

Text Attribution as Continuous Trajectories: Paired AUC Metrics and Mid-Layer Allocation-Effect Decomposition

Zihan Gu^{1,2}, Xinyan Zhang³, Tianyi Zhang⁴, Han Zhang⁵, Hua Zhang^{1,2, *}, Yue Hu^{1,2}

¹Institute of Information Engineering, Chinese Academy of Sciences

²School of Cyber Security, University of Chinese Academy of Sciences ³Independent Researcher

⁴State Grid Jiangxi Information & Telecommunication Company ⁵Shanghai Jiao Tong University

{guzihan, huyue, zhanghua}@iie.ac.cn

Abstract

Attribution aims to identify which input tokens are responsible for a language model’s predictions, yet its evaluation in NLP remains dominated by discrete Top- k removal tests that poorly reflect global importance and interaction effects. In this work, we reformulate text attribution evaluation as a continuous functional over attribution-induced trajectories and introduce two complementary AUC-based metrics tailored to autoregressive models. *Perturbation AUC* measures token necessity via progressive embedding-level corruption with a stable logit-based signal, while *Recovery AUC* evaluates sufficiency by tracking the restoration of internal representations in hidden-state space. Together, these metrics subsume Top- k tests and enable principled, ordering-sensitive assessment of attribution quality. Using these diagnostics, we find that many existing attribution methods remain close to random rankings in the studied settings and often fail to recover first-order interaction effects. Motivated by this failure mode, we propose a mid-layer attribution framework that uses an empirically observed layer-wise saturation pattern to decompose attribution into information allocation and causal effect. Across the evaluated models and tasks, our method achieves the best or near-best performance in most settings under Perturbation and Recovery AUC, yielding more faithful and interaction-aware explanations of language model behavior. Code is available at <https://github.com/Qihuai27/Mid-Layer-Attr>.

1 Introduction

Attribution is a central problem in explainable artificial intelligence (XAI), aiming to identify which components of an input most strongly influence a model’s decision (Dwivedi et al., 2023; Madsen et al., 2022). Across modalities, most attribution methods reduce this goal to producing an

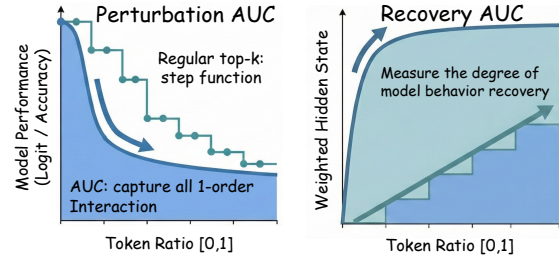


Figure 1: **Perturbation and Recovery AUC.** Continuous deletion- and insertion-style metrics for textual attribution.

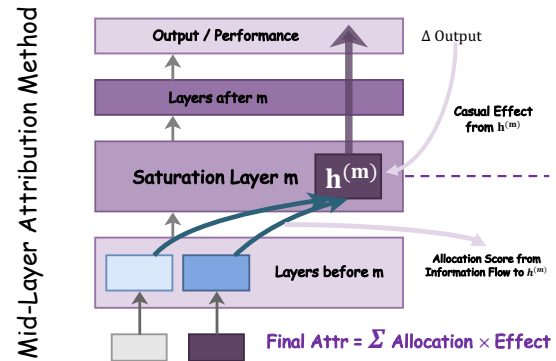


Figure 2: **Mid-layer hybrid attribution via allocation and effect.**

importance ranking over input factors (Selvaraju et al., 2020; Sundararajan et al., 2017; Scott et al., 2017). While this formulation is shared, attribution research in vision and language has diverged substantially in how such rankings are evaluated (Neely et al., 2022; Chan et al., 2022). In NLP, a substantial body of work has analyzed the reliability of attribution through diagnostic studies, faithfulness tests, and interaction analyses (Atanasova et al., 2020; Hooker et al., 2019; Atanasova et al., 2023; Jumelet and Zuidema, 2023; Mohebbi et al., 2023); nevertheless, the evaluation of attribution quality itself remains largely confined to discrete perturbation metrics.

Visual attribution is supported by a mature evaluation framework based on Insertion and Deletion AUC metrics (Petsiuk et al., 2018), which inter-

*Corresponding Author

pret a model’s prediction as a continuous functional of progressively revealed or removed information (Figure 1). In contrast, textual attribution is predominantly evaluated using discrete Top- k removal tests (DeYoung et al., 2020; Vafa et al., 2021), where the highest- or lowest-ranked tokens are masked and the resulting change in model behavior is observed. These tests provide only sparse local probes of an underlying continuous process.

This discrepancy is structural rather than incidental. Top- k metrics approximate a continuous attribution-induced trajectory using a finite set of evaluation points, analogous to approximating a measurable function by a simple function. As a result, they discard interaction effects between tokens, including first-order interactions, precisely the phenomena attribution methods are intended to capture (Janizek et al., 2021; Jumelet and Zuidema, 2023). From this perspective, AUC-style metrics strictly subsume Top- k evaluation as a degenerate case. The main obstacle to adopting AUC metrics in language is that text models lack a well-defined insertion process, and accuracy-based curves are unstable due to the discrete nature of the metric.

We address these issues by introducing two continuous evaluation metrics tailored to autoregressive language models, each targeting a distinct limitation of existing practices. *Perturbation AUC* can be viewed as a continuous generalization of traditional deletion-based Top- k tests: by replacing discrete accuracy probes with logit trajectories, it corrects the sparsity and trajectory blindness inherent in pointwise perturbation metrics. However, as a purely outcome-based criterion, Perturbation AUC primarily captures token *necessity* and remains insufficient to assess whether a ranking is *sufficient* to carry the model’s internal representations.

To address this complementary limitation, we introduce *Recovery AUC*, a representation-space insertion metric that measures how efficiently ranked tokens restore the model’s final hidden state from a fully masked baseline (Figure 1). By operating directly on hidden states, Recovery AUC provides a stable notion of sufficiency that is unavailable to output-based perturbation metrics. Together, Perturbation AUC and Recovery AUC form a paired evaluation framework that jointly characterizes necessity and sufficiency, yielding a more complete and modality-consistent assessment of textual attribution.

Using these metrics, we find that many existing

attribution methods, including gradient-based (Sundararajan et al., 2017; Ancona et al., 2018) and forward-flow-based (Abnar and Zuidema, 2020; Wang et al., 2023b) approaches, yield rankings that remain close to random orderings in several evaluated settings. Causal mediation analysis (Meng et al., 2022), while capable of identifying dominant tokens, typically produces sparse and unstable rankings that do not generalize across the input.

Motivated by these observations, we propose a hybrid attribution method that combines forward information flow decomposition with causal mediation analysis (Figure 2). We observe that in large language models, forward information accumulation saturates at an early layer: layers before this point primarily allocate and weight information, while subsequent layers compute task-specific features. We therefore attribute tokens by combining their forward flow contributions to the saturation layer with their causal effects on the output thereafter. This allocation–effect decomposition yields numerically stable, interaction-aware attribution scores without requiring combinatorial search. Experiments show that the proposed method achieves the best or near-best performance in most evaluated settings under Perturbation and Recovery AUC, providing more faithful explanations of language model behavior.

Our contributions can be summarized as follows:

- **Continuous attribution metrics.** We introduce Perturbation AUC and Recovery AUC, a paired set of continuous evaluation metrics for textual attribution in autoregressive language models. Perturbation AUC generalizes deletion-based Top- k tests to a continuous necessity criterion, while Recovery AUC complements it with a representation-level sufficiency measure.
- **Hybrid attribution method.** We propose a hybrid attribution framework that combines forward information flow decomposition with causal mediation analysis, exploiting a layer-wise saturation property of large language models to produce numerically stable and interaction-aware token attributions.
- **Unified empirical evaluation.** We conduct extensive experiments across models and baselines, showing that the proposed metrics reveal limitations of existing methods and that our approach achieves the best or near-best performance in most evaluated settings under Perturbation and Recovery AUC.

2 Problem Formulation and AUC-Based Evaluation

2.1 Text Attribution as an Ordering Problem

Let $x = (x_1, \dots, x_n)$ denote an input token sequence, and let $M(\cdot)$ be a pretrained language model. We denote the model’s specific output for a target prediction (e.g., logits) as $f(x) \in \mathbb{R}$, and its internal representations as $h(x)$. A text attribution method aims to identify which tokens in x are most influential to the model’s decision.

Crucially, regardless of their internal mechanisms, existing attribution methods ultimately induce an *ordering* over input tokens. We therefore formalize text attribution as a mapping

$$\mathcal{A} : x \mapsto \pi \in \mathfrak{S}_n,$$

where $\mathfrak{S}_n$ denotes the set of permutations over $\{1, \dots, n\}$, and $\pi(i)$ indicates the rank of token x_i in descending order of importance. Throughout this paper, we use the terms *ordering* and *ranking* interchangeably.

From this perspective, any real-valued attribution score is auxiliary; only the induced ordering π is observable and directly comparable. This separation allows us to rigorously assess whether π faithfully reflects the model’s decision boundary, independent of the scoring magnitude.

2.2 Evaluation as a Functional over Trajectories

To evaluate an ordering π , we analyze the evolution of model behavior under progressive information removal or restoration. Let $\mathcal{P}_t^\pi(x)$ denote a perturbation operator parameterized by a continuous variable $t \in [0, 1]$, modifying the input according to π . This defines a continuous trajectory $\gamma_\pi(t) = \mathcal{P}_t^\pi(x)$.

We treat attribution quality as a functional of this trajectory:

$$\mathcal{E}(\pi) = \int_0^1 \Phi(M(\gamma_\pi(t))) dt,$$

where $\Phi(\cdot)$ is a scalar metric derived from the model state, such as logit probability or representation similarity. This formulation emphasizes that evaluation is inherently continuous and depends solely on the ordering π , isolating the attribution method’s ranking quality from its implementation details.

2.3 Limitations of Top- k Evaluation

Most existing benchmarks rely on Top- k removal, effectively evaluating $\mathcal{E}(\pi)$ at a sparse set of discrete points. Formally, this corresponds to approximating the integral in Eq. (2) with coarse step functions. While computationally cheap, this discretization has two critical flaws relevant to interpretability.

First, it ignores **synergistic interactions**. Two tokens may individually have negligible impact but jointly determine the prediction. Top- k metrics, which remove tokens in isolation or rigid blocks, often fail to capture these combinatorial effects. Second, discrete removal creates **out-of-distribution (OOD)** inputs that violate the linguistic manifold. Sudden deletion of tokens can introduce high-frequency noise that dominates the model’s response, obscuring the true causal effect of the features. In contrast, AUC-style evaluation subsumes Top- k as a limiting case while providing a smoother, more robust measure of global importance.

2.4 AUC-Based Evaluation Metrics

Building on Section 2.2, we propose two metrics: *Perturbation AUC* (logit-space necessity) and *Recovery AUC* (representation-space sufficiency).

Notation. Let x be an input of length T . We discretize the evaluation trajectory into K steps. At step k , the ratio of processed tokens is $x_k = m_k/T$, where $m_k = \text{round}(k \cdot T/K)$.

2.4.1 Perturbation AUC (Necessity)

Perturbation AUC evaluates the *necessity* of highly ranked tokens by progressively corrupting them and measuring the resulting change in the model’s output logit.

