Korea's Trillion-Dollar Sovereign AI Investment: Nvidia Wins, Hynix Loses

Korea hosts a Squid Games, National AI Tournament, the best non-Chinese open source model gets eliminated, why Nvidia needs open source, implications for Hynix and Samsung

MAX KAN, RAY WANG, AND DYLAN PATEL

SEP 02, 2026 · PAID

70 2 Share

Every day, businesses and governments around the world are becoming increasingly reliant on America's frontier models. Startup CEOs already can't imagine running their companies without AI, and it won't be long until the same is true for every other organization in the world.

At the same time, it's become abundantly clear that access to frontier models is at the mercy of Anthropic, OpenAI, and the United States government. Fable 5 was temporarily banned by the USG, and GPT 5.6 and Astra were similarly delayed. Both models have cyber, bio, and other safety safeguards that, though well-intentioned, often prevent good users from completing harmless tasks. Given recent security incidents and general worries about increasingly powerful AI, it is extremely likely that frontier model usage will only become more restricted from here. In fact, we believe it's plausible **OpenAI/Anthropic will eventually stop offering API access entirely for their most capable models.**

Open source seems like the obvious solution to all these dependency concerns, but it's far from a silver bullet. First, all the "open source" licenses are starting to become increasingly restrictive. As just one example, any "model as a service business" making over $20M a year must negotiate a separate agreement with Moonshot to serve Kimi K3. Second, just because a lab open sources their models today doesn't guarantee they'll continue doing so in the future. Imagine if in 2028, you're stuck with 2027 level intelligence for some extremely high-value use case because no relevant model is open source and everything has to stay on-prem for security reasons. It's kind of like having your Fable request downgraded to Opus except 100x worse.



Open source token volumes have increased significantly over the past month.
Source: OpenRouter

The only way to fully address these concerns is to pretrain your own model that runs on your own GPUs. Until now, this has largely been a question for companies: do the potential benefits justify the enormous investment required? **Soon, however, this same calculation will confront every major nation state.**

Today, we'll do a deep dive on South Korea's sovereign AI efforts. Beyond being home to two of the most important companies in the AI supply chain, Korea also has a long tradition of technological self-reliance (as any foreigner that's had to use Naver Maps knows) and is currently the clear leader in sovereign AI.

From there, we'll discuss the two major implications for investors: why Nvidia is sovereign AI's largest supporter and why Korea's ambitions may not be aligned with Samsung and Hynix shareholders.

Korean Squid Games - National AI Tournament: attempting to build a domestic frontier model

In June 2025, the Korean government announced "독자 AI 파운데이션 모델", or the "Independent AI Foundation Model" project. As the name suggests, the goal is to develop a model that Korean organizations can train, modify, and operate without depending on a foreign AI lab.

Perhaps the most interesting thing about the project is its structure. Rather than selecting a single national champion upfront, the government is hosting a tournament. All participants are provided subsidies for the three pillars of AI—compute, data, and researchers—and evaluated every 6 months. At each stage, losers are eliminated and have their resources reallocated towards the winners.

The competition began with 15 consortiums. Ten passed the initial document review, and five were selected in August 2025: Naver Cloud, LG AI Research, SK Telecom, NC AI, and Upstage. Most readers are likely entirely unaware of any of these companies' AI efforts, but some of them are surprisingly credible. SKT, LG, and Naver all pre-trained LLMs pre-ChatGPT for example.

The exact subsidies varied by team. For the first round, the government rented ~3000 H100 equivalents from SKT and Naver and distributed them to the other 3 contestants. They also spent ~$45M USD on data. Most of this was paid to Korean companies for things like books, news articles, and video broadcasts, which—along with some Korean government records—were shared among all the contestants. However, they also gave each company ~$2M to buy data themselves. Lastly for researchers, the government offered each company ~$1.4M to try recruiting overseas talent, but only Upstage took them up on the offer.

Surviving teams will be given additional resources as the tournament progresses, but the total government budget of ~$350M is still a rounding error compared to US labs. However, as we'll explain later in this article, this is just a small portion of Korea's

planned AI investment. Additionally, their results so far show that training a decent model from scratch may be cheaper than most think.

The original plan was to go from 5 teams to 4 to 3 and then finally 2, with each round lasting roughly 6 months. The two winners would be selected at the end of 2026, and the government would give them additional resources to scale up their models throughout 2027.

