REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of AI Image Generation Models: Methodology, Estimates, and Justification

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

Date: 25 March 2026 | Version: 1.4 | Classification: For publication — subject to stated caveats

Important note: *No AI image generation provider has published per-image inference energy data. The only measured data comes from open-source models (Stable Diffusion family) running on research hardware. Commercial models — DALL-E 3, GPT-4o Image, Gemini Image, and Midjourney — are black boxes excluded from every published energy study. All estimates are presented as ranges to reflect this substantial uncertainty.*

Scope notice: *This document estimates operational inference emissions only (Scope 2, electricity consumption during model inference). It does not constitute a lifecycle assessment and excludes: embodied emissions of hardware manufacturing; model training energy; upstream supply chain emissions (Scope 3); water consumption; and electronic waste. The estimates cover electricity consumed by GPU servers conducting inference, supporting infrastructure (CPUs, memory, networking), and data centre overhead (cooling, power distribution), expressed via Power Usage Effectiveness (PUE).*

1. Executive Summary

This document provides carbon intensity estimates for AI image generation models, expressed in grams of CO_2 equivalent per image (gCO_2e/image) at standard quality (1024×1024 pixels). The primary anchor is the Chung et al. measurement of Stable Diffusion 3 Medium at 2,282 joules (0.634 Wh) total per image on NVIDIA H100 hardware, with models being scaled using image generation time as a proxy for energy consumption and then adjusted by additional factors (see section 5 for methodology.) Estimates are reported as Clean Energy Adjustment Factor (CEAF)-adjusted figures for decision support, with location-based gross figures retained for Scope 2 transparency.

Model	Location-Based Range(gCO_2e/image, Estimate)	Location-Based Central(gCO_2e/image, Estimate)	CEAF %	CEAF-Adjusted Central(gCO_2e/image, Estimate)	Confidence
Gemini Image (Imagen)	0.06–0.14	0.10	66%	0.03	Low
Stable Diffusion 3	0.15–0.45	0.18	50%	0.09	Medium
DALL-E 3 (standard)	0.68–0.99	0.78	50%	0.39	Low
GPT-4o Image Generation	1.13–2.48	1.63	50%	0.82	Low
DALL-E 3 HD	0.77–1.26	0.90	50%	0.45	Low
Midjourney v6	2.71–5.41	4.05	0%	4.05	Very Low

Table 1: Summary estimates. All figures at 1024×1024, standard quality. Location-based uses per-provider grid intensity (see Table 2). CEAF reflects provider clean energy investment: Google 66% hourly CFE[1]; Microsoft Azure 50% (annual match); Midjourney 0% (undisclosed).[2] Sorted by CEAF-adjusted central estimate, greenest first. All commercial estimates derived via generation-time scaling from the Chung/MIT SD3 anchor (Section 5.0).

[1]Google 2025 Environmental Report: hourly Carbon-Free Energy (CFE) rose to 66% in 2024. https://sustainability.google/google-2025-environmental-report/
[2]Microsoft, "A milestone achievement in our journey to carbon negative," Feb 2026. 100% annual renewable match in 2025. https://blogs.microsoft.com/blog/2026/02/18/a-milestone-achievement-in-our-journey-to-carbon-negative/

2. What Has Been Measured

Chung & Chowdhury / ML.Energy / MIT Technology Review (May 2025): The strongest anchor. Measured Stable Diffusion (SD) 3 Medium (2B parameters) on H100 GPUs. At standard quality (1024×1024, 25 denoising steps): 1,141 joules GPU energy. Applying a 2× overhead multiplier (cooling, CPUs, networking) yields 2,282 joules total = 0.634 Wh per image. At 50 steps (high quality): 4,402 joules = 1.22 Wh.[3][4][5]

Luccioni et al. (2024, ACM FAccT): Tested 88 models on A100 GPUs. Image generation median: 1.35 kWh per 1,000 images (1.35 Wh/image); mean 2.907 kWh. Stable Diffusion XL (SDXL) was highly carbon intensive at approximately 1,594gCO_2e/1k. Only open-source models tested.[6]

Bertazzini et al. (2025, University of Florence): 17 diffusion models. Resolution doubling increases energy 1.3–4.7×. U-Net models consume less than Transformer-based. Prompt length has no significant impact.[7]

HuggingFace AI Energy Score: SDXL on H100: 1,640 Wh/1k images vs A100: 11,410 Wh — a 7× hardware generation difference.[8]

What has NOT been measured: No published study has measured DALL-E 3, GPT-4o Image, Gemini Image, or Midjourney. Every figure for these models is an estimate.

