REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of OpenAI ChatGPT Models: Methodology, Estimates, and Justification

Prepared by: **InferenceCarbon.ai Reference Team**

Date: 17 March 2026 | Version: 1.3 | Classification: For publication — subject to stated caveats

Important note: *There is severely limited publicly available data on the gCO_2e impact of OpenAI's models. Unlike Google, OpenAI has not published a technical paper on inference energy. Additionally, OpenAI has shifted from a single cloud provider (Microsoft Azure) to a multi-cloud infrastructure spanning Azure, Oracle, and AWS, complicating carbon estimation. This document shows the methodology behind our estimates for OpenAI's ChatGPT model family. All estimates are presented as ranges to reflect the substantial underlying uncertainty.* ***We welcome verified real-world data from OpenAI, Microsoft, Oracle, or AWS that will help us improve estimates.***

Scope notice: *This document estimates operational inference emissions only. It does not constitute a lifecycle assessment and excludes embodied emissions of hardware, model training, and upstream supply chain.*

1. Executive Summary

This document provides carbon intensity estimates for OpenAI's ChatGPT model family, expressed in grams of CO_2 equivalent per 1,000 tokens (gCO_2e/1k tokens). No published source provides per-model figures. Our estimates are derived from Sam Altman's June 2025 disclosure that the "average ChatGPT query uses about 0.34 watt-hours,"[1] cross-validated against Epoch AI's independent estimate of 0.30 Wh[2] and Jegham et al.'s simulation figure of 0.42 Wh for a short GPT-4o prompt.[3]

The analysis uses location-based carbon accounting as the default (Section 8), with a Clean Energy Adjustment Factor (CEAF) to derive supplementary CEAF-adjusted figures for decision support (Section 9). All estimates are presented as ranges (low/central/high) because the anchor's token-count assumption alone produces a 2.5× spread. This document has been subjected to adversarial review (Section 14).

Our best estimate of the gCO_2e of OpenAI's models is shown in Table 1, below. The "Adjusted Central" value shows gCO_2e/1k tokens adjusted for OpenAI's/OpenAI's suppliers' clean energy use. Note that the ongoing shift from a Microsoft Azure-based infrastructure to a more heterogeneous mix may result in an increase in the amount of Carbon emitted per 1,000 tokens (Section 8).

[1]Sam Altman, "The Gentle Singularity," personal blog, 10 June 2025. https://blog.samaltman.com/the-gentle-singularity
[2]Josh You, "How much energy does ChatGPT use?" Epoch AI Gradient Updates, 7 February 2025. https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use
[3]Jegham, N. et al., "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference," arXiv:2505.09598, May 2025 (updated November 2025). https://arxiv.org/abs/2505.09598

Model	Location-Based Range (gCO_2e/1k tokens)	Central Scenario	CEAF %	Adjusted Central	Confidence
GPT-4.1 nano	0.15 – 0.23	0.19	0.50	**0.10**	High (Jegham sim.)
GPT-4.1 mini	0.34 – 0.49	0.42	0.50	**0.21**	High (Jegham sim.)
GPT-4.1	0.53 – 1.09	0.81	0.50	**0.40**	High (Jegham sim.)
GPT-4o	0.31 – 0.47	0.39	0.50	**0.20**	High (Jegham sim.)
GPT-4o-mini	0.41 – 0.66	0.53	0.50	**0.27**	High (Jegham sim.)
o4-mini (high)	2.02 – 4.73	3.38	0.50	**1.69**	Med (Jegham sim.)
GPT-5 (high)	0.25 – 1.30	0.65	0.50	**0.33**	V. Low (est.)
GPT-5.4 (xhigh)	0.20 – 3.40	1.50	0.50	**0.75**	V. Low (est.)

