Who is positioned to win AI inference?

Verdict

The AI-inference market produces multiple structurally durable winners through 2028, not a single monopolist. Confidence 0.60.

NVIDIA retains the largest absolute share of inference revenue (~80% in 2026 [S1]) and the deepest software moat (CUDA), and its acquisition of Groq for US$20bn in late 2025 extended its position into the speed-inference niche [S5]. But the inference share NVIDIA holds today is structurally lower than its training share, the slope is downward, and three independent forces compress it through 2028: (1) hyperscaler captive silicon (Google TPU v7, AWS Trainium 3, MS Maia 200, Meta MTIA) growing at 44.6% CAGR (compound annual growth rate) [S2][S3] takes hyperscaler-own workloads off NVIDIA; (2) AMD's MI355X / MI400 family is competitive on per-TCO inference and has won AWS + Meta as committed buyers [S6]; (3) specialist wedges (Cerebras for high-throughput, SambaNova for agentic / long-context) remain real but small. The most-cited analyst projection — that NVIDIA's *inference-specific* share could fall from >90% to 20-30% by 2028 [S2] — is the upper bound of the bear case; the more honest base case sits at 50-60% inference share in 2028, with NVIDIA still the largest single beneficiary in absolute dollars on a US$255bn 2030 inference TAM (total addressable market) [S7].

Confidence: 0.60 — moderate. This is a "multiple winners" verdict where the candidate moats (NVIDIA's CUDA, hyperscaler captive economics, AMD's open-stack alternative, Cerebras's wafer (thin polished disc of semiconductor (silicon, glass, III-V) on which chips are built)-scale) each defend a different slice of demand. The verdict does NOT meet the asymmetric-upside calibration (>=0.65 conviction requires >=2 independent paths to a 2–5×+ outcome) because the candidate winners are correlated via the broader AI capex (capital expenditure) cycle — if AI capex stalls, all of them stall together.

Decision boundaries

Specific, observable signals that would change the verdict. Falsifiable in 18 months.

- *(+)* If NVIDIA's inference market share is disclosed at >=75% through Q4 2027 (i.e., hyperscaler captive silicon under-delivers) -> conviction in "multiple winners" *drops* by ~0.10 (the "NVIDIA monopoly" case strengthens).
- *(+)* If hyperscaler in-house silicon is publicly disclosed as >=40% of the named hyperscaler's accelerator capex by Q4 2027 -> conviction in "multiple winners" rises by ~0.10 (the displacement is real and on track).
- *(+)* If AMD MI400 takes >=10% share of new AI-accelerator deployments by Q4 2027 -> conviction rises by ~0.05.
- *(+)* If a specialist (Cerebras, SambaNova, post-acquisition Groq inside NVIDIA) is disclosed at >=US$1bn ARR with >=3 non-UAE / non-SoftBank customers -> conviction in "multiple wedge winners" rises by ~0.05.
- *(-)* If CUDA-based deployment friction is disclosed as the binding constraint on >=2 hyperscaler-captive silicon ramps (i.e., they can't port frontier models off CUDA fast enough) -> conviction in "NVIDIA holds longer" rises (multiple-winners thesis weakens) by ~0.10.
- *(-)* If AI capex is publicly cut by >15% across >=2 hyperscaler guides -> the whole framework softens; conviction drops by ~0.10 (the rising tide that floats multiple winners ebbs).
- *(-)* If NVIDIA's gross margin holds >=75% through FY2027 -> indicates pricing power is intact and competitive pressure is overstated; conviction drops by ~0.05.