Let e_i denote the embedding of token x_i . At step k , we perturb the embeddings of the top- m_k tokens in π as

$$e_i^{(k)} = (1 - \lambda)e_i + \lambda b + \sigma \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I),$$

where b is a baseline embedding and $\lambda \in [0, 1]$ controls the perturbation strength. All remaining token embeddings are left unchanged. Feeding the perturbed sequence into the model yields a logit vector $\ell^{(k)}$ at a target position, with $\ell^{(0)}$ denoting the unperturbed logits.

We quantify output preservation using cosine similarity

$$r_k = \cos(\ell^{(0)}, \ell^{(k)}),$$

which serves as the ordinate of the perturbation curve. The raw Perturbation AUC is computed via trapezoidal integration:

$$\text{AUC}_{\text{pert}}^{\text{raw}} = \sum_{k=0}^{K-1} \frac{1}{2} (r_k + r_{k+1}) (x_{k+1} - x_k).$$

To provide a common within-setting reference, we calibrate this metric against a random ordering baseline AUC_{rand} :

$$\text{AUC}_{\text{pert}} = \frac{\text{AUC}_{\text{pert}}^{\text{raw}} - \text{AUC}_{\text{rand}}}{1 - \text{AUC}_{\text{rand}}} + 1.$$

Interpretation. This normalization centers the random-ranking reference at 1.0 within each model–input setting; it is not intended to make raw scores universally commensurate across models. A value of $\text{AUC}_{\text{pert}} < 1$ indicates that the method successfully prioritizes necessary tokens (causing a faster logit drop than random), while a value > 1 suggests performance worse than random guessing. Lower values represent stronger necessity attribution.

2.4.2 Recovery AUC (Sufficiency)

As systematically analyzed by Ju et al. (2022), most existing attribution evaluation metrics in NLP rely on outcome perturbations (e.g., deletion or masking) and thus primarily measure the *destructiveness* of token removal, rather than whether the identified tokens are *sufficient* to support the model’s prediction. Such perturbation-based AUC metrics are therefore difficult to interpret as faithful attribution measures and are highly sensitive to the specific modification strategy.

To address this limitation, we introduce *Recovery AUC*, a sufficiency-based metric designed as a textual analogue of insertion AUC widely used in vision. Our key premise is that, for a frozen model, recovering the internal representations of the original input is a necessary condition for recovering its original output behavior. Consequently, attribution quality can be evaluated by how effectively the top-ranked tokens restore the model’s hidden states, rather than by how much damage their removal causes.

Formally, let H_{full} denote the final hidden states of the original input, and H_{base} those obtained from a fully masked baseline. We define a layer-wise distance metric

$$d(A, B) = \sum_{i=1}^T w_i \cos(A[i], B[i]),$$

where $w_i = 1/T$ by default. At step k , we restore the top- m_k tokens according to the attribution ranking, yielding hidden states H_k . The recovery ratio is defined as

$$R_k = 1 - \frac{d(H_k, H_{\text{full}})}{d(H_{\text{base}}, H_{\text{full}})}.$$

The Recovery AUC is computed as the integral of R_k over k . Higher values indicate that the top-ranked tokens form a compact and sufficient subset that captures the majority of the semantic information required to reconstruct the model’s internal representations.

2.5 Diagnosing Existing Attribution Methods.

We first use the proposed AUC metrics to diagnose common attribution baselines on LLAMA-3.2-3B-INSTRUCT (Meta, 2024) over three datasets: IOI (Wang et al., 2023a), Counterfact (Meng et al., 2022), and LongRA (Hong et al., 2025). We evaluate classical methods: Attention Rollout (Abnar and Zuidema, 2020), DePASS (Hong et al., 2025), Integrated Gradients (Sundararajan et al., 2017) Grad×Input (Hassani et al., 2017) and Leave-One-Out (LOO) (DeYoung et al., 2020).

Metric	Method	IOI	Counterfact	LongRA
P-AUC (Nec.)	Attention Rollout	1.20 ± 0.10	1.12 ± 0.23	1.04 ± 0.21
	DePASS	1.11 ± 0.08	0.95 ± 0.22	0.89 ± 0.18
	Integrated Gradients	1.04 ± 0.12	0.99 ± 0.22	0.86 ± 0.19
	Grad × Input	1.02 ± 0.11	0.92 ± 0.24	0.87 ± 0.14
	Leave-One-Out	0.88 ± 0.11	0.83 ± 0.34	0.77 ± 0.14
R-AUC (Suff.)	Attention Rollout	0.41 ± 0.03	0.37 ± 0.03	0.44 ± 0.02
	DePASS	0.38 ± 0.03	0.36 ± 0.04	0.40 ± 0.03
	Integrated Gradients	0.39 ± 0.04	0.35 ± 0.05	0.40 ± 0.05
	Grad × Input	0.38 ± 0.03	0.35 ± 0.04	0.41 ± 0.03
	Leave-One-Out	0.41 ± 0.03	0.38 ± 0.04	0.43 ± 0.04

Table 1: Evaluation of representative attribution methods on LLAMA-3.2-3B-INSTRUCT. P-AUC denotes Perturbation AUC (necessity), where lower values indicate stronger necessity. R-AUC denotes Recovery AUC (sufficiency), where higher values indicate stronger sufficiency. Results are reported as mean ± standard deviation across runs.

Table 1 summarizes the results. For **Perturbation AUC** (P-AUC; necessity), lower values indicate that removing top-ranked tokens quickly degrades the model, hence stronger necessity. Across all datasets, most methods exhibit relatively large P-AUC values (e.g., Attention Rollout reaches 1.20 on IOI), indicating weak necessity under destructive perturbation. Among the listed baselines, LOO is consistently the strongest necessity signal (lowest P-AUC: 0.88, 0.83, 0.77 on IOI/Counterfact/LongRA), while attention-based rankings are generally less effective (higher P-AUC).

For **Recovery AUC** (R-AUC; sufficiency), higher values indicate that restoring top-ranked tokens reconstructs internal representations more effectively. Unlike P-AUC, R-AUC varies only mildly across methods and datasets: all baselines cluster in a narrow band around 0.35–0.45. Even the best entries in Table 1 (e.g., 0.44 on LongRA) remain far from a regime where a small set of tokens can reliably recover the full hidden-state representation. This suggests that existing attribution methods capture, at best, limited sufficiency structure in LLM internal activations.

Taken together, Table 1 exposes a consistent diagnostic gap: current baselines provide weak improvements in both (i) ordering-sensitive destructive evaluation (P-AUC) and (ii) constructive recovery evaluation (R-AUC).

2.6 Limits of First-Order Attribution: Greedy vs. Optimal Selection

Most existing attribution methods assign *token-wise* importance scores and implicitly assume that multi-token explanations can be constructed by greedy aggregation. Under this assumption, a top- k explanation is obtained by independently selecting the k highest-scoring tokens. However, this procedure is only optimal when evidence contributions are additive and independent. In the presence of interaction effects, greedy selection can deviate substantially from the optimal token set.

To explicitly test this assumption, we conduct a first-order validation experiment on the Indirect Object Identification (IoI) task. IoI is a canonical benchmark in circuit tracing and mechanistic interpretability research (Wang et al., 2023a), where the model’s decision is known to depend on a small set of interacting tokens—most commonly an entity token and its associated relational cue. Extensive circuit analyses have established that the top-2 tokens driving the prediction constitute a verified ground truth, and the task exhibits clear second-order interaction effects between these tokens. This makes IoI an ideal controlled testbed: the correct token pair is not a matter of interpretation but a structurally validated property of the model. For each of 100 randomly sampled IoI instances, we compare the top-2 token set produced by each attribution method against a greedy oracle. The greedy oracle serves as an upper bound: it performs combinatorial search over token pairs and has been shown to recover the ground-truth interacting pair on IoI, thereby confirming that faithful attribution is fea-

sible in principle on this task. Nevertheless, the greedy oracle is computationally prohibitive in general due to its exponential search space.

We evaluate agreement with the greedy oracle under four progressively relaxed criteria: *Exact Match*, where both tokens and their order match; *Set Match*, where the selected token set matches regardless of order; *Top-1 Match*, where the most important token matches; and *Partial Match*, where at least one token overlaps. (The Token Shapley (Xiao et al., 2025) method was added here because the computational cost of calculating only the first two tokens is not high.)

Method	Exact	Set	Top-1	Partial
Leave-One-Out	20.0%	20.0%	96.0%	96.0%
DePASS	3.0%	4.0%	48.0%	84.0%
Token Shapley	2.0%	4.0%	27.0%	66.0%
Grad $\times$ Input	1.0%	1.0%	5.0%	25.0%
Grad SAM	1.0%	3.0%	1.0%	37.0%
Attention Rollout	0.0%	0.0%	1.0%	17.0%

Table 2: Agreement between attribution-based greedy top-2 selection and a greedy oracle on 100 IoI instances. Exact and Set measure strict set-level alignment, while Top-1 and Partial capture weaker forms of agreement.

As shown in Table 2, existing attribution methods exhibit strikingly low agreement with the greedy oracle under strict criteria. Exact and Set Match rates are near zero for most methods, indicating a failure to recover even the correct pair of interacting tokens. Even Top-1 Match rates vary widely across methods, and high Partial Match rates often mask the fact that only a single salient token is identified while its interacting counterpart is missed. We have included a visual example of this first-order interaction effect in the Appendix F.1.

These results demonstrate that current attribution methods struggle to capture even first-order interaction effects. In contrast, the greedy oracle reliably identifies the semantically critical token pair on IoI, typically corresponding to an entity token and its relational cue, which aligns with both human intuition about maximal semantic disruption and the task’s design principles.

This failure of greedy aggregation provides a concrete explanation for the limitations of Top- k -based evaluation. When attribution scores are unreliable at the first-order level, discrete selection metrics cannot faithfully assess attribution quality. By integrating model behavior along continuous perturbation and recovery trajectories, AUC-based evaluation exposes these interaction failures directly, rather than obscuring them behind isolated

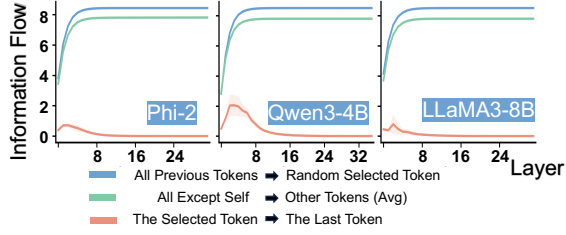


Figure 3: **Layer-wise information flow saturation on SST-2 (Socher et al., 2013)**. Cumulative forward flow statistics across three representative LLMs. The curves track information aggregated from context into random anchor tokens (**blue**), ambient background flow among non-anchors (**green**), and direct flow from anchors to the final prediction position (**red**). A consistent phase transition is visible: context aggregation saturates rapidly in early layers, while direct flow to the output remains negligible, indicating a structural separation between representation building and downstream reasoning.

selection tests.

3 Methodology

3.1 Motivation: Alignment with Causal Topology

Our diagnostic analysis reveals a fundamental mismatch: existing output-centric methods fail to capture the continuous interaction effects inherent to autoregressive generation. We argue that accurate attribution requires inspecting the model state where information allocation is decided, rather than where it is merely executed. To this end, we introduce a framework leveraging the *layer-wise saturation* property. By identifying the topological boundary where context integration stabilizes, we decouple information allocation from downstream reasoning, enabling stable and interaction-aware attribution.