Table 1: Summary estimates. GPT-4.1 and GPT-4o models simulated by Jegham et al. (arXiv:2505.09598); GPT-5 models estimated via TPS scaling. All figures in gCO_2e/1k tokens, location-based. CEAF = 0.50 for Microsoft Azure, based on verified 100% annual renewable matching (2025), discounted for lack of published hourly CFE data. Since OpenAI now also runs inference on Oracle and AWS infrastructure (Section 3.5), the true blended CEAF may be materially lower. CEAF-adjusted figures are supplementary decision-support metrics, not substitutes for the location-based physical estimate.

2. What Has Been Published

OpenAI is not as transparent as Google regarding inference energy data. The key disclosures are:

Sam Altman blog post (10 June 2025): "The average query uses about 0.34 watt-hours."[4] Published in a wide-ranging personal blog post titled "The Gentle Singularity." No methodology, no model specification, no breakdown by modality. The figure preceded GPT-5's release (7 August 2025) and almost certainly reflects a traffic-weighted average across all ChatGPT models, input lengths, and modalities.

Epoch AI (February 2025): Independent first-principles estimate of ~0.30 Wh for a GPT-4o query with 500 output tokens, assuming ~100B active parameters (MoE), H100 GPUs at 10% utilisation, 70% power factor, and ~1,500 W per GPU inclusive of data centre overhead.[5] Explicitly acknowledged as "pessimistic" (erring toward higher costs); true figure could be as low as 0.1 Wh.

Jegham et al. (May 2025, arXiv:2505.09598): Infrastructure-aware benchmarking of 30 models. Simulated GPT-4o at 0.42 Wh (short prompt) and 1.79 Wh (long prompt). Found GPT-4o-mini consumed slightly more energy per query than GPT-4o despite being smaller, attributing this to deployment on less efficient A100 hardware.[6] Ranked Claude-3.7 Sonnet highest in eco-efficiency (0.886).

University of Rhode Island AI Lab (August 2025): Estimated GPT-5 at 18.35 Wh average per medium-length response (~1,000 tokens), with some reaching 40 Wh.[7] Methodology: response times multiplied by assumed hardware power consumption. Widely cited but highly indirect, and critically conflates hidden reasoning tokens with visible output.

No other OpenAI disclosure exists. The company has not published a sustainability report, per-model energy data, or Scope 1/2/3 emissions.[8]

[7]The Guardian, "OpenAI will not disclose GPT-5's energy use," 9 August 2025; Tom's Hardware, "ChatGPT 5 power consumption could be 8× higher," 14 August 2025. https://www.tomshardware.com/tech-industry/artificial-intelligence/chatgpt-5-power-consumption-could-be-as-much-as-eight-times-higher-than-gpt-4

[8]Earth911, "Your AI Carbon Footprint: What Every Query Really Costs," March 2026. Notes Anthropic has not reported Scope 1, 2, or 3 emissions. https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/

3. Anchor Derivation

3.1 Which model was simulated?

Altman's 0.34 Wh refers to the "average ChatGPT query" as of June 2025. At that time, GPT-4o was the default model for all ChatGPT subscribers.[9] GPT-4o-mini was the fallback for free-tier users. The o-series reasoning models were available but less popular. We treat the figure as GPT-4o-weighted, but acknowledge it is a fleet average that includes multimodal queries, tool use, variable input lengths, and free-tier fallback behaviour.

3.2 The critical unknown: token count

Altman did not disclose the token count of the "average query." However, Jegham et al. (arXiv:2505.09598) provide direct energy simulations for GPT-4o and many other OpenAI models, largely superseding the need to derive per-token estimates from the Altman anchor. We now use Jegham's simulated values as the primary data source for models they cover, and retain the Altman anchor only for cross-validation. Epoch AI assumed 500 output tokens;[10] Chiang et al. found 261 tokens average in chatbot conversations.[11] We use 300 as a central assumption, consistent with our Gemini methodology, but present the full sensitivity:

Assumed tokens	Wh/1k tokens	gCO_2e/1k tokens	Note
200	1.70	0.63	Upper bound
300 (central)	1.13	0.42	Consistent with Gemini
400	0.85	0.31	Moderate responses
500 (Epoch AI)	0.68	0.25	Lower bound

Table 2: Sensitivity of anchor to token-count assumption. The 2.5× range is not a secondary sensitivity — it is a structural uncertainty that dominates most model-to-model differences reported in this document.

