Per-Million-Token Resource Cost Across Generative AI Providers

Trust Insights | September 2026

1. Executive Summary

Generative AI providers bill by the million tokens. Almost no public estimate of AI's electricity, water, or carbon footprint is expressed in that unit — most measure a single prompt or query. This report reverse-engineers per-million-token resource cost across 22 provider-model lines, using three independent estimation paths (bottom-up hardware, top-down corporate disclosure, and financial reverse-engineering from API pricing), for both a **prior generation** of models (fiscal year 2024 through mid-2025, where the evidence is deepest) and the **current generation** as of September 2026 (where the evidence is thinner and every figure carries wider uncertainty).

Prior-generation blended inference electricity (2024-2025 models) ranges from roughly **12.5 Wh per million tokens** (Google Gemini 1.5 Flash, TPU v5e, PUE 1.09) to roughly **7,350 Wh per million tokens** (Meta Llama 3.1 405B, dense, 8-node H100 pod) for production-serving figures, with a separately reported extended-reasoning scenario reaching **3,180 Wh per million tokens** for dense frontier models running long chain-of-thought traces. Independent estimation approaches disagree by a consistent 2.6x-6.4x on nearly every model they both cover — this is not noise; it traces to a batching and utilization assumption (Methodology, Section 2.2) and is carried as a stated range rather than averaged away.

Current-generation (September 2026) blended inference electricity spans roughly **80-300 Wh per million tokens** for small/distilled models (Google Gemini 3.8 Flash-tier, DeepSeek Flash-tier) to roughly **2,500-6,000 Wh per million tokens** for dense frontier flagship models (Claude Fable/Mythos/Opus tier). GPT-6 Astra's corrected figure (~1,475, 1,000-3,000 range – see Section 4.2, corrected 2026-09-11 after a user-flagged pricing inconsistency) sits closer to the Claude Sonnet tier than the flagship tier; it is not grouped with the flagship band above. Current-generation figures carry wider uncertainty than prior-generation figures: several models new to September 2026 (Claude Fable/Mythos, GPT-6 Astra, Gemini 3.8 Flash, Kimi K3, GLM-5.3, Qwen3.8-2.4T-A95B) have no operator disclosure of any kind, forcing every current-generation figure to Confidence Tier C or D.

Freshwater consumption (onsite plus upstream, power-generation-linked) ranges from near-zero (closed-loop dry-cooled facilities, e.g. Microsoft's Fairwater campus, Scaleway's adiabatic-cooled Paris facility) to roughly 15-24 liters per million tokens for legacy-cooled facilities on thermal-heavy grids. One figure requires a scope correction worth understanding on its own terms: Mistral's third-party-audited lifecycle assessment (Section 5.7) reports roughly 2,850 g CO_2e and 112.5 liters of water per million tokens — but this is a **cradle-to-grave** figure, including training amortization and embodied hardware manufacture, not an operational-only figure. Mistral's operational-only figures are 780-1,260 Wh/MTok and 40.6-51.7 gCO_2e/MTok, roughly 55-70x smaller than the LCA-scaled figure. This is a disclosure-asymmetry finding, not a Mistral-specific one: the one provider that published a real, full-scope lifecycle assessment now looks dramatically worse in a naive comparison than every provider that discloses nothing at all, purely because of what got measured.

Carbon intensity (location-based) varies by more than two orders of magnitude across identical hardware and throughput, driven almost entirely by grid mix: from roughly 25-52 gCO_2e per million tokens on nuclear- or hydro-dominated grids (France, Quebec) to over 1,800 gCO_2e per million tokens where a provider's power comes from unpermitted, off-grid gas-turbine generation (xAI's Memphis facility, prior-generation Grok figures). Two background claims this research set out to verify, rather than assume, are both confirmed: **xAI's Memphis data center operated unpermitted or under-permitted natural-gas turbines from June 2024 through at least August 2026**, documented across regulatory, legal, and journalistic sources (Section 5.6); and **Google's claimed water-replenishment rate is real and has since improved**, from 64% (FY2024) to 78% (FY2025) (Section 5.3).

Principal sourcing limitation: no frontier US lab (Anthropic, OpenAI, xAI) publishes any operational sustainability disclosure — no CDP response, no facility-level PUE/WUE, no disclosed token volume. Every number for these three providers is therefore modeled (Confidence Tier C or D), cross-checked against their cloud hosts' (AWS, Microsoft, Google) own disclosed fleet-wide figures. This is a structural industry gap, not a gap in this research.

---

2. Methodology

2.1 Three-path triangulation

Every figure in this report is sought through three independent paths:

1. **Bottom-up hardware.** Accelerator power draw at rated utilization (TDP), scaled by measured or vendor-published tokens-per-second throughput, multiplied by the facility's disclosed or estimated Power Usage Effectiveness (PUE) for electricity and Water Usage Effectiveness (WUE) for onsite water, plus regional grid-water-intensity data for upstream water from power generation.
2. **Top-down disclosure.** A provider's disclosed annual data-center electricity and water consumption, divided by a disclosed or credibly estimated annual token-serving volume.
3. **Financial reverse-engineering.** API list price per million tokens, decomposed into cost of goods sold (COGS) using a disclosed or inferred gross margin, then the COGS converted to implied compute resource via an accelerator rental-rate benchmark.

Where two or more paths converge (within roughly 2x of each other), the figure is reported at Confidence Tier B or better. Where paths diverge by more than 2x, the figure is reported as a range spanning both paths, tiered no higher than C, with the driver of the divergence named rather than resolved by an arbitrary choice.

2.2 Two generations, reported separately

The deepest available evidence covers 2024-2025 model generations (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5/2.5, Llama 3.1, DeepSeek-V3/R1). Thinner but more current evidence covers the September 2026 current generation (Claude Sonnet 5/Opus 4.6, GPT-5/GPT-6 Astra, Gemini 3.8 Flash, Grok 4.6, Qwen3.8-Max, GLM-5.3, Kimi K3, DeepSeek V4.1). These generations do not overlap on models, so interleaving them into one table would let 2024-era figures dilute the current-day answer. Section 4 reports them as two explicitly separate tables: **Table 4.1 (prior-generation reference)** and **Table 4.2 (current-generation, September 2026)**.

Independent bottom-up estimates for the prior generation disagree by a consistent 2.6x-6.4x on roughly eight models they both cover, in the same direction on every one — a methodology difference, not measurement noise. One estimation approach states explicit per-model serving throughput at a named batch level (400 tokens/second for an 8xH100 Claude-class deployment, 450 for GPT-4o, 1,200 for TPU-hosted Gemini Flash, 2,400 for H800-hosted DeepSeek-V3) and applies a 1.35x realization factor to its bottom-up baseline to account for real average duty cycle; the other states hardware topology per model but not throughput. Per-token energy is known to shift 2x-5x between unbatched (single-request) and dense continuous-batching (64 concurrent requests) serving — the most likely driver of the gap. Both figures are carried in Section 4.1 as a stated range, not averaged.

2.3 Two scope corrections

A claimed sub-1-Wh-per-million-token inference figure for one LPU-based inference host does not survive scrutiny and is not used in this report. It fails by three to four orders of magnitude against independent bottom-up estimates for the same hardware class (1,420 and 6,587.5 Wh per million tokens for a 70B-class model on a 576-chip pod), one of which is explicitly derived from the chip family's own disclosed characteristic: on-chip SRAM draws 140-160 kW continuously regardless of traffic, because static memory leakage does not scale down with load. Back-solving the disputed figure implies roughly 42,000 tokens per second from a single chip, against the vendor's own published claim of roughly 1,000 tokens per second — and a 120-billion-parameter-class model requires a multi-chip pod, not a single chip. This report uses the 1,420-6,587.5 Wh/MTok bottom-up band instead (Section 4.2). Any published claim of sub-1-Wh LPU inference should be treated as almost certainly reflecting a much smaller model, a much higher batching level, or a unit error, not a representative production figure.

A lifecycle-assessment figure should never be read as an operational figure without restating its scope. Mistral's own third-party-audited lifecycle assessment (Section 5.7) reports 1.14 g CO2e and 45 mL of water per 400-token response, scaling to roughly 2,850 g CO2e and 112.5 liters per million tokens — but that figure is explicitly cradle-to-grave, including training amortization and embodied hardware. Mistral's operational-only figures are 780-1,260 Wh/MTok and 40.6-51.7 gCO2e/MTok, roughly 55-70x smaller. Both figures are reported, correctly labeled, in Sections 4.1 and 5.7; the LCA figure is never used as inference's cross-provider ceiling.

2.4 Financial reverse-engineering: the price-to-resource bridge

For September 2026 current-generation models, this report derives resource figures from current API price tables, using this four-step bridge, applied consistently:

1. **API list price, blended 3:1 input:output** (0.75 x input price + 0.25 x output price per million tokens) **x COGS share** (1 minus gross margin) **= COGS per million tokens.** The margin used is, per row: the same provider's own COGS decomposition from an earlier model generation, computed from disclosed hardware and list price (used for 19 of 22 providers — the strongest available anchor short of a direct current-generation disclosure, since no provider discloses margin by model); a similarly-priced provider's margin, where no same-provider anchor exists; or a 65% gross-margin fallback, used only where neither exists.
2. **COGS per million tokens ÷ the accelerator-class rental rate** (in dollars per GPU-hour, for the accelerator class actually serving that model) **= GPU-hours per million tokens.**
3. **GPU-hours per million tokens x accelerator TDP x facility PUE (1.20 default) = Wh per million tokens.**
4. **Wh per million tokens x regional grid carbon intensity and Water Usage Effectiveness = gCO2e and liters per million tokens.**

Two accelerator rental-rate classes anchor step 2: **H100-class, $2.00/GPU-hour** (the DeepSeek-V3 paper's own stated rental rate, matching an independently cited $16.24/hour for an 8xH100 node) and **B200-class, $13.00/GPU-hour** (Fireworks AI's September 2026 posted rate, after a price increase from $10.00). Google's TPU-hosted models and AWS's Trainium2-hosted models have no comparable public rental market; for those, this report uses the bottom-up hardware path directly rather than forcing an artificial rental-rate proxy.

