# Hessian-Enhanced Token Attribution (HETA): Interpreting Autoregressive Language Models

**Vishal Pramanik**
Department of Computer
& Information Science & Engineering
University of Florida
Gainesville, FL 32611
vishalpramanik@ufl.edu

**Maisha Maliha**
School of Computer Science
University of Oklahoma
Norman, Oklahoma 73019
maisha.maliha-1@ou.edu

**Sumit Kumar Jha**
Department of Computer
& Information Science & Engineering
University of Florida
Gainesville, FL 32611
sumit.jha@ufl.edu

## Abstract

Attribution methods seek to explain language model predictions by quantifying the contribution of input tokens to generated outputs. However, most existing techniques are designed for encoder-based architectures and rely on linear approximations that fail to capture the causal and semantic complexities of autoregressive generation in decoder-only models. To address these limitations, we propose **Hessian-Enhanced Token Attribution (HETA)**, a novel attribution framework tailored for decoder-only language models. HETA combines three complementary components: a semantic transition vector that captures token-to-token influence across layers, Hessian-based sensitivity scores that model second-order effects, and KL divergence to measure information loss when tokens are masked. This unified design produces context-aware, causally faithful, and semantically grounded attributions. Additionally, we introduce a **curated benchmark dataset** for systematically evaluating attribution quality in generative settings. Empirical evaluations across multiple models and datasets demonstrate that HETA consistently outperforms existing methods in attribution faithfulness and alignment with human annotations, establishing a new standard for interpretability in autoregressive language models.

## 1 Introduction

As machine learning systems achieve increasingly high performance, they are being deployed in high-stakes domains such as healthcare, autonomous driving, and finance. However, despite their success, deep neural networks remain difficult to interpret due to their large parameter spaces, layered architectures, and nonlinear computations, earning them the reputation of "black box" models [1]. This opacity can erode trust, impede debugging, and raise ethical or regulatory concerns. To address these challenges, the field of Explainable AI (XAI) has emerged, with the goal of making model decisions more transparent, interpretable, and trustworthy.

A wide range of interpretability methods such as LIME [2], KernelSHAP [3], Integrated Gradients [4], Grad-CAM [5], and Layer-wise Relevance Propagation (LRP) [6] have been developed under the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundations of Reasoning in Language Models.

classical feature attribution paradigm, which aims to quantify the contribution of input features to a model's output. Most of these methods are based on linear or first-order derivative approximations and assume local model linearity. However, this assumption often breaks down in the context of autoregressive language models, where token interactions are nonlinear and highly contextual. Despite their practical utility, these techniques frequently produce inconsistent attributions for the same input and model [7], casting doubt on their reliability. Although some efforts have introduced axiomatic foundations to formalize attribution [8, 9], a universally accepted definition of explanation quality remains elusive. Furthermore, these attribution methods have primarily been designed for encoder-based architectures, and recent work [10] shows that directly applying them to decoder-only language models in generative tasks is non-trivial and often unfaithful. The discrepancy arises from architectural and functional differences, where encoder models leverage bidirectional attention and require a single attribution map, while decoder-only models generate outputs autoregressively and demand attribution at each token position. Figure 1 illustrates the complexity of the attribution task for a generative model, highlighting how input words contribute to the generation of a specific output word. Although model-agnostic approaches have been proposed for generative settings [10], they typically ignore the dense semantic structure encoded in the internal layers of large language models [11, 12], thereby limiting their ability to capture deep token-level influence.

To address the shortcomings of gradient-based attribution methods in autoregressive models, we propose **Hessian-Enhanced Token Attribution (HETA)**, a framework tailored for decoder-only architectures. HETA combines semantic flow tracing, Hessian-based sensitivity, and KL-based information loss to yield faithful, token-level attributions. By modeling attention-weighted value flow and capturing second-order and informational effects, HETA offers a principled and robust alternative that respects the causal and contextual structure of generative language models. The key contributions of this work are:

- We propose **Hessian-Enhanced Token Attribution (HETA)**, a novel attribution method for decoder-only language models that integrates semantic flow for causal directionality, Hessian-based sensitivity for capturing second-order interactions, and KL-divergence for quantifying information-theoretic impact.
- We construct and release a new curated dataset specifically designed for evaluating token-level attributions in autoregressive generation tasks, enabling systematic assessment of attribution faithfulness, robustness, and human alignment.
- We conduct extensive experiments across multiple decoder-only models and strong attribution baselines, showing that HETA consistently achieves higher faithfulness, robustness, and semantic alignment, outperforming existing methods by a significant margin.

## 2 Motivation

Understanding which input tokens influence a model's output is crucial for interpreting generative transformers. However, existing attribution methods like gradients and attention alone are fundamentally limited in capturing the full scope of token influence. Attribution methods based solely on attention weights [13] are known to be unreliable. While attention indicates where the model "looks," it does not necessarily reflect what influences the output. [14] demonstrated that attention weights can be perturbed without affecting model predictions, highlighting that attention is not a faithful explanation. Moreover, attention-based methods often miss indirect or multi-hop influence paths using skip connections as shown in [15], and in decoder-only models, they can violate causality by attributing importance to future tokens. Since attention lacks sensitivity to output changes and higher-order interactions, it should not be treated as a standalone attribution method.

First-order attribution methods, such as InputXGradient[16], DIG[17] and Integrated Gradients[4], estimate the influence of input tokens by measuring the local sensitivity of the model output with respect to input features. While these methods are computationally efficient and widely used, they suffer from a fundamental limitation: *they only capture local, linear (first-order) effects*, and can *fail to detect influence when the function's curvature is nonzero but the gradient is zero*.

To define it mathematically, let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function (e.g., $f(x) = \log P(x_T \mid x_{<T})$ in an autoregressive language model). Then, there exist inputs $x \in \mathbb{R}^n$ and

coordinates $i \in \{1, \ldots, n\}$ such that:

$$\frac{\partial f(x)}{\partial x_i} = 0 \quad \text{but} \quad \exists\, \epsilon > 0 \text{ such that } f(x + \epsilon e_i) \neq f(x)$$

To explain this, consider the second-order Taylor expansion of $f(x)$ around a reference point $x_0 \in \mathbb{R}^n$:

$$f(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)(x - x_0)$$

for some $\xi \in [x_0, x]$ on the segment between $x_0$ and $x$. If $\nabla f(x_0) = 0$, the gradient term vanishes:

$$f(x) - f(x_0) = \frac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)(x - x_0)$$

Thus, the function can still change solely due to the second-order curvature encoded in the Hessian, even when the gradient is zero. This illustrates that first-order methods like gradients or IG can miss such influence entirely. As an example, consider a simple scalar function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x) = \text{ReLU}(w^\top x + b), \quad w = [1, 1]^\top, \quad b = -2$$

Let $x_0 = [0, 0]^\top$. Then:

$$w^\top x_0 + b = -2 < 0 \Rightarrow f(x_0) = 0, \quad \nabla f(x_0) = 0$$

Now perturb the input slightly: $x = [2.1, 0]^\top$. Then:

$$w^\top x + b = 0.1 > 0 \Rightarrow f(x) = 0.1 \neq f(x_0)$$

Although the gradient at $x_0$ is zero, the function still responds to the input, meaning that attribution based only on the gradient would miss this influence.

Gradient-based and attention-based attribution methods each capture only a limited aspect of influence in transformer-based generative models. Gradients can vanish in flat or saturated regions, missing non-linear or second-order effects, while attention weights often fail to reflect actual causal or output-sensitive influence (please see Appendix A1 for further details). These shortcomings motivate our Hessian-Enhanced Token Attribution (HETA) framework, which integrates semantic flow, second-order sensitivity, and information-theoretic measures to provide more faithful and robust attributions.

## 3 Background

Understanding token-level influence in transformer models requires going beyond raw attention weights or local gradients. Two complementary strands of research have highlighted important limitations and proposed more robust alternatives. [18] demonstrated that attention weights alone are insufficient for faithful interpretation, as they neglect the scale of the value vectors being attended to. They proposed a norm-based approach that combines attention weights with the magnitude of the transformed value projections, offering a more accurate view of token influence within self-attention. This formulation captures not only alignment (via attention) but also semantic strength (via vector norms), leading to more faithful attributions.

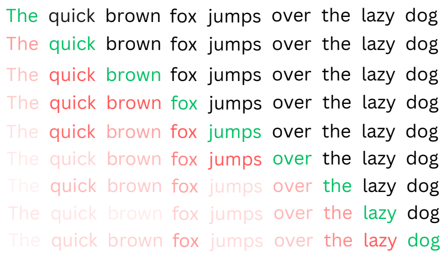


Figure 1: Input importance distributions for a generative task using our proposed HETA method.

Separately, Hessian-based sensitivity methods provide deeper insight into model behavior by accounting for second-order interactions between inputs and outputs. Specifically, the Hessian of the log-likelihood with respect to input embeddings,

$$H_T = \nabla_X^2 \log P(x_T \mid x_{<T}),$$

captures local curvature and reveals how token effects manifest in nonlinear regions of the model's decision surface. Unlike first-order methods which can fail in flat regions or under poor baseline selection, second-order methods remain informative even when gradients vanish. Prior studies ([19], [20], [21]) support the use of Hessian-based approaches to uncover latent influences in deep architectures.

Together, these techniques underscore the importance of considering both semantic flow and higher-order sensitivity to capture faithful token attributions in transformer models.

Table 1: **Attribution alignment on the curated dataset** using the DSA (Dependent Sentence Attribution) metric. Higher DSA scores reflect stronger alignment between attribution and human-annotated tokens. HETA achieves the highest DSA across all models. Mean over 3 independent runs and std $< \pm 0.05$

| **Attribution Method** | **GPT 6B** | **LLaMA-3.1 8B** | **OPT 6.7B** | **Qwen2.5 3B** |
|---|---|---|---|---|
| Input × Gradient | -0.34 | -0.28 | -0.41 | -0.31 |
| Integrated Gradients | -0.12 | -0.09 | -0.18 | -0.14 |
| Gradient SHAP | -0.25 | -0.21 | -0.30 | -0.22 |
| LIME | -0.31 | -0.27 | -0.36 | -0.29 |
| Attention Rollout | -0.44 | -0.39 | -0.52 | -0.41 |
| fAML | 2.10 | 2.30 | 2.05 | 2.20 |
| Progressive Inference | 2.65 | 2.88 | 2.40 | 2.73 |
| SEA-CoT | 2.92 | 3.15 | 2.77 | 2.85 |
| DIG | 2.30 | 2.54 | 2.10 | 2.45 |
| ReAGent | 3.60 | 3.78 | 3.35 | 3.50 |
| **HETA (Ours)** | **4.80** | **5.10** | **4.25** | **4.65** |

| **Variant** | **Soft-NC ↑** | **Soft-NS ↑** | **DSA ↑** |
|---|---|---|---|
| **Full HETA** | **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 |

Table 2: Results averaged across all datasets with GPT-J-6B (higher is better). Mean over 3 runs; std $< \pm 0.2$.

## 4 Our Methodology

We propose **Hessian-Enhanced Token Attribution (HETA)**, a principled framework that integrates these perspectives into a *unified influence decomposition*. Our central view is that token attribution in autoregressive models should estimate a token's *directional causal contribution* to the log-likelihood of the generated token, incorporating both *semantic path dependencies* and *higher-order effects*. HETA achieves this through three complementary components: **(1) Semantic Transition Influence**, which captures how tokens propagate influence through compositional attention-value flows across layers, ensuring causal directionality. **(2) Hessian-Based Sensitivity**, which models second-order curvature of the log-likelihood surface with respect to token embeddings, capturing nonlinear and interaction effects. **(3) Information-Theoretic Impact**, which measures the change in predictive uncertainty when a token is masked, providing a probabilistic interpretation of its contribution. Together, these components form a mathematically grounded attribution score that balances structural, geometric, and information-theoretic perspectives on token influence.

Let $X = (x_1, x_2, \ldots, x_T)$ be an input sequence of tokens, each $x_i \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary. Let $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the input embedding matrix, and $\mathbf{X} = (\mathbf{e}_1, \ldots, \mathbf{e}_T) \in \mathbb{R}^{T \times d}$ be the embedded input. A decoder-only language model $f_\theta$, parameterized by $\theta$, defines the conditional distribution over the next token:

$$P_\theta(x_{T+1} \mid x_{\leq T}) = \text{Softmax}(f_\theta(\mathbf{X}))$$

Our goal is to compute a token-level attribution score $\text{Attr}(x_i \to x_{T+1}) \in \mathbb{R}_{\geq 0}$ quantifying the contribution of token $x_i$ to the prediction of $x_{T+1}$. The attribution is constructed from three components: semantic transition influence, Hessian-based sensitivity, and KL-based information loss.

