REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of Anthropic's Claude Models: Methodology, Estimates, and Justification

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

Date: 16 March 2026 | Version: 1.4 | Classification: For publication — subject to stated caveats

Important note: *Anthropic has published no per-query energy data, no sustainability report, and no Scope 1/2/3 emissions disclosure that we have identified as of March 2026.[1] The estimates in this document rely on a single empirical anchor (Jegham et al., 2025)[2] supplemented by throughput-based scaling.[3] Anthropic's inference infrastructure spans at least four providers across three chip architectures, creating a blended infrastructure problem that cannot be resolved without provider disclosure.[4] Confidence is LOW across all estimates. All figures should be read as scenario-based modelled ranges, not measured values.* ***We invite Anthropic to publish data that allows for more confident estimates.***

Scope notice: *This document covers operational inference emissions only. It excludes training, embodied hardware emissions, water consumption, and end-of-life disposal. Estimates represent grams of CO_2 equivalent (gCO_2e) per 1,000 tokens of model output, using location-based Scope 2 accounting as the primary metric.[5] CEAF (Clean-Energy Adjustment Factor)-adjusted figures are provided as a supplementary decision-support metric, not for offsetting calculations.*

1. Executive Summary

Anthropic's Claude model family presents the most challenging case in the InferenceCarbon reference series. Unlike Google (which published measured per-query energy data[6]) and OpenAI (where CEO Sam Altman disclosed an average query figure[7]), Anthropic has made no public energy or carbon disclosure that we have discovered.[8]

Our sole empirical anchor is the Jegham et al. (2025) benchmarking study, which simulated Claude 3.7 Sonnet at 0.950 ± 0.040 Wh per short query (100 input + 300 output tokens).[9] This figure, combined with throughput data from Artificial Analysis,[10] forms the basis for all estimates.

A key complication is that Anthropic operates a multi-cloud infrastructure spanning AWS, Google Cloud, Microsoft Azure, and Fluidstack across three chip architectures (NVIDIA GPUs, Google TPUs, Amazon

[1]Earth911 (March 2026). "Your AI Carbon Footprint: What Every Query Really Costs." https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/

[2]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[3]Artificial Analysis (2026). LLM Leaderboard and model benchmarks. https://artificialanalysis.ai/models/

[4]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[5]GHG Protocol (2025). Scope 2 Guidance — upcoming revisions. https://ghgprotocol.org/blog/upcoming-scope-2-public-consultation-overview-revisions

[6]Google (August 2025). "Measuring the Environmental Impact of AI Inference." arXiv:2508.15734. https://arxiv.org/abs/2508.15734

[7]Altman, S. (June 2025). "The Gentle Singularity." https://blog.samaltman.com/the-gentle-singularity

[8]Earth911 (March 2026). "Your AI Carbon Footprint: What Every Query Really Costs." https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/

[9]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[10]Artificial Analysis (2026). LLM Leaderboard and model benchmarks. https://artificialanalysis.ai/models/

Trainium).[11][12][13] Jegham et al. assumed AWS-only infrastructure. If inference runs partly on Google TPUs, both the energy per query (Wh) and the carbon intensity factor (CIF) could differ. This creates two distinct uncertainty dimensions — hardware-serving-efficiency uncertainty and carbon-factor uncertainty — that compound on top of the anchor's own measurement uncertainty.

External adversarial review characterised this paper as "a stronger estimation memo [than the ChatGPT version], but still not a robust measurement framework" and noted that it "outruns its evidence when it turns a hardware-contaminated estimate into neat per-model defaults." We accept this characterisation. All estimates in this paper should be read as modelled scenario ranges under stated assumptions, not as measured carbon intensities.

Summary of Estimates (AWS infrastructure assumption)

Model	Location-Based Range (gCO_2e/1k tokens, Estimated)	Location-Based Central Scenario (gCO_2e/1k tokens, Estimated)	CEAF%	Adjusted Central Scenario (gCO_2e/1k tokens, Estimated)	Confidence
Claude Haiku 4.5	0.20 – 0.50	0.31	50%	0.16	Low
Claude Sonnet 4.6	0.35 – 0.80	0.52	50%	0.26	Low
Claude Opus 4.6	0.50 – 1.00	0.72	50%	0.36	Low
Claude Sonnet 4.6 (thinking, max)	0.75 – 4.50	1.55	50%	0.78	Very Low
Claude Opus 4.6 (thinking, max)	1.00 – 6.00	2.00	50%	1.00	Very Low

Note: Ranges incorporate anchor uncertainty, GEC range, and throughput-proxy error. They do not incorporate infrastructure-pathway uncertainty (the possibility that TPU or Trainium inference has a fundamentally different Wh profile). Central scenarios assume AWS NVIDIA H200/H100 hardware per Jegham et al. See Section 5 for infrastructure-adjusted scenarios and Section 10 for the split uncertainty framework.

