REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of Google Gemini Models: Methodology, Estimates, and Justification

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

Date: 13 March 2026 | Version: 1.4 | Classification: For publication — subject to stated caveats

Important note: *There is limited publicly available data on the gCO_2e impact of LLMs. This paper shows the methodology behind our estimates for Google's Gemini model family. We welcome verified real-world data from LLM providers that will help us improve our model. In particular, we invite Google to continue publishing data on AI inference energy and emissions, and to expand this to report on a per-model basis.*

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 Google's Gemini 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 Google's August 2025 technical paper on Gemini inference environmental impact[1] and scaled across models using throughput-based proxies.

The analysis starts with **location-based carbon accounting** (Section 7), with a **Clean Energy Adjustment Factor (CEAF)** to derive supplementary CEAF-adjusted figures for decision support (Section 8). This acknowledges Google's verified clean energy investments without abandoning physical accuracy. This gives a **CEAF-Adjusted gCO_2e** figure that we use as our primary measure.

This document has been subjected to adversarial review (Section 10).

Summary of recommended estimates (gCO_2e per 1,000 tokens):

	2.5 Flash-Lite	3.1 Flash-Lite	3 Flash	3 Pro	3.1 Pro	Confidence
Gross (location, gCO_2e)	0.25	0.25	**0.40**	0.50	0.50	Med to Low
CEAF (Google) %	0.66	0.66	0.66	0.66	0.66	Med to Low
CEAF-Adj (gCO_2e)	**0.09**	**0.09**	**0.14**	**0.17**	**0.17**	Med to Low

Table 1: Summary estimates. Thinking-mode figures excluded; see Section 6. CEAF = Clean Energy Adjustment Factor based on Google's verified 66% hourly carbon-free energy (2024). CEAF-adjusted figures are a supplementary decision-support metric acknowledging Google's verified clean energy investment.

2. What Has Been Published

[1]Elsworth, C. et al., *Measuring the environmental impact of delivering AI at Google Scale*, arXiv:2508.15734, August 2025. Data collected May 2025. Available: https://arxiv.org/abs/2508.15734

In August 2025, Google published the first production-environment measurement of AI inference energy and emissions, covering the full serving stack of its Gemini AI assistant.[2] The paper reports the following for the *median Gemini Apps text prompt*, from data collected in May 2025:

Metric	Comprehensive	Narrow (GPU only)
Energy per median prompt	**0.24 Wh**	0.10 Wh
Carbon (market-based)	0.03 gCO_2e	0.02 gCO_2e
Carbon (location-based)	**≈0.09 gCO_2e (see below)**	Not reported
Water consumption	0.26 mL	0.12 mL

Table 2: Google's published per-prompt metrics. "Comprehensive" includes accelerator power, host CPU/DRAM, idle capacity, and data centre overhead (PUE 1.09). See fn 4 for location-based derivation. See Section 7 for discussion of market-based vs location-based accounting.

The 3× difference between market-based (0.03) and location-based (0.09) reflects Google's Power Purchase Agreements, which reduce its market-based intensity to approximately 125 gCO_2e/kWh versus a location-based fleet average of approximately 375 gCO_2e/kWh.[3] (See Section 7 for a discussion of this point.)

3. Anchor Derivation

3.1 Which model was measured?

The paper does not name the model, referring only to the "median Gemini Apps text prompt." Google released Gemini 2.0 Flash as the Gemini Apps default on 30 January 2025.[4] Gemini 2.5 Flash was announced at Google I/O on 20 May 2025 but did not reach general availability until 17 June 2025. The May 2025 data was therefore most likely collected from **Gemini 2.0 Flash**, with the possibility that some late-May traffic was served by an early preview of 2.5 Flash. For this analysis, we treat the anchor as Gemini 2.0 Flash.

3.2 The critical unknown: median prompt token count

Google does not disclose the token count of the median prompt.[5] We use 300 total tokens (input + output) as our central assumption, based on the following: Gemini Apps is a general-purpose chatbot whose median interaction is a moderate-length question with a paragraph-length response; Google chose the median for robustness to outliers, implying a right-skewed distribution; and the independent Greenly analysis modelled a 400-token response as a comparator, suggesting this is at or above the median for typical chatbot use.