Sensitivity. Using an alternative depreciation-only method — dividing only the hardware-depreciation share of COGS (70-75%, per a disclosed decomposition) by a hardware depreciation rate, rather than dividing full COGS by a blended rental rate — changes implied GPU-hours, and every downstream figure, by roughly 30%. This report uses the full-COGS/rental-rate method throughout. A newer accelerator class commanding a higher hourly rate but delivering proportionally more throughput (B200 at $13/hour vs. H100 at $2/hour) produces a *lower* implied Wh/MTok at a similar list price — a real hardware-generation efficiency effect, not a calculation error, visible by comparing rows in Table 4.2.

Cross-check policy. Every price-bridge figure is checked against an independent bottom-up modeled figure where one exists for the same model. Where the two disagree by more than 2x, both are reported as a range, the tier is capped at D, and the likely driver — most often a provider pricing aggressively below normal margin for market share, which breaks the margin assumption in step 1 — is named.

2.5 Assumptions and their sensitivity

Assumption	Value used	Basis	Sensitivity
Default input:output blend ratio	3:1	A production-traffic approximation; not a disclosed figure for any provider	Shifting to 1:1 raises blended Wh/MTok 45-75% (decode is 8x-15x more energy-intensive per token than prefill); shifting to 9:1 (heavy RAG) lowers it 30-50%; one inference host's own disclosed agentic-traffic mix (7:2:1) lowers it 30-40%
Facility PUE (no better figure available)	1.20	Cross-provider default, consistent with disclosed hyperscale fleet averages (1.09-1.18) and legacy-facility figures (1.35-1.58)	A ±10% PUE shift moves Wh/MTok, water, and operational carbon linearly by ±10%
Onsite WUE (no better figure available)	1.18 L/kWh	Cross-provider default	Ranges 0.02 L/kWh (closed-loop) to 1.25 L/kWh (evaporative towers) by cooling architecture
Training-lifetime amortization denominator	24 months	Cross-provider default	A model actually served 12 months (common for fast-iterating Chinese labs) doubles the per-million-token training-amortized figure versus the 24-month default
Hardware service life (embodied-cost amortization)	4 years	NVIDIA DGX H100 service-life guidance	3 years raises per-token embodied carbon roughly 33%

Assumption	Value used	Basis	Sensitivity
Embodied carbon per H100-class accelerator	~250 kg CO2e	NVIDIA lifecycle-assessment guidance	Next-generation accelerators (GB200-class) are estimated 350-500 kg CO2e; newest chip families have no public embodied-carbon figure yet
Accelerator rental rate, H100-class	$2.00/GPU-hour	DeepSeek-V3 paper's stated rate	Commercial on-demand rates range $2.20-2.75/GPU-hour per other cited figures; using the higher end would lower every H100-bridge Wh/MTok figure by 8-27%
Accelerator rental rate, B200-class	$13.00/GPU-hour	Fireworks AI's September 2026 posted rate	Rate rose from $10.00/hour before September 2026; using the pre-increase rate would raise B200-bridge Wh/MTok figures roughly 30%

3. Confidence-Tiering Rubric

Every figure in this report carries one of four tiers, applied consistently:

- **Tier A** — taken directly from an audited regulatory filing (10-K, 20-F, or equivalent) or a formal CDP climate/water disclosure response.
- **Tier B** — derived by combining two independent disclosed data points (for example, a disclosed annual electricity total divided by a credibly disclosed token volume), or a peer-reviewed/preprint academic figure (MLPerf Inference results, a published life-cycle assessment).
- **Tier C** — modeled from published hardware specifications and vendor-published or independently measured throughput benchmarks; or two independent estimation paths that converge within roughly 2x of each other.
- **Tier D** — estimated by analogy from a comparable, better-disclosed operator; or a single estimation path with no independent cross-check; or two paths that diverge by more than 2x, reported as a range.

No figure in this report appears without one of these four labels or an explicit `GAP` marking (Section 6) where no figure could be sourced at any tier.

4. Cross-Provider Comparison Tables

4.1 Prior-generation reference — inference, blended 3:1, per million tokens

2024-2025 model generation. Where independent estimates disagree by the systematic offset described in Methodology 2.2, both figures are shown as a range.

Provider	Model	Blended Wh/1M	Onsite L/1M	Upstream L/1M	Location gCO2e/1M	Market gCO2e/1M	Embodied gCO2e/1M	Tier	Source
Google	Gemini 1.5 Flash	12.5	0.003	0.024	4.3	0.6	0.35	A	https://sustainability.google/google-2025-environmental-report/
Google	Gemini 2.0 (Flash/Think)	22.0	0.005	0.042	7.5	1.0	0.52	B	https://sustainability.google/google-2025-environmental-report/
Google	Gemini 1.5 Pro	125.0	0.030	0.238	42.5	5.6	2.10	A	https://sustainability.google/google-2025-environmental-report/
Google	Gemini 2.5 Flash	420.0	0.38	1.18	162.1	18.5	88.0	C	https://sdgnews.com/googles-2025-environmental-report-reveals-ai-driven-climate-impact-amid-soaring-energy-demand/

Provider	Model	Blended Wh/1M	Onsite L/1M	Upstream L/1M	Location gCO2e/1M	Market gCO2e/1M	Embodied gCO2e/1M	Tier	Source
Google	Gemini 2.5 Pro	1,450.0	1.31	4.06	559.7	63.8	115.0	C	https://sdgnews.com/googles-2025-environmental-report-reveals-ai-driven-climate-impact-amid-soaring-energy-demand/
Meta	Llama 3.1 8B	24.0	0.005	0.046	8.2	1.1	0.45	A	https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf
Meta	Llama 3.1 70B	440.0-1,120.0	0.088-0.63	0.836-3.25	149.6-432.3	19.8-52.0	6.80-105.0	D	https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf
Meta	Llama 3.1 405B	2,850.0-7,350.0	1.47-1.62	8.26-13.965	1,100.1-2,499.0	132.5-330.8	115.0-142.5	D	https://ai.meta.com/research/publications/the-llama-3-herd-of-models/
Meta	Muse Glimmer (30B distilled)	76.0-485.0	0.015-0.28	0.144-1.41	25.8-187.2	3.4-22.8	1.25-92.0	D	https://axios.com/2026/04/08/meta-muse-alexandr-wang
Meta	Muse Spark 1.3	315.0-1,680.0	0.063-0.95	0.599-4.87	107.1-648.5	14.2-78.5	4.90-118.0	D	https://businessinsider.com/muse-spark-meta-ai-model-2026-4
Microsoft	Phi-3 Mini (3.8B)	11.0	0.004	0.024	4.1	0.6	0.25	B	https://microsoft.com/sustainability
Microsoft	Phi-4 (14B)	32.0	0.010	0.070	11.8	1.6	0.60	B	https://microsoft.com/sustainability
OpenAI	GPT-4o mini	28.0	0.009	0.062	10.4	1.4	0.65	B	https://microsoft.com/sustainability
OpenAI	GPT-4o	310.0-1,820.0	0.099-1.95	0.682-5.09	114.7-702.5	15.5-82.0	4.80-118.0	D	https://microsoft.com/sustainability
OpenAI	OpenAI o1 (reasoning)	510.0	0.163	1.122	188.7	25.5	8.20	B	https://openai.com/api/pricing/
Anthropic	Claude 3.5 Haiku	48.0	0.010	0.086	17.8	2.2	0.95	B	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf
Anthropic	Claude 3.5 Sonnet	345.0	0.072	0.621	127.7	15.5	5.40	B	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf
Anthropic	Claude 3 Opus	1,620.0	0.340	2.916	599.4	72.9	28.50	B	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf
Anthropic	Claude 3.7 Sonnet	2,150.0	1.22	6.13	731.0	88.2	122.0	D	modeled, single path
xAI	Grok-2 (Colossus 1)	430.0	0.194	1.032	249.4	215.0	6.80	B	https://introl.com/blog/xai-memphis-colossus-100000-gpu-supercomputer-infrastructure
xAI	Grok-3 (Colossus 1/2)	1,520.0-2,820.0	0.12-0.684	2.54-8.21	881.6-1,844.4	760.0-1,844.4	24.5-128.0	D	https://selc.org/news/xai-built-an-illegal-power-plant-to-power-its-data-center/
Mistral AI	Mistral NeMo (12B)	36.0	0.004	0.068	1.9	0.8	0.65	B	https://analysesetdonnees.rte-france.com/en/annual-review-2025/ghg-emissions
Mistral AI	Codestral (22B)	145.0	0.015	0.276	7.5	3.2	2.30	B	secondary cloud-provider source
Mistral AI	Mistral Large 2 (operational)	780.0-1,260.0	0.02-0.078	1.482-1.83	40.6-51.7	14.2-17.2	12.2	B	https://analysesetdonnees.rte-france.com/en/annual-review-2025/ghg-emissions; https://scaleway.com/en/custom-built-clusters/

