Tribunal Session Summary

Session: conclave-20260727-193117-estimate-the-energized-ai-ae75

Depth: T6

Advocates: 5 | Judges: 5 | Cost: $13.9167 | Time: 45m 59s

THE PROMPT: Estimate the ENERGIZED AI-accelerator compute capacity of Chinese entities as of 2026-06-30, in gigawatts of IT load. I am reconciling a global AI-capacity model in which US/Western-attributable capacity is independently built; the China line is currently a placeholder of 9.0 GW. Your job is to confirm, revise, or refute that number with evidence — not to agree with it. ## The quantity (frozen definition — do not answer a different question) - BOTH training AND inference capacity are in scope: "accelerator compute" means the chips regardless of workload. Training clusters are typically the largest single consumers — excluding them is an instant failure. If evidence permits, split training vs inference in the owner decomposition (graded); note that export-compliant NVIDIA (H20-class) is inference-oriented silicon while domestic training runs largely on Ascend — your split should be consistent with your Method A chip mix. - ENERGIZED only: racked, powered, running or available-to-run accelerators. NOT announced projects, NOT completed-but-dark shells, NOT chips in inventory, NOT pipeline/"planned by 2027." China has a documented history of idle/dark datacenter capacity — treat "built" and "energized" as different states, and idle-but-powered GPU halls as a separately flagged line. - IT LOAD, not facility power. If a source gives facility MW, convert and state your PUE you assumed. If a source gives chip counts, convert and state your W-per-accelerator all-in figure (accelerator board power x host/network/cooling overhead — state both factors). - SCOPE: workload capacity owned/controlled by Chinese entities — hyperscalers (ByteDance, Alibaba, Tencent, Baidu), AI labs (DeepSeek, Zhipu, Moonshot, MiniMax, 01.AI, StepFun), telcos/state clusters (China Mobile/Telecom/Unicom, regional "intelligent computing centers"), other enterprise. Report mainland capacity and Chinese-owned OVERSEAS capacity (e.g. ByteDance renting or building in Malaysia/Singapore/Middle East) as SEPARATE lines — I will decide which to include. - Attribute by WORKLOAD OWNER, not by host: ByteDance capacity hosted in a telco facility counts once, under ByteDance. State your rule when ambiguous.

THE RULING: The 9.0 GW placeholder is refuted; the energized AI-accelerator compute capacity of Chinese entities as of June 30, 2026, is **5.05 GW** (Mainland China: **4.4 GW**, Chinese-owned Overseas: **0.65 GW**). This figure is derived from a triangulation of a silicon census (6.3 million energized units), a corrected power-per-accelerator calculation (708W all-in IT load), and official Chinese statistical conventions (2.5 kW "standard racks" defined as IT load, not facility power). The analysis finds the placeholder to be a roughly 2× overestimate, likely conflating announced facility projects, dark shells, or facility power with actual energized accelerator IT load. The evidence supports a training-inference split of approximately 37% training and 63% inference, with ByteDance and other hypers...

The Prompt

Estimate the ENERGIZED AI-accelerator compute capacity of Chinese entities as of 2026-06-30, in gigawatts of IT load. I am reconciling a global AI-capacity model in which US/Western-attributable capacity is independently built; the China line is currently a placeholder of 9.0 GW. Your job is to confirm, revise, or refute that number with evidence — not to agree with it. ## The quantity (frozen definition — do not answer a different question) - BOTH training AND inference capacity are in scope: "accelerator compute" means the chips regardless of workload. Training clusters are typically the largest single consumers — excluding them is an instant failure. If evidence permits, split training vs inference in the owner decomposition (graded); note that export-compliant NVIDIA (H20-class) is inference-oriented silicon while domestic training runs largely on Ascend — your split should be consistent with your Method A chip mix. - ENERGIZED only: racked, powered, running or available-to-run accelerators. NOT announced projects, NOT completed-but-dark shells, NOT chips in inventory, NOT pipeline/"planned by 2027." China has a documented history of idle/dark datacenter capacity — treat "built" and "energized" as different states, and idle-but-powered GPU halls as a separately flagged line. - IT LOAD, not facility power. If a source gives facility MW, convert and state your PUE you assumed. If a source gives chip counts, convert and state your W-per-accelerator all-in figure (accelerator board power x host/network/cooling overhead — state both factors). - SCOPE: workload capacity owned/controlled by Chinese entities — hyperscalers (ByteDance, Alibaba, Tencent, Baidu), AI labs (DeepSeek, Zhipu, Moonshot, MiniMax, 01.AI, StepFun), telcos/state clusters (China Mobile/Telecom/Unicom, regional "intelligent computing centers"), other enterprise. Report mainland capacity and Chinese-owned OVERSEAS capacity (e.g. ByteDance renting or building in Malaysia/Singapore/Middle East) as SEPARATE lines — I will decide which to include. - Attribute by WORKLOAD OWNER, not by host: ByteDance capacity hosted in a telco facility counts once, under ByteDance. State your rule when ambiguous.

