REFERENCE DOCUMENT

Estimating the Inference Carbon Intensity of AI Audio 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

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, supporting infrastructure (CPUs, memory, networking), and data centre overhead (cooling, power distribution), expressed via Power Usage Effectiveness (PUE).*

Important note: *Audio model energy data is extremely limited. The only published ASR energy study is El Bahri et al. (2025), a conference paper testing Whisper Base and Google Speech-to-Text (STT) on ~22 hours of audio. No published study has measured the energy consumption of any Text-to-Speech (TTS) model. All estimates in this document use a generation-time proxy methodology consistent with the InferenceCarbon image models reference.*

1. Executive Summary

This document provides carbon intensity estimates for AI audio models, covering both Automatic Speech Recognition (ASR, i.e. transcription) and TTS (i.e. voice generation). Estimates use a generation-time proxy methodology: energy = effective server power (380W, from Chung/MIT) × processing time, with hardware corrections for TPU efficiency and CEAF adjustments for clean energy. All estimates are reported as Clean Energy Adjustment Factor (CEAF)-adjusted figures for decision support, with location-based gross figures for Scope 2 transparency.

Model	Location-Based Range(gCO_2e/min)	Location-BasedCentral(gCO_2e/min)	CEAF %	CEAF-Adjusted Central(gCO_2e/min)	Confidence
Gemini Audio Input	0.007–0.019	0.012	66%	0.004	Low
Whisper Turbo	0.03–0.08	0.047	50%	0.023	Medium-Low
Whisper Large v3	0.16–0.47	0.29	50%	0.15	Medium
OpenAI TTS	0.12–0.39	0.20	50%	0.10	Low
ElevenLabs	0.20–0.59	0.31	0%	0.31	Low
OpenAI TTS HD	0.23–0.78	0.39	50%	0.20	Low

Table 1: Summary estimates. Location-based at per-provider grid intensity (see Section 3). CEAF: Google 66% (hourly CFE); OpenAI/Azure 50% (annual match); ElevenLabs 0% (undisclosed infrastructure). Sorted by CEAF-adjusted central. All estimates derived via generation-time proxy (Section 4.1).

2. What Has Been Measured

El Bahri, Kouissi & Achkari Begdouri (2025): The only published ASR energy study. Presented at SCGT'2025, published in MDPI Computer Sciences & Mathematics Forum. Compared OpenAI Whisper Base (74M parameters, run locally on RTX A1000 laptop GPU) with Google STT (cloud API) on ~22.2 hours of Urdu audio. Whisper Base local energy: 0.43–0.53 kWh (average ~0.49 kWh across three measurement tools: PowerAPI, PyJoules, CodeCarbon). Google STT cloud energy: 0.35 kWh (via Google Cloud Carbon Footprint tool). Google STT was ~51% faster in processing time (122 min vs 238 min) and produced ~42% fewer emissions. Note: This compares cloud vs local deployment, not architecture vs architecture on equivalent hardware.[1]

[1]El Bahri, J., Kouissi, M. & Achkari Begdouri, M. (2025) "Comparative Analysis of Energy Consumption and Carbon Footprint in ASR Systems." Comput. Sci. Math. Forum 10(1), 6. Whisper Base (74M params): 0.43-0.53 kWh local; Google STT cloud: 0.35 kWh. 22.2 hr Urdu audio on RTX A1000. https://doi.org/10.3390/cmsf2025010006

No published study has measured the energy of any TTS model — not OpenAI TTS, not ElevenLabs, not any other voice synthesis system.

3. Provider Infrastructure and Grid Context

Provider	Model(s)	Infrastructure	Primary Data Centre Locations	CEAF %
Google	Gemini Audio	Google Cloud TPUs	US (Oregon, Iowa), EU, APAC	66%
OpenAI	Whisper, TTS, TTS HD	Microsoft Azure	US East (Virginia), US Central	50%
ElevenLabs	ElevenLabs TTS	Undisclosed	Unknown	0%

Table 2: Provider infrastructure. Google CEAF from published 66% hourly CFE (Google 2025 Environmental Report). OpenAI/Azure from 100% annual renewable match (Microsoft, Feb 2026). ElevenLabs: no disclosure.

4. Anchor and Scaling Methodology

4.1 Generation-Time Proxy

Consistent with the InferenceCarbon image models reference, we use API processing time as the primary energy proxy. During active inference, GPU power draw is approximately constant, so energy ≈ power × time. For ASR models, the proxy is the Real-Time Factor (RTF): how many times faster than real-time the model processes audio. For TTS models, the proxy is generation time: how many seconds of compute are needed to produce one minute of output audio.

