From Silicon to Services

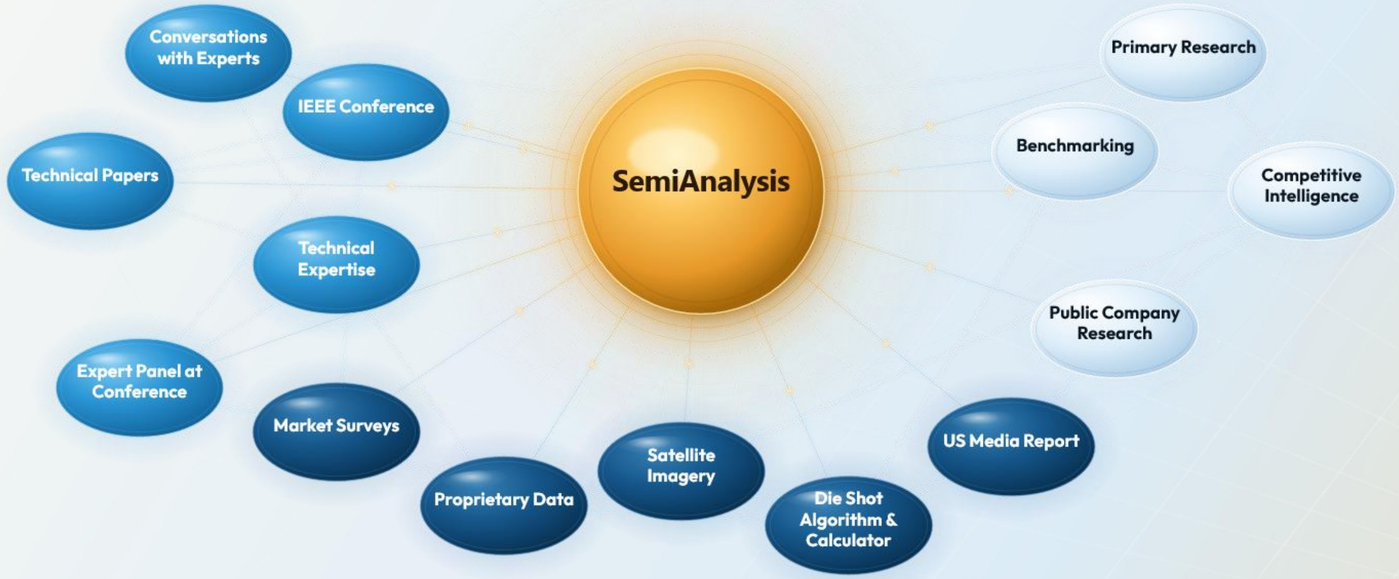
Dylan Patel

Founder & CEO, SemiAnalysis

Who is SemiAnalysis?

Independent semiconductor and AI research, 80+ conferences a year, bottoms-up models

- SemiAnalysis is the best-of-breed Semiconductor and AI Research boutique. Our coverage spans the entire AI Infrastructure supply chain, from semiconductor fabrication essentials to cutting-edge AI Models, software, and infrastructure for them.
- Product-first Approach
  - Analysts have extensive experience following and working in the Semiconductor and AI industry, allowing us to delve into technical complexities. We attend 80+ conferences annually and interact with industry professionals to understand technological details and the merits of products and services.
- Connecting the Dots
  - No other independent research house spends most of its time in the field, seeking out and connecting publicly available data points and applying deep technical expertise to build bottoms-up models that drive actionable conclusions and insights.