3.3 Per-token derivation

$$\textit{Energy: } 0.24\ Wh \div 300 \times 1{,}000 = 0.80\ Wh \textit{ per 1,000 tokens}$$

$$\textit{Location-based carbon: } 0.09\ gCO_2e \div 300 \times 1{,}000 = 0.30\ gCO_2e \textit{ per 1,000 tokens}$$

This figure (0.30 gCO_2e/1k tokens, location-based) is the anchor for all subsequent scaling. We believe that it applies to the **Gemini 2.0 Flash model as of May 2025**, not to any specific later model.

[2]Google Cloud Blog, *Measuring the environmental impact of AI inference*, 21 August 2025. Available: https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference/
[3]Google, *Carbon free energy for Google Cloud regions*, https://cloud.google.com/sustainability/region-carbon. Updated February 2025.
[4]Wikipedia, *Gemini (language model)*, https://en.wikipedia.org/wiki/Gemini_(language_model). Confirms: "On January 30, 2025, Google released Gemini 2.0 Flash as the new default model." Gemini 2.5 Flash reached GA on June 17, 2025.
[5]Google does not disclose the token count of the "median Gemini Apps text prompt." See Greenly, *AI Efficiency: A Gemini Reality Check*, August 2025. Available: https://greenly.earth/en-us/leaf-media/data-stories/ai-efficiency-a-gemini-reality-check-with-greenly

4. Scaling to Other Google Models

4.1 Why pricing is not a reliable energy proxy

API pricing is the most accessible data for comparing models, but it is a poor proxy for energy consumption. Gemini 3.1 Flash-Lite costs 3.75× more than the anchor but runs *faster* (256 vs ~250 tokens/second), demonstrating that pricing includes capability premiums unrelated to energy.[6] Using raw price ratios would yield absurd overestimates (e.g. Gemini 3 Pro at 9.0 gCO_2e/1k tokens).

4.2 Approach: throughput-based scaling with generational efficiency correction

We use output throughput (tokens per second) as the primary proxy, on the basis that slower output at similar power draw implies more compute per token. We then apply a Generational Efficiency Correction (GEC) to account for hardware and architectural improvements.

Crucially, all Gemini models since 1.5 use sparse Mixture-of-Experts (MoE) architecture.[7] The anchor (2.0 Flash) is MoE, and all target models are MoE. The difference between generations is not dense-to-MoE but rather improvements in sparsity (newer models activate proportionally fewer parameters per token[8]), hardware efficiency (newer TPU generations), and serving optimisations (speculative decoding, distillation).[9]

The formula is:

$$\textit{Model estimate} = \textit{Anchor} \times (\textit{Anchor throughput} \div \textit{Model throughput}) \times \textit{GEC}$$

The GEC is set at **0.8** for Gemini 3.x models (i.e., newer models use ~80% of the energy throughput alone suggests) and **0.9** for 2.5 models (smaller generational gap). This is conservative: actual efficiency gains may be larger, but we prefer to overstate rather than understate emissions.

Caveat: *Architectural improvements embodied within Google's latest Ironwood TPUs and Mixture-of-Experts architecture may justify a significantly lower GEC. We invite Google to publish appropriate data: we will be pleased to adjust our estimates accordingly.*

4.3 Scaling data and results

Model	Output $/M	Throughput (t/s)	Inv. throughput ratio	GEC	Scaling factor	Gross gCO_2e/1k

[6]Pricing: Google AI Developer documentation, https://ai.google.dev/gemini-api/docs/pricing (updated March 2026). Throughput: Artificial Analysis, https://artificialanalysis.ai/models/

[7]Gemini 2.0 Flash Model Card (April 2025): "The Gemini 2.0 series builds upon the sparse Mixture-of-Experts (MoE) Transformer architecture used in Gemini 1.5." Available: https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf. Gemini 3 Pro Model Card confirms same architecture: https://huggingface.co/datasets/multimodalart/google-gemini-3-pro-pre-release-model-card

[8]B. Dickson, *What (I think) makes Gemini 3 Flash so good and fast*, TechTalks, 22 December 2025. Available: https://bdtechtalks.substack.com/p/what-i-think-makes-gemini-3-flash. Suggests >1T total parameters but only 5–30B active per inference.

