Globlex
Securities Co., Ltd.
20 December 2025

Code Red & Silicon Velocity: The 2026 AI Landscape

- The era of the "General Purpose God Model" has ended.
- Google's low-cost TPU system vs. NVIDIA's high-velocity Blackwell performance.
- We project a 2026 "Functional Duopoly", Google's Utility + NVIDIA's Infrastructure.

The New Rules of Late-2025 AI

The AI landscape of late 2025 has exited its "honeymoon phase" and entered a period of brutal industrial warfare. The year was defined by two seismic events: a "Code Red" panic in software and a "Blackwell Gap" in hardware. The delay of NVIDIA's next-generation Blackwell chips—caused by the physics-defying requirements of liquid-cooled data centers—forced the industry to innovate out of necessity, birthing a new generation of "Reasoning" models that saved AI scaling from stagnation. We argue that the era of the "General Purpose God Model" is over; the new battleground is defined by **Unit Economics** (who can generate intelligence cheapest) and **Silicon Velocity** (who can deploy infrastructure fastest).

Why Users Don't Chase Benchmarks

In the model layer, the monopoly is dead. The release of **Gemini 3 Pro** in November 2025 shattered the perception of OpenAI's invincibility, triggering an immediate counter-strike with **GPT-5.2**. However, our analysis shows that users are no longer migrating to the "smartest" model, but to the one that fits their specific workflow—creating a "Tri-Polar" market. Power users are now splitting their time: using **Claude** for reliable coding, **Gemini** for creative lateral thinking, and **ChatGPT** for deep business logic. Intelligence has become a commodity; utility is the new moat.

Google's Cost Weapon vs NVIDIA's Velocity Moat

Beneath the surface, a "Silicon Cold War" has erupted between Google and NVIDIA. Google has unveiled its most dangerous weapon, and it isn't a chatbot—it's a cost structure. By vertically integrating its new **TPU v7 (Ironwood)**, Google has become the lowest-cost producer of intelligence, utilizing a "Sovereign Advantage" to price competitors out of the market. In response, NVIDIA has abandoned its historical two-year cycle for an aggressive one-year "Velocity Moat," accelerating its **Rubin** architecture to render custom silicon obsolete before it can even be deployed.

The 2026 Playbook: Survive the SaaS Trap, Pick Your Lane

Looking ahead to 2026, we predict the emergence of a "Functional Duopoly." The market will bifurcate into two distinct lanes: the **Premium Lane** (dominated by NVIDIA) for frontier training and reasoning agents where performance is non-negotiable, and the **Efficiency Lane** (dominated by Google TPUs) for the massive volume of everyday inference where cost-per-token rules supreme. The "SaaS Trap" looms large for legacy software companies, who risk margin destruction if they cannot navigate this new cost reality.

Own Both Engines: Google's Utility + NVIDIA's Infrastructure

A Rising Tide for Two Giants. We view this not as a zero-sum war, but as a dual-engine expansion of the AI economy. We recommend with a **BUY** on **GOOG80** with TP THB 5.10), favoring its unique "Sovereign" economics and margin safety. Simultaneously, we recommend a **BUY** on **NVDA80** TP THB 33.50), identifying it as the inescapable "Arms Dealer" for the high-performance frontier. Investors should own both: one to capture the *utility* of the AI era, and the other to capture its *infrastructure*.

Analyst

Suwat Sinsadok, CFA, FRM, ERP
suwat.s@globlex.co.th,
+662 687 7026

Assistant Analyst

Peerayu Sirivorawong

Code Red & Silicon Velocity: The 2026 AI Landscape

To understand late-2025 Generative AI, you only need three ideas: 1) AI progress is limited by chips, 2) new chips require new data centers, and 3) when chips are delayed, software must find other ways to improve.

1. AI is a compute business - Modern AI is trained and run on massive clusters of specialized chips. More compute generally means better models—either because you can train bigger models, train longer, or run more "thinking" at inference time.

2. A "new chip generation" is not plug-and-play - When NVIDIA releases a new generation (e.g., H100 → B200), it's not like swapping a CPU in a PC. These systems change power needs, cooling, rack design, networking, and even how data centers are built. So "the chip exists" does not mean "the industry can use it tomorrow."