2026 | Confidential | All Rights Reserved

What We Build

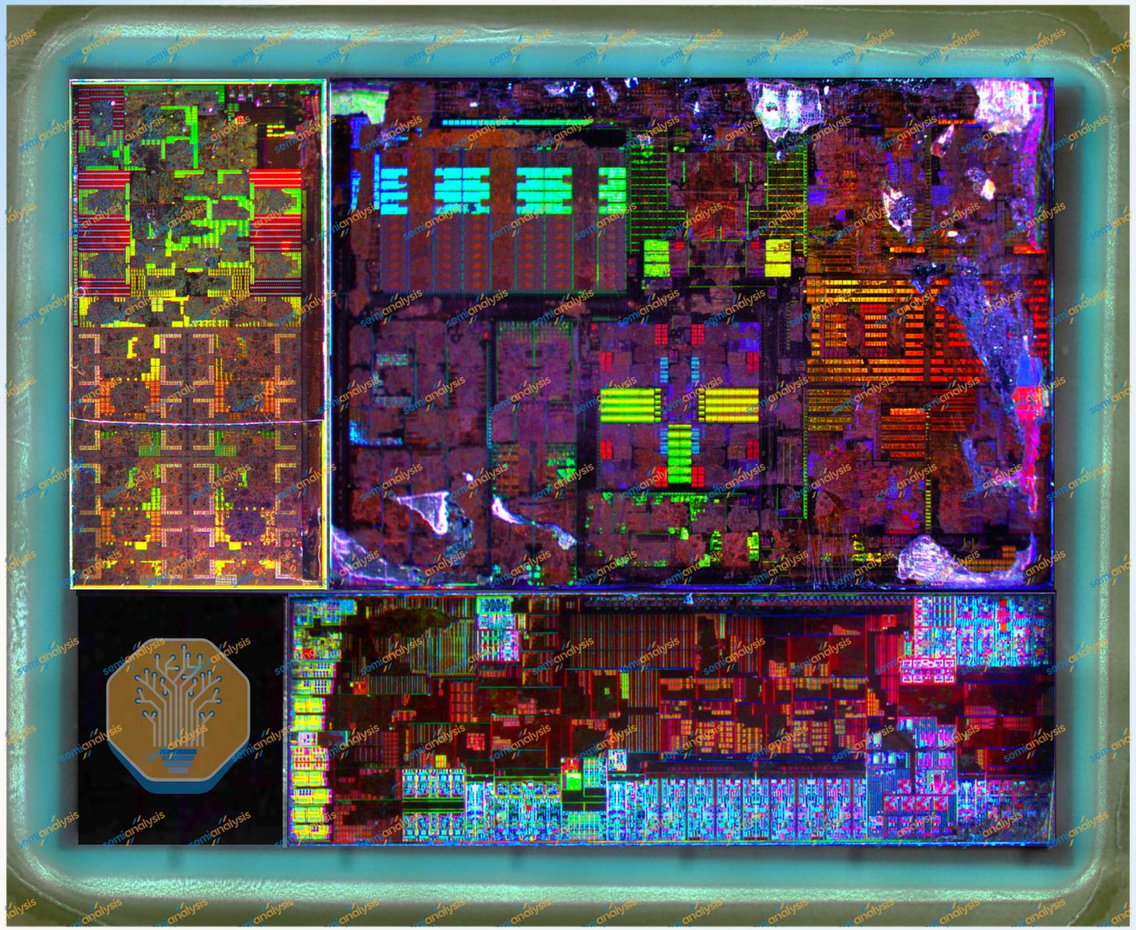
Fourteen institutional models, two open benchmarks, one rating system, one free price index

MODELS	BENCHMARKS	RATINGS	DATA
Tokenomics	InferenceX	ClusterMAX 3.0, September 2026	GPU Spot-Contract Composite Indexes, hourly
Datacenter – Industrials, Energy, China Datacenter	AgentX	323 providers tracked	H100 1-year contract index, free since April 2026
AI Supply - Accelerator & HBM, Memory Model	EndpointX (first edition coming)	77 reviewed	H100 composite on Bloomberg, ticker SAH100SC
AI Cloud TCO	SandboX (upcoming)	19 Medallions	All other GPUs available to clients
AI Networking			
Wafter Fab Equipment			
Core Research			
Foundry			
Compute Capital and Markets (launching)			
Chip Book			
Space DC TCO Model			
VR NVL72 Model			

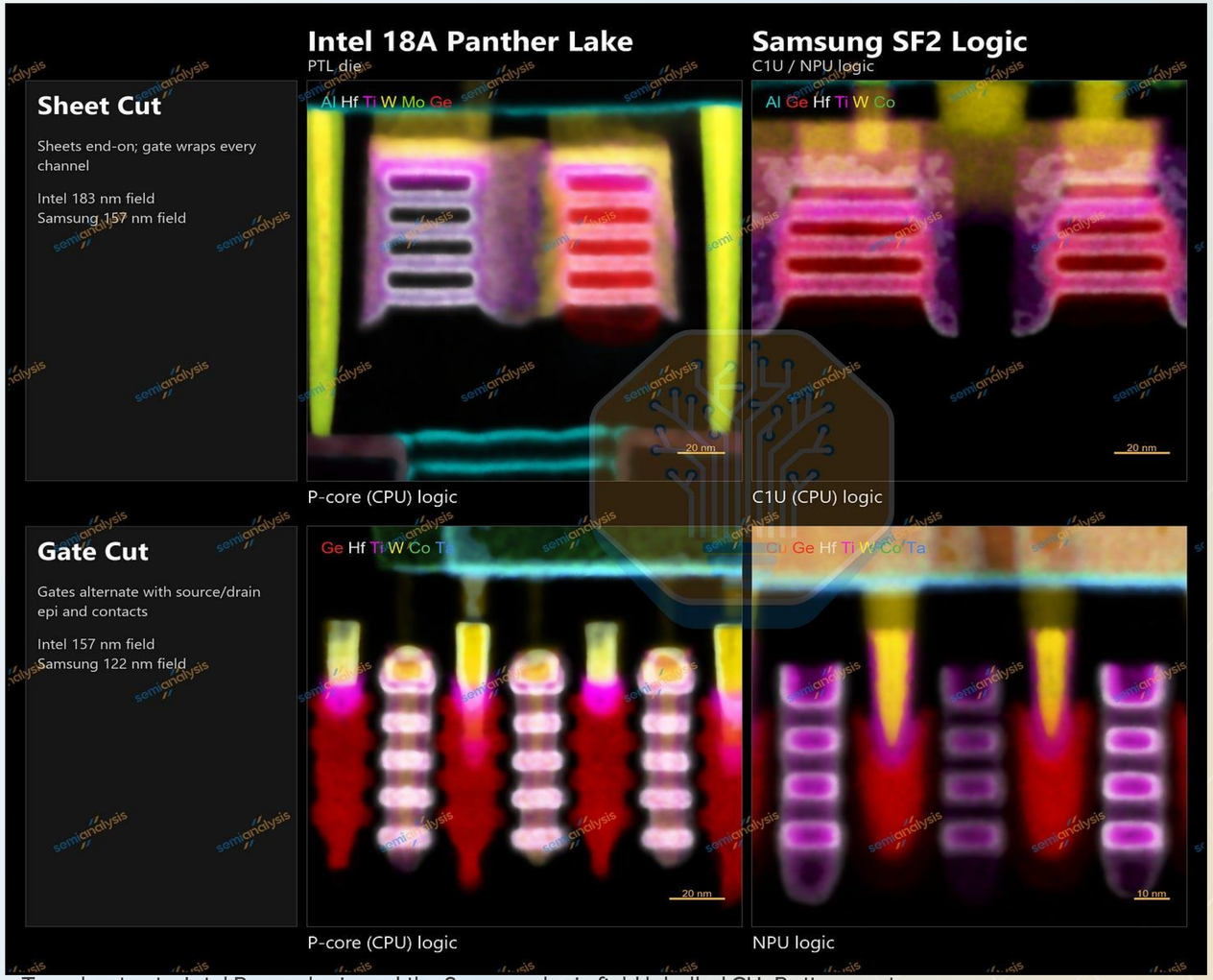
Source: semianalysis.com, inferencex.semianalysis.com, clustermax.ai, gpu-index.semianalysis.com, 30 Sep 2026