2. What Has Been Published

Anthropic's disclosure position is the least well-developed of any major AI provider (this is perhaps understandable for a company only founded in 2021, however it needs to be noted):

No energy data: Anthropic has published no per-query energy consumption figure for any Claude model. There is no equivalent to Altman's 0.34 Wh disclosure[14] or Google's measured 0.24 Wh median prompt figure.[15]

No sustainability report: Earth911 confirmed in March 2026 that Anthropic has not reported Scope 1, 2, or 3 emissions in any public filing.[16]

[11]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[12]CNBC (October 2025). "Google and Anthropic announce cloud deal worth tens of billions of dollars." https://www.cnbc.com/2025/10/23/anthropic-google-cloud-deal-tpu.html

[13]Introl (December 2025). "Anthropic's $50 Billion Data Center Plan." https://introl.com/blog/anthropic-50-billion-data-center-plan-december-2025

[14]Altman, S. (June 2025). "The Gentle Singularity." https://blog.samaltman.com/the-gentle-singularity

[15]Google (August 2025). "Measuring the Environmental Impact of AI Inference." arXiv:2508.15734. https://arxiv.org/abs/2508.15734

[16]Earth911 (March 2026). "Your AI Carbon Footprint: What Every Query Really Costs." https://earth911.com/business-policy/your-ai-carbon-footprint-what-every-query-really-costs/

No infrastructure transparency: While Anthropic has announced partnerships with AWS, Google Cloud, Azure, and Fluidstack,[17][18][19] it has not disclosed which providers handle inference versus training, the traffic split across providers, hardware configurations per model tier, or clean energy arrangements beyond each cloud provider's general commitments.

Partial infrastructure signals: Anthropic has confirmed that it uses "three chip platforms — Google's TPUs, Amazon's Trainium, and NVIDIA's GPUs"[20] and that AWS remains its "primary training partner and cloud provider."[21] Claude models are available for inference via Amazon Bedrock and Google Cloud Vertex AI. The routing of direct API and consumer traffic is unknown.

3. Anchor Derivation

3.1 Source: Jegham et al. (2025)

The anchor is drawn from Jegham et al. (arXiv:2505.09598v6),[22] an infrastructure-aware benchmarking framework that simulated the environmental footprint of LLM inference across 30 models in commercial data centres. The study uses a Monte Carlo simulation (10,000 correlated samples per model) combining API latency and throughput data from Artificial Analysis[23] with published GPU specifications and region-specific environmental multipliers.

Infrastructure assumption caveat: Jegham et al. assumed all Claude models run on AWS with DGX H200/H100 hardware (PUE (Power Usage Effectiveness) 1.14, on-site WUE (Water Usage Effectiveness) 0.18 L/kWh, CIF (Carbon Intensivity Factor) 0.287 $kgCO_2e/kWh$).[24] This was a reasonable assumption at the time (study published May 2025, last revised November 2025), but may not reflect the current infrastructure mix following the October 2025 Google Cloud TPU expansion[25][26] and subsequent Azure and Fluidstack deals.[27] The energy-per-query (Wh) figures are calibrated to NVIDIA GPU power draw and would differ on TPUs or Trainium. We retain Jegham as anchor but flag this as a significant source of unquantified uncertainty.

---

[17]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[18]CNBC (October 2025). "Google and Anthropic announce cloud deal worth tens of billions of dollars." https://www.cnbc.com/2025/10/23/anthropic-google-cloud-deal-tpu.html

[19]Introl (December 2025). "Anthropic's $50 Billion Data Center Plan." https://introl.com/blog/anthropic-50-billion-data-center-plan-december-2025

[20]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[21]CNBC (October 2025). "Google and Anthropic announce cloud deal worth tens of billions of dollars." https://www.cnbc.com/2025/10/23/anthropic-google-cloud-deal-tpu.html

[22]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[23]Artificial Analysis (2026). LLM Leaderboard and model benchmarks. https://artificialanalysis.ai/models/

[24]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[25]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[26]Google Cloud Press (October 2025). "Anthropic to Expand Use of Google Cloud TPUs and Services." https://www.googlecloudpresscorner.com/2025-10-23-Anthropic-to-Expand-Use-of-Google-Cloud-TPUs-and-Services

[27]Introl (December 2025). "Anthropic's $50 Billion Data Center Plan." https://introl.com/blog/anthropic-50-billion-data-center-plan-december-2025