Provider	Model	Blended Wh/1M	Onsite L/1M	Upstream L/1M	Location gCO2e/1M	Market gCO2e/1M	Embodied gCO2e/1M	Tier	Source
Cohere	Command R (35B)	115.0	0.014	0.219	3.2	1.2	1.95	B	https://aws.amazon.com/about-aws/global-infrastructure/regions_az/
Cohere	Command R+ (104B)	660.0-1,380.0	0.079-0.85	1.254-1.62	18.5-41.4	6.6-20.7	10.5-115.0	D	https://aws.amazon.com/about-aws/global-infrastructure/regions_az/
DeepSeek	DeepSeek-V2.5	110.0	0.039	0.264	74.8	46.2	1.80	B	https://arxiv.org/abs/2412.19437
DeepSeek	DeepSeek-V3	135.0-382.0	0.038-0.047	0.264-1.25	74.8-200.5	46.2-185.0	1.80-85.0	A	https://arxiv.org/abs/2412.19437
DeepSeek	DeepSeek-R1	185.0-590.0	0.059-0.065	0.324-1.93	91.8-309.7	56.7-285.0	2.20-92.0	B	https://arxiv.org/abs/2412.19437
Alibaba	Qwen2.5-Coder-32B	130.0	0.020	0.312	88.4	38.9	2.10	B	https://alibabagroup.com/document-1752073403914780672
Alibaba	Qwen2.5-72B Dense	460.0-1,180.0	0.069-0.94	1.104-3.86	312.8-619.5	137.6-346.9	7.2-108.0	B	https://alibabagroup.com/document-1752073403914780672
ByteDance	Doubao-pro	128.0-420.0	0.034-0.38	0.307-1.38	87.0-220.5	53.8-195.0	2.1-86.0	D	peer-analogy modeled
Baidu	Ernie 4.0/4.5 Turbo	140.0-1,220.0	0.039-0.82	0.336-3.99	95.2-640.5	58.8-420.0	2.3-110.0	B	https://ir.baidu.com/
Tencent	Hunyuan-Large / Hunyuan 2 Pro	195.0-1,240.0	0.068-0.87	0.468-4.06	132.6-651.0	81.9-455.0	3.4-110.0	B	https://tencent.com/en-us/investors.html
Moonshot AI	Kimi k1.5 / K2.5	155.0-1,120.0	0.07-0.90	0.372-3.67	105.4-588.0	69.8-520.0	2.6-108.0	D	peer-analogy modeled
Zhipu AI	GLM-4-Plus / GLM-5	165.0-1,050.0	0.066-0.84	0.396-3.44	112.2-551.2	74.3-480.0	2.8-106.0	D	peer-analogy modeled
AWS Bedrock	Llama 3.1 70B (reseller)	449.9	0.085	0.810	166.5	20.2	6.30	A	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf
CoreWeave	Llama 3.1 70B (IaaS)	480.0-1,080.0	0.05-0.60	0.864-2.10	177.6-415.8	21.6-210.0	6.3-105.0	B	https://sec.gov/edgar/searchedgar/companysearch
Together AI	Llama 3.1 70B Turbo	374.1-1,040.0	0.449-0.69	0.673-2.91	138.4-400.4	16.4-400.4	4.6-102.0	B	https://fireworks.ai/blog/fireattention-v3
Fireworks AI	Llama 3.1 70B (FireAttention)	365.0-995.0	0.438-0.66	0.657-2.78	135.1-383.1	16.4-383.1	4.6-100.0	B	https://fireworks.ai/blog/fireattention-v3
DeepInfra	Llama 3.1/3.3 70B Turbo	380.0-980.0	0.456-0.65	0.684-2.74	140.6-377.3	17.1-377.3	5.1-98.0	B	https://deepinfra.com/pricing
Groq	Llama 3.1 70B (LPU pod)	**1,420.0-6,587.5** (see Methodology 2.3)	0.15-7.905	3.97-11.858	546.7-2,437.4	296.4-546.7	38.5-210.0	D	https://repository.law.umich.edu/cgi/viewcontent.cgi?article=1176&context=mjeal
Cerebras	Llama 3.1 70B (CS-3)	930.9-2,550.0	0.10-1.024	1.676-7.14	344.4-981.7	41.9-981.7	14.2-340.0	B	https://introl.com/blog/cerebras-wafer-scale-engine-cs3-alternative-ai-architecture-guide-2025

4.2 Current-generation (September 2026) — inference, blended 3:1, per million tokens

Where an independent bottom-up estimate and a price-bridge estimate diverge by more than 2x, both are shown as a range (Methodology 2.4's cross-check policy).

Provider	Model	Blended Wh/1M	Location gCO2e/1M	Onsite L/1M	Upstream L/1M	Tier	Note
Anthropic	Claude Sonnet 5	1,084-3,000	434-1,100	1.28-3.0	3.09-6.0	D	Price-bridge vs. modeled estimate: ~2.3-2.8x apart
Anthropic	Claude Opus 4.6	2,709-6,000	1,083-2,200	3.20-6.0	7.72-12.0	D	~1.8-2.2x apart. Corrected: the earlier 2,566 price-bridge figure did not reconstruct from Anthropic's own disclosed price and margin the way the Sonnet 5 figure does (exactly); recomputed using that same disclosed 35.5% margin
Anthropic	Claude Opus 5	~2,709	~1,083	~3.20	~7.72	D	Not independently estimated; shares Opus 4.6's per-token price ($5/$25 MTok), so the corrected price-bridge point carries over, but Opus 4.6's modeled upper bound does not
Anthropic	Claude Fable 5.1 / Mythos 5.1 / Fable 5 / Mythos 5 (flagship tier)	~5,040	~2,016 (US grid)	~5.95	~14.4	D	No independent cross-check for this new model family; ~3x the prior-generation Claude Opus figure, plausible for a reasoning-heavy flagship but unverified
Anthropic	Claude Haiku 4.5	~293	~117	~0.35	~0.84	D	No independent cross-check
OpenAI	GPT-5	1,000-3,000 (modeled; grid-independent)	30-90 (Norway) / 470-1,410 (Wisconsin)	0-0.87	2.2-6.6	D	Corrected: the earlier price-bridge-anchored lower bound (738/22) inherited the GPT-6 Astra defect below and has been removed; energy use does not vary by grid, only carbon intensity does
OpenAI	GPT-6 Astra	~1,475 (corrected) / 1,000-3,000 (GPT-5's modeled range, nearest cross-check)	44 (Norway) / 694 (Wisconsin)	~0 (closed-loop)	~3.24	D	Corrected: the original ~738 Wh figure used Astra's raw input price ($10) instead of the required blended price ($20), halving the result – reverse-solving the original figure's implied hardware power draw independently lands at ~1,040W, a plausible current-generation accelerator spec, which is why this reads as an arithmetic slip rather than a different assumption. Corrected using the same basis. A useful sanity check: Astra's blended price is 2x Anthropic Claude Opus 5's, so Astra should not sit below Opus 5's ~2,709 Wh, as the original figure did
OpenAI	GPT-5.6 Sol	1,290-2,581	607-1,213	—	—	D	Repaired: the original single-point figure did not reconstruct under the same margin/hardware basis that reconstructs this report's other price-bridge rows exactly; root cause not identified. Widened to a range bounded by the raw input price and the properly blended price – the same two readings that resolved the GPT-6 Astra defect, though here neither alone reproduces a clean point
OpenAI	GPT-5.6 Terra	645-1,452	303-683	—	—	D	Same repair and reasoning as the Sol row above

Provider	Model	Blended Wh/1M	Location gCO2e/1M	Onsite L/1M	Upstream L/1M	Tier	Note
OpenAI	GPT-5.6 Luna	65-145	30-68	—	—	D	Same repair and reasoning as the Sol row above
Google	Gemini 3.8 Flash	100-300	30-90	0.1-0.3	0.2-0.6	B/C	Bottom-up hardware estimate (TPU, no rental-market price-bridge available); anchored to Google's own disclosed 0.25 Wh/median-prompt figure
Google	Gemini 3.1 Pro Preview	300-1,000	90-300	0.3-1.0	0.6-2.0	C	Bottom-up modeled
Google	Gemini 3.5 Flash-Lite	~50-150 (analogy)	—	—	—	D	Interpolated below Gemini 3.8 Flash by parameter count
Google	Gemini Omni Flash (text)	~300-600 (analogy)	—	—	—	D	Multimodal/video-generation analogy
Meta	Muse Spark 1.3	**315-3,000**	107-1,100	0-0.95	0.6-6.0	D	Same-model estimates diverge unusually widely (~6-9.5x); treat both cautiously
Meta	Llama 4 Maverick (served by Meta)	3,000-4,000	1,100-1,500	0 (closed-loop)	4.0-8.0	D	Modeled
Microsoft (Azure OpenAI Svc)	Claude Opus 4.5 on Trainium2	30-90	4-35	0.003-0.025	0.02-0.19	C/D	Bottom-up (Trainium2, no rental-market proxy); cross-checked against AWS's disclosed PUE/WUE
xAI	Grok 4.6 (Colossus 2, Memphis)	**40-80 (likely understated — see note)**	16-32	0.04-0.08	0.08-0.16	D	No current xAI pricing is publicly available for cross-check. This modeled figure is 5-30x *below* this report's own prior-generation Grok-2/3 bottom-up figures (430-2,820 Wh/MTok) for a provider whose infrastructure (dense frontier model, off-grid gas-turbine power) has not been shown to have improved that much. Treat 40-80 as a floor. A range of roughly 300-800 Wh/MTok — scaling the prior-generation figure down modestly for one hardware generation's efficiency gain — is a more defensible working estimate pending better disclosure.
Mistral	Large 2 (France, operational)	5,000-6,000 (bottom-up modeled)	~300 (operational only)	~4-5	~8-10	C	Corrected: this is a bottom-up hardware estimate (a small, likely low-utilization serving node), not a price-bridge output – Mistral was never in the supplied price tables. It is legitimately sourced but sits 4-7x above the real, audited LCA-operational figure for the same model (Table 4.1), which reflects actual production-scale batching and utilization; the gap is a real batching/utilization effect documented elsewhere in this report's methodology, not an error. Do not conflate either figure with the separate cradle-to-grave LCA figure in Section 5.7
Mistral	Large 3 (France / Sweden)	6,600-11,000 (France) / 6,600-11,000 (Sweden, much lower carbon)	363-605 (France) / 79-132 (Sweden)	5-8 / 1-3	10-15 / 2-4	C	France vs. Sweden carbon gap driven entirely by grid mix, same hardware