3. The 2025 bottleneck wasn't a model problem—it was an infrastructure problem. In late 2025, the world wanted more compute faster, but the transition into next-gen systems was slower than expected. That created what we call the "Blackwell Gap": demand for next-gen performance existed, but deployment lagged**.**

4. The workaround was "Reasoning" - When new hardware is constrained, labs can still improve outputs by making models "think longer" per question (test-time compute) and by using training methods that reward correct reasoning steps (e.g., Reinforcement learning with verified rewards). In simple terms: if you can't buy a faster engine yet, you tune the software to drive smarter with the old engine.

"Reinforcement learning with verified rewards" means you train the AI like a player in a game: it gets points when its answer can be checked as correct (like a math answer matching the solution, or code passing tests). If it's wrong, it gets zero—so the model learns to reason in ways that actually work, not just sound convincing.

With that context, the next section explains why the H100→B200 transition was unusually difficult—and why "Reasoning" became the bridge that kept AI progress accelerating despite the hardware delay.

(For readers who want the full infrastructure setup behind this cycle, we previously laid it out in our report *'Everything You Need to Know About the Trillion-Dollar Wave'* on AI and data centers—here.)

The Landscape of Generative AI (Late 2025)

The landscape of AI in late 2025 has moved beyond simple competition and into a high-stakes war of **Unit Economics** and **Silicon Velocity**. The year was defined by the **"Blackwell Gap"**—a period where a hardware delay forced a software revolution in **"Reasoning,"** which eventually saved the AI scaling momentum.

The transition from NVIDIA's **Hopper (H100)** to **Blackwell (B200)** proved to be **the most difficult product transition in semiconductor history**. It required a total overhaul of data centers—necessitating liquid cooling, floor reinforcement for 3,000lb racks, and massive 130kW power draws. This infrastructure complexity caused a significant delay in the deployment of next-gen compute.

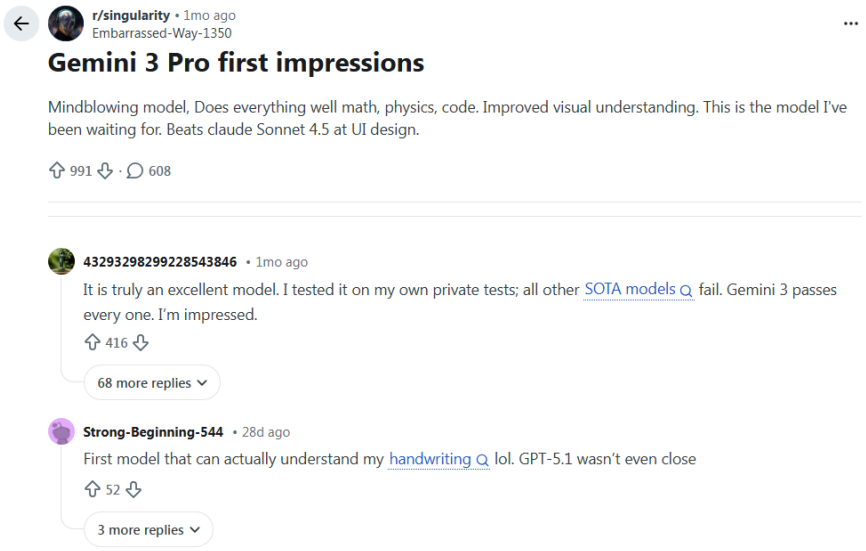
How "Reasoning" Saved AI: Had "Reasoning" (System 2 thinking / Test-Time Compute) not emerged, AI progress would have stalled for 18 months.

The Bridge: While waiting for Blackwell, research labs leveraged two new scaling laws—**Reinforcement Learning with Verified Rewards (RLVR)** and **Test-Time Compute**—to squeeze massive gains out of existing "old" chips (H100s).

The End of the Intelligence Monopoly

The market landscape shifted fundamentally on **20 Nov 2025**, with the launch of **Gemini 3 Pro**. This release served as "scientific proof" of the continued validity of pre-training scaling laws. Gemini 3 did more than just compete; its native multimodal understanding and viral creative features—specifically the **Nano Banana Pro** image engine—triggered a "Code Red" at OpenAI.

