ENERGY LANDSCAPES OF TRUTHFULNESS IN LLM ATTENTION

Sohum Mehta*
Illinois Mathematics and Science Academy
smehta@imsa.edu

Shiv Narayan Rai*
Illinois Mathematics and Science Academy
srai@imsa.edu

ABSTRACT

We study whether large language model (LLM) truthfulness is reflected in an internal "energy landscape" derived from modern Hopfield networks and associative memory theory. Leveraging the established connection between transformer attention and continuous Hopfield retrieval (Ramsauer et al., 2021; Millidge et al., 2022), we operationalize a Hopfield-style energy proxy directly from pre-softmax attention logits under teacher-forced evaluation. We evaluate 300 TruthfulQA questions (Lin et al., 2022), pairing each prompt with a truthful reference answer (`best_answer`) and a false reference answer (`incorrect_answers[0]`). On Qwen2.5-0.5B-Instruct, truthful answers exhibit systematically lower mean Hopfield energy (paired Cohen's $d_z = -0.27$, Wilcoxon $p < 0.001$), and larger retrieval margins ($d_z = +0.24$, $p < 0.001$). A lightweight logistic regression probe over 15 energy-derived features achieves 0.66 AUC (5-fold stratified CV with question-grouped folds), indicating that attention-logit energy carries non-trivial truthfulness signal without fine-tuning the base model (only a linear probe on frozen features). Layer-wise trajectories show that energy separation emerges primarily in early layers (0–8), consistent with early blocks implementing the critical associative pattern-matching step. Together, these findings connect associative memory energy proxies to practical truthfulness discrimination and suggest a concrete mechanistic direction for hallucination monitoring.

1 INTRODUCTION

Hallucination and imitative falsehoods are persistent failure modes of LLMs. A common informal description is that models "retrieve" plausible but incorrect content rather than the correct facts. This view becomes mechanistically concrete under the modern Hopfield interpretation of attention, where attention acts as associative retrieval over key–value memories (Ramsauer et al., 2021; Millidge et al., 2022). This paper asks a targeted question:

> **Do pre-softmax attention logits carry an energy signature that differentiates truthful from false candidate answers under the same prompt?**

We compute Hopfield-style energy and related retrieval metrics from pre-softmax attention logits and test whether they separate truthful versus false *reference* answers on TruthfulQA (Lin et al., 2022) under teacher forcing. Our analysis is complementary to representation-level truth probes (Burns et al., 2023) and to memory-augmentation approaches that improve groundedness (Modarressi et al., 2024).

Contributions.

1. **Associative-memory lens on truthfulness.** We analyze truthfulness using Hopfield-style energy proxies computed from attention logits.
2. **Pre-softmax metric suite.** We compute energy, entropy, margin, and cross-layer drift metrics from inference-time attention-logit hooks, requiring no LLM fine-tuning.

*Equal contribution.

3. **Empirical separation.** On 300 TruthfulQA prompt-level pairs, we find significant paired differences for energy, entropy, margin, and an energy-slope feature ($d_z \approx 0.24$–0.32; Wilcoxon $p < 0.001$).
4. **Lightweight probe.** A logistic regression probe over 15 features reaches 0.66 ± 0.02 AUC (5-fold CV; question-grouped).

2 METHOD

2.1 ATTENTION AS ASSOCIATIVE RETRIEVAL

Modern continuous Hopfield networks provide an energy-based view of associative retrieval (Ramsauer et al., 2021). For query token t, attention logits are

$$A_t = \frac{Q_t K^\top}{\sqrt{d_k}} + \text{mask}, \qquad \alpha_t = \text{softmax}(A_t). \tag{1}$$

Following the Hopfield framing, we define a lightweight energy proxy

$$E_t = -\text{logsumexp}(A_t). \tag{2}$$

This proxy summarizes the scale and competition structure of logits. We treat the direction of attention entropy as an open question: concentrated attention (low entropy) may reflect decisive retrieval, while broader attention (high entropy) may reflect distributed grounding over multiple relevant cues.