3.3 Per-token derivation

Energy: From Jegham: GPT-4o short prompt (400 tokens) = 0.423 Wh. Per 1,000 tokens: 0.423 ÷ 400 × 1,000 = 1.06 Wh per 1,000 tokens

Location-based carbon: 1.06 × 0.370 gCO_2e/Wh = **0.39 gCO_2e per 1,000 tokens**[12]

Anchor range: 0.31–0.47 gCO_2e/1k tokens (±1 SD from Jegham). This figure applies to GPT-4o as simulated by Jegham et al. on Azure infrastructure, using a US-weighted grid intensity of 370 gCO_2e/kWh.

3.4 Cross-validation

Three independent sources bracket the anchor. They do not share the same workload definition, token basis, or system boundary, so this is directional consistency, not strict convergence:

Altman: 0.34 Wh/query (fleet average, unknown model/tokens/boundary)

Epoch AI: 0.30 Wh/query (GPT-4o, 500 tokens, server-only, first-principles)[13]

Jegham et al.: 0.42 Wh/query (GPT-4o, ~300 tokens, infrastructure-aware, includes PUE)[14]

[9]OpenAI, "Introducing GPT-5," 7 August 2025. States GPT-5 with thinking performs better than o3 with 50–80% fewer output tokens. https://openai.com/index/introducing-gpt-5/
[11]Chiang, W.-L. et al., "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," arXiv:2403.04132, March 2024. Found average response length of 261 tokens.
[12]EPA eGRID 2022: US national average 373 gCO_2/kWh. https://www.epa.gov/egrid

All three fall within 0.30–0.42 Wh. The ~40% spread is consistent with different system boundaries. The order of magnitude is reliable; the precise per-token figure remains uncertain by 2–3×.[15]

3.5 Infrastructure context: the multi-cloud shift

A critical development since Altman's June 2025 disclosure is that OpenAI has moved from a single-provider to a multi-cloud infrastructure. From 2019 to mid-2025, Microsoft Azure was OpenAI's exclusive cloud provider.[16] This exclusivity has now ended. As of March 2026, OpenAI operates inference and training workloads across at least three major providers:

Microsoft Azure ($250B commitment, ~6 years): Remains the largest provider by dollar volume and retains exclusive OpenAI API access through 2030. Still hosts the majority of ChatGPT inference as of early 2026.

Oracle / Stargate ($300B, 5 years): The flagship Abilene, Texas campus has been operational since September 2025, with two buildings running Nvidia GB200 racks for both training and inference workloads.[17] Six additional buildings are scheduled for mid-2026, bringing total capacity to 1.2 GW.[18]

Amazon AWS ($38B, 7 years): Signed November 2025 with immediate deployment. Full capacity targeted by end 2026. Provides access to hundreds of thousands of Nvidia GPUs.[19]

CoreWeave ($22.4B through 2029): Primarily training compute. Operational.[20]

This has three consequences for the estimates in this document:

First, the anchor remains valid. Altman's 0.34 Wh figure (June 2025) predates both the Oracle Stargate operational launch (September 2025) and the AWS deal (November 2025). At the time of disclosure, Azure was still the exclusive provider. The anchor therefore maps to Microsoft Azure infrastructure.

Second, the grid carbon intensity may now be higher than assumed. We use a US-weighted average of 370 gCO_2e/kWh based on Azure's data centre concentration in Virginia, Iowa, and Arizona. However, Oracle's Stargate flagship is in Abilene, Texas, on the ERCOT grid, which has a carbon intensity of approximately 350–420 gCO_2e/kWh depending on wind conditions. If a material share of current ChatGPT inference runs on Oracle in Texas, the effective grid intensity may be higher than our assumption, not lower. This means our location-based estimates may slightly understate current emissions.