Steel-manned bull case

The strongest version of "multiple structurally durable winners":

1. NVIDIA's inference share is unwinding from a different starting point than its training share. Training is where the CUDA ecosystem is deepest — every frontier-lab model lands on CUDA first; the porting cost to alternative architectures is a quarter or more of engineering time per model. Inference is fundamentally different: the model is fixed, the workload is well-understood, and the question is cost-per-token at production scale. That's where NVIDIA's moat is the weakest. SemiAnalysis

InferenceMAX benchmarks have NVIDIA Blackwell B200 falling from US$0.11 to US$0.02 per million tokens in two months on software optimisation alone [S4] — impressive, but it also shows the metric IS economic and IS subject to rapid optimisation, exactly the metric where custom ASICs out-compete general-purpose GPUs.

2. The hyperscaler captive-silicon ramp is structurally different from the 2018-22 in-house silicon hype. This time the chips are shipping at volume. Google's internal TPU forecast for the next cycle has been revised from 2m units to ~4m [S3]. AWS Trainium 2 is the back-end for Anthropic's training; Trainium 3 ships in 2026-27 at 3nm. Microsoft Maia 200 claims 3× the FP4 of Trainium 3. Meta MTIA is ramping. The four hyperscalers committed US$650bn+ in AI infrastructure for 2026 alone [S4]; even if NVIDIA captures 80%, the remaining 20% is US$130bn — a market the size of NVIDIA's total revenue in FY2022.

3. AMD has graduated from "credible #2 narrative" to "design-win reality." The MI355X is competitive with NVIDIA HGX B200 on per-TCO inference for small / medium LLMs [S6]; the MI400 in H2 2026 with HBM4 (432GB) targets NVIDIA's VR200 NVL144 rack-scale. AWS was the title sponsor at AMD Advancing AI 2025; Meta is committed to MI355X + MI400 [S6]. That's not pre-announcement narrative — that's real production volume going to AMD because hyperscalers want a second source and have spent >=18 months porting their stacks.

4. The specialist wedges are narrow but real. Cerebras's OpenAI 750MW deal (US$10bn+ through 2028) [S8] anchors a workload pattern (single-user latency-sensitive inference) where NVIDIA can't compete on speed. SambaNova's SN50 (1.6 PFLOPS FP16, 10T parameter / 10M context support) won SoftBank for agentic workloads [S8]. NVIDIA's response — acquiring Groq for US$20bn in late 2025 [S5] — is an admission that the speed-inference niche is real and that NVIDIA preferred to buy rather than compete.

5. Inference is now 2/3 of AI compute [S2], and the AI inference market is projected at US$106bn (2025) -> US$255bn (2030) at 19% CAGR [S7]. The pie is big enough that a NVIDIA at 50-60% share and AMD + hyperscalers + specialists at 40-50% combined still makes NVIDIA the largest single beneficiary in absolute dollars — but no longer a monopolist.

Steel-manned bear case

The strongest version of "NVIDIA monopolises through 2028+":

1. CUDA is genuinely irreplaceable in production deployment. The hyperscaler captive silicon promise — "we'll port everything to Trainium / TPU / Maia" — has been promised every year since 2020 and consistently under-delivered. Anthropic's Trainium deployment took 18+ months to reach material volume; Microsoft's Maia is still tied to specific OpenAI workloads. New model architectures (next attention type, next MoE pattern) land on CUDA first; alternative architectures are perpetually 6-12 months behind. The 80% NVIDIA share isn't because hyperscalers want NVIDIA — it's because the porting cost is real, and software-engineering capacity at hyperscalers is finite.

2. NVIDIA's economic-floor is dropping faster than competitors can catch up. Blackwell B200 dropped to US$0.02 per million tokens in two months on software alone [S4]. The Rubin generation (2026-27) is a further architectural step. Custom ASICs improve on a 24-month silicon cadence; NVIDIA improves on a 6-month software cadence. The "AMD MI355X is competitive at TCO" point is true for *current* Blackwell; by the time MI400 ships in H2 2026, Blackwell-Ultra and pre-Rubin software optimisation may have reset the cost floor again.