2.2 METRICS FROM PRE-SOFTMAX LOGITS

For each head and query position (restricted to answer-token positions), we compute:

$$\textbf{Energy: } E_t = -\text{logsumexp}(A_t) \tag{3}$$

$$\textbf{Entropy: } H_t = -\sum_i \alpha_{t,i} \log \alpha_{t,i} \tag{4}$$

$$\textbf{Margin: } M_t = A_t[\text{top1}] - A_t[\text{top2}] \tag{5}$$

$$\textbf{KL drift: } D_t^\ell = \text{KL}\big(\alpha_t^\ell \,\|\, \alpha_t^{\ell+1}\big)\,. \tag{6}$$

We average each metric across heads and answer-token positions. We then derive layer-aggregated features (mean across layers, slope across layers, and select summaries) to form 15 total features for the probe. In Hopfield derivations that include a query-norm term, we drop it as approximately constant under paired comparisons sharing the same prompt.

2.3 EXPERIMENTAL SETUP

Dataset: TruthfulQA generation split (Lin et al., 2022); 300 questions. **Pairing:** teacher-forced evaluation of (i) `best_answer` (truthful) and (ii) `incorrect_answers[0]` (false). **Model:** Qwen2.5-0.5B-Instruct. **Logging:** pre-softmax attention logits captured via eager attention and inference-time hooks. **Controls:** max 64 answer tokens; deterministic seeds.

2.4 ANALYSIS

We report paired Cohen's d_z and Wilcoxon signed-rank tests. For classification, we train a logistic regression probe on the 15 frozen features using 5-fold stratified CV with question-grouped folds (preventing within-question leakage between truthful/false pairs). For layer-wise plots, per-layer p-values are treated as descriptive (see Limitations).

3 RESULTS

3.1 LAYER-WISE ENERGY SEPARATION

Figure 1 shows the paired difference (truthful minus false) across layers for energy and entropy. Energy separation emerges early, with the largest differences concentrated in layers 0–8, suggesting early blocks contain the strongest associative retrieval signature.

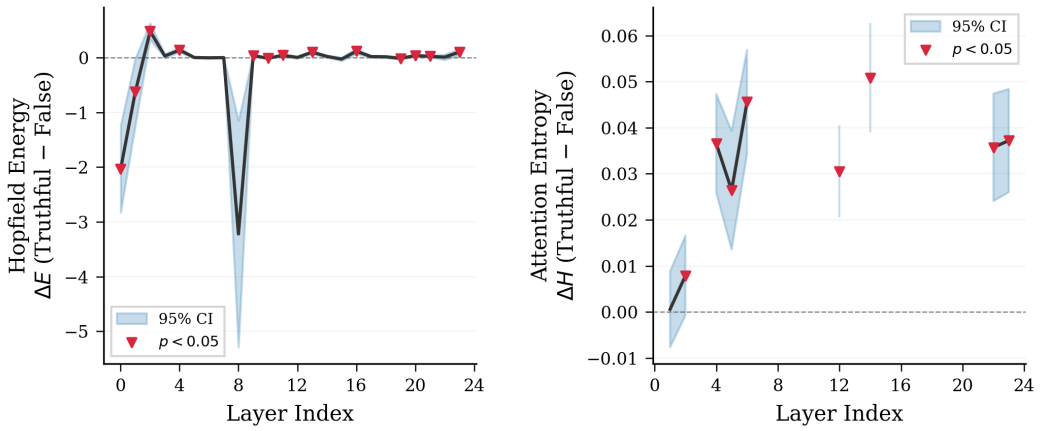

Figure 1: **Layer-resolved separation.** Paired difference (truthful minus false) in Hopfield energy (left) and attention entropy (right) across layers on 300 TruthfulQA pairs. Energy separation concentrates in early layers (0–8).

