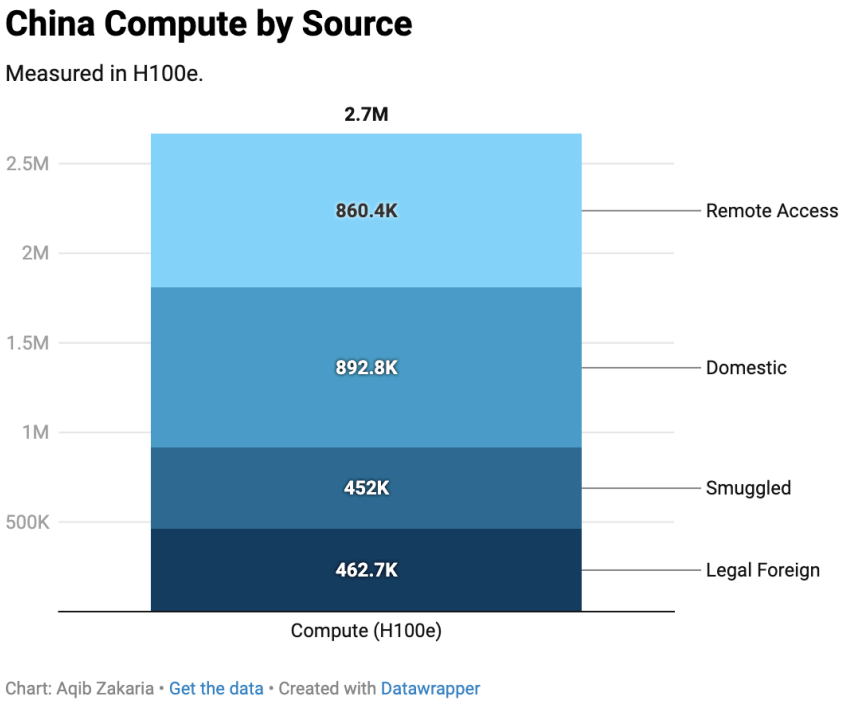
China’s AI Compute Demand

1. Introduction

1.1 Purpose

This document presents a demand-side back-of-the-envelope calculation of China’s AI compute consumption. It is a companion to a separate supply-side analysis by Aqib Zakaria tallying China’s installed chip inventory (legal Nvidia purchases, domestic production, and smuggled hardware).

The supply-side calculation arrived at **~2.7 million H100-equivalent GPUs**.



The supply-side method is higher fidelity. The demand-side method pursued here carries substantially more uncertainty and guesswork but could provide a useful second method: rather than asking ‘how many chips does China have,’ it asks ‘how many chips does China need.’

1.2 Two Pools, One Framework

- Inference pool (H100e running continuously): H100-equivalent GPUs that must be active at any moment to serve China’s AI workloads — chatbots, enterprise APIs, video generation, recommendation systems, VLMs, and more. ‘Continuously running’ means averaged across the full day and year, accounting for overnight troughs and daytime peaks.
- Training pool (dedicated H100e in training clusters): A separate count of chips reserved primarily for model training. Training is episodic: a cluster runs flat-out for weeks, then

sits idle, so a point-in-time snapshot systematically undercounts it. Calculated from annual GPU-hours consumed, then converted to an implied dedicated cluster size. These chips are predominantly but not exclusively distinct from inference chips — high-interconnect hardware (H800s, smuggled H100s/B100s) dominates training, while H20s dominate inference, but repurposing across pools is common.

Logistics: This BOTEC estimates an early 2026 annualized run-rate. Section 5 projects forward to end-2026 across a range of growth scenarios (2×, 5×, and 10×). All H100e figures use INT8 TOPS as the benchmark (H100 SXM5 = 3,958 INT8 TOPS).

Scope: This framework counts datacenter-class inference and training demand only. Edge inference is excluded, not because it is negligible but because it runs on hardware outside the H100e reference class. The supply-side estimate this BOTEC is compared against also counts only datacenter-class chips, so the comparison remains apples-to-apples.

