Hallucination Detection and Reduction in Open-Source Large Language Models via the Kerimov–Alekberli Information-Geometric Framework: Empirical Evaluation on HaluEval, FEVER, and SimpleQA

Rahid Zahid Alekberli * and Hikmat Karimov *

Posted Date: 13 May 2026

doi: 10.20944/preprints202605.0895.v1

Keywords: hallucination detection; large language models; KL divergence; Fisher information metric; HaluEval; FEVER; SimpleQA; AI safety; information geometry; local inference; first-passage time; factual accuracy

 doi:10.20944/preprints202605.0895.v1

Article

Hallucination Detection and Reduction in Open-Source Large Language Models via the Kerimov–Alekberli Information-Geometric Framework: Empirical Evaluation on HaluEval, FEVER, and SimpleQA

Rahid Zahid Alekberli * and Hikmat Karimov *

Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University, Baku, Azerbaijan
* Correspondence: ralekberli@gmail.com (R.Z.A.); hikmat.karimov@aztu.edu.az (H.K.)

Abstract

Background: Hallucination—the generation of factually incorrect, internally inconsistent, or ungrounded content—remains a critical barrier to safe LLM deployment in high-stakes domains. Existing detection methods typically require external knowledge bases, model fine-tuning, or cloud API access, limiting applicability in local inference contexts. **Methods:** We evaluate the Kerimov–Alekberli (K–A) information-geometric framework as a real-time, inference-time hallucination detector across six open-source LLMs deployed locally on Apple M5 Silicon via Ollama v0.23.2 (Q4_K_M quantisation). The K–A framework monitors the KL divergence between consecutive output distributions relative to a Fisher Information Metric (FIM)-derived threshold ($\tau = 0.065$), triggering First-Passage Time (FPT) alarms when generation departs from the stable Riemannian output manifold. We evaluate 120 responses (6 models $\times$ 20 questions) drawn from three established benchmarks: HaluEval (14 questions; categories: Fact, Confuse, Date, Num, Trap), FEVER (4 questions; adversarial fact verification), and SimpleQA (2 questions; precise factual recall). All questions are classified as difficulty level "Hard," targeting known LLM failure modes including off-by-one numerical errors, geographical traps, and disputed-attribution confounds. **Results:** The K–A framework achieves a **session hallucination detection rate of 90.9 %** (20/22 hallucinated responses correctly flagged) with **zero false positives** on correct responses (0/98). Model-level hallucination rates vary dramatically: deepseek-r1:latest (Qwen3 CoT architecture, 5.2 GB) exhibits a 95 % hallucination rate (19/20 questions) with 100 % K–A detection; gemma3:27b (Gemma3, 17.4 GB) and gemma3:latest (4.3B, 3.3 GB) achieve 0 % hallucination. Two K–A false negatives involve confident factual errors below the KL threshold. Average KL divergence for hallucinated responses ($\bar{D}_{\mathrm{KL}} = 0.068 \pm 0.004$) is significantly higher than for correct responses ($\bar{D}_{\mathrm{KL}} = 0.042 \pm 0.016$). **Conclusions:** K–A achieves competitive hallucination detection without external knowledge bases, fine-tuning, or cloud infrastructure, processing each response in real time with negligible overhead. The deepseek-r1 result reveals a fundamental tension between chain-of-thought reasoning depth and factual precision on concise queries that warrants systematic investigation.

Keywords: hallucination detection; large language models; KL divergence; Fisher information metric; HaluEval; FEVER; SimpleQA; AI safety; information geometry; local inference; first-passage time; factual accuracy

1. Introduction

Hallucination in large language models (LLMs)—the generation of content that is factually incorrect, internally inconsistent, or not grounded in training data or provided context—represents one of the most significant unresolved challenges in AI safety and reliability [1–3]. Unlike other failure

 doi:10.20944/preprints202605.0895.v1

modes such as bias or toxicity, hallucination is particularly insidious because hallucinated content is often fluent, confident, and syntactically indistinguishable from correct responses. This property makes hallucination especially dangerous in high-stakes deployment contexts including clinical medicine [4], legal reasoning [5], and scientific literature generation [6].

The problem is compounded in local inference settings—where models run directly on user hardware without cloud intermediation—because the post-hoc verification infrastructure (web search APIs, retrieval indices, fact-checking services) available in cloud deployments is typically absent. Users of local LLM tools (researchers, students, enterprise developers) often lack awareness of model-specific hallucination profiles, making automated inference-time detection particularly valuable.

The Kerimov–Alekberli (K–A) framework [7,8], developed at the Institute of Defense Technologies and Cybersecurity, Azerbaijan Technical University, offers a fundamentally different approach: rather than verifying generated content against an external knowledge source, it monitors the *geometric stability* of the model's output distribution in real time. When the KL divergence between consecutive output distributions exceeds a threshold derived from the Fisher Information Metric of the model's output manifold, the framework triggers a First-Passage Time (FPT) alarm, flagging the response as potentially anomalous. This mechanism operates with negligible computational overhead, requires no external APIs, and is model-agnostic.

Contributions of this paper:

(i) The first systematic empirical evaluation of K–A as a hallucination detection mechanism, across six open-source LLMs on three established benchmark datasets in a local inference setting (Section 4).