However, there was some unexpected drama at the end of round 1!

The 5 teams were judged 40% based on benchmark scores, 35% on expert review, and 25% on user testing. NC finished last, which means the other 4 teams should've moved on, but the government decided to disqualify Naver as well. This is because they used vision and audio encoders from Alibaba's Qwen model family.

In Naver's defense, these are separate neural networks whose outputs are then fed into the main LLM, and the Korean government was originally unclear about what contestants could vs couldn't use from open source. It was only after the submissions were made that the government decided foreign architectures were allowed, but even auxiliary components must be trained and developed after initializing the weights.

Naver maintained that they'd already developed their own encoders and could replace the Qwen pieces at any time, but the government was predictably stringent and refused to reinstate them.

Now short 1 team, the Korean government decided to host a supplemental competition to find a new 4th contestant and ultimately selected Motif Technologies, a neolab that spun out of Moreh, a Korean AI infra software company—in February 2025.

All 4 companies open sourced their newest models in July and the results for the most recent round were announced on August 18th. But before we spoil the results, let's take a look at the models.

So how good are their models?

Here's an overview of the 4 models:

semianalysis

KOREA SOVEREIGN COMPUTE

Korea's sovereign AI models

All four teams shipped a new open-weight MoE in late July 2026 — each with a different attention architecture

COMPANY	MODEL	ACTIVE PARAMS	TOTAL PARAMS	ATTENTION TYPE
LG AI Research	K-EXAONE 2.0 750B-A37B	37B	750B	Hybrid sliding-window + global GQA
SK Telecom	A.X K2	33B	688B	MLA + top-k sparse attention
Upstage	Solar Open 2 250B-A15B	15B	250B	Hybrid linear (KDA) + softmax GQA
Motif	Motif-3	13B	314B	Grouped differential latent attention

Source: SemiAnalysis

And here's how they compare on the benchmarks:

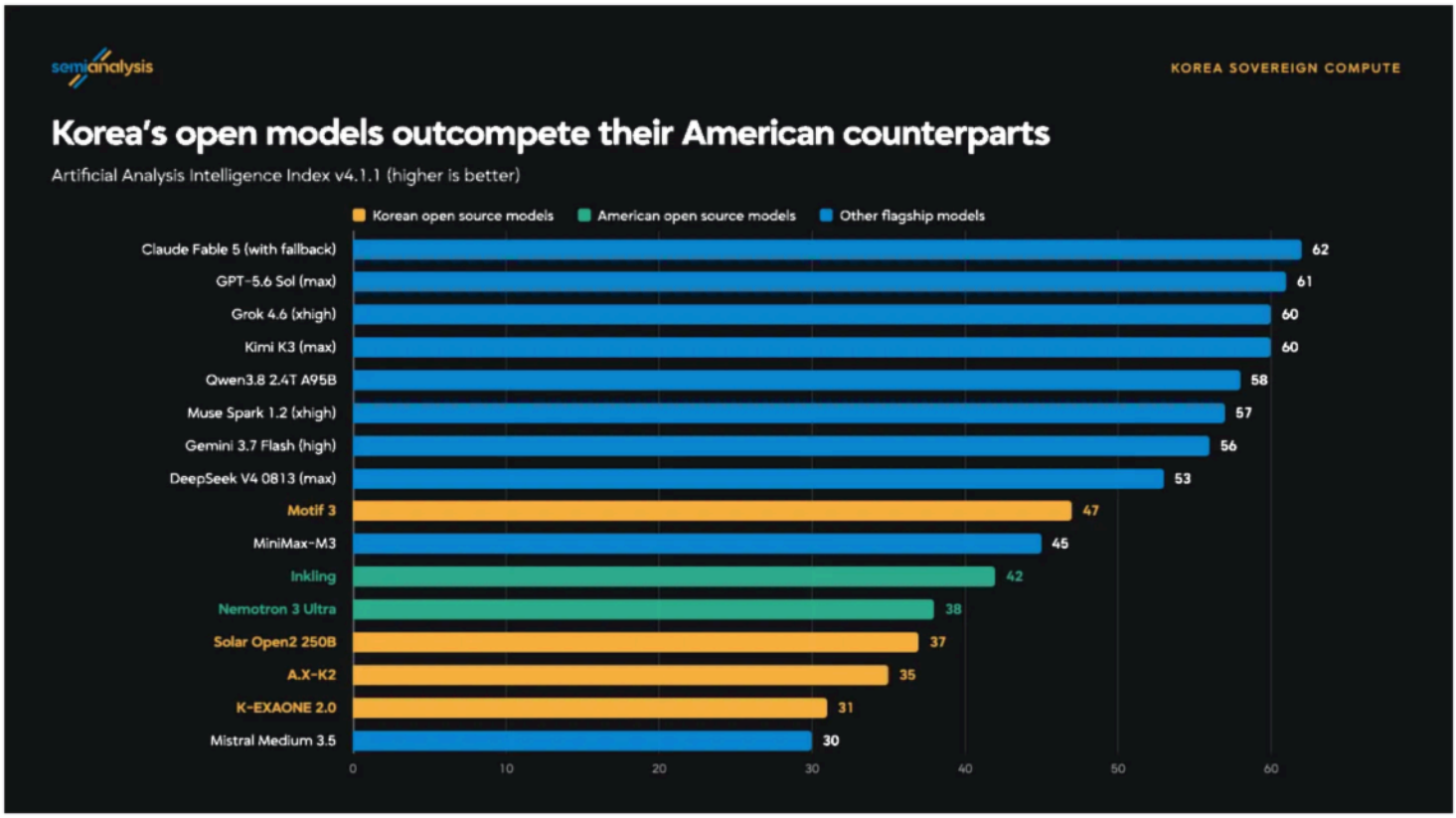
BENCHMARK	Motif-3 Motif Technologies 5 wins	Solar Open 2 250B-A15B Upstage 4 wins	K-EXAONE 2.0 750B-A37B LG AI Research 0 wins	A.X K2 SK Telecom (A.X AI) 2 wins
AGENTIC · 2 BENCHMARKS				
Terminal-Bench 2.1	74.9% Source	–	43.8% Source	–
τ³-Bench Banking	35.3% Source	19.6% Source	14.2% Source	–
CODING · 2 BENCHMARKS				
LiveCodeBench v6	–	92.4% Source	–	84.0% Source
SWE-Bench Verified	76.2% Source	70.4% Source	68.2% Source	–
GENERAL · 4 BENCHMARKS				
Artificial Analysis Long Context Reasoning	72.3% Source	62.3% Source	56.2% Source	66.0% Source
IFBench	78.2% Source	80.0% Source	72.6% Source	75.9% Source
KMMLU-Pro	–	78.4% Source	69.1% Source	80.5% Source
MMLU-Pro	–	86.2% Source	83.5% Source	–
MATH · 1 BENCHMARK				
AIME 2026	–	95.7% Source	92.3% Source	97.1% Source
REASONING · 1 BENCHMARK				
Humanity's Last Exam	37.0% Source	28.8% Source	18.3% Source	27.8% Source
SCIENCE · 1 BENCHMARK				
GPQA Diamond	83.4% Source	86.3% Source	82.2% Source	85.6% Source

Showing 11 benchmarks of 49 where at least one selected model has a reported score; a dash marks a model with no comparable number. Scores are limited to audited primary sources. Methodology varies across benchmarks.

Source: SemiAnalysis Tokenomics Model

As you can see, the two startups (Motif and Upstage) significantly outperformed teams backed by bona fide chaebols. The fact that their models are less than half the size is even more impressive.

On Artificial Analysis' Intelligence Index, Motif 3 is 10 points ahead of Upstage and by far the best Korean model. What's even more notable, however, is the fact that it's comfortably ahead of both Inkling and Nemotron 3 Ultra—the two best American open source models today.



Source: Artificial Analysis, SemiAnalysis

Motif is a sub 30 person startup that released their first pre-trained model, which was only 2.6B total parameters, last June. They've raised just $17M and have access to a mere 768 B200s (< 2MW).

In contrast, Thinking Machines is the hottest neolab around, raised $2B at $12B post for their seed, and signed a 1GW+ compute deal with Nvidia which was accompanied by additional funding. As for Nvidia itself, the resources at its disposal hardly need mentioning.