Exhibit 1: Gemini 3 Pro user's impression

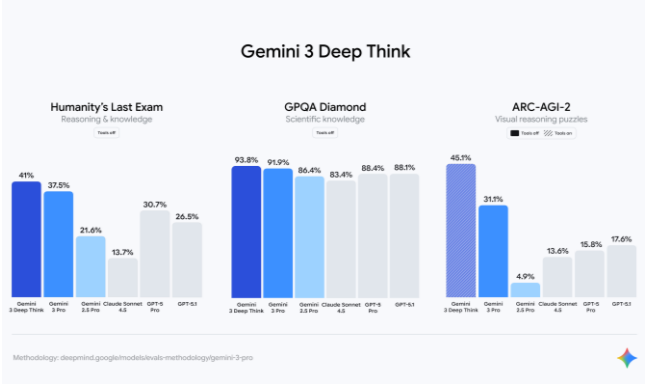


Sources: Reddit

OpenAI responded with the emergency release of **GPT-5.2** on **11 Dec 2025**. This high-speed "Leapfrog Cycle" proved that raw intelligence is now a commodity. As of today, **the "best" model** is often simply whichever one was **released most recently.**

In our view, Generative AI will **no longer be a monopoly** dominated by a single player as ChatGPT did in 2023-2024. Instead, the market is fragmenting based on **specialized utility**.

Exhibit 2: Gemini 3 Deep Think performance



Sources: Gemini

Exhibit 3: ChatGPT 5.2 Thinking performance

	OpenAI		Anthropic	Google
	GPT-5.2 Thinking	GPT-5.1 Thinking	Claude Opus 4.5	Gemini 3 Pro
SWE-Bench Pro (Software engineering)	• 55.6%	50.8%	52.0%	43.3%
GPQA Diamond (Science questions, No tools)	• 92.4%	88.1%	87.0%	91.9%
CharXiv Reasoning (Scientific figure questions, No tools)	• 82.1%	67.0%	—	81.4%
FrontierMath (Advanced mathematics, Tier 1–3, Tier 4)	• 40.3% 14.6%	31.0% 12.5%	— —	37.6% • 18.8%
AIME 2025 (Competition math, No tools)	• 100.0%	94.0%	92.8%	95.0%
ARC-AGI-1 (Abstract reasoning)	• 86.2%	72.8%	80.0%	75.0%
ARC-AGI-2 (Abstract reasoning)	• 52.9%	17.6%	37.6%	31.1%
GDPval (Knowledge work tasks)	• 70.9%	38.8% GPT-5	59.6%	53.5%

Sources: OpenAI

From “Smartest” to “Best Fit”

Current market sentiment suggests that users are no longer asking "Which AI is smarter?" but rather "Which AI fits my specific workflow?" Qualitative data from power users confirms that the leading models have developed distinct "personalities" and strengths, leading to a fragmentation of the user base.

The "Logic vs. Lateral Thinking" Split Users are finding that OpenAI and Google have optimized for different cognitive approaches. As user *head_leek_880* noted in a comparison of the two flagship models:

> *"GPT-5.2 is very good at following direction... I used it to analyze a contract [and it] laid out all the details and even did all the calculation for me. Then Gemini [3 Pro] started picking up the intention of the contract is not to offer service but to mitigate the risk... Gemini 3 has a 'weird' ability to reframe problems and look at things from different angles."*

This highlights a key divergence: **GPT-5.2** excels at execution and compliance, while **Gemini 3** excels at strategic reframing and lateral insight.

The Tri-Polar Market (The "Best Tool" Era) Feedback indicates a solidification of a "Big Three" dynamic, where each firm holds a distinct territory:

1. **Claude Opus 4.5:** The specialist for coding, front-end UI, and architectural reliability.
2. **Gemini 3 Pro:** The leader for images, lateral creativity, and "thinking outside the box."
3. **GPT-5.2:** The engine for logical business processes, extensive reasoning chains, and large-context analysis.

Importantly, this segmentation is not just internet commentary. We reached the **same conclusion** through an **interview:** a software developer at a Nasdaq-100 company independently confirmed that, in day-to-day production work, Claude is the preferred choice for coding tasks—citing higher reliability and a smoother workflow when shipping real code. This supports our view that **“best model”** is becoming **task-specific**, and that switching behavior will increasingly be partial (multi-model routing) rather than full migration to a single provider.

The Silicon Cold War - Infrastructure (2026 Outlook)

Context Primer: Think of an AI company like a bakery selling bread (Intelligence).