Effective server power: 380W, derived from the Chung/MIT measurement of Stable Diffusion 3 on H100 (0.634 Wh in ~6 seconds = 380W effective, including the 2× overhead multiplier for cooling, CPUs, and networking). This figure is used consistently across the InferenceCarbon series.[2]

The 5-step derivation for each model: Step 1: Effective server power = 380W. Step 2: Compute time per minute of audio (from RTF for ASR, from generation time for TTS). Energy = 380W × compute_time ÷ 3600. Step 3: Hardware correction factor (Google TPU ×0.5; same GPU class ×1.0; unknown ×1.0). Step 4: Location-based carbon = Wh × 0.370 gCO_2e/Wh. Step 5: CEAF-adjusted = location-based × (1 − CEAF%).

4.2 Cross-Validation: El Bahri et al. (2025)

El Bahri et al. (2025) tested Whisper Base (74M parameters) on a local RTX A1000 laptop GPU, processing ~22.2 hours of Urdu audio. Three measurement tools gave Whisper Base local energy of 0.43–0.53 kWh (average ~0.49 kWh), or 0.37 Wh per minute of audio. Google STT cloud processing of the same audio consumed 0.35 kWh total (via Google Cloud Carbon Footprint tool).

Cross-validation: Our Whisper Large v3 estimate (0.79 Wh/min on H100) is ~2× the El Bahri Whisper Base figure (0.37 Wh/min on laptop). Whisper Large is ~20× larger than Base but runs on vastly more powerful hardware (H100 vs RTX A1000). The same order of magnitude provides directional validation.

5. Per-Model Estimates

All estimates use the generation-time proxy (Section 4.1) with 380W effective server power and per-provider grid carbon intensity (OpenAI/Azure: 350 gCO_2e/kWh; Google/GCP: 375 gCO_2e/kWh; ElevenLabs: 370 gCO_2e/kWh US-weighted default). Ranges reflect uncertainty in processing speed (RTF or generation time) and, where applicable, hardware correction factors.

[2] Chung, J.-W. & Chowdhury, M. (2026) "Where Do the Joules Go? Diagnosing Inference Energy Consumption." arXiv:2601.22076. 380W effective server power derived from SD3 Medium measurement. https://arxiv.org/abs/2601.22076

5.1 Whisper Large v3 (Automatic Speech Recognition (ASR))

OpenAI's flagship speech recognition model (1.55B parameters). Runs on OpenAI API (Microsoft Azure, H100/A100). Whisper Large v3 processes audio at approximately 8× real-time on production H100 hardware (derived from Whisper Turbo being 6× faster than Large, with Turbo at ~50× real-time on standard GPU deployment).[3][4]

Step 1: Effective server power = 380W (from Chung/MIT, Section 4.1).

Step 2: Compute time = 60s ÷ 8 (RTF) = 7.5 seconds per minute of audio. Energy = 380W × 7.5 ÷ 3600 = 0.79 Wh/min.

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

Step 4: Location-based carbon = 0.79 × 0.350 (OpenAI/Azure) = 0.28 gCO_2e/min.

Step 5: CEAF-adjusted (Microsoft Azure, Tier 1.5, CEAF 50%) = 0.29 × (1 − 0.50) = 0.15 gCO_2e/min.

Range: RTF 5–15× → 4.0–12.0s compute → 0.16–0.47 gCO_2e/min.

Confidence: Medium — RTF well-documented for Whisper variants; hardware assumptions standard.

5.2 Whisper Turbo (ASR)

Optimised variant reducing decoder layers from 32 to 4 (809M parameters). Achieves 6× faster inference than Large v3 while maintaining accuracy within 1–2% (OpenAI, Northflank benchmarks).[5] Groq reports 216× real-time on specialised hardware; standard GPU deployment estimated at ~50× real-time.

Step 1: Effective server power = 380W.

Step 2: Compute time = 60s ÷ 50 (RTF) = 1.2 seconds per minute of audio. Energy = 380W × 1.2 ÷ 3600 = 0.13 Wh/min.

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

Step 4: Location-based carbon = 0.13 × 0.350 (OpenAI/Azure) = 0.046 gCO_2e/min.

Step 5: CEAF-adjusted (Azure, 50%) = 0.047 × 0.50 = 0.023 gCO_2e/min.

Range: RTF 30–90× → 0.03–0.08 gCO_2e/min.

Confidence: Medium-Low — RTF varies significantly by deployment optimisation. Production deployments typically achieve much higher throughput through batching (Turbo's reduced 4-layer decoder allows significantly higher batch sizes per GPU), which would reduce per-request energy below our estimate.