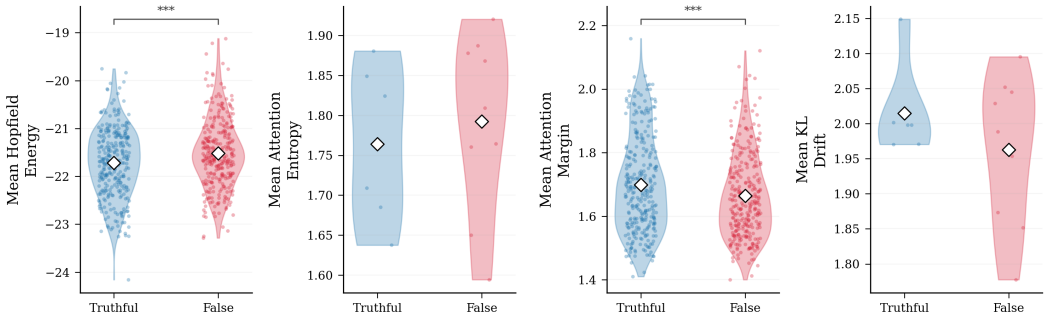

Figure 2: **Feature distributions.** Four-panel violin plot comparing *mean* Hopfield energy, mean attention entropy, mean retrieval margin, and mean KL drift between truthful and false candidate answers (teacher forced).

3.2 Feature distributions and summary statistics

Figure 2 compares distributions for four representative metrics. Table 1 summarizes paired effect sizes and probe performance.

Interpretation. Lower mean energy for truthful answers supports an energy-based separation. The finding that truthful answers have higher entropy suggests truthfulness may rely on integrating multiple cues (distributed retrieval) rather than collapsing to a single narrow association. The steeper energy slope, together with early-layer separation, localizes the effect to early blocks. KL drift is not supported by this proxy in our setting.

4 Related Work

Transformers as associative memories. Modern Hopfield networks formalize attention as single-step associative retrieval with an energy function (Ramsauer et al., 2021). Universal Hopfield Networks generalize this retrieval/energy view beyond dot-product similarity, motivating multi-metric analyses (Millidge et al., 2022). **Truthfulness benchmarks.** TruthfulQA targets imitative falsehoods and provides reference answers for controlled comparison (Lin et al., 2022). **Truth signals in internals.** CCS extracts truth-related directions in activations without labeled supervision (Burns et al., 2023); our approach instead uses attention-logit energy proxies. **Memory and groundedness.** Explicit read–write memory modules can improve factual use of stored knowledge (Modarressi et al.,

Metric	paired d_z	Wilcoxon p	Direction
Energy (mean)	-0.27	< 0.001	truthful lower
Energy (slope)	+0.32	< 0.001	truthful steeper
Entropy (mean)	+0.26	< 0.001	truthful higher
Margin (mean)	+0.24	< 0.001	truthful larger
KL drift (mean)	-0.11	0.80	n.s.
Probe AUC	0.66 ± 0.02 (5-fold CV; 15 features)		

Table 1: **Summary statistics and probe performance.**

2024), aligning with the hypothesis that retrieval quality is mechanistically linked to truthfulness.
Uncertainty-based detection. Semantic-uncertainty approaches detect hallucinations via meaning-level dispersion (Farquhar et al., 2024); our metrics are single-pass signals from attention logits under teacher forcing.

5 Discussion and Conclusion

We presented an energy-proxy analysis of attention logits motivated by modern Hopfield networks and tested whether these proxies separate truthful from false candidate answers under paired, teacher-forced conditions. Across 300 TruthfulQA prompt-level pairs, we find significant differences in mean Hopfield energy, entropy, margin, and energy slope. A lightweight linear probe on 15 frozen features reaches 0.66 AUC, showing that attention-logit energy carries measurable truthfulness signal without fine-tuning the base model.

Limitations. (1) Teacher forcing vs. free generation: decoding-time dynamics may differ. (2) Truthfulness proxy: false candidates are benchmark-provided incorrect references, not necessarily model-generated hallucinations. (3) Scope: one dataset and one 0.5B model. (4) Multiple comparisons: per-layer significance markers should be interpreted cautiously without correction.

Future work. Monitor energy proxies during autoregressive decoding; test energy-guided decoding; evaluate cross-dataset and cross-scale generalization; and study how retrieval augmentation shifts the energy landscape.

References

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, et al. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (Long Papers)*, pp. 3214–3252, may 2022. doi: 10.18653/v1/2022.acl-long.229.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 15561–15583, 2022.

Ali Modarressi, Abdullatif Köksal, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Memllm: Finetuning llms to use an explicit read-write memory. *CoRR*, abs/2404.11672, 2024. doi: 10.48550/ARXIV.2404.11672.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David

Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations (ICLR)*, 2021.