3.2 Layer-wise Forward Information Flow and Saturation

We characterize *forward information flow* as the cumulative propagation of token information through attention and residual connections (Abnar and Zuidema, 2020; Wang et al., 2023b). Let $x = (x_1, \dots, x_T)$ be the input sequence processed by a transformer of depth L . Detailed definitions of flow computation and statistics are provided in Appendix C.1 and C.2.

The Saturation Phenomenon. Figure 3 illustrates the evolution of information flow across representative LLMs. We observe a recurring empirical transition in the evaluated transformer models:

1. *Rapid Aggregation* ($l < m$): In early layers, anchor tokens aggressively aggregate information from the context. This "in-flow" increases strictly and saturates rapidly (blue curves).
2. *Stable Propagation* ($l \geq m$): Once saturation is reached, the representation stabilizes. Notably, direct flow to the final prediction position (red curves) remains negligible until the very last layers.

The Critical Boundary m . This phenomenon suggests a structural separation: early layers function as a *contextual encoder* that builds a sufficient latent representation, while later layers operate on this integrated state to perform reasoning. We define the saturation layer m as the earliest depth where aggregation flow stabilizes (formal criterion in Appendix C.3).

We use layer m as a practical vantage point for attribution. At this boundary, information allocation has largely stabilized in our measurements, while the signals have not yet undergone the full downstream mixing of later layers. We therefore use the cumulative flow matrix $R^{(m)}$ as an empirical allocation map; Appendix G.2 tests the sensitivity of this choice.

3.3 Mid-Layer Attribution via Allocation–Effect Decomposition

Algorithm 1: Mid-Layer Attribution via Allocation–Effect Decomposition

Input: Input $x = (x_1, \dots, x_T)$; target position t ; model f ; saturation layer m

Output: Attribution scores $\{\text{attr}(i)\}_{i=1}^T$

Compute cumulative rollout $R^{(m)}$;

for $i = 1$ **to** T **do**

$s_1(i, \cdot) \leftarrow \text{Normalize}(R^{(m)}[:, i])$;

$p_{\text{orig}} \leftarrow g(x, t)$;

for $j = 1$ **to** T **do**

$p_{\text{mod}} \leftarrow g(\text{do}(h_j^{(m)} = 0), t)$;

$s_2(j) \leftarrow \max(0, p_{\text{orig}} - p_{\text{mod}})$;

for $i = 1$ **to** T **do**

$\text{attr}(i) \leftarrow \sum_{j=1}^T s_2(j) s_1(i, j)$;

return $\{\text{attr}(i)\}$

Based on the saturation property, we define attribution via a two-stage decomposition aligned with the model’s internal structure: *allocation*, which measures how input tokens are routed to the mid-layer, and *effect*, which measures the causal influence of mid-layer representations on the target prediction.

Let $h_j^{(m)} \in \mathbb{R}^d$ denote the hidden state at layer m and position j . Algorithm 1 specifies the resulting attribution procedure.

Clustered Variant. To reduce intervention cost and account for redundancy in mid-layer representations, we further introduce a clustered variant that performs causal intervention at the cluster level while preserving the allocation–effect structure. Full details are provided in Appendix D.2.

4 Experiments

4.1 Setup

Models. We evaluate attribution methods on three representative instruction-tuned large language models spanning different architectures and scales: LLaMA-3.2-3B (Meta, 2024), LLaMA-3.1-8B (Dubey et al., 2024), and Qwen-3-4B (Yang et al., 2025). All models are evaluated in a zero-shot setting without task-specific fine-tuning.

Datasets and Tasks. Experiments are conducted on three datasets capturing complementary attribution challenges: (i) IOI (Wang et al., 2023a) (841 samples), probing indirect object identification and long-range dependency resolution; (ii) Counterfact (Meng et al., 2022) (16,715 samples), evaluating factual recall and localized causal attribution; and (iii) LongRA (Hong et al., 2025) (573 samples), focusing on attribution robustness under long-context reasoning. Dataset construction, preprocessing, and evaluation splits are detailed in Appendix E.

Baselines. We compare our method against a diverse set of commonly used text attribution techniques, including attention-based methods (Attention Rollout (Abnar and Zuidema, 2020), DePASS (Hong et al., 2025)), gradient-based methods (Integrated Gradients (Sundararajan et al., 2017), Grad×Input (Hassani et al., 2017), GradSAM (Barkan et al., 2021)), and the perturbation-based Leave-One-Out baseline (DeYoung et al., 2020). All methods reported in Table 3 are *real-time* attribution methods, producing token-level importance scores with at most a small constant number of forward or backward passes.

To ensure scalability and fair comparison on large models and long sequences, we restrict the main evaluation to real-time methods. More expensive approaches based on exhaustive or combinatorial analysis (e.g., Token Shapley (Xiao et al.,

2025)) are therefore excluded from the main table and evaluated separately on small-scale settings (Appendix D.3).

Metric Computation. For efficiency and consistency, the main evaluation computes Perturbation AUC and Recovery AUC over the top ten ranked tokens, providing a computationally focused view of the early attribution trajectory. Appendix G.3 compares this protocol with the full curve and shows that the top-ten range does not systematically favor the proposed method.

4.2 Results

Table 3 reports attribution performance across models and datasets under the proposed AUC-based evaluation. Several consistent patterns emerge.

Existing methods struggle under AUC evaluation. Across all three models and tasks, most existing attribution methods perform close to random ranking under Perturbation AUC, and achieve limited Recovery AUC. This includes both attention-based methods (e.g., Attention Rollout, DePASS) and gradient-based methods (e.g., Integrated Gradients, Grad × Input). These results indicate that high attribution scores produced by such methods do not reliably correspond to tokens that are necessary or sufficient for maintaining model behavior under progressive perturbation.

Causal perturbation improves necessity but not ranking quality. Leave-One-Out and related perturbation-based baselines achieve lower P-AUC in some settings, reflecting improved necessity. However, their Recovery AUC remains limited, suggesting that while such methods can identify a small set of critical tokens, they fail to produce a coherent global ranking over the input sequence.

Mid-layer attribution performs best or near-best in most settings. Our proposed Mid-Layer Attribution achieves the best or near-best performance in most evaluated model–dataset settings under P-AUC and R-AUC. Complete P-AUC confidence intervals are reported in Appendix G.1.

Effect of clustering. The clustered variant (Mid-Layer Attribution++) further improves stability and, in several cases, achieves the highest Recovery AUC, particularly on longer-context tasks. This suggests that grouping redundant mid-layer representations helps capture interaction effects while reducing intervention noise.

Model	Method	IOI Task		Counterfact Task		LongRA Task	
		P-AUC ($\downarrow$)	R-AUC ($\uparrow$)	P-AUC ($\downarrow$)	R-AUC ($\uparrow$)	P-AUC ($\downarrow$)	R-AUC ($\uparrow$)
LLaMA-3.2 (3B, Instruct)	Attention Rollout	1.20 $\pm$ 0.10	0.41 $\pm$ 0.03	1.12 $\pm$ 0.23	0.37 $\pm$ 0.03	1.04 $\pm$ 0.21	0.44 $\pm$ 0.02
	DePASS	1.11 $\pm$ 0.08	0.38 $\pm$ 0.03	0.95 $\pm$ 0.22	0.36 $\pm$ 0.04	0.89 $\pm$ 0.18	0.40 $\pm$ 0.03
	Integrated Gradients	1.04 $\pm$ 0.12	0.39 $\pm$ 0.04	0.99 $\pm$ 0.22	0.35 $\pm$ 0.05	0.86 $\pm$ 0.19	0.40 $\pm$ 0.05
	Grad $\times$ Input	1.02 $\pm$ 0.11	0.38 $\pm$ 0.03	0.92 $\pm$ 0.24	0.35 $\pm$ 0.04	0.87 $\pm$ 0.14	0.41 $\pm$ 0.03
	Grad SAM	0.81 $\pm$ 0.10	0.44 $\pm$ 0.03	0.86 $\pm$ 0.23	0.36 $\pm$ 0.04	0.73 $\pm$ 0.12	0.48 $\pm$ 0.02
	Leave-One-Out	0.88 $\pm$ 0.11	0.41 $\pm$ 0.03	0.83 $\pm$ 0.34	0.38 $\pm$ 0.04	0.77 $\pm$ 0.14	0.43 $\pm$ 0.04
	Mid-Layer Attribution	0.80 $\pm$ 0.09	0.45 $\pm$ 0.03	0.75 $\pm$ 0.23	0.39 $\pm$ 0.04	0.69 $\pm$ 0.14	0.48 $\pm$ 0.02
	Mid-Layer Attribution++	0.84 $\pm$ 0.10	0.44 $\pm$ 0.03	0.76 $\pm$ 0.22	0.39 $\pm$ 0.04	0.74 $\pm$ 0.18	0.45 $\pm$ 0.02
LLaMA-3.1 (8B, Instruct)	Attention Rollout	1.32 $\pm$ 0.15	0.34 $\pm$ 0.02	1.28 $\pm$ 0.18	0.35 $\pm$ 0.05	1.19 $\pm$ 0.24	0.39 $\pm$ 0.03
	DePASS	1.02 $\pm$ 0.18	0.35 $\pm$ 0.03	0.98 $\pm$ 0.19	0.33 $\pm$ 0.05	0.87 $\pm$ 0.19	0.39 $\pm$ 0.03
	Integrated Gradients	0.99 $\pm$ 0.19	0.35 $\pm$ 0.05	0.95 $\pm$ 0.21	0.33 $\pm$ 0.06	0.94 $\pm$ 0.22	0.40 $\pm$ 0.03
	Grad $\times$ Input	0.99 $\pm$ 0.21	0.32 $\pm$ 0.04	0.99 $\pm$ 0.22	0.31 $\pm$ 0.05	0.84 $\pm$ 0.20	0.37 $\pm$ 0.03
	Grad SAM	0.80 $\pm$ 0.20	0.33 $\pm$ 0.04	0.87 $\pm$ 0.26	0.29 $\pm$ 0.05	0.61 $\pm$ 0.28	0.40 $\pm$ 0.03
	Leave-One-Out	0.71 $\pm$ 0.21	0.34 $\pm$ 0.03	0.90 $\pm$ 0.23	0.33 $\pm$ 0.05	0.79 $\pm$ 0.19	0.38 $\pm$ 0.03
	Mid-Layer Attribution	0.70 $\pm$ 0.20	0.36 $\pm$ 0.04	0.75 $\pm$ 0.24	0.33 $\pm$ 0.05	0.66 $\pm$ 0.32	0.41 $\pm$ 0.03
	Mid-Layer Attribution++	0.71 $\pm$ 0.18	0.37 $\pm$ 0.04	0.75 $\pm$ 0.23	0.33 $\pm$ 0.05	0.75 $\pm$ 0.25	0.42 $\pm$ 0.02
Qwen-3 (4B, Instruct)	Attention Rollout	1.32 $\pm$ 0.11	0.19 $\pm$ 0.01	1.16 $\pm$ 0.21	0.20 $\pm$ 0.02	1.20 $\pm$ 0.21	0.18 $\pm$ 0.01
	DePASS	0.96 $\pm$ 0.17	0.31 $\pm$ 0.05	1.01 $\pm$ 0.20	0.27 $\pm$ 0.06	0.89 $\pm$ 0.16	0.31 $\pm$ 0.04
	Integrated Gradients	0.95 $\pm$ 0.19	0.29 $\pm$ 0.04	0.97 $\pm$ 0.24	0.29 $\pm$ 0.06	0.95 $\pm$ 0.17	0.30 $\pm$ 0.04
	Grad $\times$ Input	0.89 $\pm$ 0.19	0.28 $\pm$ 0.04	0.94 $\pm$ 0.26	0.28 $\pm$ 0.06	0.81 $\pm$ 0.17	0.27 $\pm$ 0.03
	Grad SAM	0.83 $\pm$ 0.15	0.33 $\pm$ 0.03	0.88 $\pm$ 0.27	0.30 $\pm$ 0.05	0.61 $\pm$ 0.14	0.30 $\pm$ 0.04
	Leave-One-Out	0.62 $\pm$ 0.23	0.27 $\pm$ 0.04	0.84 $\pm$ 0.36	0.28 $\pm$ 0.08	0.80 $\pm$ 0.20	0.25 $\pm$ 0.04
	Mid-Layer Attribution	0.77 $\pm$ 0.17	0.48 $\pm$ 0.05	0.76 $\pm$ 0.25	0.45 $\pm$ 0.09	0.57 $\pm$ 0.22	0.48 $\pm$ 0.08
	Mid-Layer Attribution++	0.77 $\pm$ 0.16	0.49 $\pm$ 0.05	0.75 $\pm$ 0.26	0.45 $\pm$ 0.09	0.51 $\pm$ 0.19	0.53 $\pm$ 0.04

Table 3: **Comprehensive comparison of cross-model attribution performance.** P-AUC columns report means $\pm$ standard deviation; complete 95% confidence intervals are provided in Appendix G.1. R-AUC columns report mean $\pm$ standard deviation across examples.