The "Renter" (OpenAI): They have the best recipes (Software), but they rent their kitchen from Microsoft Azure and buy their ovens from NVIDIA. Every loaf of bread they bake has a "rent tax" built into the cost.

The "Owner" (Google): They own the building, they build their own ovens (TPUs), and they even own the power plant. Because they pay no rent to NVIDIA or Microsoft, they can sell their bread for **$0.20** while it costs OpenAI **$1.00** to make the same loaf.

To understand the “Silicon Cold War” in 2026, you only need four simple ideas:

1. AI is expensive to run, not just expensive to invent: Building a great model is one thing. Running it for millions of users every day is another. Most of the money is spent on “inference”—the cost of generating answers, images, and videos at scale
2. Chips decide the cost per answer: Every AI response has a real cost: electricity, servers, cooling, and the chips doing the work. So, the core business question becomes: “How cheaply can you produce one useful AI output (one token)?
3. There are two ways to get compute: own it or rent it: Some companies build their own “AI factories” (design chips + run data centers). Others rent compute from NVIDIA-based clouds. Owning tends to be cheaper long-term; renting is faster and more flexible—but usually higher cost.

4. Speed is a weapon because the frontier moves fast: In AI, being early matters. If one chip platform improves every year, and another takes 2–3 years to redesign, the fast platform can stay ahead even if the slower one is cheaper.

   With that context, 2026 splits into two economic models: “low-cost sovereigns” who can produce AI cheaply because they own the stack (Google + TPUs), versus “high-performance renters” who pay premium costs to access the very best performance (OpenAI/Anthropic on NVIDIA). And NVIDIA’s defense is not to compete on price—it’s to compete on velocity.

While the headlines of 2025 focused on the "Model Wars" between Gemini and GPT-5.2, the decisive battles were fought in the data center. As we look toward 2026, the industry is bifurcating into two distinct economic realities: the **Low-Cost Sovereigns** (Google) and the **High-Performance Renters** (OpenAI/Anthropic).

The "Economic Oxygen" Strategy: Google’s TPU Advantage

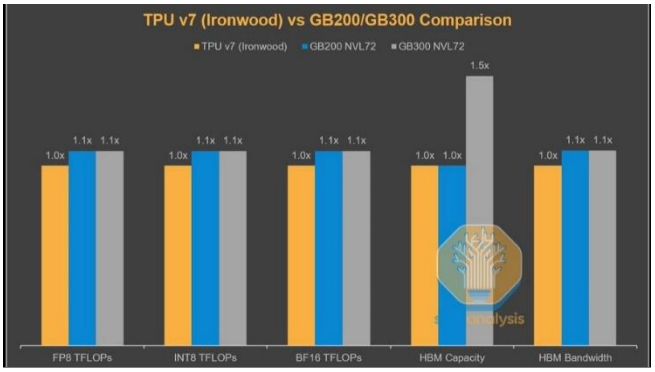
In 2025, Google unveiled its most dangerous weapon, and it was not a chatbot—it was a cost structure. By vertically integrating its stack with **TPU v7 (Ironwood)**, Google has become the lowest-cost producer of intelligence in the world.

The Sovereign Advantage: Unlike OpenAI, which must pay the "NVIDIA Tax" (70%+ gross margins) and the "Cloud Tax" (Microsoft Azure/Oracle margins), Google pays cost.

Sucking the Oxygen: As noted by investor Gavin Baker, Google is utilizing this advantage to "suck the economic oxygen out of the room." By pricing Gemini 3 Pro and Flash at aggressive rates (often near zero margin), they effectively starve competitors of capital.

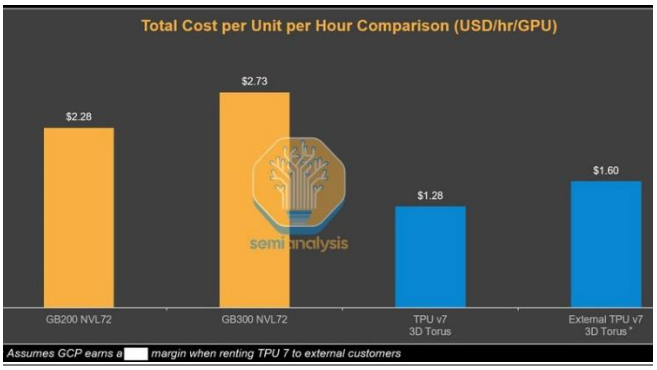
This creates a crisis for companies like OpenAI. If the market price for a "Reasoning Token" drops to match Google’s TPU economics, OpenAI—a high-cost producer using rented NVIDIA GPUs—faces a margin crisis. This was the true driver of the "Code Red" in late 2025: OpenAI realized they could not compete on price without their own hardware (Project Stargate).

Exhibit 4: Google’s TPU v7 vs Nvidia’s Blackwell



Sources: SemiAnalysis

Exhibit 5: Cost per unit per hour, TPU v7 vs Blackwell



Sources: SemiAnalysis

NVIDIA’s Response: The "Velocity Moat"

For years, analysts predicted that custom ASICs (like TPUs and AWS Trainium) would eventually kill NVIDIA’s margins. In 2025, NVIDIA responded not by lowering prices, but by increasing speed.

The One-Year Cadence: NVIDIA abandoned its historical two-year release cycle. By accelerating the roadmap from **Blackwell (2025)** to **Rubin (2026)**, NVIDIA created a "Velocity Moat.

Outrunning the ASIC: Designing a custom chip like the TPU takes 24–36 months. By the time Google optimizes software for TPU v7 (targeting Blackwell performance), NVIDIA has already released Rubin.

The threat to Google is not just speed; it is **Memory Bandwidth**. Rubin introduces **HBM4 memory**, which solves the 'Reasoning Bottleneck.' If future models need to 'think' for 5 minutes, they need massive bandwidth that current TPUs lack. Rubin is designed to win the 'Thinking War,' not just the 'Training War.'

The Silicon Rivalry – TPU vs. NVIDIA Blackwell

The rivalry between Google and NVIDIA in late 2025 is defined by a paradox: **Google currently holds a "temporary advantage"** in deployment, but NVIDIA is positioned for a generational leap that may fundamentally alter the strategic calculus by 2026.

To understand the current hardware gap, industry insiders liken the chips to eras of fighter aviation:

NVIDIA Hopper (H100): The **P-51 Mustang** (World War II era). Reliable and widely used, but aging.

Google TPU v6/v7 (2024/2025): The **F-4 Phantom** (reliable, cost-effective workhorses). A Jet Age leap over Hopper, giving Google a temporary but distinct performance lead in the current "gap" period.

NVIDIA Blackwell (B200): The **F-35 Lightning**. A superior 5th-generation platform that has taken longer to deploy but will dominate once airborne.

Why Firms Choose TPU: You do not always need an F-35 to deliver mail. For inference and training "good enough" models, the cost efficiency of TPUs is superior.

The TPU v7 is not just faster than the Hopper (H100); it is an **'Assembly Line Jet.'** It lacks the flexibility of NVIDIA chips (it can't do everything), but for the specific path of 'flying in a straight line' (Inference), it is unbeatable on price. NVIDIA is the 'Dogfighter'—flexible enough to handle any new math researchers invent.

The Transition Timeline: Why 2026 is the "Blackwell Year"

While Blackwell racks are physically deploying as of late 2025, software reality lags behind hardware delivery.

History suggests it takes **6 to 12 months** for a new chip generation to truly outperform a "finely tuned" previous generation. For Blackwell specifically, engineers estimate a **6 to 9-month ramp** to unlock its full potential.

Model Deployment: Because scaled Blackwell deployments only began in Q3/Q4 2025, the first AI models *actually trained* entirely on Blackwell are not expected to release until **early 2026**

Once Blackwell and its **GB300** iteration (which is drop-in compatible) are fully operational, the users of these chips (OpenAI, Meta, xAI) are expected to become the new **"low-cost producers,"** effectively neutralizing Google's current economic advantage.

Exhibit 6: Google TPU vs NVIDIA Hopper (H100) and Blackwell (B200)