2026 | Confidential | All Rights Reserved

STEEL



Intel Core Ultra 7 365 (Panther Lake). Source: SemiAnalysis



Top: sheet cuts, Intel P-core logic and the Samsung logic field labelled CU. Bottom: gate cuts, Intel P-core logic and Samsung NPU logic. Cut directions are matched; cell functions and drive targets differ. Scale bars per panel. Source: SemiAnalysis

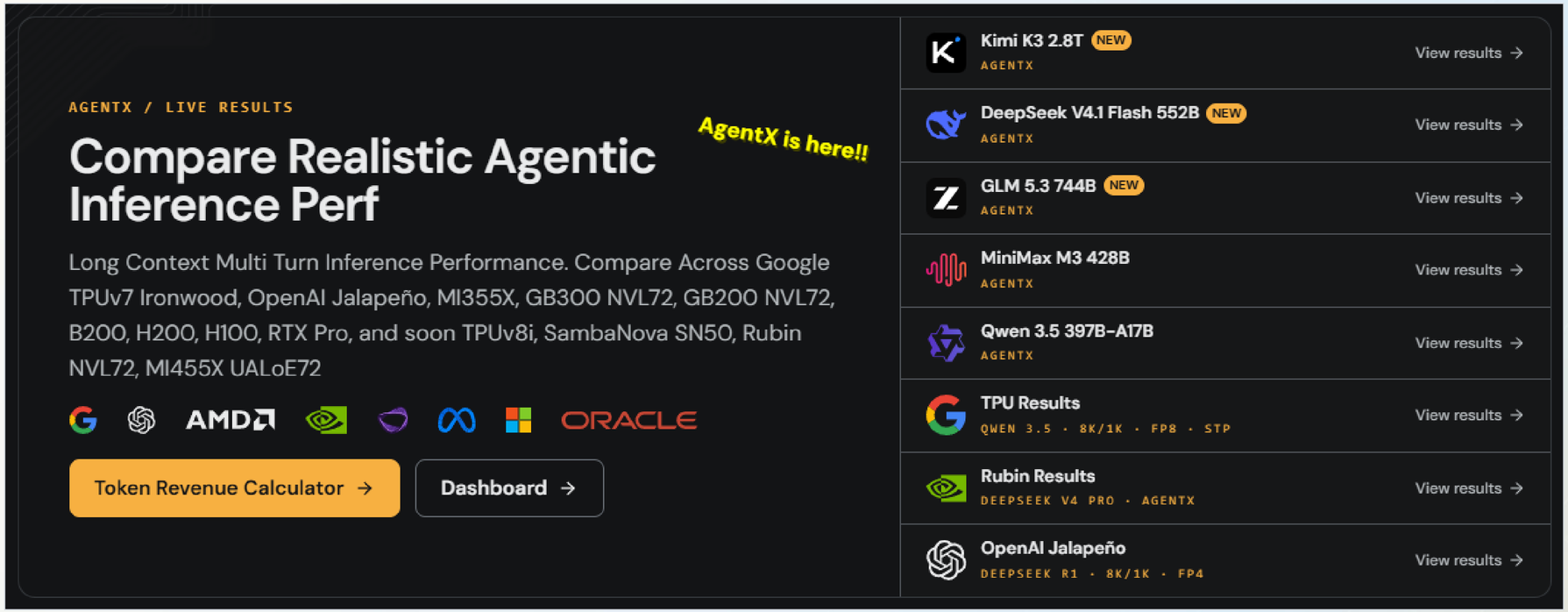
2026 | Confidential | All Rights Reserved

InferenceX

semianalysis

One Open Benchmark, Every Accelerator

NVIDIA, AMD, Google TPU and OpenAI silicon, rerun on public GitHub Actions



13 accelerator SKUs

~19 AgentX models

10+ serving stacks

1,806 GitHub stars

Data: inferencex.semianalysis.com homepage and About page, captured Oct 6, 2026; every data point links to its public GitHub Actions run