[9]The 33× energy and 44× carbon reductions (May 2024–May 2025) are reported in both the Google Cloud Blog (fn 2) and the technical paper (fn 1).

2.0 Flash (anchor)	$0.40	~250 (est.)	1.00	—	1.00	**0.30**
2.5 Flash-Lite	$0.40	256	0.98	0.9	0.88	**0.25**
3.1 Flash-Lite	$1.50	256	0.98	0.8	0.78	**0.25**
3 Flash	$3.00	159	1.15–1.57	0.8	0.92–1.26	**0.40**
3 Pro	$12.00	122	2.05	0.8	1.64	**0.50**
3.1 Pro	$12.00	122	2.05	0.8	1.64	**0.50**

Table 3: Scaling data and results. Throughput from Artificial Analysis (fn 8). GEC = Generational Efficiency Correction. Gemini 3 Flash uses reasoning-mode throughput (159 t/s) for central estimate. All results rounded to nearest 0.05.

4.4 Sensitivity to the Generational Efficiency Correction

The GEC is the area of greatest legitimate technical uncertainty in this analysis. Our recommended GEC of 0.8 is deliberately conservative. See the adversarial review discussion in Section 10 for arguments that the true figure may be 0.5–0.6, hypothetically arguing architectural sparsity improvements. We acknowledge that there is possible variability here, and the table below shows how the estimates may change:

Model	GEC 0.8 (recommended)	GEC 0.7	GEC 0.6 (Google range)	Plausible range
3 Flash	**0.40**	0.35	0.30	0.30–0.40
3 Pro / 3.1 Pro	**0.50**	0.45	0.35	0.35–0.50

Table 3a: Sensitivity of Gemini 3.x estimates to the GEC. Our recommended GEC (0.8) produces the figures used throughout this document. If Google publishes verifiable per-model energy data supporting a lower GEC, we will update accordingly.

5. Per-Model Estimates

Combining the anchor derivation (Section 3), scaling methodology (Section 4), and CEAF (Section 8):

Model	Gross (location)	CEAF	CEAF-Adj $gCO_2e/1k$	Conf.	Notes
2.5 Flash-Lite	0.25	0.66	**0.09**	Med	Similar tier to anchor
3.1 Flash-Lite	0.25	0.66	**0.09**	Med	Faster than anchor; capability premium in price
3 Flash	0.40	0.66	**0.14**	Med-Low	Current Gemini Apps default (Mar 2026)
3 Flash (thinking)	0.60–4.40	0.66	0.20–1.50	Low	Varies by thinking level; see Section 6
3 Pro	0.50	0.66	**0.17**	Low	Provisional. ~1.7× the anchor via throughput scaling.
3.1 Pro	0.50	0.66	**0.17**	Low	Provisional. Same throughput as 3 Pro.

Table 4: Recommended per-model estimates. CEAF-adjusted figures are supplementary decision-support metrics acknowledging Google's clean energy investment; the gross location-based figure is the primary physical metric. Confidence levels discussed in Section 11.

6. Special Case: Thinking Models

Gemini 3 Flash with thinking enabled generates hidden "thinking tokens" before producing visible output.[10] The energy cost is real but invisible. Google's thinking_level parameter (minimal, low, medium, high) controls this:

Thinking level	Est. thinking:visible ratio	Multiplier	Gross gCO_2e/1k visible tokens
Minimal	~0.5:1	1.5×	0.60
Low	~2:1	3×	1.20
Medium	~5:1	6×	2.40
High	~10:1	11×	4.40

Table 5: Estimated thinking multipliers. These are estimates requiring validation. Where the API exposes thinking token counts, InferenceCarbon should use actual counts rather than these estimates.