Source: SemiAnalysis

Put differently, **a tiny Korean startup you've likely never heard of trained—from scratch and on a shoestring budge—the best non-Chinese open source model in the world.**

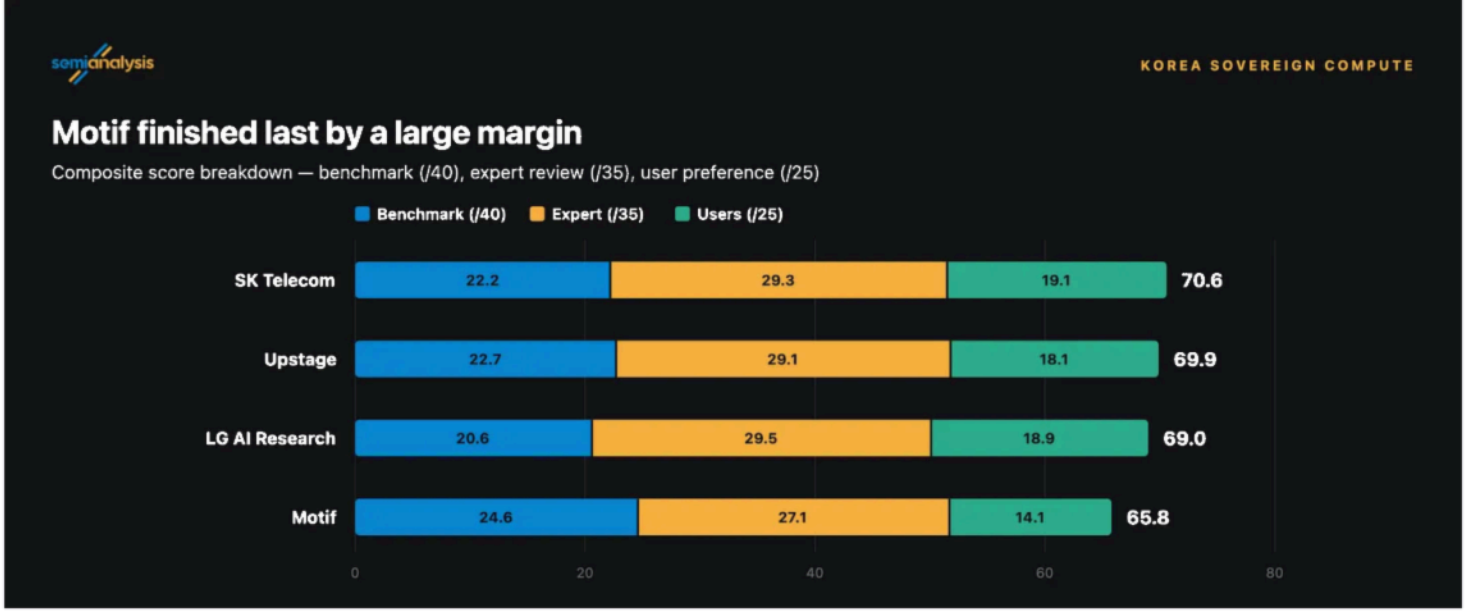
The point here is not to diss Thinking Machines or Nvidia. Training open source models is neither of their core businesses. Instead, we want to highlight that training a near-SOTA open source model from scratch is likely less resource intensive than most think. There's no substitute for focus, and the total compute cost to train Motif 3, including experimentation, was just ~$15M USD at today's prices. This is obviously well within budget for every major nation state.

Questionable judgement from the Korean government

We've previously written in-depth about the many issues with benchmarks, but they still tend to be directionally correct. **A 10+ point gap on something like the Artificial Analysis Intelligence Index has almost always corresponded to a step change improvement in model capability** (e.g. GPT 4o to o1 or Opus 4.8 to Fable 5).

With this in mind, we were shocked to see that **Motif was the one company eliminated in the latest round of competition.** Maybe you could argue that the model is benchmaxxed and thus didn't deserve to be first, but finishing dead last is baffling.

Remember that competitors were scored on three categories: 40% benchmarks, 35% expert review, and 25% user testing. Although Motif scored the best on benchmarks, they ranked last for the other two.



Source: SemiAnalysis

The exact methodology for these last two sections is frustratingly opaque. For user testing, we have no idea what tasks people actually used the model for or what criteria they considered to judge the models. Additionally, none of the testing was blind, so it's possible users had some bias towards the more famous companies.

