ICLR 2026

Spilled Energy in Large Language Models

A training-free, energy-based framework for hallucination detection in LLMs

Adrian R. Minut · Hazem Dewidar · Iacopo Masi

Sapienza University of Rome · OmnAI Lab · GLADIA

github.com/OmnAI-Lab/spilled-energy/

The Hallucination Problem

What are Hallucinations?

Hallucinations are any form of error produced by an LLM, including:

- Factual mistakes
- Biased outputs
- Commonsense reasoning failures
- Unverifiable claims

Limitations of Existing Methods

Probe Classifiers

Require training per task; don't generalize across datasets

Logit Confidence

Weak baseline; degrades with instruction tuning

p(true)

Marginally better than random guessing (~51%)

→ **We need a training-free, task-agnostic hallucination detector**

Core Idea: LLMs as Energy-Based Models

Energy-Based Models (EBMs)

EBMs assign a scalar energy to each configuration: lower energy = higher likelihood.
The probability distribution is: $\mathrm{p(x)} = \exp(-\mathrm{E(x)}) \,/\, \mathrm{Z}$ where Z is the partition function.

1 Autoregressive LLMs

LLMs model $p(x_{N:1})$ as a chain of conditional probabilities: $\prod_{i=1}^{N} p(x_i|x_{i-1:1})$

Each step is a softmax classifier over the vocabulary V.

2 Reinterpret as EBMs

Each softmax classifier can be recast as an EBM ratio: $\log p(x_i|x_{i-1:1}) = -\mathrm{E}_{\mathrm{i:1}}^{\ell} + \mathrm{E}_{\mathrm{i-1:1}}^{m}$

3 Discover the Spill

Across consecutive steps, **these two energy quantities should be equal**, but they differ in practice since there is **no training constraint**. We call this discrepancy *spilled energy*.

Core Idea: LLMs as Energy-Based Models

Energy-Based Models (EBMs)

EBMs assign a scalar energy to each configuration: lower energy = higher likelihood.
The probability distribution is: $\mathrm{p}(\mathrm{x}) = \exp(-\mathrm{E}(\mathrm{x})) \,/\, \mathrm{Z}$ where Z is the partition function.

Language Modeling as Chain of Probabilities

$$\mathrm{p}(x_{\mathrm{i:1}}) = \exp(-\mathrm{E}(x_{\mathrm{i:1}})) \,/\, \mathrm{Z}$$

$$p(x_{N:1}) = \prod_{i=1}^{N} p(x_i|x_{i-1:1}) = p(\mathrm{x}_1)p(x_2|x_1)p(x_3|x_2,x_1) \ldots$$

$$\log \mathrm{p}(\mathrm{x}_3,\mathrm{x}_2,\mathrm{x}_1) = \log \mathrm{p}(\mathrm{x}_1) + \log p(x_2|x_1) + \log p(x_3|x_2,x_1)$$

$$\log \mathrm{p}(\mathrm{x}_3,\mathrm{x}_2,\mathrm{x}_1) = -\mathrm{E}(\mathrm{x}_1) - \log \mathrm{Z} - \mathrm{E}(\mathrm{x}_2,\mathrm{x}_1) - \log \mathrm{Z} + \mathrm{E}(\mathrm{x}_1) + \log \mathrm{Z} - \mathrm{E}(\mathrm{x}_3,\mathrm{x}_2,\mathrm{x}_1) - \log \mathrm{Z} + \mathrm{E}(\mathrm{x}_2,\mathrm{x}_1) + \log \mathrm{Z} = -\mathrm{E}(\mathrm{x}_3,\mathrm{x}_2,\mathrm{x}_1) - \log \mathrm{Z}$$

Core Idea: LLMs as Energy-Based Models

Energy-Based Models (EBMs)

EBMs assign a scalar energy to each configuration: lower energy = higher likelihood.
The probability distribution is: $\mathrm{p}(\mathrm{x}) = \exp(-\mathrm{E}(\mathrm{x})) \,/\, \mathrm{Z}$ where Z is the partition function.