Third, the CEAF becomes substantially harder to defend as a single number. The CEAF of 0.50 in this document is based on Microsoft's verified 100% annual renewable matching. Oracle has not published comparable clean energy disclosure for Stargate — its public materials state only a commitment to "grid-friendly and responsible energy practices," which is effectively Tier 3 language. Amazon's AWS has its own renewable energy programme but operates at a different scale and transparency level from Microsoft. Without knowing the inference traffic split across providers, we cannot calculate a defensible blended CEAF. The true blended CEAF is almost certainly lower than 0.50, meaning our CEAF-adjusted figures may understate actual emissions. See Section 9 for the detailed CEAF discussion.

This multi-cloud shift strengthens the case for leading with location-based gross figures (Layer 2), which are less affected by the provider mix because all three operate primarily in the US with broadly similar grid intensities. The CEAF-adjusted figures (Layer 3) should be treated with additional caution.

---

[15]Towards Data Science, "Let's Analyze OpenAI's Claims About ChatGPT Energy Use," June 2025. Analyses Altman's 0.34 Wh claim against Epoch AI and Jegham data. https://towardsdatascience.com/lets-analyze-openais-claims-about-chatgpt-energy-use/

[16]Built In, "OpenAI's $1T Infrastructure Plan," November 2025. Documents the shift from Microsoft-exclusive to multi-cloud (Azure $250B + Oracle $300B + AWS $38B + CoreWeave $22.4B). https://builtin.com/articles/openai-cloud-deals

[17]OpenAI–Oracle Stargate: $300B over 5 years. Abilene TX campus operational September 2025 with GB200 racks running training and inference. CNBC, 23 September 2025.
https://www.cnbc.com/2025/09/23/openai-first-data-center-in-500-billion-stargate-project-up-in-texas.html

[18]Oracle Stargate Abilene: 2 buildings operational, 6 more by mid-2026, total 1.2 GW on ERCOT grid (Texas). Tom's Hardware, March 2026. https://www.tomshardware.com/tech-industry/oracle-and-openai-scrap-planned-600mw-abilene-expansion

[19]OpenAI–AWS deal ($38B, 7 years), signed November 2025. Full capacity targeted by end 2026. SiliconANGLE, 3 November 2025. https://siliconangle.com/2025/11/03/openai-inks-38b-ai-infrastructure-deal-aws/

4. Scaling to Other OpenAI Models

4.1 Approach and caveats

We scale from the GPT-4o anchor using output throughput (tokens per second) from Artificial Analysis[21] as the primary proxy. Slower throughput at similar power draw implies more compute per token. We use API pricing[22] as a secondary sanity check for commercial segmentation only — pricing is a business signal, not a physics signal.

Known confounds in throughput-based scaling include: hardware generation and binning (different models may run on different GPU types); GPU count (smaller models may run on fewer GPUs than larger ones); batch size and scheduler behaviour; KV-cache and memory bandwidth effects; routing policies (GPT-5 uses adaptive model routing); visible vs hidden token generation; and product settings that change latency targets. The resulting estimates should be read as ranges.

4.2 Generational Efficiency Correction (GEC)

We apply a GEC range of 0.80–0.90 for GPT-5.x models relative to GPT-4o (10–20% efficiency gain). This is based on[23]OpenAI's statement that GPT-5 achieves comparable performance to o3 with 50–80% fewer output tokens[24]. We use midpoint 0.85 for central estimates, although technologies such as GPT-5's adaptive routing may mean that this should be lower.

