LUMO INSIGHTS

Wednesday, 27 May 2026

# Faster Models, Trillion-Dollar Moves, and an 80-Year Problem Solved

## Executive Summary

In the past week, AI's capability frontier leapt forward on multiple fronts. Google launched a high-speed Gemini model that breaks performance-speed tradeoffs, an OpenAI system cracked a math problem unsolved for 80 years, and industry leaders hit trillion-dollar milestones. Each development carries strategic implications for enterprises preparing for the next wave of AI-driven change.

## Faster Models at Scale: Speed Over Size

The foundation model capability frontier advanced significantly this week with the launch of new, high-speed AI models designed for real-world deployment. At Google's I/O conference, the company unveiled Gemini 3.5 Flash, a state-of-the-art language model that delivers frontier-level performance at unprecedented speed. Crucially, it can handle up to one million tokens of context — meaning it can process entire books or extensive datasets in a single query, enabling analysis and operations at scales previously impossible.

Google's strategy with Gemini 3.5 Flash is telling: rather than simply chasing the biggest model, the focus is on rapid, broad deployment. CEO Sundar Pichai has stressed the importance of, in his words, "stay at the frontier, but prioritize models cheap and fast enough to deploy across products used by billions." In practice, that means building AI not just for benchmarks but for immediate integration into widely used services. Indeed, 3.5 Flash is already accessible via Google's search engine, mobile apps, and cloud platform. Over one billion users now have the option to use generative AI in everyday search. This marks a strategic shift where distribution and speed become key competitive advantages: Google is leveraging its vast user base and infrastructure to weave advanced AI into routine workflows.

For enterprise leaders, the emergence of ultrafast models like Gemini 3.5 signals an inflection point in AI tool adoption. These models drastically reduce latency and cost barriers. For example, 3.5 Flash's usage pricing is around 25% lower than the previous generation and notably cheaper per query than rival offerings. By making high-caliber AI responses nearly instantaneous and more affordable, Google (and others following suit) are removing friction for business use cases – from real-time customer service bots to on-the-fly data analysis. Slow, expensive model runs have often limited how deeply organizations integrate AI into their operations. That obstacle is rapidly shrinking; tasks that used to take specialists days can now be handled by AI in hours or minutes, allowing companies to scale AI-driven processes without exorbitant cloud bills.

Notably, this “faster and cheaper” approach doesn’t mean sacrificing capability. Google’s latest model shows that speed and quality can go hand-in-hand. This has been corroborated by other players: for instance, Alibaba’s newly released Qwen 3.7 Max (announced this week at a tech summit) reportedly matched leading Western models like Claude on certain complex coding benchmarks while also offering a one-million-token context and competitive pricing. The takeaway: the cutting edge in AI is no longer just about building bigger models—it’s about making advanced AI accessible at scale. In the coming months, expect the competition to increasingly center on how quickly and widely these tools can be deployed, which in turn will accelerate their impact on every industry.

## Breakthrough Reasoning: AI Solves the Unsolvable

Last week also delivered a striking proof of how advanced AI reasoning has become: an OpenAI-developed AI system solved a mathematical problem that had stumped humans for 80 years. The model autonomously cracked the Erd  2 “unit distance” geometry conjecture, a famous problem first posed in 1946. This marks the first time an AI has independently discovered a proof to a major unsolved math challenge – an achievement verified by outside mathematicians. The AI wasn’t a special-purpose theorem prover, but a general-purpose next-generation language model being tested on open research questions. In tackling a problem that eluded experts for decades, the AI introduced a novel proof strategy by bringing in ideas from a different branch of mathematics that no human had tried. One Fields Medal-winning mathematician has hailed the result as a 'milestone in AI mathematics.'

Leading researchers say this breakthrough shows that modern AI systems can achieve levels of deep, creative reasoning beyond what we’ve seen before. One expert noted these models are now producing 'original, ingenious ideas' and carrying them out to completion – not just acting as helpers to human experts. Importantly, the AI’s success suggests it can connect knowledge across domains and handle very long, complex chains of logic. Such capabilities have clear implications beyond theoretical math: similar AI reasoning engines could be applied to intractable problems in fields like drug discovery, material science, or logistics. Essentially, the frontier of AI intelligence is moving from pattern recognition into true problem-solving.

For business leaders, this is a wake-up call that AI is evolving from a support tool into an autonomous innovation engine. Organizations that tap into these advanced reasoning models could find solutions to problems previously thought “unsolvable” in their industries – whether optimizing a supply chain in novel ways or devising new product strategies. The competitive gap may soon arise not just from who has AI, but who has AI smart enough to develop new breakthroughs. Enterprises should begin partnering with AI research initiatives or investing in advanced model capabilities aimed at their toughest challenges. The coming wave of AI won’t just answer questions faster; it may well start asking – and answering – the questions no one realized we should be tackling.