Overall, these results indicate that, in the studied settings, aligning attribution with the internal organization of transformer computation often yields more faithful attributions under the proposed AUC metrics. Appendix G.2 further evaluates sensitivity to the selected mid-layer.

4.3 Qualitative Results

We use the IoI task to demonstrate the effectiveness of the Mid-Layer Attribution method. As shown in Figure 4, Mid-Layer Attribution accurately pinpoints the two tokens identified by the IoI mechanism, confirming that our method effectively localizes key tokens. The direct benefit of this performance is the numerical advantage in P-AUC and R-AUC—driven by the superiority of the early-stage curves (see Appendix F.2), which further demonstrates that our metrics can indeed distinguish the quality of different attribution methods.

5 Related Work

Attribution Metrics. Attribution fidelity has long been central to interpretability. In vision, insertion and deletion AUC were popularized by RISE (Petsiuk et al., 2018), exploiting the spatial

continuity of images. For text, however, insertion-style evaluation is less direct because language is discrete and order-sensitive. Consequently, text attribution has mainly relied on output perturbation tests (Nguyen, 2018; DeYoung et al., 2020; Chan et al., 2022). Such tests provide useful diagnostics but can conflate attribution quality with model robustness and capture only a narrow notion of causal contribution (Jacovi and Goldberg, 2020; Ju et al., 2022). Later work incorporates attribution scores into evaluation or aligns attribution with downstream behavior (Zhao and Aletras, 2023; Chrysostomou and Aletras, 2021; Yin et al., 2022). In contrast, we formulate evaluation as continuous logit- and representation-level trajectories, yielding paired AUC metrics for necessity and sufficiency in autoregressive language models.

Attribution Methods. Existing methods include gradient-based explanations (Hassani et al., 2017; Sundararajan et al., 2017; Scott et al., 2017; Chattopadhyay et al., 2018; Barkan et al., 2021; Zhao et al., 2024b,a) and Shapley-value-based approaches (Shapley, 1953; Xiao et al., 2025). For language models, practical attribution often favors efficient forward-flow or perturbation-based methods requiring only a small number of model eval-

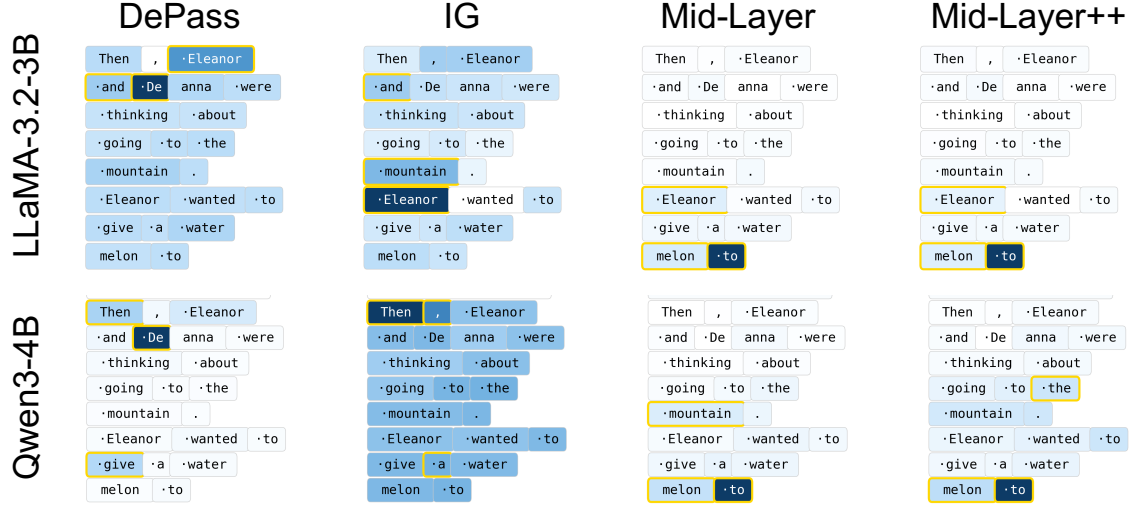


Figure 4: **High signal localization (Qwen3-4B and LLaMA-3.2-3B).** Visualization of attribution for an IOI sample. The gradient integration methods and the top three choices of DePASS exhibit a high degree of uncertainty; in contrast, the Mid-Layer Attribution choice is invariant across models and is the only method that selects the relational anchors "to" and "Eleanor".

uations (DeYoung et al., 2020; Vafa et al., 2021; Abnar and Zuidema, 2020; Zhao and Shan, 2024; Hong et al., 2025; Zhang et al., 2025). These methods scale to long sequences and autoregressive generation, but they usually produce token-wise rankings and are not designed to recover interaction structure. By contrast, many visual attribution methods use search or optimization to identify structured regions (Petsiuk et al., 2021; Dwivedi et al., 2023; Gu et al., 2026a,c), which is harder to transfer to text due to combinatorial cost. Our method instead combines mid-layer information allocation with causal effect estimation, targeting interaction-aware attribution without exhaustive search.

Faithfulness and Reliability of NLP Attribution. A substantial line of NLP work shows that attribution methods can be unstable and mutually inconsistent across gradient-, attention-, and perturbation-based explanations (Atanasova et al., 2020; Neely et al., 2022; Jain and Wallace, 2019; Wiegrefe and Pinter, 2019; Bastings and Filippova, 2020). This has motivated faithfulness diagnostics based on retraining ablations, counterfactual editing, adversarial sensitivity, and unified evaluation frameworks (Hooker et al., 2019; Atanasova et al., 2023; Kamahi and Yaghoobzadeh, 2024; Manna and Sett, 2024; Sun et al., 2025; Zaman and Srivastava, 2025), as summarized by recent surveys (Madsen et al., 2022; Lyu et al., 2024; Singh et al., 2024). Related analyses further question fixed Top- k evaluation (Kamp et al., 2023, 2025), emphasize feature interactions and informa-

tion flow (Jumelet and Zuidema, 2023; Choudhury et al., 2023; Janizek et al., 2021; Mohebbi et al., 2023; Kobayashi et al., 2020; Ferrando et al., 2022; Ferrando and Voita, 2024), and study explanation faithfulness in autoregressive LLMs (Madsen et al., 2024; Dehghanighobadi et al., 2025; Wang and Atanasova, 2025; Tutek et al., 2025; Lanham et al., 2023; Lan et al., 2025; Yin and Neubig, 2022; Sarti et al., 2023; Yu et al., 2024). Our work builds on these concerns but shifts the focus from comparing static score sets or fixed Top- k subsets to evaluating full ranking-induced trajectories and deriving interaction-aware token attributions.

6 Conclusion

This paper challenges the prevailing reliance on discrete Top- k removal for attribution evaluation, arguing that it can obscure feature interactions. We introduce **Perturbation AUC** and **Recovery AUC** to assess feature necessity and sufficiency along continuous trajectories. In the studied settings, this lens shows that many existing attribution rankings remain close to random baselines under global evaluation. We therefore propose a mid-layer attribution method based on an empirically observed information-saturation pattern, separating feature allocation from downstream causal effects. Our approach achieves the best or near-best fidelity in most evaluated settings without combinatorial search. These results support continuous AUC-based evaluation as a useful complement to discrete attribution tests for LLMs.

Limitations

First, while our AUC metrics offer a stable comparison framework, they remain **relative measures** of ordering quality. The field-wide lack of ground-truth explanations prevents absolute verification of correctness. Second, our method relies on the presence of a distinct information saturation layer. Although empirically stable in our experiments, the precise location of this layer may shift depending on the model architecture or fine-tuning strategy, potentially requiring calibration when applied to new model families.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62372448.

References

- Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *ACL*.
- Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*.
- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *ACL*, pages 283–294.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *EMNLP*, pages 3256–3274.
- Oren Barkan, Edan Haulon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. 2021. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2882–2887.
- Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.
- Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In *ACL*, pages 5029–5038.
- Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pages 839–847.
- Sagnik Ray Choudhury, Pepa Atanasova, and Isabelle Augenstein. 2023. Explaining interactions between text spans. In *EMNLP*, pages 12709–12730.
- George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *ACL*, pages 477–488.
- Zahra Dehghanighobadi, Asja Fischer, and Muhammad Bilal Zafar. 2025. Can LLMs explain themselves counterfactually? In *EMNLP*, pages 7787–7815.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*, pages 4443–4458.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models.
- Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and 1 others. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9):1–33.
- Javier Ferrando, Gerard I Gállego, and Marta R Costajussà. 2022. Measuring the mixing of contextual information in the transformer. In *EMNLP*, pages 8698–8714.
- Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *EMNLP*.
- Zihan Gu, Ruoyu Chen, Junchi Zhang, Yue Hu, Hua Zhang, and Xiaochun Cao. 2026a. Phasewin search framework enable efficient object-level interpretation. In *CVPR*.
- Zihan Gu, Ruoyu Chen, Junchi Zhang, Li Liu, Xiaochun Cao, and Hua Zhang. 2026b. [PhaseWin: An efficient search algorithm for faithful visual attribution](#). *Preprint*, arXiv:2606.18008. Extended version.
- Zihan Gu, Jiayu Wang, Hua Zhang, and Yue Hu. 2026c. A good initialization is all you need for faithful visual attribution. *arXiv preprint arXiv:2607.06726*.
- Hamed Hassani, Mahdi Soltanolkotabi, and Amin Karbasi. 2017. Gradient methods for submodular maximization. *Advances in Neural Information Processing Systems*, 30.