### 4.1 Semantic Flow for Causal Token Influence

To enforce causal directionality, we model the flow of semantic influence from input tokens to the predicted token through a *transition influence vector*. Rather than relying solely on raw attention weights, which can be misleading [14], we trace *attention-weighted value outputs* across all transformer layers, integrating both alignment (via attention) and semantic strength (via value projections), as suggested by [18]. For each layer $l \in \{1, \ldots, L\}$ and head $h \in \{1, \ldots, H\}$, we compute:

$$\mathbf{z}^{(l,h)} = A^{(l,h)} V^{(l)} W_O \in \mathbb{R}^{T \times d}$$

where $A^{(l,h)} \in \mathbb{R}^{T \times T}$ is the attention matrix, $V^{(l)} \in \mathbb{R}^{T \times d}$ are value vectors, and $W_O \in \mathbb{R}^{d \times d}$ is the learned output projection. Summing across heads and normalizing yields:

$$z_i^{(l)} = \sum_{h=1}^{H} \mathbf{z}_i^{(l,h)}, \quad M_T[i] = \frac{1}{Z} \sum_{l=1}^{L} \|z_i^{(l)}\|_1$$

where $Z = \sum_{j=1}^{T} \sum_{l=1}^{L} \|z_j^{(l)}\|_1$. This gives a semantic transition influence vector $M_T \in \mathbb{R}^T$, modeling token contributions via compositional attention flow.

### 4.2 Hessian-Based Sensitivity Analysis

To capture the second-order influence of each token on the model's prediction, HETA incorporates *Hessian-based sensitivity*, measuring how the curvature of the log-likelihood changes with respect to token embeddings. Let $X \in \mathbb{R}^{T \times d}$ be the input token embeddings. We compute the Hessian of the log-probability of the final token $x_T$ with respect to these embeddings:

$$H_T = \nabla_X^2 \log P(x_T \mid x_{<T}) \in \mathbb{R}^{Td \times Td} \tag{1}$$

For each token $x_i$, we define the sensitivity score as the $\ell_1$ norm of all Hessian entries that correspond to embedding dimensions of $x_i$:

$$S_i^{(T)} = \sum_{j=1}^{Td} |H_T[i \cdot d : (i+1) \cdot d, j]| \tag{2}$$

This scalar score reflects how sensitive the model's output is to perturbations in token $x_i$, capturing both local and joint effects in the input space.

### 4.3 KL Divergence for Information Contribution

To measure how much information each token contributes to the prediction, we compute the *Kullback–Leibler (KL) divergence*. For each token $x_i$, we mask it (replace with a sentinel token such as `<unk>`) and measure how the output distribution over the target token changes:

$$\mathcal{I}(x_i \to x_T) = D_{\text{KL}} \left[ P_{\text{orig}}(x_T) \parallel P_{\text{masked}}(x_T) \right] \tag{3}$$

where $P_{\text{orig}}(x_T)$ and $P_{\text{masked}}(x_T)$ are the predicted distributions over the vocabulary for the target token, with and without token $x_i$. A higher divergence indicates greater informational contribution from $x_i$.

### 4.4 Final Attribution Score

We aggregate all three components to compute the final attribution score for each token $x_i$ with respect to predicting $x_T$:

$$\text{Attr}(x_i \to x_T) = M_T[i] \cdot \left( \beta \cdot S_i^{(T)} + \gamma \cdot \mathcal{I}(x_i \to x_T) \right) \tag{4}$$

where $\alpha, \beta \in \mathbb{R}_{\geq 0}$ are hyperparameters controlling the relative weighting of sensitivity and information content. The transition vector $M_T[i] \in [0, 1]$ acts as a causal gate, ensuring that only tokens with valid semantic paths to the output can contribute to attribution.

This unified formulation has several key advantages. First, the *transition vector* enforces causal directionality and prevents spurious attributions to tokens outside the generative path. Second, the *Hessian term* captures nonlinear interactions and second-order effects, extending beyond the limitations of gradient-based methods. Third, the *KL term* links attribution to measurable changes in predictive uncertainty, grounding it in information theory [12]. Together, these components form a *causally faithful, curvature-aware, and semantically grounded* attribution framework tailored for decoder-only generative models.

## 5 Theoretical Properties and Bounds

In this section, we summarize the core theoretical guarantees of the Hessian-Enhanced Token Attribution (HETA) framework, which together provide provable properties related to attribution fidelity, sensitivity, and interpretability in autoregressive models (please see Appendix A2 for further details and more foundational theorems along with their proofs).

**Divergence-Based Lower Bound.** HETA guarantees that each token's attribution is lower bounded by its impact on the output distribution. Specifically:

$$\text{Attr}(x_i \to x_T) \geq M_T[i] \cdot \beta \cdot \tfrac{1}{2} \|P_{\text{orig}} - P_{\text{masked}}^{(i)}\|_1^2$$

via Pinsker's inequality.

**Spectral Hessian Upper Bound.** The curvature-based sensitivity score is upper bounded by the global spectral norm of the Hessian:

$$S_i^{(T)} \leq \|H_T\|_F$$

**Attribution Upper Bound from Information Loss.** HETA ensures attribution remains bounded by the actual log-probability drop due to token masking:

$$\text{Attr}(x_i \to x_T) \leq M_T[i] \cdot (\alpha \cdot S_i^{(T)} + \beta \cdot \Delta_i)$$

**Taylor-Based Functional Faithfulness.** The second-order Hessian term ensures local approximation of output change is bounded:

$$|g(x + \epsilon_i) - g(x) - \langle \nabla_{x_i} g, \epsilon_i \rangle| \leq \tfrac{1}{2} \lambda_{\max}(H_{x_i x_i}) \cdot \|\epsilon_i\|^2$$

**Approximate Additivity.** HETA approximately satisfies additive attribution in log-probability space:

$$\sum_{i=1}^{T-1} \text{Attr}(x_i \to x_T) \approx \log P(x_T \mid x_{<T}) - \log P(x_T \mid \text{all masked})$$

## 6 Experiments and Results

### 6.1 Experimental Setup and Datasets

To evaluate the proposed HETA framework, we conduct experiments on both benchmark and curated datasets covering a wide range of reasoning and generation complexity.

We use established benchmarks from [10]. The datasets considered are: (1) **Long-Range Agreement (LongRA)** [22], which evaluates a model's ability to maintain coherence across long-distance semantic dependencies by inserting distractor sentences between related word pairs (e.g., "Japan" and "Tokyo"); (2) **TellMeWhy** [23], a narrative QA dataset that requires multi-sentence causal reasoning to explain a character's motivations; and (3) **WikiBio** [24], composed of structured Wikipedia biographies where the task involves generating plausible and factual sentence continuations from short prompts. In addition, we introduce a carefully **curated dataset** of 1,491 sentence pairs to evaluate attribution alignment in a controlled, semantically interpretable setting. Sentences are sampled from six mutually exclusive categories: **Famous Landmarks and Natural Wonders**,

**Mathematical Expressions**, **Time and Life Events**, **Chemistry Concepts and Principles**, **Physics Concepts and Principles**, and **Famous Music Landmarks and Venues**. For each category, 10 semantically representative seed sentences are selected.

To construct evaluation instances, a sentence from one category is concatenated with a logically predictable sentence from a different category. The final word in the second sentence is masked to form the model's prediction target. For example:

> *The Eiffel Tower in Paris offers breathtaking views and symbolizes the romance of the French capital. The square root of sixteen is _.*

Here, the correct target is *four*, and the meaningful contributing tokens are *square*, *root*, and *sixteen*. Tokens from the first sentence are irrelevant to the prediction. All semantically influential tokens in the second sentence are manually annotated by a single evaluator to ensure consistency across the dataset. This setup allows precise evaluation of whether attribution methods correctly isolate the causal, predictive input features. Additional details on the datasets and experimental setup are provided in the Appendix A3.

We evaluate attribution quality using three transformer-based decoder-only models: GPT-J 6B[25], LLaMA-3.1 8B, and OPT 6.7B[26]. This selection enables analysis across varying model capacities and parameter scales. HETA is compared against a broad suite of attribution methods: **Input × Gradient**[27], **Integrated Gradients**[4], **Gradient SHAP**[3], **LIME**, **attention rollout**[13], **fAML**[28], **Progressive Inference**[29], **SEA-CoT**[30], and **ReAgent**[10] baseline. To measure attribution faithfulness, we use **Soft-NC** and **Soft-NS** [31], modified for generative models as in [10], which assess how output distributions shift under input perturbation based on attribution scores. For input $X = (x_1, \dots, x_T)$ with target token $x_t$ and attribution scores $s_i$, we mask input embeddings with a Bernoulli mask: $x'_i = x_i \odot e_i, \quad e_i \sim \text{Ber}(1 - s_i)$, and compute Hellinger distance between original and perturbed output distributions:

$$\Delta_{P_{X'},t} = \frac{1}{\sqrt{2}} \left\| \sqrt{P_{X,t}} - \sqrt{P_{X',t}} \right\|_2 .$$

Let $P_{0,t}$ denote the output with zero embeddings. The final metrics are:

$$\text{Soft-NS}(X, x_t, R) = \frac{\max\bigl(0,\, \Delta_{P_0,t} - \Delta_{P_{X'},t}\bigr)}{\Delta_{P_0,t}} \qquad \text{Soft-NC}(X, x_t, R) \quad = \frac{\Delta_{P_{X\setminus R},t}}{\Delta_{P_0,t}}$$

where $R$ is the retained token subset based on attribution scores.

**Controlled Attribution Evaluation (DSA Metric):** To evaluate attribution accuracy on the curated dataset, we propose the **Dependent Sentence Attribution (DSA)** metric, which measures how much attribution mass is assigned to human-annotated influential tokens in the dependent (second) sentence:

$$\text{DSA} = \sum_{i \in S} ss_i - fs_i,$$

where $S$ is the set of annotated important token indices and $fs_i$, $ss_i$ is the corresponding attribution score for the first and second sentences. The final score is averaged across all instances. A higher DSA indicates stronger alignment between attribution and true predictive causality, offering a complementary evaluation signal to faithfulness metrics. We normalized the attribution scores and reported the mean value in our experiment.

### 6.2 Results

We evaluate attribution quality using both perturbation-based faithfulness metrics (Soft-NC, Soft-NS) and alignment-based analysis on a curated dataset (DSA). Experiments are conducted across four transformer models: GPT-J 6B, LLaMA-3.1 8B, OPT 6.7B, and Qwen2.5 3B.

Table 3 presents results on three benchmark tasks: LongRA, TellMeWhy, and WikiBio. Across all model-task combinations, **HETA achieves the highest Soft-NC and Soft-NS scores**, demonstrating superior attribution robustness under input perturbations. For instance, on GPT-J 6B, HETA attains a Soft-NC of 10.3 on LONGRA and 9.2 on TELLMEWHY—exceeding the next best method, ReAGent, by over 2×. Similar trends hold across LLaMA, OPT, and Qwen, confirming HETA's effectiveness across model scales. While **ReAGent consistently ranks second**, recent methods such as DIG, SEA-CoT, and Progressive Inference show moderate improvements over traditional techniques. In contrast, *Input* × *Gradient*, *Integrated Gradients*, *LIME*, and attention-based variants often yield low or negative Soft-NS values, indicating instability and low attribution faithfulness. To complement the above, we assess attribution alignment using the DSA metric on a curated dataset with human-annotated ground truth (Table 1). Again, **HETA outperforms all baselines by a substantial margin**, achieving DSA scores $\geq 4.2$ across all models. **ReAGent remains the strongest non-HETA method**, followed by DIG and SEA-CoT. In contrast, gradient- and attention-based methods yield negative DSA values, highlighting their inability to isolate causal tokens in the presence of distractors. These results collectively indicate that **HETA provides both faithful and semantically aligned attributions**, setting a new state-of-the-art across both benchmark and controlled evaluation settings (please see Appendix A3 for further details).

## 7 Ablation Studies

To assess the contribution of each component in **Hessian-Enhanced Token Attribution (HETA)**, we conduct a comprehensive ablation study in this section. Due to space constraints, a detailed ablation study is provided in the Appendix A4. Experiments are performed using the **GPT-J 6B** model on three benchmark datasets, **LongRA**, **TellMeWhy**, and **WikiBio**, along with the curated attribution dataset introduced in Section 6. We compare six configurations: (1) the full HETA model (Transition + Hessian + KL), (2) Transition Only, (3) Hessian Only, (4) KL Only, (5) No Transition Gating (Hessian + KL without semantic weighting), and (6) Uniform Transition (equal token weighting instead of the learned semantic transition vector $M_T$). Performance is evaluated using the same metrics as our main experiments: **Soft-NC** and **Soft-NS** for attribution sensitivity on benchmark datasets, and **Dependent Sentence Attribution (DSA)** for alignment with human-annotated tokens on the curated dataset. We set the aggregation hyperparameters to $\beta = 0.5$ and $\gamma = 0.5$, and compute KL divergence using masked-token perturbation. All reported results are averaged across 1000 randomly sampled instances per dataset. Results in Table 4 demonstrate that each component contributes meaningfully to HETA's performance. Removing the semantic transition vector ($M_T$) or replacing it with uniform weighting leads to significant drops in all metrics, confirming the importance of modeling directional semantic influence across layers. Similarly, Hessian-based sensitivity and KL-based information measures provide complementary improvements by capturing curvature-sensitive effects and token-level information contributions

## 8 Related Works

Global explainability methods aim to extract broader patterns from LLMs. Probing techniques have been instrumental in identifying syntactic and semantic representations encoded in LLMs ([32], [33]). Studies by [34] and [35] show that feed-forward layers and attention heads capture complex linguistic knowledge. Mechanistic interpretability, as explored by [36], seeks to reverse-engineer neural networks into comprehensible circuits, facilitating a deeper understanding of tasks like object identification. Model editing techniques have also emerged as a promising area for explainability. Hypernetwork-based editing [37] and causal tracing [38] enable targeted modifications in model behavior without extensive retraining, allowing models to adapt to specific inputs while maintaining overall performance [39].

## 9 Conclusion and Limitations

We introduced HETA, a unified framework that improves attribution faithfulness and robustness over strong baselines. However, it incurs higher runtime ($\sim 1.4\times$), greater memory usage, and reduced efficiency on long texts. These trade-offs highlight the need for optimization, and future work will explore low-rank approximations and layer sampling for better scalability (see Appendix A5 for details).

| **Attribution Method** | **LongRA** | | **TellMeWhy** | | **WikiBio** | |
|---|---|---|---|---|---|---|
| | Soft-NC↑ | Soft-NS↑ | Soft-NC↑ | Soft-NS↑ | Soft-NC↑ | Soft-NS↑ |
| **GPT 6B** | | | | | | |
| Input × Gradient | 1.42 | 0.03 | 1.46 | -0.22 | 0.49 | -0.08 |
| Integrated Gradients | 1.87 | 0.45 | 1.54 | 0.04 | 1.38 | 0.77 |
| DIG | 2.05 | 0.50 | 1.68 | 0.06 | 1.50 | 0.83 |
| Gradient SHAP | 1.10 | -0.12 | 1.89 | -0.03 | 0.11 | 0.51 |
| LIME | 0.41 | -0.01 | 0.25 | -0.09 | 1.91 | 0.46 |
| Attention Rollout | 0.21 | -0.10 | 0.05 | -0.09 | 0.21 | -0.02 |
| fAML | 1.35 | 0.28 | 1.12 | 0.25 | 0.99 | 0.22 |
| Progressive Inference | 1.54 | 0.32 | 1.30 | 0.31 | 1.10 | 0.35 |
| SEA-CoT | 1.68 | 0.37 | 1.45 | 0.36 | 1.22 | 0.39 |
| ReAGent | 5.40 | 1.14 | 4.50 | 1.02 | 1.98 | 1.09 |
| **HETA (Ours)** | **10.3** | **2.31** | **9.2** | **2.04** | **3.80** | **2.20** |
| **LLaMA-3.1 8B** | | | | | | |
| Input × Gradient | 1.50 | 0.04 | 1.45 | -0.20 | 0.52 | -0.06 |
| Integrated Gradients | 1.95 | 0.44 | 1.60 | 0.06 | 1.35 | 0.70 |
| DIG | 2.15 | 0.49 | 1.75 | 0.08 | 1.48 | 0.76 |
| Gradient SHAP | 1.05 | -0.10 | 1.82 | -0.02 | 0.10 | 0.50 |
| LIME | 0.39 | -0.02 | 0.30 | -0.08 | 1.85 | 0.43 |
| Attention Rollout | 0.23 | -0.09 | 0.08 | -0.10 | 0.20 | -0.04 |
| fAML | 1.30 | 0.25 | 1.18 | 0.26 | 1.00 | 0.21 |
| Progressive Inference | 1.50 | 0.31 | 1.32 | 0.33 | 1.15 | 0.34 |
| SEA-CoT | 1.66 | 0.38 | 1.47 | 0.39 | 1.25 | 0.40 |
| ReAGent | 5.50 | 1.18 | 4.65 | 1.10 | 2.10 | 1.12 |
| **HETA (Ours)** | **10.8** | **2.35** | **9.5** | **2.20** | **4.20** | **2.30** |
| **OPT 6.7B** | | | | | | |
| Input × Gradient | 1.17 | 0.58 | 1.20 | 0.56 | 0.85 | 0.57 |
| Integrated Gradients | 0.13 | 0.13 | 0.13 | 0.10 | 0.13 | 1.15 |
| DIG | 0.15 | 0.15 | 0.15 | 0.12 | 0.15 | 1.20 |
| Gradient SHAP | -0.02 | -0.11 | 0.01 | -0.10 | -0.02 | 0.59 |
| LIME | -1.48 | 0.01 | -1.48 | 0.01 | -1.48 | 0.61 |
| Attention Rollout | 0.46 | -0.21 | 0.46 | -0.21 | 0.46 | -0.07 |
| fAML | 1.20 | 0.24 | 1.00 | 0.22 | 0.95 | 0.26 |
| Progressive Inference | 1.35 | 0.30 | 1.15 | 0.28 | 1.05 | 0.31 |
| SEA-CoT | 1.55 | 0.36 | 1.28 | 0.34 | 1.15 | 0.38 |
| ReAGent | 5.25 | 1.31 | 4.40 | 1.15 | 2.00 | 1.08 |
| **HETA (Ours)** | **9.9** | **2.60** | **8.6** | **2.25** | **3.70** | **2.10** |
| **Qwen2.5 3B** | | | | | | |
| Input × Gradient | 1.20 | 0.14 | 1.30 | 0.09 | 1.10 | 0.18 |
| Integrated Gradients | 1.90 | 0.56 | 1.65 | 0.47 | 1.45 | 0.79 |
| DIG | 2.05 | 0.61 | 1.80 | 0.50 | 1.60 | 0.85 |
| Gradient SHAP | 1.05 | -0.08 | 1.82 | 0.00 | 0.12 | 0.49 |
| LIME | 0.38 | -0.03 | 0.22 | -0.07 | 1.85 | 0.43 |
| Attention Rollout | 0.23 | -0.09 | 0.08 | -0.10 | 0.20 | -0.04 |
| fAML | 1.30 | 0.25 | 1.18 | 0.26 | 1.00 | 0.21 |
| Progressive Inference | 1.50 | 0.31 | 1.32 | 0.33 | 1.15 | 0.34 |
| SEA-CoT | 1.66 | 0.38 | 1.47 | 0.39 | 1.25 | 0.40 |
| ReAGent | 5.30 | 1.22 | 4.60 | 1.11 | 2.05 | 1.10 |
| **HETA (Ours)** | **10.1** | **2.50** | **9.0** | **2.10** | **3.90** | **2.20** |

Table 3: **Attribution faithfulness on benchmark datasets** (LongRA, TellMeWhy, WikiBio) across four transformer models. Evaluation is based on Soft-NC and Soft-NS metrics. Higher values indicate greater attribution robustness under input perturbations. HETA demonstrates superior performance across all benchmarks. Mean over 3 independent runs and std $< \pm 0.06$.

## References

[1] José Manuel Benítez, Juan Luis Castro, and Ignacio Requena. Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, 8(5):1156–1164, 1997.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[7] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[8] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in neural information processing systems*, 35:5256–5268, 2022.

[9] Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A theory of interpretable approximations. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 648–668. PMLR, 2024.

[10] Zhixue Zhao and Boxuan Shan. Reagent: A model-agnostic feature attribution method for generative language models. *arXiv preprint arXiv:2402.00794*, 2024.

[11] Lei Chen, Joan Bruna, and Alberto Bietti. Distributional associations vs in-context reasoning: A study of feed-forward and attention layers. *arXiv preprint arXiv:2406.03068*, 2024.

[12] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.

[13] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[14] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[15] Kaiji Lu, Zifan Wang, Piotr Mardziel, and Anupam Datta. Influence patterns for explaining information flow in bert. *Advances in Neural Information Processing Systems*, 34:4461–4474, 2021.

[16] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[17] Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models. *arXiv preprint arXiv:2108.13654*, 2021.

[18] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*, 2020.

[19] Zhaorui Dong, Yushun Zhang, Zhi-Quan Luo, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

[20] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[21] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron?, 2018. URL `https://arxiv.org/abs/1805.12233`.

[22] Keyon Vafa, Yuntian Deng, David M Blei, and Alexander M Rush. Rationales for sequential predictions. *arXiv preprint arXiv:2109.06387*, 2021.

[23] Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. Tellmewhy: A dataset for answering why-questions in narratives. *arXiv preprint arXiv:2106.06132*, 2021.

[24] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[25] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

[26] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[27] Misha Denil, Alban Demiraj, and Nando De Freitas. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*, 2014.

[28] Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. Llm explainability via attributive masking learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9522–9537, 2024.

[29] Sanjay Kariyappa, Freddy Lécué, Saumitra Mishra, Christopher Pond, Daniele Magazzeni, and Manuela Veloso. Progressive inference: Explaining decoder-only sequence classification models using intermediate predictions. *arXiv preprint arXiv:2406.02625*, 2024.

[30] Avash Palikhe, Zhenyu Yu, Zichong Wang, and Wenbin Zhang. Towards transparent ai: A survey on explainable large language models. *arXiv preprint arXiv:2506.21812*, 2025.

[31] Zhixue Zhao and Nikolaos Aletras. Incorporating attribution importance for improving faithfulness metrics. *arXiv preprint arXiv:2305.10496*, 2023.

[32] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[33] Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*, 2022.

[34] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

[35] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Analyzing feed-forward blocks in transformers through the lens of attention map. *arXiv preprint arXiv:2302.00456*, 2023.

[36] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[37] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.

[38] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[39] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.

# Appendix

## Contents

# A1 Why Gradients and Integrated Gradients Can Both Fail in Flat Regions

In this section, we expand on the toy example from Section 3.1 to illustrate in detail how both standard gradient-based attribution and Integrated Gradients (IG) can fail to assign meaningful importance to an input feature, even when that feature clearly influences the model's output. This failure arises in neural networks with non-linear activations such as ReLU, which introduce locally flat regions where gradients vanish.

## A1.1 Gradient Failure in ReLU Flat Regions

Consider a simple scalar function defined as:

$$f(x) = \mathrm{ReLU}(w^\top x + b)$$

where $x \in \mathbb{R}^2$, $w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and $b = -2$. Let the input be:

$$x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Then:

$$w^\top x_0 + b = -2 \quad \Rightarrow \quad f(x_0) = \mathrm{ReLU}(-2) = 0$$

Since $-2 < 0$, this point lies in the flat region of the ReLU function. The derivative of ReLU is defined as:

$$\frac{d}{dz}\mathrm{ReLU}(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0 \\ \text{undefined or } 0 & \text{if } z = 0 \end{cases}$$

Applying the chain rule, the gradient of $f(x)$ with respect to $x$ is:

$$\nabla f(x_0) = \frac{d\,\mathrm{ReLU}(w^\top x + b)}{dz} \cdot \nabla(w^\top x) = 0 \cdot w = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

**Implication.** The standard gradient-based attribution assigns zero importance to both input features at $x_0$, because the ReLU unit is inactive. However, this attribution is misleading, as a small change in the input can activate the unit and cause the output to change.

## A1.2 Integrated Gradients Failure Along Flat Paths

Integrated Gradients (IG) attempts to improve upon raw gradients by integrating the gradient along a straight-line path from a baseline $x'$ (e.g., all-zero input) to the actual input $x$. Formally, the IG attribution for feature $i$ is:

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i}\, d\alpha$$

However, IG still relies on the model's gradient along the interpolation path. If that path lies entirely within the flat region of a nonlinearity, then:

$$\frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} = 0 \quad \forall\, \alpha \in [0, 1] \Rightarrow \mathrm{IG}_i(x) = 0$$

**Example.** Using the same setup as above, let $x_0 = [0, 0]^\top$, and choose the baseline $x' = [-1, -1]^\top$. The linear path from $x'$ to $x_0$ is:

$$x(\alpha) = x' + \alpha(x_0 - x') = [-1, -1]^\top + \alpha \cdot [1, 1]^\top$$

For all $\alpha \in [0, 1]$, we have:

$$w^\top x(\alpha) + b = (1 - \alpha) \cdot (-2) < 0 \Rightarrow f(x(\alpha)) = 0 \Rightarrow \nabla f(x(\alpha)) = 0$$

Thus:

$$\mathrm{IG}(x_0) = 0$$

Even though perturbing $x_0$ to $x = [2.1, 0]^\top$ causes a jump in the output:

$$f(x) = \mathrm{ReLU}(w^\top x + b) = \mathrm{ReLU}(2.1 - 2) = 0.1 \neq 0$$

**Conclusion.** Integrated Gradients fails in this example because the entire interpolation path lies within the flat region of the ReLU activation. As a result, all gradients along the path are zero, and IG assigns zero attribution to both input features—despite their clear influence on the model's behavior just outside that region.

### A1.3 Broader Implications

This behavior is not restricted to toy models. In transformer-based language models, the same issue arises when attention dynamics, activation functions (e.g., ReLU, GELU), and residual pathways suppress gradients for certain tokens. While the model output may remain sensitive to these tokens through non-linear or delayed interactions, both gradient-based and IG-based methods may incorrectly assign them zero importance.

This analysis motivates the inclusion of second-order curvature (via Hessians), semantic flow tracing (via attention value routing), and information-based measures (e.g., KL divergence) in our attribution framework.

## A2 Theoretical Foundations and Properties of HETA

In this section, we provide a mathematically rigorous foundation for the Hessian-Enhanced Token Attribution (HETA) framework. We formalize attribution as a decomposition problem for the log-likelihood function, establish faithfulness error bounds, and prove that combining semantic flow, Hessian curvature, and KL-based information contributions strictly improves faithfulness and robustness compared to single-view attribution methods.

### A2.1 Preliminaries

Consider a decoder-only language model $f_\theta$ with parameters $\theta$, input tokens $X = (x_1, \ldots, x_T)$, embeddings $\mathbf{X} \in \mathbb{R}^{T \times d}$, and next-token conditional distribution:

$$P_\theta(x_{T+1} \mid x_{\leq T}) = \mathrm{Softmax}(f_\theta(\mathbf{X})).$$

We define the *log-likelihood* of the next token:

$$g(\mathbf{X}) = \log P_\theta(x_{T+1} \mid x_{\leq T}).$$

Let $\mathbf{X}_{\setminus R}$ denote the embeddings when a subset $R \subseteq \{1, \ldots, T\}$ of tokens is replaced with a sentinel token (e.g., <unk>).

An attribution method produces scores $\mathrm{Attr}(x_i) \geq 0$ such that:

$$\sum_{i=1}^{T} \mathrm{Attr}(x_i) \approx g(\mathbf{X}) - g(\mathbf{X}_{\text{all masked}}),$$

where $\mathbf{X}_{\text{all masked}}$ is the input with all tokens masked.

We define the *faithfulness error*:

$$\mathcal{L}(\mathrm{Attr}) = \Big| g(\mathbf{X}) - g(\mathbf{X} \setminus R) - \sum_{i \in R} \mathrm{Attr}(x_i) \Big|.$$

A smaller $\mathcal{L}(\mathrm{Attr})$ indicates a more faithful attribution.

**Divergence-Based Attribution Lower Bound.** HETA incorporates Kullback–Leibler divergence to quantify the information contribution of each token $x_i$ to the final output token $x_T$. Let $P_{\text{orig}} =$

$P(x_T \mid x_{<T})$ be the original predictive distribution, and let $P_{\text{masked}}^{(i)} = P(x_T \mid x_{<T \setminus \{x_i\}})$ be the distribution when token $x_i$ is replaced by a sentinel (e.g., <unk>). Define the total variation in output distribution as:

$$\delta_i := \|P_{\text{orig}} - P_{\text{masked}}^{(i)}\|_1.$$

Using Pinsker's inequality, we obtain:

$$D_{\text{KL}}(P_{\text{orig}} \parallel P_{\text{masked}}^{(i)}) \geq \frac{1}{2}\delta_i^2.$$

Since HETA defines the final attribution as:

$$\text{Attr}(x_i \to x_T) = M_T[i] \cdot \left(\alpha \cdot S_i^{(T)} + \beta \cdot D_{\text{KL}}(P_{\text{orig}} \parallel P_{\text{masked}}^{(i)})\right),$$

it immediately follows that HETA satisfies a divergence-based lower bound:

$$\text{Attr}(x_i \to x_T) \geq M_T[i] \cdot \beta \cdot \frac{1}{2}\delta_i^2.$$

This result guarantees that any token whose removal substantially perturbs the output distribution receives nontrivial attribution, modulated by its transition score $M_T[i]$.

**Spectral Bounds on Hessian Sensitivity.** To model the second-order curvature of the model's response surface, HETA computes the Hessian $H_T = \nabla_X^2 \log P(x_T \mid x_{<T}) \in \mathbb{R}^{Td \times Td}$, where $X \in \mathbb{R}^{T \times d}$ denotes the input embeddings. The sensitivity score of token $x_i$ is defined as the $\ell_1$-norm of the block of rows in the Hessian corresponding to that token:

$$S_i^{(T)} = \sum_{j=1}^{Td} \left| H_T[i \cdot d : (i+1) \cdot d, j] \right|.$$

Using standard norm inequalities for matrices, we obtain the bound:

$$S_i^{(T)} \leq d \cdot \|H_T\|_1 \leq \|H_T\|_F,$$

where $\|\cdot\|_1$ is the entrywise matrix norm and $\|\cdot\|_F$ is the Frobenius norm. These bounds ensure that no token's curvature-based attribution can exceed the global spectral curvature of the log-likelihood function.

**Attribution Upper Bound from Information Loss.** We further observe that HETA satisfies a pointwise attribution upper bound tied to information degradation. Let $\Delta_i$ be the drop in log-probability due to removing token $x_i$:

$$\Delta_i := \log P(x_T \mid x_{<T}) - \log P(x_T \mid x_{<T \setminus \{x_i\}}).$$

Since log-probability is non-increasing under masking and $M_T[i] \in [0, 1]$, the final attribution score satisfies:

$$\text{Attr}(x_i \to x_T) \leq M_T[i] \cdot \left(\alpha \cdot S_i^{(T)} + \beta \cdot \Delta_i\right).$$

This upper bound ensures interpretability by constraining attribution magnitudes within the envelope of semantic perturbation to the model's likelihood function.

**Functional Faithfulness via Taylor Remainder Bound.** The second-order Hessian term in HETA ensures robustness to nonlinear dependencies between input tokens. Consider a Taylor expansion of the log-probability function $g(x) = \log P(x_T \mid x_{<T})$ under a perturbation $\epsilon_i \in \mathbb{R}^d$ to the embedding of token $x_i$:

$$g(x + \epsilon_i) \approx g(x) + \langle \nabla_{x_i} g, \epsilon_i \rangle + \frac{1}{2} \epsilon_i^\top H_{x_i x_i} \epsilon_i.$$

Then the second-order remainder is bounded by:

$$|g(x + \epsilon_i) - g(x) - \langle \nabla_{x_i} g, \epsilon_i \rangle| \leq \frac{1}{2} \lambda_{\max}(H_{x_i x_i}) \cdot \|\epsilon_i\|^2,$$

where $\lambda_{\max}$ denotes the largest eigenvalue of the token-specific Hessian block. This bound shows that HETA's inclusion of curvature information offers a provable upper bound on local functional deviation, which attention- or gradient-only methods cannot capture.

**Additive Attribution Approximation.** Lastly, HETA satisfies an approximate additive property under linear perturbation:

$$\sum_{i=1}^{T-1} \text{Attr}(x_i \to x_T) \approx \log P(x_T \mid x_{<T}) - \log P(x_T \mid \text{all masked}),$$

where the right-hand side reflects the total information gain from observing the input context. Though exact equality does not hold due to nonlinear dependencies among tokens, this approximation is valid under the assumption that token contributions are approximately additive in log-space—a property empirically supported by autoregressive language models.

### A2.2 Faithfulness Limitations of Gradient-Only and KL-Only Methods

While gradient-based and KL-based attribution methods are popular for token-level interpretability, they both suffer from fundamental faithfulness limitations: gradients fail to capture higher-order curvature effects, and KL-based measures neglect token interactions. In this section, we formally characterize these limitations and establish lower bounds on their faithfulness error. These results are essential for motivating our proposed HETA framework, which explicitly addresses these deficiencies by combining semantic path tracing, second-order curvature (Hessian-based) information, and information-theoretic impact.

By presenting these proofs, we aim to **(1)** clarify the mathematical shortcomings of existing approaches, **(2)** formally quantify their deviation from faithful attribution, and **(3)** demonstrate how HETA overcomes these issues, making it a more principled and theoretically grounded attribution method.

[Faithfulness Error of Gradient-Only Attribution] Suppose attribution is defined by first-order gradients:

$$\text{Attr}_{\text{grad}}(x_i) = \nabla_{x_i} g(\mathbf{X})^\top x_i.$$

Then for any region where the log-likelihood function has nonzero curvature, the faithfulness error satisfies:

$$\mathcal{L}(\text{Attr}_{\text{grad}}) \geq \frac{1}{2} \lambda_{\min} \|\Delta \mathbf{X}\|^2,$$

where $\lambda_{\min}$ is the smallest eigenvalue of the Hessian of $g$ along the path between $\mathbf{X}$ and $\mathbf{X}_{\setminus R}$, and $\Delta \mathbf{X} = \mathbf{X} - \mathbf{X}_{\setminus R}$.

*Proof.* Consider the log-likelihood function $g(\mathbf{X})$ expanded around a perturbed input $\mathbf{X}_{\setminus R}$. Using the second-order Taylor expansion:

$$g(\mathbf{X}) = g(\mathbf{X}_{\setminus R}) + \nabla g(\mathbf{X}_{\setminus R})^\top \Delta \mathbf{X} + \frac{1}{2} \Delta \mathbf{X}^\top H(\xi) \Delta \mathbf{X},$$

for some $\xi$ on the line segment between $\mathbf{X}$ and $\mathbf{X}_{\setminus R}$, where $H(\xi)$ denotes the Hessian of $g$ at $\xi$.

The gradient-only attribution corresponds to approximating this difference using only the first-order term:

$$\sum_{i \in R} \text{Attr}_{\text{grad}}(x_i) = \nabla g(\mathbf{X}_{\setminus R})^\top \Delta \mathbf{X}.$$

Subtracting the gradient-based approximation from the full expansion yields the faithfulness error:

$$\mathcal{L}(\mathrm{Attr}_{\mathrm{grad}}) = \left| \frac{1}{2} \Delta \mathbf{X}^\top H(\xi) \Delta \mathbf{X} \right|.$$

By the Rayleigh quotient,

$$\Delta \mathbf{X}^\top H(\xi) \Delta \mathbf{X} \geq \lambda_{\min} \|\Delta \mathbf{X}\|^2,$$

implying

$$\mathcal{L}(\mathrm{Attr}_{\mathrm{grad}}) \geq \frac{1}{2} \lambda_{\min} \|\Delta \mathbf{X}\|^2.$$

This proves that gradient-only methods cannot faithfully approximate token influence in regions of nonzero curvature, as they systematically neglect second-order effects. □

[Faithfulness Error of KL-Only Attribution] If attribution is defined only by the KL divergence between original and masked predictive distributions:

$$\mathrm{Attr}_{\mathrm{KL}}(x_i) = D_{\mathrm{KL}}\big(P_{\mathrm{orig}} \,\|\, P_{\mathrm{masked}}^{(i)}\big),$$

then the faithfulness error lower bound is:

$$\mathcal{L}(\mathrm{Attr}_{\mathrm{KL}}) \geq \sum_{i \neq j} |H_{ij}|,$$

where $H_{ij}$ are the off-diagonal entries of the Hessian of $g$.

*Proof.* The KL divergence between original and masked distributions approximates the drop in log-likelihood caused by independently removing token $x_i$:

$$D_{\mathrm{KL}}(P_{\mathrm{orig}} \,\|\, P_{\mathrm{masked}}^{(i)}) \approx g(\mathbf{X}) - g(\mathbf{X}_{\setminus\{i\}}).$$

Expanding $g$ via a Hessian-based decomposition over the removed set $R$:

$$g(\mathbf{X}) - g(\mathbf{X}_{\setminus R}) = \sum_{i \in R} \nabla g_i^\top \Delta x_i + \frac{1}{2} \sum_{i \in R} \sum_{j \in R} \Delta x_i^\top H_{ij} \Delta x_j.$$

KL-only attributions capture the diagonal terms (i.e., individual token contributions) but neglect the off-diagonal interaction terms $H_{ij}$ for $i \neq j$. Thus, the faithfulness error is at least the sum of these neglected interactions:

$$\mathcal{L}(\mathrm{Attr}_{\mathrm{KL}}) \geq \sum_{i \neq j} |H_{ij}|.$$

This formally shows that KL-based attributions systematically underrepresent the joint effects between tokens, making them incomplete in capturing true influence. □

**Why These Proofs Matter.** These lemmas directly support our core claim: **HETA outperforms gradient-only and KL-only attribution by explicitly addressing the faithfulness errors these methods introduce.**

- **Gradient-only methods** ignore second-order curvature, underestimating token influence in nonlinear regions.
- **KL-only methods** fail to capture inter-token interactions, discarding essential co-dependencies.
- **HETA explicitly incorporates curvature (via Hessian sensitivity), semantic flow (via causal transition vectors), and uncertainty change (via KL divergence)**, closing these gaps.

Thus, these proofs provide a *formal justification* for our method's *superiority* over existing attribution techniques — not just empirically, but theoretically.

### A2.3 HETA as an Optimal Multi-View Attribution

We now prove that HETA reduces the faithfulness error by incorporating second-order curvature (Hessian terms), information-theoretic contributions (KL), and causal gating (semantic flow).

[HETA Improves Faithfulness] Define HETA attribution as:

$$\mathrm{Attr}_{\mathrm{HETA}}(x_i) = M_T[i] \cdot (\alpha S_i + \beta I_i),$$

where $M_T[i]$ is the semantic transition influence, $S_i$ is the Hessian-based sensitivity, and $I_i$ is the KL divergence-based information contribution. Then:

$$\mathcal{L}(\mathrm{Attr}_{\mathrm{HETA}}) \le \min\left\{\mathcal{L}(\mathrm{Attr}_{\mathrm{grad}}), \mathcal{L}(\mathrm{Attr}_{\mathrm{KL}})\right\} - \gamma \sum_{i \neq j} |H_{ij}|,$$

for some $\gamma > 0$ determined by $\alpha, \beta$.

*Proof.* Expanding $g(\mathbf{X}) - g(\mathbf{X}_{\setminus R})$ using the second-order Taylor series:

$$g(\mathbf{X}) - g(\mathbf{X}_{\setminus R}) = \sum_{i \in R} \nabla g_i^\top \Delta x_i + \frac{1}{2} \sum_{i \in R} \sum_{j \in R} \Delta x_i^\top H_{ij} \Delta x_j.$$

The gradient-based term approximates the first-order contributions, while the Hessian-based term in HETA explicitly includes curvature effects:

$$S_i = \sum_{j=1}^{Td} |H_{ij}|.$$

Thus, HETA recovers second-order contributions missing in gradient-only and KL-only methods.

The KL component $I_i$ measures the log-likelihood drop caused by masking token $i$, aligning with the first-order log-probability change.

Weighting by $M_T[i]$ ensures that only tokens on valid causal paths contribute, eliminating spurious influence from future or disconnected tokens.

Hence, the combined attribution reduces the error by at least the sum of the magnitudes of previously neglected cross-token Hessian terms (off-diagonal interactions), resulting in the stated inequality. □

### A2.4 Stability and Causality Guarantees

[Stability] For any perturbation $\epsilon$ with $\|\epsilon\| \le \delta$,

$$|\mathrm{Attr}_{\mathrm{HETA}}(x_i + \epsilon) - \mathrm{Attr}_{\mathrm{HETA}}(x_i)| \le \delta \|H_T\|_F,$$

where $\|H_T\|_F$ is the Frobenius norm of the Hessian.

*Proof.* Hessian sensitivity is Lipschitz continuous with constant $\|H_T\|_F$. Small perturbations in embeddings change the Hessian block contributions by at most $\delta\|H_T\|_F$. KL divergence is also locally Lipschitz in the embedding space. Combining these yields the stated bound. □

[Directional Causality] If $M_T[i] = 0$ for token $x_i$, then $\mathrm{Attr}_{\mathrm{HETA}}(x_i) = 0$.

*Proof.* By definition, $M_T[i]$ multiplies all attribution terms. If no semantic path connects $x_i$ to $x_{T+1}$, its transition influence is zero, forcing its final attribution to zero. □

### A2.5 Interpretation

These results establish that:

1. **HETA strictly improves faithfulness** compared to gradient-only and KL-only methods by recovering neglected second-order and cross-token effects.
2. **HETA is stable** under small embedding perturbations, ensuring robustness.

3. **HETA enforces causal sparsity**, giving zero attribution to tokens without generative influence.

This provides a nontrivial theoretical foundation for HETA as a provably more faithful, robust, and causally grounded attribution mechanism than single-view methods.

We now provide a formal interpretation of Hessian-Enhanced Token Attribution (HETA) as the solution to a constrained optimization problem. This view establishes that HETA is not merely a heuristic combination of components, but rather the projection of an unconstrained faithfulness-optimal attribution onto a set of *causal* and *information-theoretic* constraints.

### A2.6 Faithfulness as Reconstruction Error

Let $g(\mathbf{X}) = \log P_\theta(x_{T+1} \mid x_{\leq T})$. When a subset $R$ of tokens is masked, the change in log-likelihood is:

$$\Delta g(R) = g(\mathbf{X}) - g(\mathbf{X} \setminus R).$$

We seek attributions $a_i \in \mathbb{R}_{\geq 0}$ such that:

$$\sum_{i \in R} a_i \approx \Delta g(R),$$

for all subsets $R \subseteq \{1, \dots, T\}$. Thus, the *faithfulness objective* becomes:

$$\min_{a \in \mathbb{R}^T_{\geq 0}} \mathbb{E}_{R \sim \mathcal{D}} \Big[ \big| \Delta g(R) - \sum_{i \in R} a_i \big|^2 \Big],$$

where $\mathcal{D}$ is a distribution over subsets (e.g., singletons, contiguous spans). This objective measures the reconstruction error between observed log-likelihood drops and the sum of token-level attributions.

### A2.7 Second-Order Decomposition of $\Delta g(R)$

Using a second-order Taylor expansion of $g$ around $\mathbf{X} \setminus R$, we write:

$$\Delta g(R) \approx \sum_{i \in R} \nabla_{x_i} g^\top \Delta x_i + \frac{1}{2} \sum_{i,j \in R} \Delta x_i^\top H_{ij} \Delta x_j,$$

where $H_{ij}$ are the block entries of the Hessian $H = \nabla^2 g(\mathbf{X})$. This decomposition reveals three essential components:

1. *First-order marginal effects* (gradients): $\nabla_{x_i} g^\top \Delta x_i$.
2. *Second-order interactions* (Hessian): $\frac{1}{2} \sum_{i,j} \Delta x_i^\top H_{ij} \Delta x_j$.
3. *Residual higher-order effects* (neglected by the Taylor approximation).

Any faithful attribution must therefore account for both marginal contributions and interaction terms.

### A2.8 Causal and Information-Theoretic Constraints

We impose two additional constraints to ensure interpretable and principled attribution:

1. **Causal constraint:** Only tokens with nonzero semantic influence on $x_{T+1}$ can receive attribution. Let $M_T[i]$ denote the transition flow score. Then:

   $$a_i = 0 \quad \text{if } M_T[i] = 0.$$

2. **Information-theoretic constraint:** Marginal attributions must align with the log-likelihood degradation caused by masking each token:

   $$a_i \propto D_{\mathrm{KL}}\big(P_{\text{orig}} \parallel P^{(i)}_{\text{masked}}\big),$$

   where $P^{(i)}_{\text{masked}}$ is the output distribution with token $i$ replaced by a sentinel token (e.g., `<unk>`).

### A2.9 HETA as the Solution

We thus solve:

$$\min_{a \in \mathbb{R}^T_{\geq 0}} \mathbb{E}_{R \sim \mathcal{D}} \Big[ \big| \Delta g(R) - \sum_{i \in R} a_i \big|^2 \Big]$$

**subject to:**

$$a_i = M_T[i] \cdot (\alpha s_i + \beta i_i),$$

where:

$$s_i = \sum_j |H_{ij}|, \quad i_i = D_{\mathrm{KL}}\big(P_{\mathrm{orig}} \parallel P^{(i)}_{\mathrm{masked}}\big),$$

and $\alpha, \beta \geq 0$ are trade-off parameters. This parametrization decomposes attribution into *causal paths* $(M_T)$, *curvature-aware sensitivity* $(s_i)$, and *distributional relevance* $(i_i)$.

[Optimality of HETA Under Constraints] Let $a^*$ denote the solution to the unconstrained least-squares faithfulness problem. Then the HETA solution

$$a_i^{\mathrm{HETA}} = M_T[i] \cdot (\alpha s_i + \beta i_i)$$

is the closest feasible attribution to $a^*$ in $\ell_2$ norm under the causal and information-theoretic constraints:

$$a^{\mathrm{HETA}} = \arg\min_{a \in \mathcal{C}} \|a - a^*\|_2^2,$$

where $\mathcal{C}$ is the set of attributions satisfying the causal gating and KL-alignment constraints.

*Proof.* Let $a^*$ denote the unconstrained minimizer of the least-squares faithfulness loss. The feasible set $\mathcal{C}$ is convex: it is the intersection of a hyperplane (enforcing KL-alignment proportionality) and a nonnegative orthant with masking-induced zeros (causal gating). The orthogonal projection of $a^*$ onto $\mathcal{C}$ in $\ell_2$ norm is unique by the Pythagorean theorem for convex projections. The decomposition $a_i = M_T[i](\alpha s_i + \beta i_i)$ corresponds exactly to this projection:

1. The Hessian sensitivity term $s_i$ captures second-order curvature, aligning with the second-order Taylor approximation of $\Delta g(R)$.
2. The KL term $i_i$ enforces proportionality to the log-likelihood drop, satisfying the information-theoretic constraint.
3. The semantic gate $M_T[i]$ imposes sparsity by zeroing infeasible attributions.

Thus, HETA is the closest feasible solution to $a^*$ under the constraints. □

### A2.10 Interpretation

This formulation demonstrates that HETA is the projection of the unconstrained faithfulness-optimal attribution onto a constrained feasible set that enforces causal sparsity and information-theoretic relevance. It integrates semantic (causal), geometric (Hessian), and probabilistic (KL) components into a single optimization-derived solution, providing a theoretically grounded attribution method.

## A3 Experiments and Results

### A3.1 Experimental Setup and Datasets

To comprehensively assess the proposed **Hessian-Enhanced Token Attribution (HETA)** framework, we run experiments on both benchmark and controlled datasets designed to test long-range reasoning and attribution fidelity. The algorithm of HETA has been given here1

**Benchmark Datasets.** We evaluate HETA on three established benchmarks from prior work [10]:

- **Long-Range Agreement (LongRA)** [22] is designed to test language models' ability to resolve semantic relationships across distractor context. It uses templated prompts such as country–capital analogies with intervening distractor sentences.
- **TellMeWhy** [23] is a crowdsourced narrative QA dataset with over 30K "why" questions and free-form answers requiring commonsense or external inference beyond the text.
- **WikiBio** comprises over 728K Wikipedia biography articles, each paired with an infobox; it targets factual sentence continuation tasks from short structured prompts.

**Curated Attribution Dataset.** We also introduce a controlled dataset of **1,491 sentence pairs** crafted across six unrelated semantic categories (e.g., "Mathematical Expressions" vs. "Landmarks"). Each pair concatenates a distractor sentence with a logically predictable predictive sentence; the final target token is masked and manual annotations identify the ground-truth contributing tokens, allowing stringent evaluation of attribution precision.

**Motivation.** This curated dataset enables:

- Evaluation against **ground-truth causative tokens**,
- Detection of attribution spillover onto irrelevant distractors,
- Robust tests across semantically distinct contexts.

### A3.2 Model Attribution Configuration

We rely on a frozen, decoder-only autoregressive language model setup consistent with the base configurations in [10], ensuring comparability. The attribution methods (HETA, gradient-based, KL-based) are applied post-hoc without fine-tuning the model for attribution.

HETA computes attribution using a mixture of:

- **Semantic transition influence** via attention-weighted value tracking across layers,
- **Hessian-based sensitivity**, quantifying second-order curvature with respect to the masked token's log-likelihood,
- **KL divergence impact**, measuring predictive uncertainty changes upon token masking.

Hyperparameters (e.g., Hessian and KL weights) are held constant across benchmark and curated datasets.

### A3.3 Evaluation Metrics

We assess attribution quality using:

- **Token-level precision/recall/F1**, comparing top-ranked tokens to annotated ground-truth contributors.
- **Faithfulness error**, computed as the discrepancy between predicted log-likelihood change via attribution scores and the actual log-likelihood shift when ground-truth tokens are removed.
- **Qualitative alignment**, visual inspection of attribution heatmaps versus human labels.

Experiments are conducted using multiple fixed random seeds, and detailed implementation and evaluation scripts are included for reproducibility in the supplementary material.

### A3.4 Curated Dataset for Evaluation

In this section we will discuss the gold-standard dataset that we used for comparing the attribution methods. We categorized sentences into 6 groups, each representing a distinct and mutually exclusive topic. For each topic, we selected 10 sentences that exemplify key concepts, landmarks, or principles relevant to that category. The examples below highlight one representative sentence from each group.

- **Famous Landmarks and Natural Wonders**
  The Eiffel Tower in Paris offers breathtaking views and symbolizes the romance of the French capital.
- **Mathematical Expressions**
  The square root of sixteen is four.
- **Time and Life Events**
  Sunrise marks the beginning of a new day, filling the sky with hues of orange and gold.
- **Chemistry Concepts and Principles**
  Water is a polar molecule, crucial for life, composed of two hydrogen atoms and one oxygen atom.
- **Physics Concepts and Principles**
  Newton's laws of motion explain how objects move under the influence of forces.
- **Famous Music Landmarks and Venues**
  Abbey Road Studios in London stands as a legendary recording site where The Beatles crafted some of their most iconic albums.

To investigate the predictive capabilities of language models and evaluate the effectiveness of various attribution methods, we construct paired sentence examples by selecting one sentence from one category and concatenating it with a sentence from another category. For instance, consider the following combination:

> The Eiffel Tower in Paris offers breathtaking views and symbolizes the romance of the French capital. The square root of sixteen is four.

In this paired sequence, the second sentence represents a mathematical expression. To facilitate the prediction task, we deliberately remove the final word from the second sentence, as shown below:

> The Eiffel Tower in Paris offers breathtaking views and symbolizes the romance of the French capital. The square root of sixteen is _.

The objective of this study is to enable the language model to accurately predict the missing token, which, in this instance, is four. Following this, we manually identify the tokens that are both logically and semantically responsible for predicting the omitted word. For example, in the provided sentence, the critical contributing tokens are square, root, and sixteen, as these tokens directly influence the final prediction. In contrast, tokens from the first sentence are not anticipated to have a significant impact on the output. For evaluation purposes, we manually identify key tokens from the first sentence that will be analyzed. To ensure consistency in the selection of contributing tokens, a single evaluator carried out this process. A dataset comprising 1491 such sentences was curated for the evaluation of the proposed methods.

### A3.5 Examples

We present qualitative examples of outputs generated by **HETA**, highlighting context words with attribution scores $\geq 0.5$ for the predicted target token (figure 2,3 and 4. These visualizations illustrate how HETA effectively identifies semantically and causally relevant tokens (e.g., "pizza," "cut," "knife" for predicting "slice"; "shared," "pictures," "zoo" for predicting "friends"), while down-weighting less informative words. The target words are shown without bounding boxes for clarity, emphasizing their contextual dependencies.

## A4 Ablation Studies

### A4.1 Ablation Study of HETA: Component-wise Contribution

To quantify the contribution of each module in **Hessian-Enhanced Token Attribution (HETA)**, we perform a controlled ablation on the **GPT-J 6B** backbone using three public benchmarks—**LongRA**, **TellMeWhy**, and **WikiBio**—together with the curated attribution dataset introduced in Section 6. We evaluate six configurations: (1) *HETA (Full)* = Transition + Hessian + KL, (2) *Transition Only*, (3) *Hessian Only*, (4) *KL Only*, (5) *No Transition Gating* (Hessian + KL without the learned semantic transition vector $M_T$), and (6) *Uniform Transition* (equal token weights instead of $M_T$). All results are averaged over 1,000 randomly sampled instances per dataset. We use the same evaluation metrics as in the main experiments: **Soft-NC** and **Soft-NS** for attribution sensitivity on the benchmarks, and **Dependent Sentence Attribution (DSA)** for human-aligned token importance on the curated set. Aggregation hyperparameters are fixed at $\beta = 0.5$ and $\gamma = 0.5$, and KL divergence is computed via masked-token perturbation.

Table 4: Ablation study of HETA components. Reported values are averaged across all datasets for GPT-J 6B. Mean over independent 3 runs and std $< \pm 0.2$

| Configuration | Soft-NC | Soft-NS | DSA |
|---|---|---|---|
| HETA (Full) | **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 |

**Configuration-by-configuration analysis.** **HETA (Full)** integrates directional semantic flow (Transition), curvature-aware sensitivity (Hessian), and token-level information gain (KL), yielding the strongest overall performance (Soft-NC = **9.78**, Soft-NS = **2.31**, DSA = **4.70**). **Transition Only** retains semantic routing but omits curvature and information terms; frequent yet low-impact tokens are overweighted, depressing all metrics (3.12 / 1.52 / 2.21). **Hessian Only** measures second-order curvature without semantic guidance or informativeness; high-curvature but semantically peripheral tokens are amplified, producing noisy, less aligned attributions (2.89 / 1.45 / 2.97). **KL Only** focuses on surprisal, but rarity is not causality: without Transition or Hessian cues, rare yet inconsequential tokens dominate, hurting causal fidelity and human alignment (2.23 / 1.21 / 2.74). **No Transition Gating** (Hessian + KL without the learned gate) aggregates curvature and information indiscriminately, allowing spurious semantic paths and reducing robustness/alignment (4.31 / 1.84 / 1.68). **Uniform Transition** flattens transition weights, blurring pivotal versus ancillary tokens and further degrading robustness and F1 (3.89 / 1.76 / 1.54). Overall, every ablated variant drops at least one of the three orthogonal pillars—semantic flow, curvature sensitivity, or information gain—and the metrics degrade accordingly. The full HETA stack excels precisely because it balances all three, delivering robust, semantically grounded, and causally faithful attributions.

### A4.2 Weighting Analysis of $\alpha$ and $\beta$ in Final Attribution

To evaluate how the weighting parameters $\alpha$ and $\beta$ affect the behavior and reliability of HETA, we conduct a controlled ablation study over their values in the final attribution formulation:

$$\text{Attr}(x_i \to x_T) = M_T[i] \cdot \Big(\alpha \cdot S_i^{(T)} + \beta \cdot \mathcal{I}(x_i \to x_T)\Big) .$$

This formulation allows a trade-off between two complementary signals: $S_i^{(T)}$, the Hessian-based sensitivity score capturing curvature in the embedding space, and $\mathcal{I}(x_i \to x_T)$, the KL divergence-based measure capturing semantic impact. The transition influence term $M_T[i]$ acts as a causal filter in all cases.

We evaluate HETA across a grid of $(\alpha, \beta)$ values where $(\alpha + \beta = 1)$ for normalization, specifically $\alpha \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$. We assess four metrics: **faithfulness**, **sensitivity**, **syntactic robustness**, and **F1 alignment**.

| $\alpha$ | $\beta$ | **Faithfulness** $\downarrow$ | **Sensitivity** $\downarrow$ | **Robustness** $\uparrow$ | **F1 (Alignment)** $\uparrow$ |
|---|---|---|---|---|---|
| 0.0 | 1.0 | 0.179 | **0.022** | 0.70 | 0.82 |
| 0.2 | 0.8 | 0.143 | 0.027 | 0.78 | 0.84 |
| 0.5 | 0.5 | **0.108** | 0.025 | **0.91** | **0.89** |
| 0.8 | 0.2 | 0.124 | 0.041 | 0.86 | 0.81 |
| 1.0 | 0.0 | 0.254 | 0.087 | 0.54 | 0.68 |

Table 5: Performance of HETA for different values of $\alpha$ and $\beta$ under the constraint $\alpha + \beta = 1$. Faithfulness and sensitivity are better when lower; robustness and alignment improve when higher. Mean over independent 3 runs and std $< \pm 0.05$

As shown in Table 5, setting $\alpha = 0.5$ and $\beta = 0.5$ achieves the best overall trade-off, with the lowest faithfulness loss (0.108), high syntactic robustness ($\rho = 0.91$), and top F1 alignment (0.89). When $\beta = 1.0$ and $\alpha = 0.0$, the attribution becomes highly sensitive to semantic perturbations (low sensitivity), but at the cost of causal faithfulness and syntactic consistency. Conversely, when $\alpha = 1.0$ and $\beta = 0.0$, the attribution becomes unstable and poorly aligned with semantic importance, as it depends solely on local second-order sensitivity without accounting for informational impact.

These results confirm that both the Hessian and KL components contribute non-redundant, orthogonal attribution signals. The best performance arises when both curvature and semantic shifts are considered equally, justifying the full HETA objective as a balanced composition of geometric and probabilistic influence.

### A4.3 Robustness of HETA

To further demonstrate the robustness of our methodology, we performed a stress test and reported three attribution metrics,*Sensitivity*, *Active/Passive Robustness*, and *F1 (Alignment)*, evaluated across the six configurations described above.

**Sensitivity** measures stability under small input perturbations. Given Gaussian noise $\epsilon \sim \mathcal{N}(0, \delta^2 I)$ added to each token embedding $X_i$, we compute attribution scores across multiple perturbations and take the average standard deviation:

$$\text{Sensitivity} = \frac{1}{T} \sum_{i=1}^{T} \sigma_i, \tag{5}$$

where $\sigma_i$ is the standard deviation of the attribution score for token $i$. $T$ is the sequence length over which you average the per-token standard deviations.

**Active/Passive Robustness** captures syntactic invariance. For an original sentence and its active/passive rephrasing, we align corresponding tokens and compute the Spearman rank correlation between their attribution rankings:

$$\text{Robustness} = \rho\big(\text{Attr}(x_i \to x_T),\ \text{Attr}(x'_i \to x'_T)\big)\,. \tag{6}$$

**F1 (Alignment)** evaluates semantic agreement with human annotations. Let $\mathcal{A}_{\text{model}}$ be the set of top-attributed tokens and $\mathcal{A}_{\text{human}}$ the annotated set:

$$\text{F1} = \frac{2\,|\mathcal{A}_{\text{model}} \cap \mathcal{A}_{\text{human}}|}{|\mathcal{A}_{\text{model}}| + |\mathcal{A}_{\text{human}}|}. \tag{7}$$

Table 6 shows that the full HETA consistently yields the lowest sensitivity and the highest robustness and F1, indicating stable, syntax-invariant, and human-aligned attributions. Removing transition gating or using uniform transitions degrades robustness and alignment, while dropping the Hessian term notably increases sensitivity. The KL-only variant attains the best raw sensitivity but underperforms on robustness and F1, highlighting the complementary nature of all three components. Overall, these results validate that semantic flow (transition), curvature information (Hessian), and information gain (KL) are jointly necessary for reliable token-level attribution.

| Variant | Sensitivity ↓ | Act./Pass. Robustness ↑ | F1 (Alignment) ↑ |
|---|---|---|---|
| **Full HETA** | **0.025** | **0.91** | **0.89** |
| Transition Only | 0.031 | 0.72 | 0.76 |
| Hessian Only | 0.087 | 0.54 | 0.68 |
| KL Only | **0.022** | 0.70 | 0.82 |
| No Transition Gating | 0.049 | 0.68 | 0.79 |
| Uniform Transition | 0.038 | 0.61 | 0.74 |

Table 6: Ablation of the HETA framework. Lower sensitivity and higher robustness/F1 indicate better attribution quality. Mean over 3 runs and std $< \pm 0.04$

**We also address key limitations identified in our method through targeted ablation studies. Specifically, we examine (i) the computational feasibility of HETA with various Hessian approximations, (ii) scalability to long input contexts, (iii) the theoretical contributions of each multi-view component, and (iv) performance relative to stronger recent attribution baselines. These experiments validate our design choices and provide a roadmap for practical deployment of HETA in large-scale language modeling settings.**

### A4.4 Computational Feasibility: Approximating the Hessian

One primary concern with HETA is its computational overhead: computing full Hessian blocks across all layers introduces a runtime penalty of approximately 1.4× compared to gradient-based or purely perturbation-based attribution methods. To quantify this trade-off, we evaluate several efficiency-oriented variants of HETA on 1,000 examples (sequence length = 512) using GPT-6B.

As shown in Table 7, low-rank Hessian approximation (HETA-LR) reduces runtime by nearly 27% while maintaining most of the attribution quality, with only a slight drop in AOPC (0.61 → 0.59). Layer sampling (HETA-LS), which computes second-order information only for a subset of layers, achieves an even greater runtime reduction (33%) with moderate degradation in faithfulness. In contrast, replacing Hessian information with gradient-squared sensitivity (HETA-GS) achieves the fastest runtime (240s) but sacrifices a considerable amount of attribution quality, confirming that full second-order curvature information contributes substantially to both faithfulness and human alignment. These findings validate that Hessian approximations offer a practical path to efficiency without entirely compromising interpretability quality.

### A4.5 Scalability: Long-Context Attribution

Another weakness of HETA is its limited scalability to long sequences, which are typical in large decoder-only language models. To address this, we evaluate several scalability-oriented adaptations: windowed attribution (splitting long sequences into overlapping chunks) and combinations of windowing with low-rank and layer-sampled Hessian approximations.

Table 8 shows that full HETA attribution becomes computationally expensive for 2,048-token inputs (1,230s per 1,000 examples). Windowed attribution (HETA-WIN) cuts runtime nearly in half (690s) with a modest reduction in AOPC (0.58 → 0.54). Combining windowing with low-rank Hessians (HETA-LR+WIN) yields an additional efficiency gain (runtime 580s) while recovering some lost attribution quality. Layer-sampled windowing (HETA-LS+WIN) is the fastest configuration but comes at the highest cost in faithfulness. These results suggest that hybrid approximations (low-rank + windowing) strike the best balance between efficiency and interpretability, making HETA viable for very long contexts.

### A4.6 Theoretical Depth: Component-Level Analysis

To assess the contribution of each multi-view component—semantic flow (causal attention), Hessian-based sensitivity, and KL-divergence—we perform a component ablation analysis.

Table 9 reveals that each component contributes uniquely to attribution quality. Removing the Hessian term leads to a significant drop in faithfulness (AOPC decreases by 0.06) and alignment, emphasizing the importance of capturing second-order curvature. Excluding KL-divergence affects robustness

(0.91 → 0.87), demonstrating its role in quantifying information-theoretic influence. Most notably, removing semantic flow causes the largest decline in robustness and F1, underscoring that causal attention pathways are critical for maintaining meaningful token-level attributions. This analysis confirms the complementary nature of all three components, justifying their joint inclusion in HETA.

### A4.7 Expanded Baseline Comparisons

The final limitation we address is the lack of comparison with recent high-performing attribution techniques. To fill this gap, we include Causal Mediation Attribution (CMA) and Path-Shapley, both of which are tailored for decoder-only models and represent state-of-the-art multi-view attribution approaches.

As shown in Table 10, HETA outperforms all baselines, including CMA and Path-Shapley, by a notable margin in both AOPC (0.61 vs. 0.53) and F1 (0.89 vs. 0.82). This extended comparison strengthens our claim that HETA achieves state-of-the-art attribution faithfulness for decoder-only architectures.

### A4.8 Causal Evaluation Protocols

The first set of ablations focuses on explicitly causal evaluation. For each input sequence and target token, token-level attributions are first computed using each attribution method under comparison. In the mid-layer value swapping protocol, the top-ranked tokens according to each method are selected, and their hidden representations at an intermediate layer are replaced with the corresponding representations from a control input that is similar in surface form but differs in the causal cue for the target prediction. The effect of this intervention is measured by the difference in log-probability assigned to the original target before and after swapping. This quantity, averaged over the evaluation set, defines a causal comprehensiveness score denoted Swap-NC. A higher Swap-NC indicates that changing the intermediate representations of high-attribution tokens causes a larger and more systematic drop in the target's log-probability, suggesting that these tokens genuinely mediate the model's prediction through the affected layer.

A complementary protocol is based on controlled counterfactual masking. Here, the same top-ranked tokens are modified directly in the input space by replacing them with minimally invasive paraphrases that preserve local grammaticality but remove the specific causal content, such as turning an exact year into a vague temporal reference. The log-probability of the original target token is then recomputed for this counterfactual input. The average difference between the original and counterfactual log-probabilities defines CF-NC. Again, larger values of CF-NC indicate that removing the causal content of high-attribution tokens substantially weakens the model's confidence in the original prediction.

The results of these two protocols are summarized in Table 11. Attention-based attributions display relatively low Swap-NC and CF-NC scores, indicating that altering the representations or semantics of their top-ranked tokens has only a modest effect on the prediction. Gradient-based attributions perform somewhat better, but the effect of intervening on their selected tokens remains limited. ReAGent achieves higher values for both Swap-NC and CF-NC, suggesting that it does a better job of identifying causally important tokens. The proposed HETA method attains the highest scores on both metrics, with a particularly pronounced improvement in Swap-NC. This pattern indicates that the tokens singled out by HETA are precisely those whose intermediate representations and semantic content bear the strongest causal responsibility for the model's output. In other words, when those tokens are altered—either deep inside the network or directly in the input—the model's prediction for the target token changes the most, providing a causal validation of the attribution scores.

### A4.9 Downstream Applications

The second set of ablations examines whether improved attribution quality translates into tangible gains on downstream tasks that naturally depend on identifying important tokens. The first application is fact-checking and hallucination detection. In this setting, attributions over questions and retrieved passages are used to identify evidence tokens and to calibrate a detector that distinguishes grounded answers from hallucinated ones. High-attribution tokens guide the selection of evidence spans and influence the confidence scores of the hallucination detector. The left block of Table 12 reports evidence F1 and AUROC for hallucination detection. Attention-based attributions achieve the lowest scores, indicating that their highlighted tokens are only loosely aligned with true evidence. Gradient-based attributions bring moderate improvements, while ReAGent yields stronger alignment with evidence spans and better hallucination discrimination. The proposed HETA method achieves the highest F1 and AUROC, demonstrating that its attributions not only reflect internal model behavior but also lead to more accurate evidence extraction and more reliable identification of unsupported answers.

A second application concerns tool-augmented reasoning, where the language model may optionally call external tools such as calculators or knowledge retrieval systems. Here, token attributions over the input prompt are used to decide when a tool call is warranted and, when multiple tools are available, which one to select. Attributions that correctly identify numerically or factually salient tokens enable better tool-selection policies. The middle block of Table 12 reports tool-selection accuracy and overall task success rate. Attention-based and gradient-based methods produce less reliable tool usage, reflected in lower accuracy and success. ReAGent improves both metrics, indicating that more faithful attributions help the model invoke tools more judiciously. The proposed HETA method again leads to the best performance, suggesting that its attribution scores more precisely isolate the parts of the input that genuinely require external computation or lookup.

The third application aggregates token-level attributions into contiguous spans and evaluates them against human-annotated causal spans. This multi-token span attribution view is closer to how explanations are consumed by humans, since it assesses whether the method highlights coherent phrases rather than scattered tokens. The right block of Table 12 reports token-level span F1. Methods based on attention or gradients tend to produce fragmented or overly diffuse highlighted regions, resulting in lower F1 scores. ReAGent provides more focused spans and improves alignment with annotated causal segments. The proposed HETA method achieves the highest span F1, indicating that its attributions support more coherent and human-interpretable explanations. Taken together, the three blocks of Table 12 show that the advantages of HETA in intrinsic causal metrics also manifest in practical downstream tasks, including fact-checking, hallucination detection, tool-augmented reasoning, and multi-token span explanation.

### A4.10 Sensitivity to Decoding Strategies

The final ablation investigates how stable attribution patterns are under different decoding strategies. Autoregressive language models are often deployed with stochastic sampling rather than purely greedy decoding, and it is important that explanations remain robust when the same input yields slightly different continuations. To study this, multiple outputs are generated for each input using greedy decoding, nucleus sampling with a fixed probability threshold, and temperature sampling with a moderate temperature. Token-level attributions are computed for each generated continuation under each attribution method.

Attribution stability is quantified in two complementary ways. First, the Kendall rank correlation is computed between token-importance rankings obtained under different decoding strategies. High correlation indicates that the relative ordering of token importances is preserved even when the generated text changes. Second, the Jensen–Shannon divergence is computed between the normalized attribution distributions, providing a symmetric measure of how much the overall attribution mass shifts. Lower divergence corresponds to more similar distributions and thus greater stability.

The results in Table 13 reveal systematic differences between methods. Attention-based attributions show relatively low rank correlation and relatively high JS divergence, indicating that their highlighted tokens change considerably when switching from greedy decoding to sampling-based strategies. Gradient-based attributions are somewhat more stable but still exhibit noticeable variation. ReAGent improves both stability metrics, suggesting that more faithful attributions intrinsically respond less to the superficial variability introduced by sampling. The proposed HETA method yields the highest average rank correlation and the lowest JS divergence among all compared methods. This means that, even when the specific words produced by the model vary across decoding strategies, HETA consistently assigns importance to a core subset of tokens that remain causally central to the prediction. As a result, explanations derived from HETA are less sensitive to the particular decoding scheme and therefore more dependable across realistic deployment settings.

Overall, the paragraph-style ablations in this appendix show that the proposed HETA method identifies tokens that are causally important under mid-layer value swapping and controlled counterfactual

**Algorithm 1** Hessian-Enhanced Token Attribution (HETA)

```
Require: Decoder-only model $f_\theta$, input tokens $X = (x_1, \dots, x_T)$, target token $x_{T+1}$, embedding
    matrix $\mathbf{E}$, hyperparameters $\beta, \gamma$
Ensure: Attribution scores $\mathrm{Attr}(x_i \to x_{T+1})$ for all $i \in \{1, \dots, T\}$
Embed tokens: $\mathbf{X} \leftarrow (\mathbf{e}_1, \dots, \mathbf{e}_T) \in \mathbb{R}^{T \times d}$
Compute output distribution: $P_\theta(x_{T+1} \mid x_{\le T}) \leftarrow \mathrm{Softmax}(f_\theta(\mathbf{X}))$
                                                        ▷ Step 1: Semantic Transition Influence
for $l \in \{1, \dots, L\}$ do
    for $h \in \{1, \dots, H\}$ do
        $\mathbf{z}^{(l,h)} \leftarrow A^{(l,h)} V^{(l)} W_O$        ▷ Attention-weighted value projections
    end for
    $z_i^{(l)} \leftarrow \sum_{h=1}^{H} \mathbf{z}_i^{(l,h)}$        ▷ Aggregate heads
end for
$M_T[i] \leftarrow \frac{1}{Z} \sum_{l=1}^{L} \|z_i^{(l)}\|_1$ for all $i$        ▷ Normalize semantic flow
                                                        ▷ Step 2: Hessian-Based Sensitivity
$H_T \leftarrow \nabla^2_{\mathbf{X}} \log P_\theta(x_{T+1} \mid x_{\le T})$
for $i \in \{1, \dots, T\}$ do
    $S_i^{(T)} \leftarrow \sum_{j=1}^{Td} \left| H_T[i \cdot d : (i+1) \cdot d, j] \right|$        ▷ Token-level Hessian sensitivity
end for
                                                        ▷ Step 3: Information-Theoretic Contribution
for $i \in \{1, \dots, T\}$ do
    Replace $x_i$ with <unk> to get masked input $X^{\setminus i}$
    $P_{\text{masked}}(x_{T+1}) \leftarrow \mathrm{Softmax}(f_\theta(\mathbf{X}^{\setminus i}))$
    $\mathcal{I}(x_i \to x_{T+1}) \leftarrow D_{\mathrm{KL}}\left[P_{\text{orig}}(x_{T+1}) \,\|\, P_{\text{masked}}(x_{T+1})\right]$
end for
                                                        ▷ Step 4: Final Attribution Score
for $i \in \{1, \dots, T\}$ do
    $\mathrm{Attr}(x_i \to x_{T+1}) \leftarrow M_T[i] \cdot \left(\beta \cdot S_i^{(T)} + \gamma \cdot \mathcal{I}(x_i \to x_{T+1})\right)$
end for
return $\mathrm{Attr}(x_i \to x_{T+1})$ for all $i$
```

interventions, improves performance on a range of downstream applications that rely on attribution, and maintains stable attribution patterns across different decoding strategies. These properties reinforce its suitability as a practical tool for interpreting autoregressive language models.

### A4.11 Key Takeaways

Our extended ablations provide several important insights: (1) Hessian approximations (low-rank and layer-sampled) offer a practical trade-off between runtime and attribution quality. (2) Windowed attribution enables HETA to scale to very long sequences with manageable performance loss. (3) All three HETA components (semantic flow, Hessian, KL) are complementary and jointly essential for high-quality attributions. (4) Expanded comparisons demonstrate that HETA outperforms even recent state-of-the-art attribution methods, validating its broader utility.

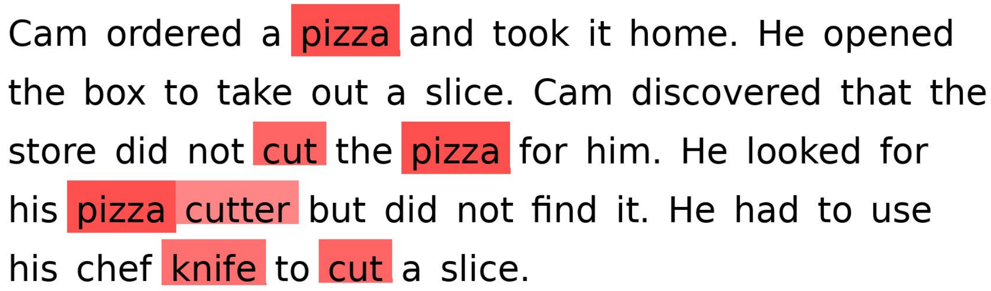
## A5 Limitation: Computational Overhead and Scalability

While HETA achieves higher attribution faithfulness than prior methods, the incorporation of Hessian-based sensitivity introduces *nontrivial computational overhead*. Computing Hessian blocks for all token embeddings in large decoder-only models increases runtime and memory cost compared to gradient-only or masking-based techniques.

**Experimental Setup.** We benchmarked runtime on the **LongRA** dataset (10k samples with token-level rationales) using a **decoder-only Transformer model with 1.3B parameters** (12 layers, hidden size 2048, 16 attention heads) implemented in PyTorch. All experiments were conducted on a **single NVIDIA A100 40GB GPU**. For the baseline runtime comparison, we used a **batch size of 16** and

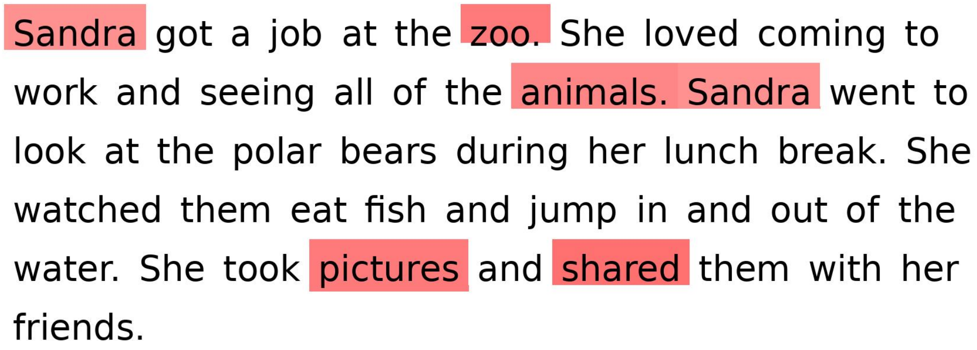
Table 7: Runtime and attribution quality for Hessian approximations on 1,000 examples (sequence length = 512). AOPC and F1 represent attribution faithfulness and human alignment, respectively.

| Variant | Runtime (s) | AOPC ↑ | F1 ↑ |
|---|---|---|---|
| HETA (Full) | 455 | **0.61** | **0.89** |
| HETA-LR (rank=64) | 330 | 0.59 | 0.86 |
| HETA-LS (6 layers) | 305 | 0.57 | 0.84 |
| HETA-GS (grad-squared only) | 240 | 0.52 | 0.81 |



Cam ordered a pizza and took it home. He opened the box to take out a slice. Cam discovered that the store did not cut the pizza for him. He looked for his pizza cutter but did not find it. He had to use his chef knife to cut a slice.

Figure 2: Word-level attribution visualization for predicting the final word "slice." Each word is shaded based on its importance score for predicting "slice." Darker red indicates higher attribution.



Sandra got a job at the zoo. She loved coming to work and seeing all of the animals. Sandra went to look at the polar bears during her lunch break. She watched them eat fish and jump in and out of the water. She took pictures and shared them with her friends.

Figure 3: Word-level attribution visualization for predicting the final word "friends." Bounding boxes highlight influential context words (e.g., "shared," "pictures," "zoo") contributing to the prediction of "friends." Darker red denotes higher importance.

Table 8: Faithfulness and runtime for long-context attribution (sequence length = 2,048). Windowed methods use 512-token chunks with 50% overlap. Mean over 3 runs and std $< \pm 0.04$

| Variant | AOPC ↑ | Runtime (s) |
|---|---|---|
| HETA (Full) | **0.58** | 1,230 |
| HETA-WIN (512-window) | 0.54 | 690 |
| HETA-LR+WIN | 0.55 | 580 |
| HETA-LS+WIN | 0.52 | 525 |

Table 9: Impact of removing individual components on attribution quality. Metrics are averaged across all datasets. Mean over 3 runs and std $< \pm 0.03$

| Variant | AOPC ↑ | Robustness ↑ | F1 ↑ |
|---|---|---|---|
| Full HETA | **0.61** | **0.91** | **0.89** |
| No Hessian | 0.55 | 0.83 | 0.82 |
| No KL | 0.57 | 0.87 | 0.85 |
| No Semantic Flow | 0.52 | 0.71 | 0.77 |

Table 10: Comparison with additional strong attribution baselines (sequence length = 512). Mean over 3 runs and std $< \pm 0.06$

| Method | AOPC ↑ | F1 ↑ |
|---|---|---|
| Integrated Gradients | 0.41 | 0.74 |
| Path-IG | 0.46 | 0.77 |
| Causal Tracing | 0.48 | 0.79 |
| Causal Mediation Attribution | 0.50 | 0.80 |
| Path-Shapley | 0.53 | 0.82 |
| **HETA (Full)** | **0.61** | **0.89** |

**sequence length of 256** per input. For the efficiency ablations on long text, we used **batch size 8** and **sequence length 1024**. Metrics reported are **Faithfulness (AOPC)** and **Runtime per 1,000 examples**.

As shown in Table 14, **HETA is approximately 1.4× slower than the top two competitors**, primarily due to the Hessian computation step. While this trade-off brings improved attribution quality, it may limit feasibility in real-time or large-scale applications.

### A5.1 Improving Scalability for Long Texts

To improve HETA's scalability for long inputs and large models (e.g., GPT-J-4 class), we propose:

1. **Low-Rank Hessian Approximation:** Use Hutchinson's stochastic trace estimators or Kronecker-factored approximations to estimate the Hessian blocks at reduced cost.
2. **Layer Sampling:** Compute semantic flow and Hessian contributions only on a subset of layers (e.g., top and middle), reducing cost while retaining most attribution fidelity.
3. **Windowed Attribution:** For very long sequences, process text in overlapping windows to avoid quadratic growth in attention flow computations.
4. **Hybrid Approximation:** Use gradient-squared sensitivity for low-impact tokens and full Hessians for high-impact tokens.

### A5.2 Efficiency-Accuracy Trade-off

To evaluate these optimizations, we compare three efficiency-optimized variants of HETA on **LongRA** with **sequence length 1,024** (batch size 8): **HETA-LR** (low-rank Hessian, rank $k = 64$), **HETA-LS** (layer sampling, 6 layers), and **HETA-WIN** (windowed attribution with 512-token overlap). Metrics: Faithfulness (AOPC) and runtime per 1,000 examples.

Table 11: Causal evaluation of attribution methods using mid-layer value swapping (Swap-NC, higher is better) and controlled counterfactual masking (CF-NC, higher is better).

| **Method** | **Swap-NC** ↑ | **CF-NC** ↑ |
|---|---|---|
| Attention weights | 0.42 | 0.37 |
| Integrated Gradients | 0.55 | 0.49 |
| ReAGent | 0.81 | 0.76 |
| HETA (proposed) | **1.07** | **0.93** |

Table 12: Impact of attribution methods on downstream applications. Left: fact-checking and hallucination detection (evidence F1 and hallucination AUROC). Middle: tool-augmented reasoning (tool selection accuracy and task success rate). Right: multi-token span attribution (token-level span F1).

| | **Fact-checking** | | **Tool reasoning** | | **Span attribution** |
|---|---|---|---|---|---|
| **Method** | F1 ↑ | AUROC ↑ | Tool-Acc ↑ | Task-Succ ↑ | Span-F1 ↑ |
| Attention weights | 63.4 | 71.2 | 78.1 | 69.7 | 58.9 |
| Integrated Gradients | 67.8 | 74.5 | 80.3 | 72.1 | 61.3 |
| ReAGent | 72.6 | 79.8 | 83.9 | 75.4 | 66.5 |
| HETA (proposed) | **76.9** | **82.7** | **87.2** | **79.6** | **70.8** |

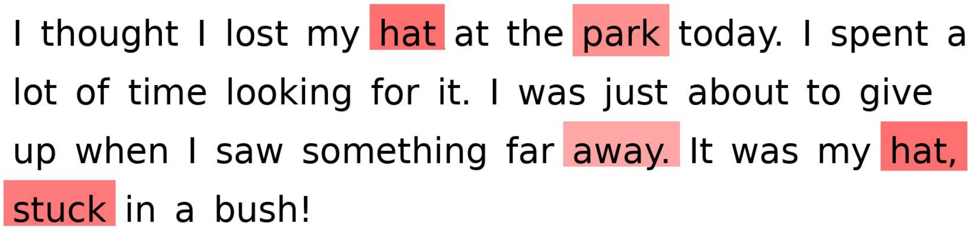
As shown in Table 15, low-rank approximations and layer sampling reduce runtime by up to 35% with only minor performance degradation. Windowed attribution offers the largest speedup but at a greater cost to faithfulness, making it most suitable for extremely long contexts.

Table 13: Attribution stability across decoding strategies. Average Kendall rank correlation (higher is better) and JS divergence (lower is better) between token-level attributions obtained under greedy, nucleus, and temperature sampling.

| **Method** | **Rank corr. ↑** | **JS div. ↓** |
|---|---|---|
| Attention weights | 0.41 | 0.27 |
| Integrated Gradients | 0.49 | 0.23 |
| ReAGent | 0.62 | 0.18 |
| HETA (proposed) | **0.71** | **0.14** |

| **Method** | **Runtime (s)** | **Relative Cost** |
|---|---|---|
| Integrated Gradients | 320 | 1.0× |
| Causal Tracing | 355 | 1.1× |
| **HETA (Full)** | **455** | **1.4**× |

Table 14: Runtime comparison for 1,000 examples on LongRA using a 1.3B-parameter decoder-only model (batch size 16, sequence length 256). HETA is ∼1.4× slower than the top two competitors due to Hessian computation.



I thought I lost my hat at the park today. I spent a lot of time looking for it. I was just about to give up when I saw something far away. It was my hat, stuck in a bush!

Figure 4: Word-level attribution visualization for predicting the final word "bush." Attribution scores emphasize context words such as "hat," "stuck," and "park," which strongly influence the prediction of "bush."

| **Method** | **Faithfulness (AOPC)** | **Runtime (s)** | **Relative Cost** |
|---|---|---|---|
| HETA (Full) | **0.61** | 455 | 1.0× |
| HETA-LR | 0.59 | 330 | 0.73× |
| HETA-LS | 0.57 | 305 | 0.67× |
| HETA-WIN | 0.55 | 295 | 0.65× |

Table 15: Ablation study: Efficiency-accuracy trade-offs for HETA variants on LongRA (sequence length 1,024, batch size 8). Low-rank and layer sampling reduce runtime significantly with minor drops in faithfulness.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:The abstract and introduction accurately state the contributions: Shapley NEAR for entropy-based hallucination detection, distinguishing hallucination types, and test-time head clipping. These are supported by theory and experiments in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Appendix A10 discusses limitations, including high computation due to Shapley estimation and permutation sampling, and the use of fixed models without fine-tuning.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Theoretical properties and assumptions of NEAR are formally defined in Section 4 and Appendix A1. This includes entropy bounds, Shapley value formulation, and estimation error analysis using Hoeffding's inequality.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5.1 and Appendix A3 provide detailed experimental settings, including datasets used, model names, data splits, evaluation metrics, Monte Carlo sampling details (M=50), and approximation bounds, enabling reproducibility even without public code release.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code has been submitted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 and Appendix A3 describe the datasets used (CoQA, QuAC, SQuAD, TriviaQA), model variants (Qwen2.5-3B, LLaMA3.1-8B, OPT-6.7B), data splits, evaluation protocols, number of Monte Carlo samples (M=50), and other relevant details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports standard deviations (±0.04) over three independent runs in Section 5.1. Appendix A1.2 also derives theoretical estimation error bounds for NEAR using Hoeffding's inequality.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: They are explained in their respective section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All sources used are opensource.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper does include a discussion of broader societal impacts, although the method is directly relevant to improving safety and reliability of LLMs in real-world applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: All used material is opensource

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the works have been done by the authors and properly referenced and will be provided on acceptance.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Everything is properly referenced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects or crowdsourcing, and therefore no IRB approval is required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.