3.2 Simulated Energy Data

Jegham et al. simulated the energy footprint of three Claude models across three prompt configurations using Monte Carlo methods, combining API throughput data with published GPU specifications:[28]

Model	Short (Wh)	Medium (Wh)	Long (Wh)
Claude 3.7 Sonnet	0.950 ± 0.040	2.989 ± 0.201	5.671 ± 0.302
Claude 3.5 Sonnet	0.973 ± 0.066	3.638 ± 0.256	7.772 ± 0.345
Claude 3.5 Haiku	0.975 ± 0.063	4.464 ± 0.283	8.010 ± 0.338

Short = 100 input + 300 output tokens; Medium = 1,000 + 1,000; Long = 10,000 + 1,500. All in watt-hours with standard deviation.

Claude 3.7 Sonnet is consistently the most energy-efficient of the three models. Claude 3.5 Haiku consumed slightly more energy than the larger 3.5 Sonnet for short prompts, likely reflecting slower API throughput at the time of measurement rather than inherent model inefficiency.[29]

3.3 Deriving gCO_2e per 1,000 Tokens

We select Claude 3.7 Sonnet as the primary anchor: the most recent model and the baseline from which current models descend.

Anchor query: 0.950 Wh for 400 tokens (100 input + 300 output)

Energy per 1,000 tokens: 0.950 ÷ 400 × 1,000 = 2.375 Wh/1k tokens

Location-based carbon (AWS CIF 0.287): 2.375 × 0.287 = 0.682 gCO_2e/1k tokens

Unlike the ChatGPT paper (where the Altman figure's token count was ambiguous), our anchor has a known token configuration. The primary measurement uncertainty comes from the Jegham methodology itself (±4.2% standard deviation on the short-query figure), not from token-count ambiguity.

Prefill/decode asymmetry caveat: *We acknowledge that different queries with different patterns of input and output tokens and different context windows vary significantly in energy per token, and this is particularly the case with thinking-mode workloads. This is not currently captured in our model.*

4. Scaling to Other Models

4.1 Method: Inverse Throughput Ratio

Consistent with the series, we use the inverse throughput ratio as the primary scaling proxy. Slower throughput implies more compute time per token and therefore more energy per token.

Formula: Model estimate = Anchor energy × (Anchor TPS (Tokens per Second) ÷ Model TPS) × GEC (Generational Efficiency Correction, see below.)

Adversarial review caveat: This formula is workable as a rough heuristic but only if compared workloads are close enough in architecture, batching, and provider path. The anchor throughput comes

[28]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[29]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

from Amazon Bedrock[30] while current model throughput comes from Anthropic's first-party API,[31][32][33] meaning the comparison may embed provider-side serving differences. Throughput is confounded by batch size, speculative decoding, KV-cache implementation, latency targets, chip architecture, and hidden reasoning tokens. Once AWS GPU, Google TPU, and Trainium paths are all in the picture, identical TPS no longer implies comparable joules per token.

4.2 Throughput Data

Model	Output TPS	Source	Mode
Claude 3.7 Sonnet (anchor)	54.5	Artificial Analysis (Amazon Bedrock)	Non-reasoning
Claude Opus 4.6	43.7	Artificial Analysis (Anthropic API)	Non-reasoning
Claude Opus 4.6	55.9	Artificial Analysis (Anthropic API)	Adaptive reasoning, max
Claude Sonnet 4.6	60.2	Artificial Analysis (Anthropic API)	Adaptive reasoning, max
Claude Haiku 4.5	89.3	Artificial Analysis (Anthropic API)	Non-reasoning

Sources: Claude 3.7 Sonnet[34]; Opus 4.6[35]; Sonnet 4.6[36]; Haiku 4.5.[37]

4.3 Generational Efficiency Correction (GEC)

The GEC accounts for hardware and software efficiency improvements between model generations. For Claude 4.x relative to 3.7 Sonnet: GEC range 0.65–0.95, midpoint 0.80 reflecting unknown architecture changes including possible serving optimisations such as KV-cache compression and speculative decoding. For Haiku 4.5 (from the 3.5 Haiku anchor): GEC range 0.60–0.90, midpoint 0.75 (widened from 0.75–0.85).

Adversarial review caveat: These GEC bands are sensible placeholders but lack first-principles justification for their specific values. The chosen numbers may be reasonable, but the paper's conclusions are sensitive to them. The entire estimation chain is: (1) anchor, (2) throughput rescaling, (3) GEC correction. Step 1 is solid for an estimate. Steps 2 and 3 are where most uncertainty lives.