Important caveat: *These multipliers assume thinking tokens are as computationally expensive as visible output tokens. In practice, thinking tokens may be processed in optimised batches with lower per-token overhead, meaning the true energy multiplier could be sub-linear relative to the token ratio. For example, a 10:1 thinking-to-visible ratio might produce an energy multiplier of 6–8× rather than 11×. Additionally, Google has argued that thinking-phase processing could be spatially shifted to regions with cleaner energy, further reducing its carbon impact. These figures should therefore be treated as upper-bound estimates. We will refine them when published data on thinking token energy costs becomes available.*

7. Why Location-Based Accounting Should Be Preferred

A critical decision in this analysis is whether to use market-based or location-based carbon accounting. Google's headline figure of 0.03 gCO_2e per prompt uses market-based accounting; the location-based equivalent is approximately 0.09 — a 3× difference.[11] This section sets out the case for why location-based figures should be adopted as a default, while reporting market-based figures as a secondary metric.

Location-based accounting reflects actual physical emissions. The location-based method calculates emissions using the average carbon intensity of the power grid where electricity is actually consumed.[12] It answers: "What emissions did this electricity consumption physically cause?" Market-based accounting allows companies to reduce reported emissions through contractual instruments — Renewable Energy Certificates, Power Purchase Agreements, and Guarantees of Origin — regardless of whether those instruments change the physical electricity supply at the point of consumption.

The divergence is large and growing in the AI sector. An analysis of five major tech companies found location-based emissions from data centres approximately 662% higher than market-based figures.[13] Microsoft's location-based Scope 2 more than doubled from 2020 to 2024 while its market-based Scope 2 actually fell.[14]

[10]Google Blog, *Introducing Gemini 3 Flash*, 7 January 2026. Available: https://blog.google/products/gemini/gemini-3-flash/
[11]The location-based figure of ≈0.09 gCO_2e is derived by applying location-based grid intensity to the 0.24 Wh energy figure. See K. Kloster, *Is Google's Reveal of Gemini's Impact Progress or Greenwashing?*, Towards Data Science, 22 August 2025. Available: https://towardsdatascience.com/is-googles-reveal-of-geminis-impact-progress-or-greenwashing/
[12]GHG Protocol Scope 2 revision (public consultation October 2025 – January 2026) proposes hourly matching. See: https://ghgprotocol.org/scope-2-guidance. IFRS S2 requires only location-based Scope 2. See: https://ghgprotocol.org/blog/upcoming-scope-2-public-consultation-overview-revisions
[13]The Guardian analysis (September 2024) found location-based emissions ~662% higher than market-based. See also FAS, *Measuring AI's Energy/Environmental Footprint*, June 2025. Available: https://fas.org/publication/measuring-and-standardizing-ais-energy-footprint/
[14]Microsoft location-based Scope 2 doubled (4.3M to ~10M tonnes, 2020–2024) while market-based fell. See A. Ferrara, *Big Tech's 2025 Sustainability Reports*, Internet Policy Review, 2025. Available: https://policyreview.info/articles/news/big-techs-2025-sustainability-reports/2027

Regulatory momentum is toward location-based primacy. IFRS S2 requires only location-based Scope 2. The GHG Protocol's Scope 2 revision is tightening market-based criteria with mandatory hourly matching and deliverability requirements.[15]

Investors increasingly question market-based claims. In the 2025 proxy season, shareholder proposals at Amazon, Meta, and Alphabet explicitly questioned whether renewable energy procurement strategies remain credible as AI energy demands accelerate.[16] Even Google's own published research acknowledges that current market-based practices "have come under increasing criticism."[17]

Location-based is the appropriate default for a carbon transparency tool. InferenceCarbon exists to help users understand the actual environmental impact of their AI usage. A user in a coal-heavy grid region should not be told their query produced 0.14 gCO_2e/1k tokens when the physical reality is closer to 0.25–0.50. Location-based figures provide the honest, conservative default. The CEAF framework (Section 8) provides the principled mechanism for crediting genuine provider investment without abandoning physical accuracy.