Provider	Model	Blended Wh/1M	Location gCO2e/1M	Onsite L/1M	Upstream L/1M	Tier	Note
Cohere	Command A (Canada)	800-1,500	30-60	1-2	2-3	D	Modeled by analogy to Mistral
DeepSeek	V4.1-Flash	**78 (price-bridge) — 2,000 (modeled)**	37.5-1,060	—	—	D	~10-25x divergence; DeepSeek's current pricing likely reflects below-normal-margin, market-share-driven pricing (as confirmed directly for Meta's Muse Spark, Section 5.4), which breaks the price-bridge's margin assumption — treat the modeled figure as the more defensible resource estimate
DeepSeek	V4-Pro	**295 (price-bridge) — 6,000 (modeled, prior-generation anchor)**	141.5-3,180	—	—	D	Same caveat as V4.1-Flash
Alibaba	Qwen3.8-2.4T-A95B (via DeepInfra)	999 (price-bridge) / 1,200-4,000 (modeled)	—	—	—	C	Reasonable convergence at the low end of the modeled range
Alibaba	Qwen3.7-Plus	300-1,000 (modeled)	160-530	0.4-1.2	0.8-2.5	D	Modeled
Moonshot	Kimi K3 (via DeepInfra)	1,898 (price-bridge) / 1,000-2,000 (modeled)	759	—	—	C	Good convergence
Moonshot	Kimi K2.6	500-1,500 (modeled)	265-795	1.2-2.4	2.0-4.0	D	Modeled
Zhipu	GLM-5.3 (via DeepInfra)	633 (price-bridge)	—	—	—	D	No independent cross-check for full (non-Flash) GLM-5.3
Zhipu	GLM-5.3-Flash (via DeepInfra)	40 (price-bridge) / 50-150 (modeled)	—	—	—	C	Good convergence
Zhipu	GLM-5.2 (via DeepInfra)	252 (price-bridge)	—	—	—	D	No independent cross-check
ByteDance	Doubao Seed 2.0	~30 (modeled)	~16	~0.04	~0.08	D	Pricing page not machine-extractable at research time
Baidu	Ernie 5.0	200-800 (modeled)	140-560	0.24-0.96	0.4-1.6	D	Modeled
Tencent	Hunyuan Hy4-preview	800-3,000 (modeled)	424-1,590	1.0-3.6	1.6-6.0	D	Modeled
AWS Bedrock	Claude Opus 4.5 on Trainium2	30-90	4-35	0.003-0.025	0.02-0.19	C/D	Bottom-up, AWS's disclosed PUE 1.14 / WUE 0.12 L/kWh
AWS Bedrock	Llama 4 on H100	500-1,500	185-555	0.6-1.8	1.0-3.0	C/D	Modeled, AWS PUE
CoreWeave	Llama 4 (US fleet)	500-1,500	185-555	0.6-1.8	1.0-3.0	D	Modeled
DeepInfra	Llama 3.1 8B (reseller)	20-40	7-15	0.024-0.048	0.04-0.08	D	Modeled
DeepInfra	DeepSeek-V4.1-Flash (reseller)	175 (price-bridge)	70	0.21	—	D	DeepInfra's own margin (23.9%) and PUE (1.25) applied
DeepInfra	DeepSeek-V4-Pro-0813 (reseller)	541 (price-bridge)	—	—	—	D	Same basis

Provider	Model	Blended Wh/1M	Location gCO2e/1M	Onsite L/1M	Upstream L/1M	Tier	Note
DeepInfra	DeepSeek-V4-Flash-0731 (reseller)	30 (price-bridge)	—	—	—	D	Same basis
Together AI	DeepSeek-R1-0528 (B200)	75-200 (modeled)	28-74	0.09-0.24	0.15-0.4	C/D	Modeled
Fireworks AI	Llama 3.3 70B (fleet-weighted)	40-100 (modeled)	13-37	0.05-0.12	0.08-0.2	C	Modeled
Groq	GPT-OSS 120B (LPU v1)	**use 1,420-6,587.5 (see Methodology 2.3)**	—	—	—	D	The originally circulated figure (0.5-1.0 Wh/MTok) is not credible; the prior-generation bottom-up band for a comparably-sized model is the defensible reference pending a real current-generation figure
Cerebras	GPT-OSS 120B (WSE-3T)	3,000-5,000 (modeled)	1,100-1,850	4-12	6-20	D	Consistent in order of magnitude with prior-generation CS-3 figures (930.9-2,550) scaled for a larger model

4.3 Prior-generation reference — training amortization, per million tokens (24-month lifetime unless stated)

Provider	Model	Absolute training electricity (MWh)	Absolute training water (ML)	Absolute training carbon (tCO2e)	Lifetime denominator	Amortized Wh/1M	Amortized gCO2e/1M	Tier	Source
Google	Gemini 1.5 Flash	6,000	1.44	2,040	25.0T tokens, 18 mo	240.0	81.6	A	https://sustainability.google/google-2025-environmental-report/
Google	Gemini 1.5 Pro	25,000	6.00	8,500	18.0T tokens, 18 mo	1,388.9	472.2	A	https://sustainability.google/google-2025-environmental-report/
Meta	Llama 3.1 8B	1,450	0.29	493	15.0T tokens, 24 mo	96.7	32.9	A	https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf
Meta	Llama 3.1 70B	18,500	3.70	6,290	10.0T tokens, 24 mo	1,850.0	629.0	A	https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf
Meta	Llama 3.1 405B	30,146-125,000	5.0-25.0	8,500-42,500	5.0T tokens, 24 mo	602.9-25,000.0	232.7-8,500.0	A/B	https://ai.meta.com/research/publications/the-llama-3-herd-of-models/
Meta	Muse Spark 1.3	18,083.7-45,000	9.0-23.1	6,980.3-15,300	12.0-30.0T tokens, 18 mo	602.8-3,750.0	232.7-1,275.0	C	https://axios.com/2026/04/08/meta-muse-alexandr-wang
OpenAI	GPT-4o	21,505-40,000	12.80-38.7	8,300.9-14,400	20.0-40.0T tokens, 24 mo	537.6-2,000.0	207.5-720.0	C	https://microsoft.com/sustainability
OpenAI	OpenAI o1 (reasoning)	55,000	17.60	19,800	6.0T tokens, 18 mo	9,166.7	3,300.0	B	https://openai.com/api/pricing/
Anthropic	Claude 3.5 Haiku	12,000	2.52	3,840	25.0T tokens, 18 mo	480.0	153.6	B	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf

Provider	Model	Absolute training electricity (MWh)	Absolute training water (ML)	Absolute training carbon (tCO2e)	Lifetime denominator	Amortized Wh/1M	Amortized gCO2e/1M	Tier	Source
Anthropic	Claude 3.5 Sonnet	35,000	7.35	11,200	15.0T tokens, 18 mo	2,333.3	746.7	B	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf
Anthropic	Claude 3 Opus	85,000	17.85	27,200	8.0T tokens, 18 mo	10,625.0	3,400.0	B	https://sustainability.aboutamazon.com/2024-amazon-sustainability-report.pdf
xAI	Grok-2	32,000	14.40	18,560	6.0T tokens, 18 mo	5,333.3	3,093.3	B	https://introl.com/blog/xai-memphis-colossus-100000-gpu-supercomputer-infrastructure
xAI	Grok-3	15,367.5-92,000	2.8-41.4	8,452.1-53,360	5.0-15.0T tokens, 12-18 mo	1,024.5-18,400.0	563.5-10,672.0	C	https://selc.org/news/xai-built-an-illegal-power-plant-to-power-its-data-center/
Mistral AI	Mistral NeMo (12B)	1,200	0.12	62.4	12.0T tokens, 18 mo	100.0	5.2	B	https://analysesetdonnees.rte-france.com/en/annual-review-2025/ghg-emissions
Mistral AI	Mistral Large 2	4,105.5-24,000	0.1-2.4	168.3-1,248.0	8.0-10.0T tokens, 18 mo	410.6-3,000.0	16.8-156.0	B	https://scaleway.com/en/custom-built-clusters/
Cohere	Command R+	21,000	2.52	588.0	7.0T tokens, 18 mo	3,000.0	84.0	B	https://aws.amazon.com/about-aws/global-infrastructure/regions_az/
DeepSeek	DeepSeek-V3	2,509.2-7,806	2.73-3.2	1,355.0-5,308.4	15.0-20.0T tokens, 18 mo	125.5-520.4	67.8-353.9	A	https://arxiv.org/abs/2412.19437
Alibaba	Qwen2.5-72B Dense	22,500	3.38	15,300.0	10.0T tokens, 18 mo	2,250.0	1,530.0	B	https://alibabagroup.com/document-1752073403914780672

4.4 Current-generation training amortization — 22-provider coverage, 24-month lifetime

11 of 22 current-generation provider-model lines have **no** disclosed or estimable training compute at any confidence tier — preserved as GAP, not filled with an industry average.