3. Provider Infrastructure and Grid Context

Carbon estimates depend critically on where inference runs and the grid intensity at that location.

Provider	Model(s)	Infrastructure	Primary Data Centre Locations	Grid CIF(gCO_2e/kWh)	PUE	CEAF Tier	CEAF %
Google	Gemini/Imagen	Google Cloud TPUs	US (Oregon, Iowa, Virginia), EU, APAC	375	1.10	Tier 1	66%
OpenAI	DALL-E 3, GPT-4o	Microsoft Azure H100/A100	US East (Virginia), US Central, EU	350	1.12	Tier 1.5	50%
Stability AI	SD3 (API)	AWS	US (Oregon, Virginia)	350	1.12	Tier 2	50%
Midjourney	v6, v7	Unknown	Unknown	Unknown	Unknown	Tier 3	0%

Table 2: Provider infrastructure context. Grid CIF = Carbon Intensity Factor, US-weighted average for the provider's known data centre mix. Google's 66% CEAF is derived from verified hourly CFE data. Microsoft's 50% CEAF is derived from 100% annual renewable match (2025), discounted for lack of hourly CFE data. Midjourney has disclosed no infrastructure information.

[3]Chung, J.-W. & Chowdhury, M. (2026) "Where Do the Joules Go? Diagnosing Inference Energy Consumption." University of Michigan / ML.Energy. arXiv:2601.22076. https://arxiv.org/abs/2601.22076
[4]MIT Technology Review, "We did the math on AI's energy footprint," 20 May 2025. SD3 Medium: 1,141 J GPU, 2,282 J total at 1024×1024, 25 steps, H100. https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/
[5]MIT Technology Review, Methodology companion, 20 May 2025. GPU→total 2× multiplier; hardware: NVIDIA H100. https://www.technologyreview.com/2025/05/20/1116331/ai-energy-demand-methodology/
[6]Luccioni, A.S., Jernite, Y. & Strubell, E. (2024) "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" ACM FAccT '24. Median image gen: 1.35 kWh/1k images; SDXL: 11.49 kWh/1k. A100 hardware. https://doi.org/10.1145/3630106.3658542
[7]Bertazzini, G. et al. (2025) "The Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation." Univ. Florence. 17 models, resolution 1.3–4.7× scaling. arXiv:2506.17016. https://arxiv.org/abs/2506.17016
[8]HuggingFace AI Energy Score leaderboard. SDXL on H100: 1,640 Wh/1k images vs A100: 11,410 Wh/1k. https://huggingface.co/spaces/AIEnergyScore/Leaderboard

4. Anchor Derivation

4.1 Primary anchor: SD3 Medium on H100

The Chung/MIT measurement provides the only rigorously measured figure for a current-generation image model:[9]

GPU energy: 1,141 joules (SD3 Medium, 1024×1024, 25 steps, H100)

Total with 2× overhead: 2,282 joules = 0.634 Wh

Location-based carbon: 0.634 Wh × 0.370 gCO_2e/Wh = **0.23 gCO_2e/image**

CEAF-adjusted (Stability AI, CEAF 50%): 0.23 × 0.50 = **0.12 gCO_2e/image**

4.2 Cross-validation

Luccioni et al. (2024) measured a median of 1.35 Wh/image across 88 models on A100 GPUs. The H100 is 2–3× more energy-efficient per FLOP than A100 (HuggingFace AI Energy Score: SDXL 11.41 Wh on A100 vs 1.64 Wh on H100 — a 7× difference). Adjusting conservatively (÷2 to ÷3): 0.45–0.68 Wh. The Chung anchor of 0.634 Wh falls squarely within this range, providing strong cross-validation.

HuggingFace: SDXL on H100 at 1.64 Wh. SD3 Medium at 0.634 Wh being ~2.6× more efficient is plausible given architectural improvements (Rectified Flow vs standard diffusion).

