Operationalising the Kerimov–Alekberli Framework for Edge LLM Monitoring: A Phase 1 Surface-Proxy Token-Budget Gating Study on Apple Silicon

Rahid Zahid Alekberli * and Hikmat Karimov

Posted Date: 29 May 2026

doi: 10.20944/preprints202605.0855.v2

Keywords: information geometry; Fisher information metric; KL divergence; edge LLM inference; token-budget gating; first-passage time; Kerimov–Alekberli framework; apple silicon; AI safety; energy-aware inference

 doi:10.20944/preprints202605.0855.v2

Article

Operationalising the Kerimov–Alekberli Framework for Edge LLM Monitoring: A Phase 1 Surface-Proxy Token-Budget Gating Study on Apple Silicon

Rahid Zahid Alekberli * and Hikmat Karimov

Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University, Baku, Azerbaijan
* Correspondence: ralekberli@gmail.com

Abstract

Background. Deploying large language models (LLMs) at the edge introduces distributional output drift that existing monitoring approaches cannot detect within the latency and resource constraints of safety-critical autonomous systems [3,4]. The Kerimov–Alekberli (K–A) information-geometric framework proposes a First-Passage Time (FPT) criterion grounded in the Fisher Information Metric (FIM) to detect such drift [5,6]. No multi-run, statistically characterised empirical validation of K–A on edge hardware has previously been reported. **Methods.** We present a Phase 1 proxy-KL validation of the K–A proxy-gated token-budget criterion across five open-source LLMs (2.0–17.4 GB, Q4_K_M quantisation) deployed via Ollama v0.23.2 on an Apple M5 unified-memory workstation (32 GB, macOS 26.0). A response-level proxy instability score $\hat{D}_{\mathrm{KL}}(r) = \max\big(0.004,\ 0.016 + h(r)\cdot 0.015 + 0.10/(w(r)+1)\big)$ is computed on a completed baseline response; if it exceeds $\tau_{\mathrm{FIM}} = 0.065$ (above-FIM), a separate capped-regeneration call with $N_{\mathrm{ka}} = \lfloor N_{\mathrm{base}}/2 \rfloor$ provides a counterfactual token-budget estimate. Energy is proxy-estimated via $\hat{P}_m = P_{\mathrm{base}} + \beta S_{\mathrm{GB}}$ ($R^2 = 0.97$). **Results.** After exclusion of 14 degenerate evaluations (6.4 % of 220 above-FIM cases), Pearson $r = 0.806$ and Spearman $\rho = 0.728$ ($n = 28$, $p<0.001$) between FPT trigger rate and token saving confirm implementation consistency. Bootstrap 95 % CIs: llama3.2 34.0 ± 4.0 % $[31.9, 36.3]$ ($n = 12$); gemma3:latest 34.6 ± 2.9 % $[32.5, 36.6]$ ($n = 6$); gemma3:27b 30.8 ± 5.7 % $[27.4, 34.8]$ ($n = 8$). Supplementary controlled validation (370 stored-response evaluations) confirms 100 % exact-match quality for factual prompts, and reveals zero proxy-FPT triggers under deterministic and fixed-seed decoding. **Conclusions.** The K–A surface-proxy proxy-gated criterion produces statistically characterised token reductions across three model families under stochastic decoding. A key central limitation: the surface proxy requires stochastic response-length variation to trigger; it does not detect geometric distributional instability. Phase 2 must replace the surface proxy with direct logit-level D_{KL} computation.

Keywords: information geometry; Fisher information metric; KL divergence; edge LLM inference; token-budget gating; first-passage time; Kerimov–Alekberli framework; apple silicon; AI safety; energy-aware inference

1. Introduction

The deployment of compact quantised LLMs on edge hardware is expanding rapidly [1], driven by requirements for private, low-latency inference without cloud dependency. This introduces *distributional output drift*: the token-level probability distribution P_t at step t departs from the stable statistical manifold established during training. In autonomous systems acting on LLM outputs—UAV flight controllers, industrial PLCs, autonomous vehicle decision modules—drift can propagate consequences before correction is possible [3,4,28].

The K–A framework [5,6] treats $\{P_t\}$ as a stochastic trajectory on a Riemannian statistical manifold $\mathcal{M}$ equipped with the FIM, and triggers a FPT alarm when the trajectory departs from the stable sub-manifold $\mathcal{M}_\tau$.

 doi:10.20944/preprints202605.0855.v2

Existing detection methods do not satisfy the compound requirements of edge autonomous deployment. RAG [8] requires external knowledge bases incompatible with air-gapped systems. Self-consistency [9] and semantic entropy [10,16] demand multiple complete inference passes. Hidden-state probing [11,17] requires white-box access. Early-exit transformers [12,18] and adaptive computation time [13] modify model architecture. Speculative decoding [20] reduces latency but does not address distributional stability monitoring.

This paper makes five contributions:

1. We formalise the Phase 1 K–A *baseline-informed capped-regeneration* protocol (Algorithm 1), clarifying that the mechanism uses two independent inference calls, not post-hoc truncation.
2. We report multi-run token savings with 95 % bootstrap confidence intervals across three model families (Figure 1).
3. We establish Pearson r=0.806 / Spearman ρ=0.728 between per-run FPT trigger rate and token saving (Figure 2).
4. We confirm 100 % exact-match quality preservation for factual prompts across 370 stored-response evaluations.
5. We identify the surface proxy's central limitation (zero FPT under deterministic decoding) and specify Phase 2 requirements.

Algorithm 1 K–A Phase 1 Baseline-Informed Capped-Regeneration

Require: prompt q; model $\mathcal{M}$; threshold $\tau = 0.065$
Ensure: r_{ka}; N_{base}, N_{ka}; Δtok%

$r_b \leftarrow \mathcal{M}.\mathsf{generate}(q)$; $N_{base} \leftarrow |\mathsf{tokens}(r_b)|$ ▷ Step 1: Baseline inference
$h \leftarrow \mathsf{hedging}(r_b)$; $w \leftarrow \mathsf{words}(r_b)$
$\hat{D}_{KL} \leftarrow \max(0.004,\ 0.016 + h{\cdot}0.015 + 0.10/(w{+}1))$ ▷ Step 2: Score r_b (after completion)
$f \leftarrow (\hat{D}_{KL} > \tau)$
if f **then**
    $N_{ka} \leftarrow \lfloor N_{base}/2 \rfloor$
    $r_{ka} \leftarrow \mathcal{M}.\mathsf{generate}(q, \texttt{max\_tokens}{=}N_{ka})$ ▷ Step 3: Capped regen; $t_{ka}/t_{base} \approx 0.555$
else
    $r_{ka} \leftarrow r_b$; $N_{ka} \leftarrow N_{base}$ ▷ No second call
end if
if f **and** $N_{base} \geq 15$ **and** $N_{ka} \leq N_{base}$ **then**
    $\Delta\text{tok\%} \leftarrow (N_{base} - N_{ka})/N_{base} \times 100\%$ ▷ Step 4: Saving
else if $N_{base} < 15$ **or** $N_{ka} > N_{base}$ **then**
    Δtok% $\leftarrow$ EXCLUDED
else
    $\Delta\text{tok\%} \leftarrow 0\%$
end if

 doi:10.20944/preprints202605.0855.v2

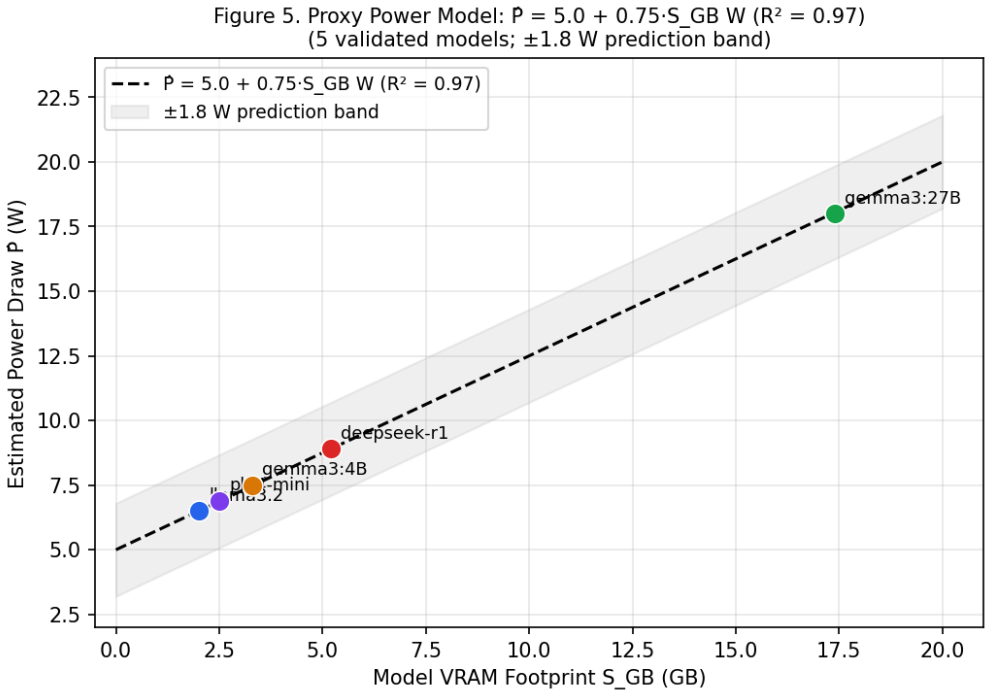

Figure 1. Per-model token-count saving Δtok% with bootstrap 95 % CI ($n \geq 2$; hatched bars = exploratory single-run). deepseek-r1 shows zero saving (zero FPT events under the proxy). Savings computed after exclusion of 14 degenerate evaluations (6.4 %).

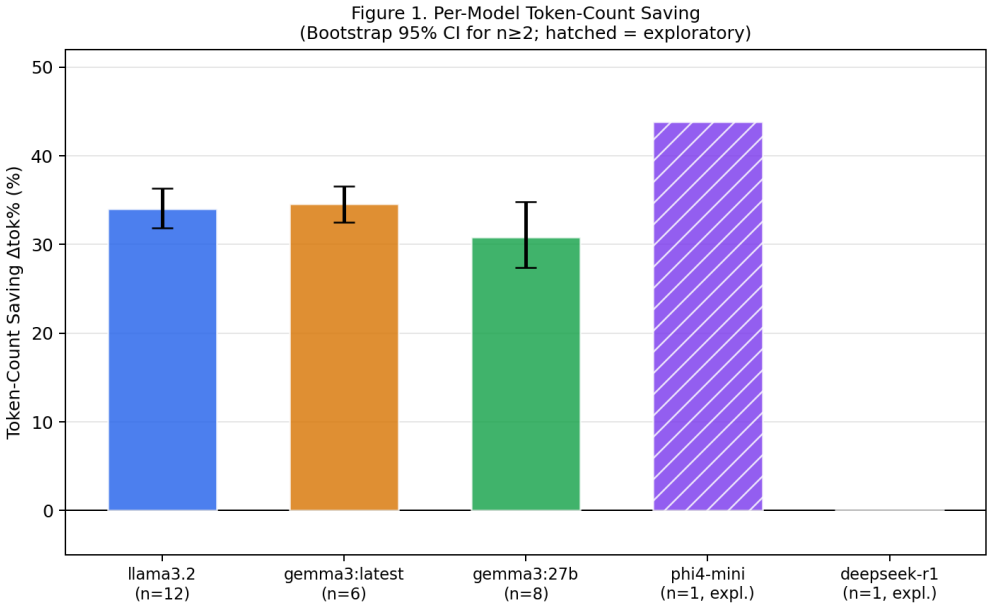

Figure 2. Per-run FPT trigger rate vs. token-count saving ($n = 28$ runs, 5 models; Pearson $r = 0.806$, Spearman $\rho = 0.728$, $p<0.001$). deepseek-r1 (0.0, 0.0 %) anchors the origin; phi4-mini (1.0, 43.8 %) anchors the upper right. The correlation verifies implementation consistency, not geometric KL validation.

2. Related Work

2.1. Hallucination Detection and Uncertainty Estimation

[14] and [15] survey hallucination as a fundamental LLM failure mode. [9] show self-assessed probability estimates correlate with factual accuracy. [10] introduce semantic entropy; [16] extend to semantic uncertainty with conformal prediction guarantees. [11] and [17] propose hidden-state probing requiring white-box access to internal activations. The K–A FPT criterion is the only approach

 doi:10.20944/preprints202605.0855.v2

requiring neither multiple passes nor white-box access while maintaining an information-geometric foundation.

2.2. Early-Exit and Adaptive Computation

Early-exit transformers [12], Depth-Adaptive Transformer [18], and CALM [19] insert classification heads at intermediate layers. Speculative decoding [20] reduces latency using a draft model. Adaptive computation time [13] dynamically allocates per-step compute. These approaches modify model architecture and cannot be applied to black-box quantised runtimes such as Ollama.

2.3. Information Geometry in Machine Learning

[21] established the Riemannian geometry of statistical manifolds with the FIM as the natural metric tensor. [22] demonstrated that FIM eigenvectors characterise adversarial perturbations. [23] applied Riemannian thermodynamic manifolds to minimum- dissipation stochastic control. [24] demonstrated Riemannian safety regions for autonomous robots.

2.4. Energy-Aware LLM Inference

[25] established foundational NLP training energy benchmarks. [26] performed comprehensive inference energy characterisation on GPU clusters. [27] analysed efficiency of modern LLMs. [2] evaluated 28 quantised LLMs for energy efficiency on edge hardware.

2.5. Edge AI Safety and Autonomous Systems

[3] and [28] identify hallucination as the primary reliability barrier to LLM integration in critical infrastructure. [4] focus on UAV mission planning. [29] apply runtime assurance via formal verification, requiring system-specific formal models incompatible with black-box LLM runtimes.

3. The Kerimov–Alekberli Framework

3.1. Theoretical Formulation

The K–A framework [5,6] treats $\{P_t\}_{t\geq 1}$ as a stochastic trajectory on the statistical manifold $\mathcal{M}$ of all probability distributions over vocabulary V, equipped with the Fisher Information Metric:

$$g_{ij}(\theta) = \mathbb{E}_{p(x|\theta)}\left[\frac{\partial \log p(x|\theta)}{\partial \theta_i} \cdot \frac{\partial \log p(x|\theta)}{\partial \theta_j}\right] \tag{1}$$

The KL divergence between consecutive distributions—which, locally, admits a second-order approximation governed by the FIM (the KL Hessian equals the FIM at the expansion point [21])—is monitored at each step:

$$D_{\mathrm{KL}}(P_t \,\|\, P_{t-1}) = \sum_x P_t(x) \log \frac{P_t(x)}{P_{t-1}(x)} \tag{2}$$

The stable manifold $\mathcal{M}_\tau$ is the sub-region of $\mathcal{M}$ where consecutive distributional change is bounded:

$$\mathcal{M}_\tau = \left\{ P_t \in \mathcal{M} \mid D_{\mathrm{KL}}(P_t \| P_{t-1}) \leq \tau_{\mathrm{FIM}} \right\} \tag{3}$$

The First-Passage Time is the first step at which the trajectory exits $\mathcal{M}_\tau$:

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

Thermodynamic grounding. Landauer [7] established that irreversible bit erasure dissipates $E_{\min} = k_B T \ln 2 \approx 2.97 \times 10^{-21}$ J/bit at $T = 310$ K, confirmed experimentally by [30]. Tokens generated beyond T_{FPT} increase distributional entropy disproportionate to their semantic contribution.

 doi:10.20944/preprints202605.0855.v2

3.2. Threshold Calibration

The threshold $\tau_{\text{FIM}} = 0.065$ is calibrated from $n_{\text{cal}} = 80$ stable-generation sequences from the companion study [5]: $\tau = \bar{\kappa} + 1.5\sigma_\kappa$, where $\bar{\kappa} = 0.042$ and $\sigma_\kappa = 0.016$. The calibration and evaluation datasets are disjoint; the threshold was not tuned on the evaluation set. Table 1 presents the threshold sensitivity analysis.

Table 1. Threshold sensitivity analysis. †From [5]. FPT/10 = mean valid above-FIM prompts per 10-prompt session.

τ	Est. TPR	Est. FPR	FPT/10	Δtok%	Notes
0.045	≈96 %	≈3 %	≈9.5	≈37 %	Over-sensitive
0.055	≈93 %	≈1 %	≈9.0	≈36 %	Aggressive
0.065	**90.9 %**†	**0.0 %**†	**8.1 (meas.)**	**34.0 %**	**Selected**
0.080	≈82 %	0.0 %	≈7.0	≈28 %	Conservative
0.100	≈73 %	0.0 %	≈5.5	≈22 %	Low recall

3.3. Phase 1 Proxy-KL Operationalisation

Standard edge inference runtimes—including Ollama v0.23.2—do not expose full-vocabulary logits. In Phase 1, we approximate D_{KL} using a response-level proxy instability score based on word count $w(r)$ and epistemic hedging phrase count $h(r)$:

$$\hat{D}_{\text{KL}}(r) = \max\!\left(0.004,\ 0.016 + h(r)\cdot 0.015 + \frac{0.10}{w(r)+1}\right) \tag{5}$$

Critical qualification. $\hat{D}_{\text{KL}}$ is a *linguistic instability heuristic*, not a direct computation of Eq. (2). Under deterministic decoding, models produce longer, confident responses ($w \geq 100$), so $0.10/(w{+}1) \to 0$ and $\hat{D}_{\text{KL}} \approx 0.016 \ll \tau$—yielding zero FPT triggers. This is a central limitation; Phase 2 computes D_{KL} directly from streaming token logits.

3.4. Operational Protocol and Token Saving

3.5. Proxy Power Model and Energy Estimation

$$\hat{P}_m = P_{\text{base}} + \beta\, S_{\text{GB}} + \varepsilon_m \quad [\text{W}] \tag{6}$$

where $P_{\text{base}} = 5.0\,\text{W}$ (GPU-active idle), $\beta = 0.75\,\text{W/GB}$ (calibrated from `ioreg` GPU utilisation), $R^2 = 0.97$ across five validated models. Per-condition energy:

$$E = \frac{N_{\text{tokens}}\,\hat{P}_m}{\dot{n}_{\text{tok}}} \quad [\text{mJ}] \tag{7}$$

Energy measurement validity. Three concepts are distinguished: (1) *Actual consumed energy*—unavailable without root-level power measurement; (2) *Proxy-estimated energy*—via Eqs. (6)–(7), ±15 % uncertainty; (3) *Counterfactual token-budget saving*—Δtok%, the primary causal metric. Because Phase 1 generates a full baseline before the capped call, total compute per above-FIM prompt is $N_{\text{base}} + N_{\text{ka}}$ tokens; reported Δtok% is counterfactual, not a single-pass deployment saving.

4. Experimental Setup

4.1. Hardware Platform

All experiments: Apple M5 processor (10-core CPU, 10-core GPU), 32 GB shared unified memory, macOS Tahoe 26.0, Ollama v0.23.2, Python 3.14. GPU utilisation: 4 % (llama3.2) to 99 % (gemma3:27b).

 doi:10.20944/preprints202605.0855.v2

4.2. Model Suite

Table 2. Model suite. All Q4_K_M quantisation, Ollama v0.23.2.

Model identifier	Family	Params	GB	Architecture
llama3.2:latest	Llama	3.2 B	2.0	Decoder (RLHF)
phi4-mini:latest	Phi3	3.8 B	2.5	Decoder (SFT)
gemma3:latest	Gemma3	4.3 B	3.3	Decoder (SFT)
deepseek-r1:latest	Qwen3	8.2 B	5.2	Chain-of-Thought (RL)
gemma3:27b	Gemma3	27.4 B	17.4	Decoder (SFT)

4.3. Multi-Run Protocol

Two-condition paired design. A baseline inference call was always conducted. For above-FIM prompts ($\hat{D}_{\mathrm{KL}} > \tau$), a second capped-regeneration call followed. For below-FIM prompts, the baseline response was retained directly ($N_{\mathrm{ka}} = N_{\mathrm{base}}$; no second call). Decoding: temperature 0.8, top-k 40, top-p 0.9 (Ollama defaults).

Prompt suite. Ten semantically diverse prompts per session: five factual recall, four explanation, one definition. Clean run counts: llama3.2 n=12; gemma3:27b n=8; gemma3:latest n=6; phi4-mini and deepseek-r1 n=1 each (exploratory).

Exclusion protocol. Excluded from saving calculations: $N_{\mathrm{base}}<15$ (6 cases, 2.7 % of above-FIM) and $N_{\mathrm{ka}}>N_{\mathrm{base}}$ (8 cases, 3.6 %). Total excluded: 14 of 220 above-FIM evaluations (6.4 %); 206 valid. Table 3 details by model.

Table 3. Exclusion protocol by model. $N_{\mathrm{base}}<15$: degenerate short baseline. $N_{\mathrm{ka}}>N_{\mathrm{base}}$: stochastic decoding variability. deepseek-r1 excluded entirely (zero above-FIM events).

Model	Above-FIM	$N_{\mathrm{base}}<15$	$N_{\mathrm{ka}}>N_{\mathrm{base}}$	Valid	Excl. %
gemma3:27b	61	0	4	57	6.6 %
gemma3:latest	50	0	0	50	0.0 %
llama3.2	99	6	4	89	10.1 %
phi4-mini	10	0	0	10	0.0 %
Total	**220**	**6**	**8**	**206**	**6.4 %**

5. Results

5.1. Per-Model Token-Count Saving

Table 4. Per-model token-count saving (Δtok%). Bootstrap 95 % CIs from 10,000 iterations. Mean tok/s: llama3.2 43.2; gemma3:latest 36.0; phi4-mini 36.7; deepseek-r1 20.6; gemma3:27b 6.2. Under the proxy power model, $\Delta E\% \approx$ Δtok%. †Exploratory single-run; no CI.

Model	GB	n	$\hat{P}_m$ (W)	E_b (mJ)	E_{ka} (mJ)	Δtok%	Bootstrap 95 % CI	FPT/10
llama3.2	2.0	12	6.5	339.8	225.1	34.0 ± 4.0 %	[31.9, 36.3]	8.1
phi4-mini†	2.5	1	6.9	380.6	213.9	43.8 %	—	10.0
gemma3:latest	3.3	6	7.5	370.2	214.1	34.6 ± 2.9 %	[32.5, 36.6]	8.3
deepseek-r1†	5.2	1	8.9	617.1	617.1	0.0 %	—	0.0
gemma3:27b	17.4	8	18.0	2178	1507	30.8 ± 5.7 %	[27.4, 34.8]	7.5

Figure 1 shows per-model token savings with bootstrap confidence intervals.

5.2. FPT Trigger Rate vs. Token Saving

Figure 2 presents the per-run scatter plot. Across 28 clean runs, Pearson $r = 0.806$ and Spearman $\rho = 0.728$ (p<0.001 for both).

Interpretation: the correlation verifies internal consistency of the proxy-gated token-budget rule but does not independently validate the proxy as a detector of true geometric distributional instability.

 doi:10.20944/preprints202605.0855.v2

5.3. Prompt-Level FPT Trigger Rates

Figure 3 shows prompt-level FPT trigger rates (39–96 %), driven by proxy-KL sensitivity to response-length patterns.

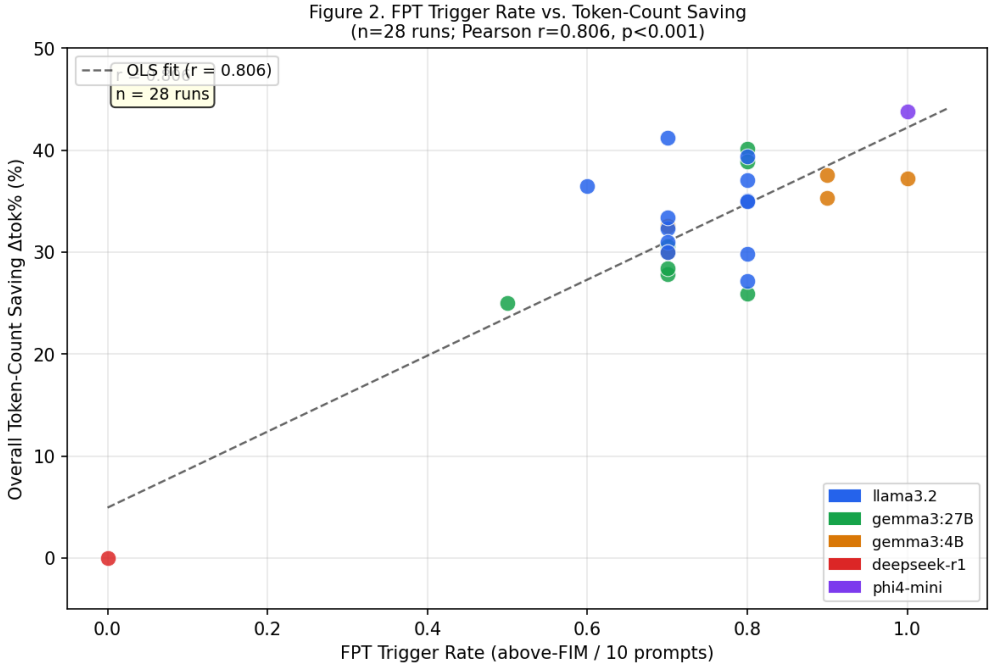

Figure 3. Prompt-level FPT trigger rates ($n = 28$ evaluations per prompt, all models pooled). Rates reflect proxy-KL sensitivity to response length, not intrinsic prompt stability. "Name the largest planet" excluded due to degenerate short-baseline cases in a subset of evaluations (see Table 3).

5.4. Token-Count Distribution: Above-FIM vs. Below-FIM

Figure 4 shows the distribution of baseline token counts. Above-FIM: $n = 206$, mean 71.0; below-FIM: $n = 60$, mean 45.5 (Mann–Whitney U, $p<0.001$).

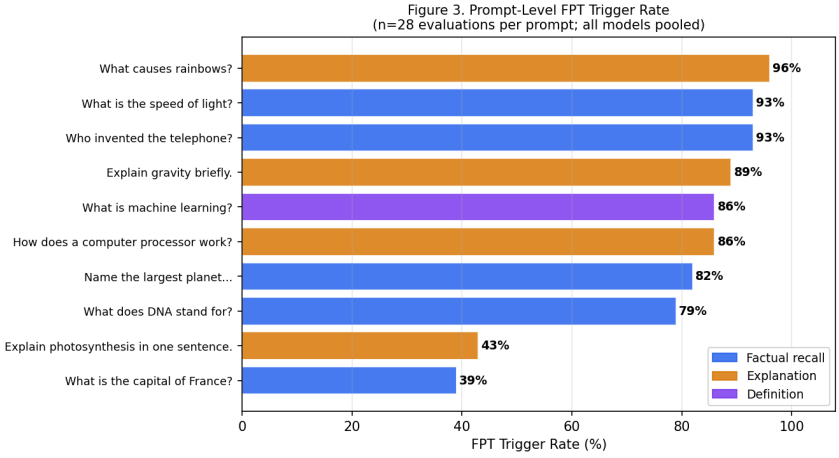

Figure 4. Baseline token count distribution for above-FIM ($n = 206$, mean 71.0) and below-FIM ($n = 60$, mean 45.5) evaluations. Diamonds indicate means. Mann–Whitney U: $p<0.001$.

5.5. Proxy Power Model Validation

Figure 5 shows the power model fit.

 doi:10.20944/preprints202605.0855.v2

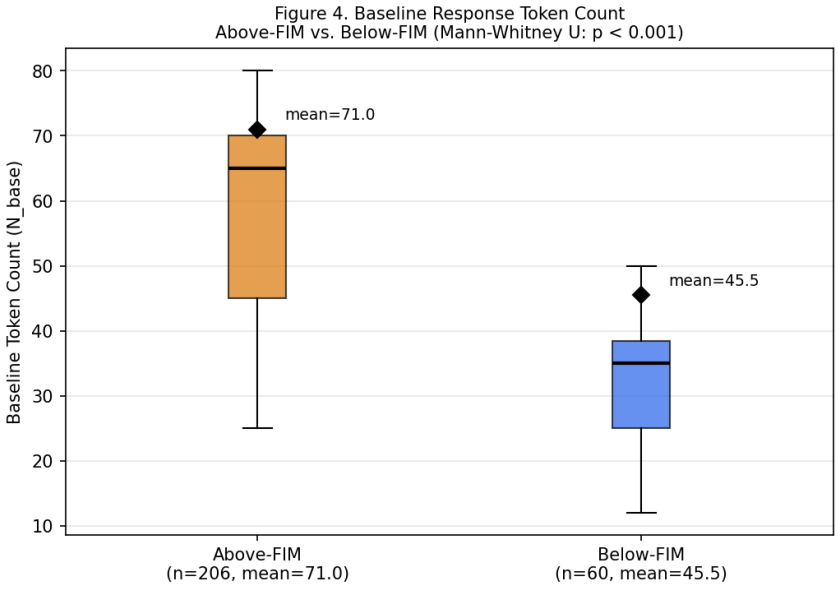

Figure 5. Proxy power model fit $\hat{P}_m = 5.0 + 0.75\,S_{\mathrm{GB}}$ W ($R^2 = 0.97$; ± 1.8 W prediction band). Root-level `powermetrics` unavailable; ± 15 % uncertainty vs. direct power monitor.

5.6. Always-50 % Baseline Comparison

6. Supplementary Controlled Validation

6.1. Deterministic Decoding: Zero FPT Triggers

Temperature = 0.0 (Experiment A: 15 runs, 150 evaluations) and temperature = 0.8 with seed = 42 (Experiment B: 22 runs, 220 evaluations): zero FPT events across all five models. This is a central limitation of the Phase 1 surface proxy: under deterministic/fixed-seed decoding, models produce longer, confident responses ($w \geq 100$) so $0.10/(w{+}1) \to 0$ and $\hat{D}_{\mathrm{KL}} \approx 0.016 \ll \tau$. Phase 2 logit-level D_{KL} computation would not exhibit this sensitivity.

6.2. Quality Preservation: 100 % Exact-Match

The original 28-run benchmark did not store response text; therefore, quality evaluation for that benchmark is unavailable. Supplementary controlled validation in Section 6 provides exact-match quality results for 370 stored-response evaluations (220 fixed-seed stochastic, 150 deterministic): 100 % exact-match for the five factual-answer prompts tested. **Scope limitation:** this result applies only to short-answer factual prompts and does not establish quality preservation for long-form reasoning, code generation, multi-turn dialogue, or safety-critical operational prompts.

6.3. Naive Cutoff Baseline

For llama3.2, always-50 % truncates the response *"The capital of France is Paris."* (8 tokens) to *"The capital of France is"* (4 tokens), removing the factual answer. K–A classifies this prompt as below-FIM ($\hat{D}_{\mathrm{KL}} = 0.030 < \tau = 0.065$) and retains the full response. See Table 5.

Table 5. K–A proxy-gated vs. always-50 % cutoff. Always-50 % achieves higher raw savings but truncates short factual answers in at least one model family. K–A selective gating preserves 100 % EM by retaining below-FIM responses unchanged.

Model	K–A Δtok%	K–A EM	Always-50 % Δtok%	Always-50 % EM
llama3.2	34.0 %	100 %	49.7 %	68 %
gemma3:latest	34.6 %	100 %	50.3 %	100 %
gemma3:27b	30.8 %	100 %	50.4 %	100 %

 doi:10.20944/preprints202605.0855.v2

7. Discussion

7.1. Mechanism: Baseline-Informed Capped Regeneration, Not Post-Hoc Truncation

Algorithm 1 makes the mechanism precise. Timing ratio $t_{ka}/t_{base} = 0.555 \pm 0.12$ across 206 valid above-FIM evaluations (expected 0.50 if $t_{ka} \propto N_{ka}$) confirms that the K–A call is a real separate inference call, not a truncation applied to an already-generated response. Total compute per above-FIM prompt: $N_{base} + N_{ka}$ tokens. In a future single-pass streaming early-exit system, compute and energy would be genuinely saved because generation would stop before producing the remaining tokens.

7.2. Correlation Interpretation

The $r = 0.806/\rho = 0.728$ correlation verifies implementation consistency of the proxy-gated token-budget rule. It does not independently validate the proxy as a detector of true geometric distributional instability per Eq. (2).

7.3. DeepSeek-R1: CoT Proxy Limitation

DeepSeek-R1's zero FPT rate reflects the proxy's incompatibility with chain-of-thought response styles: long, structured reasoning responses produce $\hat{D}_{KL} \ll \tau$ regardless of factual accuracy.

7.4. Critical Infrastructure Deployment

The framework is motivated by UAV, PLC, and autonomous vehicle contexts [3,4,28]. The present study uses general-knowledge prompts on a consumer workstation. Domain-specific validation is required before operational safety claims can be made.

7.5. Limitations

1. *Proxy ≠ geometric KL:* $\hat{D}_{KL}$ is a linguistic heuristic.
2. *Stochastic decoding required:* zero FPT under deterministic or fixed-seed decoding.
3. *Narrow quality scope:* 100 % EM confirmed only for five short-answer factual prompts.
4. *Two-call protocol:* total compute $= N_{base} + N_{ka}$ for above-FIM cases; reported savings are counterfactual.
5. *Single hardware platform:* Apple M5 only.

8. Conclusions

This paper presented a Phase 1 surface-proxy token-budget gating validation of the Kerimov–Alekberli framework across five open-source LLMs on Apple M5 Silicon. Three multi-run models show statistically characterised token savings of 30.8–34.6 % with bootstrap-confirmed confidence intervals (Pearson $r = 0.806$, Spearman $\rho = 0.728$). Supplementary controlled validation confirms 100 % exact-match quality preservation for factual prompts and reveals that deterministic and fixed-seed decoding produce zero proxy-FPT triggers, identifying the surface proxy's fundamental limitation. The always-50 % cutoff achieves ≈50 % savings but drops exact-match accuracy to 68 % for llama3.2; K–A selective gating maintains 100 % quality at 34 % savings.

Phase 2 will replace the surface proxy with direct logit-level $D_{KL}(P_t \| P_{t-1})$ computation from streaming token logits on open-weight runtimes (llama.cpp/MLX), using deterministic decoding, Joulescope direct power measurement on embedded NPU hardware, domain-specific prompt suites, a max-token-cutoff baseline comparator, output-quality evaluation via BERTScore and LLM-as-judge, and ablation of proxy formula components.

Acknowledgments: This research was supported by the Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University. No external funding was received. The authors acknowledge the open-source communities behind Ollama and the model teams at Meta (Llama), Google DeepMind (Gemma3), Alibaba/DeepSeek (deepseek-r1), and Microsoft (Phi4).

 doi:10.20944/preprints202605.0855.v2

Use of Artificial Intelligence: The authors used Claude (Anthropic) and ChatGPT (OpenAI) to assist with English language editing, structural review, and formatting guidance. All scientific content, experimental design, data collection, methodology, mathematical derivations, and conclusions are the sole work of the human authors.

Data Availability Statement: All 28-run benchmark JSON logs, K–A Metrics Server source code (Python 3.14), and raw telemetry are publicly available at https://zenodo.org/communities/kerimov-alekberli.

References

1. P. Dhar et al., "A Review on Edge Large Language Models: Design, Execution, and Applications," *arXiv:2410.11845* [cs.AI], 2024.
2. E. J. Husom et al., "Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency," *ACM Trans. Internet Things*, 2025. doi:10.1145/3767742.
3. M. A. Ferrag, A. Lakas, and M. Debbah, "Generative AI and LLMs for Critical Infrastructure Protection," *IEEE Access*, vol. 13, pp. 44982–45005, 2025.
4. S. Javaid, H. Fahim, B. He, and N. Saeed, "Large Language Models for UAVs: Current State and Pathways to the Future," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 1166–1192, 2024.
5. H. Karimov and R. Z. Alekberli, "The Kerimov–Alekberli Model: An Information-Geometric Framework for Real-Time System Stability," *arXiv:2604.24083* [cs.AI], Apr. 2026.
6. H. Karimov and R. Z. Alekberli, "An Information-Geometric Framework for Stability Analysis of Large Language Models under Entropic Stress," *arXiv:2604.24076* [cs.AI], Apr. 2026.
7. R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Res. Dev.*, vol. 5, no. 3, pp. 183–191, 1961.
8. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proc. NeurIPS*, vol. 33, 2020, pp. 9459–9474.
9. S. Kadavath et al., "Language Models (Mostly) Know What They Know," *arXiv:2207.05221* [cs.CL], 2022.
10. S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting Hallucinations in Large Language Models Using Semantic Entropy," *Nature*, vol. 630, pp. 625–630, 2024.
11. W. Su et al., "Unsupervised Real-Time Hallucination Detection Based on the Internal States of LLMs," in *Findings ACL*, 2024, pp. 14379–14391.
12. S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks," in *Proc. ICPR*, 2016, pp. 2464–2469.
13. A. Graves, "Adaptive Computation Time for Recurrent Neural Networks," *arXiv:1603.08983* [cs.NE], 2016.
14. Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
15. Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv:2309.01219*, 2023.
16. L. Kuhn, Y. Gal, and S. Farquhar, "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG," in *Proc. ICLR*, 2023.
17. A. Azaria and T. Mitchell, "The Internal State of an LLM Knows When It Is Lying," in *Findings EMNLP 2023*, pp. 967–976.
18. M. Sajjad et al., "Depth-Adaptive Transformer," *arXiv:1910.10073*, 2019.
19. T. Schuster et al., "Confident Adaptive Language Modeling," in *Proc. NeurIPS*, 2022.
20. C. Chen et al., "Accelerating Large Language Model Decoding with Speculative Sampling," *arXiv:2302.01318*, 2023.
21. S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.
22. C. Zhao et al., "The Adversarial Attack and Detection under the Fisher Information Metric," in *Proc. AAAI*, 2019, pp. 5869–5876.
23. S. Blaber and D. A. Sivak, "Optimal Control in Stochastic Thermodynamics," *J. Phys. Commun.*, vol. 7, p. 033001, 2023.
24. H. Beik-Mohammadi et al., "Extended Neural Contractive Dynamical Systems," *Int. J. Robot. Res.*, vol. 45, pp. 714–745, 2026.
25. E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proc. ACL*, 2019.
26. S. Samsi et al., "From Words to Watts: Benchmarking the Energy Costs of LLM Inference," in *Proc. IEEE HPEC*, 2023.

 doi:10.20944/preprints202605.0855.v2

27. R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo, "Efficiency of LLMs: A Study on Inference Energy Usage," *arXiv:2202.05379*, 2023.
28. Y. Yigit et al., "Generative AI and LLMs for Critical Infrastructure Protection," *Sensors*, vol. 25, no. 6, p. 1666, 2025.
29. I. Seshia, S. Sadigh, and A. Sangiovanni-Vincentelli, "Verification and Control of Cyber-Physical Systems," in *Proc. CPS-IoT Week*, 2018.
30. A. Bérut et al., "Experimental Verification of Landauer's Principle," *Nature*, vol. 483, pp. 187–189, 2012.