Source: InferenceX, About

semianalysis

2026 | Confidential | All Rights Reserved

Backed by the Inference Ecosystem

Model labs, hyperscalers, neoclouds and serving frameworks, plus Lisa Su and Jensen Huang



Data: InferenceX supporters page (logos and named quotes), captured Oct 6, 2026

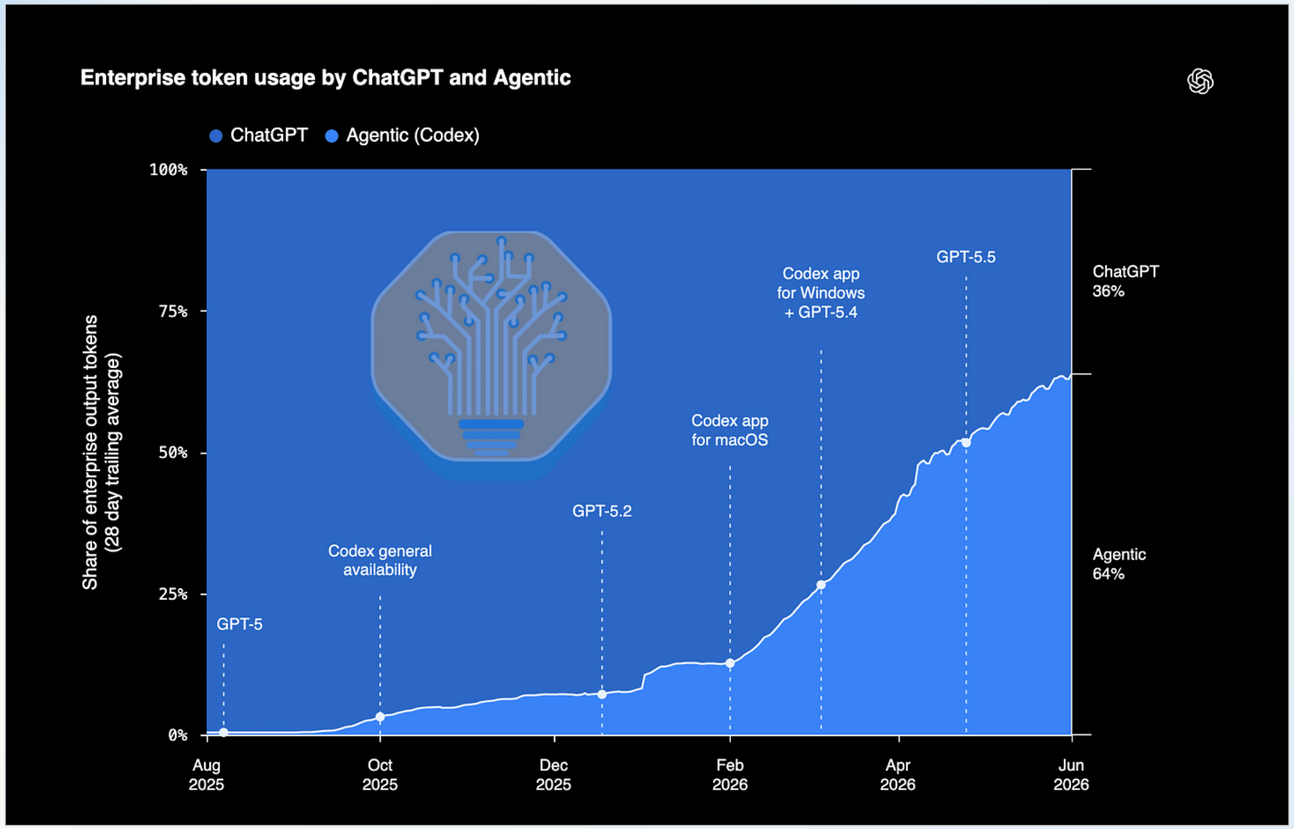
Source: InferenceX supporters

semianalysis

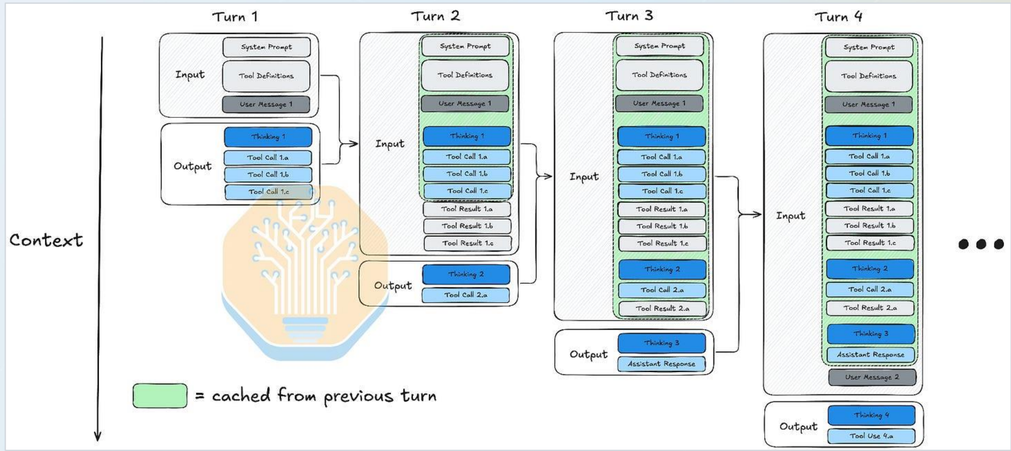
2026 | Confidential | All Rights Reserved