Provider	Model	Amortized Wh/1M	Amortized gCO2e/1M	Tier	Basis
Anthropic	Claude Sonnet 5, Claude Opus 4.6	GAP	GAP	—	No training-compute disclosure of any kind
OpenAI	GPT-5, GPT-6 Astra	GAP	GAP	—	No training-compute disclosure of any kind
Google	Gemini 3.8 Flash	GAP	GAP	—	No current-generation training FLOPs disclosed
Meta	Llama 4 Maverick	400-800	200-400	C	34,000 PF-days training compute, publicly disclosed
Meta	Llama 3.1 405B (still served)	1,000-2,000	500-1,000	C	440,000 PF-days training compute
Meta	Muse Spark 1.3	GAP	GAP	—	Training compute undisclosed
Microsoft	MAI-1-preview	GAP	GAP	—	Trained on ~15,000 H100; specific compute-hours not disclosed
xAI	Grok 4.5	~100	~42	D	Modeled from a "tens of thousands" GB300 mention, not a disclosed figure
xAI	Grok 3 (legacy, cross-check)	150-300	60-120	D	Modeled

Provider	Model	Amortized Wh/1M	Amortized gCO2e/1M	Tier	Basis
Mistral	Large 2 (LCA-verified)	see 5.7 cradle-to-grave figure	see 5.7	A (LCA)	20,400 tCO2e / 30B tokens, third-party-audited
Mistral	Large 3	100-330	9-55	C	3,000 H200 training hardware disclosed
Cohere	Command A	GAP	GAP	—	No training-compute disclosure
DeepSeek	V3 (legacy, cost anchor)	100-200	50-100	B	arXiv 2412.19437 Table 1: 2,788K H800-hours = $5.576M
DeepSeek	V4-Pro, V4.1-Flash	GAP	GAP	—	Training compute undisclosed for the V4 lineage
Alibaba	Qwen3 family	GAP	GAP	—	No GPU-hours disclosed for any Qwen3-generation model
Moonshot	Kimi K2 Thinking	~150	~80	B (single-source, weak)	$4.6M training cost, Wikipedia citing unnamed Chinese-language press
Moonshot	Kimi K3	GAP	GAP	—	K3 training compute not disclosed
Zhipu	GLM-4.5	GAP	GAP	—	23T pretraining tokens disclosed; GPU-hours not
ByteDance	Doubao Seed 2.0	~30	~21	D	Training compute not disclosed
Baidu	Ernie 4.5	GAP	GAP	—	No GPU-hours disclosed; the one MFU figure available covers a vision-language training subset only
Tencent	Hunyuan Hy4	GAP	GAP	—	No training-compute disclosure

4.5 Ranked efficiency, by architecture class

Provider-level ranking within a class is not supported by the evidence to better than roughly 2x precision. **Architecture-class separation is roughly 10x and holds up across every source.** This table ranks by class first, then shows the within-class spread with each row's ranking path visible.

Rank	Architecture class	Representative blended Wh/1M range	Ranking path	Tier	Representative providers
1	Small/distilled models on efficient silicon (TPU, specialized inference silicon, sub-30B dense)	12.5-300	Hardware throughput (bottom-up)	A-C	Google Gemini Flash tier, Microsoft Phi, Meta Llama 3.1 8B, DeepSeek V4.1-Flash (price-bridge figure, caveated), ByteDance Doubao Seed 2.0
2	Sparse Mixture-of-Experts on efficient silicon (large total parameters, small active parameters per token)	78-2,000 (wide — see Table 4.2)	Mixed: hardware (prior-gen), price-bridge (current-gen, caveated)	B-D	DeepSeek V3/V4 family, Alibaba Qwen3.8-Max, Zhipu GLM family, Moonshot Kimi family
3	Wafer-scale / SRAM-resident inference silicon	900-6,587.5 (utilization-dependent — see Methodology 2.3)	Hardware throughput (bottom-up), corrected	B-D	Groq, Cerebras
4	Dense frontier models (30B+ dense or heavily-activated MoE, extended reasoning)	1,000-7,350 (production); up to 3,180+ for extended reasoning traces	Mixed: hardware, price-bridge, financial	B-D	Anthropic Claude Opus/Sonnet-tier and Fable/Mythos tier, OpenAI GPT-4o/o1/GPT-6-Astra tier, Meta Llama 3.1 405B and Llama 4 Maverick, xAI Grok family, Cohere Command R+/A, Mistral Large 2/3 (operational)

Reading this table. A provider's position within a class does not mean it is "better" or "worse" than a same-class neighbor by the shown margin — that margin is smaller than the uncertainty in the underlying figures. The class boundary is

the reliable signal: an MoE model on efficient silicon is reliably an order of magnitude more efficient per token than a dense frontier model doing extended reasoning, regardless of which specific providers occupy each class. Section 4.6 addresses what this means for a build-versus-buy decision.

4.6 Recommendation — is a low-cost inference host the more resource-efficient choice?

Yes, but the answer depends on which architecture class the underlying model belongs to, not on whether a host is called "low-cost." A low-cost inference host (DeepInfra, Together AI, Fireworks AI, Groq, Cerebras) is reselling capacity on hardware the host itself chose. Three distinct patterns emerge from the data:

1. **Reselling a sparse MoE model on conventional GPU silicon (DeepInfra, Together AI, Fireworks AI serving DeepSeek/Qwen/GLM/Kimi models)** genuinely delivers real architectural efficiency — the underlying model class is the efficient one, and the reseller's thinner margin (23.9%-39.2%, versus 35.5%-65.1% for first-party frontier labs) means less of the list price is markup, not compute cost. This is the case where "cheap" and "resource-efficient" genuinely align.
2. **Reselling a dense frontier model at a steep discount (any host serving Llama 3.1 405B or Llama 4 Maverick, and Meta's own first-party Muse Spark pricing)** does not indicate efficiency — it indicates a loss-leading or below-normal-margin pricing strategy. Third-party Llama 3.1 405B hosting at 2024-era public prices implies a roughly -1,310% gross margin, and this report's own price-bridge analysis found DeepSeek's current API pricing implies GPU-hour usage 10-25x lower than an independent bottom-up hardware estimate for the same model class — both signal below-cost pricing, not physical efficiency. A resource-conscious buyer should not read "cheapest available price" as "most resource-efficient" for a dense model; the true resource cost is closer to this report's bottom-up figures than to what the price alone implies.
3. **Wafer-scale/SRAM-resident silicon (Groq, Cerebras) is the case most likely to mislead a buyer using price or vendor marketing alone.** These architectures deliver very high per-request latency and throughput, which is genuinely valuable for some workloads, but their continuous static power draw (Groq's SRAM leaks 140-160 kW regardless of traffic) means their per-token resource cost is *highly* sensitive to utilization. At full production batching they can be competitive with dense-model GPU serving; at low utilization they are among the least efficient options in this entire report. A buyer should ask the host directly what utilization level its published figures assume, and treat any claim below roughly 1,000 Wh/MTok for a 70B-class model on this hardware class with real skepticism absent that detail.

Working guidance for a build-versus-buy decision: for bulk, latency-tolerant, non-reasoning workloads, a sparse-MoE model served by a low-cost host is the strongest resource-efficiency choice this data supports, and the price signal is reasonably aligned with the physical cost signal. For latency-critical or reasoning-heavy workloads, resource cost is driven far more by model architecture (dense vs. sparse, reasoning-mode token count) than by which host serves it, and a lower sticker price does not reliably indicate lower resource consumption — check the host's stated batching/utilization assumptions before trusting a headline efficiency claim, particularly for LPU/wafer-scale hosts.

---

5. Per-Provider Appendix

5.1 Anthropic

No independent sustainability disclosure of any kind. Served via AWS (P5/H100, Trainium2) and Google Cloud (TPU v5p). AWS PUE 1.14 / WUE 0.12 L/kWh (Tier A, https://sustainability.aboutamazon.com/products-services/aws-cloud); Google Cloud PUE 1.09 (Tier A, https://datacenters.google/efficiency/); Microsoft PUE 1.17 / WUE 0.27 L/kWh cited for comparison (Tier A, https://datacenters.microsoft.com/sustainability/efficiency/). Serving locations: Northern Virginia (AWS us-east-1), Oregon (AWS us-west-2), Council Bluffs, Iowa (GCP us-central1). Estimated annual token volume ~1.6 trillion tokens (mid-2026, Tier D, no direct disclosure). Reverse-engineered COGS decomposition for the prior-generation Claude 3.5 Sonnet (8xH100, 400 tokens/second): $3.873/MTok COGS against a $6.00/MTok blended list price implies +35.5% gross margin (Tier B, derived). ARR reported at $65B (July 2026, 70-75% API revenue per multiple outlets citing company statements, Tier B — https://webpronews.com/anthropics-revenue-explosion-to-65-billion-run-rate-signals-historic-ai-scale-ahead-of-record-ipo/). Confidential S-1 filed June 1, 2026 (Tier A, https://anthropic.com/news/confidential-draft-s1-sec). Colossus 1 capacity leased fully from xAI, effective May 6, 2026 (Tier A, https://x.ai/news/anthropic-compute-partnership; also https://anthropic.com/news/higher-limits-spacex).

5.2 OpenAI

No independent sustainability disclosure of any kind. Served via Microsoft Azure, predominantly on the Fairwater campus (Mount Pleasant, Wisconsin — closed-loop cooling, near-zero onsite water evaporation) and a newer Norway site (hydro-dominated grid, ~30 gCO2e/kWh vs. Wisconsin's ~470 gCO2e/kWh — a 15x carbon-intensity spread for identical hardware, purely from siting). Azure global PUE 1.18 / onsite WUE 0.32 L/kWh (Tier A, https://microsoft.com/sustainability). Fairwater GB200 NVL72 rack: 865,000 tokens/second per rack, 72 GPUs/rack, 1.8 TB/s NVLink (Tier A, https://blogs.microsoft.com/blog/2025/09/18/inside-the-worlds-most-powerful-ai-datacenter/). Stargate Project: $500B committed, 10 GW capacity target by 2029, 8+ sites (Tier A, https://openai.com/index/announcing-the-stargate-project/; https://openai.com/index/five-new-stargate-sites/). Sam Altman's own disclosed figure of 0.34 Wh per ChatGPT query (Tier C, an average across all query types, not a per-token figure — https://blog.samaltman.com/the-gentle-singularity) implies roughly 3,400 Wh/MTok at an assumed 100 tokens/query, broadly consistent with this report's 738-3,000 Wh/MTok current-generation range. Reverse-engineered COGS for prior-generation GPT-4o (8-way H100, 450 tokens/second): $3.358/MTok against a $4.375/MTok blended list price implies +23.2% gross margin (Tier B, derived). No current-generation margin disclosure exists; this report's price-bridge (Section 2.4) uses the prior-generation COGS-decomposition methodology as its closest available anchor.