3. Hyperscaler captive silicon shrinks NVIDIA's pie, but only for the hyperscaler's OWN workloads. Google TPU doesn't ship to anyone but Google Cloud customers (and inside Google). AWS Trainium doesn't show up at Meta. Each hyperscaler's silicon dent is isolated to its own footprint. The merchant TAM that's left for NVIDIA — enterprise, sovereign, frontier-lab non-captive — is enormous and is the part where CUDA dominates structurally.

4. AMD's design wins are real but the actual deployed-volume gap is wide. AWS as MI300X sponsor and Meta as MI355X buyer are real, but Microsoft, Google and Oracle are still NVIDIA-only at scale. AMD's 5-7% accelerator share in 2026 [S1] is a doubling vs 2023, but the slope of growth needed to reach 20%+ by 2028 is steeper than the named design-wins support.

5. NVIDIA's Groq acquisition removes the most threatening pure-inference specialist. Cerebras is concentrated (86% UAE-linked) and contractually constrained (OpenAI MRA exclusivity); SambaNova is single-customer (SoftBank). The serious specialist wedge — Groq's deterministic LPU + best-in-class per-user throughput — is now inside NVIDIA. The remaining specialists have narrower, structurally bounded markets.

Core claims

1. [confidence: 0.65] NVIDIA's inference share is structurally lower than its training share and is compressing

- Evidence: NVIDIA's overall accelerator share is ~80-85% in 2026 (down from ~92% in 2023) [S1]. The inference-specific share is higher today (~90%+) but analysts project it falls to 20-30% by 2028 [S2]. SemiAnalysis InferenceMAX shows Blackwell B200 cost-per-token compressing rapidly [S4], indicating the inference economics curve is itself moving — not just market share.
- Counter: The 20-30% number is the *most aggressive* bear projection; the consensus base case is closer to 50-60% NVIDIA inference share in 2028. NVIDIA's economic-floor improvements may outrun ASIC (application-specific integrated circuit) ramp.
- Resolution: The direction (down) is well-supported by data; the magnitude (severe vs moderate) depends on hyperscaler captive-silicon execution. The forward 18-month visibility is reasonable; the 2028 projection has wide error bars.

2. [confidence: 0.70] Hyperscaler captive silicon is the largest single threat to NVIDIA's share

- Evidence: Custom ASICs from Google (TPU v7 Ironwood), AWS (Trainium 3), MS (Maia 200), Meta (MTIA) are growing at 44.6% CAGR [S2][S3]. Google's internal TPU forecast doubled from 2m -> 4m units for the next cycle [S3]. The four major hyperscalers committed US$650bn+ in AI infrastructure for 2026 [S4]. Hyperscaler captive silicon is collectively a larger and faster-growing threat to NVIDIA than AMD is [S1].
- Counter: Hyperscaler in-house silicon has under-delivered relative to announcements every prior cycle. Anthropic on Trainium took 18+ months; Maia is OpenAI-specific. The captive silicon is for *the hyperscaler's own* workloads, not external customers.
- Resolution: The captive silicon takes share from NVIDIA only within the hyperscaler's own footprint, not the broader merchant market. The bear case strength: hyperscalers are also the largest single block of NVIDIA's customers, so their captive deployments directly cannibalise the largest single line item on NVIDIA's revenue.

3. [confidence: 0.55] AMD is the credible merchant #2 — not a #1 challenger, but a real second source

- Evidence: MI355X competitive with NVIDIA HGX B200 on per-TCO inference for small/medium LLMs [S6]; MI400 in H2 2026 with HBM4 432GB targets VR200 NVL144 [S6]. AWS title-sponsor at AMD Advancing AI; Meta committed to MI355X + MI400 [S6]. AMD share rose from ~3% (2023) to ~5-7% (2026) [S1].
- Counter: 5-7% is a doubling, but the slope to reach 20%+ requires Microsoft, Google, Oracle to all join — none have publicly committed at MI400 scale. The MI355X is NOT competitive at frontier-model rack-scale inference [S6]; it's competitive at the small / medium LLM tier, which is a smaller TAM.
- Resolution: AMD takes meaningful share but not parity. The realistic 2028 path is ~10-15% accelerator share, anchored by AWS + Meta + open-source community, with the GB200 / VR200 frontier tier still NVIDIA's.