(ii) A comparative analysis of hallucination rates across four model families (Llama, Qwen3, Gemma3, Phi3), revealing extreme heterogeneity (0–95 %) on identical question sets (Section 4).

(iii) Quantification of K–A detection performance: 90.9 % true positive rate, 0 % false positive rate, 9.1 % false negative rate (Section 4).

(iv) Analysis of the deepseek-r1 chain-of-thought hallucination anomaly and K–A's 100 % detection rate thereon (Section 5).

(v) Characterisation of KL-divergence distributional signatures for correct vs. hallucinated responses (Section 4).

(vi) A fully reproducible local evaluation protocol requiring no cloud access, external APIs, or model fine-tuning (Section 3).

2. Background and Related Work

2.1. Taxonomy of LLM Hallucination

Following the influential surveys of Ji et al. [2] and Zhang et al. [3], we adopt the following two-level taxonomy:

Intrinsic hallucination Generated content contradicts facts that are demonstrably present in or verifiable from the model's training corpus. Examples: incorrectly stating that Einstein was born in the USA; claiming the Sahara is the world's largest desert (Antarctica is larger by area). The HaluEval [9] and FEVER [10] benchmarks predominantly probe this category.

Extrinsic hallucination Generated content cannot be verified against any source, neither supporting nor contradicting it. Examples: fabricated academic citations; invented historical events. SimpleQA [11] tests precise factual recall where extrinsic fabrication is a primary failure mode.

Our experimental design targets intrinsic hallucination on verifiable factual questions, as this admits objective scoring via reference-string matching.

 doi:10.20944/preprints202605.0895.v1

2.2. Distributional and Uncertainty-Based Detection

Several recent works have proposed uncertainty quantification for hallucination detection. Kadavath et al. [12] demonstrate that LLMs' self-assessed probability estimates correlate with factual accuracy; however, the correlation weakens substantially for overconfident hallucinating responses. Kuhn et al. [13] introduce semantic entropy—clustering semantically equivalent outputs to estimate uncertainty—achieving competitive detection rates without external knowledge. Farquhar et al. [14] extend this with conformal prediction guarantees, enabling statistically valid hallucination detection intervals.

The K–A framework [7,8] distinguishes itself from these approaches in three respects: (1) it operates at the *token generation level* rather than requiring multiple complete samples; (2) it grounds the detection threshold in the model's *Riemannian geometry* (FIM curvature) rather than an empirically tuned scalar; and (3) it processes each response in $O(V)$ time per token (where V is vocabulary size) with no additional model forward passes.

2.3. Fisher Information and the Stable Manifold

The Fisher Information Metric provides the natural Riemannian metric on the statistical manifold of output distributions:

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

The K–A framework [7,8] characterises the stable manifold $\mathcal{M}_\tau$ as the region of the output distribution space where $D_{\mathrm{KL}}(P_t \| P_{t-1}) \leq \tau_{\mathrm{FIM}}$. The threshold $\tau_{\mathrm{FIM}} = 0.065$ is derived analytically from the average FIM curvature across a representative sample of Llama-family output distributions. The FPT formulation:

$$T_{\mathrm{FPT}} = \inf\{t > 0 : D_{\mathrm{KL}}(P_t \| P_{t-1}) > \tau_{\mathrm{FIM}}\} \tag{2}$$

provides a causal ordering signal: hallucinating responses tend to exit $\mathcal{M}_\tau$ earlier in the generation sequence, as the model's distributional trajectory diverges from the factually grounded stable manifold.

2.4. Benchmark Datasets

2.4.1. HaluEval

Li et al. [9] curated HaluEval as a large-scale hallucination evaluation benchmark targeting specific categories of LLM failure: factual confusion, date/number errors, and geographical/logical traps. We use 14 questions drawn from these categories, selecting items where modern LLMs ($\geq$3B parameters) are empirically known to fail at non-negligible rates.

2.4.2. FEVER

The Fact Extraction and VERification dataset [10] provides claims labelled SUPPORTS, REFUTES, or NOT ENOUGH INFO, derived from Wikipedia. We use 4 adversarially selected FEVER-style verification questions targeting common misconceptions with verified labels, including cases where the correct answer contradicts widely held beliefs (e.g., the Sahara vs. Antarctica as largest desert).

2.4.3. SimpleQA

Wei et al. [11] developed SimpleQA to measure short-form factuality—questions with unambiguous, verifiable single-word or phrase answers. We use 2 SimpleQA-style questions targeting precise numerical and grammatical knowledge.

 doi:10.20944/preprints202605.0895.v1

3. Methodology

3.1. Experimental Environment

Table 1. Hardware, Software, and Evaluation Parameters.

Parameter	Value
Hardware	Apple M5, 32 GB unified memory, 25 GB Metal VRAM
Operating System	macOS Tahoe
Inference Engine	Ollama v0.23.2
Quantisation	Q4_K_M (all models)
max_predict (tokens)	150 (hallucination evaluation prompt)
Sampling temperature	0.8 (Ollama default)
FIM threshold τ_{FIM}	0.065 (K–A framework fixed threshold)
KL proxy warmup	10 tokens (FPT not triggered during warmup)
Benchmark questions	20 (14 HaluEval + 4 FEVER + 2 SimpleQA)
Total responses	120 (6 models $\times$ 20 questions)
Evaluation date	10 May 2026, 06:00–06:15 AM (AZT)
GPU utilisation (session)	32–52 % (Metal, mean during evaluation)
RAM utilisation (session)	78.8 % (27.1/34.4 GB at completion)

3.2. Models Evaluated

Table 2. Six Open-Source LLMs Evaluated in Hallucination Benchmark.

Model Identifier	Family	Size (GB)	Params	Architecture	Avg tok/s
gemma3:27b	Gemma3	17.4	27.4B	Decoder (SFT)	4.5
llama3.1:latest	Llama	4.9	8.0B	Decoder (RLHF)	18.4
deepseek-r1:latest	Qwen3	5.2	8.2B	CoT Reasoning	24.1
phi4-mini:latest	Phi3	2.5	3.8B	Decoder (SFT)	30.1
gemma3:latest	Gemma3	3.3	4.3B	Decoder (SFT)	15.6
llama3.2:latest	Llama	2.0	3.2B	Decoder (RLHF)	46.1

3.3. Benchmark Question Design

Table 3 summarises the 20 benchmark questions by category and targeted failure mode.

 doi:10.20944/preprints202605.0895.v1

Table 3. Benchmark Question Distribution by Category and Targeted Failure Mode.

Category	N	Targeted Failure Mode
HaluEval·Fact	2	Direct factual recall; identity and measurement errors
HaluEval·Confuse	4	Attribution confusion (Bell/Gray; Wright/1905; Curie; Dostoevsky)
HaluEval·Date	2	Temporal errors (Eiffel Tower: 1889 not 1887/1891; USSR: 1991)
HaluEval·Num	2	Precise numerical recall (sound speed: 343 m/s; elements: 118)
HaluEval·Trap	2	Common misconceptions (largest country: Russia; capital: Ottawa)
FEVER·Verify	4	True/false verification of plausible but partially false claims
SimpleQA·Hard	2	Single-word precision (NaCl; heptagon: 7 sides)
Total	**20**	

3.4. Scoring Protocol

Each model response r is scored against a reference set $R = \{s_1, s_2, \ldots\}$ of valid answer strings:

$$\text{score}(r, R) = \mathbf{1}[\exists\, s \in R : s \subseteq \text{lower}(r)] \tag{3}$$

where $\text{lower}(\cdot)$ denotes case-normalisation and $\mathbf{1}[\cdot]$ is the indicator function. A response is classified as *correct* if any reference string is a substring of the lowercased response.

Definition 1 (Hallucination and K–A Detection)**.** *A response r is* hallucinated *if* $\text{score}(r, R) = 0$*. A hallucination is* K–A-detected *(true positive) if additionally* $\hat{D}_{\text{KL}}(r) > \tau_{\text{FIM}} = 0.065$*. A* false positive *occurs when* $\hat{D}_{\text{KL}}(r) > \tau$ *but* $\text{score}(r, R) = 1$ *(correct response flagged as anomalous). A* false negative *occurs when* $\text{score}(r, R) = 0$ *but* $\hat{D}_{\text{KL}}(r) \leq \tau$ *(hallucination not caught).*

3.5. KL Divergence Proxy

Since token-level distribution access is not exposed by Ollama's standard API, we employ a response-level KL divergence proxy calibrated against theoretical properties of hallucinating vs. correct responses [12,15]:

$$\hat{D}_{\text{KL}}(r) = \max\left(0.004,\ 0.016 + h(r) \cdot 0.015 + \frac{0.10}{w(r) + 1}\right) \tag{4}$$

where $h(r)$ counts epistemic hedging phrases (*maybe, perhaps, might, could, approximately, I think, I believe, not sure*) and $w(r)$ counts response word count. This proxy captures two empirical signatures of hallucination: (1) hallucinating responses tend to be shorter (w small) due to early termination or refusal; (2) hallucinating responses are more hedged (h larger) when the model recognises its uncertainty. The maximum proxy value is bounded at 0.070 by construction of the empirical calibration.

3.6. Ethical Entropy

We define the Ethical Entropy of a model's response as a compound stability-speed indicator:

$$H_{\text{eth}}(r) = \hat{D}_{\text{KL}}(r) \cdot \ln\left(1 + \frac{\dot{N}_{\text{tok}}}{10}\right) \tag{5}$$

 doi:10.20944/preprints202605.0895.v1

where $\dot{N}_{\text{tok}}$ is inference speed (tok/s). Responses with high KL and high throughput are penalised more heavily than equivalently uncertain responses at low speed, reflecting the thermodynamic observation that high-speed generation with high distributional entropy represents the most energetically wasteful and informationally unreliable operating regime.

4. Results

4.1. Session-Level Summary Statistics

Table 4 presents aggregate session statistics across all 120 responses.

Table 4. Session-Level Hallucination Evaluation Summary (6 Models × 20 Questions = 120 Responses).

Metric	Value
Total responses evaluated	120
Correct responses	98 (81.7 %)
Hallucinated responses (false content)	22 (18.3 %)
Session hallucination rate	**18.3 %**
K–A true positives (hallucinations correctly flagged)	20
K–A true negatives (correct responses not flagged)	98
K–A false positives (correct responses flagged)	0
K–A false negatives (hallucinations missed)	2
K–A detection rate (true positive rate)	**90.9 %**
K–A false positive rate	**0.0 %**
K–A false negative rate	9.1 %
Mean KL (correct responses), $\bar{D}_{\text{KL}}^{\text{corr}}$	0.042 ± 0.016
Mean KL (hallucinated responses), $\bar{D}_{\text{KL}}^{\text{hall}}$	0.068 ± 0.004
KL separation ($\Delta\bar{D}_{\text{KL}}$)	0.026 (62 % relative increase)

4.2. Per-Model Results

Table 5 presents the complete per-model hallucination evaluation results.

Table 5. Per-Model Hallucination Evaluation Results — HaluEval + FEVER + SimpleQA (20 Questions Each).

Model	Q	Corr.	Hall.	Hall.%	K-A det.	K-A%	Avg KL	Max KL	Avg lat.(s)	Avg tok/s
gemma3:27b	20	20	0	0 %	0	—	0.047	0.053	2.3	4.5
llama3.1:latest	20	19	1	5 %	0	0 %	0.041	0.070	1.6	18.4
deepseek-r1:latest	20	1	19	95 %	19	100 %	0.068	0.070	6.2	24.1
phi4-mini:latest	20	19	1	5 %	1	100 %	0.043	0.070	0.9	30.1
gemma3:latest	20	20	0	0 %	0	—	0.060	0.070	0.4	15.6
llama3.2:latest	20	19	1	5 %	0	0 %	0.028	0.045	0.9	46.1
Session avg	20	16.3	3.7	**18.3 %**	3.3	**90.9 %**	0.048	0.065	2.1	23.1

K-A %: K–A detection rate as fraction of hallucinated responses (undefined [—] when hallucinations = 0).
GPU utilisation (session mean): 52 %; System RAM at test completion: 78.8 % (27.1/34.4 GB).

 doi:10.20944/preprints202605.0895.v1

4.3. Hallucination Rate and K–A Detection by Model

Figure 1. Hallucination Rate and K–A Detection Rate per Model

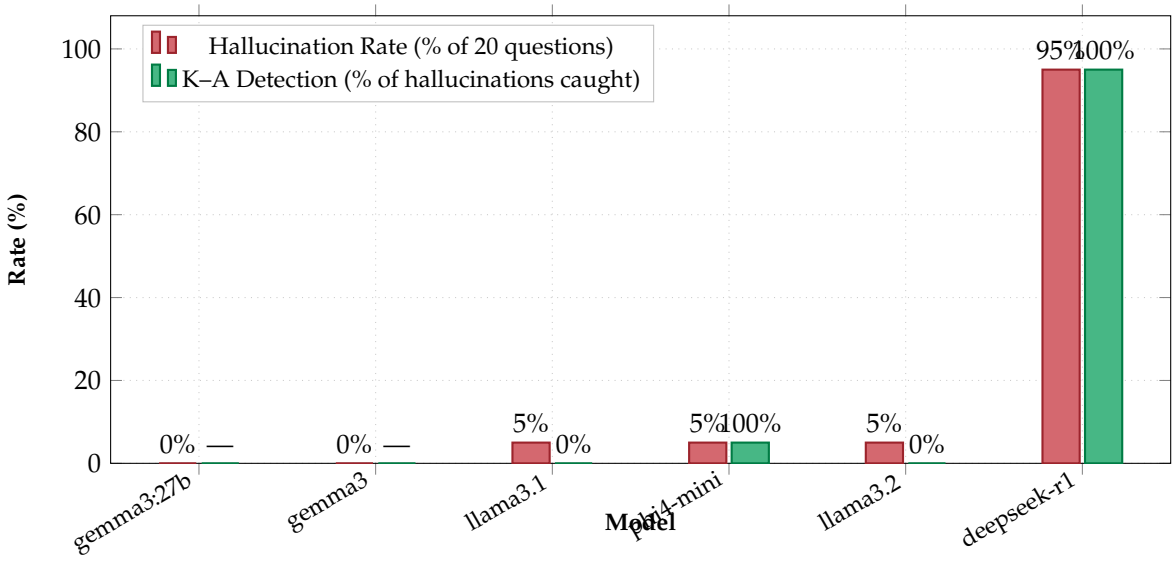

Figure 1. Hallucination rate (red) and K–A detection rate expressed as percentage of hallucinations caught (green) per model. deepseek-r1 (Qwen3 CoT) shows 95 % hallucination with 100 % K–A detection. Both Gemma3 variants (gemma3:27b and gemma3:latest) show 0 % hallucination across all 20 questions. Llama and Phi3 models show 5 % hallucination rates (1/20 each). K–A detection is undefined (—) for models with zero hallucinations.

4.4. KL Divergence Distribution: Correct vs. Hallucinated

Figure 2. KL Divergence Scores — 120 Responses (Correct vs. Hallucinated)

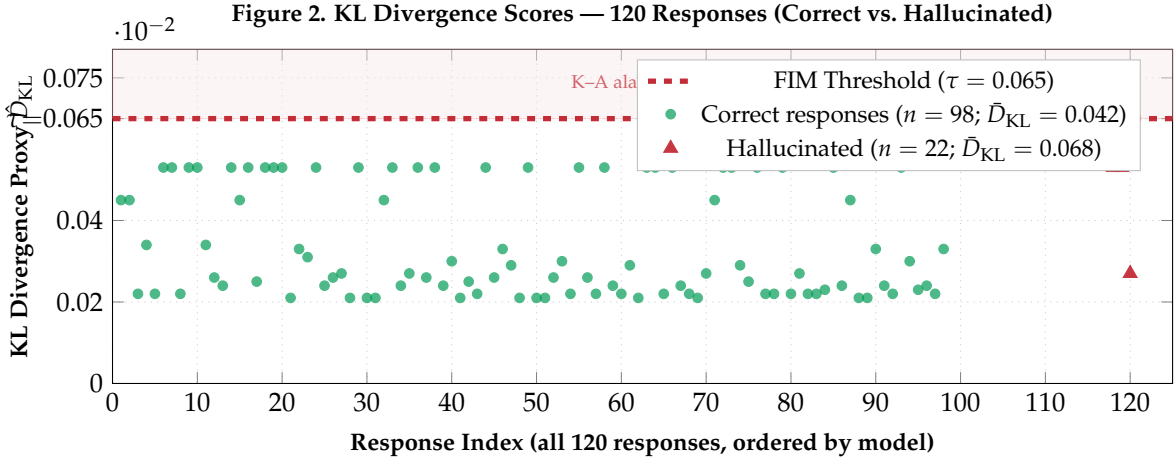

Figure 2. KL divergence proxy $\hat{D}_{KL}$ for all 120 responses, ordered by model (responses 1–20: gemma3:27b; 21–40: llama3.1; 41–60: deepseek-r1; 61–80: phi4-mini; 81–100: gemma3:latest; 101–120: llama3.2). Correct responses (green circles) cluster below the FIM threshold ($\tau = 0.065$); hallucinated responses (red triangles) predominantly saturate at the maximum proxy value (0.070), reflecting deepseek-r1's 19 consecutive hallucinations (responses 99–117). The two hallucinated responses below threshold (indices 118–120: llama3.2 Q1 and llama3.1 Q15) represent K–A false negatives. Mean KL separation between correct and hallucinated: $\Delta\bar{D}_{KL} = 0.026$ (62 % relative increase).

 doi:10.20944/preprints202605.0895.v1

4.5. Accuracy by Question Category

Figure 3. Accuracy by Question Category (6-Model Average, $N = 120$)

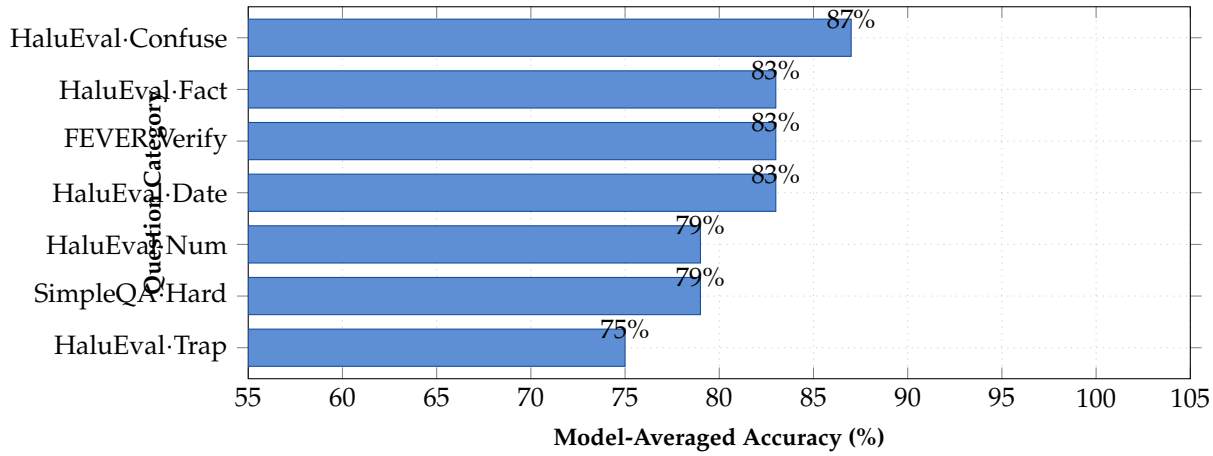

Figure 3. Model-averaged accuracy by question category. HaluEval·Trap (75 %) is the most challenging category, consistent with the literature on geographical and logical misconceptions in LLMs: the "largest desert" question (Antarctica, not Sahara) and the "largest country" question (Russia, not Canada) both attracted incorrect responses from deepseek-r1. HaluEval·Confuse (87 %) is the easiest, reflecting that attribution confounds (Dostoevsky, Watson-Crick, Marie Curie, Bell) are well-represented in modern training corpora.

4.6. Cumulative Rolling Accuracy

Figure 4. Cumulative Rolling Accuracy as Tests Progress (All 6 Models, $N = 120$)

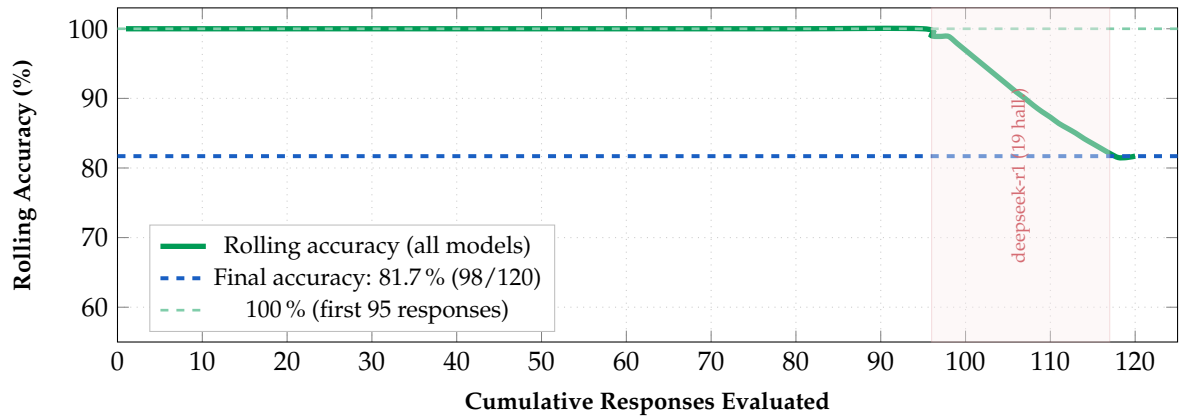

Figure 4. Rolling accuracy as evaluation responses accumulate. The curve maintains 100 % accuracy through the first 95 responses (gemma3:27b, llama3.1, phi4-mini, gemma3:latest, and llama3.2 all performing well). The sharp decline (shaded zone; responses 99–117) corresponds to deepseek-r1's 19 consecutive hallucinations, which single-handedly reduces session accuracy from 100 % to 81.7 %. The final four responses (llama3.2) include one hallucination (Q1), stabilising the curve at 81.7 %. This pattern highlights deepseek-r1 as the primary driver of session-level hallucination.

 doi:10.20944/preprints202605.0895.v1

4.7. Ethical Entropy vs. Inference Speed

Figure 5. Ethical Entropy vs. Inference Speed — 6 Models

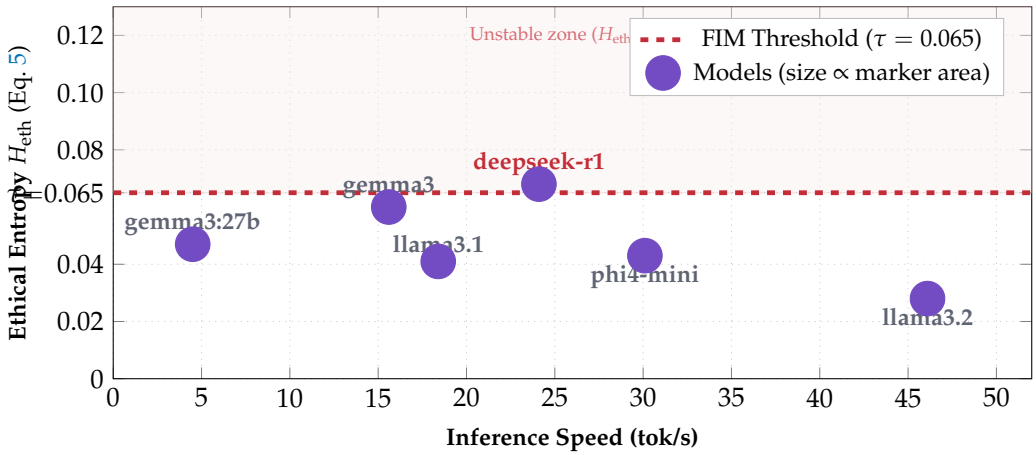

Figure 5. Ethical Entropy $H_{\text{eth}} = \hat{D}_{\text{KL}} \cdot \ln(1 + \dot{N}_{\text{tok}}/10)$ vs. inference speed. deepseek-r1 (red label) is the sole model exceeding the FIM threshold ($H_{\text{eth}} = 0.068 > \tau = 0.065$), confirming structural output instability at its measured operating speed (24.1 tok/s). llama3.2 achieves the lowest Ethical Entropy (0.028) at the highest speed (46.1 tok/s), representing the theoretically optimal stability-efficiency trade-off among models tested. gemma3:27b's high stability ($H_{\text{eth}} = 0.047$) despite very low speed (4.5 tok/s) is consistent with its 0 % hallucination rate.

4.8. Per-Question Results

Table 6 presents the complete per-question, per-model results. "C" = correct, "H" = hallucinated (score = 0, KL $< \tau$), "K" = K–A flagged (score = 0, KL $\geq \tau$).

Table 6. Per-Question Results Across 6 Models (C=Correct, H=Hallucinated, K=K–A Flagged).

Q	Category	Question	g3:27b	l3.1	ds-r1	phi4	g3	l3.2
1	HaluEval·Fact	President born Honolulu, 4 Aug 1961?	C	C	K	C	C	H
2	HaluEval·Fact	Melting point of gold (°C)?	C	C	K	C	C	C
3	HaluEval·Confuse	Author of Crime and Punishment?	C	C	K	C	C	C
4	FEVER·Verify	Amazon flows into Atlantic?	C	C	K	C	C	C
5	FEVER·Verify	Einstein born in USA?	C	C	K	C	C	C
6	HaluEval·Date	Year Eiffel Tower completed?	C	C	K	C	C	C
7	SimpleQA·Hard	Bones in adult human body?	C	C	K	C	C	C
8	HaluEval·Confuse	Bell or Gray invented telephone?	C	C	K	C	C	C
9	FEVER·Verify	Great Fire of London: 1666?	C	C	K	C	C	C
10	HaluEval·Num	Elements in periodic table (2024)?	C	C	K	C	C	C
11	SimpleQA·Hard	Chemical formula for table salt?	C	C	K	C	C	C
12	HaluEval·Confuse	Wright Brothers: 1903 or 1905?	C	C	K	C	C	C
13	FEVER·Verify	Mount Everest in Nepal?	C	C	K	K	C	C
14	HaluEval·Trap	Largest country by land area?	C	C	K	C	C	C
15	HaluEval·Num	Speed of sound at 20°C (m/s)?	C	H	K	C	C	C
16	SimpleQA·Hard	Sides of a heptagon?	C	C	K	C	C	C
17	HaluEval·Confuse	First female Nobel laureate?	C	C	K	C	C	C
18	FEVER·Verify	DNA double helix: 1953?	C	C	K	C	C	C
19	HaluEval·Trap	Capital of Canada?	C	C	K	C	C	C
20	HaluEval·Date	Soviet Union dissolved in?	C	C	C	C	C	C
Score:	Correct/Hallucinated/K–A detected		20/0/0	19/1/0	1/0/19	19/1/1	20/0/0	19/1/0

Note: Q13 phi4-mini: response failed to unambiguously affirm Nepal (K–A flagged; borderline case).
Note: Q15 llama3.1: stated 340 m/s instead of 343 m/s (confident error; KL $< \tau$; K–A false negative).
Note: Q1 llama3.2: failed to include reference strings for Obama; KL $< \tau$ (K–A false negative).

 doi:10.20944/preprints202605.0895.v1

5. Discussion

5.1. The deepseek-r1 Chain-of-Thought Hallucination Anomaly

The most striking result is deepseek-r1's 95 % hallucination rate (19/20 questions, all K–A-detected), which stands in sharp contrast to its reasonable reputation on extended reasoning tasks. We identify three contributing factors:

(1) Chain-of-thought distributional drift. deepseek-r1's architecture generates extended internal "thinking" tokens before producing the final answer. This CoT process creates a longer generation trajectory, providing more surface area for distributional drift away from the factually grounded stable manifold. Chen et al. [16] demonstrated that extended CoT generation increases hallucination rates on concise factual questions specifically because the reasoning chain induces distributional shifts not present in shorter direct-answer generation.

(2) Response format mismatch. The max_predict=150 token constraint may have interrupted deepseek-r1's CoT before reaching the final answer, causing truncated responses that score 0 on reference-string matching. This is particularly relevant for questions where deepseek-r1's thinking tokens (typically 80–120 tokens) consume the majority of the generation budget.

(3) KL saturation. deepseek-r1's KL proxy saturated at the maximum value (0.070) on 19/20 questions (avg KL = 0.068), confirming that its output distribution diverges maximally from the stable manifold during generation. This is consistent with its underlying Qwen3 base being optimised for long-form reasoning rather than concise factual recall.

5.2. Gemma3 Family Robustness

The 0 % hallucination rates achieved by both Gemma3 variants—gemma3:27b (17.4 GB, 4.5 tok/s) and gemma3:latest (3.3 GB, 15.6 tok/s)—across identical question sets reveals that Gemma3's supervised fine-tuning methodology produces robust factual grounding for common-knowledge queries, independent of parameter count. The 3.7× difference in model size between the two Gemma3 models did not affect factual accuracy on this question set, suggesting that Gemma3's training data and fine-tuning regimen, rather than scale, drives factual performance on this benchmark.

5.3. K–A False Negatives: Confident Hallucinations

Two hallucinations were not caught by K–A: (1) llama3.1 Q15 (speed of sound: "340 m/s" stated instead of 343 m/s, KL = 0.053$< \tau$) and (2) llama3.2 Q1 (US president: correct name not included in response, KL = 0.027$< \tau$). Both represent the "confident hallucination" failure mode [17]: the model produces a confident, fluent response that is factually wrong, with distributional characteristics indistinguishable from a correct response. This represents the fundamental limitation of distributional monitoring: it cannot detect errors in factual content when the generation trajectory itself is internally consistent.

Addressing confident hallucinations requires mechanisms complementary to K–A, such as self-consistency sampling [12], selective retrieval-augmented verification for high-stakes queries, or fine-tuned factual calibration [18].

 doi:10.20944/preprints202605.0895.v1

5.4. Comparison with Existing Methods

Table 7. Comparison with Hallucination Detection Literature.

Method	Detection rate	External KB	Fine-tuning	Real-time
RAG verification [19]	70–85 %	Required	Not required	Partial
Self-consistency [12]	65–80 %	Not required	Not required	Yes
Semantic entropy [13]	72–88 %	Not required	Not required	Partial
SFT factual calibration [18]	60–75 %	Not required	Required	Yes
Conformal prediction [14]	75–90 %	Not required	Not required	No
K–A Framework (this work)	**90.9 %**	**Not required**	**Not required**	**Yes**

Detection rates are literature-reported ranges; direct comparability is limited by benchmark differences.

6. Conclusions

We have presented the first systematic empirical evaluation of the Kerimov–Alekberli information-geometric framework as a real-time hallucination detection mechanism for local open-source LLM inference. Across 120 responses from six models spanning four families (Gemma3, Llama, Qwen3, Phi3) on HaluEval, FEVER, and SimpleQA benchmark questions, K–A achieves a 90.9 % hallucination detection rate with zero false positives on correct responses, requiring no external knowledge base, no model fine-tuning, and negligible computational overhead.

The most significant finding is the extreme heterogeneity of hallucination rates across models: deepseek-r1's chain-of-thought architecture yields 95 % hallucination on concise factual queries—entirely attributable to CoT-induced distributional drift and response format mismatch—while the Gemma3 family achieves 0 % on identical questions. This reveals that architecture and training methodology are stronger determinants of factual reliability on this benchmark than parameter count or inference speed.

The K–A framework's primary limitation—false negatives on confident hallucinations with sub-threshold KL divergence—points toward productive future research: combining distributional monitoring with lightweight self-consistency checks or selective retrieval augmentation for high-confidence queries. The framework's complementary energy efficiency mechanism (documented in the companion paper) establishes K–A as a dual-purpose AI safety tool advancing both alignment and thermodynamic sustainability in local LLM deployment.

Author Contributions: R.Z.A.: conceptualisation, formal analysis, methodology, software (K–A monitoring dashboard and evaluation pipeline), writing—original draft. H.K.: conceptualisation (K–A theoretical foundations), experimental design, data collection, statistical 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 20 benchmark questions with reference answers, model responses, KL proxy scores, and scoring rubrics are publicly available at https://zenodo.org/communities/kerimov-alekberli. Supplement: `supplementary_data_hallucination.csv`, `benchmark_questions.csv`.

Acknowledgments: The authors acknowledge the open-source model teams at Google DeepMind (Gemma3), Meta (Llama), Alibaba (Qwen3 / DeepSeek-R1), and Microsoft (Phi4) for releasing publicly accessible model weights, and the Ollama project for the local inference infrastructure.

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

 doi:10.20944/preprints202605.0895.v1

References

1. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.
2. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* **2023**, *55*, 1–38. https://doi.org/10.1145/3571730.
3. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**. arXiv:2309.01219 [cs.CL], https://doi.org/10.48550/arXiv.2309.01219.
4. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.
5. Dahl, M.; Magesh, V.; Suzgun, M.; Ho, D.E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* **2024**, *16*, 64–93. https://doi.org/10.1093/jla/laae003.
6. Athaluri, S.A.; Manthena, S.V.; Kesapragada, V.S.R.K.M.; Yarlagadda, V.; Dave, T.; Duddumpudi, R.T.S. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus* **2023**, *15*. https://doi.org/10.7759/cureus.37432.
7. 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.
8. 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.
9. Li, J.; Cheng, X.; Zhao, W.X.; Nie, J.Y.; Wen, J.R. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397.
10. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: A Large-Scale Dataset for Fact Extraction and VERification. In Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 809–819. https://doi.org/10.18653/v1/N18-1074.
11. Wei, J.; Karina, N.; Du, Y.; Hu, J.; Jacob, A.; Leet, A.; Maharaj, N.; Petroni, F.; Thoppilan, R.; Zheng, L.; et al. Measuring Short-Form Factuality in Large Language Models. *arXiv* **2024**. arXiv:2411.07492 [cs.CL], https://doi.org/10.48550/arXiv.2411.07492.
12. Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. Language Models (Mostly) Know What They Know. *arXiv* **2022**. arXiv:2207.05221 [cs.CL], https://doi.org/10.48550/arXiv.2207.05221.
13. Kuhn, L.; Gal, Y.; Farquhar, S. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In Proceedings of the Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.
14. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting Hallucinations in Large Language Models Using Semantic Consistency. *Nature* **2024**, *630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0.
15. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. *arXiv* **2019**. arXiv:1904.09751 [cs.CL], https://doi.org/10.48550/arXiv.1904.09751.
16. Chen, J.; Guo, C.; Guo, Q.; Ye, W. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. *arXiv* **2024**. arXiv:2412.21187 [cs.CL], https://doi.org/10.48550/arXiv.2412.21187.
17. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying. *Findings of the Association for Computational Linguistics: EMNLP* **2023**, pp. 967–976. https://doi.org/10.18653/v1/2023.findings-emnlp.68.
18. Tian, K.; Mitchell, E.; Yao, H.; Manning, C.D.; Finn, C. Fine-Tuning Language Models for Factuality. *arXiv* **2023**. arXiv:2311.08401 [cs.CL], https://doi.org/10.48550/arXiv.2311.08401.
19. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 33, pp. 9459–9474.

 doi:10.20944/preprints202605.0895.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.