[30] Artificial Analysis: Claude 3.7 Sonnet providers. https://artificialanalysis.ai/models/claude-3-7-sonnet/providers
[31] Artificial Analysis: Claude Opus 4.6. https://artificialanalysis.ai/models/claude-opus-4-6
[32] Artificial Analysis: Claude Sonnet 4.6 (adaptive). https://artificialanalysis.ai/models/claude-sonnet-4-6-adaptive
[33] Artificial Analysis: Claude Haiku 4.5. https://artificialanalysis.ai/models/claude-4-5-haiku
[34] Artificial Analysis: Claude 3.7 Sonnet providers. https://artificialanalysis.ai/models/claude-3-7-sonnet/providers
[35] Artificial Analysis: Claude Opus 4.6. https://artificialanalysis.ai/models/claude-opus-4-6
[36] Artificial Analysis: Claude Sonnet 4.6 (adaptive). https://artificialanalysis.ai/models/claude-sonnet-4-6-adaptive
[37] Artificial Analysis: Claude Haiku 4.5. https://artificialanalysis.ai/models/claude-4-5-haiku

4.4 GEC Sensitivity

Model	GEC 0.70	GEC 0.80	GEC 0.85 (central)	GEC 0.90
Haiku 4.5	0.27	0.31	0.33	0.35
Sonnet 4.6	0.43	0.49	0.52	0.56
Opus 4.6	0.60	0.68	0.72	0.77

Values are gCO_2e/1k tokens (location-based gross, AWS CIF 0.287). Under a Google Cloud CIF (~0.25), values would be approximately 13% lower.

5. Infrastructure and Carbon Intensity

5.1 Anthropic's Multi-Cloud Architecture

As of early 2026, Anthropic operates across at least four infrastructure providers:[38, 39, 40, 41]

Provider	Relationship	Hardware	Clean Energy
Amazon AWS	$8B+ investment; Project Rainier; primary training partner	Trainium 2, NVIDIA GPUs	100% annual RE match (2023); no hourly CFE
Google Cloud	$3B equity; up to 1M TPUs; >1 GW in 2026	Google TPUs (up to v7)	66% hourly CFE (2024); best transparency
Microsoft Azure	$30B commitment w/ NVIDIA + Microsoft	NVIDIA GPUs	100% annual RE; no hourly CFE
Fluidstack	$50B partnership; Texas and New York	NVIDIA GPUs (presumed)	No disclosure

Sources: Anthropic blog[42]; CNBC[43]; Introl[44]; Data Center Frontier[45]; Google Cloud Press.[46]

[38]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[39]CNBC (October 2025). "Google and Anthropic announce cloud deal worth tens of billions of dollars." https://www.cnbc.com/2025/10/23/anthropic-google-cloud-deal-tpu.html

[40]Introl (December 2025). "Anthropic's $50 Billion Data Center Plan." https://introl.com/blog/anthropic-50-billion-data-center-plan-december-2025

[41]Data Center Frontier (December 2025). "Inside Anthropic's Multi-Cloud AI Factory." https://www.datacenterfrontier.com/machine-learning/article/55335703/inside-anthropics-multi-cloud-ai-factory-how-aws-trainium-and-google-tpus-shape-its-next-phase

[42]Anthropic (October 2025). "Expanding our use of Google Cloud TPUs and Services." https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

[43]CNBC (October 2025). "Google and Anthropic announce cloud deal worth tens of billions of dollars." https://www.cnbc.com/2025/10/23/anthropic-google-cloud-deal-tpu.html

[44]Introl (December 2025). "Anthropic's $50 Billion Data Center Plan." https://introl.com/blog/anthropic-50-billion-data-center-plan-december-2025

[45]Data Center Frontier (December 2025). "Inside Anthropic's Multi-Cloud AI Factory." https://www.datacenterfrontier.com/machine-learning/article/55335703/inside-anthropics-multi-cloud-ai-factory-how-aws-trainium-and-google-tpus-shape-its-next-phase

[46]Google Cloud Press (October 2025). "Anthropic to Expand Use of Google Cloud TPUs and Services." https://www.googlecloudpresscorner.com/2025-10-23-Anthropic-to-Expand-Use-of-Google-Cloud-TPUs-and-Services

5.2 The Blended Infrastructure Problem

This multi-cloud architecture creates two distinct uncertainty dimensions:

(A) Carbon-factor uncertainty: The CIF differs across providers (AWS 0.287, Google ~0.25, Azure 0.35, Fluidstack ~0.35–0.40).[47, 48] Without knowing the traffic split, we cannot calculate a single defensible CIF. This represents approximately ±15% variation from our central estimate.

