Source: AI generated

The great split in the AI market

Gemini 3 has just been ranked as the best performing AI model, a reversal that few saw coming. A year ago, OpenAI models were leading the charts.

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Assia Driss
Syz Research Lab Team Coordinator
assia.driss@syzgroup.com

Nathan Willemin
Intern
nathan.willemin@syzgroup.com

Syz PRIVATE BANKING

Introduction

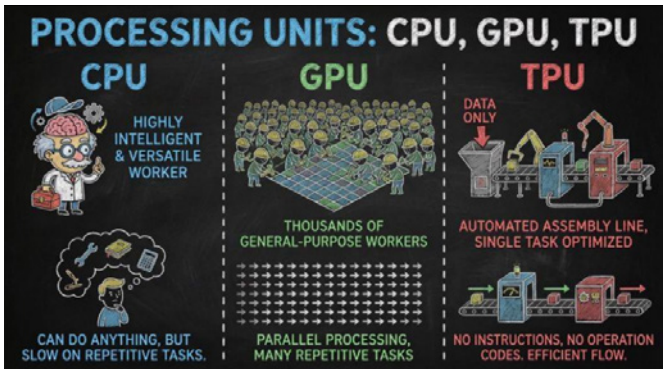
AI dominates the headlines daily: new models, unprecedented partnerships, innovative products. As ChatGPT celebrates its third anniversary, the artificial intelligence landscape has become a maze where the same players compete on multiple fronts simultaneously. Alphabet faces off against Nvidia in the chip market, while Gemini rivals ChatGPT as a language model. But how do these different battles fit together and what is the common thread in this technological war?

Explanation of CPUs, GPUs, TPUs, and NPUs

CPUs have been the core of computers since the early days of modern computing. Still today, every computer runs on a CPU. It is the main processor and works as the "general-purpose brain" of the computer. Without one, a computer can't even start. However, these chips were made for flexibility and not for repetitive calculations. Thus, in the 1990s, when video games were getting increasingly realistic, computers needed chips that could process thousands of pixels simultaneously instead of one at a time. That is what led to Graphics Processing Units (GPUs). A GPU is basically thousands of smaller, simpler cores working in parallel. While a CPU colours pixels one by one (serially), a GPU colours all of them at once (parallelly). Most computers today have both a CPU and a GPU. And the biggest producer of GPUs was Nvidia.

Then came modern AI. Training large neural networks requires mind-numbing amounts of matrix math, and that math looks almost identical to what GPUs were already designed to do for 3D graphics. Nvidia capitalised on this by adapting GPUs for datacentres, far bigger and more powerful than the ones in laptops, and gained a huge market share very quickly. GPUs, however, were still general-purpose graphics chips, not purpose-built AI chips. In 2015, Google introduced the Tensor Processing Unit (TPU), designed specifically for the operations used in neural networks. In other words, the TPU is an ASIC (application-specific integrated circuit), a chip customised for a particular use, rather than intended for general-purpose use. This specialisation enabled TPUs to offer superior cost-efficiency, energy efficiency, and high throughput at scale.

The third type of chip is the Neural Processing Unit (NPU). They work on-device—like with a smartphone—focusing on energy efficiency and real-time AI processing. It does not compete with GPUs and TPUs, as it is far from ideal when it comes to training large-scale AI models and have lower computational power. However, since they are optimised for low-power AI applications, they are primarily deployed in applications such as real-time image recognition and voice recognition.



Source: Leon Zhu & SemiVision on X

But if TPUs are better than GPUs when it comes to training and inference, why aren't all LLMs using TPUs?

Because history locked the AI world into GPUs long before TPUs arrived.

Think of it like this: by the time Google invented the TPU in 2015, the machine-learning world had already spent a decade building everything, such as research code, early neural-network libraries, university courses, or company infrastructure on top of Nvidia GPUs. And the reason wasn't just "GPUs are good." It was CUDA, Nvidia's proprietary programming platform that launched in 2007.

CUDA let developers write code that directly exploited the parallel power of GPUs. It was stable, fast, well-documented, and crucially, it worked everywhere from gaming PCs to giant datacentres. That created a flywheel, where more developers used CUDA, meaning more tools and libraries were built for CUDA, leading more companies to buy Nvidia GPUs, finally pushing even more developers toward CUDA.

By the early 2010s, the entire ecosystem was built assuming people were using Nvidia GPUs and CUDA. Universities taught it. Startups deployed on it. Cloud providers stocked warehouses full of it.

When Google dropped TPUs, they were powerful, especially for large-scale training and inference workloads. But, they were also far less flexible than GPUs in terms of the types of models and operations they could handle. Most importantly, they were late to a party Nvidia had been hosting for years, and the whole community had already built their house on CUDA. In addition, Google didn't sell TPUs. If you wanted to use them, you had to train your model on Google Cloud. Meanwhile, Nvidia GPUs were everywhere: Amazon, Microsoft, Oracle, private datacentres, university clusters. Broad availability made GPUs the default for anyone building an LLM.

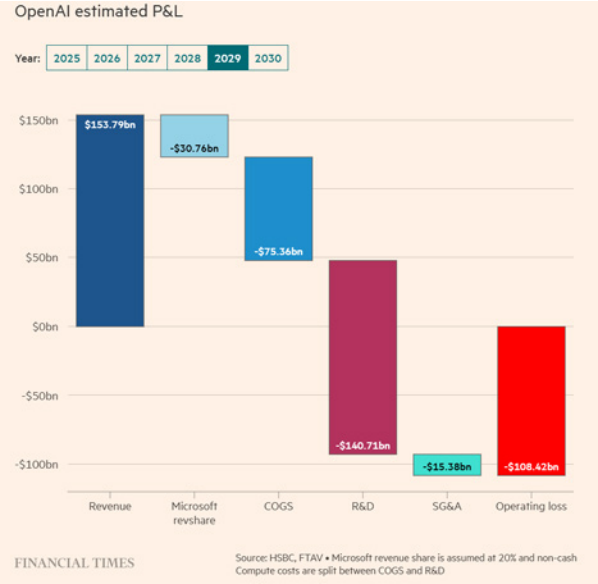
Even still, TPUs' better performance and the wish for less reliance to a single company—in this case, Nvidia—led players to start integrating Google's chips in their value chain. Anthropic recently signed a multi-billion-dollar deal with Google, gaining access to up to one million TPUs. Their models will now run across three different platforms: Nvidia's GPUs, Amazon's custom Trainium chips (also ASICs), and Google's TPUs. Meta is reportedly also in talks with Alphabet to invest billions of dollars in Google TPUs for its data centres by 2027. This would help the company diversify their chip supply beyond Nvidia's and AMD's GPUs. More generally, more companies are looking to build their own ASICs. For example, OpenAI has a new deal with Broadcom, who helped build Google's TPUs, to build its own ASICs starting in 2026.

A tale of two AI markets

The battle for AI dominance seems to have split into two competing blocks. On one side, the "Google Complex", built around Alphabet and a network of infrastructure suppliers: Broadcom, Celestica, Lumentum, TTM Technologies. On the other side lies the OpenAI Complex, centred on OpenAI and amplified by partners such as Microsoft, Nvidia, Oracle, SoftBank, AMD, and CoreWeave. Microsoft provides the financial backbone and Azure cloud infrastructure on which OpenAI trains and deploys its models, while Nvidia and AMD supply the GPUs and AI accelerators that make this compute possible. CoreWeave, backed in part by Nvidia, acts as a specialised high-performance cloud provider, leasing vast GPU

clusters to OpenAI and absorbing part of the model-training burden that Azure cannot handle alone. Oracle, for its part, has signed multi-year capacity deals to host OpenAI workloads on its own cloud, further expanding the available compute footprint. For most of the past two years, markets rewarded any association with the OpenAI Complex. A new chip-supply agreement with Nvidia, fresh cloud-capacity deals with Oracle, or a proximity deal with OpenAI routinely triggered sharp rallies as markets bet it would translate into explosive AI demand and future revenue.

