investors awaited the Fed decision; Alphabet and Microsoft report after the close, and Oracle holds an investor meeting focused on AI spending. Earlier this year Alphabet guided capex to about $195-205 billion for the year and its stock fell after the July report, while Oracle sold off after its June results over its financing needs. Figures are from secondary coverage and may use different periods. Results from today's reports were not available at time of writing.

The concept, explained simply: Capex (capital expenditure) is money spent on long-lived assets, here data centres, chips and power. Free cash flow is what remains after paying for that. When capex rises faster than the revenue it produces, free cash flow shrinks.

Why it matters: Investors now judge AI companies less on model quality and more on whether spending turns into revenue. Day 214 covered the credit side of this story; today's earnings show the equity side. Both shape how fast compute capacity can grow and what it costs to rent.

2) DeepSeek Open-Sources DeepInfer, Its Production Inference Engine

Reports say DeepSeek released DeepInfer as open source around October 20-21, describing it as the framework that serves DeepSeek-V4.1-Flash at scale in production, alongside a companion component (deepseek-recipe) for token encoding and decoding. One report says the README lists Huawei Ascend 950 NPUs as well as Nvidia SM100 GPUs as supported hardware. Reported licence and performance details were not verified.

The concept, explained simply: Inference is running a trained model to answer requests. An inference engine is the software that schedules requests, manages memory and keeps expensive chips busy. Better engines mean more answers per chip, and so a lower cost per token.

Why it matters: Training gets the headlines but inference is the recurring bill. Publishing a production-grade engine lets others copy cost-saving techniques and run on non-Nvidia hardware, which pressures prices across the industry.

3) Meta Muse Spark Open: Open Weights, Semi-Open Licence

Meta released Muse Spark Open on October 15: reportedly 816B total parameters with 42B active and a 1M-token context window, with weights on Hugging Face and hosted on SageMaker, Google Cloud, Microsoft Foundry, OpenRouter and Together AI. Meta calls it the most capable US open-weight model. Outlets describe the licence as not fully open, and one report puts it seventh among open-weight models on Artificial Analysis (score 39), behind Chinese models such as Kimi K3 and GLM 5.3 at about 60. Day 214 covered its closed siblings, Muse Wonder and Muse Wonder Full.

The concept, explained simply: Open weights means you can download the trained model and run it yourself, but that is not the same as open source. The licence decides what you may do with it, such as commercial use or redistribution, so read it before building on a model.

Why it matters: A US lab returning to open weights gives enterprises a self-hostable option with a Western provenance, but the licence limits and the ranking gap show that 'open' and 'best' are different questions.

VIRAL / OPEN-SOURCE SPOTLIGHT

DeepInfer (DeepSeek): the inference engine behind a price war. DeepInfer is open-source serving software. It is the real production code DeepSeek uses to run V4.1-Flash, rather than a demo. It is trending because inference cost is now the main competitive lever and DeepSeek has a history of using low prices to pressure rivals (in July-August OpenAI cut GPT-5.6 prices and DeepSeek answered within hours, per secondary reports). Caveats: I found no verified GitHub star

counts, so this is based on press coverage, and hardware support claims such as Ascend 950 come from one report.

MARKET SIGNAL

Two forces pull in opposite directions: hyperscalers are spending close to $200 billion a year each on capacity while open releases from DeepSeek, Meta and Moonshot push the price of using models down. Watch whether revenue growth justifies capex, and whether cheaper inference compresses margins for API sellers.

PRACTICAL TAKEAWAYS

1) Read the capex and free-cash-flow lines, not just the AI headline.

When you judge an AI vendor's staying power, check whether spending is funded by operating cash flow or by debt.

2) Benchmark inference cost, not only model quality.

Measure cost per successful task on your own workload, since serving efficiency can matter as much as the model.

3) Read the licence before adopting an open-weight model.

Open weights can still restrict commercial use, so have legal check Muse Spark Open's terms before production use.

4) Keep hardware options open.

Engines that support several chip vendors reduce lock-in, so prefer portable serving stacks where you can.

SELF-TEST

Q1. What is capex, and why do investors watch it?

A1. Spending on long-lived assets such as data centres and chips; if it outruns revenue, free cash flow shrinks.

Q2. What does an inference engine do?

A2. It schedules requests and manages memory so a model serves answers cheaply on its hardware.

Q3. Is open weights the same as open source?

A3. No; the weights can be downloadable while the licence still restricts how you use them.

Q4. Why does DeepInfer matter to prices?

A4. It shares production serving techniques and supports more hardware, lowering the cost of running models.

TOMORROW

What We're Watching: what Alphabet and Microsoft say about capex guidance and cloud growth. What We're Watching: the public text of Anthropic's S-1, which reports suggest could be filed soon. What We're Watching: Kimi K3.1 Pro's licence and weights ahead of November 4. What We're Watching: independent benchmarks of DeepInfer's efficiency claims.

TOPICS COVERED SO FAR (RECAP)

Recent days covered Claude Opus 5.5, Nvidia's Vera Rubin POD, Nemotron 3.5, GPT-6.1 Astra, Gemini 4 Argon, Claude Duet, Kimi K3.1 Pro, Meta's Muse Wonder, AI credit stress and CDS, Anthropic's containment disclosures and agent scope creep. Today added Big Tech capex scrutiny, DeepSeek's DeepInfer and Meta's Muse Spark Open.

Sources: Schwab market update (schwab.com/learn/story/stock-market-update-open), SBS News morning wrap (news.sbs.co.kr), AI Business on DeepInfer (aibusiness.com), fourweekmba.com, tech-insider.org, aiweekly.co and thenewstack.io on Muse Spark Open, officechai.com, Motley Fool and Webull earnings recaps (fool.com, webull.com), letsdatascience.com on Anthropic's S-1.