Cryptographically Attested Environmental Accounting for LLM Inference

Jasraj Budigam

Indus International School, Hyderabad, India
jasrajharikrishna.b@indusschoolhyd.com

Abstract. Large language models need verifiable environmental accounting at the point of use. We present a cryptographically attested, per-request environmental accounting framework for LLM inference. The design aligns with EAT (RFC 9711) semantics and produces COSE-signed receipts that bind energy traces and environmental factors to an invocation. We embed timestamped ElectricityMaps responses in each receipt. During our runs the API returned fallback values, so we report both fallback results and a post-hoc historical correction at the same timestamps. The fallback contrast between GB and continental zones produces an apparent 7.5× spread. Historical correction changes regional ranks and lowers mean per-request CO_2 by about 32%. Because clipping creates left-censoring, we report Tobit 0.162 J/token, Hurdle 0.198 J/token for non-zero requests, and OLS 0.140 J/token on all requests. Across 384 requests we observe a mean of 2.618 J per request and a mean ratio of 0.0215 J per token, while the regression slope captures marginal energy per token. Verification runs in 0.109 ms and signing in 1.066 ms, negligible relative to inference latency.

Keywords: Large Language Models · Environmental Impact Assessment · Cryptographic Attestation · Carbon Footprint

1 Introduction

Large language models require environmental accountability frameworks extending beyond training to inference-level quantification. Current studies focus on training emissions [1], while inference operations remain inadequately characterized [2]. Existing methods lack per-request granularity and suffer trust deficits undermining regulatory compliance [4].

We present a cryptographically attested, per-request environmental accounting framework for LLM inference producing COSE-signed receipts binding energy traces and environmental factors to invocations. Our approach leverages RFC 9711 standards [7] establishing tamper-evident environmental claims for high-risk AI systems.

2 Related Work

Environmental impact assessment focuses on training emissions [1] and lifecycle considerations [2]. Cryptographic attestation for computational claims represents emerging TEE applications [11]. RFC 9711 establishes hardware-bound frameworks [7], while energy measurement benefits from NVML advances [12].

3 Methodology

3.1 Experimental Design

We employed 384 inference requests across four electricity grids (Great Britain, Germany, France, US-CAL-CISO), two temporal conditions (02:00/14:00 UTC), three token limits (64, 128, 256), and four repetitions. Model: TinyLlama/TinyLlama-1.1B-Chat-v1.0. Operational electricity only; embodied manufacturing and end-of-life are out of scope.

3.2 Hardware Monitoring

We sample NVML and RAPL at 10 Hz. GPU attribution uses weighted SM/memory splits; CPU uses per-process counters. RAPL PACKAGE excludes DRAM. We validated the α-weighted attribution against NVML per-process accounting with minimum MAE of 2.8% at $\alpha = 0.5$ [12, 13].

3.3 Baseline and Censoring

We subtract a 30th-percentile idle baseline measured over 60 s. Negative values are clipped to zero, yielding zero-inflation. We fit Tobit and Hurdle models, and include OLS for comparability.

3.4 Environmental Factors

Carbon intensity uses ElectricityMaps at receipt timestamp. We set AWARE CF = 1.0 and WUE = 0.5 L/kWh, reporting sensitivity for WUE in [0.3, 1.2] L/kWh [15, 14]. With CF fixed at 1.0, scarcity-weighted and absolute water coincide; uncertainty is dominated by WUE in the [0.3, 1.2] L/kWh range.

3.5 Cryptographic Attestation

Attestation aligns with EAT (RFC 9711) semantics and uses COSE with Ed25519 signatures [7, **?**]. Each request generates signed receipts with energy measurements, environmental calculations, and SHA-256 provenance hashing.

Table 1: Environmental Impact Summary. Quality denotes receipt-level provenance completeness (timestamped ElectricityMaps payload + receipt signature). During our runs the API returned fallback intensities; we therefore include a timestamp-matched historical correction in Table 2. Live regionality coverage at inference time was limited because of fallback responses.