1.3 Key Finding

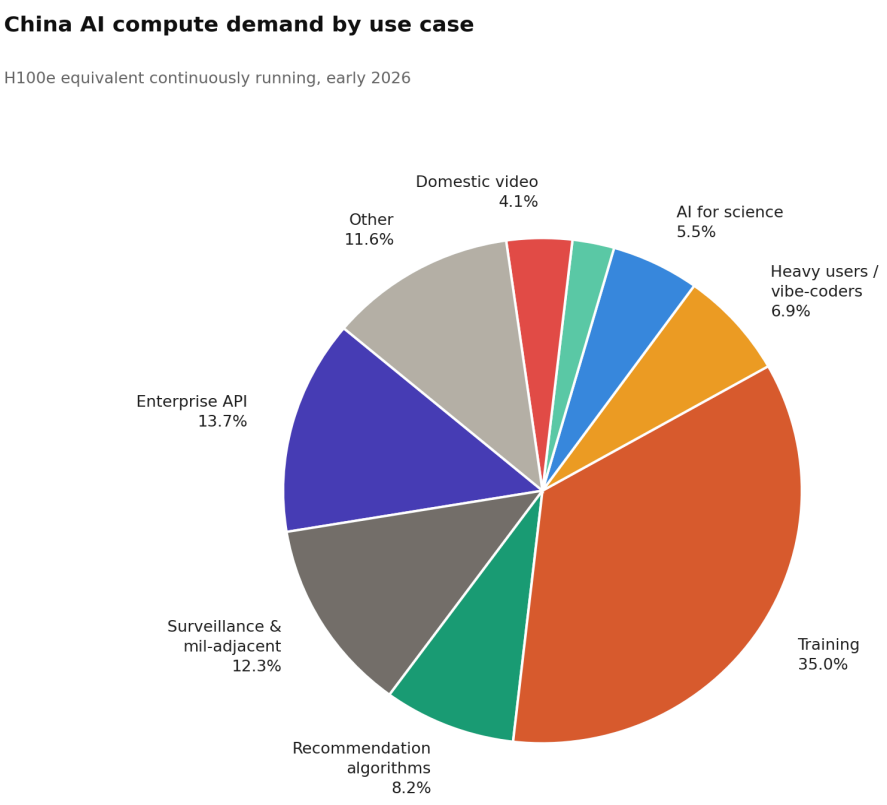
Current run-rate (early 2026): China's AI infrastructure currently requires approximately 237K H100e running continuously for inference workloads. At 55% central inference-serving utilization, this implies an inference installed base of ~431K H100e. Adding the dedicated training pool (~128K H100e) gives a central unadjusted installed-base estimate of ~558K H100e., meaning the sum total of AI chips currently in China could be **~2.8 million H100e at 20% whole-fleet utilization.** These numbers are then extrapolated to account for a range of growth scenarios (2×, 5×, and 10×) and degrees of Western cloud access usage (15-40%).

Key caveats: (1) Any individual estimate could be wildly inaccurate, but the most pivotal estimates are utilization rate, enterprise AI usage, non-frontier model training (i.e. R&D, wacky training projects), and the use of Western cloud services. Furthermore, the H100e metric obscures a qualitative constraint that raw equivalence numbers cannot capture: China is severely limited in access to the highest-end training hardware, where interconnect bandwidth and memory capacity matter most for frontier model runs. An H100e figure built largely from H20s and Ascend 910Cs is not the same as one built from B100s, even if the TOPS math balances out. China may have enough H100e for inference at scale while still facing a meaningful ceiling on how fast it can train the next generation of frontier models. Not all FLOPs are created equal.

1.4 Summary: Inference Components

Workload	H100e Continuous	Share	Notes
T1 Casual inference (domestic)	~1,000	0.4%	
T2 Professional inference (domestic)	~8,300	3.5%	Queries with a fair amount of reasoning
T3 Heavy users / vibe-coders (domestic)	~25,000	10.6%	Long context, agentic sessions

T4 Enterprise API (domestic)	~50,000	21.1%	revised to ~40% of US enterprise usage; see comparison to US usage
Other platforms / long tail (domestic)	~50	<0.1%	
Intl LLM on Chinese servers	~8,700	3.7%	DeepSeek (~80M intl MAU), Qwen intl, others; discounted 50% — Western hyperscalers serve a meaningful share of international DeepSeek inference
Intl enterprise API (Alibaba Cloud / Huawei Cloud)	~800	0.3%	~10K intl orgs; SE Asia, Middle East, Africa
Intl video — Kling (identified, closed-source)	~4,200	1.8%	7M intl MAU (60% of 12M MAU) × 2 vids/wk; all on Chinese servers
Intl video — buffer (80% of Kling)	~3,400	1.4%	Hailuo, Wan endpoints, other Chinese video tools; excludes Seedance (domestic only)
Recommendation systems (24/7)	~30,000	12.7%	Douyin, Taobao, Kuaishou; continuous. Much recsys runs on edge/specialized hardware, not H100-class silicon.
Domestic video generation	~14,900	6.3%	10M users × 5 vids/wk
Domestic image generation	~12,500	5.3%	50M users × 20 imgs/day
VLMs + multimodal inference	~7,000	3.0%	3–5× per-query cost vs. text
AV cloud-side training / simulation	~6,000	2.5%	Apollo, WeRide, Pony.ai, Huawei ADS
Miscellaneous (surveillance, mil-adjacent, industrial)	~45,000	19.0%	Surveillance AI alone is large; military usage also probably quite large
AI for science and other research	~20,000	8.4%	Protein folding, drug discovery, capability evals, etc.
TOTAL INFERENCE	**~236,850**	**100%**	
Training pool (dedicated cluster, separate)	~127,568	—	Episodic; separate pool



China AI compute demand by use case, in continuously-running H100e equivalents.