The expert review section is even more questionable. Essentially, 10 external experts graded all four models on things like "development strategy", "future plans", and "ecosystem impact". As for what these words actually mean, your guess is as good as ours, but it seems clear that some categories disproportionately benefit large companies like LG and SKT. We also heard that the expert reviewers were concerned Motif's technology could be transferred overseas via foreign investors. This is quite puzzling, as one of the requirements of the competition is that all of the technical components must be open sourced.

We think it's a real shame Motif will no longer get any government support now that they've been eliminated from the competition. The whole point of a sovereign AI program is to rally your domestic AI talent to build a high quality model. Motif demonstrated that they're capable of exactly this, and the government should be utilizing its reach and resources to help them with things like "future plans" and "ecosystem impact"—not punish them for it.

Now, **Motif may be forced to relocate out of Korea in order to acquire d compute required to continue their research**. This is perhaps the most e result for the surprise standout of your national AI tournament. Hopefu countries can do better as they pursue their own sovereign AI efforts.



A trillion dollars worth of datacenters

With that being said, South Korea's ultimate AI champion will be very well resourced. In July, they announced a monster $919B investment to build 8.4GWs by 2029 and 18.4GWs by 2035. For the first phase, SK Group, GS Group, and Naver will be building 5GW, 2.4GW, and 1GW respectively. SK Group is additionally responsible for the remaining 10GWs for phase 2.

We've already identified 3 active sites for phase 1 totaling 4.4GWs. For full details including the exact location, MW ramp, and power source, see our Datacenter Model.



Naver's 1GW site in Sejong, South Korea. Source: SemiAnalysis Datacenter Model

Although these timelines are aggressive, we believe they are achievable. The regulatory environment is favorable, and President Lee Jae-myung is explicitly framing AI infra as a legacy-defining project akin to Korea's broadband push in the 1990s. Throughout the 2000s and 2010s, Korea had uniquely cheap and fast internet, which is what enabled their local search, e-commerce, and esports industries.

To be clear, not all of this capacity is intended to support Korea's sovereign AI efforts, and much of it will likely be sold to Anthropic or OpenAI. Importantly, however, Korea will have the optionality to dramatically scale up their sovereign AI efforts if necessary, and this domestic compute ramp is something more countries outside of the US and China will likely copy.

Why Nvidia needs open source and sovereign AI

Jensen recently made his Twitter debut with a manifesto on the importance of open source AI. It was later co-signed by basically every relevant AI company except Anthropic. Here's an excerpt:

> "Our AI leadership will be judged not by one frontier AI model, but by whether the United States builds a strong, open ecosystem that diffuses into every sector. This is essential for creating opportunities for innovation and prosperity across the country. It requires expanding access to AI, encouraging competition, robust application layers, and giving Americans greater control over the technology they rely on. Open weight models—AI models that anyone can download, inspect, modify, and run on their own infrastructure—are an important part of that foundation because they make advanced AI more accessible, adaptable, and widely available."

Much of the discourse around open source AI tends to wax poetic about the democratization of intelligence and decentralization of technological power. However, look under the hood and it's often just people talking their book.