Agentic Coding at OpenAI

Codex accounts for 64% of OpenAI enterprise output tokens in the June 2026 snapshot



Data: share of OpenAI enterprise output tokens, ChatGPT vs agentic (Codex), 28-day trailing average, Aug 2025 to Jun 2026



Data: coding agents alternate model turns and tool calls. Green blocks reuse context from earlier turns

Source: SemiAnalysis, AgentX / InferenceX v3

2026 | Confidential | All Rights Reserved

AgentX Replays Real Coding Sessions

393 anonymized Claude Code sessions, selected from a corpus of 8,300 sessions



88K / 272K / 675K
p50/p90/p99 input tokens

393
sessions in AgentX v1.0

44%
sessions use subagents

1M
max context

Data: AgentX flamegraph of one session (turns, subagents, cached vs uncached input); ISL percentiles from DeepSeek V4 Pro 0813 runs; corpus 8.3K sessions, 3.4M requests, $3M+ spend (Aug 2026)

Source: SemiAnalysis, AgentX / InferenceX v3 · Hugging Face dataset

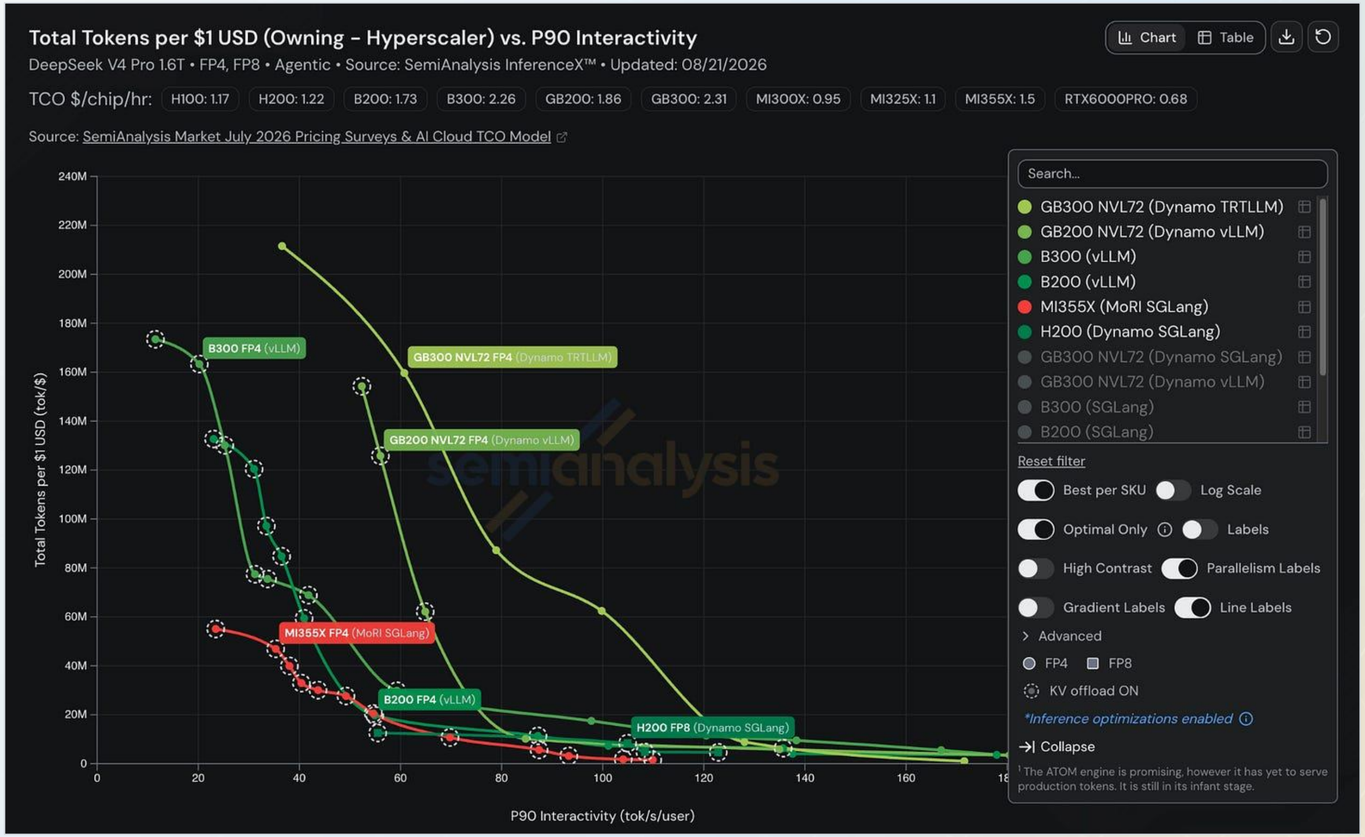
semianalysis

2026 | Confidential | All Rights Reserved

NVIDIA Leads Selected AgentX Results

August 2026 results. The advantage depends on the model, engine and latency target

- **20x+** NVIDIA vs AMD on SGLang, Qwen3.5 at 90 tok/s/user
- **Up to 5x** NVIDIA cost efficiency on SGLang, GLM 5.3 at 150 tok/s/user
- **+81.7%** output throughput from vLLM CPU KV offload
- GB300 NVL72 Dynamo leads tokens per $ on DeepSeek V4 Pro



Data: DeepSeek V4 Pro 1.6T · FP4/FP8 · AgentX · total tokens per $1 TCO (owning, hyperscaler) vs P90 interactivity · InferenceX, updated Aug 21, 2026

Source: SemiAnalysis, AgentX / InferenceX v3; Forbes

2026 | Confidential | All Rights Reserved

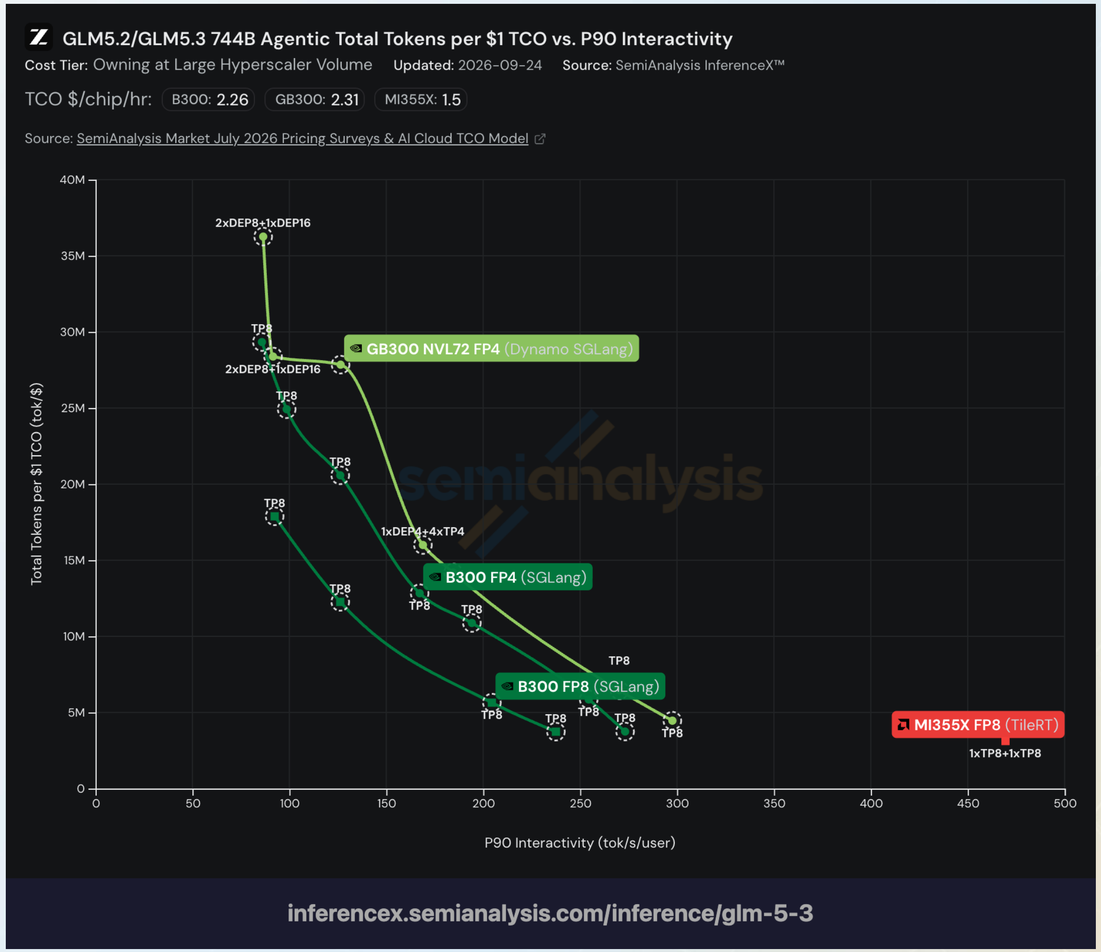
AMD + TileRT Reaches ~470 Tokens/s/User

GLM5.2 / GLM5.3 744B AgentX. P90 interactivity. September 24, 2026 benchmark view

~470 tok/s/user peak P90 interactivity on MI355X FP8

vLLM prefill + TileRT decode
The displayed GB300 curve uses Dynamo SGLang FP4

Higher TTFT in the source result
Peak streaming speed alone does not establish a cost-efficiency win

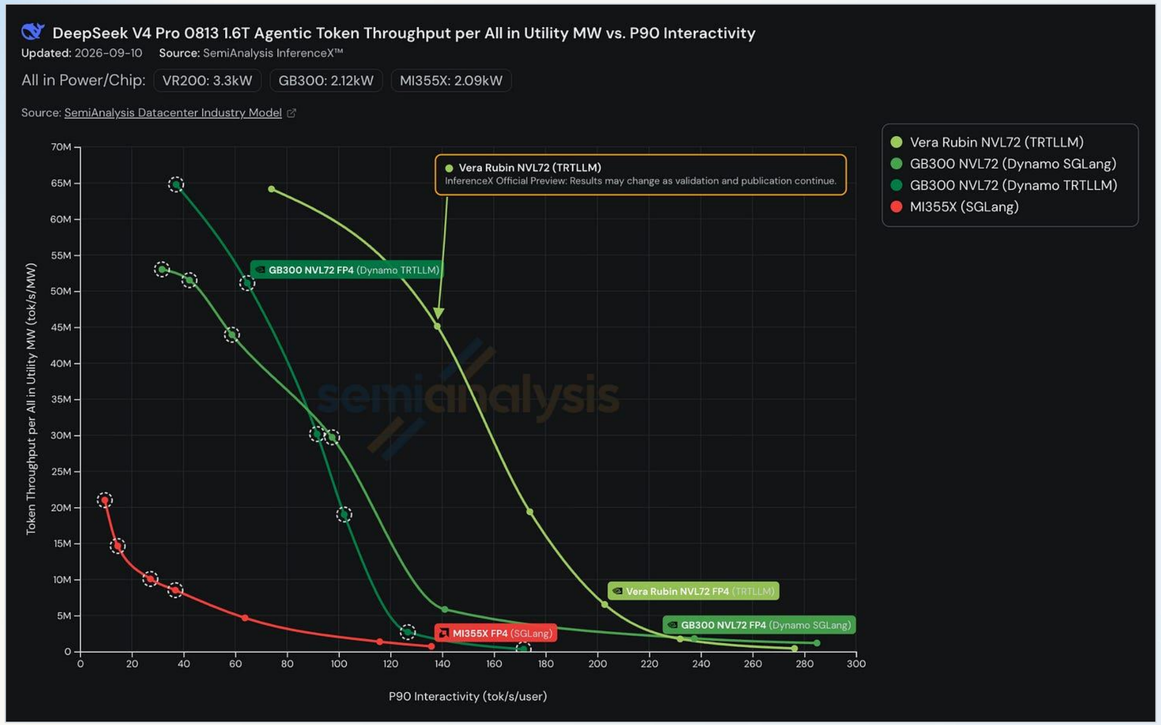


Actual benchmark graph · owning at hyperscale TCO
B300 $2.26 / GB300 $2.31 / MI355X $1.50 per chip-hour

Source: SemiAnalysis InferenceX · September 24 view, captured October 7, 2026

Rubin Delivers Up to 7.2x Tokens per MW

DeepSeek V4 Pro preview: 2.1x at 100 tok/s/user and 7.2x at 150, versus the best GB300 engine



Data: DeepSeek V4 Pro 0813 1.6T AgentX · tok/s per all-in utility MW vs P90 interactivity · modeled all-in power/chip VR200 3.3 kW, GB300 2.12 kW, MI355X 2.09 kW · updated Sep 10, 2026, preview

SemiAnalysis · InferenceX

Nvidia Rubin and GB300 at matched interactivity

Total throughput in million tok/s per all-in utility MW

DeepSeek V4 Pro 0813 1.6T · Agentic · FP4

P90 interactivity tok/s/user	GB300 NVL72 Dynamo SGLang million tok/s/MW	GB300 NVL72 Dynamo TRTLLM million tok/s/MW	Vera Rubin NVL72 TRTLLM million tok/s/MW	**Rubin / best GB300** Throughput ratio
75	37.94	44.14	64.00	1.45x
100	28.47	21.15	59.38	2.09x
125	12.01	2.97	51.55	4.29x
150	5.14	1.01	36.98	7.19x
170	3.92	0.35	21.78	5.56x
200	2.73	N/A	7.43	2.72x

Data: same run, throughput in M tok/s per MW at matched P90 interactivity · Rubin on pre-release software

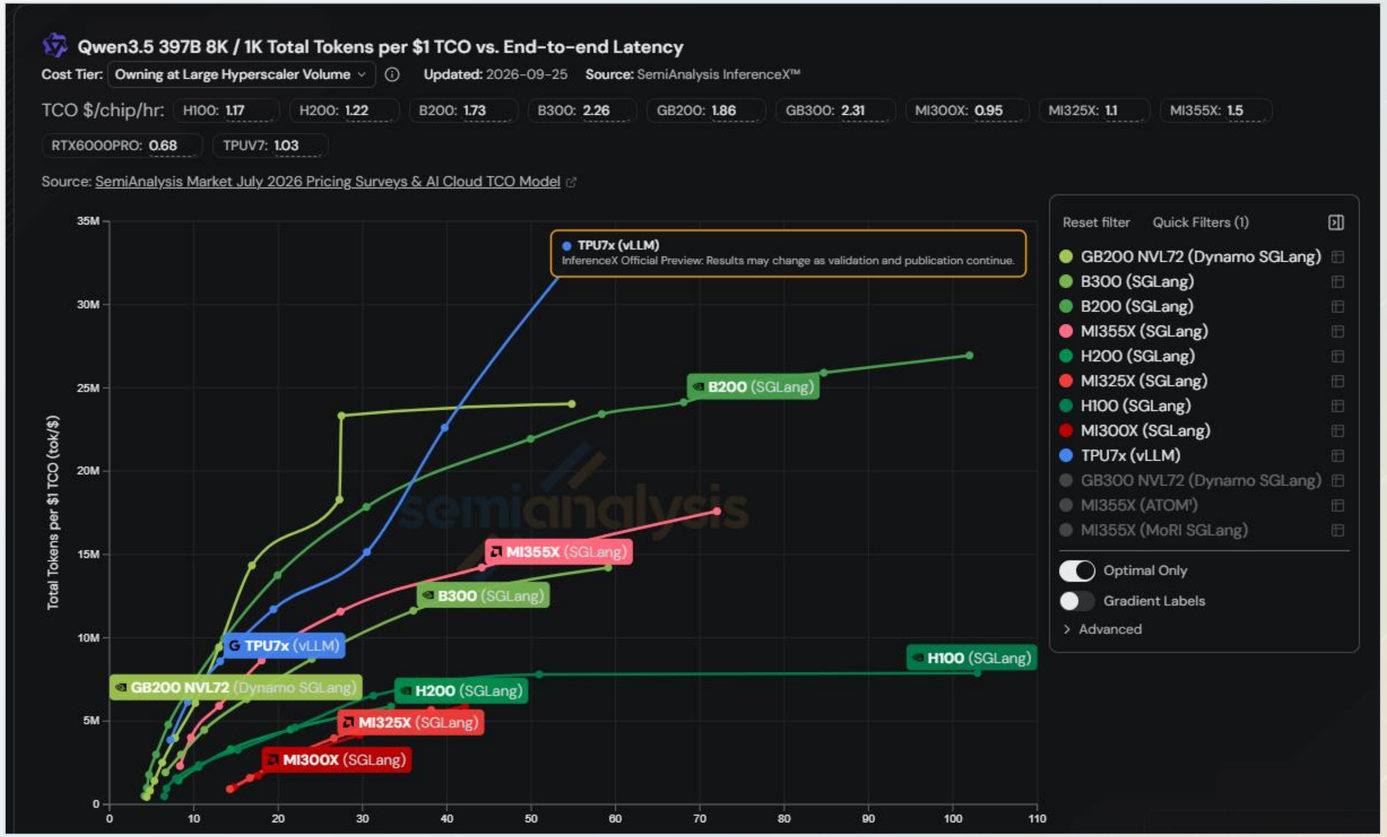
Source: SemiAnalysis, Vera Rubin NVL72 Agentic Inference

2026 | Confidential | All Rights Reserved

TPU, NVIDIA and AMD on One Chart

Qwen3.5 397B, FP8, 8K/1K. Internal TCO versus end-to-end latency

- **$1.03 per TPU chip-hour, using Google's internal TCO**
- **At low E2E latency, B200 and GB200 lead TPUv7**
- **At higher E2E latency, TPUv7 leads the displayed curves**
- **Same FP8 model and 8K/1K workload. Serving recipes differ by chip**



Data: Qwen3.5 397B · 8K/1K · FP8 · STP · total tokens per $1 TCO (internal TCO, large hyperscaler volume) vs E2E latency · dashboard updated Sep 25, 2026 · TPUv7 is a preview

Source: InferenceX dashboard; TPU InferenceX article

2026 | Confidential | All Rights Reserved

TPU Preview Scope and Next Steps

FP8 results today. Disaggregation, MTP and AgentX remain follow-up work

- **Matched precision:** The chart compares FP8 across accelerators. Blackwell FP4 results answer a different deployment question
- **Preview scope:** TPUv7 uses aggregated serving and STP. Disaggregated prefill/decode, MTP and AgentX remain future comparisons
- **Serving software:** TorchTPU supports native PyTorch with Pallas kernels. This TPU preview uses vLLM

EXTERNAL TCO, 100 TOK/S/USER

TPUv7 Ironwood, official preview
Qwen3.5 397B · FP8 · 8K/1K · STP
vLLM on TorchTPU
$0.181 per M total tokens vs $0.222 B200 and $0.276 B300

PLANNED COMPARISONS

TPUv8i: native FP4, compared FP4 vs FP4
Disaggregated prefill/decode
MTP
AgentX on TPUv7

Data: external TPU TCO at 100 tok/s/user, Qwen3.5 FP8, 8K/1K, STP. Source article: Sep 7, 2026

Source: TPU InferenceX article (Sep 7, 2026); Inferact TPU partnership post

2026 | Confidential | All Rights Reserved

What's Next on InferenceX

Coming next on the same methodology: Rubin NVL72, MI455X, TPUv8i, SN50 and Amazon Trainium3

Rubin NVL72

From preview to official results

MI455X UALoE72

Rack-scale AMD

TPUv8i

Native FP4; FP4 vs FP4 comparisons, and AgentX on TPUv7 first

SambaNova SN50

Listed "coming soon" on the InferenceX dashboard

Amazon Trainium3

Planned benchmark coverage. Performance results pending

Same open methodology for every new chip: public recipes, public CI, same TCO basis

Data: roadmap in the Oct 6 draft. Trainium3 is planned. No Trainium3 benchmark result is shown

Source: InferenceX · github.com/SemiAnalysisAI/InferenceX

2026 | Confidential | All Rights Reserved

THANK YOU

semianalysis.com

Scan to visit

semianalysis