5.3 Gemini Audio Input (ASR)

Google's speech recognition runs on custom TPUs with PUE 1.10 and 66% hourly CFE. Google Cloud Speech-to-Text is extremely fast; El Bahri found it processed 22 hours of audio in 122 minutes (2× faster than local Whisper Base). Production Google Cloud STT on TPUs is estimated at ~100× real-time or faster.

Step 1: Effective server power = 380W.

Step 2: Compute time = 60s ÷ 100 (RTF) = 0.6 seconds per minute of audio. Energy (pre-correction) = 380W × 0.6 ÷ 3600 = 0.063 Wh.

---

[3] OpenAI Whisper GitHub. Turbo model: decoder layers reduced from 32 to 4; 6x speed improvement. https://github.com/openai/whisper

[4] Groq, "Whisper Large v3 Turbo Now Available on Groq." 216x real-time speed factor on Groq hardware. https://groq.com/blog/whisper-large-v3-turbo-now-available-on-groq-combining-speed-quality-for-speech-recognition

[5] Northflank, "Best open source speech-to-text model in 2026." Whisper Turbo: 6x faster than Large v3. https://northflank.com/blog/best-open-source-speech-to-text-stt-model-in-2026-benchmarks

Step 3: Hardware correction = ×0.5 (central). Google's TPUs are 2–3× more energy-efficient per FLOP than H100 (Patterson et al. 2022). Google's latest v5p/v6 TPUs, purpose-built for Transformer operations used in ASR, may achieve even greater efficiency (possibly 3–4×, implying a correction of 0.25–0.33). As with the image paper, we use 0.5 as a conservative central estimate; our Gemini figure likely overstates Google's actual energy. Corrected energy = 0.063 × 0.5 = 0.032 Wh/min.

Step 4: Location-based carbon = 0.032 × 0.375 (Google) = 0.012 gCO_2e/min.

Step 5: CEAF-adjusted (Google, Tier 1, CEAF 66%) = 0.012 × (1 − 0.66) = 0.004 gCO_2e/min.

Range: RTF 50–200× + TPU factor 0.25–0.6 → 0.007–0.019 gCO_2e/min.

Confidence: Low — no published energy data; inferred from processing speed and Google's TPU efficiency. Note: For long-form audio, Google may parallelise across TPU pods, meaning effective power is distributed across a cluster with a different efficiency profile than a single 380W server. This adds further uncertainty to the estimate.

5.4 OpenAI TTS-1 (Text-to-Speech)

Neural voice synthesis on Microsoft Azure (H100/A100). OpenAI documents ~0.5s latency (time-to-first-byte) for TTS-1.[6] For 1 minute of speech output (~150 words, ~750 characters, within the 4,096-character API limit), total generation time is estimated at ~5 seconds (generates faster than real-time). TTS uses autoregressive or flow-based synthesis, with GPU utilisation sustained during generation.

Step 1: Effective server power = 380W.

Step 2: Generation time = 5 seconds per minute of output audio (central). Energy = 380W × 5 ÷ 3600 = 0.53 Wh/min.

Step 3: Hardware correction = ×1.0 (H100/A100, same GPU class). Note: TTS uses autoregressive or flow-based synthesis where compute intensity varies during generation (unlike diffusion, where GPU utilisation is more constant). The 380W figure likely overstates average TTS power draw, making this estimate conservative.

Step 4: Location-based carbon = 0.53 × 0.350 (OpenAI/Azure) = 0.19 gCO_2e/min.

Step 5: CEAF-adjusted (Azure, 50%) = 0.20 × 0.50 = 0.10 gCO_2e/min.

Range: Generation time 3–10s → 0.12–0.39 gCO_2e/min.

Confidence: Low — no published TTS energy data; generation time from API documentation and latency reports.

5.5 ElevenLabs (Text-to-Speech)

Premium voice synthesis service. ElevenLabs Flash v2.5 achieves 75ms inference latency; full generation of 1 minute of speech estimated at ~8 seconds. Infrastructure undisclosed — no published architecture, hardware, or cloud provider information. ElevenLabs ranks #3 on Artificial Analysis TTS leaderboard (ELO 1,108).[7]

Step 1: Effective server power = 380W (assumed GPU-class hardware).

Step 2: Generation time = 8 seconds per minute of output audio (central). Energy = 380W × 8 ÷ 3600 = 0.84 Wh/min.