1.5 Chip Conversion Factors (INT8 TOPS)

Chip	INT8 TOPS	H100e	Notes
H100 SXM5 (reference)	3,958	1.0×	Benchmark for all estimates
H800	~2,000	~0.51×	Export-controlled China variant; dominates training clusters
A800	~1,200	~0.30×	Older generation; declining share
H20	~296 (dense) / ~592 (sparse)	~0.15×	Conversion uses sparse INT8. Higher memory bandwidth (4.0 vs 3.35 TB/s) lifts effective LLM-serving throughput to ~0.18× H100e — better than raw TOPS ratio, but well below H100 capability. Dominates legal inference pool.
Ascend 910B	~512	~0.13×	Huawei domestic; large installed base
Ascend 910C	~900 (est.)	~0.23×	2025 volume shipments
Smuggled H100 SXM5	3,958	1.0×	FT: $1B+ smuggled in one quarter 2025

Smuggled B200	~9,000	~2.25×	Highest-end hardware; concentrated in frontier training after being smuggled in

2. User Headcounts

2.1 Domestic Chinese Users

Two data sources bracket the domestic user population:

- CNNIC (February 2026): 602 million generative AI users as of December 2025, a 141.7% YoY increase. This counts anyone who used an AI-powered feature at least once — an upper bound, not a measure of active compute-generating users.
- Microsoft AI Economy Institute (January 2026): Approximately 20% of China's ~900M working-age population actively using AI tools, roughly 180M active users. Stricter definition; more useful for compute estimation.

Working assumption: 180–240M regular active AI users (T2 or higher) generating meaningful per-query compute load; central estimate ~207M. T1 (300M) represents casual daily AI feature contact — a lighter-weight population that overlaps substantially with the '300–400M who touch AI features occasionally' category and contributes minimal compute per user (~1,000 tokens/day). Total individuals with any AI touchpoint: ~500M+; total individuals in the tier table generating meaningful inference load: ~207M. Platform anchors: Doubao ~60M MAU (largest app); Qwen ~30M MAU; Yuanbao ~20–30M MAU. A long-tail buffer of ~15M users is added for smaller platforms not individually tracked — Baidu Ernie variants, 360 AI, Huawei Celia/Xiaoyi, vertical AI tools (legal, medical, education), and smaller coding assistants. At T1 intensity this adds only ~50 H100e, meaning the long tail is computationally negligible.

2.2 International Users on Chinese Infrastructure

International users of Chinese-hosted AI products are a larger and more important compute category than most analyses recognize. Their queries often run on Chinese servers (though they sometimes take the open source weights and run them locally), adding meaningfully to infrastructure demand beyond domestic user counts. However, Western CSPs also service Chinese AI international workloads through their data centers.

Consumer LLM Users

Platform	Intl Users (est.)	Server Location	Notes
DeepSeek (app + web)	~80M intl MAU	China (confirmed in privacy policy)	~68% of projected 100M global MAU; India 20%, Indonesia 8%
Qwen international app	~5–10M MAU	Alibaba Cloud (mix)	Launched late 2025; growing
Kimi, MiniMax, others	~5–10M MAU	China/mixed	Smaller but real presence

Chinese LLMs via HarmonyOS	~10–15M MAU	China	Preloaded on Huawei phones in Global South

Total international consumer LLM users hitting Chinese servers: ~100–120M MAU, central estimate 110M. These users skew heavily toward casual/T1 usage — lower-income markets (India, Africa, Southeast Asia), fewer power users. Intensity: 80% T1-equivalent (800 tok/day), 15% T2-equivalent (8K tok/day), 5% T3-equivalent (100K tok/day). This is also discounted by ~50% for the amount of these workloads that are potentially serviced through Western CSPs.

International Enterprise API

A meaningful but opaque category: international organizations accessing Chinese-hosted AI APIs through Alibaba Cloud International, Huawei Cloud, or direct API access to Chinese model providers. Alibaba Cloud has significant enterprise penetration in Southeast Asia, the Middle East, and Africa. Estimated ~10,000 international enterprise organizations at ~3M tokens/org/day = ~1,600 H100e continuously, discounted 50% to ~800 H100e to account for Western hyperscaler routing. But international B2B API volume is almost entirely opaque.