5. Per-Model Estimates

5.0 Scaling Methodology: Generation-Time Proxy

For text LLMs, throughput (tokens/second) serves as the primary energy proxy. For image models, the equivalent is API generation time: longer generation at similar hardware power draw implies more energy per image. GPU power draw during diffusion inference is approximately constant (70–90% utilisation throughout denoising), so energy ≈ power × time.

Step 1 — Back-derive effective server power from the anchor: SD3 Medium on H100 consumes 0.634 Wh total (Chung/MIT, including 2× overhead). Artificial Analysis reports SD3 Medium generation time on Replicate at approximately 5.2 seconds.[10] Effective server power: P = 0.634 Wh ÷ (5.2s ÷ 3600) = 439W. This is consistent with an H100 at ~350W TDP plus data centre overhead (higher than 350W TDP due to inclusion of cooling, CPUs, networking in the 2× overhead). Note: Modern GPUs use Dynamic Voltage and Frequency Scaling (DVFS), so power draw is not strictly constant. However, during active diffusion inference, GPU utilisation is sustained at 70–90%, making ~439W a reasonable average for the active generation period. We do not claim this figure applies during idle or queuing periods.

Step 2 — For each model: Wh = 380W × (generation time in seconds) ÷ 3600 × hardware correction factor.

Step 3 — Hardware correction factors: Google TPUs: ×0.5 (TPUs are documented as 2–3× more energy-efficient per FLOP than H100; Patterson et al. 2022, Google Environmental Report 2025).[11][12] Same GPU class (H100/A100 on Azure/AWS): ×1.0. Unknown hardware (Midjourney): ×1.0 with wide uncertainty band.

Step 4 — Convert to gCO_2e: Wh × provider CIF gCO_2e/Wh (per-provider grid intensity: OpenAI/Azure: 350 gCO_2e/kWh; Google/GCP: 370 gCO_2e/kWh; Amazon (estimate across Oregon and Virginia grids): 287 gCO_2e/kWh; US-weighted default: 370 gCO_2e/kWh.)

Step 5 — Apply CEAF: Location-based × (1 − CEAF%).

[10] Artificial Analysis, Image Generation Model Benchmarks (accessed March 2026). Generation time, quality, and pricing data. https://artificialanalysis.ai/image/models

[12] Patterson, D. et al. (2022) "The Carbon Footprint of ML Training Will Plateau, Then Shrink." IEEE Computer. Google DC fleet PUE: 1.10. https://arxiv.org/abs/2204.05149

Caveat: API generation time includes queue wait, safety filtering, and network latency that do not directly consume GPU energy. These are typically small (<1s) relative to diffusion (5–30s). We do not discount for non-compute time, making estimates slightly conservative.

5.1 Stable Diffusion 3 (Anchor)

Directly measured by Chung/MIT. Stability AI API runs on Amazon Web Services (AWS), primarily US data centres (Oregon, Virginia).

Step 1: Effective server power = 439W (derived in Section 5.0 from the Chung measurement).

Step 2: Energy = 0.634 Wh (directly measured: 1,141 J GPU × 2 overhead = 2,282 J = 0.634 Wh). No scaling needed — this is the anchor.

Step 3: Hardware correction = ×1.0 (H100, same as anchor hardware).

Step 4: Location-based carbon = 0.634 × 0.287 (Amazon) = 0.18 gCO_2e/image.

Step 5: CEAF-adjusted (AWS, Tier 2, CEAF 50%) = 0.18 × (1 − 0.50) = 0.09 gCO_2e/image.

Confidence: Medium.

5.2 Gemini Image Generation (Imagen 4)

Google's image generation runs on custom Tensor Processing Units (TPUs) in data centres with PUE 1.10 and 66% hourly Carbon-Free Energy (CFE). Multiple sources report Imagen 4 Standard generation times of 3–5 seconds (Artificial Analysis[13]; WaveSpeedAI reports "3–5 seconds"; Imagen 4 Fast achieves sub-2-second generation).

Step 1: Effective server power = 439W (from anchor).

Step 2: Energy (central, 4.2s generation time) = 439W × (4.2 ÷ 3600) = 0.51 Wh (before hardware correction).