Category	Google TPU (Cloud TPU, e.g., v6/v7)	NVIDIA Hopper (H100)	NVIDIA Blackwell (B200/GB200)
What it is	Google's in-house AI chip (an "ASIC") mainly offered through Google Cloud	NVIDIA's flagship AI GPU generation (workhorse of 2023–2025)	NVIDIA's next-gen AI GPU platform (the big upgrade after H100)
Simple idea	"Cheaper AI at scale (inside Google's ecosystem)"	"The standard engine for frontier AI"	"The new high-performance engine—bigger leap, harder to deploy"
Best use case	High-volume inference (serving lots of requests), cost-optimized training	Training large models, serious inference, broad enterprise workloads	Frontier training at larger scale, heavier reasoning/inference, next wave of giant clusters
Biggest strength	Cost efficiency when you're willing to live in Google's stack	Most mature + widely supported ecosystem (tools, libraries, integrations)	Much higher performance ceiling + designed for rack-scale systems
Typical buyer motivation	"I want lowest cost per answer in Google Cloud"	"I need reliable, widely compatible AI compute now"	"I need the next performance tier and can invest in infrastructure"
Software ecosystem	Strong inside Google stack; improving portability, but more "platform-specific"	CUDA ecosystem is the industry default for developers	Same CUDA advantage, plus newer system-level stack as clusters get larger
Switching friction	Higher if your code/tools are CUDA-centric; easier if you're already on Google Cloud	Lowest friction across clouds and vendors	Similar dev experience to Hopper, but may require newer system designs
Cost structure (intuition)	Often attractive economics for large-scale serving	Premium-priced compute, but very versatile	Likely premium (or more) vs Hopper, justified by performance and scale
"Time-to-deploy" in real life	Generally easier if you're already on Google Cloud TPU capacity	Mature supply + many deployment recipes	Often slower at first: new racks, power/cooling, integration ramp
Data center demands	High, but Google provides a lot "as a service"	High, but well-understood	Highest: next-level power + cooling + rack integration for full benefits
Scaling to huge clusters	Great within Google's TPU pod architecture	Strong; widely used in big clusters	Stronger emphasis on rack-scale designs and dense interconnect
Where it can feel "best"	When you care more about serving cost/throughput than absolute peak performance	When you want the safest, most compatible choice	When you want the new frontier capability and can handle the ramp time
Main tradeoff	Less universal developer ecosystem; more tied to Google platform decisions	Cost per token can be high at massive inference scale	Early-cycle friction + infrastructure upgrades; takes time to reach full potential
The vibe	"Public transit: cheap per ride at huge volume"	"Reliable SUV: works for almost everything"	"New supercar platform: fastest, but needs the right track + garage"

Sources: Globlex Research

The Divergent Duopoly

In our view, 2026 will not be the year NVIDIA's "monopoly breaks." It will be the year the market matures into a **Functional Duopoly**. As compute TAM explodes, the industry naturally splits into **two lanes**—and this ends the "one chip fits all" era. Customers will stop buying the most expensive silicon for every workload, and instead **segment spend by task**.

Lane 1: The Premium Lane (NVIDIA = "Ferrari Dealership").

This lane is defined by workloads where "good enough" is not acceptable: frontier-model training, reasoning agents, and scientific simulations. Here, performance, interconnect, and full-stack reliability matter more than unit cost—and NVIDIA remains the default. In this lane, NVIDIA effectively retains the monopoly on "smart," because customers are paying for the highest ceiling of capability.

Lane 2: The Efficiency Lane (Google TPU = "Public Transit System").

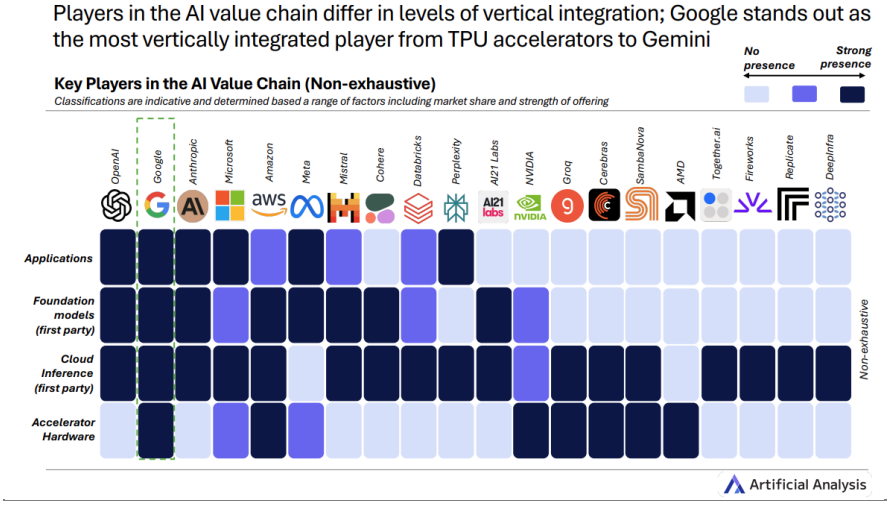
This lane is defined by scale economics: high-volume inference, internal enterprise workflows, and everyday serving where the goal is the lowest cost per token at massive throughput. For these jobs, TPUs can be "good enough" at a meaningfully better economics, allowing Google to capture the bulk volume of utility workloads.