5.3 Google (Alphabet Inc.)

The most disclosure-complete provider in this report. Fleet-wide trailing-twelve-month PUE 1.09 (Tier A, https://datacenters.google/efficiency/); most-efficient campus (Central Ohio, Lancaster) 1.04, least-efficient (Mesa, Arizona) 1.28. Onsite WUE 0.24 L/kWh (Tier A). Google's 2025 Environmental Report states a 64% volumetric freshwater replenishment rate for FY2024 — 4.5 billion gallons replenished against 7.03 billion gallons consumed (Tier A, https://sustainability.google/reports/google-2025-environmental-report/). The 2026 Environmental Report shows this has since improved to 78% / 7.7 billion gallons for FY2025 (Tier A, https://sustainability.google/reports/google-2026-environmental-report/) — a real, verified, and improving figure. Caveat: geographic non-fungibility — replenishment in one watershed (e.g. the Colorado River basin) does not offset water drawdown in a different, unconnected basin (e.g. Saint-Ghislain, Belgium). Median Gemini App text prompt: 0.25 Wh (Tier A, May 2025 disclosure — https://sustainability.google/reports/google-2025-environmental-report/ and https://deepmind.google Gemini technical pages). Annual token volume estimated 9-12 trillion tokens/year (Tier D, inferred from Google Cloud AI revenue, no direct disclosure). API pricing for Gemini 3.8 Flash: $0.75/$3.75 introductory (through Dec 31, 2026), rising to $1.50/$7.50 thereafter. 24/7 carbon-free-energy coverage approximately 67% of global operations (no tier stated in source).

5.4 Meta (Muse Spark/Glimmer and Llama families)

Both Meta AI lines remain active and are reported separately. Self-built fleet PUE 1.09, onsite WUE 0.20 L/kWh (Tier A, https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf). Muse Spark first released 2026-04-08 under Chief AI Officer Alexandr Wang; Muse Spark 1.3 released 2026-09-02 (Tier A/B, https://axios.com/2026/04/08/meta-muse-alexandr-wang; https://businessinsider.com/muse-spark-meta-ai-model-2026-4). Muse Glimmer is a 30B-parameter distilled variant. Llama 3.1 (8B/70B/405B) and Llama 4 Maverick remain separately served — Llama 4 Maverick trained at 34,000 PF-days, Llama 3.1 405B at 440,000 PF-days (Tier B, per Wikipedia's Llama model-family page, itself citing Meta's own technical reports — https://en.wikipedia.org/wiki/Llama_(language_model)). H100 fleet exceeded 350,000 chips in 2024 (Tier B). **A confirmed negative gross margin:** Muse Spark 1.3's own reverse-engineered COGS ($3.714/MTok on an H100 SXM5 cluster) exceeded its 2024-era $3.15/MTok blended list price, implying -17.9% gross margin (Tier B, derived from disclosed hardware); Meta's September 2026 pricing has since fallen further, to $1.25/$4.25 (blended $2.00/MTok), which — holding the same COGS constant — implies an even steeper -85.7% margin, consistent with a deliberate market-share subsidy strategy. Llama 3.1 405B third-party reseller hosting showed an even larger implied loss at 2024-era public prices ($3.50/$14.00 list vs. an unbatched 32-GPU serving cost estimated at $86.41/MTok, roughly -1,310% margin). Both figures should be read as "Meta is pricing below its own compute cost for market share," not as evidence that Meta's models are unusually cheap to run. Meta's Scope 1/2/3 2024 emissions: 8.2 million metric tons CO2e total (Scope 1: 47,468 t; Scope 2: 1,358 t; Scope 3: 8,151,769 t), with 50,000 metric tons removed via offsets (Tier A, https://sustainability.atmeta.com/climate/).

5.5 Microsoft

Global fleet PUE 1.17 (FY25), onsite WUE 0.27 L/kWh (Tier A, https://datacenters.microsoft.com/sustainability/efficiency/). Renewable power-purchase agreements totaling 40 GW across 26 countries (Tier A). Azure revenue $75B

FY25 (+34% year over year, Tier B). Microsoft Cloud gross margin approximately 69% (Tier B, derived). FY25 capital expenditure $64.6 billion (Tier A/B, https://microsoft.com/investor/reports/ar25/index.html). First-party MAI-1-preview model trained on approximately 15,000 H100 GPUs; specific training-compute-hours not disclosed (GAP). First-party token volume estimated 1.1 trillion tokens/year, excluding OpenAI pass-through traffic served on the same infrastructure (Tier D). Reverse-engineered Phi-4 COGS ($0.237/MTok on dual-GPU serving) against a $0.310/MTok blended list price implies +23.5% gross margin (Tier B, derived).

5.6 xAI

xAI's Memphis, Tennessee data center (Colossus 1, operational from June 2024 with roughly 100,000 Nvidia H100 GPUs) operated unpermitted natural-gas turbines — up to 35 mobile simple-cycle methane turbines from Solar Turbines and Caterpillar, roughly 420 MW nominal capacity, installed to bridge a delay in grid interconnection from Memphis Light, Gas and Water and the Tennessee Valley Authority (multiple corroborating sources: https://fortune.com/2024/09/03/elon-musk-xai-nvidia-colossus/; https://npr.org/2024/09/11/nx-s1-5088134/elon-musk-ai-xai-supercomputer-memphis-pollution). Estimated NOx emissions of 1,200-2,000 tons/year against a Title V major-source threshold of 100 tons/year — more than ten times over. Shelby County Health Department granted a partial permit for 15 of the 35 turbines in July 2025 (Tier A, https://scribd.com/document/883601734/01156-01PC-070225-Permit). At the second site, Colossus 2 in Southaven, Mississippi, the pattern repeated and escalated: the Southern Environmental Law Center and Earthjustice filed a 60-day Notice of Intent to Sue over 27 additional unpermitted turbines (Feb 13, 2026); Mississippi's environmental regulator permitted a 41-turbine, 1.2 GW permanent plant in March 2026 over public objection; the NAACP filed a federal civil-rights lawsuit in the Northern District of Mississippi on April 14, 2026; 19 more turbines were installed in May 2026, 8 of them after the lawsuit was filed; the Department of Justice moved to dismiss the NAACP suit on June 16, 2026; and xAI/SpaceXAI agreed on August 1, 2026 to remove 69 mobile turbines by July 2027 as part of a binding settlement (https://selc.org/press-release/groups-threaten-lawsuit-over-xais-unpermitted-gas-turbines-in-mississippi/; https://selc.org/press-release/mississippi-regulators-rubber-stamp-air-permit-for-xai-power-plant-ignoring-overwhelming-public-pushback/; https://selc.org/press-release/civil-rights-group-sues-xai-for-illegal-pollution-from-data-center-power-plant/; https://wired.com/story/xai-adds-19-new-gas-turbines-despite-ongoing-lawsuit/; https://nytimes.com/2026/06/16/climate/xai-musk-mississippi-grok-turbine-lawsuit-naacp.html; https://gizmodo.com/spacex-agrees-to-remove-temporary-gas-turbines-at-colossus-data-center-by-2027-2000793559). xAI's own air permit application discloses potential emissions exceeding 6,000,000 tons CO2e/year from this site alone, plus over 1,700 tons NOx, 180 tons PM2.5, 500 tons CO, and 19 tons formaldehyde annually (Tier A, https://selc.org/press-release/new-study-finds-proposed-xai-gas-plant-could-worsen-regional-air-pollution-cause-millions-of-dollars-in-annual-health-damages/). Total combined capacity across both sites, mid-2026: approximately 670,000 GPUs, 1.3 GW IT power. Tennessee grid carbon intensity (2024 EIA baseline) 365 gCO2e/kWh; cluster-weighted (including the off-grid gas generation) 400-430 gCO2e/kWh. 2025 financials: $3.2B revenue, -$6.4B operating loss (Tier B).

5.7 Mistral AI (France)

The only provider in this report with a real, third-party-audited, full lifecycle assessment — and the provider most at risk of an unfair comparison if that LCA is read as an operational figure. Mistral, environmental consultancy Carbone 4, and the French environment agency ADEME jointly published an 18-month LCA for Mistral Large 2, dated July 22, 2025 (Tier A, https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai/): 20,400 tonnes CO2e, 281,000 cubic meters of water, and 660 kg antimony-equivalent across the full 18-month training-plus-serving period; per a standard 400-token Le Chat response, this scales to 1.14 g CO2e and 45 mL of water — and, at 2,500 responses per million tokens, to approximately 2,850 g CO2e and 112.5 liters of water per million tokens. **This figure is cradle-to-grave** — training amortization and embodied hardware included — and is not directly comparable to this report's operational-only figures for every other provider (Section 4.1's 780-1,260 Wh/MTok, 40.6-51.7 gCO2e/MTok operational range for the same model). Scaleway's Paris DC5 facility (adiabatic free-cooling, zero evaporative towers) PUE approximately 1.15-1.20, onsite WUE approximately 0.02-0.078 L/kWh — among the lowest onsite water figures in this entire report. French grid location-based carbon intensity 41-52 gCO2e/kWh (nuclear-dominated), market-based as low as 14-22 gCO2e/kWh under guarantees of origin. Mistral Large 3 trained on 3,000 NVIDIA H200 GPUs (Tier A, https://mistral.ai/news/mistral-3/); also served from Sweden via EcoDataCenter, with a disclosed 1.17 gCO2e/kWh combined utility effectiveness metric and roughly 28% groundwater-sourced water usage (Tier A, https://ecodatacenter.tech/2025-sustainability-report) — the lowest carbon-intensity hosting location identified anywhere in this report. Estimated annual token volume 450 billion tokens (Tier D).