## Trillion-Dollar Economics of AI

A second theme of the week was the unprecedented scale of investment and revenue in the AI sector – underscoring that frontier AI is not just a technological phenomenon but an economic one. Anthropic, the startup behind the Claude AI models, was reported to be closing a new $30 billion funding round that would value it above $900 billion – suddenly making it the world’s most valuable privately-held AI company, ahead of even OpenAI. If finalized, this valuation leap (from around $380B just a few months ago) reflects investors’ confidence in the company’s growth – confidence reinforced

by Anthropic's revelation that it has achieved its first-ever quarterly profit on the strength of surging enterprise adoption. In fact, Anthropic projects about $10.9 billion in Q2 2026 revenue, up 130% from last quarter's $4.8B. Investors now foresee a path for the firm to reach over $50B in annual revenue within the next 18 months. Frontier AI providers are rapidly becoming as financially significant as the cloud or social media giants of the last decade.

Not to be outdone, OpenAI is accelerating toward the public markets. The ChatGPT creator confirmed it filed a confidential IPO prospectus on May 22, targeting a valuation as high as $1 trillion when it lists on the stock market later this year. For perspective, only a handful of tech companies in history have approached that threshold. Going public will bring capital for even larger AI investments and, for the first time, public transparency into OpenAI's economics – shedding light on the true costs (and profits) of cutting-edge AI. It's also a signal that these companies now feel ready to pitch themselves as stable long-term businesses, not just R&D projects burning cash. Despite heavy infrastructure spending, OpenAI is reportedly already generating around $2 billion per month in revenue (a $25B annual run-rate). The expected IPO – alongside Anthropic's massive fundraise – indicates that the race to build and monetize the most advanced AI is entering a new, more competitive phase.

The vast resources at play are also reshaping the economics of AI infrastructure. Training and running frontier models requires enormous computing power and energy. This week we learned that Anthropic is paying SpaceX about $1.25 billion every month for access to a dedicated supercomputing cluster (as revealed in SpaceX's own IPO filings) – a $45 billion commitment over several years. Meanwhile, in the energy sector, NextEra Energy's $67B deal to acquire Dominion Energy was driven largely by the need to power future AI data centers. When power companies make strategic moves to support AI workloads, it underscores that AI's growth is now constrained by physical resources as much as algorithms. For enterprises, this dynamic means the cost and availability of AI compute will remain a key strategic factor – and those with strong cloud or hardware partnerships may have an edge in securing scarce processing capacity.

Overall, the flurry of trillion-dollar valuations and industry-spanning deals sends a clear message: AI at the frontier is turning into one of the biggest economic forces of our time. For business leaders, it's a twofold signal. First, AI capabilities are becoming core to competitive advantage, driving real revenue – if your organization isn't already leveraging these tools, your competitors likely are. Second, accessing top-tier AI may increasingly depend on aligning with a major platform or ecosystem, whether through partnerships or strategic use of their cloud services. Companies need to factor the cost of AI (from talent to compute to power) into their long-range plans. The upside is that intense competition among these well-funded AI providers could drive down user costs and spur faster innovation – but it will also likely determine which platforms dominate enterprise AI in the years ahead. Choosing the right AI partners (and possibly hedging across multiple vendors to avoid lock-in) will be a critical strategic decision going forward.

## Global Competition and Open Models

The open vs. closed AI debate also saw new developments this week, intertwined with intensifying global competition. On the open-source front, smaller AI players and research groups are making notable strides by releasing advanced models for anyone to use. For example, Cohere unveiled a 218-billion-parameter model under an open license, Europe's Mistral AI open-sourced a new 128B model, and xAI (Elon Musk's startup) rolled out significant upgrades to its Grok model – including a 1-million-token context and plug-in 'skills' for custom tasks. Open models even achieved record-breaking feats, such as a prototype demonstrating a 12-million-token context window (the largest ever reported). For

enterprises, these open-source advances promise more control and customization, allowing organizations to self-host powerful AI and keep sensitive data in-house. However, they still trail the absolute cutting-edge closed models in certain areas, such as the most challenging reasoning tasks or highly specialized capabilities.