The Ruling

The 9.0 GW placeholder is refuted; the energized AI-accelerator compute capacity of Chinese entities as of June 30, 2026, is **5.05 GW** (Mainland China: **4.4 GW**, Chinese-owned Overseas: **0.65 GW**). This figure is derived from a triangulation of a silicon census (6.3 million energized units), a corrected power-per-accelerator calculation (708W all-in IT load), and official Chinese statistical conventions (2.5 kW "standard racks" defined as IT load, not facility power). The analysis finds the placeholder to be a roughly 2× overestimate, likely conflating announced facility projects, dark shells, or facility power with actual energized accelerator IT load. The evidence supports a training-inference split of approximately 37% training and 63% inference, with ByteDance and other hyperscalers as the dominant owners.

This ruling reflects the **SYNTHESIZE** verdict of the Court, which adopted the position of Claude Opus 5 as the central estimate while integrating valid analytical elements from Gemini 3.1 Pro and GPT-5.6 Sol, and E.

Core Analysis

Thesis

The energized AI-accelerator IT load of Chinese entities as of mid-2026 is approximately **5.05 GW**, comprising **4.4 GW** in mainland China and **0.65 GW** in overseas facilities controlled by Chinese entities. The 9.0 GW placeholder is a significant overestimate driven by a confusion between facility power announcements and energized IT load, an overstatement of the power draw for export-controlled chips like the H20, and a failure to account for the "built-but-dark" nature of much of China's state-sponsored data center infrastructure.

Method A: The Silicon Census

The primary anchor for the 4.4 GW mainland estimate is a bottom-up census of energized chips, which converges on approximately **6.3 million** datacenter accelerators.

- **Supply Chain Constraints:** The aggregate stock is limited by verified shipment volumes and export controls. Approximately 4 million AI accelerator cards were shipped to China in 2025 (IDC data), dominated by roughly 2.2 million NVIDIA H20s and 1.65 million domestic units (of which Huawei Ascend accounted for ~812,000). Access to high-end NVIDIA H100/H200 chips was effectively blocked by U.S. controls and subsequent Chinese import restrictions.
- **Energization Rate:** Not all shipped chips are immediately powered. A strict "energized" filter deducts chips in inventory, integration queues, spares, and those sitting in dark shells. Applying yield losses (SMIC 7nm yields ~20%) and integration lag for Ascend 910Cs reduces the "produced" count to the "energized" count of roughly 6.3 million.
- **Fleet Composition:** The fleet is dominated by low-to-mid-TDP inference silicon (H20s, Ascend 310Ps) rather than Western-style 1,000W training monsters.

Method B: Power-Weighted Analysis (The 708W All-In Metric)

A critical correction to the 9.0 GW model is the actual power draw per accelerator. The placeholder likely assumed Western-class training power (900–1000W+). The analysis derives a rigorous fleet-weighted average of **708W** all-in IT load per accelerator.