This logic has recently flipped. Over the past weeks, markets have been seeing a high-profile relationship with OpenAI as a millstone around one's neck. SoftBank, set to own about 11% of the company, dropped nearly 40% in November. Oracle had surged after its $300 billion infrastructure deal with OpenAI. Now its credit-default-swap spreads are widening as investors rethink the risk of building huge data-centre capacity for a customer whose credit profile looks far weaker than expected. Even Microsoft, despite its size and diversification, felt the pressure. HSBC now estimates that OpenAI could accumulate close to half a trillion dollars in operating losses through 2030, a figure that raises serious questions about funding sustainability in an ecosystem already burdened by heavy capex, partner debt and stretched valuations.

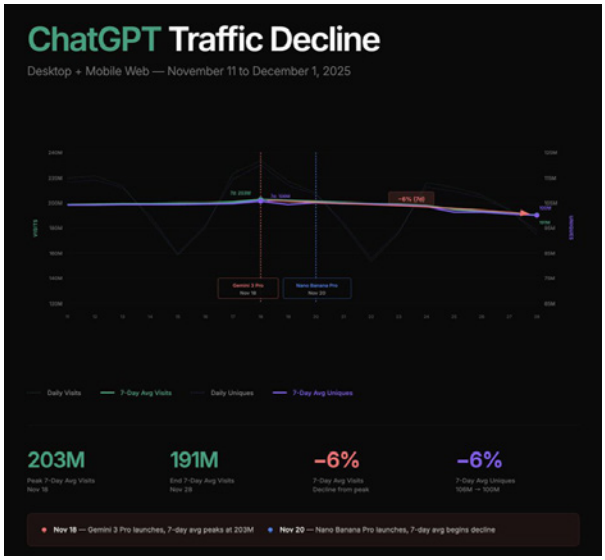


Source: Financial Times

In contrast, the Google Complex is built on disciplined AI spending, vast in-house infrastructure and steady margins. Over the past year, Alphabet has generated $151.4 billion in operating cash flow, enough to fund nearly $78 billion of capital spending, retire close to $20 billion of debt, and still return almost $70 billion to shareholders through buybacks and dividends. It is operating at a level of cash generation that very few companies in the AI race can match, with higher-quality free cash flow and less balance-sheet strain than many of the names orbiting the OpenAI ecosystem. This foundation allows Alphabet to deploy models like Gemini 3 without stretching its balance sheet. By training Gemini 3 entirely on its own TPUs, Alphabet is cutting its internal compute costs dramatically, while building a potential hardware business.

This financial strength and vertical integration are now translating directly into the model performance. Gemini 3 has climbed to the top of industry leaderboards and is now viewed as the most capable model available. In the latest LMArena rankings, Gemini 3 was the most-voted-for model. Salesforce CEO Marc Benioff did not hold back after trying it "I've used ChatGPT every day for three years. Two hours on Gemini 3, I'm not going back."

This performance shock has triggered genuine pressure inside the OpenAI Complex. Sam Altman reportedly issued a "code red" memo after Gemini 3's launch, acknowledging the rising threat posed not only by Google but also by Anthropic's Claude. SimilarWeb data shows ChatGPT traffic falling by nearly 6% in the weeks following Gemini 3's debut, dropping from 203 million to 191 million average daily visits.



Source: Former Google Deedy Das on X

Conclusion

Nvidia is still a powerhouse, and demand for its GPUs remains immense. New architectures like Blackwell and Rubin could strengthen its position again in 2026. But major cloud providers are developing their own silicon, margins are tightening, and investors are increasingly wary of the leverage and over-investment surrounding the OpenAI ecosystem. Nvidia's enduring moat remains its CUDA software and dominance in training workloads.

This is not a zero-sum contest. It is a rapidly expanding market with room for multiple winners. And right now, two centres of gravity are forming: Alphabet, with its cost-efficient integrated stack and fast-rising Gemini platform; and Nvidia, with its unmatched training and software ecosystem.

Welcome to Syzerland®

For further information

Banque Syz SA
Quai des Bergues 1
CH-1201 Geneva
T. +41 58 799 10 00
syzgroup.com

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Assia Driss
Syz Research Lab Team Coordinator
assia.driss@syzgroup.com

Nathan Willemin
Intern
nathan.willemin@syzgroup.com