Language Modeling in Practice

$$\mathrm{p}(x_{\mathrm{i:1}}) = \exp(-\mathrm{E}(x_{\mathrm{i:1}})) \,/\, \mathrm{Z}$$

$$p(x_i|x_{i-1:1}) = \frac{\exp(\theta(x_{\mathrm{i-1:1}})[id(x_i)])}{\sum_{\mathrm{k}} \exp(\theta(x_{\mathrm{i-1:1}}))[\mathrm{k}]} = [\exp(-\mathrm{E}(x_{\mathrm{i-1:1}}) + \mathrm{E}(x_{\mathrm{i:1}}))]\, {}^{\mathrm{Z}}/_{\mathrm{Z}}$$

$$\mathrm{E}^{\ell}(x_{\mathrm{i:1}}) = \theta(x_{\mathrm{i-1:1}})[id(x_i)]$$
$$\mathrm{E}^{m}(x_{\mathrm{i-1:1}}) = \log \sum_{\mathrm{k}} \exp\left(\theta\left(x_{(i-1:1)}\right)\right)[\mathrm{k}]$$

What is Spilled Energy?

Definition 4.1 — Spilled Energy $\Delta E(x_{i:1})$

The discrepancy between two energy values that, in principle, should be equal but are measured at different time steps and from different components of the LLM softmax:

$$\Delta E(x_{i:1}) = -\log \sum \exp(\theta(x_{i:1}))[k] + \theta(x_{i-1:1})[id(x_i)]$$

E^{ℓ} — Logit Energy

$$E^{\ell} = \theta(x_{i-1:1})[id(x_i)]$$

Measured at step i:
The **logit** of the **sampled token**, i.e., the numerator of the softmax.

This is the energy assigned to the chosen token sequence.

E^{m} — Marginal Energy

$$E^{m} = \log \sum_{k} \exp(\theta(x_{i:1}))[k]$$

Measured at step $i+1$:
The **log-sum-exp** of **ALL logits**, i.e., the denominator of the softmax.

This marginalizes over all possible next tokens.

Two Training-Free Metrics

Δ Spilled Energy ΔE

Two consecutive steps

Captures the discrepancy between energy values that should theoretically be equal across consecutive generation steps.

ΔE is large → model internally inconsistent → likely hallucinating.

$$\Delta E(x_{i:1}) = \theta(x_{i-1:1})[\mathrm{id}(x_i)] - \log \Sigma \exp(\theta(x_{i:1})[k])$$

E Marginal Energy E^m

Single step

Measurable at a single decoding step.
Corresponds to the log-partition function: how "spread out" the probability mass is over the entire vocabulary.

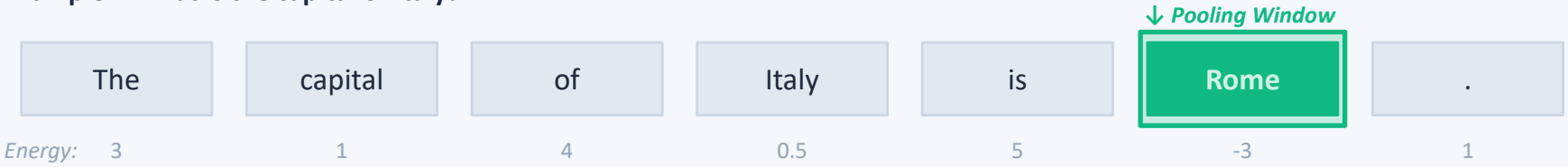
Low marginal energy → model is more certain → likely correct.

$$E^m(x_{i-1:1}) = -\log \Sigma \exp(\theta(x_{i-1:1})[k])$$

Key Insight: Exact Answer Token Localization

Detection is most reliable when restricted to the "exact answer" tokens:
those carrying the semantic content of the answer (e.g., "Rome" in "The capital of Italy is Rome").

Example: "What is the capital of Italy?"



Answer Extraction Strategies

Heuristic Matching

For classification & multiple-choice tasks with a closed label set: string matching.

LLM-based Extraction

For open-ended tasks (TriviaQA, Math):
an auxiliary Mistral-7B model extracts the short answer from the full generation.