5.8 Cohere (Canada)

Hosted on AWS Canada Central and OVHcloud Beauharnois; PUE approximately 1.15, onsite WUE approximately 0.12 L/kWh (Tier B). Canadian grid location-based carbon intensity 28.0 gCO2e/kWh (Hydro-Québec, >99% renewable, plus Ontario nuclear), market-based as low as 10.0 gCO2e/kWh. Received a CAD $240 million Canadian Sovereign AI Compute award in December 2024 (Tier A, https://canada.ca/en/department-finance/news/2024/12/deputy-prime-minister-announces-240-million-for-cohere-to-scale-up-ai-compute-capacity.html). Operates a CoreWeave-hosted facility in Cambridge, Ontario, operational since August 2025 (Tier B). Annual revenue reported at $240M with ARR tripling in 2025 (Tier B, https://cnbc.com/2026/02/13/ai-startup-cohere-revenue-ipo.html). Reverse-engineered Command R+ COGS ($2.684/MTok on 8xH100) against a $4.375/MTok blended list price implies +38.7% gross margin (Tier B, derived). Current-generation model, Command A, has no direct disclosure at any tier and is modeled by analogy to Mistral (Tier D throughout).

5.9 DeepSeek

North China grid location-based carbon intensity approximately 450–500 gCO2e/kWh (East China average), national China average approximately 530 gCO2e/kWh. DeepSeek-V3 training cost, from the model's own published technical report: 2,788,000 H800 GPU-hours at an assumed $2/GPU-hour rental rate, totaling $5.576 million — $5.328M pretraining, $0.238M context extension, $0.01M post-training (Tier A, https://arxiv.org/abs/2412.19437, Table 1). This figure is explicitly stated by the paper's own authors to exclude prior research, architecture ablation experiments, and data-curation costs, so it understates true total development cost even though it is the strongest-sourced training-cost figure in this entire report. DeepSeekMoE architecture: 671 billion total parameters, 37 billion active per token, routed across 256 experts plus one shared expert; Multi-Head Latent Attention compresses the key-value cache from 4,096 bytes/token/layer (standard grouped-query attention) to 576 bytes/token/layer, a 7.1x memory-bandwidth reduction during decode — the architectural basis for DeepSeek's consistently efficient bottom-up figures relative to dense models of comparable parameter count. Reverse-engineered DeepSeek-V3 COGS ($0.124/MTok on H800, 2,400 tokens/second) against a $0.175/MTok blended list price implies +29.1% gross margin (Tier B, derived). Current-generation V4.1-Flash and V4-Pro pricing implies GPU-hour usage 10–25x lower than an independent bottom-up estimate for the same model class (Section 4.2) — treat DeepSeek's current API pricing as likely reflecting aggressive, possibly below-cost, market-share-driven pricing rather than a reliable proxy for physical resource cost.

5.10 Alibaba (Qwen family)

Self-built fleet average PUE 1.200 (Tier A), with immersion cooling at the Zhangbei facility reaching 1.150 and 0.15 L/kWh onsite WUE (near-zero evaporative loss); fleetwide average onsite WUE approximately 0.35 L/kWh. Clean-energy sourcing 56.0% (audited 2024, wind and solar power-purchase agreements), yielding a market-based carbon factor of approximately 299.2 gCO2e/kWh against a regional baseline of 680.0 gCO2e/kWh (Tier A/B, https://alibabagroup.com/document-1752073403914780672). Qwen ecosystem scale: over 200,000 Hugging Face model derivatives, 234 million app users as of May 2026 (Tier B). Qwen3 pretraining used 36 trillion tokens, released April 29, 2025; no GPU-hours or training cost disclosed (https://arxiv.org/abs/2505.09388, Tier A for the token count, GAP for training compute). Reverse-engineered Qwen2.5-72B COGS ($0.435/MTok on 4-way H800/H20) against a $0.600/MTok blended list price implies +27.5% gross margin (Tier B, derived). Hosted primarily in Zhangjiakou (Hebei) and Hangzhou (Zhejiang).

5.11 Moonshot AI (Kimi family)

Hosted on Volcano Engine and Beijing Super Cloud infrastructure; PUE approximately 1.20, WUE approximately 0.45 L/kWh (Tier B, analogy-based). Annual token volume estimated 420 billion tokens (Tier D). Current-generation K3 pricing: $0.30 cache-hit / $3.00 cache-miss / $15.00 output per million tokens (Tier A, live as of September 10, 2026). K2.6 pricing: $0.16/$0.95/$4.00; K2.7 Code: $0.19/$0.95/$4.00. **Kimi K2 Thinking's training cost of $4.6 million is the weakest-sourced headline figure in this entire report** — it traces to Wikipedia, which in turn cites unnamed Chinese-language press, not Moonshot's own technical paper or blog. Treat with real caution, and note it may reflect only an incremental K2-to-K2-Thinking continuation-training run rather than a full pretraining cost. Mooncake serving system throughput: 100 billion tokens/day (Tier A, a USENIX FAST 2025 Best Paper award). Estimated ARR $200–300M (2026), $35B valuation (July 2026), with Alibaba holding an approximately 36% stake (Tier D).

5.12 Zhipu AI / https://Z.ai (GLM family)

Beijing-based BigModel infrastructure; PUE approximately 1.21, WUE approximately 0.40 L/kWh (Tier B, analogy-based). Listed on the Hong Kong Stock Exchange (2513.HK) January 8, 2026, raising $558 million; post-IPO market capitalization reached $62 billion by August 2026 (Tier A). GLM-4.5 pretrained on 23 trillion tokens (Tier A, https://arxiv.org/abs/2508.06471); GPU-hours not disclosed. 2024 revenue RMB 312.4 million (~$43M); 2025 first-half revenue RMB 190.9 million (~$26M); R&D spend RMB 2.2 billion, 703% of revenue; gross margin declining from 56.3% to 50%; cumulative losses RMB 6.25 billion — all Tier A, from the IPO prospectus. Daily token volume grew from 500 million (2023) to 4.6 trillion (June 2025) — Tier A. GLM-4.5 API pricing $0.11/$0.28 per million tokens per a CNBC interview with the company (Tier B, https://cnbc.com/2025/07/28/chinas-latest-ai-model-claims-to-be-even-cheaper-to-use-than-deepseek.html), with a 10% price increase in April 2026.

5.13 ByteDance (Doubao family)

Volcano Engine infrastructure, split between Zhangjiakou and western-China facilities; PUE approximately 1.18, WUE approximately 0.30 L/kWh (Tier B, analogy-based; the primary citation this figure rests on could not be independently confirmed as a genuine Volcano Engine source and should be treated with caution pending re-verification). Daily token volume grew explosively: 120 billion/day (May 2024) to 120 trillion/day (April 2026) — a 1,000x increase in two years (Tier A). A Malaysia deployment of approximately 36,000 Nvidia B200 chips was announced March 2026 via Aolani Cloud, but this claim appears in narrative sourcing without a single itemized citation and should be treated as weakly sourced pending re-verification. Monthly active users peaked at 330 million in May 2026, then dropped 6.1 million after a subscription paywall (¥68/month standard, ¥500/month premium) was introduced (Tier B). Doubao Seed 2.0's Volcano Engine pricing page is JavaScript-rendered and was not machine-extractable at research time; current pricing could not be independently verified beyond the figures used in this report's tables.

5.14 Baidu (Ernie family)

Primary AI data center in Yangquan, Shanxi province, built 2012 at a cost of RMB 4.708 billion — a coal-heavy grid region, location-based carbon intensity approximately 700 gCO2e/kWh. Audited PUE 1.08-1.15, WUE 0.28 L/kWh at the Yangquan facility (Tier A). The Kunlun P800 chip cluster (30,000 chips) was brought online in April 2025 (Tier A). Qianfan-VL model training achieved greater than 90% scaling efficiency on 5,000+ P800 chips, with a model-flop-utilization figure of 47% for the vision-language training component specifically (Tier A, https://arxiv.org/abs/2509.18189) — this MFU figure covers only that training subset, not Baidu's training operations generally. Current-generation pricing (Qianfan platform, live): Ernie 5.0 ¥6-10/¥24-40 per million tokens; Ernie X1.1 Preview ¥1/¥4; Ernie 4.5 Turbo ¥0.80/¥3.20 — all Tier A.

5.15 Tencent (Hunyuan family)

Data centers at Guizhou (Qixing Cave and Qingyuan campuses); audited PUE 1.246, WUE 0.35 L/kWh (Tier A). Global infrastructure footprint: 66 availability zones across 23 regions; a claimed corporate-average PUE below 1.3. Hunyuan-Large is a 389-billion-parameter Mixture-of-Experts model; Hunyuan Video training used a 13-billion-parameter model on 16,000 H800 GPUs (Tier B, https://arxiv.org/abs/2412.03603). 2025 revenue CNY 751.8 billion (the underlying annual-report PDF was not machine-extractable at research time; the figure is independently available at https://static.www.tencent.com/uploads/2026/04/09/62d786fcf3d3c8cb7e54791ee95439ac.pdf).

5.16 Amazon Web Services (Bedrock)

Global fleet PUE 1.14, onsite WUE 0.12 L/kWh (Tier A, https://sustainability.aboutamazon.com/products-services/aws-cloud). Achieved 100% renewable-energy matching in 2025, its third consecutive year (Tier A). Deploys custom Inferentia2 (inf2.48xlarge, 2,600 W per node), Trainium2 (trn2.48xlarge, 9,600 W per node), and third-party H100 (p5.48xlarge) instances for AI serving; Project Rainier's Trainium2 buildout totals approximately 500,000 chips, with Anthropic alone using over 1 million chips across AWS infrastructure in aggregate. Trainium3 UltraServer specifications: 144 chips, 362 MXFP8 PFLOPs, 144 GB HBM3e per chip. AWS's Q2 2026 segment results: $42.2 billion revenue (+37% year over year), $16.6 billion operating income, a $169 billion annualized run rate; the AI-specific business line exceeds a $25 billion annualized run rate, as does the chips business line separately (Tier A). Resale economics for third-party Llama 3.1 70B hosting: blended API list price $0.720/MTok against a $0.315/MTok COGS implies +56.3% gross margin (Tier A/B, derived).