Step 3: Hardware correction = ×1.0 (infrastructure unknown). This does not assume ElevenLabs is inefficient — a specialised TTS provider may well use inference-optimised hardware (e.g., L4 GPUs,

[6] OpenAI, Text-to-Speech API documentation. TTS-1 latency ~0.5s; TTS HD 2x pricing ($30 vs $15/1M chars). https://platform.openai.com/docs/guides/text-to-speech

[7] Artificial Analysis, TTS Model Leaderboard (accessed March 2026). ElevenLabs ELO 1,108 (#3). https://artificialanalysis.ai/text-to-speech/arena

custom ASICs) that draws significantly less than 380W. The ×1.0 factor represents unknown direction of error, not an assumption of inefficiency.

Step 4: Location-based carbon = 0.84 × 0.370 (US-weighted default) = 0.31 gCO_2e/min.

Step 5: CEAF-adjusted (Tier 3, CEAF 0%) = 0.31 × (1 − 0.00) = 0.31 gCO_2e/min — no adjustment; infrastructure undisclosed.

Range: Generation time 5–15s → 0.20–0.59 gCO_2e/min.

Confidence: Low — no published data; infrastructure entirely unknown.

5.6 OpenAI TTS HD (Text-to-Speech)

Higher-fidelity variant of OpenAI TTS. API pricing is 2× standard ($30/1M chars vs $15/1M chars), suggesting approximately 2× the compute per minute of output. Caveat: pricing reflects R&D, market positioning, and hardware availability, not just electricity cost. The 2× compute assumption from pricing is a rough proxy. Same Azure infrastructure as TTS-1.

Step 1: Effective server power = 380W.

Step 2: Generation time = 10 seconds per minute of output audio (2× standard TTS, based on 2× pricing). Energy = 380W × 10 ÷ 3600 = 1.06 Wh/min.

Step 3: Hardware correction = ×1.0 (same Azure H100/A100 infrastructure).

Step 4: Location-based carbon = 1.06 × 0.350 (OpenAI/Azure) = 0.37 gCO_2e/min.

Step 5: CEAF-adjusted (Azure, 50%) = 0.39 × 0.50 = 0.20 gCO_2e/min.

Range: Generation time 6–20s → 0.23–0.78 gCO_2e/min.

Confidence: Low — generation time estimated from pricing ratio only.

6. 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 recognise 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.

7. Confidence Assessment

Element	Confidence	Rationale
El Bahri Whisper measurement	Medium	Published peer-reviewed conference paper. Small dataset (22 hrs). Single hardware config.
Google STT 51% lower	Medium-Low	Single comparative study. Cloud vs local deployment differs.
TTS generation-time proxy	Low	No published TTS energy data. Generation time used as proxy (Section 5.0).
ElevenLabs estimate	Low	No published data; infrastructure unknown.
Google CEAF (66%)	Medium	Published hourly CFE data; auditable.
OpenAI/Azure CEAF (50%)	Medium-Low	Annual match verified; hourly data absent.

Table 4: Confidence assessment.

8. Limitations

1. **Single small-scale anchor.** El Bahri tested ~22 hours of audio on one hardware configuration.

2. **No TTS energy measurement exists.** TTS estimates use generation-time proxy (Section 4.1) rather than a fixed multiplier, but remain unvalidated against any direct measurement.

3. **ElevenLabs is a black box.** No architecture, hardware, or energy information published.

4. **El Bahri measured Whisper Base (not Large) on local laptop hardware (RTX A1000).** The study is used for cross-validation only, not as the primary anchor (see Section 4.2).

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

9. Response to Adversarial Review

This document has been subjected to internal adversarial review. Key findings and responses:

9.1 Critical anchor correction

v1.0 attributed a 0.35 kWh figure to "Whisper Large" processing 22 hours of audio. Review of the source paper (El Bahri et al. 2025) revealed three errors: (a) the model tested was Whisper Base (74M), not Whisper Large (1.55B); (b) the 0.35 kWh was Google STT's cloud energy, not Whisper's; (c) the "51% lower" comparison was cloud-vs-local, not architecture-vs-architecture. All v1.0 downstream calculations were therefore based on a misidentified data point. v2.0 adopts the generation-time proxy methodology consistent with the image models reference.

9.2 TTS estimates remain speculative

No published study has measured the energy consumption of any TTS model. Our TTS estimates (Sections 5.4–5.6) use generation time as a proxy, which is an improvement over v1.0's unsourced multipliers (1.8×, 2.5×) but remains speculative. The generation-time approach has the advantage of being grounded in an observable, benchmarkable metric (API latency / generation speed), but the assumption that GPU power draw during TTS is comparable to diffusion inference (~380W) may not hold for all TTS architectures. We flag all TTS estimates as Low confidence accordingly.