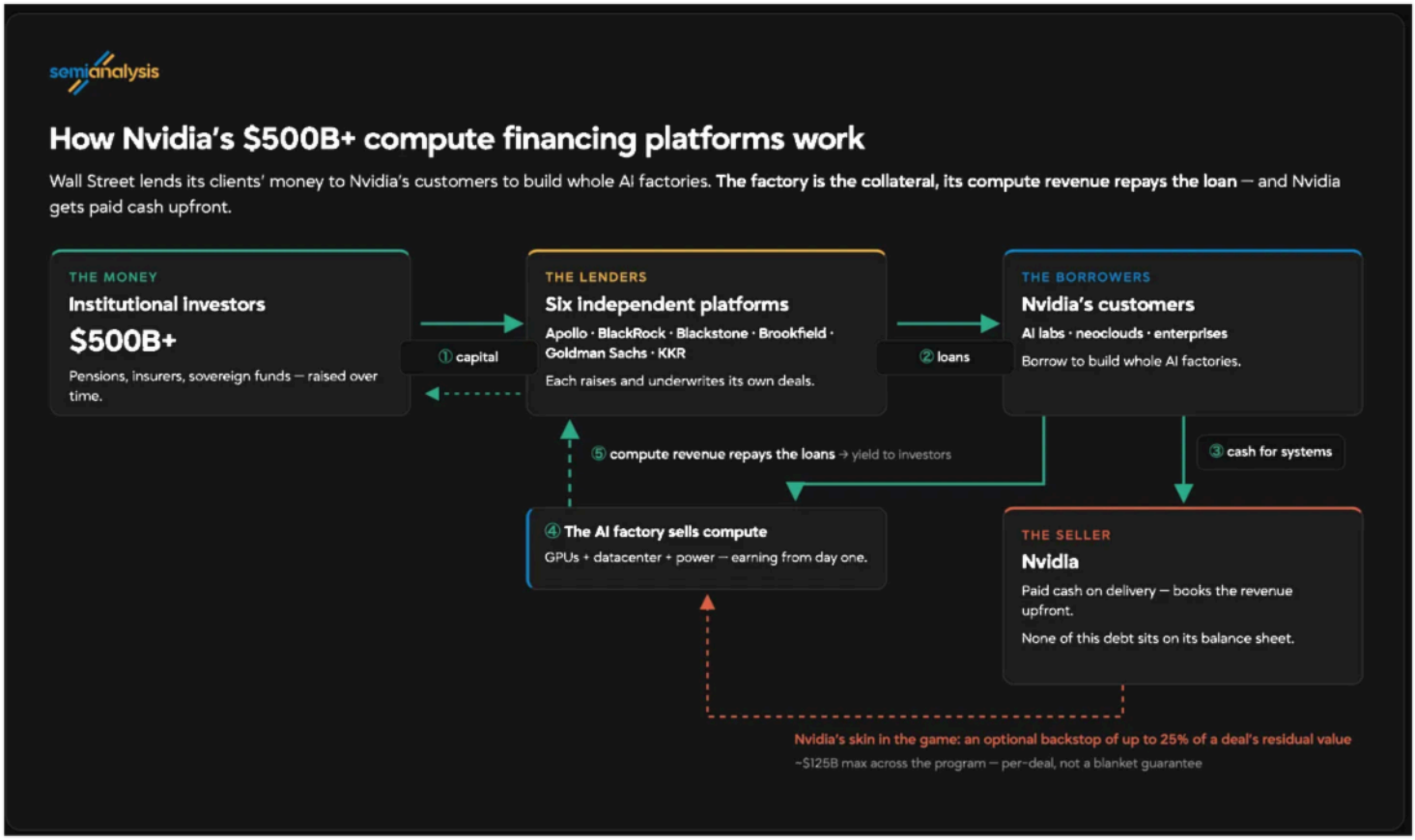
No one needs open source AI more than Nvidia. As we previously explained to Tokenomics Model subscribers, selling frontier tokens at API prices is the highest ROI use case of incremental compute. Anthropic plus OpenAI are therefore set to take a larger and larger percentage of net new GWs in the future, and Nvidia is currently on track for a world where they only have 2 real customers. Sure maybe the number's 7 if you include the hyperscalers, but that's still obviously unacceptable for the world's largest company. This is doubly true when all 7 customers except SpaceX are actively trying to build their own XPUs to cannibalize Nvidia's margins.

One of the most famous competitive strategies in tech is to "commoditize your complements". The general idea is that as something becomes cheaper and more

abundant, demand for everything that goes well with it increases. This explains why companies will often "give away" things for "free". Classic examples include 1) Google offering Chrome for free and open-sourcing Android, which increased access to the web and made search advertising more valuable and 2) Microsoft retaining the right to license MS-DOS to IBM's competitors, which helped turn the PC into a commodity hardware platform.

In Nvidia's case, AI models and applications are the complements to its GPUs. **The long-term health of their business depends on a vibrant, diverse AI ecosystem that exists outside of just Anthropic and OpenAI.**

Their recently announced $500B MOU with the world's largest capital allocators is just one example of the clever financial engineering Nvidia can do to make this vision a reality. By making "Nvidia AI Factory Compute" an "investable asset class," the SSIs and Thinking Machines of the world have a way to leverage capital from pension funds, insurance companies, and private credit firms to buy enormous amounts of compute.