Pooling Strategy

When the answer spans multiple tokens, apply min / max / mean pooling over the localized span to get a final score.

★ **Using exact answer tokens provides +24% AuROC improvement for spilled energy vs. full sequence evaluation**

Experiment 1: Synthetic Arithmetic

Setup: Multi-digit arithmetic (13-digit integers) with correct vs. deliberately wrong answers. Tested on LLaMA-3, Qwen-3, and Mistral-7B.

Easy

Offset [1,000–10,000]

Large errors — clearly wrong answers. Spilled energy cleanly separates correct vs. incorrect distributions.

Medium

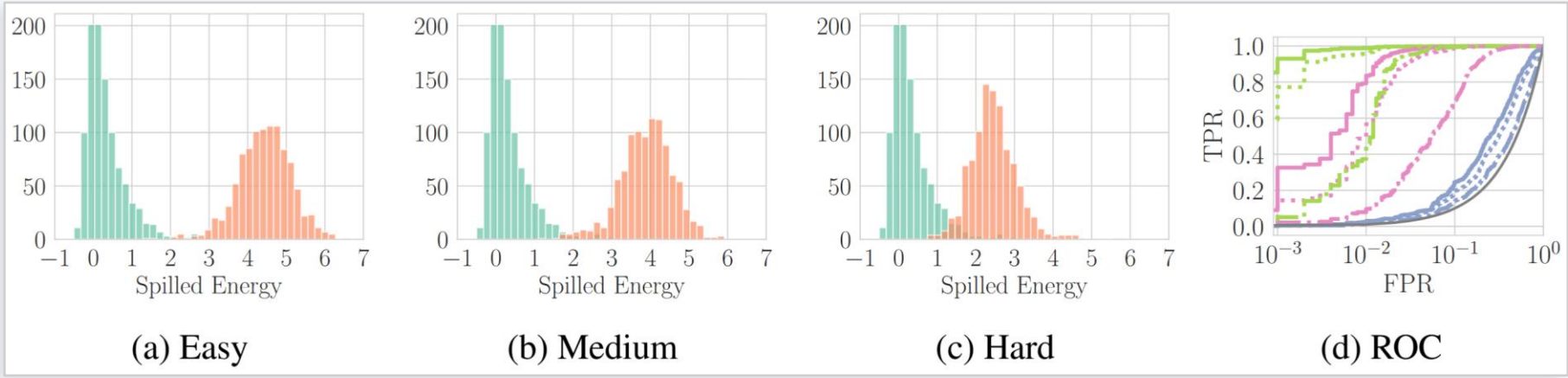
Offset [100–1,000]

Moderate errors. Spilled energy still maintains strong separation even as the gap shrinks.

Hard

Offset [1–10]

Subtle errors — hardest to detect. Spilled energy outperforms logit-based detection at this difficulty.



Experiment 2: Cross-Dataset Generalization

Probing classifiers degrade sharply out-of-distribution, while spilled energy generalizes without any retraining.

