

The Carbon Dividend of Right-Sizing

Internal study — estimated CO₂e savings from routing prompts to the smallest capable model, bounded by published per-token energy and grid carbon-intensity literature.

Generated: 2026-09-12 01:18 UTC

The Carbon Dividend of Right-Sizing

Internal study — September 11, 2026

Version: 1.1

Last updated: 2026-09-11

Abstract

Right-sizing — Elis's practice of routing each task to the smallest model tier capable of doing it, rather than defaulting every prompt to a frontier model — is documented elsewhere in this Whitepapers library as a token and cost optimization. This note asks a narrower question: what does the same token reduction imply for estimated CO₂e emissions? We combine our own measured token-avoidance ratios with externally published per-token energy and grid carbon-intensity figures to produce a bounded estimate, not a platform-measured footprint. Every energy and carbon figure below is borrowed from external published sources, cited individually; only the token-reduction percentages are our own measurements. Central estimate: routing 1 million assistant prompts through Elis's measured 80% token reduction avoids roughly 75 kg CO₂e versus serving them all on a frontier model — comparable to a few hundred kilometers of average passenger-car driving. The range across scenarios spans roughly 30x, and the reasons for that spread are the point of this paper as much as the headline number.

1. Background

Right-sizing means matching each task's actual capability requirement to the smallest model tier that can serve it, instead of defaulting every request to the most capable (and most compute-hungry) frontier model. Elis's dynamic model scorer performs this routing at runtime, per CLAUDE.md Non-Negotiable #1 (never hardcode a model) and #4 (keep orchestration dynamic) — the platform selects capabilities, and a scorer binds the concrete model only at execution time.

Elis has measured, on its own platform, that right-sizing meaningfully cuts the number of tokens a frontier model would otherwise have had to generate or process. Two independent measurements exist (see §2). This paper asks what avoiding those frontier-model tokens is plausibly worth in CO₂e, building the estimate from external energy-per-token literature rather than from any energy measurement Elis has made itself — the platform does not instrument miner power draw, and no such claim is made here.

2. Methodology

Formula.

$$\text{CO2e saved} = \text{frontier tokens avoided} \times \Delta E_{\text{per_token}} \times \text{PUE factor} \times \text{grid carbon intensity}$$

Frontier tokens avoided is a measured ratio; the other three factors are externally published estimates, combined explicitly below rather than presented as one black-box number.

Per-token energy — frontier side. Google's production measurement of its Gemini serving fleet reports a median 0.24 Wh per text prompt, already inclusive of the fleet's measured Power Usage Effectiveness (PUE ≈ 1.09) (Elkin et al., arXiv:2508.15734, Aug 2025; Google Cloud blog, "Measuring the environmental impact of AI inference"). Converting a per-prompt figure to a per-token figure requires an assumed prompt length; we use 750 tokens as the midpoint of a 500–1,000 token band, giving:

$$0.24 \text{ Wh} / 750 \text{ tokens} \approx 0.32 \text{ mWh/token (frontier central; band } 0.24\text{--}0.48 \text{ mWh/token across the } 500\text{--}1,000 \text{ token range)}$$

Two independent figures corroborate this band: Epoch AI's estimate for GPT-4o of roughly 0.3 Wh per query (IQR 0.16–0.60 Wh) (epoch.ai, "How much energy does ChatGPT use?"), and OpenAI's own stated figure of 0.34 Wh per query. All three cluster around the same order of magnitude.