(B) Hardware-serving-efficiency uncertainty: Our anchor Wh figure is calibrated to NVIDIA H200/H100 GPUs. If Claude inference runs partly on Google TPUs or Amazon Trainium, the energy per query could differ in ways we cannot quantify. Google's own measurements of Gemini inference on TPUs (0.24 Wh per median prompt)[49] suggest TPUs may offer competitive energy efficiency, but TPU energy profiles cannot be directly mapped to GPU-based measurements. This uncertainty is of unknown direction and potentially comparable magnitude to the CIF uncertainty.

Critical: These two uncertainties are independent and compound. The scenario table below varies only the CIF. It does not capture hardware-pathway uncertainty, which could shift the Wh/query in either direction. The total estimation uncertainty is therefore wider than any single table shows.

5.3 Infrastructure-Adjusted Scenarios (CIF only)

The following scenarios illustrate how the Sonnet 4.6 estimate varies under different provider splits. All hold Wh/query constant at 0.73 Wh (the NVIDIA-based derivation); only CIF changes.

Scenario	Split (AWS/GCP/Azure/FS)	Eff. CIF	Sonnet 4.6 gCO_2e/1k
A: AWS-only (Jegham)	100/0/0/0	0.287	0.52
B: Google Cloud-only	0/100/0/0	~0.25	0.46
C: Balanced	40/40/15/5	~0.28	0.51
D: High-carbon	30/10/20/40	~0.33	0.60

These scenarios are illustrative only. The actual traffic split and the hardware-pathway Wh/query impact are both unknown.

6. Per-Model Estimates

All estimates use Jegham AWS infrastructure assumptions (CIF 0.287, PUE 1.14). Ranges incorporate anchor uncertainty, GEC range, throughput-proxy error, and measurement variance. They do not incorporate infrastructure-pathway uncertainty (Section 5.2B).

6.1 Claude Haiku 4.5

Anthropic's fastest model, released October 2025.[50] Near-Sonnet-4 performance at one-third the cost and over twice the speed.[51] Derived from Claude 3.5 Haiku anchor (0.975 Wh, ~50 TPS estimated from Jegham-era Artificial Analysis data — this figure is approximate and not directly sourced) scaled to Haiku 4.5 throughput (89.3 TPS), GEC 0.80.

[47]Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[48]US EPA eGRID (2024). Emissions & Generation Resource Integrated Database. https://www.epa.gov/egrid

[49]Google (August 2025). "Measuring the Environmental Impact of AI Inference." arXiv:2508.15734. https://arxiv.org/abs/2508.15734

[50]Anthropic (October 2025). "Introducing Claude Haiku 4.5." https://www.anthropic.com/news/claude-haiku-4-5

[51]Artificial Analysis: Claude Haiku 4.5. https://artificialanalysis.ai/models/claude-4-5-haiku

Derivation: 0.975 × (50 ÷ 89.3) × 0.80 = 0.437 Wh per short query

Range: 0.20 – 0.50 gCO_2e/1k tokens (central scenario: 0.31)

Confidence: Low — two-step scaling (3.5 Haiku → 4.5 Haiku), no provider data, possible GPU-count mismatch

GPU-count sensitivity caveat: *Our estimate assumes that all models run with 8 GPUs per inference node. Haiku is a smaller model and may run on fewer GPUs in production, meaning that its actual power draw could be significantly lower.*

6.2 Claude Sonnet 4.6

Anthropic's workhorse model, released February 2026.[52][53] Approaches Opus-class performance at Sonnet pricing. Throughput 60.2 TPS (adaptive reasoning, max effort).[54] Scaled from Claude 3.7 Sonnet anchor (0.950 Wh, ~57 TPS), GEC 0.85.

Derivation: 0.950 × (54.5 ÷ 60.2) × 0.85 = 0.731 Wh per short query

Range: 0.35 – 0.80 gCO_2e/1k tokens (central scenario: 0.52)

Confidence: Low — throughput from reasoning mode; non-reasoning mode likely faster (lower estimate)

6.3 Claude Opus 4.6

Anthropic's most capable model, released February 2026.[55] Notably slower (43.7 TPS non-reasoning),[56] reflecting greater computational depth.

Derivation: 0.950 × (54.5 ÷ 43.7) × 0.85 = 1.007 Wh per short query

Range: 0.50 – 1.00 gCO_2e/1k tokens (central scenario: 0.72)

Confidence: Low — direct scaling from a different model generation

6.4 Thinking/Reasoning Variants

Both Opus 4.6 and Sonnet 4.6 support adaptive thinking mode, generating invisible "thinking tokens."[57] Artificial Analysis found Sonnet 4.6 used ~3× more output tokens than Sonnet 4.5 in max-effort mode, and Opus 4.6 used ~2.3× more.[58]

Model	Thinking multiplier	Range (gCO_2e/1k vis. tokens)	Confidence
Sonnet 4.6 (thinking, max)	~3×	0.75 – 4.50 (central: 1.55)	Very Low
Opus 4.6 (thinking, max)	~2.8×	1.00 – 6.00 (central: 2.00)	Very Low

[52]Anthropic (February 2026). "What's new in Claude 4.6." https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-6

[53]Artificial Analysis (February 2026). "Claude Sonnet 4.6: Everything You Need to Know." https://artificialanalysis.ai/articles/sonnet-4-6-everything-you-need-to-know

[54]Artificial Analysis: Claude Sonnet 4.6 (adaptive). https://artificialanalysis.ai/models/claude-sonnet-4-6-adaptive

[55]Anthropic (February 2026). "What's new in Claude 4.6." https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-6

[56]Artificial Analysis: Claude Opus 4.6. https://artificialanalysis.ai/models/claude-opus-4-6

[57]Anthropic (February 2026). "What's new in Claude 4.6." https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-6

[58]Artificial Analysis (February 2026). "Claude Sonnet 4.6: Everything You Need to Know." https://artificialanalysis.ai/articles/sonnet-4-6-everything-you-need-to-know

Critical caveat: The thinking-token ratio varies enormously by task. Simple queries may trigger minimal thinking; complex tasks may generate 10–50× the visible token count. The figures above assume the Artificial Analysis evaluation workload, which is heavily reasoning-weighted. Real-world usage at default effort levels will be much lower. These estimates exist on a continuous spectrum. Following adversarial review, we present them as scenario ranges rather than model-level figures.

7. Three-Layer Disclosure Structure

Layer	Metric	Haiku 4.5	Sonnet 4.6	Opus 4.6
Layer 1: Energy	Wh per short query	0.44	0.73	1.01
Layer 2: Location-based	gCO_2e/1k tokens (range)	0.20–0.50	0.35–0.80	0.50–1.00
Layer 3: CEAF-adjusted	gCO_2e/1k (AWS CEAF 0.50)	0.10–0.25	0.18–0.40	0.25–0.50

Layer 2 (location-based gross range) is the primary metric. Layer 3 is supplementary. Both reflect the AWS infrastructure assumption only.

8. Clean Energy Adjustment Factor (CEAF)

8.1 Per-Provider CEAF Assessment

Provider	Annual RE	Hourly CFE	CEAF Tier	CEAF
AWS	100% (2023)	Not published	Tier 2	0.50
Google Cloud	66% hourly (2024)	Published by region	Tier 1	0.66
Azure	100% (Feb 2026)	Not published	Tier 1.5	0.50
Fluidstack	No disclosure	N/A	Tier 3	0.00

Sources: AWS[59][60][61]; Google Cloud[62][63]; Amazon sustainability report.[64]

Sustainability review caveat: These CEAF values are normative scoring choices built from partial disclosures of very different kinds (Google's hourly CFE, Amazon's annual RECs (Renewable Energy Certificates), Fluidstack's silence). They are not measured residual emissions. The blended CEAF exercise is useful for illustrating governance complexity but should not be treated as a precise quantified adjustment.

[59] Amazon Sustainability (2024). "Sustainability in the Cloud." https://sustainability.aboutamazon.com/environment/the-cloud
[60] DCD (July 2024). "Amazon: All our operations now run on renewable energy." https://www.datacenterdynamics.com/en/news/amazon-all-our-operations-now-run-on-renewable-energy/
[61] The New Stack (March 2025). "Sustainability: How Did Amazon, Azure, Google Perform in 2023?" https://thenewstack.io/sustainability-how-did-amazon-azure-google-perform-in-2023/
[62] Google Cloud (2024). Region Carbon Data. https://cloud.google.com/sustainability/region-carbon
[63] Google (2025). 2025 Environmental Report. https://sustainability.google/google-2025-environmental-report/
[64] Amazon (2024). 2024 Sustainability Report. https://sustainability.aboutamazon.com/2024-report

8.2 Blended CEAF Scenarios

Scenario	Split (AWS/GCP/Az/FS)	Blended CEAF	Sonnet 4.6 CEAF-adj range
AWS-heavy	70/15/10/5	0.46	0.19–0.43
Balanced	40/40/15/5	0.52	0.17–0.38
Google-heavy	25/55/15/5	0.56	0.15–0.35
Fluidstack-heavy	30/20/10/40	0.28	0.25–0.58

For default CEAF-adjusted figures, we use CEAF 0.50 (AWS Tier 2) as a conservative baseline.

9. Why Location-Based Accounting Should Be Preferred

As argued throughout this series, location-based Scope 2 accounting should be the primary metric.[65] Market-based accounting allows companies to claim near-zero emissions via RECs, but this does not reflect the physical carbon intensity of electricity consumed at the point of inference.

This is acutely relevant for Anthropic because: (a) inference runs across multiple providers with different market-vs-location gaps; (b) Amazon's Scope 2 emissions were 2.79 $MmtCO_2e$ in 2023 despite a 100% renewable claim;[66] (c) AWS does not publish hourly CFE data; and (d) Fluidstack has no clean energy disclosure.

10. Uncertainty Framework

Following adversarial review, we present the dominant uncertainty sources in two distinct categories. These uncertainties are independent and compound.

Category	Source	Direction	Approximate magnitude	Resolvable by
A: Anchor measurement	Jegham Monte Carlo methodology	Known (±4.2% SD)	Small	Replication studies
A: GEC	Generational efficiency placeholder	Unknown	Moderate (±15%)	Provider disclosure or new measurements
A: Throughput proxy	TPS as energy proxy; confounded by batching, cache, chip	Unknown	Moderate to large	Direct energy measurement
A: Token composition	Short-query mix may not match production	Unknown	Moderate	Provider usage-mix disclosure
B: Carbon factor (CIF)	Provider/region electricity mix	Known range (±15%)	Moderate	Provider routing disclosure

[65]GHG Protocol (2025). Scope 2 Guidance — upcoming revisions. https://ghgprotocol.org/blog/upcoming-scope-2-public-consultation-overview-revisions

[66]The New Stack (March 2025). "Sustainability: How Did Amazon, Azure, Google Perform in 2023?" https://thenewstack.io/sustainability-how-did-amazon-azure-google-perform-in-2023/

B: Hardware pathway	TPU vs GPU vs Trainium Wh/query	Unknown direction	Potentially large	Provider hardware disclosure

Category A uncertainties affect the energy estimate (Wh per query). Category B uncertainties affect the carbon conversion (gCO_2e per Wh). Both must be resolved for decision-grade carbon intensity figures. Neither can be resolved without provider disclosure from Anthropic.

11. Cross-Validation

Against Jegham/Altman: Jegham's GPT-4o short-query figure (0.423 Wh) aligns within 19% of Altman's 0.34 Wh claim,[67] giving confidence in the framework's calibration.

Against pricing ratios (sanity check only): Haiku ($1/$5 MTok) is 3× cheaper than Sonnet ($3/$15), Sonnet is 40% cheaper than Opus ($5/$25). Our estimates show ~2.5× between Opus and Haiku, directionally consistent. Pricing is a business signal, not physics.

Against eco-efficiency ranking: Jegham's cross-efficiency DEA ranked Claude 3.7 Sonnet at 0.825, third-highest across 30 models (behind o3-mini 0.884, o1-mini 0.836).[68] This confirms strong intelligence-per-watt.

Against Couch's Claude Code analysis: Simon Couch (January 2026) estimated Claude Sonnet energy at ~390 Wh/MTok input and ~1,950 Wh/MTok output via pricing ratios.[69] Our blended figure of ~2,375 Wh/MTok is the same order of magnitude, though imprecise due to different input/output split assumptions.

Against Mistral LCA: Jegham's Mistral Large 2 estimate (~1.09 gCO_2e for 400 tokens) aligned within one standard deviation of Mistral's own published LCA figure (~1.14 gCO_2e for 400 tokens),[70] further validating the framework.

12. Confidence Assessment

Estimate	Confidence	Dominant uncertainty
Anchor (3.7 Sonnet)	Medium	Single source; AWS-only hardware assumption
Haiku 4.5	Low	Two-step scaling; GPU-count uncertainty
Sonnet 4.6	Low	Throughput from reasoning mode; infrastructure unknown
Opus 4.6	Low	Large throughput gap from anchor; infrastructure unknown
Thinking variants	Very Low	Token ratio highly variable by task; scenario-only
CEAF (all)	Low	Multi-cloud split unknown; range 0.28–0.56
Infrastructure pathway	Unquantified	TPU/Trainium Wh/query impact unknown

[67] Altman, S. (June 2025). "The Gentle Singularity." https://blog.samaltman.com/the-gentle-singularity

[68] Jegham, N. et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference." arXiv:2505.09598v6, November 2025. https://arxiv.org/abs/2505.09598

[69] Couch, S.P. (January 2026). "Electricity use of AI coding agents." https://www.simonpcouch.com/blog/2026-01-20-cc-impact/

[70] Mistral AI (2025). "Our Contribution to a Global Environmental Standard for AI." https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai

13. Limitations

1. Single empirical source. All anchor data comes from Jegham et al. (2025). No second independent measurement exists for cross-validation.