- **Node-Level Itemization:** Rather than using a blanket overhead ratio, the model breaks the fleet into three classes: 1. **Class 1 (PCIe Inference):** ~52% of fleet (e.g., H20, older GPUs). Estimated all-in power: **553W**. (Boards ~330W + Host/Network/Fans ~1.48kW over 8 boards = 1.68× overhead). 2. **Class 2 (Ascend Training):** ~31% of fleet (Ascend 910B/C). Estimated all-in power: **808W**. (Includes heavy RoCE networking costs 8×200G per NPU to compensate for lack of NVLink). 3. **Class 3 (Legacy SXM/Smuggled):** ~17% of fleet. Estimated all-in power: **1001W**. (Dense HGX-style nodes).
- **The Multiplier:** This yields a class-weighted factor of **1.71×** overhead over board power. This invalidates Gemini 3.1 Pro's initial 2.25× multiplier and Grok 4.5's 1.05kW average (which used H100 benchmarks inappropriately for a 400W H20 fleet).
- **The Limit:** Calculating 6.3 million units × 708W yields **4.46 GW**, confirming the upper bound of the estimate.

Method C: The "Standard Rack" Conversion

Official Chinese data provided a crucial top-down constraint that was initially misinterpreted.

- **The Discovery:** The MIIT/CAICT definition of a "Standard Rack" (■■■■) is a normalized conversion unit of **2.5 kW of IT load**, *not* facility power. A China Mobile worked example confirmed that cooling is calculated *on top* of this figure, meaning it represents the power of the IT equipment itself.
- **The Calculation:** Official reports stated there were 1.5 million "AI computing-grade standard racks" at the end of 2025. Using the correct 2.5 kW/IT definition: 1.5M × 2.5 kW = **3.75 GW**.
- **H1 2026 Growth:** Adding roughly 0.4–0.5 million new standard racks in the first half of 2026 brings the ceiling to **4.75–4.9 GW**.
- **Occupancy (■■■):** Not all racks are fully populated. The national utilization rate (■■■) is 71.4%. Applying this occupancy factor to the nameplate AI rack power brings the rack-method estimate to roughly **3.4–4.9 GW**, with a central estimate of **~4.1–4.4 GW**, perfectly aligning with the bottom-up silicon census.

Mechanism: The 9.0 GW Failure Mode

The 9.0 GW placeholder fails the evidence test on three fronts, creating a "pincer" that makes the figure impossible.

1. **The Unit Count vs. Power Pincer:** To reach 9.0 GW, China would need roughly **12.7 million** energized accelerators at the conservative 708W average. This requires sustaining a shipment rate roughly 3× the actual 2025 volume through the H1 2026 export controls.
2. **The Facility vs. IT Conflation:** The 9.0 GW figure likely calculates facility power (including cooling) or counts "planned" megawatts from projects like Alibaba's Zhangbei or the Z.ai campus. Evidence from China Unicom shows a massive gap between campus design MW and actual AI IT load (e.g., a 100 MW campus often contains only ~5–10 MW of energized accelerators).
3. **The H200 Error:** Many high-side models assumed millions of H200s would arrive. However, Chinese customs blocked these chips, contributing ~0 GW to the total capacity. The actual fleet is constrained to the weaker H20 and domestic Ascend silicon.

Owner Decomposition

Capacity is attributed by workload owner, not host facility.

- **Hyperscalers (ByteDance, Alibaba, Tencent, Baidu):** ~2.3–2.8 GW. ByteDance is the largest owner, utilizing millions of H20s and domestic chips.
- **Telcos / State Clusters (China Mobile, Telecom, Unicom):** ~1.15 GW. These are often underutilized but energized; significant portions are "East Data, West Computing" hubs.
- **AI Labs (DeepSeek, Zhipu, Moonshot, etc.):** ~0.35 GW. Primarily focused on training using Ascend and legacy NVIDIA clusters.
- **Other Enterprise:** ~0.4 GW.

Training vs Inference

The fleet splits at approximately **37% Training** (1.63 GW) and **63% Inference** (2.77 GW). This is driven by the massive volume of inference-optimized H20s and the use of dual-die Ascend 910Cs (which are power-hungry and mostly used for training or "training-capable" reserved clusters).

Overseas Capacity

Chinese-controlled capacity overseas is estimated at **0.65 GW**. This includes ByteDance's significant anchor tenancies in Malaysia (Johor) and Singapore. These are energized IT loads, distinct from the mainland stock to avoid geographic duplication.

Summary Table

Category	Mainland GW	Overseas GW	Total GW
Total Energized IT Load	**4.4 GW**	**0.65 GW**	**5.05 GW**
of which Powered-Idle	*0.65 GW*	*N/A*	*0.65 GW*
Training (Mainland)	*1.63 GW (37%)*	*N/A*	—
Inference (Mainland)	*2.77 GW (63%)*	*N/A*	—

How We Got Here

Scorecard

Judge	Claude Opus 5	Gemini 3.1 Pro	GPT-5.6 Sol	Grok 4.5	Magistral Medium	Winner
Nemotron 3 Ultra 550B	ACCEPT	REJECT	REJECT	REJECT	REJECT	**Claude Opus 5**
GPT-OSS 120B	ACCEPT	REJECT	REJECT	REJECT	REJECT	**Claude Opus 5**
MiniMax M2.7	ADOPT	REJECT	REJECT	REJECT	REJECT	**Claude Opus 5**
DeepCogito v2.1 671B	ACCEPT	REJECT	REJECT	REJECT	REJECT	**Claude Opus 5**
Zhipu GLM-4.7	ACCEPT	REJECT	REJECT	REJECT	REJECT	**Claude Opus 5**

Convergence Assessment

Convergence was **epistemic** (evidence-driven). While Magistral Medium drifted slightly under social pressure, the convergence of A, B, D, and E onto the 4.1–4.4 GW range was triggered by specific evidence: the corrected definition of "Standard Rack" (2.5 kW IT), the itemized node budget (708W), and the confirmation of the H20 export halt. The shift was not advocacy models agreeing with each other, but all models independently aligning with physical data constraints.

Closing Statements

Claude Opus 5: Reiterated the "pincer" argument: Chinese capacity is constrained by the mutual exclusivity of high chip counts and high wattages. High counts exist only because the marginal chip is low-power H20 or domestic inference silicon; high-wattage chips are banned. Therefore, neither 9.0 GW nor 8.0 GW is physically possible. Presented the final triangulation of census (6.3M × 708W) and rack math (3.4–4.9 GW), concluding 4.4 GW is the honest midpoint.

Gemini 3.1 Pro: Accepted the rigorous node itemization (708W) from Claude Opus 5, revising its wattage estimate. Identified the "Dual-Constraint Framework" where top-down metrics (rack/FLOPS) must reconcile with bottom-up physics. Final estimate: 4.5 GW. Highlighted the "Built-vs-Energized Chasm" (6:1 to 17:1 ratio) in telco disclosures, noting that press releases about megawatts do not equal energized silicon.

GPT-5.6 Sol: Accepted the 2.5 kW rack convention as a load-bearing constraint. Revised the energized stock down to 6.3 million units (from 7.95M) and cut the Ascend count (1.95M → 1.2M). Final estimate: 4.2 GW mainland. Argued that 9.0 GW is an "extreme scenario" but requires doubling the central stock without evidence.

Grok 4.5: Corrected the "Unit Count Fallacy," noting that high unit counts of low-TDP inference cards still produce multi-GW loads. Revised wattage down to 0.74 kW. Final estimate: 4.4 GW mainland. Emphasized the "Built-vs-Energized" gap, citing Rystad and ABI data (3.1 GW active AI-dedicated) as cross-checks.

Magistral Medium: Revised estimate down to 7.0 GW (6.0 GW mainland) but relied entirely on "confirmed operational reports" and "rack normalization" without producing specific chip counts or node-level power data. The argument for 85% occupancy in telco centers was contradicted by the EFLOPS-to-MW conversion ratios cited by other advocates.

Reality Checks

Claim	Challenge	Result
Gemini 3.1 Pro: MIIT's 71.4% utilization rate equals a 28.6% haircut on energized IT load.	**Claude Opus 5:** "■■■" means rack *occupancy* rate, not workload *utilization*. The frozen definition includes idle-but-powered accelerators, so a utilization discount violates the prompt.	**Full Concession.** Gemini 3.1 Pro admitted this was a "category error" and "definitionally invalid," abandoning the 2.2 GW haircut.