International Video Generation

Kling (Kuaishou) is the only Chinese video model where international compute can be quantified with a modicum of confidence: closed-source, majority international user base (36M of 60M registered users, 12M MAU globally), all inference on Chinese servers. Using the 12M MAU figure, approximately 7M are international (~60%), generating ~4,200 H100e continuously. An 80% buffer (~3,400 H100e) is applied on top to capture international load from Hailuo (MiniMax), Wan via Alibaba Cloud endpoints, and other Chinese video tools with international followings. Seedance is excluded from this buffer — it is gated behind Chinese phone verification internationally and is already captured in the domestic video bucket. Total international video: ~7,600 H100e continuously.

2.3 Domestic Enterprise API (T4)

Enterprise API is one of the largest inference categories — 50K H100e continuously, 21.1% of the total inference — and the most opaque. The estimation approach is: anchor on the US enterprise number, then take a fraction based on relative enterprise AI depth.

US anchor: OpenAI's 2025 Enterprise AI Report [pdf] discloses ~200 orgs exceeding 1T tokens/year and ~9,000 exceeding 10B tokens/year. Treating 8,800 orgs at the 10B floor and 200 at the 1T floor yields ~288T tokens/year — at 0.8M tokens/H100e-hour, that is ~41K H100e continuously. This is a floor on OpenAI alone. Adding Microsoft (15M paid Copilot seats as of January 2026, up 160% YoY), Anthropic, Google, Amazon, and non-public deployments puts plausible total US enterprise AI demand in the low hundreds of thousands of H100e.

China fraction: China's enterprise AI ecosystem is real and growing fast — Alibaba Cloud's Tongyi platform reached 300K enterprise customers in 2024 — but is earlier-stage than the US. Enterprise software integration is less deep, per-org compute budgets are lower, and no

Chinese provider has disclosed token volumes comparable to OpenAI's. There are also less suitable customers (as a proportion of population) in China than the international userbase of Western companies. I believe a reasonable prior is that China's T4 sits at roughly 40% of total US enterprise demand. At approximately 40% of US enterprise usage and using a US total well above the 41K floor, 50K H100e is the central estimate for China. The 40K–65K zone is the defensible base-case band; 80K+ requires stronger China-specific anchors than currently available.

3. Other Workloads

- Recommendation systems (30,000 H100e): Douyin, Taobao, Pinduoduo, Kuaishou, and Meituan run continuous real-time ranking inference 24/7 with no overnight drop-off. The H20 can work for recommendation use cases. However, much of this demand is served on heterogeneous infrastructure — edge inference, CPUs, older GPUs, and specialized/custom chips — rather than scarce H100/H20-class accelerators. Range: 10K–35K H100e; though, given how large and technocratized China's social media ecosystem has become, I'm going to follow my hunch and use a higher estimate of 30K.
- Domestic video generation (14,900 H100e): 10M domestic video users × 5 videos/week × 0.05 H100-hours per clip. Platforms: Kling domestic, Wan (Alibaba), Hailuo (MiniMax), Vidu (Shengshu), Seedance (ByteDance).
- Domestic image generation (12,500 H100e): 50M users × 20 images/day × 0.0003 H100-hours per image.
- Domestic VLMs and multimodal inference (7,000 H100e): 3–5× per-query cost vs. text chat due to the vision encoder. Fast-growing; likely undercounted.
- AV cloud-side (6,000 H100e): Baidu Apollo Go, WeRide, Pony.ai, and Huawei ADS run substantial simulation and retraining pipelines. Edge inference inside vehicles uses specialized low-power silicon (Horizon Robotics, Nvidia Orin) not expressible in H100e terms, nor too much compute.
- Miscellaneous (45,000 H100e): Government surveillance AI (Hikvision, SenseTime, Megvii — real-time video analytics across hundreds of millions of cameras continuously, using a mix of edge NPUs and data-center-class chips for backend aggregation and reidentification), military-adjacent AI, HPC at national labs, industrial quality control at factory scale, financial trading systems. Military AI is probably significant but completely unquantifiable. The elevated figure (vs. a naive bottom-up sum) deliberately factors in unknown unknowns — dark pools of state and military demand that leave no public trace. Range: 10K–65K H100e.
- AI for science and research (20,000 H100e): Protein structure prediction and drug discovery at biotech firms and national institutes; materials science and chemistry simulation; AI capability evaluation and red-teaming at frontier labs; scientific computing at CAICT and CAS institutes; neuroscience simulation including whole brain emulation-adjacent research. Central estimate 10K–30K H100e; central 20K.

4. The Training Pool

4.1 Framework and Two-Pool Separation

Training compute cannot be folded into the continuous H100e inference framework because training is episodic. At any random snapshot, most training clusters are between runs, making a point-in-time demand figure systematically too low. The appropriate unit is annual GPU-hours consumed, converted to an implied dedicated cluster size.

The training pool is dominated by different chip types than the inference pool — H20s are poorly suited for frontier training, so training skews toward H800s, smuggled H100s, and Ascend 910Cs. Past training runs are forensic evidence of chip presence in China, not demand to double-count; the chips remain after a run ends and are available for other workloads.