4.3 Throughput data

Model	Output t/s	Ratio vs 4o	Input $/M	Output $/M
GPT-4o (anchor)	n/a	—	$2.50	$10.00
GPT-4o-mini	36.4	0.39×	$0.15	$0.60
GPT-5 (high)	64.5	—	$1.25	$10.00
GPT-5 Mini (high)	81.8	—	$0.25	$2.00
GPT-5.4 (xhigh)	77.4	—	$2.50	$15.00
GPT-5.3 Codex	62.0	—	$1.75	$14.00
o3-mini (high)	n/a	—	$1.10	$4.40

*Table 3: Throughput and pricing data. Throughput from Artificial Analysis (fn 7). Pricing from OpenAI (fn 8). *o3-mini visible throughput is misleading; hidden reasoning tokens treated separately in Section 7.*

4.4 Per-model derivation

GPT-4o-mini: Jegham et al. simulated GPT-4o-mini at 0.577 Wh per short prompt (400 tokens), giving 0.53 gCO_2e/1k tokens — higher than GPT-4o (0.39) despite being smaller, due to deployment on less efficient A100 hardware.[25] Central estimate 0.53, range 0.41–0.66 gCO_2e/1k tokens.

GPT-5 (standard): Scaled from GPT-4o-mini (Jegham+TPS): 0.53 × (36.4/64.5) × 0.85 = 0.25 gCO_2e/1k (TPS lower bound). With reasoning overhead of ~2.5× for "high" mode, central ~0.65. GPT-5 "high" is a reasoning effort level; actual energy includes hidden reasoning tokens not captured by TPS.[26] Range: 0.25–1.30.

GPT-5 Mini: Scaled from GPT-4o-mini: 0.53 × (36.4/81.8) × 0.85 = 0.20 gCO_2e/1k (TPS lower bound). With reasoning overhead, central ~0.50. Range: 0.20–1.00.

[21]Artificial Analysis, model throughput and pricing data (accessed March 2026). https://artificialanalysis.ai/models/
[22]OpenAI API Pricing (accessed March 2026). https://openai.com/api/pricing/
[23]

GPT-5.4 and o3/o4-mini: Reasoning models with very long hidden token chains. TTFT of 169s for GPT-5.4 at high effort indicates very significant hidden compute. Treated as multiplier-based ranges in Section 7.

5. Three-Layer Disclosure Structure

Following adversarial review, all estimates are reported across three separated layers:

Layer 1 — Operational Electricity (Wh/1k tokens): Energy consumed by GPU servers, infrastructure, and data centre overhead (PUE - Power Usage Effectiveness). Most physically grounded layer.

Layer 2 — Location-Based Carbon (gCO_2e/1k tokens): Layer 1 × grid carbon intensity. Primary carbon metric for transparency reporting.

Layer 3 — CEAF-Adjusted (gCO_2e/1k tokens): Layer 2 × (1 − CEAF). Supplementary metric acknowledging provider clean energy investment. Should not be sole basis for offsetting claims without explicit disclosure.

The bridge between layers is explicit: users who disagree with the grid intensity or CEAF can substitute their own values while retaining the Layer 1 energy estimate.

6. Special Case: Thinking and Reasoning Models

GPT-5 incorporates adaptive reasoning by default and is presented by OpenAI as a unified system rather than separate models.[27] The boundary between "standard" and "thinking" is a continuous spectrum of reasoning effort, not a binary switch. Newer API documentation shows reasoning effort controls with a default of "none" for standard usage.

Epoch AI found o1 and o3-mini generated ~2.5× as many total tokens as GPT-4o for the same questions.[28] The URI estimate of 18.35 Wh implies ~40× our anchor — plausible only for maximum-effort reasoning.[29]

Jegham et al. provide simulated energy data for several reasoning models (short prompt, 400 tokens): o3: 1.18 Wh (1.09 gCO_2e/1k tokens); o3-mini: 0.67 Wh (0.62 gCO_2e/1k); o3-mini (high): 3.01 Wh (2.79 gCO_2e/1k); o1: 2.27 Wh (2.10 gCO_2e/1k); o1-mini: 0.54 Wh (0.49 gCO_2e/1k); o4-mini (high): 3.65 Wh (3.38 gCO_2e/1k). These provide direct cross-validation for the multiplier-based estimates below. For example, o3 at 1.09 gCO_2e/1k is ~2.8× the GPT-4o baseline (0.39), consistent with the "Low" thinking scenario. o3-mini (high) at 2.79 gCO_2e/1k is ~7× baseline, between "Low" and "Medium".