Per-token energy — small-model side. We deliberately do **not** use the $\sim 40\times$ energy gap Luccioni et al. measured between BLOOM-176B and a 7B model (FAccT 2024, arXiv:2311.16863) as the small-model discount, because it is a comparison of two specific historical models rather than a general large-vs-small ratio, and current serving practice narrows it further: MLPerf inference benchmarks show near-linear (not super-linear) energy scaling with parameter count at fixed hardware, and production serving techniques compress the gap further still — batching reduces per-query energy by an estimated 25–55%, and quantization techniques (e.g. FP8) enable further efficiency gains (MLPerf Power) (see arXiv:2505.09598 and arXiv:2410.12032 for the batching and quantization measurements respectively). We therefore use a conservative 1/10 reduction in per-token energy for the small model relative to the frontier model — deliberately smaller than the $\sim 40\times$ literature datapoint, so that this paper understates rather than overstates the energy benefit of right-sizing:

$$\Delta E_{\text{per_token}}(\text{central}) = 0.32 \text{ mWh/token frontier} - 0.032 \text{ mWh/token small-model} \approx 0.29 \text{ mWh/token avoided}$$

PUE handling. The Google figure above is already facility-inclusive (it bakes in Google's fleet PUE of ~1.09), so no additional PUE multiplier is applied to it. The Luccioni et al. BLOOM figure, by contrast, is an IT-load (compute-only) measurement, so the HIGH scenario below applies an explicit PUE multiplier to it rather than double-counting or omitting facility overhead.

Grid carbon-intensity scenarios. CO2e per frontier-token-avoided, by scenario:

Scen ario	Grid intensity	Source	CO2e per token avoided
LOW	384 gCO2/kWh (Ember 2024 actual, US grid average; 393 in 2023)	Ember, "US Electricity 2025 Special Report"	0.11 mg
CEN TRA L	435 gCO2/kWh (world average grid)	IEA, "Electricity 2026" (emissions)	0.125 mg
HIG H	435 gCO2/kWh, applied to an older-literature energy figure (BLOOM ~4 Wh/query IT-load, ×1.54 industry-average PUE)	Luccioni et al. FAccT 2024 (arXiv:2311.16863); Uptime Institute 2025 PUE survey	3.22 mg

LOW and CENTRAL differ from each other **only** by grid carbon intensity — same energy-per-token assumption, different regional grid mix. The LOW figure uses Ember's actual 2024 US grid-average intensity (384 gCO2/kWh, up from 393 in 2023); using the actual figure rather than an older, lower estimate **narrows, rather than widens**, the gap between LOW and the CENTRAL world-average estimate (435 gCO2/kWh). HIGH swaps in an entirely different, older, larger, less-optimized reference system (an IT-load-only measurement of a 176B-parameter model with an industry-average PUE applied on top) rather than a different grid. The resulting ~26x spread between CENTRAL and HIGH therefore reflects two compounding effects at once — the 5–20x methodology gap documented across the published energy-per-query literature, and genuine efficiency gains across model generations (smaller, more optimized serving vs. an older large monolithic model) — and is not a tight statistical confidence interval. Treat HIGH as

a deliberately conservative upper bound built from an older, worse-case reference system, not as equally likely to the CENTRAL estimate.

Savings bases. Two independent, platform-measured token-avoidance ratios are carried through side by side rather than collapsed into one number:

1. **80% basis** — measured on one real production account over 30 days, reported (not projected). At a 750-token frontier prompt, this avoids 600 tokens per prompt.
2. **44.8% basis** — fewer tokens measured across 508 of our own platform plans (the same measurement underlying the separately reported 1.8x quota-runway token ratio, p50 1.73x). At a 750-token frontier prompt, this avoids 336 tokens per prompt.

Both are our own platform's measured token ratios; neither is a projection or an externally sourced figure. See §5 for what these ratios do and do not imply.

2.7 Toward per-run granularity: tier-resolved accounting

Motivation. The platform records, for every orchestration step, which concrete model executed it and its token counts (CLAUDE.md Non-Negotiable

5 — trace everything). That trace already carries the information needed

to compute CO2e per run from the *actual* tier mix a run used, rather than from the blanket small-model assumption in §2 above, which credits every avoided frontier token to a single small-tier model regardless of which tier actually served it.