4.2 Annual Training Compute Budget

The anchor is DeepSeek V3's reported 2.788M H800-hours, adjusted upward by 1.5× for underreporting, as I believe Chinese labs have strategic incentives to minimize disclosed compute, since it reinforces the efficiency narrative and understates chip needs under export control scrutiny. The exact calculation is 2.788M × 0.51 × 1.5 = ~2.13M H100e-hours; this is rounded down to ~2.0M as a conservative simplification used consistently throughout the training table below. Other major labs are assumed less efficient than DeepSeek, costing 2–3× more for equivalent model quality. This is not because they are inferior companies, but because training V3 seemed to be abnormally efficient. The lab count is expanded to three tiers reflecting China's full pre-training ecosystem.

Component	Methodology	H100e-hours/year
Adjusted anchor per frontier LLM run	DeepSeek V3: 2.788M H800-hrs × 0.51 H100e/H800 = 1.42M; × 1.5 underreporting adjustment → 2.0M H100e-hrs per run	2.0M per run
Tier A: 5 major labs (ByteDance, Alibaba, Baidu, Huawei, Tencent)	5 × 2.5× anchor × 1.5 runs/yr × 4× overhead (ablations, RLHF, post-training, failed runs) = 5 × 2.5 × 2.0M × 1.5 × 4	~150.0M
Tier B: 10 mid-tier labs	10 × 1.0× anchor × 1.2 runs/yr = 10 × 1.0 × 2.0M × 1.2	~24.0M
Tier C: 5 state/military-adjacent actors	5 × 0.5× anchor × 1.0 run/yr; explicitly a floor	~5.0M
LLM subtotal (Tiers A + B + C)		~179.0M
Video model training (5 major models)	5 × 12M H100e-hrs per model per year	~60.0M
Image model training (10 models)	10 × 0.5M H100e-hrs per model per year	~5.0M
AV simulation training (4 companies)	4 × 1M H100e-hrs per company per year	~4.0M
Research & experimentation	Architecture experiments, scaling law runs, RL paradigm research, synthetic data; across major labs	~50.0M

TOTAL		**~298.0M H100e-hours/year**

Converting to a dedicated cluster: training clusters run approximately 4 months per year at 80% utilization when active. Available hours per chip: (4/12) × 8,760 × 0.80 = 2,336 hours/year. Dedicated training pool = 298.0M ÷ 2,336 = ~127,568 H100e. Training clusters are assumed to run at 80% utilization when active, giving (4/12) × 8,760 × 0.80 = 2,336 hours/year. This includes ~50M H100e-hours of research and experimentation compute beyond production training runs (architecture tests, scaling law experiments, RL paradigm research, synthetic data runs).

Note: LLM training (179M H100e-hrs) still exceeds video training (60M H100e-hrs) in aggregate. But individual video model runs (~12M H100e-hrs each) are larger than individual frontier LLM runs (~2–5M H100e-hrs each), reflecting the per-run compute intensity of spatiotemporal diffusion training.

5. Synthesis: Installed Base and 2026 Projection

5.1 Two-Pool Architecture

The total installed base is the sum of two predominantly distinct pools: the inference pool (serving users continuously and the training pool serving labs episodically). In practice, the boundary is not impermeable — training-class chips can be repurposed between training and inference over time. The table below shows the unadjusted installed-base totals.

5.2 Current Run-Rate: Two Approaches

Two steps are necessary to estimate the installed base.