A striking shift is the role of China in the AI landscape. Chinese tech companies and research labs are now leading in the open-model arena as some Western open efforts have slowed. On one popular platform that routes queries to various AI models, over 60% of usage is now going to Chinese-developed systems (names like Zhipu's GLM and Alibaba's Qwen). This reflects a deliberate strategy: Chinese institutions are heavily investing in advanced AI and often releasing models (or providing broad access) to gain global traction. At the same time, China's major players are fielding some of the top proprietary models; Alibaba's closed-source Qwen 3.7 Max has matched leading Western models on key benchmarks and offers competitive pricing. Such progress comes as the Chinese government tightens oversight of AI talent – reportedly even restricting top researchers at firms like Alibaba and DeepSeek from traveling abroad without approval. AI has become a matter of national strategic importance, and China's efforts are clearly elevating its position in the global AI race.

For business strategists, the open vs. closed question is becoming more complex in light of these trends. On one hand, the growing capabilities of open models offer a tempting path to reduce dependency on the big US AI providers. Companies with the means to deploy models in-house might leverage open-source AI to avoid vendor lock-in and retain control over their data. On the other hand, the best-of-breed models from firms like OpenAI, Anthropic – and now Google and Alibaba – still generally outpace open models in raw performance, especially for nuanced tasks. Enterprises will need to weigh these trade-offs: is absolute top performance worth the loss of some control and the higher cost of a proprietary service? In many cases a hybrid strategy may make sense – using closed APIs for what they do best, while integrating open models where customization and privacy are paramount. Additionally, the geopolitics of AI can't be ignored. If key open-source advancements are coming from China, organizations must consider regulatory and security factors when adopting those models. The bottom line: the AI ecosystem is diversifying rapidly, and staying competitive will require navigating a more international and heterogeneous set of AI options.

## The Next 18 Months: Strategic Imperatives

This week's developments paint a clear picture of where enterprise AI is heading in the next 6–18 months, and they highlight several strategic priorities for business leaders. **First**, expect AI capabilities to become pervasive in the tools and platforms you already use. Google's integration of Gemini 3.5 Flash into Search – along with Microsoft's ongoing rollout of Copilot across Office – means your employees and customers will increasingly have advanced AI assistance at their fingertips by default. Beyond productivity apps, industry-specific AI solutions are emerging: both OpenAI and Anthropic recently launched targeted offerings (for cybersecurity and for finance, respectively), essentially aiming to become the 'operating system' for Wall Street and for enterprise security. Companies should actively engage with these developments, whether by piloting new AI-powered features from vendors or collaborating with providers to tailor models to their domain.

**Second**, re-evaluate your talent and organizational strategy in light of AI's rapid progress. Leading tech companies – and now even established enterprises – are already restructuring to become more AI-driven. The pattern is consistent: Microsoft, Salesforce, Google, and Meta (which cut 8,000 roles last week) have all undertaken AI-driven reorganizations. This week, Intuit joined their ranks by announcing it will lay off about 3,000 employees (17% of its workforce) as it doubles down on

integrating AI into its products. These firms aren't eliminating jobs because AI has fully automated those roles; rather, AI is enabling them to achieve the same output with smaller teams.

The implication is clear: gaining an edge may involve reorganizing workflows around AI augmentation. That means investing in training your workforce to use AI tools effectively and hiring for new skill sets – for example, engineers and analysts versed in integrating AI into core processes. The next wave of competitive hiring is likely to focus on those who can meld domain expertise with AI fluency to unlock value. Companies that proactively realign roles and reskill employees to work alongside AI will be better positioned to capture efficiency gains, while those that delay may find themselves with overhead that nimbler, AI-powered competitors do not.

**Third**, keep a close watch on the evolving regulatory and geopolitical backdrop. In the U.S., a plan to impose new AI governance via executive order was abruptly shelved this week after industry lobbying – a signal that, for now, federal oversight will lag the technology's pace. This laissez-faire environment may allow faster rollout of advanced AI solutions in the short term, but it also creates uncertainty: state-level rules or European Union regulations (like the upcoming EU AI Act) could fill the gap. Meanwhile, China's moves to tighten control over AI talent and IP highlight that governments see strategic and security dimensions to AI leadership. For globally operating companies, compliance agility is key. Business strategies should incorporate flexible AI governance policies and ethical guidelines now, so that teams can adapt swiftly to new laws without stalling innovation.