Per-token energy vector. We extend the two-point (frontier / small) energy comparison in §2 to three tiers:

$E = (\text{frontier: } 0.32, \text{ small: } 0.032, \text{ nano: } 0.011) \text{ mWh/token}$

Frontier is derived as in §2. Small keeps the same conservative 1/10 reduction used throughout this paper (vs. the measured ~40x BLOOM 176B-vs-7B gap — see §2 and §4). Nano is an explicit, stated **assumption**, set at 1/3 of the small-tier figure. That assumption is consistent with, and bounded above by, the published GPT-4.1-nano figure of under 0.5 Wh per prompt ("How Hungry is AI," arXiv:2505.09598) — but that

published figure is a loose per-prompt upper bound, not a per-token measurement, so it constrains rather than confirms the 0.011 mWh/token value used here.

Formulas.

```
C02e_actual(run) = Σ over steps [ tokens_step × E(tier of executing model) × grid intensity ]

C02e_counterfactual(run) = frontier-baseline tokens × E_frontier × grid intensity

savings(run) = C02e_counterfactual(run) - C02e_actual(run)
```

The blanket method used in §2–§3 is the special case of this formula where every right-sized token is assumed to have run on the small tier; substituting 100% small-tier occupancy into C02e_actual reproduces the §3 central figure (0.125 mg CO2e per token avoided) exactly.

Illustrative table (ILLUSTRATIVE — assumed small/nano tier splits, not a measured production mix). Central, 80%-basis, per-1,000,000-prompt figure under increasing nano-tier share:

Small / nano split	Per-1M-prompt CO2e (central)	vs. blanket (100/0)
100 / 0 (blanket, §3)	75 kg	—
70 / 30	~77 kg	+2.3%
50 / 50	~78 kg	+3.8%
30 / 70	~79 kg	+5.3%

Under these conservative tier factors, the granular method shifts the central estimate only modestly (+2–5%) versus the blanket §3 figure; its main value is precision and per-run auditability rather than a materially different headline number. Because nano energy is assumed strictly lower than small energy, tier-resolved accounting can only raise, never lower, the blanket estimate.

A measured production tier mix will replace the illustrative splits above once per-step token usage records are fully populated across the fleet; until then, the published figures in this paper remain the blanket (conservative) method of §2–§3.

A known conservative bias. Published measurements put frontier *reasoning* models above 33 Wh per long prompt (same source, "How Hungry is AI," arXiv:2505.09598) — far above the 0.24 Wh/prompt central frontier figure used throughout this paper (derived from Google's Gemini serving fleet, §2). Savings estimated against a reasoning-heavy frontier baseline are therefore understated by this paper's method.

3. Results

All figures rounded to 2 significant figures.

Per 1,000,000 assistant prompts (frontier-token-avoidance basis × per-token CO2e scenario):

Basis	LOW	CENTRAL	HIGH
80% tokens avoided	66 kg	75 kg	1.9 t
44.8% tokens avoided	—	42 kg	—

(The 44.8% basis is reported at CENTRAL only, to avoid presenting six numbers where the two savings bases are meant to bracket a range, not multiply one.)

Equivalence, CENTRAL / 80% basis. 75 kg CO2e is comparable to roughly 300–440 km of average passenger-car driving in the US, using the EPA's published average passenger-vehicle emissions figure of about 400 g CO2e per mile (~250 g/km US-specific) alongside a broader global average vehicle band of 170–192 g/km; we present a range rather than one figure because the per-km factor itself varies by vehicle mix and geography (EPA, "Greenhouse Gas Emissions from a Typical Passenger Vehicle").

Illustrative organization. A 50-seat organization at 20 prompts/seat/day for 30 days is 30,000 prompts/month — **this volume assumption is illustrative, not measured**; no claim is made that any actual customer account matches this shape. At that illustrative volume:

Basis	CENTRAL	HIGH
80% tokens avoided	2.3 kg/month	58 kg/month
44.8% tokens avoided	1.3 kg/month	—