Scenario	Mult.	GPT-5 Wh	gCO_2e/1k vis.	Notes
Minimal	2×	0.70–1.40	0.85–1.70	Simple Q&A; brief chain-of-thought
Low	5×	1.75–3.50	2.15–4.30	Moderate reasoning; typical usage
Medium	10×	3.50–7.00	4.30–8.60	Complex multi-step problems
High	16×	5.60–11.2	6.90–13.8	Extended thinking; o3-equivalent

Table 5: Thinking multipliers (upper bounds). These assume thinking tokens cost the same as visible output tokens; actual energy scaling is likely sub-linear. Ranges within each scenario reflect anchor uncertainty; ranges across scenarios reflect reasoning-effort uncertainty.

Important caveat: These multipliers are upper bounds. Thinking tokens may be processed in optimised batches with lower per-token overhead. OpenAI's statement that GPT-5 uses 50–80% fewer thinking tokens than o3 suggests typical multipliers of 2–5× for most queries. We will refine these when published data on thinking token energy costs becomes available.

7. Why Location-Based Accounting Should Be Preferred

The divergence between market-based and location-based emissions for Microsoft is the most extreme of any major cloud provider. In FY2024, Microsoft reported Scope 2 of 259,090 tCO_2e (market-based) versus 9,955,368 tCO_2e (location-based) — a factor of 38×.[30, 31] A transparency tool reporting only market-based figures would understate ChatGPT's physical carbon footprint by ~97%.

Microsoft achieved 100% annual renewable energy matching in 2025[32] — a significant investment. However, as Microsoft itself has acknowledged, annual matching is an accounting methodology, not a guarantee of carbon-free electricity every hour in every location.[33] Microsoft's 100/100/0 target (100% of electricity, 100% of the time, zero carbon) is a 2030 goal.

Location-based accounting reflects actual physical emissions.[34] IFRS S2 requires location-based Scope 2. The GHG Protocol's Scope 2 revision proposes hourly matching as the standard.[35]

8. Clean Energy Adjustment Factor (CEAF)

The CEAF adjusts the gross location-based estimate downward based on the provider's verified clean energy percentage. It is a supplementary metric, not a substitute for physical emissions.

CEAF-adjusted emissions = Gross × (1 − CEAF)

CEAF eligibility tiers:

Tier 1 — Verified hourly CFE: Full CEAF applied. Google qualifies at 0.66.[36]

Tier 1.5 — Verified annual match, no hourly data: 50% of 100% annual match = CEAF 0.50. Microsoft qualifies here.[37]

Tier 2 — Annual RECs (Renewable Energy Certificates), weaker verification: 50% of claimed percentage.

Tier 3 — No disclosure: CEAF = 0%.

Provider	Tier	CEAF	Basis
Google (Gemini)	Tier 1	0.66	Published 66% hourly CFE (2024 auditable data)
Microsoft (ChatGPT)	Tier 1.5	0.50	100% annual match (2025); no published hourly CFE
Amazon (Claude)	Tier 2	0.50*	Annual RECs; partial hourly pilots; to be refined