9.3 ElevenLabs CEAF

As with Midjourney in the image paper, ElevenLabs receives 0% CEAF because it has disclosed no infrastructure information. If ElevenLabs runs on a major cloud provider (AWS, GCP, Azure), its actual footprint may be lower. The CEAF rewards disclosure; without it, we cannot verify clean energy claims.

9.4 Net impact of corrections

Whisper Large v3 central estimate increased from 0.098 to 0.29 gCO_2e/min (3× higher) because v1.0 used a much-too-low anchor. TTS estimates changed modestly: OpenAI TTS from 0.18 to 0.20, ElevenLabs from 0.25 to 0.31. The rank ordering is largely preserved. All estimates now have explicit 5-step derivations traceable to the Chung/MIT server power anchor.

9.5 External adversarial review (Gemini 2.5 Pro)

The document was reviewed by Gemini 2.5 Pro (thinking) from dual perspectives at each provider: a PhD data centre engineer and a senior sustainability executive.

9.5.1 Technical critiques accepted

Google TPU factor too conservative: The reviewer argued that v5p/v6 TPUs purpose-built for Transformer ASR may achieve 3–4× efficiency vs H100, implying a correction of 0.25–0.33 rather than our 0.5. Accepted. We have widened the TPU range to 0.25–0.6 and noted that our Gemini estimate likely overstates Google's actual energy. We retain 0.5 as a conservative central estimate.

Pod parallelism for long-form audio: Google may parallelise across TPU pods for long audio, changing the effective power profile. Accepted as a caveat; we cannot model distributed inference without provider disclosure.

TTS autoregressive power profile: The reviewer noted that TTS power draw varies during generation (unlike diffusion where GPU utilisation is constant). Accepted. The 380W figure likely overstates TTS energy, making our estimates conservative. Caveat added to Section 5.4.

Whisper Turbo batching efficiency: Production deployments batch requests, reducing per-request energy. Accepted as a caveat in Section 5.2.

ElevenLabs may use specialised hardware: If ElevenLabs uses inference-optimised chips (L4, ASICs), the 380W assumption overstates their energy. Accepted — we have clarified that ×1.0 represents unknown direction, not assumed inefficiency.

Pricing as compute proxy (TTS HD): 2× pricing does not necessarily equal 2× compute. Partially accepted. Pricing reflects business positioning, not just electricity. But it is our best available signal; caveat added to Section 5.6.

9.5.2 Sustainability critiques

Microsoft 50% CEAF "penalty": Respectfully rejected. Consistent with the full InferenceCarbon series. Annual renewable matching is an accounting methodology, not a guarantee of carbon-free electricity every hour. We commit to upgrading Microsoft to Tier 1 when fleet-wide hourly CFE data is published.

Google grid locality: The reviewer argued that Google sites audio processing in low-carbon regions (Oregon, Iowa) and that US-weighted 370 gCO_2e/kWh overstates their grid intensity. Noted but rejected for consistency. We cannot know which data centre handles any given API call. The US-weighted average is applied consistently across all providers in the series.

ElevenLabs 0% CEAF "discrimination": The reviewer argued that ElevenLabs likely uses a major cloud provider and inherits its renewable portfolio. Noted but maintained. The CEAF rewards disclosure, not guesswork. If ElevenLabs discloses its infrastructure, we will assign an appropriate CEAF. We acknowledge that ElevenLabs is a younger company and that sustainability reporting lags operational maturity.

Google's 66% hourly CFE as "physical reality": Google argues that hourly matching is the actual electrons, not just a discount. We agree that Google's Tier 1 status is the highest in our framework, and their 66% figure is applied directly — we do not discount it. The remaining 34% reflects the physical reality that Google's grid is not 100% carbon-free every hour.

9.5.3 Net impact

No central estimates changed. Gemini's range widens downward (TPU factor). All TTS estimates are flagged as likely conservative (autoregressive power draw < diffusion). Whisper Turbo and ElevenLabs may benefit from production optimisations not captured by our model. The reviewer's characterisation of 380W as derived from "an open-source SD3 anchor" is fair — we accept that applying a diffusion-derived power figure to speech models is a cross-modality approximation.

10. Key Sources

Primary empirical source: El Bahri et al. (2025). All TTS estimates are engineering extrapolations with no independent anchor. Pricing verified from OpenAI, Google, and ElevenLabs official sources (March 2026). All sources cited in footnotes with URLs.