	Pool	HotpotQA	HotpotQA-WC	IMDB	Math	MNLI	Movies	TriviaQA	Winobias	Winogrande	Average
LLaMA-Instruct Dubey et al. (2024)											
p(true)	—	58.31±0.32	51.66±1.05	50.72±1.20	49.53±2.16	52.33±0.98	59.30±0.85	45.99±0.51	45.47±1.58	48.33±0.68	51.29±04.86
Orgad et al.	Mean	66.56±9.10	59.00±8.14	*69.78*±14.76	*66.56*±17.04	60.56±12.53	66.44±8.06	63.22±11.11	**67.33**±11.97	**58.00**±7.79	64.16±03.90
Logit E^{ℓ}	Max	72.85±2.12	91.11±1.52	42.08±5.07	57.81±3.82	25.52±3.00	43.97±1.38	68.89±1.96	39.95±2.41	49.40±2.16	54.62±18.97
Marginal E^{m}	Max	*76.72*±1.38	30.74±3.45	**85.63**±2.39	27.08±5.06	**89.90**±1.25	**96.17**±0.63	*80.13*±1.87	57.67±2.94	47.47±1.83	*65.72*±24.39
Marginal E^{m}	Min	75.91±1.62	**97.57**±0.75	14.37±2.39	**70.55**±2.43	61.21±3.24	72.21±1.60	73.38±1.86	47.19±2.71	53.98±2.30	62.93±21.89
Spilled ΔE_s	Max	53.65±1.40	36.28±2.99	55.80±4.32	35.44±3.41	58.81±2.58	70.30±1.49	48.70±2.44	36.53±2.98	44.32±1.70	48.87±11.26
Spilled ΔE	Min	**85.98**±1.09	*93.00*±1.61	47.66±4.06	65.58±3.02	*73.95*±1.97	*89.34*±1.04	**87.07**±1.33	*60.72*±2.74	*55.11*±2.05	**73.16**±15.64
LLaMA Dubey et al. (2024)											
p(true)	—	52.83±0.71	49.33±0.86	52.30±0.58	58.63±1.26	53.78±0.70	60.76±0.69	62.94±0.51	50.02±1.24	53.47±0.54	54.90±04.77
Orgad et al.	Mean	61.22±9.95	56.78±8.70	**72.67**±13.91	*69.67*±15.07	60.33±13.77	*64.00*±8.40	66.44±8.20	**60.89**±12.60	**53.56**±4.36	*62.84*±05.71
Logit E^{ℓ}	Max	53.47±2.13	49.02±1.79	48.27±1.32	57.38±6.09	91.76±0.91	57.42±1.43	52.77±2.58	50.74±1.51	51.17±1.83	56.89±12.70
Marginal E^{m}	Max	*78.00*±1.30	*76.90*±1.09	48.29±1.16	68.77±8.33	10.93±1.42	**80.70**±1.98	67.49±1.69	51.91±2.32	*51.28*±2.47	59.36±20.69
Marginal E^{m}	Min	58.39±2.79	59.20±1.95	*51.71*±1.16	34.13±8.78	*97.42*±0.51	50.37±2.43	*69.88*±1.40	49.05±2.20	49.00±2.30	57.68±16.75
Spilled ΔE_s	Min	77.75±1.52	79.44±2.05	43.39±1.82	*72.87*±6.10	**99.97**±0.08	61.56±2.95	77.55±1.62	*52.34*±2.57	48.17±1.62	68.12±17.15
Spilled ΔE	Min	**79.04**±1.78	**80.83**±1.87	43.22±1.67	**74.36**±5.54	**99.97**±0.08	61.97±2.81	**78.54**±1.57	*52.11*±2.58	48.21±1.62	**68.69**±17.48

Key Results & Findings

73.2%
Avg AuROC
LLaMA-3-Instruct

77.5%
Avg AuROC
Mistral-Instruct

+24%
Boost from
exact answer tokens

0
Additional
training needed

Instruction Tuning Helps

While fine-tuning degrades logit confidence (54.6%), spilled energy improves with instruction tuning (68.7% → 73.2% for LLaMA-3).

Known Limitation

False positives occur on semantically uninformative tokens (punctuation, sentence-initial words) — confirming the need for exact answer localization.

Generalizes Across Models

Results hold for LLaMA-3, Mistral, Gemma 1B & 4B — both pretrained and instruction-tuned — without any architecture-specific modifications.

Min Pooling is Best

Across all methods and models, min pooling over the exact answer span consistently outperforms max, mean, and last-token strategies.

Conclusion

Spilled Energy in Large Language Models — ICLR 2026

01 New EBM Perspective

The LLM softmax is reinterpreted as an Energy-Based Model, revealing a principled connection between autoregressive generation and energy dynamics.

02 Training-Free Detection

Spilled energy requires no additional training or probing classifiers — it reads directly from output logits, making it instantly deployable.

03 Superior Generalization

Consistently outperforms probe classifiers under cross-dataset evaluation (73% vs 62%) and surpasses logit confidence across all tested LLMs.

04 Future Direction

Open questions: real-time deployment at scale, deeper understanding of false positives, and extension to multi-modal and longer-context generation.

Code: github.com/OmnAI-Lab/spilled-energy/ · Paper: arXiv:2602.18671