[30]Microsoft, 2025 Environmental Sustainability Report (FY2024 data), May 2025. Location-based Scope 2: 9,955,368 tCO_2e; market-based: 259,090 tCO_2e. https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report/
[31]Tracenable, Microsoft GHG Emissions Dashboard. Confirms Scope 2 location-based of 9.96 $MtCO_2e$ for 2024. https://tracenable.com/company/microsoft/ghg-emissions
[32]Microsoft, "A milestone achievement in our journey to carbon negative," Official Microsoft Blog, 18 February 2026. Confirms 100% annual renewable energy matching achieved in 2025. https://blogs.microsoft.com/blog/2026/02/18/a-milestone-achievement-in-our-journey-to-carbon-negative/
[33]Data Centre Magazine, "How Microsoft Hit 100% Renewable Electricity Match for 2025," February 2026. Notes annual matching is "an accounting methodology rather than a guarantee of carbon-free electricity every hour." https://datacentremagazine.com/news/microsoft-matches-100-percent-renewable-electricity-2025
[34]GHG Protocol Scope 2 revision (public consultation October 2025 – January 2026) proposes hourly matching. https://ghgprotocol.org/blog/upcoming-scope-2-public-consultation-overview-revisions
[35]A. Ferrara, "Not greenwashing, but still… A closer look at big tech's 2025 sustainability reports," Internet Policy Review, 2025. Microsoft's location-based Scope 2 more than doubled 2020–2024. https://policyreview.info/articles/news/big-techs-2025-sustainability-reports/2027
[36]Google 2025 Environmental Report: hourly CFE rose from 64% to 66% in 2024. https://sustainability.google/google-2025-environmental-report/

Chinese providers	Tier 3	0.00	No meaningful disclosure

*Table 6: CEAF by provider. *Amazon CEAF provisional; to be refined in the Claude reference document. We commit to upgrading Microsoft to Tier 1 when fleet-wide hourly CFE data is published.*

8.1 The multi-cloud CEAF problem

As documented in Section 3.5, OpenAI now operates across multiple cloud providers with different clean energy profiles. The CEAF of 0.50 in this document applies to the Microsoft Azure portion of inference. The other providers would receive different CEAF classifications:

Microsoft Azure: Tier 1.5, CEAF = 0.50 (100% annual match, no hourly data).

Oracle Stargate: Tier 2–3, CEAF = 0.00–0.25 (no published clean energy data for Stargate; "grid-friendly" language only).

Amazon AWS: Tier 2, CEAF ≈ 0.50 (AWS has committed to 100% renewable by 2025; actual hourly delivery unverified).

A hypothetical 60/25/15 Azure/Oracle/AWS traffic split would produce a blended CEAF of approximately 0.37, significantly below our stated 0.50. We do not know the actual split, and it is changing as Oracle and AWS capacity comes online throughout 2026.

At this point in time we retain CEAF = 0.50 as the stated figure, representing the Azure-only baseline from which the anchor was derived. Users should understand that the CEAF-adjusted figures are likely optimistic relative to OpenAI's current multi-cloud reality, and that the true blended CEAF may be materially lower. This will be subject to ongoing review.

9. Cross-Validation

Google/Gemini: Our Gemini 2.0 Flash anchor of 0.30 gCO_2e/1k tokens is ~40% below GPT-4o's 0.42. Consistent with GPT-4o being a larger model (~200B total params[38]), the US grid being marginally dirtier than Google's mix, and H100 GPUs vs Google's custom TPUs.

Mistral LCA: ~1 gCO_2e per page of text.[39] Our GPT-4o figure translates to ~0.32 gCO_2e per page (at ~750 tokens/page) — below Mistral, consistent with more optimised deployment.

Jegham eco-efficiency: GPT-4o scored 0.762, ranking among the more efficient proprietary models.[40]

10. Confidence Assessment

Element	Confidence	Rationale
Altman anchor (0.34 Wh)	Medium	CEO disclosure; unaudited, no methodology; fleet average
GPT-4o as anchor model	Medium-Low	Likely dominant at time of disclosure; not confirmed
Token count (300)	Low	Undisclosed; plausible range 200–500; drives 2.5× variation

[38]Epoch AI previously estimated GPT-4o has ~200B total parameters, likely MoE with ~100B active. See fn 2.

US grid intensity (370 g/kWh)	High	EPA eGRID; well-established
Throughput scaling	Medium-Low	Valid directionally; confounded by hardware, batching
GEC (0.80–0.90)	Low	Directionally correct; no published migration data
GPT-5 base estimate	Low	Inferred; no direct measurement from OpenAI
Thinking multipliers	Low	Upper bounds; varies enormously by query
Microsoft CEAF (0.50)	Medium-Low	Annual match verified; hourly data absent; Azure-only
Blended multi-cloud CEAF	Very Low	Traffic split across Azure/Oracle/AWS unknown; Oracle undisclosed

Table 7: Confidence assessment. Reflects both evidence strength and residual uncertainty.

11. Limitations

1. Unaudited anchor. The entire framework rests on a CEO blog disclosure with no published methodology or independent verification.

2. Token count dominates uncertainty. A 2.5× range from this assumption alone is larger than most model-to-model differences.

3. Fleet average treated as single-model. Altman's figure is a traffic-weighted average including multimodal, tool use, and free-tier fallback.

4. GPT-5 estimates are speculative. No disclosed parameter count, architecture, or energy data. GPT-5's adaptive reasoning introduces a continuous spectrum of energy cost.

5. Hardware heterogeneity. OpenAI's hardware allocation is opaque (A100 for some models, H100/H200 for others).

6. Inference-only scope. Excludes embodied emissions, training, and supply chain. Microsoft's Scope 3 (15.1 $MtCO_2e$, FY2024) dwarfs operational emissions.[41]

7. Pre-GPT-5 anchor. The 0.34 Wh figure predates GPT-5 by two months and does not reflect current ChatGPT traffic.

8. Multi-cloud infrastructure. Since late 2025, OpenAI runs inference across Azure, Oracle Stargate, and AWS in unknown proportions. The CEAF of 0.50 applies only to Azure; Oracle has published no comparable clean energy data. The true blended CEAF is likely lower, meaning CEAF-adjusted figures may understate emissions. The location-based gross figures are more robust to this uncertainty.

12. Recommendation

These estimates represent our best assessment from publicly available data as of March 2026. They should be used as follows:

For user-facing carbon reporting: Present the **CEAF-Adjusted gCO_2e** figure as the primary metric, representing the physical carbon impact of inference appropriately adjusted for the provider's verified Carbon-Free Energy%.

For decision support: Use the CEAF-adjusted figure as a supplementary metric, with explicit disclosure of the adjustment method and its limitations. It should not be used as the sole basis for customer-facing carbon labels or offsetting claims without independent verification.

For provider comparison: The CEAF-Adjusted gCO_2e enables fair comparison across providers; the CEAF rewards genuine investment in clean energy.

We will update these estimates as new data becomes available, and particularly welcome per-model energy disclosures from OpenAI.

13. Response to Adversarial Review

This document has been subjected to adversarial review by ChatGPT 5.4 Thinking from dual perspectives: a PhD technologist specialising in data centre technologies, and a senior sustainability professional. This section summarises key arguments and our responses.

13.1 Technical critiques accepted

False precision: v1 presented point estimates appearing more precise than the evidence warranted. v3 replaces all with ranges.

Anchor fragility: v1 understated uncertainty in deriving a single-model per-token figure from a fleet-average CEO blog disclosure. v3 puts the 2.5× token-count sensitivity front and centre.

GEC under-argued: v3 presents GEC as a range (0.80–0.90) with explicit rationale.

GPT-5 routing heterogeneity: v3 explicitly states GPT-5 is a unified system with continuous reasoning-effort spectrum.

13.2 Sustainability critiques accepted

"Net" terminology: v1 used "net" for CEAF-adjusted values and recommended them for offsetting. v3 renames to "CEAF-Adjusted," introduces three-layer disclosure, and repositions Layer 3 as supplementary.

Lifecycle scope: v3 adds explicit scope notice and strengthens inference-only caveats.

Offsetting recommendation: v3 removes the recommendation to use CEAF-adjusted figures for offsetting, replacing it with a more conservative "decision support" framing.

13.3 Net impact

Central estimates unchanged. Primary impact is structural: uncertainty foregrounded, ranges replace points, layers separated, CEAF-adjusted repositioned as supplementary. The reviewer's summary: "a credible exploratory estimate, not a robust measurement framework." We accept this characterisation.

