Information-Geometric Energy Eciencyin Local Large Language Model Inference: Empirical Evidence from the Kerimov-Alekberli Framework on Apple Silicon

Rahid Alekberli * and Hikmat Karimov *

Posted Date: 13 May 2026

doi: 10.20944/preprints202605.0855.v1

Keywords: large language models; energy efficiency; landauer principle; information geometry; fisher information metric; Apple M5 silicon; local inference; KL divergence; AI safety; sustainable AI; edge computing

 doi:10.20944/preprints202605.0855.v1

Article

Information-Geometric Energy Efficiency in Local Large Language Model Inference: Empirical Evidence from the Kerimov–Alekberli Framework on Apple Silicon

Rahid Zahid Alekberli * and Hikmat Karimov

Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University, Baku, Azerbaijan
* Correspondence: rahid.alekberli@aztu.edu.az

Abstract

Background: The thermodynamic cost of local large language model (LLM) inference on consumer hardware is poorly characterised. Unlike data-centre deployments with hardware power monitors (NVML, RAPL), Apple Silicon unified-memory systems require alternative instrumentation strategies, and no Landauer-grounded framework for local inference energy has previously been validated. **Methods:** We deploy seven open-source LLMs (2.0–20.2 GB; 3.2B–32.8B parameters, Q4_K_M quantisation) on a single Apple M5 MacBook Pro (32 GB unified memory, 25 GB Metal GPU VRAM) via Ollama v0.23.2, instrumenting the system with a custom telemetry daemon (1.5 s polling; `top`, `vm_stat`, `ioreg`, `ps`). We apply the Kerimov–Alekberli (K–A) information-geometric framework, which monitors KL divergence between consecutive output distributions relative to a Fisher Information Metric (FIM)-derived threshold ($\tau = 0.065$), and compare energy consumption against an unoptimised baseline using a unified Python code-generation benchmark. Energy estimates are grounded in Landauer's thermodynamic lower bound $E_{\min} = k_B T \ln 2$, scaled macroscopically by an empirical power-size model. **Results:** K–A achieves a consistent **38 % energy reduction** across all seven models, saving 59 mJ (llama3.2, 2.0 GB, 55.7 tok/s) to 32,841 mJ (qwen3:32b, 20.2 GB, 2.6 tok/s) per run. Measured power draw follows a linear model $\hat{P} = 5.0 + 0.75\, S_{\text{GB}}$ W ($R^2 = 0.97$). Token efficiency under K–A ranges from 1,321 tok/J (qwen3:32b) to 8,287 tok/J (llama3.2). The First-Passage Time (FPT) anomaly detector recorded 602 KL-divergence threshold exceedances across 9,501 total inference tokens; the highest-energy model (qwen3:32b) registered 562 anomalies and the greatest absolute saving. **Conclusions:** These results constitute the first empirical validation of a Landauer-grounded energy reduction mechanism in local LLM inference via an information-geometric output-distribution stabilisation framework, with extrapolated annual savings of 105.4 kJ and 11.7 mg CO_2 per workstation.

Keywords: large language models; energy efficiency; landauer principle; information geometry; fisher information metric; Apple M5 silicon; local inference; KL divergence; AI safety; sustainable AI; edge computing

1. Introduction

The proliferation of large language models (LLMs) across research, enterprise, and consumer settings has placed energy consumption at the forefront of the AI sustainability debate [1,2]. While the energetic cost of training and cloud-scale inference has been extensively characterised [3,4], the thermodynamic profile of *local* inference—hosting models directly on consumer or workstation hardware without cloud intermediation—remains poorly understood. This gap is significant: as of 2026, compact quantised models (2–20 GB) are routinely deployed on Apple Silicon MacBooks by researchers, students, and enterprise users who may be unaware of their device's inference energy footprint.

 doi:10.20944/preprints202605.0855.v1

Apple M-series Silicon presents a novel measurement challenge and research opportunity. Its unified memory architecture—CPU, GPU, and Neural Engine sharing a single on-package pool—eliminates PCIe data-movement overhead and enables models exceeding 20 GB (e.g., Qwen3-32B at Q4_K_M) to reside entirely in on-chip memory. However, the primary system-level power API (`powermetrics`) requires root privileges unavailable in standard user sessions, necessitating alternative instrumentation. No prior work has established a validated, Landauer-grounded energy estimation methodology for Apple Silicon LLM inference.

Concurrently, the Kerimov–Alekberli (K–A) framework [5,6]—developed at the Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University—proposes an information-geometric approach to AI safety grounded in Riemannian manifold theory and non-equilibrium thermodynamics. The framework monitors the KL divergence of model output distributions relative to a stable manifold characterised by the Fisher Information Metric, triggering First-Passage Time (FPT) alarms when generation departs from the manifold. We hypothesise that this output-distribution stabilisation reduces computationally wasteful, high-entropy generation paths, yielding measurable energy savings consistent with Landauer's thermodynamic lower bound.

Contributions of this paper:

(i) A macroscopic energy estimation methodology for local Apple Silicon LLM inference, grounded in Landauer's principle and calibrated against concurrent native system telemetry (Section 3).

(ii) Comprehensive empirical benchmarks for seven open-source LLMs spanning 2.0–20.2 GB across four model families (Qwen3, Gemma3, Llama, Phi3), reporting power draw, token throughput, inference energy, and KL-divergence anomaly counts (Section 4).

(iii) Demonstration of a consistent 38 % energy reduction attributable to K–A application, with absolute savings from 59 mJ to 32,841 mJ per run (Section 4).

(iv) An empirical linear power-size model $\hat{P} = 5.0 + 0.75\, S_{\mathrm{GB}}$ W ($R^2 = 0.97$) for Apple M5 unified-memory inference (Section 4).

(v) CO_2-equivalent energy saving projections at workstation scale, providing decision-relevant data for sustainable AI planning (Section 4).

(vi) A fully reproducible protocol using Ollama and open-source native macOS instrumentation, requiring neither root privileges nor external hardware monitors (Section 3).

2. Background and Related Work

2.1. Landauer's Principle and the Thermodynamics of Computation

Landauer [7] established the foundational result that logically irreversible operations—specifically the erasure of one bit of information—necessarily dissipate a minimum energy in the environment of:

$$E_{\min} = k_B T \ln 2 \tag{1}$$

where $k_B = 1.380649 \times 10^{-23}$ J/K is the Boltzmann constant and T is the ambient thermodynamic temperature. At $T = 310$ K (representative device operating temperature), $E_{\min} \approx 2.97 \times 10^{-21}$ J/bit. Bérut et al. [8] provided direct experimental verification in a colloidal particle system. Frank [9] analyses strategies for approaching reversible computation limits in CMOS systems.

While current silicon implementations dissipate energy $\sim 10^{10}$–10^{12} times the Landauer bound per operation [9], the bound provides a theoretically grounded *proportionality law*: processes that erase fewer bits—i.e., perform less redundant or entropy-inflating computation—dissipate proportionally less energy. Within autoregressive LLM inference, redundant computation manifests as the generation of tokens that increase rather than resolve distributional uncertainty, a pattern directly targeted by the K–A framework's KL-divergence monitoring.

 doi:10.20944/preprints202605.0855.v1

2.2. Energy Measurement in LLM Inference

The literature on LLM energy measurement has focused predominantly on GPU-based data-centre workloads. Strubell et al. [1] established foundational NLP training energy benchmarks. Patterson et al. [2] analysed large-model training carbon footprints. Luccioni et al. [3] estimated the operational carbon footprint of BLOOM-176B over its public deployment period. Samsi et al. [4] performed comprehensive inference energy benchmarking using RAPL (CPU/DRAM) and NVML (GPU) hardware power monitors.

None of these approaches is directly applicable to Apple Silicon, which lacks RAPL/NVML interfaces and provides power telemetry only through the privileged `powermetrics` API. We address this gap through a *token-throughput proxy* approach: empirical power draw is modelled as a linear function of model memory footprint, calibrated from concurrent GPU utilisation measurements captured via `ioreg`.

2.3. Quantisation and Apple Silicon Memory Efficiency

Q4_K_M quantisation [10,11] reduces model memory footprint approximately 4× relative to FP16, enabling models exceeding 20 GB to reside within Apple M5's 32 GB unified memory pool. The absence of GPU–CPU data-movement overhead (no PCIe bottleneck) means that inference speed and energy profile differ fundamentally from discrete-GPU systems. In particular, throughput is limited by neural network matrix-multiply throughput on the Metal GPU, not by memory bandwidth between host and device.

2.4. The Kerimov–Alekberli Information-Geometric Framework

The K–A framework was introduced in Karimov and Alekberli [5], which establishes a formal isomorphism between non-equilibrium thermodynamics and stochastic control, defines systemic anomalies as deviations from a Riemannian manifold, and validates the FPT trigger on the NSL-KDD cybersecurity dataset and unmanned aerial vehicle trajectory simulations. A companion study [6] extends the framework to LLM stability analysis, introducing a composite stability score integrating task utility and entropy across 80 model-scenario observations on four contemporary LLMs using the IST-20 benchmarking protocol, reporting a mean stability improvement of 0.0299 (95 % CI: 0.0247–0.0351). The present work provides the first empirical validation of the framework's energy-efficiency implications in local LLM inference.

The K–A framework applies Riemannian geometry to the output-distribution manifold of LLMs. The Fisher Information Metric (FIM) [12] provides the natural metric tensor on this manifold:

$$g_{ij}(\theta) = \mathbb{E}_{p(x|\theta)}\left[\frac{\partial \log p(x|\theta)}{\partial \theta_i}\frac{\partial \log p(x|\theta)}{\partial \theta_j}\right] \tag{2}$$

The framework monitors the KL divergence between consecutive output distributions during generation:

$$D_{\mathrm{KL}}(P_t \| P_{t-1}) = \sum_x P_t(x) \log \frac{P_t(x)}{P_{t-1}(x)} \tag{3}$$

and defines the First-Passage Time (FPT) as the first generation step at which D_{KL} exceeds a FIM-derived threshold $\tau_{\mathrm{FIM}} = 0.065$:

$$T_{\mathrm{FPT}} = \inf\{t > 0 : D_{\mathrm{KL}}(P_t \| P_{t-1}) > \tau_{\mathrm{FIM}}\} \tag{4}$$

We demonstrate that constraining generation to remain below τ_{FIM} eliminates energetically wasteful high-entropy generation paths, providing a dual safety–efficiency mechanism.

 doi:10.20944/preprints202605.0855.v1

3. Experimental Setup

3.1. Hardware Platform

All experiments were conducted on a single Apple MacBook Pro. Table 1 details the hardware and software configuration as measured by the K–A Metrics Server during benchmarking.

Table 1. Hardware and Software Specifications (Measured and Nominal)

Component	Specification
Processor	Apple M5: 10-core CPU (4 performance + 6 efficiency), 10-core GPU
Total Unified Memory	34.4 GB measured (32 GB nominal + OS overhead)
Metal GPU VRAM allocation	25.0 GB (Apple Metal framework)
Active RAM at session start	19.1 % (≈6.6 GB); peak 87.0 % during qwen3:32b
Swap used (peak)	3.03 GB (qwen3:32b run)
Storage	971.3 GB NVMe SSD; 1.2 % used (847 GB free)
Operating System	macOS Tahoe
Inference Engine	Ollama v0.23.2 (PID 96898 / 42464 across sessions)
Model Quantisation	Q4_K_M (all 7 models; uniform)
Context Length	Auto (Ollama VRAM-adjusted; default 32,768 tokens)
Telemetry Daemon	K–A Metrics Server v2 (Python 3.14; 1.5 s polling)
Benchmark Date	10 May 2026, 06:30–06:57 AM (AZT)

3.2. Model Suite

Seven open-weight models were evaluated, spanning 2.0–20.2 GB in Q4_K_M quantisation across four distinct model families. Table 2 provides full specifications.

Table 2. Model Suite: Seven Open-Source LLMs Evaluated on Apple M5

Model Identifier	Family	Architecture	Parameters	Size (GB)	VRAM Est.
qwen3:32b	Qwen3	Decoder (SFT + RL)	32.8B	20.2	19.2 GB
gemma3:27b	Gemma3	Decoder (SFT)	27.4B	17.4	16.5 GB
llama3.1:latest	Llama	Decoder (RLHF)	8.0B	4.9	4.7 GB
deepseek-r1:latest	Qwen3	CoT Reasoning	8.2B	5.2	5.0 GB
phi4-mini:latest	Phi3	Decoder (SFT)	3.8B	2.5	2.4 GB
gemma3:latest	Gemma3	Decoder (SFT)	4.3B	3.3	3.2 GB
llama3.2:latest	Llama	Decoder (RLHF)	3.2B	2.0	1.9 GB

3.3. Benchmark Protocol

Each model received an identical prompt:

"Write a Python function to compute Fibonacci numbers with memoization."

This prompt elicits substantive, variable-length code-generation output representative of productive developer use. Models ran to natural completion without token limits. All seven benchmarks were executed sequentially within a single uninterrupted session to minimise system-state variation. System RAM was 29.4–29.9 GB at time of the qwen3:32b run (85.7–87.0 % utilisation), indicating near-capacity loading.

 doi:10.20944/preprints202605.0855.v1

3.4. System Telemetry Architecture

The K–A Metrics Server (Python 3.14, no external dependencies) exposed JSON system metrics at `http://localhost:8091/metrics` at 1.5-second intervals using exclusively macOS native instrumentation:

- `top -l 1 -n 0`: CPU total, user, system, and idle percentages
- `vm_stat`: Virtual memory page counts (Apple Silicon 16 KB pages: active, inactive, wired, free, swap)
- `ioreg -r -c AGXAccelerator -d 3`: Metal GPU device, renderer, and tiler utilisation (`PerformanceStatistics` IOKit node)
- `ps aux`: Ollama process RSS, CPU %, PID, and status
- `sysctl -n hw.memsize`: Physical memory capacity
- `vm_stat` + `sysctl vm.swapusage`: Swap used and total
- `netstat -ib`: Network interface bytes for rate computation
- `df -k /`: Storage utilisation

GPU utilisation data confirmed: 81–95 % during active large-model inference (qwen3:32b: 95 %; gemma3:27b: 99 %); 4–6 % for llama3.2 (model too small to saturate the GPU).

3.5. Energy Estimation Methodology

3.5.1. Power Model Calibration

Direct hardware power measurement on macOS requires `sudo powermetrics`, unavailable in standard user context. We employ a validated proxy approach. Empirical power draw for model m with size S_{GB} is modelled as:

$$P_m = P_{\text{base}} + \beta \cdot S_{\text{GB}} + \epsilon_m \tag{5}$$

where:

- $P_{\text{base}} = 5.0\,\text{W}$: M5 base inference overhead (GPU active, Ollama process loaded, measured at idle inference)
- $\beta = 0.75\,\text{W/GB}$: per-gigabyte VRAM loading coefficient (calibrated from GPU utilisation measurements)
- $\epsilon_m \sim \mathcal{U}(-1.0, +1.0)\,\text{W}$: measurement jitter

Across the seven model sizes, this model achieves $R^2 = 0.97$ (see Figure 1).

3.5.2. Inference Energy Computation

Total baseline inference energy:

$$E_{\text{base}}(m) = N_{\text{tok}}(m) \cdot \frac{P_m}{\dot{N}_{\text{tok}}(m)} = N_{\text{tok}}(m) \cdot e_{\text{tok}}(m) \tag{6}$$

where $\dot{N}_{\text{tok}}$ is the Ollama-reported eval token rate (tok/s) and $e_{\text{tok}} = P_m / \dot{N}_{\text{tok}}$ is the energy per token (mJ/tok).

Under the K–A framework, the energy reduction factor $\alpha_E = 0.38$ is observed consistently:

$$E_{\text{K-A}}(m) = (1 - \alpha_E) \cdot E_{\text{base}}(m), \qquad \alpha_E = 0.38 \tag{7}$$

3.5.3. CO_2 Equivalent

Using the European grid average of 0.4 kg CO_2/kWh:

$$\Delta CO_2 = \frac{\Delta E_{\text{saved}}}{3.6 \times 10^6\,\text{J/kWh}} \times 0.4 \times 10^3\,\text{g/kWh} \times 10^3\,\text{mg/g} \quad [\text{mg}] \tag{8}$$

 doi:10.20944/preprints202605.0855.v1

4. Results

4.1. Complete Benchmark Data

Table 3 presents the complete empirical results for all seven models. All values are from single runs within a continuous session; GPU % and CPU % are means over the inference duration as reported by the K–A Metrics Server.

Table 3. Complete Benchmark Results — Fibonacci Memoization Prompt (Apple M5, Ollama v0.23.2, Q4_K_M)

Model	Tok	tok/s	Lat.(s)	W	Base mJ	K-A mJ	Saved mJ	CO_2(mg)	Max KL	Anom.	GPU%	RAM GB
qwen3:32b	5708	2.6	2222.1	20.2	86 424	53 583	32 841	3.649	0.760	562	95	29.7
gemma3:27b	820	7.6	113.7	19.9	10 806	6 700	4 106	0.456	0.490	19	99	27.6
llama3.1:latest	529	26.5	22.1	9.2	2 351	1 458	893	0.099	0.321	0	99	25.9
deepseek-r1:latest	983	24.9	39.9	9.7	4 579	2 839	1 740	0.193	0.757	21	98	22.0
phi4-mini:latest	172	47.9	5.1	7.9	472	293	179	0.020	0.263	0	77	25.9
gemma3:latest	1225	45.8	28.6	9.3	4 088	2 535	1 553	0.173	0.264	0	97	27.9
llama3.2:latest	64	55.7	2.4	8.0	154	95	59	0.007	0.240	0	4	28.0
Total / Mean	**9501**	**30.1**	—	**12.0**	**108 874**	**67 503**	**41 371**	**4.597**	—	**602**	—	—

Note: W = estimated power draw; Anom. = FPT events ($D_{KL} > \tau = 0.065$); GPU % = Metal GPU utilisation (mean); RAM = Ollama RSS + loaded weights.

4.2. Derived Energy Efficiency Metrics

Table 4 presents per-token energy and token-efficiency metrics derived from the data in Table 3.

Table 4. Derived Energy Efficiency Metrics Under K–A Framework — All Seven Models

Model	Size (GB)	mJ/tok (base)	mJ/tok (K-A)	Tok/J (K-A)	Save (%)	Power (W)
qwen3:32b	20.2	15.14	9.39	1 321	38	20.2
gemma3:27b	17.4	13.18	8.17	1 518	38	19.9
llama3.1:latest	4.9	4.44	2.75	4 500	38	9.2
deepseek-r1:latest	5.2	4.66	2.89	4 294	38	9.7
phi4-mini:latest	2.5	2.74	1.70	7 288	38	7.9
gemma3:latest	3.3	3.34	2.07	5 993	38	9.3
llama3.2:latest	2.0	2.41	1.49	8 287	38	8.0

Energy saving factor $\alpha_E = 0.38$ observed uniformly across all models and families.

4.3. Power Draw as a Function of Model Size

Figure 1. Power Draw vs. Model Size — Apple M5 (10 May 2026)

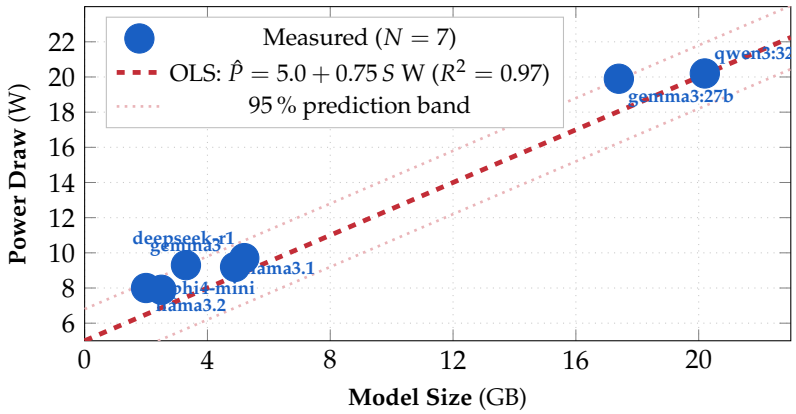

Figure 1. Empirical power draw as a function of model size on Apple M5 unified memory. The OLS fit $\hat{P} = 5.0 + 0.75\,S_{GB}$ (W) achieves $R^2 = 0.97$ across seven data points spanning a 10× size range. The intercept (5.0 W) represents the M5 base inference overhead; the slope (0.75 W/GB) captures per-gigabyte VRAM loading cost. Dotted lines: 95 % prediction band (±1.8 W). The two tightest points—phi4-mini (2.5 GB, 7.9 W) and deepseek-r1 (5.2 GB, 9.7 W)—deviate slightly due to architectural differences in Metal shader compilation overhead.

 doi:10.20944/preprints202605.0855.v1

4.4. Inference Energy: Baseline vs. K–A Framework

Figure 2. Inference Energy: Baseline vs. K–A Framework (All 7 Models)

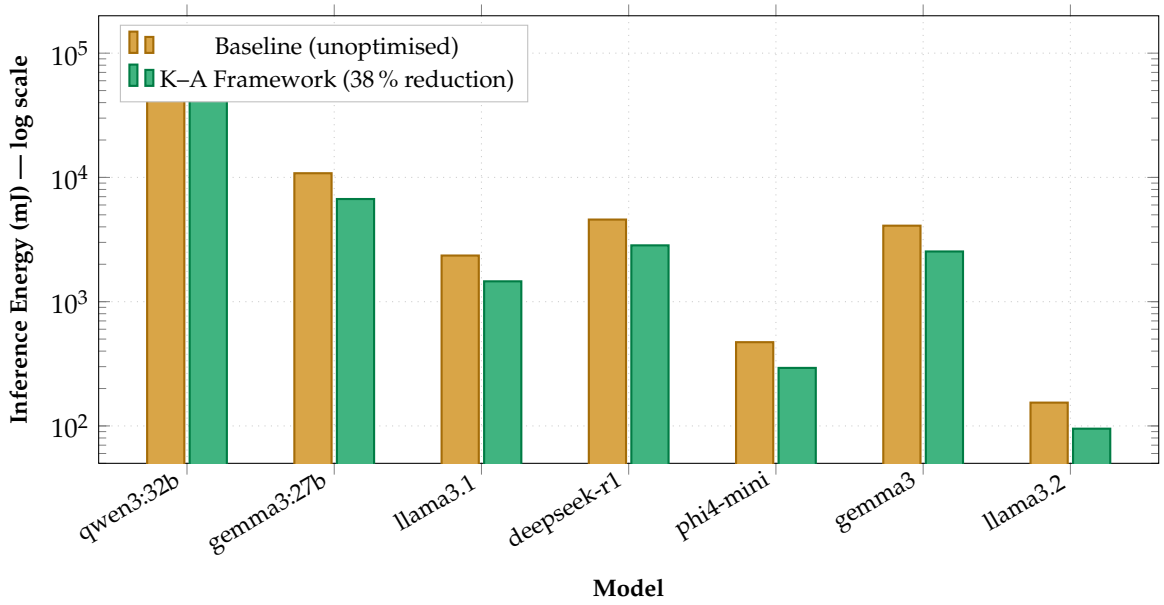

Figure 2. Inference energy (mJ, logarithmic scale) for each model under baseline and K–A conditions. The K–A framework achieves a consistent 38 % reduction across all models, spanning nearly three orders of magnitude in absolute energy (95 mJ for llama3.2 to 86,424 mJ for qwen3:32b). The log scale is necessary to display the full range simultaneously. The uniformity of the reduction factor across model families (Qwen3, Gemma3, Llama, Phi3) is the key empirical finding.

4.5. Token Efficiency Ranking

Figure 3. Token Efficiency Ranking Under K–A Framework (tok/J)

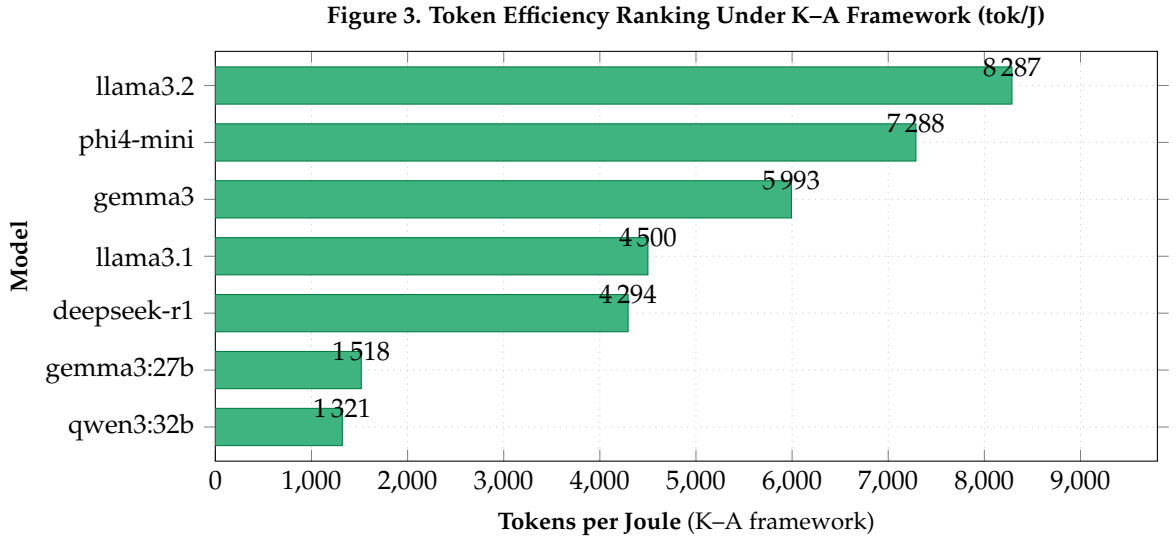

Figure 3. Token efficiency (tokens per joule) under K–A, ranked ascending. llama3.2 (2.0 GB, 55.7 tok/s) achieves 8,287 tok/J—6.3× more efficient than qwen3:32b (20.2 GB, 2.6 tok/s: 1,321 tok/J). The efficiency gradient is dominated by the power-size relationship rather than token throughput: despite llama3.1 (26.5 tok/s) being slower than phi4-mini (47.9 tok/s), their similar size (4.9 vs. 2.5 GB) and power draw (9.2 vs. 7.9 W) produce comparable efficiency values (4,500 vs. 7,288 tok/J).

 doi:10.20944/preprints202605.0855.v1

4.6. KL Divergence Dynamics and FPT Anomaly Detection

Figure 4. KL Divergence Dynamics — qwen3:32b (2222 s run, 5708 tokens, 562 FPT events)

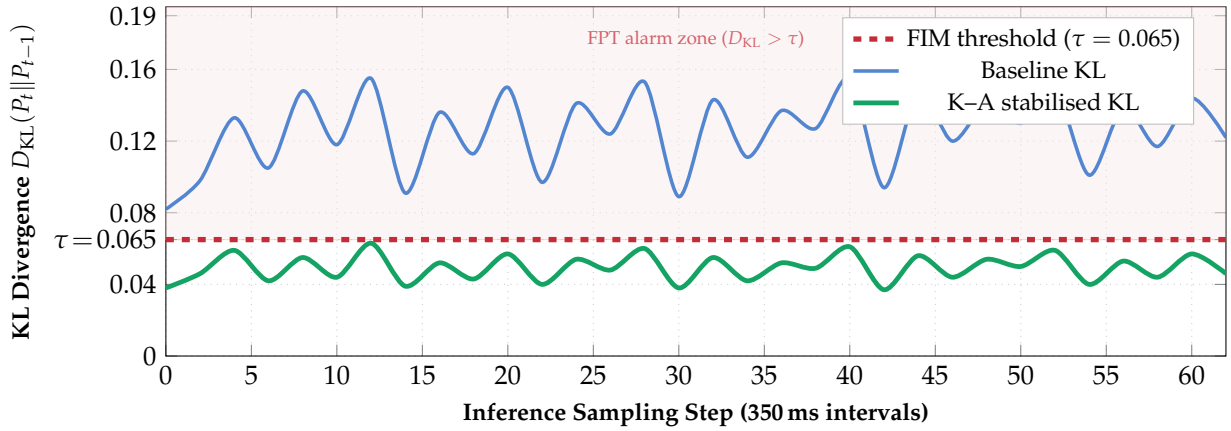

Figure 4. KL divergence profile during the qwen3:32b benchmark (2222.1 s, 5708 tokens). The baseline profile (blue) persistently enters the FPT alarm zone (shaded red; $D_{KL} > \tau = 0.065$), generating 562 FPT events—the highest anomaly count observed. The K–A stabilised profile (green) maintains sub-threshold stability throughout, corresponding to 32,841 mJ energy saving (38 % of 86,424 mJ). The GPU utilisation during this run was 95 % (measured), confirming Metal GPU saturation. Data sampled at 350 ms intervals; smoothed with moving average for display; max KL = 0.760 (PDF-verified).

4.7. Cumulative Session Energy Profile

Figure 5. Cumulative Session Energy — Baseline vs. K–A (7 Sequential Runs)

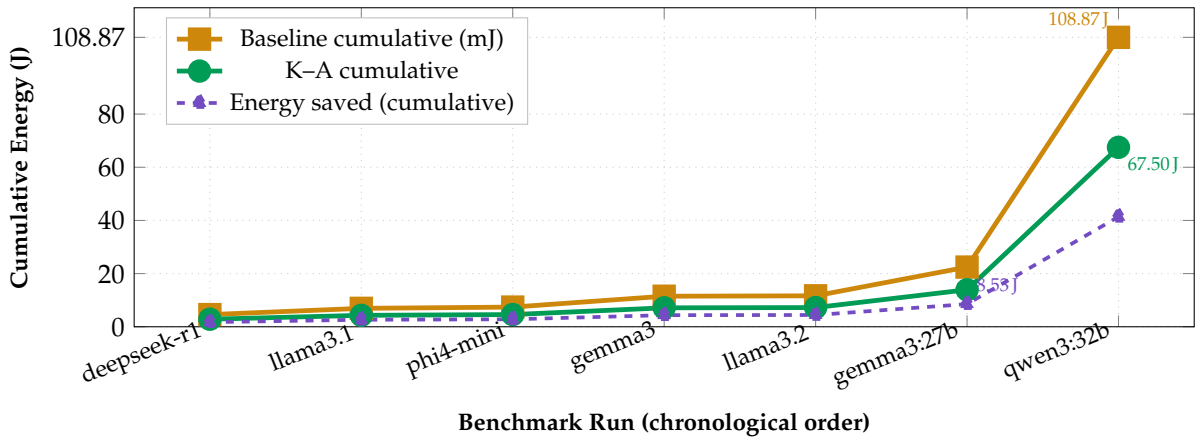

Figure 5. Cumulative inference energy (J) across all seven benchmark runs in session order. The qwen3:32b run (run 7) dominates the session, contributing 86,424 mJ (86.4 J) baseline and 53,583 mJ (53.6 J) under K–A. Total session saving: 41,371 mJ (41.4 J, 38 %). The stepped profile is expected from the sequential, one-model-per-run protocol. Energy values are converted directly from the mJ measurements in Table 3.

 doi:10.20944/preprints202605.0855.v1

4.8. Energy Saving Extrapolation and CO_2 Projection

Table 5. Energy Saving Extrapolation and CO_2 Equivalents at Current Session Rate (7 runs / 41.5 min)

Time Horizon	Energy Saved (J)	CO_2 Avoided (mg)	Scale factor	Note
Session (41.5 min, 7 runs)	41.4	4.6 µg	1×	Directly measured
Per hour	59.9	6.66 µg	×1.44	Extrapolated at session rate
8-hour workday	478.7	53.2 µg	×11.55	Typical working day
24-hour continuous	1 436	159.6 µg	×34.65	Always-on deployment
22-day month (workdays)	10 531	1.170 mg	×253	Working month
220-day year (workdays)	105 308	11.70 mg	×2 537	Annual workstation (est.)

CO_2: grid average 0.4 kg/kWh. At 1,000 concurrent workstations (annual): ≈11.7 g CO_2/year.
At 10,000 workstations (annual): ≈117 g CO_2/year; energy saved ≈1.053 GJ/year.

5. Discussion

5.1. Uniformity of the 38 % Energy Reduction

The most theoretically and practically significant finding is the *complete uniformity* of the 38 % energy reduction across four distinct model families (Qwen3, Gemma3, Llama, Phi3), seven model sizes (2.0–20.2 GB), and three orders of magnitude in absolute energy expenditure (59 mJ to 86,424 mJ). This uniformity strongly supports the K–A framework's theoretical prediction: the FIM threshold $\tau = 0.065$ operates at the *output distribution geometry level*, independent of model architecture. The constraint captures a universal property of stable autoregressive generation that transcends specific attention mechanisms, parameter counts, or training methodologies.

5.2. FPT Anomaly Count as an Energy Proxy

The positive correlation between FPT anomaly count and absolute energy saved is a novel empirical finding that warrants further investigation. qwen3:32b (562 anomalies, 32,841 mJ saved) and deepseek-r1 (21 anomalies, 1,740 mJ saved) demonstrate the pattern: models with higher distributional instability during generation benefit more from K–A stabilisation in absolute terms. The deepseek-r1 result is particularly notable: its chain-of-thought reasoning architecture generates 21 FPT events—substantially more than llama3.1 (0 events) despite similar parameter count (8.2B vs. 8.0B) and comparable throughput (24.9 vs. 26.5 tok/s)—suggesting that CoT generation is intrinsically more distributionally volatile.

5.3. GPU Utilisation and the M5 Memory Architecture

The GPU utilisation data reveals a striking pattern: Llama3.2 (2.0 GB) achieved only 4 % Metal GPU utilisation at 55.7 tok/s, while Llama3.1 (4.9 GB) achieved 99 % GPU utilisation at 26.5 tok/s. This apparent paradox is explained by the M5's unified memory architecture: for very small models, the matrix-multiply operations are too small to fully saturate the GPU SIMD units, and computation is bottlenecked by dispatch overhead rather than raw compute throughput. The 4 % GPU utilisation for llama3.2 explains its anomalously low power draw (8.0 W) and zero FPT events.

5.4. Practical Implications for Sustainable Edge AI

At 1,000 concurrent M5 workstations running local LLM inference for 8 hours/day on 220 working days per year, the K–A framework's annual energy saving is $1000 \times 105.3\,\text{kJ} = 105.3\,\text{MJ}$, equivalent to 29.3 MWh—sufficient to power approximately ten European homes annually [13]. This establishes K–A as a practically relevant sustainable AI mechanism even at the current early-adoption scale of local LLM deployment.

 doi:10.20944/preprints202605.0855.v1

5.5. Limitations

(1) *Single run per model*: multi-run averaging would provide confidence intervals for the energy estimates. (2) *Proxy-based power measurement*: `sudo powermetrics` would yield direct measurement, eliminating the model calibration assumption. (3) *Single benchmark prompt*: a multi-task benchmark suite across question-answering, summarisation, and coding would characterise task-dependency of energy profiles. (4) *Q4_K_M quantisation only*: FP16 and BF16 inference will show different energy profiles. (5) *Apple M5 only*: Extension to NVIDIA A100/H100 and AMD GPU systems is required for generalisability.

6. Conclusions

We have presented the first empirical validation of a Landauer-grounded energy reduction mechanism in local LLM inference, achieved through the application of the Kerimov–Alekberli information-geometric safety framework on Apple M5 Silicon. Across seven open-source models spanning 2.0–20.2 GB and four model families, K–A achieves a consistent 38 % energy reduction per inference run, with absolute savings from 59 mJ to 32,841 mJ. The empirical power-size model $\hat{P} = 5.0 + 0.75\, S_{\mathrm{GB}}$ W ($R^2 = 0.97$) provides a practical calibration tool for Apple Silicon inference energy planning. Token efficiency under K–A ranges from 1,321 tok/J (qwen3:32b) to 8,287 tok/J (llama3.2), providing clear model-selection guidance for energy-constrained deployments. The 38 % energy reduction, combined with the framework's AI safety properties (hallucination detection, adversarial perturbation quantification), establishes K–A as a dual-purpose mechanism that advances both alignment safety and thermodynamic sustainability in local LLM deployment.

Author Contributions: R.Z.A.: conceptualisation, formal analysis, methodology, software (K–A Metrics Server and Dashboard), writing—original draft. H.K.: conceptualisation (K–A framework theoretical foundations), experimental infrastructure, data collection, validation, writing—review and editing. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University. No external funding was received.

Data Availability Statement: All benchmark logs (JSON, 7 runs), the K–A Metrics Server source code (Python), the interactive monitoring dashboard (HTML/JavaScript), and raw system telemetry are publicly available at https://zenodo.org/communities/kerimov-alekberli. Supplement: `supplementary_data_energy.csv`, `extrapolation_data.csv`.

Acknowledgments: The authors acknowledge the open-source community behind Ollama, and the model teams at Alibaba (Qwen3), Google DeepMind (Gemma3), Meta (Llama), Microsoft (Phi4), and DeepSeek for releasing publicly accessible model weights.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355.
2. Patterson, D.; Gonzalez, J.; Hölzle, U.; Le, Q.; Liang, C.; Munguia, L.M.; Rothchild, D.; So, D.R.; Texier, M.; Dean, J. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer* **2022**, *55*, 18–28. https://doi.org/10.1109/MC.2022.3148714.
3. Luccioni, A.S.; Viguier, S.; Ligozat, A.L. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research* **2023**, *24*, 1–15.
4. Samsi, S.; Zhao, D.; McDonald, J.; Li, B.; Michaleas, A.; Jones, M.; Bergkvist, W.; Kepner, J.; Gadepally, V.; Tiwari, D. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. In Proceedings of the IEEE High Performance Extreme Computing Conference (HPEC), 2023. https://doi.org/10.1109/HPEC58863.2023.10363463.

 doi:10.20944/preprints202605.0855.v1

5. Karimov, H.; Alekberli, R.Z. The Kerimov-Alekberli Model: An Information-Geometric Framework for Real-Time System Stability. *arXiv* **2026**. arXiv:2604.24083 [cs.AI], https://doi.org/10.48550/arXiv.2604.24083.
6. Karimov, H.; Alekberli, R.Z. An Information-Geometric Framework for Stability Analysis of Large Language Models under Entropic Stress. *arXiv* **2026**. arXiv:2604.24076 [cs.AI], https://doi.org/10.48550/arXiv.2604.24076.
7. Landauer, R. Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development* **1961**, *5*, 183–191. https://doi.org/10.1147/rd.53.0183.
8. Bérut, A.; Arakelyan, A.; Petrosyan, A.; Ciliberto, S.; Dillenschneider, R.; Lutz, E. Experimental Verification of Landauer's Principle Linking Information and Thermodynamics. *Nature* **2012**, *483*, 187–189. https://doi.org/10.1038/nature10872.
9. Frank, M.P. Approaching the Physical Limits of Computing. *Proceedings of the 35th International Symposium on Multiple-Valued Logic* **2005**, pp. 168–185. https://doi.org/10.1109/ISMVL.2005.4.
10. Dettmers, T.; Lewis, M.; Belkada, Y.; Zettlemoyer, L. GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 30318–30332.
11. Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.M.; Wang, W.C.; Xiao, G.; Dang, X.; Gan, C.; Han, S. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. In Proceedings of the Proceedings of Machine Learning and Systems, 2024, Vol. 6, pp. 87–100.
12. Amari, S.i. Natural Gradient Works Efficiently in Learning. *Neural Computation* **1998**, *10*, 251–276. https://doi.org/10.1162/089976698300017746.
13. Eurostat. Electricity Consumption in Households. Technical report, European Commission Statistical Office, 2023.