Step 3: Hardware correction = ×0.5 (central). Google TPUs are documented as 2–3× more energy-efficient per FLOP than H100 GPUs (Patterson et al. 2022, Google Environmental Report 2025). Google's latest v5p TPUs are liquid-cooled and purpose-built for the sparse operations used in modern diffusion, suggesting the true efficiency advantage may be even larger (possibly 3–4×, implying a correction of 0.25–0.33). We use 0.5 as a conservative central estimate; the true figure may be lower, meaning our Gemini estimate likely overstates Google's actual energy consumption. Corrected energy = 0.51 × 0.5 = 0.26 Wh.

Step 4: Location-based carbon = 0.26 × 0.370 (Google) = 0.10 gCO_2e/image.

Step 5: CEAF-adjusted (Google, Tier 1, CEAF 66%) = 0.10 × (1 − 0.66) = 0.03 gCO_2e/image.

Range: Generation time 3–5s, TPU factor 0.25–0.6 → 0.15–0.37 Wh → 0.06–0.14 gCO_2e (location-based).

Confidence: Low — no published energy data; inferred from generation time and Google's documented TPU efficiency.

5.3 DALL-E 3 (Standard)

Pure diffusion model on Microsoft Azure (H100/A100), primarily US East (Virginia) and US Central data centres. DALL-E 3 generation times are widely reported at 15–22 seconds for standard 1024×1024 output (Artificial Analysis[14]; industry sources report 15–25s typical).

Step 1: Effective server power = 439W (from anchor).

Step 2: Energy (central, 18.3s generation time) = 439W × (18.3 ÷ 3600) = 2.23 Wh.

Step 3: Hardware correction = ×1.0 (H100/A100, same GPU class as anchor).

Step 4: Location-based carbon = 2.23 × 0.350= 0.78 gCO_2e/image.

Step 5: CEAF-adjusted (Microsoft Azure, Tier 1.5, CEAF 50%) = 0.78 × (1 − 0.50) = 0.39 gCO_2e/image.

Range: Generation time 15–22s → 1.83–2.68 Wh → 0.68–0.99 gCO_2e (location-based).

Confidence: Low — generation time from API benchmarks, not direct energy measurement.

5.4 GPT-4o Image Generation / GPT Image 1.5

Hybrid architecture: an autoregressive transformer generates ~1,024 visual tokens, which are decoded by a diffusion head into pixels.[15] This dual architecture (LLM forward pass + diffusion decoding) takes longer than pure diffusion. Runs on Microsoft Azure (H100/A100), primarily US data centres. Generation times vary widely: industry sources report 15–60+ seconds depending on prompt complexity and quality setting (CreateVision AI reports 60–180s for GPT-4o at high quality; standard API calls are typically 15–30s for GPT Image 1/1.5). Artificial Analysis reports 38.3s median generation time. We use 38.3s as the central estimate.

Step 1: Effective server power = 439W (from anchor).

Step 2: Energy (central, 38.3s generation time) = 439W × (38.3 ÷ 3600) = 4.67 Wh.

Step 3: Hardware correction = ×1.0 (H100/A100, same GPU class as anchor). Note: the hybrid LLM+diffusion architecture consumes GPU power at a similar rate throughout both stages, so no additional architectural correction is applied; the longer generation time already captures the additional compute.

Step 4: Location-based carbon = 4.67 × 0.350 (OpenAI/Azure) = 1.63 gCO_2e/image.

Step 5: CEAF-adjusted (Microsoft Azure, Tier 1.5, CEAF 50%) = 1.63 × (1 − 0.50) = 0.82 gCO_2e/image.

Range: Generation time 25–55s → 3.05–6.71 Wh → 1.13–2.48 gCO_2e. The wide range reflects genuine uncertainty in how much of the wall-clock time is active GPU compute.

Confidence: Low.

5.5 DALL-E 3 HD

Higher resolution output (1024×1792). Pixel count is 1.75× the 1024×1024 baseline. Bertazzini et al. (2025) found resolution doubling increases energy 1.3–4.7×; sub-linear scaling suggests ~1.45× for 1.75× pixels. API pricing is 2× standard ($0.08 vs $0.04), suggesting roughly proportional compute. Derived from DALL-E 3 standard rather than directly from the anchor.

Step 1: Effective server power = 439W (from anchor).