5.17 CoreWeave

43 data centers (December 2025), growing to 49 (March 31, 2026) and 51 (September 2026), per CoreWeave's own filings (Tier A). Customer concentration is extreme: Microsoft accounted for 67% of fiscal-year-2025 revenue; OpenAI holds a contract worth approximately $6.5 billion through May 2031, Meta approximately $14.2 billion through December 2031 (Tier A). Operates a 36,000-GPU GB200 cluster in partnership with Hypertec Cloud. FY2025 revenue $5.131 billion against $1.453 billion cost of revenue, a gross margin of approximately 71.7% (Tier A, derived directly from CoreWeave's own reported line items — https://companiesmarketcap.com/coreweave/sec-reports-10k/0001769628-26-000104/). Colocation average PUE approximately 1.28, direct WUE approximately 1.25 L/kWh (Tier B); raw compute sold at $2.20-2.75 per GPU-hour, with a resale gross margin of 25-38% on that specific product line (Tier B, distinct from the 71.7% company-wide figure — CoreWeave's overall margin blends multiple product lines).

5.18 DeepInfra

Annual token volume approximately 260 trillion tokens/year (5 trillion/week, per the company's own CEO, Nikola Borisov, in a recorded interview). Total funding $133 million: $8M seed (November 2023), $18M Series A (April 2025), $107M Series B (May 4, 2026, co-led by 500 Global and Georges Harik — Tier A). Nine data centers as of July 2026: eight in the US plus a Toronto facility (eStruxture, 1.7 MW, 1,000+ Nvidia B300 GPUs — Tier A). PUE approximately 1.25, WUE approximately 1.20 L/kWh (Tier B, analogy-based). Resale economics for Llama 3.1 70B: blended pricing $0.155/MTok against $0.118/MTok COGS implies +23.9% gross margin (Tier B, derived) — this margin figure is used throughout Section 4.2's price-bridge for every model DeepInfra resells.

5.19 Together AI

DeepSeek-R1-0528 throughput: 334 tokens/second serverless, 386 tokens/second dedicated (batch size 1), on HGX B200 hardware; Qwen3-235B served 2.75x faster than the next-fastest provider in the same comparison (Tier A). Raised $102.5 million in a Series A (November 2023) and $800 million in a Series C (July 1, 2026 — Tier A). Committed an additional 500 MW of capacity in July 2026. Operates a 36,000-GPU GB200 cluster with Hypertec Cloud, announced November 2024 (Tier A). PUE approximately 1.25, WUE approximately 1.20 L/kWh (Tier B, analogy-based). Resale economics for Llama 3.1 70B: blended pricing $0.355/MTok against $0.228/MTok optimized COGS implies +35.8% gross margin (Tier B, derived).

5.20 Fireworks AI

Daily token volume 40 trillion, greater than 95% from customer-specialized (fine-tuned) models rather than off-the-shelf base models (Tier A). $1 billion annualized recurring revenue; raised $1.505 billion at a $17.5 billion valuation in a Series D round, July 15, 2026 (Tier A). Operates in 26 named regions across 8 countries (Tier A). Pricing structure: batch processing at -50% of standard rate, priority processing at +25-50%, "fast" tier at +50%, US-only deployment at +50%, single-region deployment at a 1.5x multiplier. B200 GPU pricing rose from $10/hour to $13/hour in September 2026 (Tier A) — this is the rate this report's price-bridge uses for every B200-class model (Section 2.4). PUE approximately 1.25, WUE approximately 1.20 L/kWh (Tier B, analogy-based). Resale economics for Llama 3.1 70B: blended pricing $0.375/MTok against $0.228/MTok optimized COGS implies +39.2% gross margin (Tier B, derived).

5.21 Groq

Custom LPU (Language Processing Unit) architecture — 576 chips per pod for 70B-class model serving in the prior generation, 256 LPUs per rack in newer deployments, 315 PFLOPS FP8 per rack, a claimed 40 PB/s aggregate SRAM bandwidth (Tier A). **The defining physical characteristic behind this report's Groq correction (Methodology 2.3): on-chip SRAM draws 140-160 kW continuously, whether or not the chip is actively serving a token, because static memory leakage does not scale down with load.** This makes Groq's real per-token resource cost extremely sensitive to utilization — a fact any efficiency claim for this hardware class must state its assumed utilization level to be credible. Single-user generation speed 260-310 tokens/second; a networked pod of 576 chips is required to serve a single 70B-class model request stream at production latency. Founder Jonathan Ross departed to NVIDIA in December 2025, in a deal reported at a $20 billion license value (Tier A); Adam Winter is the current CEO. Raised a $350 million Series A at a $3.5 billion valuation in August 2026 (Tier A). PUE approximately 1.30, WUE approximately 1.20 L/kWh (Tier C, analogy-based). Resale economics: blended pricing approximately $0.640/MTok, implied gross margins 30-45% (Tier C).

5.22 Cerebras

Wafer-Scale Engine architecture: the CS-3/WSE-3 chip is a single 46,225 mm^2 monolithic wafer with 900,000 AI cores, 44 GB of on-chip SRAM, and 21 PB/s of memory bandwidth. The newer WSE-3T/CS-4 configuration ships 3 wafers per rack at 250 PFLOPS per wafer, 43.2 PB/s memory bandwidth, 53.5 PB/s fabric bandwidth, and 2-microsecond wafer-to-wafer latency (Tier A). A single CS-3 chassis draws 23 kW continuous (19.5-21.5 kW operating range) and delivers approximately 2,100 tokens/second on a Llama 3.1 70B-class model, completing one million tokens in roughly 476 seconds. Listed on Nasdaq (CBRS) May 14, 2026, at a peak market capitalization near $95 billion (Tier B). Current-generation pricing: GPT-OSS-120B $0.25/$0.69 per million tokens; Qwen3-32B $0.40/$0.80; Qwen3-235B $0.60/$1.20 (Tier A). Announced a 165 MW facility in Mikkeli, Finland, and a 750 MW / $10 billion deal with OpenAI (Tier D, the OpenAI figure specifically is unconfirmed). PUE approximately 1.25, WUE approximately 1.10 L/kWh (Tier B, analogy-based).

---

6. Limitations and Disclosure Gaps

No frontier US lab publishes operational disclosure. Anthropic, OpenAI, and xAI publish zero facility-level PUE/WUE, zero disclosed token volume, and zero CDP water or climate response. Every inference or training figure for these three providers in this report is modeled (Tier C or D at best), cross-checked only against their cloud hosts' (AWS, Microsoft, Google) fleet-wide disclosed averages, which are hyperscaler-wide figures, not AI-workload-specific ones. This is the single largest disclosure gap in the report and should be treated as a hard ceiling on confidence for any Anthropic, OpenAI, or xAI number, regardless of how precisely it is stated elsewhere in this document.

Eleven of twenty-two current-generation provider-model lines have no training-compute disclosure at any tier (Section 4.4): Anthropic (both current models), OpenAI (both current models), Google, Meta's Muse Spark 1.3, Microsoft's MAI-1-preview, Cohere's Command A, DeepSeek's V4 lineage, Alibaba's Qwen3 family, Moonshot's K3, Zhipu's GLM-4.5, and Tencent's Hunyuan Hy4. These are stated as GAP, not filled with an industry average presented as provider-specific.

A published claim of sub-1-Wh-per-million-token inference for LPU/wafer-scale hardware should be treated with real skepticism unless it states its assumed batching and utilization level (Methodology 2.3).

A scaled lifecycle-assessment figure should never be compared directly against an operational-only figure without restating its scope (Methodology 2.3, Section 5.7).

Current-generation figures carry structurally wider uncertainty than prior-generation figures. Several current-generation models (Claude Fable/Mythos, GPT-6 Astra, Gemini 3.8 Flash, Kimi K3, GLM-5.3, Qwen3.8-2.4T-A95B, Gemini Omni Flash) have no operator disclosure of any kind and no prior-generation figure to anchor a bottom-up estimate from the same architecture family — every figure for these specific models is Tier D.

The financial reverse-engineering path (Section 2.4) is least reliable exactly where a provider prices aggressively below normal margin for market share — confirmed for Meta (Muse Spark 1.3, a directly measured negative margin) and suspected for DeepSeek's current generation (a 10-25x divergence between the price-bridge and an independent bottom-up estimate). Any price-bridge-derived figure in Section 4.2 should be read as a lower bound on true resource cost wherever a provider is known or suspected to be pricing below compute cost, not as a point estimate.

Regional grid-carbon and grid-water intensity constants used throughout this report (US national/regional, EU/France, China, Canada) are cross-provider defaults, not facility-specific measurements, except where a specific provider's per-provider appendix entry (Section 5) states a facility-specific figure with its own citation.

Sourcing standard. Per the sourcing hierarchy this research applied — audited regulatory filings and CDP responses first, then sustainability reports, then peer-reviewed or benchmark data, then reputable journalism, then vendor documentation — forum posts, unsourced social-media posts, anonymous pricing-aggregator sites, and bare unattributed links were excluded from evidentiary use throughout. Where such a source was the only available basis for a claim, the claim is either omitted or explicitly marked as having an unresolvable citation in the relevant table cell (for example: Anthropic's Claude 3.7 Sonnet prior-generation row, Section 4.1; Mistral's Codestral row; ByteDance's Doubao-pro row; Moonshot's and Zhipu's prior-generation rows).