8. Clean Energy Adjustment Factor (CEAF)

While location-based accounting should be the default for reporting gross carbon intensity, it is important to recognise genuine, verifiable investments by LLM providers in clean energy infrastructure. Without such recognition, a provider that has invested billions in carbon-free energy would appear identical to one that has made no investment — and users seeking to offset their AI emissions could double-count emissions already addressed by the provider.

The **Clean Energy Adjustment Factor (CEAF)** adjusts the gross location-based estimate downward based on the provider's verified hourly carbon-free energy (CFE) percentage. This is distinct from simply adopting the provider's market-based figure, because: (1) it starts from physical reality; (2) it credits only *hourly* CFE matching, not annual unbundled RECs (Renewable Energy Certificates)[18]; and (3) it aligns with the GHG Protocol's proposed move toward hourly matching.[19]

$$\textit{CEAF-adjusted emissions} = \textit{Gross} \times (1 - \textit{CEAF})$$

$$\textit{where CEAF} = \textit{provider's verified fleet-wide hourly CFE\%}$$

For Google: Fleet-wide hourly CFE = 66% (2024).[20] Therefore: CEAF-adjusted = Gross × 0.34. The CEAF-adjusted figure is a supplementary decision-support metric that acknowledges Google's clean energy investment while preserving the location-based gross figure as the primary physical metric.

[15]GHG Protocol, *Scope 2 Standard Advances*, July 2025. Available: https://ghgprotocol.org/blog/scope-2-standard-advances-isb-approves-consultation-market-and-location-based-revisions. Google's LCA paper (https://arxiv.org/abs/2502.01671) recognises market-based practices "have come under increasing criticism."
[16]Sustainalytics, *Can Big Tech Keep Its Climate Commitments as Data Centers Scale?*, 2025. Available: https://www.sustainalytics.com/esg-research/resource/investors-esg-blog/can-big-tech-keep-its-climate-commitments-as-data-centers-scale
[17]Schneider, I. et al., *Life-cycle emissions of AI hardware: A cradle-to-grave approach and generational trends*, arXiv:2502.01671, 2025. Acknowledges that market-based practices "have come under increasing criticism." Available: https://arxiv.org/abs/2502.01671
[18]The GHG Protocol's proposed Scope 2 revision proposes hourly matching as the standard for market-based instruments. The CEAF framework anticipates this by using hourly CFE% as the basis for provider credit. See fn 11 and fn 20.
[19]GHG Protocol, Scope 2 Guidance (public consultation October 2025 – January 2026) proposes hourly matching as the standard for market-based instruments. Available: https://ghgprotocol.org/scope-2-guidance
[20]Google 2025 Environmental Report: hourly CFE rose from 64% to 66% in 2024. Available: https://sustainability.google/google-2025-environmental-report/

CEAF eligibility tiers:

Tier 1 — Verified hourly CFE: Provider publishes auditable hourly CFE% data (e.g. Google). Full CEAF applied.

Tier 2 — Annual matching only: Provider claims 100% renewable via annual RECs/PPAs but no hourly data (e.g. Amazon, Meta). Reduced CEAF at 50% of claimed percentage.

Tier 3 — No disclosure: No clean energy data published. CEAF = 0%. Full gross figure applies.

Note on data currency: The CEAF for Google (66%) is based on the most recent published fleet-wide hourly CFE figure, from the 2025 Environmental Report covering calendar year 2024.[21] Google's year-on-year improvement has been modest (64% to 66%, a 2 percentage point gain). It is possible that this figure may now be significantly higher due to new CFE projects activated in 2025–2026, however we have seen no published, auditable data confirming this. **We commit to updating the CEAF for each provider when new auditable data is published.** If Google's 2026 Environmental Report shows higher CFEs, the CEAF-adjusted figures will decrease accordingly.

9. Cross-Validation

OpenAI/ChatGPT: Sam Altman stated the average ChatGPT query uses 0.34 Wh[22] versus Google's 0.24 Wh. Gemini's 30% lower figure is plausible given custom TPU hardware and MoE architectures, but these are not directly comparable ("average" vs "median," different models, unknown token counts).

Mistral AI: Mistral's externally-audited LCA reports 1.14 gCO_2e for a 400-token response (≈2.85 gCO_2e/1k tokens) from Mistral Large 2, including training amortisation.[23] Higher than our estimates, consistent with including training and a denser architecture.

Jegham et al.: The benchmarking study of 30 LLMs found a median of 0.42 Wh for a short GPT-4o query[24] and consumption spanning 65× between the most and least efficient models. Our estimates fall at the efficient end, consistent with Google's custom infrastructure.

BLOOM (Luccioni et al., 2023): Demonstrated that embodied and idle emissions roughly doubled the dynamic GPU figure.[25] Google's comprehensive methodology similarly shows the full-stack figure (0.24 Wh) is 2.4× the GPU-only figure (0.10 Wh).

10. Response to Adversarial Review

This analysis has been subjected to two rounds of adversarial review by Google's Gemini model, adopting the personas of a PhD computer scientist specialising in data centre energy and a senior sustainability analyst. This section evaluates each argument on its merits.

10.1 Technical Arguments

Claim: The GEC of 0.8 is too conservative; internal data suggests 35–40% energy reduction from sparsity alone (implying GEC of 0.5–0.6).

[21]Google, 2025 Environmental Report, covering calendar year 2024. Available: https://sustainability.google/reports/google-2025-environmental-report/
[22]Sam Altman, "The Gentle Singularity," 10 June 2025. Available: https://blog.samaltman.com/the-gentle-singularity
[23]Mistral AI LCA, with Carbone 4 and ADEME, July 2025. 1.14 gCO_2e for 400-token response. Available: https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai
[24]Jegham, N. et al., *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference*, arXiv:2505.09598, May 2025. Available: https://arxiv.org/abs/2505.09598
[25]Luccioni, A.S. et al. (2023), *Estimating the Carbon Footprint of BLOOM*, JMLR, 24(253), pp.1–15. Available: https://jmlr.org/papers/v24/23-0069.html

Verdict: The strongest technical argument, partially accepted. We acknowledge this is the area of greatest legitimate disagreement. The claim is directionally plausible: Gemini 3 models are "ultra-sparse," potentially activating only 5–30 billion of over 1 trillion total parameters per inference.[26] However, the "internal data" cited has not been published and is unverifiable. If Google publishes per-model energy data supporting a lower GEC, we will update immediately. Meanwhile, Section 4.4 now provides a sensitivity analysis showing the impact of GEC values from 0.6 to 0.8. Our recommended GEC of 0.8 is deliberately conservative for a transparency tool.

Claim: Ironwood (TPU v7) has a fundamentally different energy profile; throughput-based scaling assumes a linear power-throughput relationship that doesn't hold.

Verdict: Directionally valid but unverifiable. No published Ironwood energy-per-token benchmarks exist. Our throughput-based approach is the best available public proxy that we have been able to identify. The GEC partially captures hardware improvements.

Claim: Thinking tokens are processed in optimised batches with lower per-token overhead; the multipliers are too high.

Verdict: Plausible, accepted in principle. We have updated Section 6 to acknowledge that thinking multipliers should be treated as upper-bound estimates, with the true energy scaling likely sub-linear relative to the token ratio.

Claim: Fixed overheads amortise differently at higher throughput.

Verdict: Weak. Google's own paper measures energy per prompt in production, with all overheads already amortised across actual production workloads.

10.2 Carbon Accounting Arguments

Claim: The CEAF uses 2024 data (66%) but estimates apply to 2026 models; actual CFE may be >75%.

Verdict: Fair methodological point, addressed. We now commit to updating the CEAF when new auditable data is published (Section 8). However, Google's year-on-year improvement was modest (64% → 66%). Projecting >75% for 2026 is aspirational, not evidenced. Transparency tools must use verified data, not projections.