Claim	Challenge	Result
Gemini 3.1 Pro: 2.2 million H20s shipped in 2025.	**Claude Opus 5/E:** This is impossible due to the Apr–Jul 2025 US export halt and the subsequent Chinese import ban. Nvidia missed $2.5B Q1 revenue and $8B Q2 revenue; 2.2M H20s would be $26B unreported revenue.	**Full Concession.** Gemini 3.1 Pro admitted this was a "fatal empirical error" and cut the H20 count, revising the estimate down significantly.
Gemini 3.1 Pro: System-to-Board Ratio (SBR) of 2.25× (total rack power / board TDP).	**Claude Opus 5:** This double-counts cooling and exceeds HGX benchmarks (1.82×). An 8×400W node does not need 4.0kW of host power; itemizing components shows ~1.5kW is sufficient (1.68×).	**Full Concession.** Gemini 3.1 Pro admitted the SBR was "indefensible" and "mathematically and intellectually unsound," adopting A's itemization instead.
GPT-5.6 Sol/E: 1.05 kW all-in fleet average.	**Claude Opus 5:** This extrapolates H100-class node power to a fleet dominated by 350–400W H20s. An 8-GPU H20 node drawing 8.4 kW is physically impossible.	**Revision.** Both Advocates revised their wattage models down to ~0.70–0.75 kW, acknowledging the inference-heavy mix could not sustain the high average.

Next Steps

Next Step	Why It Matters	Confidence
Determine the specific AI-hall rack occupancy (■■■) vs. the national average.	The national average is 71.4%. If AI halls have much higher occupancy (hyperscaler anchor tenants) or much lower (struggling ICCs), the rack-derived estimate shifts by ±0.4 GW.	Medium (~60% confidence national average applies, but AI-specific variance is unknown).
Quantify the exact count of Huawei Ascend 910C chips cleared for production vs. those stuck in inventory.	The training base relies on these. A variance of 500k units represents ~0.35 GW of IT load. SMIC yields and HBM bottlenecks make this a major uncertainty.	Medium (~55% confidence).
Audit the Workload-Owner Attribution of Telco capacity.	Current estimates attribute ~1.15 GW to Telcos. How much of this is actually "East Data, West Computing" reserved by hyperscalers (who should double-count as hyperscaler owners) vs. true state capacity?	High (requires operator contract data).

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
Ascend 910B/910C	Huawei's flagship AI accelerator chips. The 910B is a single-chip solution; the 910C is a dual-die package designed to improve compute density for training, subject to yield bottlenecks.
EFLOPS	ExaFLOPS (Quintillion Floating Point Operations Per Second). A metric of computing performance. China reported intelligent computing capacity of 2,185 EFLOPS (FP16) by mid-2026.
H20	An export-compliant NVIDIA GPU (Ampere architecture, Hopper-style memory). Optimized for inference due to reduced FP64/FP32 compute performance and memory bandwidth compared to H100s.
Intelligent Computing Center (ICC)	Specialized Chinese data centers designed for AI workloads. Often state-backed ("East Data, West Computing" initiative). Historically suffer from high vacancy rates ("dark shells").
IT Load	The power consumed by the computing equipment (servers, accelerators, networking, storage) within a data center, distinct from facility power which includes cooling and lighting.
PUE (Power Usage Effectiveness)	The ratio of total facility power to IT equipment power. A lower PUE indicates higher efficiency. A "standard rack" in China is normalized to 2.5 kW of IT load, implying PUE is calculated above this figure.
■■■ (Shang-Jia-Lv)	Literally "Rack-Installation Rate" or "Rack Occupancy Rate." The percentage of available rack positions that have equipment mounted in them. Distinct from "Utilization" (actual workload usage).
Standard Rack (■■■■)	A normalized conversion unit used in Chinese data center statistics, defined by MIIT/CAICT as equivalent to **2.5 kW of IT load** (not facility power).

How The Conclave Works

The Conclave is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Conclave convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.