Bottom line for 2026: it's not a zero-sum war—it is lane expansion

We expect **both NVIDIA and Google to grow**, but for different reasons. NVIDIA grows because the **ceiling of intelligence keeps rising**—as long as the world demands smarter models, demand for peak performance platforms (Rubin-class systems) persists. Google grows because the **floor of utility keeps widening**—as AI spreads into daily products, the demand for cheap inference accelerates

Exhibit 7: Player in AI value chain

Players in the AI value chain differ in levels of vertical integration; Google stands out as the most vertically integrated player from TPU accelerators to Gemini



Sources: Artificial Analysis

Alphabet (GOOG80): The Sovereign Play We recommend a **BUY** on GOOG80 with a TP of **THB5.10 ($322, USD/THB at 31.5**). Our conviction rests on Alphabet's accelerating AI momentum and its uniquely integrated stack. Unlike peers, Google controls every layer of the value chain: the data flywheel (Search/YouTube), the infrastructure (Google Cloud/TPUs), the OS (Android), and the application layer (Gemini/Workspace). This vertical integration creates durable monetization and significant operating leverage as AI utility scales.

Nvidia (NVDA80): The Infrastructure Play We recommend a **BUY** on NVDA80 with a TP of **THB33.50 ($213, USD/THB at 31.5).** Nvidia is no longer just a chipmaker; it is extracting infrastructure rent across the entire AI value chain. From the data center (Blackwell) to the edge, Nvidia remains the most critical enabler in the AI economy.

GENERAL DISCLAIMER

Analyst Certification

Suwat Sinsadok, Register No. 020799, Globlex Securities Public Company Limited

The opinions and information presented in this report are those of the Globlex Securities Co. Ltd. Research Department. No representation or warranty in any form regarding the accuracy, completeness, correctness or fairness of opinions and information of this report is offered by Globlex Securities Co. Ltd. Globlex Securities Co. Ltd. Accepts no liability whatsoever for any loss arising from the use of this report or its contents. This report (in whole or in part) may not be reproduced or published without the express permission of Globlex Securities Co. Ltd.

RECOMMENDATION STRUCTURE

Stock Recommendations

Stock ratings are based on absolute upside or downside, which we define as (target price* - current price) / current price.

BUY: Expected return of 10% or more over the next 12 months.
HOLD: Expected return between -10% and 10% over the next 12 months.
REDUCE: Expected return of -10% or worse over the next 12 months.

Unless otherwise specified, these recommendations are set with a 12-month horizon. Thus, it is possible that future price volatility may cause temporary mismatch between upside/downside for a stock based on market price and the formal recommendation.
* In most cases, the target price will equal the analyst's assessment of the current fair value of the stock. However, if the analyst doesn't think the market will reassess the stock over the specified time horizon due to a lack of events or catalysts, then the target price may differ from fair value. In most cases, therefore, our recommendation is an assessment of the mismatch between current market price and our assessment of current fair value.

Sector Recommendations

Overweight: The industry is expected to outperform the relevant primary market index over the next 12 months.
Neutral: The industry is expected to perform in line with the relevant primary market index over the next 12 months.
Underweight: The industry is expected to underperform the relevant primary market index over the next 12 months.

Country (Strategy) Recommendations

Overweight: Over the next 12 months, the analyst expects the market to score positively on two or more of the criteria used to determine market recommendations: index returns relative to the regional benchmark, index sharpe ratio relative to the regional benchmark and index returns relative to the market cost of equity.

Neutral: Over the next 12 months, the analyst expects the market to score positively on one of the criteria used to determine market recommendations: index returns relative to the regional benchmark, index sharpe ratio relative to the regional benchmark and index returns relative to the market cost of equity.

Underweight: Over the next 12 months, the analyst does not expect the market to score positively on any of the criteria used to determine market recommendations: index returns relative to the regional benchmark, index sharpe ratio relative to the regional benchmark and index returns relative to the market cost of equity.