Claim: Thinking-phase processing can be spatially shifted to cleaner regions since it happens before output begins.

Verdict: Architecturally plausible but unverified. This is a novel argument. The thinking phase could in principle be routed to clean-energy regions without affecting user-facing latency. However, no evidence has been published that Google does this. The fleet-wide CEAF already captures aggregate CFE benefits.

Claim: "Gross location-based" emissions don't exist in reality for Google users; the net figure is the most accurate atmospheric-level impact.

Verdict: A philosophical position, not an analytical error. The atmosphere responds to physical emissions from the grid, not to contractual arrangements. If Google's data centre draws 100 kWh from a grid that is 34% fossil-fuelled, 34 kWh of fossil electricity was consumed regardless of PPA contracts. Our dual-reporting approach (gross for physical reality, CEAF-adjusted for decision support) captures both dimensions. The

[26] See fn 8. B. Dickson, *What (I think) makes Gemini 3 Flash so good and fast*, TechTalks, 22 December 2025: "ultra-sparse" architecture with >1T total parameters but only 5–30B active per inference. Available: https://bdtechtalks.substack.com/p/what-i-think-makes-gemini-3-flash

gross figure enables fair cross-provider comparison; the CEAF-adjusted figure rewards genuine investment.

10.3 Net Impact on Estimates

The adversarial review has led to three refinements: (1) a GEC sensitivity analysis in Section 4.4 showing the impact of more aggressive efficiency assumptions; (2) softened language around thinking token multipliers as upper bounds in Section 6; and (3) an explicit commitment to update the CEAF when new data is published in Section 8. The central estimates and overall methodology are unchanged. The strongest technical argument (GEC should be lower) is acknowledged as the area of greatest uncertainty, with a clear pathway to resolution: published data.

11. Confidence Assessment

Element	Confidence	Basis	Key uncertainty
Anchor energy (0.24 Wh)	**High**	Measured in production by Google	Point-in-time snapshot (May 2025)
Anchor carbon (0.03/0.09)	**High**	Energy × published grid intensities	Market vs location: 3× difference
Token count (300)	**Medium**	Reasoned estimate; Greenly used 400	Undisclosed. Range 200–500.
Anchor model (2.0 Flash)	**Medium-High**	Default Jan–May 2025 per release notes	Possible late-May 2.5 Flash traffic
Flash-Lite scaling	**Medium**	Similar throughput to anchor	Anchor throughput estimated
Gemini 3 Flash	**Medium-Low**	Throughput + GEC, consistent with anchor	GEC (0.8) estimated. Throughput varies.
Pro models	**Low**	~1.7× extrapolation from anchor	No measured Pro data
Thinking estimates	**Low**	Observed token ratios	No published data
CEAF (Google 66%)	**Medium-High**	Published by Google, Electricity Maps data	Fleet avg may differ from AI workloads
Location-based preference	**High**	IFRS S2, GHG Protocol direction	GHG revision not final until 2027

Table 6: Confidence assessment. Reflects both evidence strength and residual uncertainty.

12. Limitations

1. Inference only: All figures exclude training amortisation, which may be significant.

2. Single anchor: All estimates derive from one data point (May 2025). A second anchor would substantially improve confidence.

3. Undisclosed variables: Median token counts, model identity, and per-model energy breakdowns are not published by Google.

4. Generational Efficiency Correction: The GEC (0.8/0.9) is estimated. It may understate efficiency gains (making figures conservative) or overstate them.

5. Scope: Covers Scope 1, 2, and partial Scope 3. Excludes external networking, end-user devices, and broader supply chain.

6. CEAF limitations: Based on fleet-wide hourly CFE, not AI-workload-specific. Actual inference CFE may differ.

7. Thinking models: Token ratio estimates are provisional and require validation from published API data.

13. 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 production data becomes available, and particularly welcome per-model energy disclosures from Google. The methodology established here will be applied separately to Claude (Anthropic), ChatGPT (OpenAI), DeepSeek, Mistral and other providers.