- Xiangyu Hong, Che Jiang, Kai Tian, Biqing Qi, Youbang Sun, Ning Ding, and Bowen Zhou. 2025. Depass: Unified feature attributing by simple decomposed forward pass. In *NeurIPS*.
- Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, pages 9737–9748.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *ACL*, pages 4198–4205.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *NAACL*, pages 3543–3556.
- Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54.
- Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. 2022. Logic traps in evaluating attribution scores. In *ACL*, pages 5911–5922.
- Jaap Jumelet and Willem Zuidema. 2023. Feature interactions reveal linguistic structure in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8697–8712.
- Sepehr Kamahi and Yadollah Yaghoobzadeh. 2024. Counterfactuals as a means for evaluating faithfulness of attribution methods in autoregressive language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 452–468.
- Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2023. Dynamic top-k estimation consolidates disagreement between feature attribution methods. In *EMNLP*, pages 6190–6197.
- Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2025. Learning from sufficient rationales: Analysing the relationship between explanation faithfulness and token-level regularisation strategies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing in the Asia-Pacific Region (IJCNLP-AAACL)*.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *EMNLP*, pages 7057–7075.
- Tian Lan, Jinyuan Xu, Xue He, Jenq-Neng Hwang, and Lei Li. 2025. Attention consistency for LLMs explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1736–1750.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.
- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural NLP: A survey. *ACM Comput. Surv.*, 55(8):1–42.
- Supriya Manna and Niladri Sett. 2024. Faithfulness and the notion of adversarial sensitivity in NLP explanations. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 193–206.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *NeurIPS*.
- AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December.
- Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupala, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3378–3400.
- Michael Neely, Stefan F Schouten, Maurits Bleeker, and Ana Lucic. 2022. A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. In *HAI2022: Augmenting Human Intellect*, volume 354 of *Frontiers in Artificial Intelligence and Applications*, pages 60–78. IOS Press.
- Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *ACL*, pages 1069–1078.
- Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, page 151.
- Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. 2021. Black-box explanation of object detectors via saliency maps. In *CVPR*, pages 11443–11452.
- Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. In *ACL*, pages 421–435.
- M Scott, Lee Su-In, and 1 others. 2017. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774.

- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359.
- Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.
- Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *ACL*.
- Jingyi Sun, Pepa Atanasova, and Isabelle Augenstein. 2025. Evaluating input feature explanations through a unified diagnostic evaluation framework. In *NAACL*, pages 10559–10577.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328.
- Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *EMNLP*, pages 9935–9960.
- Keyon Vafa, Yuntian Deng, David Blei, and Alexander M Rush. 2021. Rationales for sequential predictions. In *EMNLP*, pages 10314–10332.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *ICLR*.
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. In *EMNLP*, pages 9840–9855.
- Yingming Wang and Pepa Atanasova. 2025. Self-critique and refinement for faithful natural language explanations. In *EMNLP*, pages 8481–8507.
- Sarah Wiegrefe and Yuval Pinter. 2019. Attention is not not explanation. In *EMNLP*, pages 11–20.
- Yingtai Xiao, Yuqing Zhu, Sirat Samyoun, Wanrong Zhang, Jiachen T Wang, and Jian Du. 2025. Token-shapley: Token level context attribution with shapley value. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3882–3894.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report.
- Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2022. On the sensitivity and stability of model interpretations in nlp. In *ACL*, pages 2631–2647.
- Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *EMNLP*, pages 184–198.
- Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the parametric knowledge of language models: A unified framework for attribution methods. In *ACL*, pages 8173–8186.
- Kerem Zaman and Shashank Srivastava. 2025. A causal lens for evaluating faithfulness metrics. In *EMNLP*, pages 29425–29449.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS*.
- Yihua Zhang, Changsheng Wang, Yiwei Chen, Chongyu Fan, Jinghan Jia, and Sijia Liu. 2025. The fragile truth of saliency: Improving llm input attribution via attention bias optimization. In *NeurIPS*.
- Chenyang Zhao, Janet H Hsiao, and Antoni B Chan. 2024a. Gradient-based instance-specific visual explanations for object specification and object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B Chan. 2024b. Gradient-based visual explanation for transformer-based clip. In *ICML*, pages 61072–61091.
- Zhixue Zhao and Nikolaos Aletras. 2023. Incorporating attribution importance for improving faithfulness metrics. In *ACL*, pages 4732–4745.
- Zhixue Zhao and Boxuan Shan. 2024. Reagent: A model-agnostic feature attribution method for generative language models. *arXiv preprint arXiv:2402.00794*.

A Theoretical Justification for AUC Metrics

A.1 Perturbation AUC as a Generalization of Discrete Metrics

We posit that Perturbation AUC is not merely an alternative metric, but a generalization that subsumes discrete evaluation methods. Let $\mathcal{T} = \{(k, s_k) \mid k \in [0, T]\}$ be the perturbation trajectory induced by a ranking π , where s_k is the model response after removing top- k tokens. Prior metrics can be mapped to specific geometric properties of $\mathcal{T}$:

- **Top- k Removal:** Corresponds to a single point evaluation s_k at a fixed k . This ignores the path taken to reach state k .
- **Leave-One-Out (LOO):** Equivalent to sampling the trajectory at $k = 1$ for distinct single-token masks.
- **Faithfulness / Sufficiency:** Often measures the local gradient $\Delta s_k = s_k - s_{k-1}$ or end-point differences $s_T - s_0$.
- **Random Baselines:** Corresponds to comparing the integral of π against the expected integral of a uniform permutation distribution.

By defining $\text{AUC}_{\text{pert}} = \frac{1}{T} \sum s_k$, we aggregate these isolated geometric features into a global measure of **necessity**. It captures the robustness of the model’s prediction under continuous information removal.

A.2 Separating Output Robustness from Internal Sufficiency

While Perturbation AUC measures *output degradation* (Necessity), it fails to guarantee that the identified tokens are mechanistically responsible for the prediction. In deep autoregressive models, information redundancy allows the model to compensate for missing tokens, maintaining high output probabilities even when causal drivers are removed.

To address this, **Recovery AUC** targets **sufficiency** at the representation level. We define the recovery ratio R_k based on the restoration of internal hidden states H :

$$R_k = 1 - \frac{d(H_k, H_{\text{full}})}{d(H_{\text{base}}, H_{\text{full}})}.$$

Unlike output logits, the internal state H encodes the full context required for future generation. High

AUC_{rec} implies that the subset of tokens effectively compresses the causal history of the generation process.

A.3 Complementarity

The two metrics form a rigorous dual evaluation:

1. **Perturbation AUC (Necessity):** Does removing these tokens break the prediction?
2. **Recovery AUC (Sufficiency):** Does keeping *only* these tokens reconstruct the mental state?

A faithful attribution method must maximize both simultaneously, ensuring that the ranked tokens are both necessary for the output and sufficient to represent the underlying reasoning process.

B Implementation Details

All experiments were conducted on NVIDIA RTX 4090 and RTX 5090 GPUs. We distinguish between *gradient-based* and *non-gradient-based* attribution methods due to their different memory and computation characteristics.

Hardware and Memory Consumption. For the LLaMA-8B family, gradient-based methods require approximately ~ 31 GB GPU memory, while non-gradient-based methods require ~ 16 GB. For 3B/4B models, the corresponding memory footprints are ~ 16 GB (gradient-based) and ~ 8 GB (non-gradient-based). Gradient-based experiments were primarily executed on RTX 5090, whereas non-gradient-based methods were executed on both RTX 4090 and RTX 5090. Except for Token Shapley, non-gradient-based methods complete a full attribution pass in ~ 1 s per sample, while gradient-based methods take ~ 3 s.

Metric Implementation. For Perturbation AUC, we adopt three fixed random seeds and generate three independent random masking trajectories as a randomized baseline, reporting the mean score to ensure fair comparison. We do not report per-sample variance of the random baseline, because in large-scale experiments such variance cancels out under global averaging. Unless otherwise specified, all reported variances in the paper denote *across-sample* variance, reflecting the stability of attribution methods with respect to distributional variation across inputs.

Numerical computations were performed in bfloat16 precision to balance range and stability. Unless otherwise specified, hyperparameters

for baseline methods follow their original mathematical formulations or standard library defaults.

C Details of Information Flow and Saturation

C.1 Forward Information Flow

We formalize forward information flow based on attention rollout (Abnar and Zuidema, 2020) and information flow (Wang et al., 2023b; Gu et al., 2026b), adapted to autoregressive transformers with causal masking.

For each layer $l \in \{1, \dots, L\}$, let

$$A^{(l)} \in \mathbb{R}^{H \times T \times T}$$

denote the attention tensor with H heads. We first aggregate attention across heads:

$$\bar{A}^{(l)} = \frac{1}{H} \sum_{h=1}^H A_h^{(l)}.$$

To account for residual connections, we mix attention with the identity:

$$\tilde{A}^{(l)} = \alpha I + (1 - \alpha) \bar{A}^{(l)},$$

where $\alpha \in [0, 1]$ controls the residual contribution. Unless otherwise specified, we set $\alpha = 0.5$.

The cumulative forward information flow up to layer l is defined recursively as

$$\begin{aligned} R^{(0)} &= I, \\ R^{(l)} &= \tilde{A}^{(l)} R^{(l-1)}. \end{aligned}$$

Equivalently,

$$R^{(l)} = \tilde{A}^{(l)} \tilde{A}^{(l-1)} \dots \tilde{A}^{(1)}.$$

The entry $R^{(l)}[i, j]$ represents the total amount of information flowing from token position j to position i through all attention and residual paths up to layer l .

C.2 Layer-wise Flow Statistics

To summarize layer-wise forward information flow, we define three statistics computed from the cumulative flow matrix $R^{(l)}$.

Anchor selection. Let $P \subset \{1, \dots, T\}$ denote a set of anchor positions, sampled uniformly at random. Let t denote the final prediction position.

Aggregation flow ($S_{\text{agg}}^{(l)}$). This statistic measures how much information anchor tokens aggregate from all previous tokens:

$$\begin{aligned} S_{\text{agg}}^{(l)} &= \frac{1}{|C_{\text{agg}}|} \sum_{(i,j) \in C_{\text{agg}}} R^{(l)}[i, j], \\ C_{\text{agg}} &= \{(p_k, j) \mid p_k \in P, j < p_k\}. \end{aligned}$$

Background flow ($S_{\text{bg}}^{(l)}$). This statistic measures average information flow among non-anchor tokens:

$$\begin{aligned} S_{\text{bg}}^{(l)} &= \frac{1}{|C_{\text{bg}}|} \sum_{(i,j) \in C_{\text{bg}}} R^{(l)}[i, j], \\ C_{\text{bg}} &= \{(i, j) \mid j < i\} \setminus C_{\text{agg}} \setminus C_{\text{out}}. \end{aligned}$$

Output flow ($S_{\text{out}}^{(l)}$). This statistic measures information flowing from anchor tokens to the final prediction position:

$$\begin{aligned} S_{\text{out}}^{(l)} &= \frac{1}{|C_{\text{out}}|} \sum_{(i,j) \in C_{\text{out}}} R^{(l)}[i, j], \\ C_{\text{out}} &= \{(t, p_k) \mid p_k \in P\}. \end{aligned}$$

These three statistics correspond to the blue, green, and red curves in Figure 3, respectively.

C.3 Saturation Criterion

We define the saturation layer m as the earliest layer at which forward information flow stabilizes across all statistics. Formally, let $S_x^{(l)}$ denote any of the flow statistics with $x \in \{\text{agg}, \text{bg}, \text{out}\}$. We select m as the smallest layer satisfying

$$\begin{aligned} \frac{|S_x^{(l)} - S_x^{(l-1)}|}{|S_x^{(l-1)}| + \delta} &< \varepsilon, \\ \forall x, \forall l &\geq m, \end{aligned}$$

where ε is a small threshold and $\delta > 0$ avoids numerical instability.

Empirically, the resulting m is highly stable across random anchor selections, datasets, and model architectures, supporting its interpretation as an intrinsic structural boundary.

C.4 More Results About Information Saturation

To further examine the robustness of the proposed saturation phenomenon, we conduct a systematic analysis across multiple model architectures and datasets.