Source: SemiAnalysis

Sovereign AI is another huge opportunity for Nvidia to diversify their customer base. SK Group has already committed to building 2 GWs of Rubin as part of their 5GW

buildout, and Naver has signed on for another 200MW. **We expect Jensen to continue announcing similar deals with additional countries.**

Datacenter Bet and Memory

From a sovereign AI perspective, Korea's investment in government- and corporate-owned data centers and GPU clusters makes complete sense. A sovereign model ultimately requires sovereign infrastructure: domestic control over where the model is trained, how it is deployed, and who can access the underlying compute.

Korea's challenge is that it still lacks a scaled domestic provider of leading-edge accelerators. Rebellions and FuriosaAI are making some progress, but they remain objectively behind Nvidia and AMD in deployment scale, software maturity, and commercial availability. Hyperscaler-developed alternatives such as Google's TPUs and AWS's Trainium are not domestically controlled either.

The buildout of data centers essentially comes down to two things: capital and compute. Capital is the easier problem in our view. The Korean government has been prioritizing AI investment over the past few years, while Naver, SK Telecom, Samsung, SK Hynix, and other leading domestic tech companies are committing significant resources to data centers, AI infrastructure, and model development. Compute on the other hand, is another story. Korea can finance land, power, cooling, and buildings, but it cannot manufacture a globally competitive accelerator ecosystem just like the majority of countries in the world—or guarantee access to the latest platforms—simply by spending more money.

If Korea's objective is merely to establish a domestic AI ecosystem, a conventional supplier relationship with Nvidia may be sufficient. If it wants to become one of the world's leading sovereign AI players, however, it will need something deeper: greater roadmap visibility, close technical integration, reliable access to new compute platforms, and support for developing the surrounding domestic ecosystem.

This is why Nvidia's expanded relationships with Samsung and SK Hynix matter in this context. Samsung reportedly plans to build an Nvidia-powered AI factory using more than 50,000 GPUs, while SK Telecom's announced 2GW DSX AI factory is expected to deploy Vera Rubin systems powered by SK Hynix HBM4. While the public announcement does not disclose preferential GPU allocation or guaranteed volumes, deeper strategic alignment should strengthen Korea's ability to plan and deploy

leading-edge compute—something far more difficult than simply constructing the data centers that house it.

Against this backdrop, we believe Nvidia's plans extend beyond just acquiring a large sovereign compute customer. As we previously explained to Memory Model subscribers, we believe Korea's infrastructure buildout also gives Nvidia an opportunity to deepen its relationship with SK Hynix, one of its most important memory suppliers and a key affiliate within the broader SK Group. We believe Nvidia has likely secured a long-term agreement for SOCAMM, along with favorable HBM pricing and large volume commitments for 2027, although the final negotiation result could vary from our current estimate.

Expanding Beyond Memory Across the AI Infrastructure Stack Comes at a Cost

Our latest understanding on this topic suggests that SK Hynix's 2027 HBM4 pricing could range from approximately $3.0–3.3/Gb for Nvidia, versus $3.7–4.1/Gb for other Tier 1 customers, including AMD, Broadcom, AWS, and Google, and even higher for other customers. For Nvidia, we therefore expect "only" a ~70% YoY price increase, meaningfully below the increases faced by other customers. For more detail on our HBM pricing assumptions, please refer to our Accelerator and HBM Model.

Meanwhile we estimate Samsung will receive significantly higher pricing. We estimate Samsung's price at approximately $3.5–3.9/Gb for HBM4 in 2027, compared to approximately $3.0–3.4/Gb for SK hynix—implying a ~15% premium on average. Although Nvidia should still retain meaningful bargaining power and may secure Samsung supply below the prices paid by other leading GPU and ASIC customers as we forecast, we do not expect Samsung to concede to the same extent as SK Hynix.

Why is Nvidia getting such a good deal from SK Hynix, especially when memory suppliers should have more bargaining power in a time when HBM is so tight? This comes back to SK Group's AI datacenter ambitions. SK Group needs access to Nvidia GPUs to push this initiative forward. We believe that there is an implicit understanding whereby SK Hynix offers a sweetheart deal on HBM to Nvidia, and Nvidia then gives SK Group priority GPU allocation.