4. Sensitivity & limitations

- **The CENTRAL-to-HIGH spread (~26x) is not a confidence interval.** It mixes a documented 5–20x methodology gap in the published energy-per-query literature with real efficiency differences between an older large monolithic model and current optimized serving. Treat HIGH as a deliberately conservative (i.e., large) upper bound, not as an equally likely outcome to CENTRAL.
- **The small-model energy discount (1/10) is conservative by design.** The literature datapoint most directly comparable (Luccioni et al.'s BLOOM-176B vs. 7B comparison) suggests up to ~40x; we use 10x because MLPerf scaling data and production batching/quantization gains compress the practical gap, and because using a smaller discount than the literature maximum keeps this estimate on the conservative side.
- **PUE is applied once, correctly scoped.** The Google frontier figure is facility-inclusive already; the Luccioni-derived HIGH-scenario figure is IT-load-only and has an explicit PUE multiplier applied. Conflating the two — e.g. applying a PUE multiplier to the already-inclusive Google figure — would double-count facility overhead.
- **Results are grid-mix and serving-configuration dependent.** Batching, quantization, and the operator's actual PUE can each shift outcomes by tens of percent in either direction; the scenarios above bracket a plausible range rather than pin down a single true value.
- **Prompt-length assumption.** The 750-token midpoint (500–1,000 band) used to convert Google's per-prompt figure to a per-token figure is an assumption, not a measurement of Elis's own traffic mix.

5. Caveats

The 80% and 44.8% figures are measured token ratios from our own platform — the 80% figure from one real production account over 30 days (reported, not projected); the 44.8%-fewer-tokens / 1.8x quota-runway figure from 508 of our own platform plans (p50 1.73x, max 2.76x). All energy and carbon-intensity factors are external published estimates, not our measurements (Google, Epoch AI, OpenAI-stated, Luccioni et al. FAccT 2024, IEA 2025, Uptime Institute 2025). Estimates are inference-only: they exclude training-run amortization and embodied hardware carbon — embodied (manufacturing/hardware) carbon adds an estimated 10–30% on top of

operational emissions depending on server lifespan; the cited Nature Scientific Reports study applies a conservative 2x multiplier to operational emissions as a bounding assumption (Nature Scientific Reports, 2024) — so total footprints are higher than inference-only figures. Results are grid-mix and serving-configuration dependent (batching, quantization technique, and PUE shift outcomes by tens of percent), and published energy-per-query estimates span 5–20x by methodology — hence ranges rather than a point estimate. Water usage is not standardized across sources and is not estimated here. Cost figures do not transfer from token ratios: in our first controlled benchmark, estimated model cost fell 36.3% while visible token volume rose 40.7%. No dollar figure is used in this CO2e calculation. The per-tier breakdowns in §2 are illustrative, built on conservative assumed energy ratios, not a measured production tier mix; this document will be updated when the measured distribution is available.

6. References

1. Elkin et al., "Measuring the environmental impact of AI inference," arXiv:2508.15734 (Aug 2025). <https://arxiv.org/abs/2508.15734>
2. Google Cloud, "Measuring the environmental impact of AI inference." <https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference>
3. Epoch AI, "How much energy does ChatGPT use?" <https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use>
4. Luccioni, Jernite, Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" FAccT 2024, arXiv:2311.16863. <https://arxiv.org/abs/2311.16863>
5. IEA, "Electricity 2026" (emissions). <https://www.iea.org/reports/electricity-2026/emissions>
6. Ember, "US Electricity 2025 Special Report." <https://ember-energy.org/latest-insights/us-electricity-2025-special-report/>
7. Google, "Efficiency — Data Centers." <https://datacenters.google/efficiency/>
8. Uptime Institute, 2025 Global Data Center Survey (PUE). <https://uptimeinstitute.com>
9. Nature Scientific Reports, "Embodied carbon and AI hardware lifecycle" (2024). <https://www.nature.com/articles/s41598-024-76682-6>
10. "Energy-efficient inference serving via batching" analysis. <https://arxiv.org/pdf/2505.09598>
11. "Quantization for efficient LLM inference" analysis. <https://arxiv.org/pdf/2410.12032>

12. US EPA, "Greenhouse Gas Emissions from a Typical Passenger Vehicle."

<https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle>