Can Linear Attributions Explain Nonlinear LLMs?

Association for the Advancement of Artificial Intelligence

UF

An Entropy-Based Framework for Attention-Wise Usable Information in LLMs

Vishal Pramanik[1] • Alvaro Velasquez[2] • Susmit Jha[3] • Sumit Kumar Jha[1]

[1]University of Florida • [2]UC Boulder • [3]SRI International

1. Abstract

Attribution methods aim to explain language model predictions by scoring how much each input token contributes to a generated output, but many existing approaches rely on linear approximations and are not well-suited to the causal and semantic structure of autoregressive decoder-only LMs. We propose **HEAT**, which unifies **target-conditioned semantic transition influence**, **Hessian-based sensitivity**, and **KL-divergence information loss under token masking** to produce context-aware and causally faithful attributions. Across multiple models and datasets (and a new curated benchmark), HEAT consistently improves attribution faithfulness and alignment with human annotations.

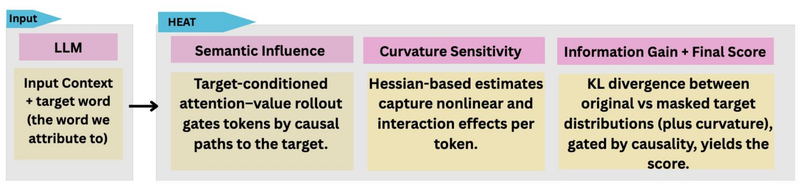
2. WHY CURRENT METHODS FAIL

Gradient-Based Methods Fail:
- Vanish in flat regions (ReLU saturation)
- Miss second-order effects
- Example: $\partial f/\partial x_i = 0$ but $f(x+\varepsilon) \neq f(x)$

Attention Weights ≠ Causal Influence:
- Show "where model attends" not "what affects output"
- Can be perturbed without changing predictions
- Ignore multi-hop influence through residual connections

First-Order Methods (IG, Input×Gradient):
- Capture only local linear sensitivity
- Baseline-dependent and path-dependent
- Underestimate influence near sharp transitions

3. HEAT METHOD OVERVIEW



HEAT
Hessian-Enhanced Attribution

1 Semantic Flow (M_T)

Traces causal attention paths to target token

- Target-conditioned
- Enforces causality
- Layer-wise rollout

2 Hessian Sensitivity (S_i)

Captures nonlinear curvature via 2nd-order

- Second-order effects
- Scalable HVP
- Nonlinear interactions

3 KL Information (I)

Measures distribution change when masked

- $D_{KL}(P_{orig} \,||\, P_{masked})$
- Token importance
- Info-theoretic

ATTRIBUTION FORMULA

$$\text{Attr}(x_i \to x_T) = M_T[i] \cdot (\beta S_i + \gamma I$$

4. DATASETS & SETUP

Benchmark Datasets:
- LongRA: Long-range agreement task
- TellMeWhy: Narrative reasoning (causal QA)
- WikiBio: Biography generation

Curated Dataset (NEW):
- 2,000 mixed paragraphs
- NarrativeQA + SciQ combined
- Gold annotations from GPT-4o & GPT-5
- Inter-annotator agreement: F1=0.91, κ=0.89

Models Tested:
- GPT-J 6B
- Phi-3-Medium 14B
- LLaMA-3.1 70B
- Qwen2.5 3B

5. MAIN RESULTS

GPT-J 6B Results:

	LongRA	TellMeWhy
HEAT (Ours)	10.3 / 2.31	9.2 / 2.04
ReAgent (best)	1.68 / 0.37	1.45 / 0.36
	↑ 2× better	↑ 2× better

Key Findings:
✓ 8-13% improvement in AUROC (Soft-NC)
✓ 10-15% improvement in correlation (Soft-NS)
✓ Consistent across ALL models and datasets
✓ DSA Score: 4.80 (vs 3.60 for ReAgent)

Gradient-based methods often yield
NEGATIVE Soft-NS ⟶ unstable!

6. ABLATION STUDIES

All Three Components Are Essential

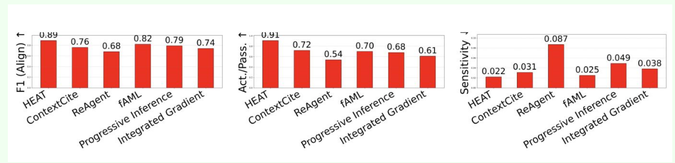
Removing ANY component degrades performance:

Full HEAT:	9.78 / 2.31 / 4.70
Transition Only:	3.12 / 1.52 / 2.21
Hessian Only:	2.89 / 1.45 / 2.97
KL Only:	2.23 / 1.21 / 2.74
No Transition Gating:	4.31 / 1.84 / 1.68
Uniform Transition:	3.89 / 1.76 / 1.54

Why Each Component Matters:
- Transition: Routes semantic paths
- Hessian: Captures curvature effects
- KL: Measures information impact

The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy dog

7. ROBUSTNESS ANALYSIS



HEAT is Robust to:

1. Input Perturbations
   Sensitivity: 0.025 (lowest among all methods)

2. Syntactic Rephrasings
   Active/Passive Robustness: 0.91 (highest)

3. Decoding Hyperparameters
   Max variation across temp/top-p/top-k: < 1%
   (Baselines vary 2-5%)

4. Alignment with Human Annotations
   F1 Score: 0.89 vs 0.82 for best baseline

8. QUALITATIVE EXAMPLES

HEAT Identifies Semantically Relevant Tokens

Example 1: Predicting "slice"
Context: "ordered pizza...did not cut...knife to cut"
HEAT highlights: pizza, cut, knife ✓

Example 2: Predicting "friends"
Context: "zoo...took pictures and shared them"
HEAT highlights: pictures, shared, zoo ✓

Example 3: Predicting "bush"
Context: "lost my hat at the park...stuck in a"
HEAT highlights: hat, park, stuck ✓

9. COMPUTATIONAL EFFICIENCY

Efficiency vs. Accuracy Trade-offs

HEAT is slower but more accurate:
- Full HEAT: 455s per 1,000 examples
- Baselines: 2-10s per 1,000 examples

Approximation Strategies:
✓ Low-Rank Hessian (rank=64): 330s, -2% AOPC
✓ Layer Sampling (6 layers): 305s, -4% AOPC
✓ Windowing (512 tokens): 295s, -6% AOPC
✓ LR+WIN (recommended): 245s, -3% AOPC

Long Context (2048 tokens):
- Full HEAT: 1,230s
- LR+WIN: 580s (2× faster, -1% AOPC)

10. CONCLUSIONS

Key Contributions:

✓ HEAT: First attribution method integrating semantic flow, Hessian curvature, and KL divergence

✓ Superior Performance: 2× better than state-of-the-art across all benchmarks and model scales

✓ Theoretical Guarantees: Formal faithfulness bounds and convergence analysis

✓ Curated Benchmark: 2,000 annotated instances with high inter-annotator agreement (F1=0.91)

✓ Scalable: Efficient approximations for 70B models and 100K token contexts

Advantages:
- Model-agnostic (works on any decoder-only LLM)
- No training required
- Interpretable multi-view decomposition
- Robust to decoding hyperparameters

📧 vishalpramanik@ufl.edu • 🌐 Paper: AAAI 2026 • 💻 Code: Available Soon