Step 2: Energy (central, 21.1s generation time from Artificial Analysis) = 439W × (21.1 ÷ 3600) = 2.57 Wh. Note: This direct generation-time approach replaces the v1.1 method of scaling from DALL-E 3 standard × 1.45 resolution factor. The results are closely consistent (2.57 vs 2.61 Wh), validating the resolution-scaling heuristic.

Step 3: Hardware correction = ×1.0 (same Azure infrastructure as DALL-E 3 standard).

Step 4: Location-based carbon = 2.57 × 0.350 (OpenAI/Azure) = 0.90 gCO_2e/image.

Step 5: CEAF-adjusted (Microsoft Azure, Tier 1.5, CEAF 50%) = 0.90 × (1 − 0.50) = 0.45 gCO_2e/image.

Range: Generation time 17–28s → 2.07–3.41 Wh → 0.77–1.26 gCO_2e.

Confidence: Low.

[15]OpenAI, "Introducing 4o Image Generation," March 2025. Architecture: tokens → [transformer] → [diffusion] → pixels. https://openai.com/index/introducing-4o-image-generation/

5.6 Midjourney v6

Most opaque model. No API, no published architecture, no energy data, no cloud provider disclosure. Midjourney's workflow is unique among major providers: every prompt generates a 2×2 grid of 4 preview images (not optional), taking 30–60 seconds. The user then selects one image to upscale to full resolution — a separate computation step taking approximately 30–60 additional seconds. Three preview images are discarded. No other major provider (DALL-E 3, GPT-4o Image, Gemini/Imagen, Stable Diffusion) works this way — all generate one final image per request. Midjourney v7 uses 2× the GPU time of v6.[16]

Step 1: Effective server power = 439W (from anchor). Assumed GPU-class hardware; wide uncertainty band applied.

Step 2: Energy (per final delivered image, v6). Two stages: (a) Grid generation: 45s (central) × 439W ÷ 3600 = 5.49 Wh (produces 4 preview images; all compute consumed regardless of which is selected). (b) Upscale of selected image: 45s (central) × 439W ÷ 3600 = 5.49 Wh. Total energy per final image = 5.49 + 5.49 = 10.95 Wh.

Step 3: Hardware correction = ×1.0 (infrastructure unknown; assumed GPU-class). Note: this assumption cuts both ways. If Midjourney uses specialised or newer hardware (e.g., H200), the true energy could be lower. If they use older, less efficient GPUs (e.g., A100) or consumer-grade hardware to reduce costs, the energy could be significantly higher. The 1.0 factor represents maximum ignorance, not confidence.

Step 4: Location-based carbon = 10.95 × 0.370 (non-specific US) = 4.05 gCO_2e/image.

Step 5: CEAF-adjusted (Tier 3, CEAF 0%) = 4.05 × (1 − 0.00) = 4.05 gCO_2e/image — no adjustment; Midjourney has disclosed no clean energy information.

Range: Grid 30–60s + upscale 30–60s = total 60–120s effective → 7.32–14.63 Wh → 2.71–5.41 gCO_2e. For Midjourney v7 (2× GPU time): 5.41–10.82 gCO_2e/image.

Confidence: Very Low — no published information of any kind; infrastructure entirely unknown; workflow analysis from public documentation only.

6. Resolution Multipliers

Multipliers scale carbon estimates relative to the 1024×1024 baseline. Derived from sub-linear pixel-count scaling (pixel_ratio^0.85), cross-referenced against Bertazzini et al.'s empirical 1.3–4.7× range for resolution doubling.[17]

Resolution	Multiplier	Pixel Ratio	Assessment
512×512	0.40×	0.25×	Slightly generous vs formula (0.28×). Recommend 0.30×.
1024×1024	1.00×	1.00×	Baseline. Matches measured conditions.
1024×1792	1.60×	1.75×	Reasonable sub-linear scaling.
1792×1024	1.60×	1.75×	Same pixel count as above.
2048×2048	3.50×	4.00×	Upper end of empirical range. Defensible.

Table 3: Resolution multipliers.

[16]Midjourney subscription: Basic $10/mo ~200 images ($0.05/image). Standard $30/mo ~900 fast ($0.033). https://docs.midjourney.com/hc/en-us/articles/27870484040333-Comparing-Midjourney-Plans

7. Diffusion Steps Multipliers

Energy scales near-linearly with denoising steps. The 50-step ratio (1.93×) is directly measured by Chung/MIT; intermediate values are linear interpolations.[18][19]

Steps	Multiplier	Basis
15	0.65×	Linear interpolation (15/25 ≈ 0.60, rounded to 0.65)
20	0.85×	Linear interpolation (20/25 = 0.80, rounded to 0.85)
25	1.00×	Baseline. Directly measured: 2,282 J.
30	1.20×	Linear interpolation (30/25 = 1.20)
50	1.93×	Directly measured: 4,402 J / 2,282 J = 1.93×

Table 4: Diffusion step multipliers. Source: Chung & Chowdhury / ML.Energy, reported in MIT Technology Review May 2025.

8. Location-Based vs Clean Energy Adjustment Factor (CEAF)-Adjusted Accounting

This document reports estimates in two layers, following the GHG Protocol Scope 2 framework:

Location-based (gross): Energy consumed × grid carbon intensity at the data centre location. This reflects the physical carbon impact of the electricity drawn from the grid. It is the primary metric for Scope 2 transparency reporting and is required by IFRS S2.

CEAF-adjusted: Location-based × (1 − CEAF). The Clean Energy Adjustment Factor (CEAF) is a discount applied to reflect the provider's verified clean energy investment. It acknowledges that providers purchasing or generating renewable electricity reduce the marginal grid emissions attributable to their operations. CEAF-adjusted figures are the primary decision-support metric in the InferenceCarbon calculator, as they reward genuine clean energy procurement and better reflect the net carbon impact for comparative purposes.

CEAF = Location-based × (1 − CEAF)

CEAF tiers: Tier 1 — Verified hourly Carbon-Free Energy (CFE): Full percentage applied. Tier 1.5 — Verified annual renewable match, no hourly data: 50% of 100%. Tier 2 — Annual Renewable Energy Certificates (RECs), weaker verification: 50% of claimed percentage. Tier 3 — No disclosure: CEAF = 0%.

The CEAF-adjusted figure is presented as the primary estimate in this document. The location-based gross figure is retained as the physical baseline and is available for organisations requiring unadjusted Scope 2 reporting.

9. Confidence Assessment

Element	Confidence	Rationale
SD3 anchor (0.634 Wh)	Medium	Directly measured on H100. 2× overhead standard but unverified for diffusion.
A100 → H100 efficiency	Medium	H100 is 2–3× more efficient per FLOP. Well-established.
US grid intensity (370 g/kWh)	High	EPA eGRID. Well-established.
Commercial overhead (1.5–3×)	Low	No published data on API overhead vs raw inference.
GPT-4o visual token count	Low	Architecture partially disclosed; ~1,024 tokens speculative.
Midjourney hardware	Very Low	No published information of any kind.
Google CEAF (66%)	Medium	Published hourly CFE data; auditable.
Microsoft CEAF (50%)	Medium-Low	Annual match verified; hourly data absent.

Table 6: Confidence assessment.

10. Limitations

1. **Single hardware anchor.** The entire framework rests on one model (SD3 Medium) measured on one GPU (H100).

2. **No commercial model measurement exists.** DALL-E 3, GPT-4o Image, Gemini Image, and Midjourney have never been measured by any independent researcher.

3. **The 2× overhead multiplier is unvalidated for diffusion.** Diffusion models have different memory and compute profiles from LLMs.

4. **Architecture heterogeneity.** GPT-4o Image (autoregressive + diffusion hybrid), DALL-E 3 (pure diffusion), Midjourney (unknown) are fundamentally different.

5. **Quality settings matter enormously.** Commercial APIs may use 20–100 denoising steps internally, and this is not disclosed.

6. **Inference-only scope.** Excludes embodied emissions, training, and supply chain.

7. **Midjourney preview-then-upscale workflow. Midjourney generates 4 preview images before the user selects one for upscaling. This means each final image requires ~2× the generation time of a single-image model, plus 75% wasted preview compute. No other major provider works this way.7. CEAF uncertainty. Midjourney's 0% CEAF may understate their actual clean energy usage if they run on a major cloud provider (AWS, GCP, or Azure) and inherit that provider's renewable energy commitments. Without disclosure, we cannot verify this.** Midjourney's 0% CEAF may understate their actual clean energy usage if they run on a major cloud provider.