Model	Dataset	# Layers	m_{agg}
Phi-2	SST-2	32	6
Phi-2	AG News	32	6
Qwen3-4B	SST-2	36	7
Qwen3-4B	AG News	36	7
LLaMA-3.2-3B-Instruct	SST-2	28	6
LLaMA-3.2-3B-Instruct	AG News	28	6
LLaMA-3-8B-Instruct	SST-2	32	6
LLaMA-3-8B-Instruct	AG News	32	6

Table 4: **Aggregation saturation layers across models and datasets.** For each model–dataset pair, we report the total number of transformer layers and the earliest layer at which aggregation flow $S_{\text{agg}}^{(l)}$ reaches a stable plateau according to the criterion in Section C.3.

Specifically, we construct cumulative forward information flow using attention rollout, as defined in Section C.1, and track the layer-wise evolution of the aggregation statistic $S_{\text{agg}}^{(l)}$. For each configuration, we record the earliest layer at which $S_{\text{agg}}^{(l)}$ enters a stable plateau according to the saturation criterion in Section C.3. We denote this layer as the aggregation saturation point m_{agg} .

Experiments are conducted on four representative autoregressive language models, spanning different parameter scales and architectural families, and evaluated on two standard text classification datasets (SST-2 (Socher et al., 2013) and AG News (Zhang et al., 2015)). For each model–dataset pair, we report the total number of layers and the corresponding saturation layer. The complete results are summarized in Table 4.

Across all settings, we observe a consistent early saturation behavior: aggregation flow stabilizes after a small number of layers, typically within the first 20%–25% of the network depth. Notably, the saturation layer remains highly stable across datasets for a fixed model, suggesting that information aggregation saturation is primarily governed by architectural properties rather than task-specific factors.

These results provide additional empirical evidence that forward information flow exhibits an intrinsic structural boundary, beyond which additional layers contribute marginally to further aggregation. This observation motivates our later design choices for layer selection in attribution and intervention experiments.

Selection of the saturation layer m . Our layer-wise analysis (Table 4) indicates that information aggregation typically stabilizes early (e.g., layers 6–8 for the LLaMA-3 families). We use a uni-

form, conservative setting of $m = 10$ in the main experiments, placing the intervention downstream of the observed boundary and avoiding instance-specific layer tuning. The local sweep in Appendix G.2 shows that performance does not collapse for nearby choices, while also making clear that no single layer is universally optimal.

D Details of Mid-Layer Attribution

D.1 Mid-Layer Attribution (V1): Full Formulation

Let m be the saturation layer.

Allocation.

$$\begin{aligned}\hat{s}_1(i, j) &= R^{(m)}[j, i], \\ s_1(i, j) &= \frac{\hat{s}_1(i, j)}{\sum_{k=1}^T \hat{s}_1(i, k)}.\end{aligned}$$

Effect.

$$s_2(j) = \max\left(0, g(x, t) - g(\text{do}(h_j^{(m)} = 0), t)\right).$$

Attribution.

$$\text{attr}(i) = \sum_{j=1}^T s_2(j) s_1(i, j).$$

D.2 Clustered Mid-Layer Attribution (V2)

Let $\mathcal{C}$ denote clusters of mid-layer positions.

Clustering.

$$\begin{aligned}\text{sim}(i, j) &= \frac{h_i^{(m)} \cdot h_j^{(m)}}{\|h_i^{(m)}\| \|h_j^{(m)}\|}, \\ (i, j) \in c &\iff \text{sim}(i, j) > \mu + k\sigma.\end{aligned}$$

Cluster effect.

$$s_2(c) = \max\left(0, g(x, t) - g(\text{do}(h_j^{(m)} = 0, \forall j \in c), t)\right).$$

Attribution.

$$\text{attr}(i) = \sum_{c \in \mathcal{C}} s_2(c) \max_{j \in c} s_1(i, j).$$

Dataset	Method	P-AUC ↓	R-AUC ↑
IOI	Attention Rollout	1.199 ± 0.095	0.408 ± 0.026
	DePASS	1.109 ± 0.076	0.382 ± 0.032
	Integrated Gradients	1.044 ± 0.117	0.386 ± 0.038
	Grad × Input	1.017 ± 0.108	0.377 ± 0.034
	Grad SAM	0.814 ± 0.102	0.444 ± 0.032
	Leave-One-Out	0.876 ± 0.106	0.412 ± 0.031
	Mid-Layer Attribution	0.802 ± 0.093	0.449 ± 0.027
Counterfact	Mid-Layer Attribution++	0.837 ± 0.096	0.437 ± 0.033
	Attention Rollout	1.122 ± 0.229	0.368 ± 0.035
	DePASS	0.952 ± 0.218	0.358 ± 0.042
	Integrated Gradients	0.987 ± 0.218	0.353 ± 0.055
	Grad × Input	0.921 ± 0.235	0.348 ± 0.044
	Grad SAM	0.857 ± 0.232	0.359 ± 0.039
	Leave-One-Out	0.826 ± 0.336	0.379 ± 0.039
	Mid-Layer Attribution	0.747 ± 0.227	0.385 ± 0.043
	Mid-Layer Attribution++	0.762 ± 0.223	0.390 ± 0.044

Table 5: **Comparison of attribution methods including Token Shapley on LLaMA-3.2-3B-Instruct (100 samples).** Lower Perturbation AUC (P-AUC) indicates better necessity; higher Recovery AUC (R-AUC) indicates better sufficiency. Results are reported as mean ± standard deviation.

Dataset	Method	P-AUC ↓	R-AUC ↑
IOI	Attention Rollout	1.351 ± 0.096	0.191 ± 0.008
	DePASS	0.968 ± 0.148	0.306 ± 0.046
	Integrated Gradients	0.947 ± 0.208	0.295 ± 0.042
	Grad × Input	0.844 ± 0.201	0.288 ± 0.039
	Grad SAM	0.759 ± 0.199	0.331 ± 0.030
	Leave-One-Out	0.602 ± 0.269	0.272 ± 0.044
	Mid-Layer Attribution	0.700 ± 0.159	0.483 ± 0.057
Counterfact	Mid-Layer Attribution++	0.712 ± 0.166	0.491 ± 0.050
	Attention Rollout	1.205 ± 0.211	0.204 ± 0.021
	DePASS	0.995 ± 0.255	0.264 ± 0.053
	Integrated Gradients	0.924 ± 0.250	0.293 ± 0.059
	Grad × Input	0.902 ± 0.275	0.276 ± 0.062
	Grad SAM	0.796 ± 0.397	0.297 ± 0.051
	Leave-One-Out	0.773 ± 0.466	0.284 ± 0.088
	Mid-Layer Attribution	0.708 ± 0.250	0.442 ± 0.095
	Mid-Layer Attribution++	0.727 ± 0.235	0.447 ± 0.084

Table 6: **Comparison of attribution methods including Token Shapley on Qwen3-4B (100 samples).** Token Shapley is evaluated on the same subset for reference. Results are reported as mean ± standard deviation.

D.3 Mid-Layer Attribution vs Token-Shapley on Small Experiment Set

As a complementary reference, we compare mid-layer attribution with token-level Token Shapley on a restricted set of 100 samples from IOI and Counterfact. Token Shapley is conceptually grounded in marginal contribution, but its exponential computational cost makes large-scale evaluation infeasible in practice.

Tables 5 and 6 show that Mid-Layer Attribution or its enhanced variant achieves lower Perturbation AUC and higher Recovery AUC than Token Shapley in the evaluated small-scale settings. In particular, Token Shapley exhibits weak necessity: many top-ranked tokens can be removed with limited impact on model behavior, resulting in rel-

atively high P-AUC values. This suggests that token-level marginal contribution does not reliably capture minimal causal support in autoregressive language models.

The performance gap is especially pronounced on Counterfact, where attribution noise is amplified by the need for precise causal localization. In contrast, mid-layer attribution remains stable under both necessity- and sufficiency-based metrics, indicating that evaluating attribution at the representation level is better aligned with the internal computation of large language models. Overall, Token Shapley serves as a useful but limited diagnostic baseline, while mid-layer attribution provides a more effective and scalable alternative under the proposed AUC metrics.

E Details of Datasets

Case 0

Indirect Object Identification

"Then, Eleanor and Deanna were thinking about going to the mountain. Eleanor wanted to give a watermelon to"

Answer:
Deanna

Case 1

Indirect Object Identification

"After the lunch, Mildred and Eleanor went to the market. Eleanor gave a watermelon to"

Answer:
Mildred

Case 2

Indirect Object Identification

"When Paula and Lee got a plum at the bar, Lee decided to give it to"

Answer:
Paula

Figure 5: **IOI dataset examples.** The target answer token is highlighted in brackets.

E.1 Why These Three Datasets

We use three datasets that stress complementary facets of attribution in autoregressive LLMs. As shown in Figures 5–7, each dataset provides a distinct supervision signal (a single target token) while inducing qualitatively different dependency structures in the prompt.

IOI (syntactic/positional binding). IOI prompts are constructed to elicit indirect-object identification under competing entities. The correct next-token prediction hinges on localized, syntactic binding and position-sensitive cues (e.g., which name occupies the recipient role), making IOI a controlled testbed for whether an attribution method can correctly localize *short-range* causal evidence rather than dispersing mass across semantically plausible distractors.

CounterFact (factual relation retrieval). CounterFact-style prompts query relational knowledge (subject–relation–object) and require the model to retrieve a specific attribute (e.g., nationality, religion, instrument). Here attribution is expected to concentrate on subject mentions and relation triggers, probing whether token importance aligns with *knowledge access* rather than surface-form heuristics.

LongRA (long-context robustness). LongRA examples embed the decisive clue within a longer narrative context. Despite the target being a single next token, the relevant evidence can be separated from the prediction point by substantial context, stressing whether attribution remains stable under *long-range* dependencies and potential attention sink/recency biases.

E.2 Unified View for Attribution Evaluation

Across all three datasets, each instance defines a prompt and a single target token to be predicted. Let the tokenized prompt be $x_{1:T}$ and the target be y . Attribution methods output scores s_i over prompt tokens x_i , which we subsequently evaluate under perturbation- and recovery-based protocols. IOI, CounterFact, and LongRA differ primarily in the effective dependency span and the type of evidence (structural binding vs. factual retrieval vs. long-context cue integration), thereby providing complementary coverage for diagnosing attribution faithfulness in LLMs.

Case 0

Relation: P103 | ParaRel: 2796

Subject:
Danielle Darrieux

"The mother tongue of Danielle Darrieux is"

Answer:
French

Case 1

Relation: P140 | ParaRel: 19501

Subject:
Edwin of Northumbria

"The official religion of Edwin of Northumbria is"

Answer:
Christianity

Case 2

Relation: P1303 | ParaRel: 13704

Subject:
Toko Yasuda

"Toko Yasuda plays the instrument called"

Answer:
guitar

Figure 6: CounterFact examples (relation querying). The expected answer token is highlighted in brackets.

Case 0

Long-context Reasoning | Category: capital

Entity pair:
a = Hanoi, b = Vietnam

"When my flight landed in Vietnam, I converted my currency and slowly fell asleep. (I had a terrifying dream about my grandmother, but that's a story for another time). I was staying in the capital,"

Target token:
Hanoi

Case

Long-context Reasoning | Category: capital

Entity pair:
a = Kabul, b = Afghanistan

"When my flight landed in Afghanistan, I converted my currency and slowly fell asleep. (I had a terrifying dream about my grandmother, but that's a story for another time). I was staying in the capital,"

Target token:
Kabul

Case 2

Long-context Reasoning | Category: capital

Entity pair:
a = Bangkok, b = Thailand

"When my flight landed in Thailand, I converted my currency and slowly fell asleep. (I had a terrifying dream about my grandmother, but that's a story for another time). I was staying in the capital,"

Target token:
Bangkok

Figure 7: LongRA examples (long-context prompts). The target token to be predicted is highlighted in brackets.

F Visualization

F.1 Qualitative Analysis of 1-order Causal Structures

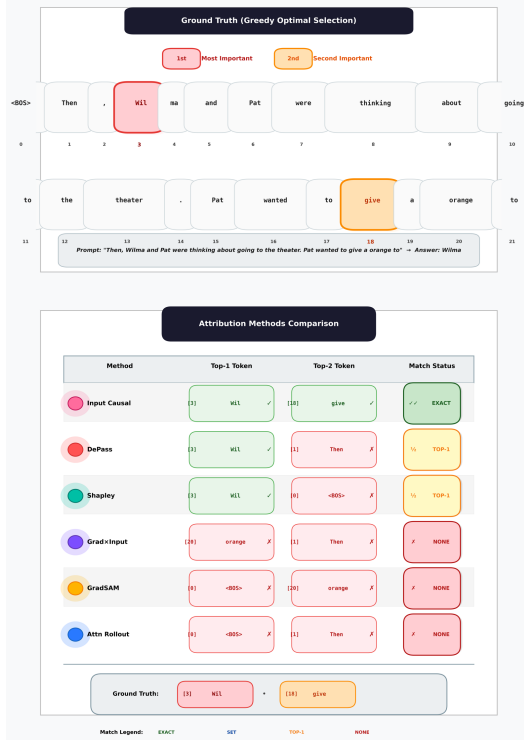


Figure 8: **Example of exact 1-order matching.** The greedy optimal path comprises a referential entity (Top-1) and a relational token (Top-2). This pair represents the minimal sufficient statistic for the prediction. Only the input-causal attribution recovers both positions correctly.

We present a detailed qualitative analysis to illustrate how different attribution methods behave under the strict requirement of *1-order matching*. This metric evaluates whether the top-2 tokens identified by a method align with the greedy optimal selection induced by the perturbation trajectory. By examining the specific semantic roles of the selected tokens, we uncover a fundamental disparity between causal necessity and heuristic salience.

The Composition of the Optimal Baseline Across the visualized examples in Figures 8 to 11, a consistent topological pattern emerges in the ground-truth greedy selection. The optimal subset is characterized by a precise coupling of a **Subject Entity** (typically the Top-1 token) and a **Relational Connector** (typically the Top-2 token, such as a preposition or verb). This structure reflects the intrinsic nature of autoregressive text attribution. To reconstruct a prediction faithfully, the model re-

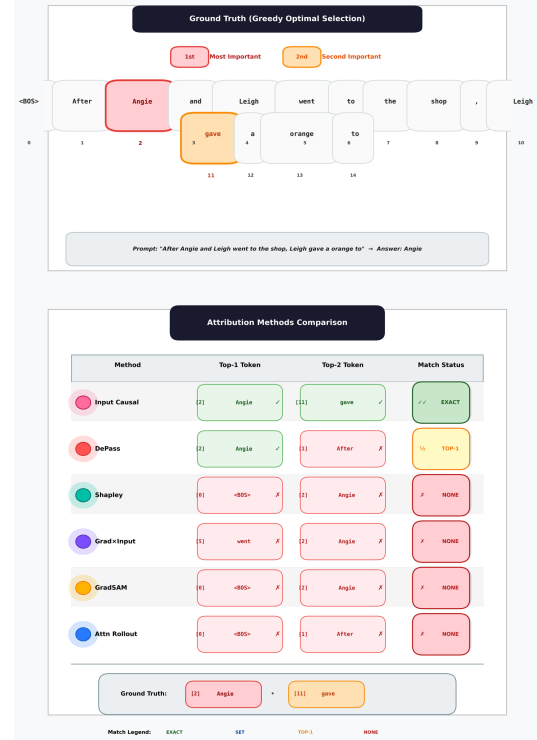


Figure 9: **Structural collapse in baselines.** While multiple methods achieve Top-1 correctness by finding the entity, they fail at the Top-2 position. The second-ranked token collapses to a structurally salient but causally irrelevant token, indicating a failure to model the necessary relational dependency.

quires two distinct types of information. First, it needs content information to anchor the semantic context, which is provided by the entity. Second, it needs structural information to direct the causal flow toward the specific target, which is provided by the connector. For instance, in the query asking for the location of a specific landmark, the entity provides the subject, while the preposition restricts the probability mass to a locative context. This observation suggests that faithful attribution must capture not just isolated salient features but the **causal structure** that binds them.

Systematic Failure of Saliency Heuristics The lower panels of the figures reveal that while our proposed input-causal attribution consistently recovers this (Entity, Relation) dyad, baseline methods exhibit systematic structural failures. Gradient-based and attention-based methods often succeed in identifying the Top-1 entity, as these tokens possess high distinctiveness and gradient norms. However, they struggle significantly with the Top-2 position. Instead of identifying the causal connector, these methods frequently assign high importance to struc-

turally prominent but causally irrelevant tokens, such as sentence-initial markers (<BOS>) or punctuation. This failure mode indicates that standard salience maps operate primarily as feature detectors rather than interaction modelers. They can locate "what" is important but fail to determine "how" these features interact to form a prediction. Consequently, they achieve high Top-1 accuracy but collapse when the evaluation demands the preservation of ordered causal interactions.

Visual Evidence The following figures provide concrete evidence of this phenomenon. Figure 8 demonstrates an exact match where the causal structure is perfectly preserved. Figures 9 through 11 showcase various lexical realizations where base-lines correctly spot the entity but miss the relational syntax, thereby failing to satisfy the 1-order matching criterion.

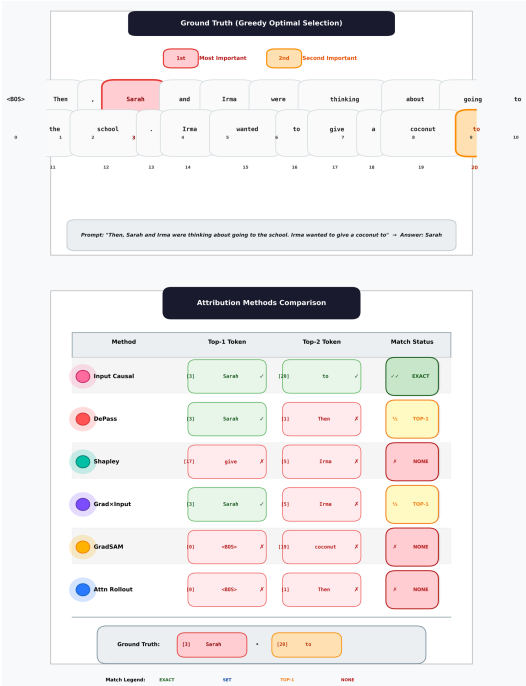


Figure 10: **Failure of gradient and attention Methods.** Despite correctly identifying the primary entity, standard methods do not preserve the interaction structure required for 1-order matching. This results in a ranking that is semantically plausible but causally insufficient.

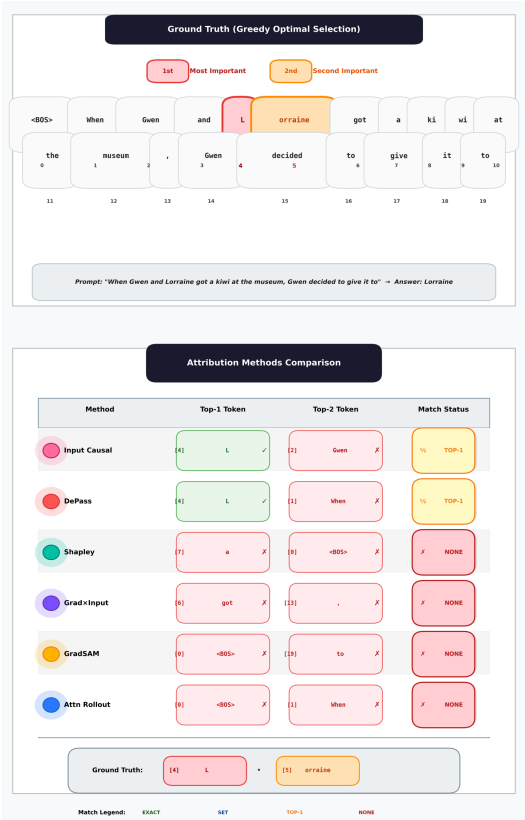


Figure 11: **Consistency across variations.** Another example showing the same qualitative pattern under different lexical conditions. The persistence of this error mode suggests that the observed failures are systematic limitations of the baselines rather than instance-specific anomalies.

F.2 Comparative Visualization of Attribution Landscapes

This section conducts a controlled visual comparison to isolate the impact of model representations on attribution patterns. Figures 12 and 13 visualize the attribution salience maps and corresponding AUC trajectories for *identical* inputs (an IOI-style query) and identical perturbation protocols, varying only the backbone model (**Qwen3-4B** vs. **LLaMA-3.2-3B**).

Sparsity vs. Diffusivity. The comparison reveals a striking divergence in how models represent causal signals:

- **Qwen3-4B (Fig. 12):** Exhibits high **signal localization**. Mid-layer attributions (ours) successfully concentrate probability mass on the sparse subset of causally decisive tokens (e.g., the specific entity). In contrast, gradient-based baselines suffer from noise, distributing importance broadly across the context.

- **LLaMA-3.2-3B (Fig. 13):** Exhibits high **representational diffusivity**. Saliency maps are noticeably more entropic across all methods, with a weaker separation between the causal signal and the background context.

Correlation with AUC Dynamics. These topological differences directly govern the metric performance. In Qwen3-4B, the sharp localization translates into superior efficiency: our method achieves rapid Recovery AUC (R-AUC) saturation with minimal tokens. Conversely, in LLaMA-3.2-3B, the diffuse representations lead to slower recovery rates and narrower gaps between methods in Perturbation AUC (P-AUC).

Implication. This analysis demonstrates that attribution quality is not merely a property of the explainer, but is intrinsically modulated by the **representational structure** of the target model. Faithful evaluation must therefore jointly inspect visual sparsity and continuous restoration dynamics (AUC) to distinguish between method failure and model uncertainty.