2. No provider disclosure. Anthropic has published no energy, emissions, or sustainability data.

3. Multi-cloud infrastructure unknown. Anthropic uses AWS, Google Cloud, Azure, and Fluidstack with three chip architectures. The traffic split is undisclosed.

4. Hardware architecture mismatch. The anchor assumes NVIDIA GPU inference. TPU and Trainium inference may differ in ways we cannot quantify.

5. Throughput as proxy. TPS is confounded by batching, KV-cache, speculative decoding, and infrastructure-level efficiency.

6. Model class assumptions. Jegham classified all Claude models as "Large" (8 GPUs). Haiku may use fewer GPUs in production.

7. Current models not measured. No current-generation Claude model has been directly measured by any study.

8. CEAF approximation. Blended CEAF ranges 0.28–0.56 depending on traffic split (±30%).

9. Operational inference only. Training, embodied hardware, water, and Scope 3 emissions excluded.

10. GEC under-justified. The 0.80–0.90 range is a reasonable placeholder without first-principles derivation.

14. Response to Adversarial Review

An external adversarial review was conducted by GPT-5.4 Thinking (OpenAI) on 13 March 2026, using dual personas: a PhD technologist specialising in data-centre technologies, and a senior sustainability professional. The review was conducted on v1.1 of this paper.

Summary verdict: "A stronger estimation memo than the [ChatGPT] version, but still not a robust measurement framework" that "outruns its evidence when it turns a hardware-contaminated estimate into neat per-model defaults."

We accept this characterisation.

Three changes insisted upon and accepted:

1. Range-led, not point-led. Restructured all summary and per-model tables so ranges are the headline, with central values as scenario references within them. Implemented throughout v1.2.

2. Split uncertainty into two visible categories. Added Section 10 (Uncertainty Framework) explicitly separating Category A (energy/Wh uncertainties: anchor, GEC, throughput proxy, token composition) from Category B (carbon/CIF uncertainties: provider mix, hardware pathway). These are independent and compound.

3. Rewrite recommendation. Section 15 no longer recommends single default display values. It recommends displaying the full range with the central scenario clearly labelled as a modelled estimate, not a measured value.

Additional changes: Thinking-mode estimates reframed as scenario ranges rather than model-level figures. GEC acknowledged as under-justified placeholder. Throughput-proxy limitations elevated from footnote caveats to inline adversarial review caveats.

15. Recommendation

For InferenceCarbon.ai's carbon transparency tool, we recommend displaying the full estimated range for Anthropic Claude models, with the central scenario clearly identified as a modelled estimate under AWS infrastructure assumptions:

Model	Display range (gCO_2e/1k tokens)	Central scenario	Label
Haiku 4.5	0.20 – 0.50	0.31	Modelled estimate — no provider data
Sonnet 4.6	0.35 – 0.80	0.52	Modelled estimate — no provider data
Opus 4.6	0.50 – 1.00	0.72	Modelled estimate — no provider data

The label "Modelled estimate — no provider data" should be visible in any context where these numbers appear. The CEAF-adjusted range should be available as a supplementary layer with methodology disclosure and a note on the blended infrastructure uncertainty.

We commit to updating these estimates when: (a) Anthropic publishes energy or sustainability data; (b) new independent measurements become available; (c) the infrastructure split becomes known; or (d) updated throughput data materially changes scaling ratios.

16. Key Sources

All sources are cited in footnotes throughout. Principal references:

Source	Used for
Jegham et al. (2025), arXiv:2505.09598v6	All anchor data — Claude 3.7/3.5 Sonnet, 3.5 Haiku energy
Artificial Analysis (2026)	Throughput for all Claude models
Earth911 (March 2026)	No Anthropic emissions disclosure confirmed
Anthropic blog (Oct 2025)	Multi-cloud strategy; TPU commitment
CNBC / Google Cloud Press (Oct 2025)	Infrastructure deal details
Introl / Data Center Frontier (Dec 2025)	Fluidstack, Azure, Project Rainier
Amazon Sustainability (2024)	AWS clean energy, WUE, PUE, Scope 2
The New Stack (2025)	Amazon market-vs-location analysis
EPA eGRID (2024)	US grid carbon intensity
GHG Protocol Scope 2 (2025)	Location-based accounting methodology
Google (2025), arXiv:2508.15734	Cross-validation of Jegham framework
Altman, S. (June 2025)	Cross-validation of GPT-4o anchor
Couch, S.P. (Jan 2026)	Independent Claude energy estimation

Mistral AI (2025)	LCA cross-validation

End of document. Version 1.2, 13 March 2026.

Prepared by InferenceCarbon.ai Reference Team