Finally, perhaps the most important imperative is to proactively explore and invest in cutting-edge AI capabilities now. The headline stories of this week – from lightning-fast models to machines solving decades-old problems – all point to one conclusion: the AI capability frontier is advancing at an extraordinary pace. In 18 months, the tools we consider "frontier" today may be commonplace in our competitors' operations. Organizations that experiment early with large-scale models, multimodal AI, and autonomous agents will be better positioned to leverage them as they mature. Craft a clear AI strategy that identifies where these emerging capabilities could transform your business – whether it's supercharging customer experiences, uncovering insights in data, or creating new AI-driven products. With industry leaders pouring billions into innovation and racing to embed AI in every facet of work, the winners of this next phase will be the companies that act decisively to ride the wave of AI-driven change – not those left scrambling to catch up later.

## Key Statistics

- Anthropic projects $10.9 /billion in Q2 2026 revenue (up 130% QoQ), marking its first profitable quarter ([www.buildfastwithai.com](https://www.buildfastwithai.com/blogs/ai-news-today-may-25-2026#:~:text=revealed%20it%20is%20on%20track,25%20billion%20every%20month%20through)).
- Anthropic's latest funding round (~$30 /billion) values it above $900 /billion, surpassing OpenAI's last $852 /billion valuation ([www.forbes.com](https://www.forbes.com/sites/jonmarkman/2026/05/04/anthropics-900b-funding-round-set-to-surpass-openai/#:~:text=than%20%24900%20billion,PROMOTED%20The%20numbers%20around%20the)).
- OpenAI's confidential IPO filing (May 22, 2026) targets a ~$1 /trillion valuation, with an estimated $60 /billion raise ([aitoolsrecap.com](https://aitoolsrecap.com/Blog/openai-ipo-2026-valuation-timeline-what-investors-need-to-know#:~:text=OpenAI%20is%20filing%20a%20confidential,timeline%2C%20valuation%2C%20revenue%2C%20and%20risks)) ([aitoolsrecap.com](https://aitoolsrecap.com/Blog/openai-ipo-2026-valuation-timeline-what-investors-need-to-know#:~:text=advising%20Target%20listing%3A%20September%E2%80%93November%202026,of%20revenue)).
- Google's Gemini 3.5 Flash outputs 4× faster than other leading models and can process up to 1 /million tokens at once ([blog.google](https://blog.google/innovation-and-ai/models-and-research/

gemini-models/gemini-3-5/#:~:text=and%20coding%20model%20yet%2C%20outperforming,used%20to%20take%20a%20developer)) ([aitoolsrecap.com](https://aitoolsrecap.com/Blog/ai-news-may-27-2026#:~:text=The%20headline%20benchmark%20numbers%3A%2076.2,1%20million%20token%20context%20window)).

- An OpenAI model solved an 80-year-old geometry conjecture—the first autonomous AI solution to a major open math problem ([techcrunch.com](https://techcrunch.com/2026/05/20/openai-claims-it-solved-an-80-year-old-math-problem-for-real-this-time/#:~:text=posted%20on%20X,That%20has%20implications%20for%20biology)).
- Chinese AI models (e.g. Alibaba's Qwen, Zhipu's GLM) now account for over 60% of usage on the OpenRouter platform ([www.buildfastwithai.com](https://www.buildfastwithai.com/blogs/ai-news-today-may-25-2026#:~:text=Is%20Claude%20Cowork%3F%20The%202026,in%20professional%20editors%20without%20switching)).

KEY TAKEAWAY

Frontier AI capabilities are accelerating and being monetized on an unprecedented scale. From ultra-fast new models to trillion-dollar valuations, a global AI race is underway. Leaders must move now to embed these advances into their strategy—or risk being left behind.

## Sources


Gemini 3.5: Frontier intelligence with action – Google Blog (May 19, 2026)
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/

OpenAI claims it solved an 80-year-old math problem – TechCrunch
https://techcrunch.com/2026/05/20/openai-claims-it-solved-an-80-year-old-math-problem-for-real-this-time/

Anthropic's $900B Funding Round Set To Surpass OpenAI – Forbes
https://www.forbes.com/sites/jonmarkman/2026/05/04/anthropics-900b-funding-round-set-to-surpass-openai/

China Clamps Down on Overseas Travel for AI Talent at Alibaba, DeepSeek – International Business Times
https://www.ibtimes.sg/china-clamps-down-overseas-travel-ai-talent-alibaba-deepseek-86961

Introducing Command A+ (Open-Source 218B Model) – Cohere Blog
https://cohere.com/blog/command-a-plus

Intuit to lay off over 3,000 employees to refocus on AI – TechCrunch
https://techcrunch.com/2026/05/20/intuit-to-lay-off-over-3000-employees-to-refocus-on-ai/

AI News Today (May 25, 2026): Top AI Stories & Headlines – BuildFastWithAI
https://www.buildfastwithai.com/blogs/ai-news-today-may-25-2026