11. Response to Adversarial Review

This document has been subjected to adversarial review by Gemini 2.5 Pro (thinking) from dual perspectives: a PhD data centre engineer at each model provider, and a senior sustainability professional at each provider. This section summarises the key challenges and our responses.

11.1 Technical critiques

Constant power assumption (DVFS): The reviewer correctly noted that modern GPUs use Dynamic Voltage and Frequency Scaling, so power draw is not strictly constant at 380W. Accepted. However, during active diffusion inference, GPU utilisation is sustained at 70–90%, making ~439W a reasonable average for the generation period. We have added a note to Section 5.0 acknowledging DVFS and clarifying that the 380W figure applies to active inference only.

TPU efficiency factor too conservative (Google): The reviewer argued that Google's v5p liquid-cooled TPUs may be even more efficient than our ×0.5 correction implies, comparing our anchor to "an unoptimized open-source stack." Accepted. We have widened the TPU range to 0.25–0.6 and noted that our central Gemini estimate likely overstates Google's actual energy. This is a case where our methodology errs on the side of caution for the provider, not against them.

Midjourney hardware could be worse than assumed: The reviewer noted that if Midjourney runs on older A100s or less efficient spot instances, our 9.50 Wh estimate might undercount their energy. Accepted. We have added a note to Section 5.6 that the ×1.0 hardware factor represents maximum ignorance — the true figure could be higher or lower.

Latency ≠ compute 1:1: The reviewer noted that API generation time includes queue wait, safety filtering, and network latency. Already addressed. Section 5.0 explicitly caveats that non-compute overheads are typically small (<1s) relative to diffusion (5–30s) and that we do not discount for them, making estimates slightly conservative.

11.2 Sustainability critiques

Microsoft 50% CEAF is a "penalty": The reviewer argued that Microsoft's massive Power Purchase Agreements (PPAs) are inadequately reflected by a 50% CEAF. Respectfully rejected. The CEAF methodology is consistent across the entire InferenceCarbon series. Annual renewable matching is an accounting methodology, not a guarantee of carbon-free electricity every hour in every location — a point Microsoft itself has acknowledged. The GHG Protocol Scope 2 revision (public consultation 2025) proposes hourly matching as the standard. We commit to upgrading Microsoft to Tier 1 when fleet-wide hourly CFE data is published. Until then, 50% is a reasonable interim recognition of genuine investment.

Midjourney 0% CEAF assumes they "don't care": The reviewer argued that if Midjourney is hosted on a major cloud provider, they inherit that provider's green grid. Noted but maintained. The CEAF is designed to reward disclosure and investment, not to guess at hidden arrangements. If Midjourney operates on AWS or GCP, their actual footprint may be significantly lower than our estimate. However, without disclosure, we cannot verify this — and the CEAF framework must be applied consistently. We invite Midjourney to publish infrastructure information that would allow a non-zero CEAF. We have added a note to Limitation 7 acknowledging this possibility.

Scope 3 and "wasted" preview compute: The reviewer noted that our Scope 2-only boundary misses embodied hardware emissions, and that Midjourney's preview grid may save energy by reducing re-runs. On Scope 3: acknowledged and explicitly stated in the Scope Notice. On preview efficiency: this is an interesting behavioural argument but unverifiable. Our methodology counts the compute actually consumed per final delivered image, which is the appropriate boundary for operational emissions accounting.

11.3 Net impact

Central estimates are modestly affected: Gemini's range widens downward (TPU factor), Midjourney's uncertainty widens in both directions. No central estimate changes. Primary impact is methodological: DVFS acknowledged, hardware uncertainty made explicit, CEAF methodology defended. The reviewer's characterisation of the generation-time proxy as "a clever hack for an outsider" is fair — we accept that this is an estimation framework, not a measurement framework, and that all commercial model estimates carry Low or Very Low confidence ratings accordingly.

12. Key Sources

The primary empirical anchor is Chung & Chowdhury (2026), reported in MIT Technology Review (May 2025). Cross-validation from Luccioni et al. (2024) and Bertazzini et al. (2025). Commercial estimates from architectural analysis and pricing proxies. All sources cited in footnotes with URLs.