Even so, a ~70% is big step-up in Nvidia's HBM costs versus 2026. While this is below the increases Nvidia is likely to face from Samsung (+98%) and Micron (+89%), it will

still materially increase the HBM component of its platform BOM. For other customers, we expect substantial price increases across all three memory vendors, although purchase volumes and existing supplier relationships could provide some degree of discount.

Nvidia will remain the largest single HBM customer in the market, and especially so for SK Hynix. While Nvidia holds importance as both a customer and technology partner, arguably Nvidia deserves pricing concessions in exchange for greater volume certainty. However, in this environment, it is the suppliers who hold the power. If Nvidia doesn't want to pay up, there are plenty of other customers who would absorb any available supply such as Broadcom, Google, AMD and Amazon. This is why we believe SK Hynix has left money on the table.

While the strategic benefits flow to SK Group, this nevertheless comes at the expense of SK Hynix's profitability. This becomes increasingly important as investors anticipate substantial earnings, free-cash-flow generation, and cash accumulation at SK Hynix over the next several years under the backdrop of moderating overall memory pricing in 2027. A more muted HBM pricing trajectory would constrain upside to earnings and free cash flow versus the opposite scenario, potentially reducing the capital available for core memory investment, dividends, and share repurchases. Some shareholders are therefore likely to scrutinize the costs and benefits of the broader SK–Nvidia partnership and the implications toward SK Hynix.

Beyond the complexity regarding data center and memory collaboration with Nvidia, we believe another factor behind the pricing difference between the two companies is product performance. We have written extensively about the challenges SK Hynix has encountered with HBM4 in our Accelerator and HBM Model, as well as the material improvements Samsung has demonstrated in HBM4 performance. Samsung's HBM4 uses a 1c DRAM process and SF4 for the logic base die. This is superior to (and more expensive than) SK Hynix's combination of 1b and N12.

The difference of HBM4 is also notable from a profitability perspective. Historically, SK Hynix has enjoyed much higher HBM margins than Samsung because of its superior front-end and back-end yields. On the cost side, SK Hynix benefits from using a cheaper 1b DRAM process and TSMC's N12 logic base die, though this comes at the expense of performance as noted above.

We nevertheless expect the margin gap between SK Hynix and Samsung to narrow materially in 2027, at least for HBM supplied to Nvidia, as pricing becomes less favorable for SK Hynix. Based on our latest estimates, the margin difference between SK Hynix's and Samsung's Nvidia-bound HBM could contract from approximately teens percentage points in 2026 to single-digit percentage points in 2027, driven primarily by the difference in their respective HBM price increases. This is built upon the fact that Samsung's front-end HBM yield has seen material improvement compared to its HBM3E 12-hi products, which were lagging behind SK Hynix and Micron in terms of yield maturity and market share.

With Samsung's stronger 2027 HBM pricing, improving product competitiveness, and execution in commodity DRAM should produce a more favorable incremental earnings and free-cash-flow trajectory than SK Hynix under our current assumptions. This could also leave Samsung in a stronger relative position to fund future investment while maintaining dividends and share repurchases.

That said, we still expect both companies to deliver strong shareholder returns given significant free cash generation, although the magnitude and form of those returns may differ. For example, SK Hynix recently announced an enhanced shareholder-return policy, raising its target from "within 50% of cumulative free cash flow" to "over 50%." The company plans to return capital through a combination of treasury share buybacks and cancellations and cash dividends, while also considering measures to expand its dividend program, including fixed and special dividends. The company said it will announce the specific scale and mix of buybacks, cancellations, and dividends alongside its 3Q26 earnings.

Samsung also announced a major shareholder-return program estimated at KRW 90–110 trillion for 2026. The plan includes approximately KRW 30 trillion in dividends, a KRW 15 trillion share buyback for employee compensation, and additional dividends and share repurchases or cancellations to be finalized in early 2027. Given the substantial free cash flow the two Korean memory makers are expected to generate, we would not be surprised to see further buybacks and enhanced shareholder returns from Samsung and SK Hynix, with Micron likely to follow suit.



Recommend SemiAnalysis to your readers

Bridging the gap between the world's most important industry, semiconductors, and business.



70 Likes · 2 Restacks

Previous

Discussion about this post

Comments **Restacks**



Write a comment...

© 2026 Dylan Patel · Privacy · Terms · Collection notice
Substack is the home for great culture