Metric	Mean	95th %ile	As-measured	Quality
Energy per request (J)	2.618	14.751	2.618	100%
Energy per token (J)	0.0215	0.109	0.0215	100%
CO_2 per request (mg)	0.237	1.447	0.237	100%*
CO_2 per token (mg)	0.00194	0.0122	0.00194	100%*
Water per request ($\times 10^{-7}$ L)	3.64	20.5	3.64	100%
Water per token ($\times 10^{-9}$ L)	2.99	15.2	2.99	100%

Table 2: Carbon Intensity: Fallback vs Historical Correction

Region	Fallback (g/kWh)	Historical (g/kWh)
GB	**66**	**66**
DE	**494**	**672**
FR	**494**	**156**
US-CAL-CISO	**494**	**387**

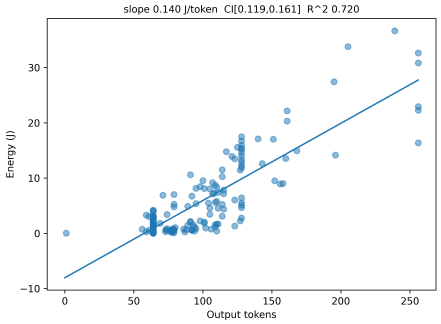
4 Results and Discussion

4.1 Energy Consumption and Scaling

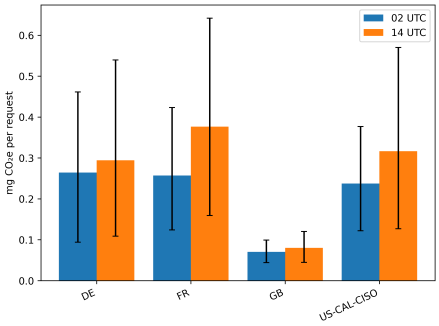
Results revealed mean energy consumption 2.618 J per request. Average energy per token is a ratio over all requests. The regression slope is the marginal increase per additional token. Mean ratio 0.0215 J/token; OLS 0.140 J/token (all requests), Tobit 0.162 J/token (censored at zero), and Hurdle 0.198 J/token (non-zeros). Linear regression explains 72% of variance ($R^2 = 0.720$).

Outlier analysis identified four requests exceeding 99th percentile, all algorithmic tasks consuming 30.8 to 36.6 J, representing 12-14× mean consumption.

AWARE CFs were set to 1.0 in this prototype, so scarcity-weighted and absolute water are identical; WUE dominates uncertainty.

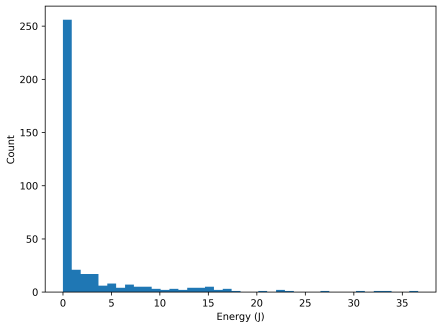


(a) Linear energy-token relationship

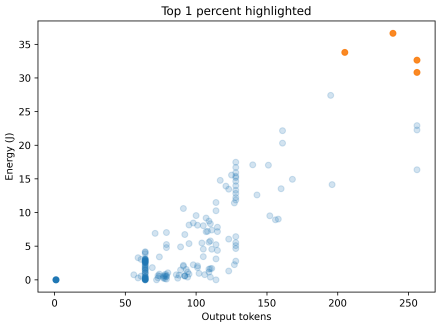


(b) Regional CO_2 variations (fallback intensities)

Fig. 1: Key results: (a) $R^2 = 0.720$, slope 0.140 J/token OLS, and (b) 7.5× regional carbon variation.



(a) Energy distribution histogram



(b) High-energy outliers

Fig. 2: Distribution analysis: (a) right-skewed distribution, (b) algorithmic outliers.

4.2 Regional Impact and Performance

Using fallback intensities, Great Britain at 66 g/kWh vs continental regions at 494 g/kWh implies a 7.5× spread. With timestamp-matched historical data, mean emissions drop by ≈32% and regional rankings change among the continental grids (GB remains lowest while DE/FR/CAISO reorder relative to each other, consistent with Table 2).

Cryptographic performance: Verification 0.109 ms (9,174/s single core). Signing 1.066 ms (850 signatures/s); both under 0.2% of TinyLlama latency. Benchmarked on Xeon Gold 6342 @ 2.80 GHz using PyNaCl.

4.3 Governance Applications

Framework enables regulatory compliance through tamper-evident receipts with hardware-backed key management, geographic optimization for carbon-aware deployment, and algorithmic efficiency monitoring. The 7.5× regional variation demonstrates load balancing opportunities [16].

5 Limitations

Limits include small model scale, grid-data gaps requiring post-hoc correction, WUE assumptions, and RAPL PACKAGE excluding DRAM. Future work includes larger models and external power measurement.

6 Conclusion

We establish cryptographically attested environmental assessment for LLM inference. Key findings: 7.5× regional carbon variation, predictive energy modeling (0.162 J/token Tobit, 0.140 J/token OLS, $R^2 = 0.720$), negligible cryptographic overhead. Framework provides foundations for carbon pricing and regulatory compliance.

7 Appendix

Receipt schema (CBOR): {issuer_kid, ts, nonce, model_id, prompt_hash, tokens_out, energy_series_10Hz, energy_J, grid_region, grid_gCO2_kWh, CO2_mg, WUE_L_kWh, water_L, u95_energy, u95_CO2, u95_water, provenance_sha256, sig}.

The verifier checks signature and issuer key, enforces a monotonic timestamp/nonce to prevent replay, re-integrates the 10 Hz energy series to verify energy_J within sensor error, and recomputes CO_2 and water before accepting the claim.

References

1. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. In: ACL, pp. 3645–3650 (2019)
2. Luccioni, A.S., Viguier, S., Ligozat, A.L.: Estimating the carbon footprint of BLOOM, a 176B parameter language model. JMLR **24**(253), 1–15 (2023)
3. Raza, A., Suh, S., Mehrotra, S., Dutt, N.: How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprints of Large Language Model Inference. arXiv preprint arXiv:2505.09598 (2025)
4. Henderson, P., Hu, J., Romoff, J., et al.: Towards systematic reporting of ML energy and carbon footprints. JMLR **21**(248), 1–43 (2020)
5. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green AI. Commun. ACM **63**(12), 54–63 (2020)

6. Ulissi, S., Dumit, A., Joyce, P.J., et al.: Criteria for Credible AI-assisted Carbon Footprinting Systems. arXiv preprint arXiv:2509.00240 (2024)
7. Mandyam, G., Lundblade, L., Ballesteros, M., O'Donoghue, J.: The Entity Attestation Token (EAT). RFC 9711. IETF (2024)
8. United Nations Environment Programme: Artificial intelligence end-to-end: Environmental impact assessment. Policy Brief (2024)
9. Abdelaal, A., Ismail, R., Kureshi, I.: Machine Learning Algorithms for Predicting Energy Consumption in Educational Buildings. Complexity **2024**, 6812425 (2024)
10. Lang, S., Haneda, M., Steubing, B.: A simplified machine learning product carbon footprint calculation method. Cleaner Environmental Systems **12**, 100254 (2024)
11. Ménétrey, J., Pasin, M., Felber, P., Schiavoni, V.: Attestation mechanisms for TEEs demystified. ACM Comput. Surv. **54**(11s), 1–36 (2022)
12. NVIDIA Corporation: NVIDIA Management Library (NVML) Documentation (2024)
13. ML.Energy: Measuring GPU Energy: Best Practices (2023)
14. Boulay, A.M., Bare, J., Benini, L., et al.: WULCA consensus model for water scarcity footprints (AWARE). Int. J. Life Cycle Assess. **23**(2), 368–378 (2018)
15. ElectricityMaps: API Documentation (2024)
16. Green Software Foundation: Software Carbon Intensity (SCI) Specification (2024)
17. García-Martín, E., Rodrigues, C.F., Riley, G., Grahn, H.: Estimation of energy consumption in machine learning. J. Parallel Distrib. Comput. **134**, 75–88 (2019)
18. Anthony, L.F., Kanding, B., Selvan, R.: Carbontracker: Tracking and predicting carbon footprint of training deep learning models. arXiv:2007.03051 (2020)
19. Dodge, J., Prewitt, T., des Combes, R.T., et al.: Measuring carbon intensity of AI in cloud instances. In: FAccT '22 (2022)
20. International Organization for Standardization: ISO 14040:2006 Environmental management - Life cycle assessment - Principles and framework (2006)
21. Luccioni, A.S., Jernite, Y., Strubell, E.: Power Hungry Processing: Watts driving AI deployment cost? In: FAccT '24 (2024)
22. Lang, J., Rünger, G.: High-resolution power profiling of GPU functions using model-specific registers. In: Euro-Par 2013, pp. 215–226. Springer (2013)
23. Weaver, V.M., Johnson, M., Kasichayanula, K., et al.: Measuring energy and power with PAPI. In: ICPPW '12, pp. 262–268. IEEE (2012)
24. Schaad, J.: CBOR Object Signing and Encryption (COSE). RFC 8152. IETF (2017)