4. [confidence: 0.60] Specialists have real but bounded wedges; NVIDIA's Groq acquisition is the key structura

- Evidence: NVIDIA acquired Groq for US$20bn in late 2025 [S5] — the strongest possible signal that NVIDIA itself views the speed-inference niche as both real and not adequately covered by Blackwell. Cerebras at US$510m FY2025 with OpenAI 750MW deal [S8]; SambaNova's SN50 won SoftBank [S8].
- Counter: The remaining specialists are concentrated (Cerebras 86% UAE; SambaNova essentially SoftBank-anchored). The Groq acquisition removed the most threatening pure-play; the survivors have narrow markets.
- Resolution: Specialists carve real wedges but combined likely <5% of inference TAM by 2028. NVIDIA's acquisition strategy suggests further consolidation: SambaNova or Cerebras post-IPO are reasonable M&A (mergers & acquisitions) candidates.

5. [confidence: 0.55] The "multiple winners" outcome requires AI capex to continue at its current trajectory

- Evidence: US$650bn+ committed AI infrastructure for 2026 across the four major hyperscalers [S4]. AI inference TAM US$106bn (2025) -> US$255bn (2030), 19% CAGR [S7]. Inference is now 2/3 of AI compute [S2].
- Counter: The hyperscaler capex commitments are guidance, not booked spend. A 2027 AI-cycle correction (frontier-lab over-investment, training-economics softening) could pull capex back hard. In that scenario, ALL the candidate winners face a contracting pie simultaneously.
- Resolution: The "multiple winners" verdict is conditional on the AI capex super-cycle continuing through 2028. If it doesn't, the verdict flips to "NVIDIA dominates a contracting market" — fewer winners but a worse environment for all.

Open questions

- [confidence: 0.4] What is the *actual* share of hyperscaler accelerator capex going to in-house silicon today — would need a T1 source: hyperscaler 10-K segment disclosure or capex breakdown.
- [confidence: 0.3] How much of NVIDIA's FY2026 revenue is "training" vs "inference" — would need a T1 source: NVIDIA Data Center segment breakout (NVIDIA does not currently disclose this split publicly).
- [confidence: 0.4] What is the realised cost-per-token gap between Trainium 3 / TPU v7 and Blackwell at parity workload — would need a T2 source: SemiAnalysis InferenceMAX cross-platform run.
- [confidence: 0.4] Does AMD MI400 actually compete at the GB200 NVL72 / VR200 NVL144 frontier rack-scale tier, or stays at the small/medium LLM tier — would need a T1 source: AMD MI400 commercial launch + a benchmark pass.
- [confidence: 0.3] Will NVIDIA acquire SambaNova or Cerebras post-IPO — would need event-driven evidence: announced M&A or strategic position changes.

Sources

Numbered references. Each entry carries its tier — T1 primary record / T2 quality secondary / T3 supplemental / T4 single-source flag.

- [S1] [T2] Silicon Analysts, "NVIDIA AI GPU Market Share 2026: ~80% of AI Accelerators" (overall accelerator share data; AMD share 5-7%; hyperscaler captive silicon framing) — https://siliconanalysts.com/analysis/nvidia-ai-accelerator-market-share-2024-2026
- [S2] [T2] Introl Blog, "Custom Silicon Inflection 2026: Hyperscaler ASICs vs NVIDIA GPU" (custom ASIC 44.6% CAGR; inference 2/3 of AI compute; analyst projection of NVIDIA inference share falling to 20-30% by 2028) — https://introl.com/blog/custom-silicon-inflection-2026-hyperscaler-asics-nvidia-gpu
- [S3] [T2] Tom's Hardware, "The custom AI ASIC state of play (May 2026) — Broadcom deals, Google TPUs, Meta MTIA & beyond" (Google internal TPU forecast 2m -> 4m units; AWS Trainium 3 specs; MS Maia 200 vs Trainium 3) — https://www.tomshardware.com/tech-industry/semiconductors/custom-ai-asics-examined-from-broadcom-to-mtia
- [S4] [T2] NVIDIA Perspectives, "Blackwell Breaks Ground: 15× Savings & Token Cost Revolution in 2026 AI" (Blackwell B200 US$0.11 -> US$0.02 cost-per-million-tokens in 2 months; US$650bn 2026 hyperscaler AI infra commitment) — https://perspectives.nvidia.com/real-cost-ai-scale-hyperscaler-accelerator-economics-2026/
- [S5] [T2] Fortune, "After Nvidia's Groq deal, these are the AI chip startups sitting pretty—and one aiming to disrupt" (NVIDIA US$20bn Groq acquisition late 2025) — https://fortune.com/2026/01/05/nvidia-groq-deal-ai-chip-startups-in-play/
- [S6] [T2] SemiAnalysis, "AMD Advancing AI: MI350X and MI400 UALoE72, MI500 UAL256" (MI355X TCO competitiveness vs HGX B200; MI400 H2 2026 rack-scale; AWS + Meta design wins) — https://newsletter.semianalysis.com/p/amd-advancing-ai-mi350x-and-mi400-ualoe72-mi500-ual256
- [S7] [T2] Spheron Blog, "Token Factory on GPU Cloud: Maximize Tokens per Watt for AI Inference Revenue (2026 Guide)" (AI inference TAM US$106bn 2025 -> US$255bn 2030 at 19.2% CAGR) — https://www.spheron.network/blog/token-factory-gpu-cloud-tokens-per-watt-guide/
- [S8] [T2] IntuitionLabs, "Cerebras vs SambaNova vs Groq: AI Chip Comparison (2025)" (specialist comparison; OpenAI-Cerebras 750MW deal; SambaNova SN50 SoftBank; Groq-NVIDIA acquisition context) —

https://intuitionlabs.ai/articles/cerebras-vs-sambanova-vs-groq-ai-chips
- [S9] [T2] SemiAnalysis, "AMD vs NVIDIA Inference Benchmark: Who Wins? Performance & Cost Per Million Tokens" (independent per-token cost comparison) — https://newsletter.semianalysis.com/p/amd-vs-nvidia-inference-benchmark-who-wins-performance-cost-per-million-tokens
- [S10] [T2] NVIDIA Technical Blog, "NVIDIA Blackwell Ultra Sets New Inference Records in MLPerf Debut" (GB300 NVL72 throughput-per-MW data, 50× vs Hopper) — https://developer.nvidia.com/blog/nvidia-blackwell-ultra-sets-new-inference-records-in-mlperf-debut/
- [S11] [T2] The Next Platform, "The Second Time Will Be The IPO Charm For Cerebras" (Cerebras IPO setup, market positioning) — https://www.nextplatform.com/compute/2026/04/22/the-second-time-will-be-the-ipo-charm-for-cerebras/5218651
- [S12] [T1] NVIDIA Corporation, "Blackwell InferenceMAX Benchmark Results" (NVIDIA primary disclosure of Blackwell inference performance + cost trajectory) — https://blogs.nvidia.com/blog/blackwell-inferencemax-benchmark-results/
- [S13] [T2] TheNextWeb, "Google launches Ironwood TPU and previews eighth-gen split into training and inference chips at TSMC 2nm" (Google TPU v7 Ironwood specs, training/inference split strategy) — https://thenextweb.com/news/google-ironwood-tpu-inference-cloud-next

---

Doctrine: see /principles. Calibration: see /conviction.

Appendix — methodology & sources

Generated by AutoLab (thesis mode) on 2026-06-02. The loop iteratively scouts the weakest point, researches it, red-teams it, and integrates the findings; . Headline confidence 0.60.