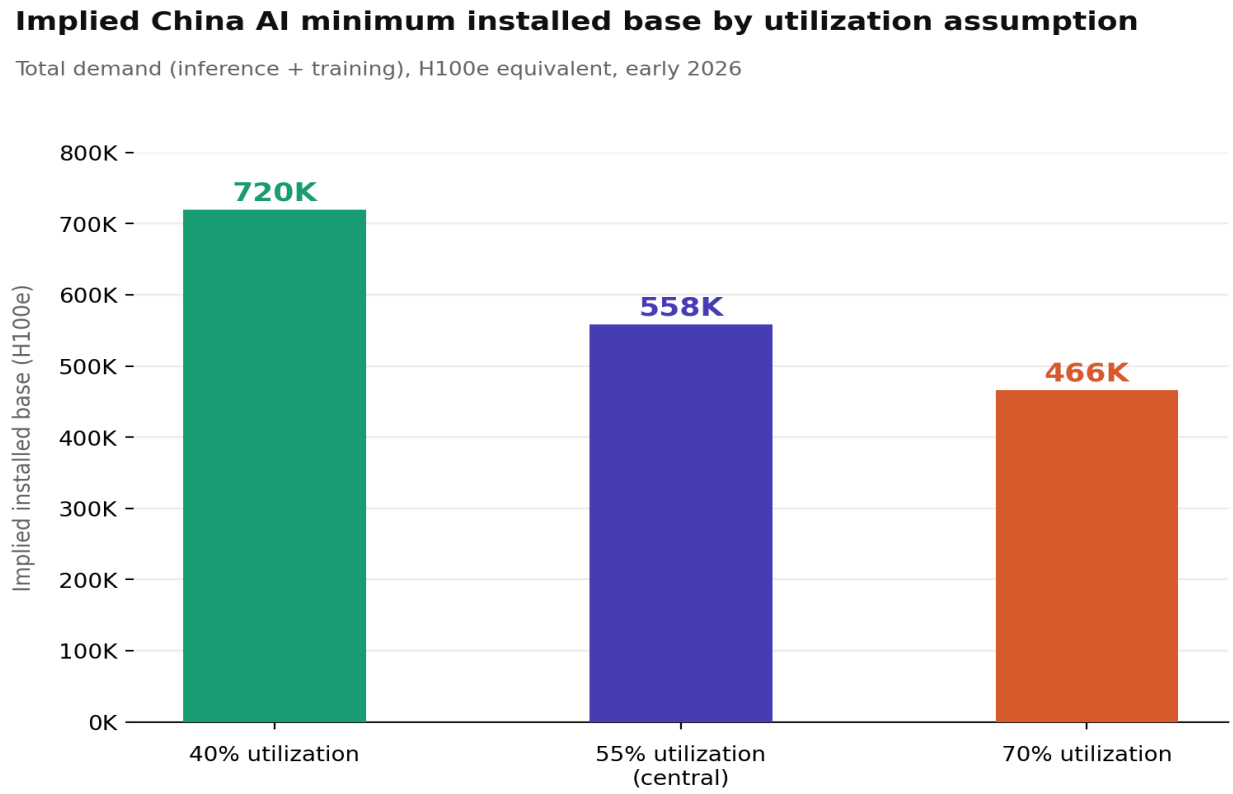
Step A — Bottom-up (first principles): Build from user headcounts and workload estimates to derive continuous H100e demand (~237K), then convert to installed base via an inference-serving utilization rate. Central estimate uses 55% utilization — the midpoint of the 40–70% plausible range I estimate based on conversations with experts and the common wisdom of the research on the topic — giving ~431K inference installed base plus ~128K training = ~558K total. **This rate measures what fraction of deployed inference chips are processing requests at any given moment** — distinct from whole-fleet utilization. Consumer inference averages ~35–45% due to factors like overnight troughs, enterprise batch runs higher, blended central estimate is 55% — the midpoint of the 40–70% range shown in Table 4. This is a debated parameter: hyperscalers who know what they are doing could get inference clusters approaching 60–70%, but, in aggregate, I believe 55% represents a reasonable middle ground. The range shown in Table 4 (40%–70%) captures the plausible spread.

Step B — Top-down (whole-fleet): The fraction of all purchased chips actively serving any workload at a given moment. This is the more relevant framing for export control policy, since it directly connects procurement to utilization. Whole-fleet utilization is structurally much lower than inference-serving utilization because it includes: chips recently procured but not yet deployed, newly deployed clusters ramping up (often taking 6–12 months to reach target

utilization), training hardware sitting idle between runs, and reserve capacity maintained for surge demand, amongst other factors. This parameter is highly debated, but the range used here (10–30%, central 20%) reflects the reality that a significant fraction of China's rapidly-expanding AI infrastructure is at any given moment in transit, configuration, or ramp-up rather than serving active workloads.

Step A results:

Pool	H100e Continuously Running	Installed Base @ 40% util	Installed Base @ 55% util (central)	Installed Base @ 70% util
Inference (all workloads)	~236,850	~592,125	~430,636	~338,357
Training (dedicated cluster)	~127,568 (separate metric)	~127,568	~127,568	~127,568
TOTAL (unadjusted)	**—**	**~719,693**	**~558,204**	**~465,925**

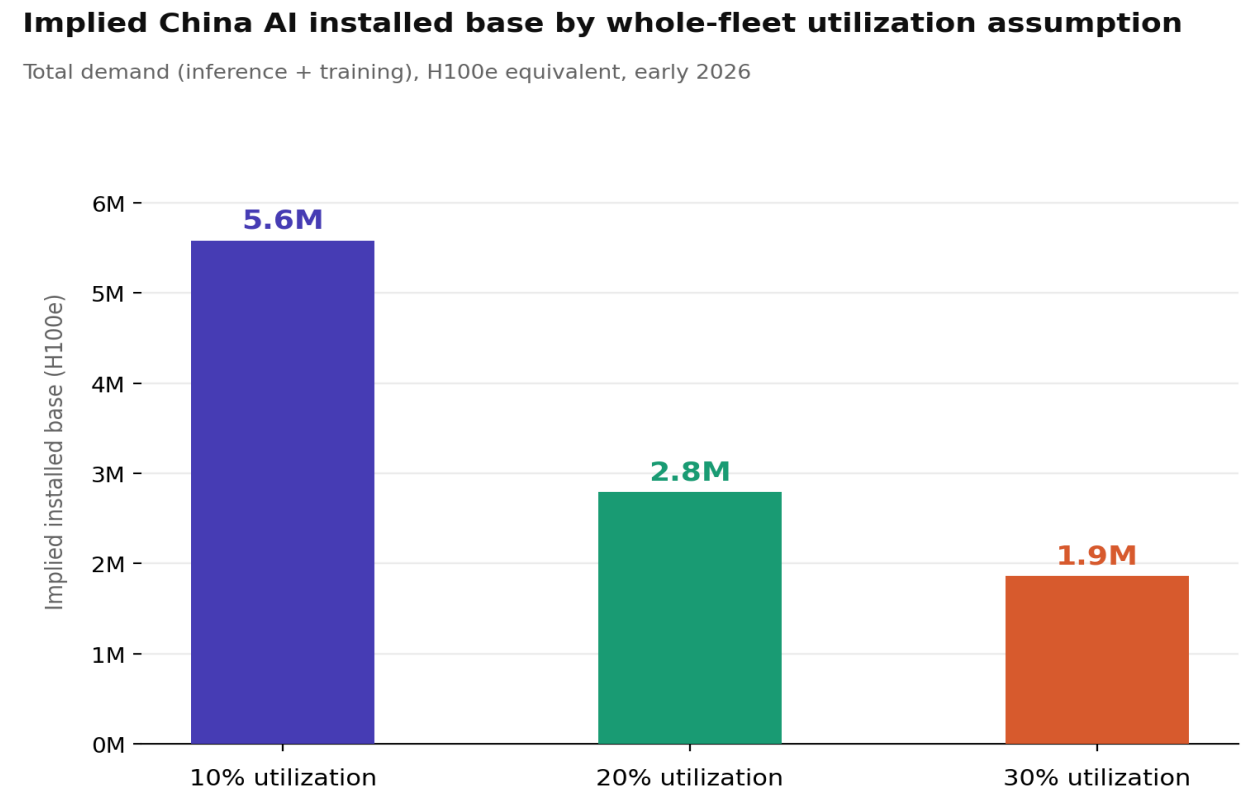


Implied minimum installed base across utilization assumptions. 55% is the central estimate.

Step B cross-check (what our bottom-up implies about the supply-side):

Rather than anchoring on any specific supply-side estimate, the table below inverts the question: given our bottom-up central case of ~558K H100e (at 55% inference util) installed, what total Chinese chip stock would that imply at different whole-fleet utilization assumptions? This lets readers calibrate against whatever supply-side figure they find credible.

Whole-Fleet Utilization Assumption	Implied Total Supply-Side (H100e)
10%	~5,582,040
20%	~2,791,020
30%	~1,860,680



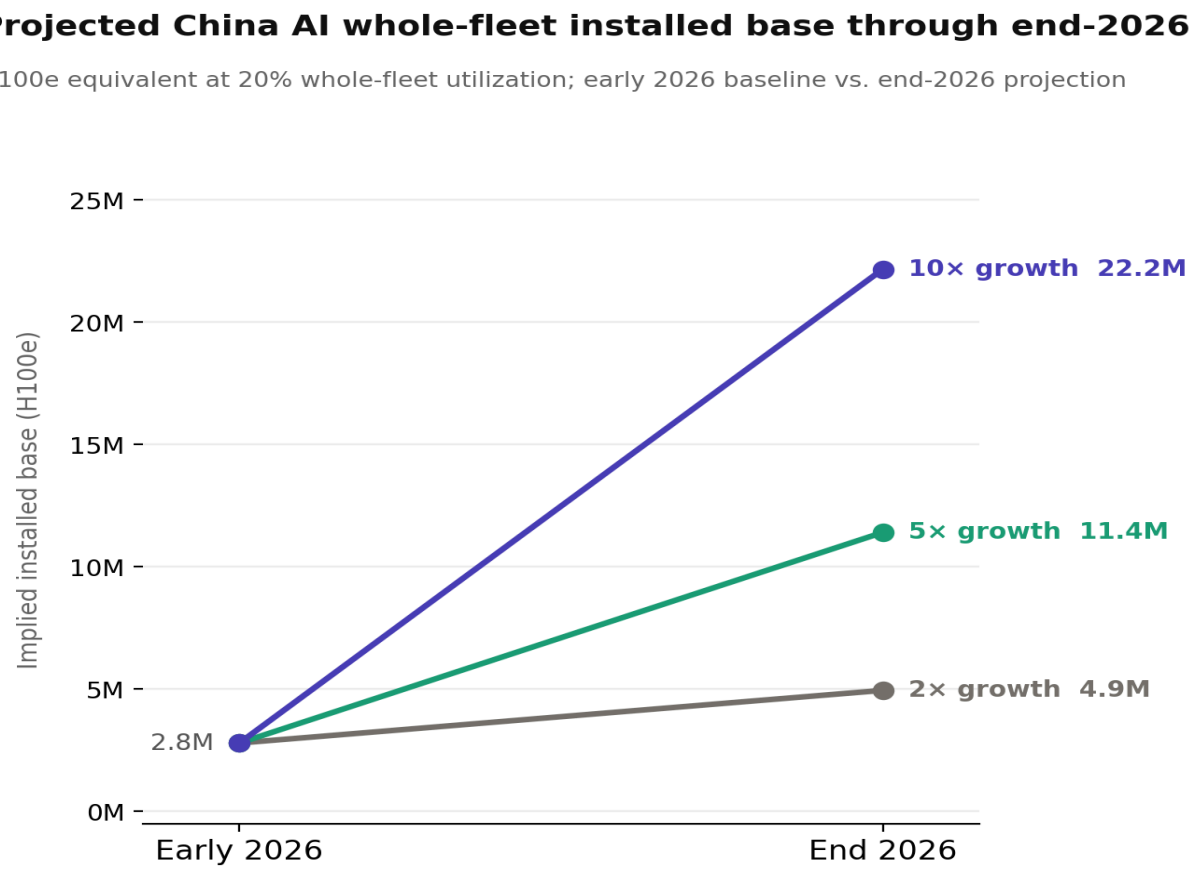
Implied total Chinese AI chip stock under three whole-fleet utilization assumptions. Central estimate at 20% (~2.8M H100e).

5.3 Full-Year 2026 Projected Demand

Global AI token consumption has grown approximately 5× annually. Three growth scenarios are shown: 2×, 5×, and 10×. Notably, at 10x growth, China would need more compute than the entire world's capacity currently.

> ⚠ *Growth multipliers are applied to the inference pool only — training demand is held flat as a roughly stable dedicated cluster (~127,568 H100e) that does not scale proportionally with inference usage.*

Growth Scenario	H100e Installed Base @ 55% util	H100e Whole-Fleet @ 20% util
Current run-rate (early 2026)	~558,204	~2,791,020
2× growth	~988,840	~4,944,200
5× growth	~2,280,748	~11,403,740
10× growth	~4,433,928	~22,169,640



Projected whole-fleet installed base through end-2026 under 2×, 5×, and 10× growth scenarios. H100e equivalent at 20% utilization.

5.4 Western CSP Discount

Scope note: All demand figures in this BOTEC represent total Chinese AI compute consumption regardless of where it physically runs. But a meaningful but uncertain fraction of Chinese AI demand runs on Western cloud infrastructure rather than Chinese-owned chips. I therefore decided to apply a 15–40% discount. The 15% floor reflects the fact that inference serving is unlikely to route heavily through Western cloud for political and data-sovereignty reasons, even if training does. The 40% ceiling reflects growing evidence of large-scale Chinese

use of non-Chinese compute for training workloads — particularly in Southeast Asia, where such arrangements are legal provided the chip owner is not headquartered in China. ByteDance is reportedly Oracle's largest customer; their largest joint cluster in Southeast Asia may reach ~250,000 B300-equivalents (~560K H100e) by summer 2026. Alibaba has also reportedly trained Qwen models on Nvidia GPUs in Southeast Asia.

Growth Scenario	Undiscounted Installed 55% util	After 15% CSP Discount	After 40% CSP Discount
Current run-rate	~558,204	~474,473	~334,922
2× growth	~988,840	~840,514	~593,304
5× growth	~2,280,748	~1,938,636	~1,368,449
10× growth	~4,433,928	~3,768,839	~2,660,357

Growth Scenario	Undiscounted Whole Fleet 20% util	After 15% CSP Discount	After 40% CSP Discount
Current run-rate	~2,791,020	~2,372,367	~1,674,612
2× growth	~4,944,200	~4,202,570	~2,966,520
5× growth	~11,403,740	~9,693,179	~6,842,244
10× growth	~22,169,640	~18,844,194	~13,301,784

6. Questions to Munch On

- What fraction of Chinese AI training or inference workloads are routed through Western CSPs?
- What are GPU utilization rates for inference and training infrastructure? This is the highest-leverage parameter in the installed base calculation, but a subject of much disagreement.
- What share of Western AI companies' enterprise API customers exceed 1 trillion tokens/year? This calibrates the T4 estimate, which accounts for a large share of AI demand.
- Are there significant AI use cases not captured here — industrial AI at factory scale, financial trading, defense — that would materially change the estimate?
- How much AI R&D is the Chinese government dedicating to the PLA?