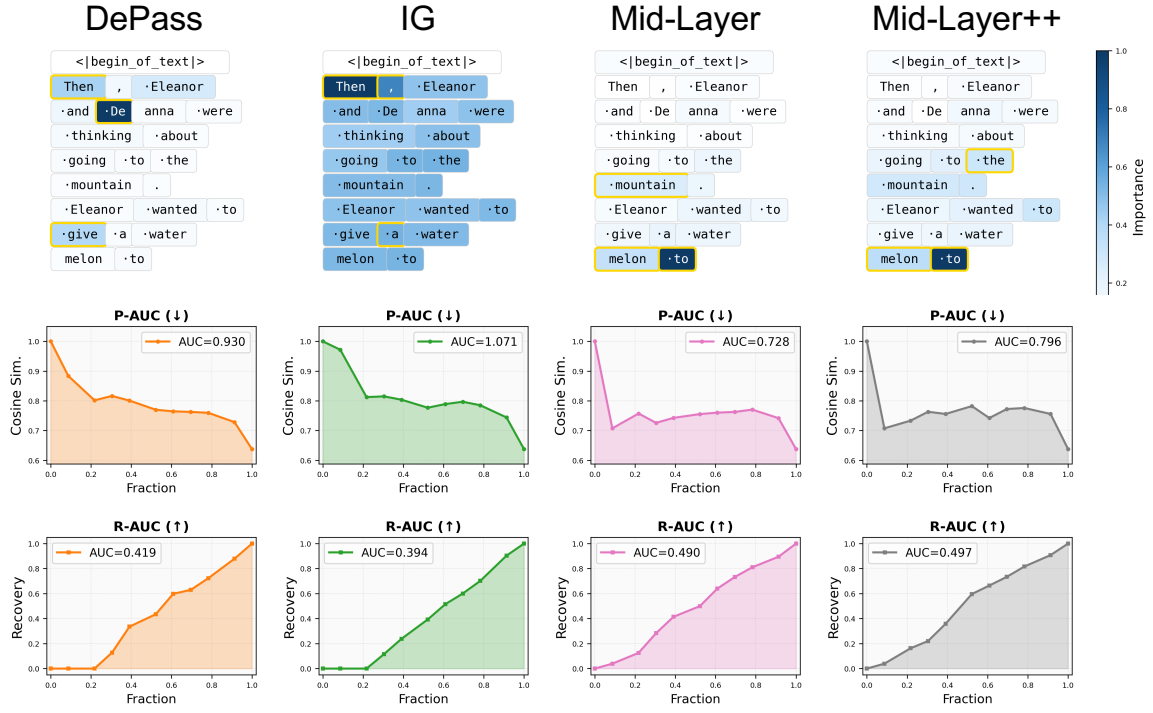


Figure 12: **High signal localization (Qwen3-4B)**. Visualization of attribution for an IOI sample. **Top**: Normalized salience maps. **Middle/Bottom**: P-AUC and R-AUC trajectories. Our mid-layer method (Ours) produces a highly **sparse** ranking that pinpoints causal tokens, leading to rapid representation recovery (high R-AUC) compared to noisy gradient baselines.



Figure 13: **High diffusivity (LLaMA-3.2-3B)**. Visualization for the *identical* sample on a different backbone. Salience distributions are markedly more **diffuse** (entropic) across all methods. This representational variance results in flatter AUC curves and reduced separability between methods, highlighting the model-dependence of attribution fidelity.

G Uncertainty, Robustness, and Component Controls

G.1 Confidence Intervals for the Main P-AUC Results

Table 7 reports two-sided 95% confidence intervals over per-example normalized P-AUC scores, computed as $\bar{x} \pm 1.96s/\sqrt{n}$, where the example is the sampling unit. The random-ranking reference is 1 and lower values are better. These intervals quantify uncertainty in the reported mean, whereas the standard deviations in Table 3 describe heterogeneity across examples.

G.2 Sensitivity to the Mid-Layer Choice

We evaluate $m \in \{6, 7, 8, 9, 10\}$ on 100 examples per dataset for LLaMA-3.2-3B and Qwen3-4B. Table 8 reports P-AUC with 95% confidence intervals, and Table 9 reports R-AUC point estimates. Performance does not collapse near the empirical saturation region, although the best exact layer varies across settings. The experiment therefore supports local robustness of the allocation–effect decomposition, not a claim that a single layer is universally optimal.

G.3 Full-Curve Versus Top-Ten Evaluation

Table 10 compares the main top-ten protocol with full-curve evaluation on 100 examples per dataset. Top-ten P-AUC is sometimes lower and sometimes higher than full-curve P-AUC, and the same is true for R-AUC. Thus, the top-ten protocol is a computationally focused view of the high-ranked, early trajectory rather than a range that systematically improves our scores. We do not claim that the two protocols always preserve method ordering.

G.4 Allocation and Effect Component Controls

The main comparison already contains component-aligned controls for the two terms in Mid-Layer Attribution. Attention Rollout propagates cumulative attention flow without a downstream causal intervention and therefore isolates the allocation-side signal. Leave-One-Out ranks tokens through direct input interventions without the mid-layer allocation map and therefore isolates an effect-dominant causal signal. Mid-Layer Attribution combines mid-layer allocation with downstream intervention effects, while MLA++ additionally groups redundant mid-layer representations. These baselines are functional component controls rather than separately tuned variants of our method.

Across Table 3, Attention Rollout often has weak P-AUC, whereas Leave-One-Out is competitive on necessity in several settings but has more limited recovery. The combined method is best or near-best in most settings, particularly for recovery. This comparison supports the role of combining allocation and effect without introducing an additional component ablation protocol.

Model	Method	IOI	CounterFact	LongRA
LLaMA-3.2-3B	Attention Rollout	1.1991 [1.1927, 1.2056]	1.1218 [1.1183, 1.1253]	1.0403 [1.0233, 1.0572]
	DePASS	1.1092 [1.1041, 1.1144]	0.9519 [0.9486, 0.9552]	0.8929 [0.8779, 0.9079]
	Integrated Gradients	1.0443 [1.0364, 1.0521]	0.9867 [0.9834, 0.9900]	0.8625 [0.8467, 0.8782]
	Grad $\times$ Input	1.0173 [1.0100, 1.0246]	0.9209 [0.9173, 0.9245]	0.8707 [0.8590, 0.8825]
	Grad SAM	0.8145 [0.8076, 0.8213]	0.8570 [0.8535, 0.8605]	0.7268 [0.7166, 0.7370]
	Leave-One-Out	0.8757 [0.8685, 0.8828]	0.8259 [0.8208, 0.8310]	0.7667 [0.7554, 0.7779]
	Mid-Layer Attribution	0.8024 [0.7960, 0.8087]	0.7474 [0.7439, 0.7508]	0.6890 [0.6775, 0.7005]
	Mid-Layer Attribution++	0.8366 [0.8300, 0.8431]	0.7621 [0.7587, 0.7654]	0.7436 [0.7289, 0.7583]
LLaMA-3.1-8B	Attention Rollout	1.3157 [1.3056, 1.3259]	1.2773 [1.2745, 1.2800]	1.1915 [1.1719, 1.2111]
	DePASS	1.0176 [1.0053, 1.0300]	0.9765 [0.9736, 0.9794]	0.8699 [0.8541, 0.8857]
	Integrated Gradients	0.9912 [0.9785, 1.0038]	0.9536 [0.9504, 0.9568]	0.9374 [0.9193, 0.9555]
	Grad $\times$ Input	0.9908 [0.9767, 1.0048]	0.9860 [0.9828, 0.9893]	0.8395 [0.8232, 0.8558]
	Grad SAM	0.7982 [0.7844, 0.8119]	0.8727 [0.8688, 0.8766]	0.6149 [0.5916, 0.6382]
	Leave-One-Out	0.7086 [0.6944, 0.7227]	0.8966 [0.8931, 0.9002]	0.7862 [0.7708, 0.8016]
	Mid-Layer Attribution	0.7002 [0.6869, 0.7135]	0.7487 [0.7451, 0.7522]	0.6621 [0.6356, 0.6885]
	Mid-Layer Attribution++	0.7140 [0.7017, 0.7263]	0.7535 [0.7499, 0.7570]	0.7491 [0.7282, 0.7699]
Qwen3-4B	Attention Rollout	1.3183 [1.3109, 1.3257]	1.1610 [1.1578, 1.1641]	1.2045 [1.1874, 1.2217]
	DePASS	0.9622 [0.9512, 0.9731]	1.0106 [1.0076, 1.0136]	0.9007 [0.8878, 0.9136]
	Integrated Gradients	0.9470 [0.9338, 0.9601]	0.9679 [0.9643, 0.9716]	0.9488 [0.9347, 0.9629]
	Grad $\times$ Input	0.8865 [0.8737, 0.8992]	0.9440 [0.9400, 0.9479]	0.8111 [0.7974, 0.8247]
	Grad SAM	0.8292 [0.8190, 0.8393]	0.8792 [0.8751, 0.8834]	0.6067 [0.5949, 0.6185]
	Leave-One-Out	0.5937 [0.5792, 0.6082]	0.8217 [0.8163, 0.8270]	0.6853 [0.6679, 0.7027]
	Mid-Layer Attribution	0.7718 [0.7608, 0.7829]	0.7543 [0.7504, 0.7581]	0.5545 [0.5364, 0.5726]
	Mid-Layer Attribution++	0.7620 [0.7507, 0.7733]	0.7536 [0.7496, 0.7576]	0.4097 [0.3940, 0.4255]

Table 7: Normalized P-AUC means with 95% confidence intervals over examples. The random-ranking reference is 1 and lower values are better.

Model	Dataset	Method	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
LLaMA-3.2-3B	IOI	MLA	0.7770 [0.7734,0.7806]	0.7861 [0.7791,0.7930]	0.8008 [0.7810,0.8206]	0.8082 [0.7942,0.8223]	0.8144 [0.8038,0.8249]
LLaMA-3.2-3B	IOI	MLA++	0.8375 [0.8167,0.8584]	0.8507 [0.8172,0.8841]	0.8521 [0.8420,0.8621]	0.8512 [0.8199,0.8825]	0.8603 [0.8479,0.8727]
LLaMA-3.2-3B	CounterFact	MLA	0.6885 [0.6346,0.7424]	0.7134 [0.6503,0.7765]	0.7082 [0.6748,0.7415]	0.6992 [0.6713,0.7272]	0.7041 [0.6488,0.7594]
LLaMA-3.2-3B	CounterFact	MLA++	0.7141 [0.6433,0.7849]	0.6969 [0.6457,0.7482]	0.7146 [0.6848,0.7444]	0.7026 [0.6098,0.7953]	0.7035 [0.6684,0.7386]
LLaMA-3.2-3B	LongRA	MLA	0.6527 [0.6430,0.6623]	0.6215 [0.5559,0.6871]	0.6132 [0.5787,0.6477]	0.6220 [0.5869,0.6571]	0.6075 [0.5370,0.6781]
LLaMA-3.2-3B	LongRA	MLA++	0.6684 [0.6260,0.7108]	0.7257 [0.6840,0.7674]	0.6544 [0.5843,0.7246]	0.6312 [0.5527,0.7098]	0.6191 [0.5495,0.6886]
Qwen3-4B	IOI	MLA	0.7006 [0.6866,0.7145]	0.7067 [0.6437,0.7697]	0.7064 [0.6734,0.7393]	0.7253 [0.6864,0.7641]	0.7149 [0.6743,0.7556]
Qwen3-4B	IOI	MLA++	0.6832 [0.6628,0.7036]	0.7833 [0.7494,0.8172]	0.7061 [0.6802,0.7320]	0.7175 [0.6925,0.7425]	0.7151 [0.6998,0.7304]
Qwen3-4B	CounterFact	MLA	0.7465 [0.6319,0.8611]	0.7620 [0.6674,0.8566]	0.7481 [0.6607,0.8355]	0.7397 [0.6295,0.8498]	0.7440 [0.6859,0.8022]
Qwen3-4B	CounterFact	MLA++	0.7519 [0.6514,0.8524]	0.7730 [0.7041,0.8418]	0.7330 [0.6044,0.8616]	0.7279 [0.6144,0.8414]	0.7352 [0.6655,0.8049]
Qwen3-4B	LongRA	MLA	0.4525 [0.4199,0.4851]	0.4484 [0.4002,0.4967]	0.4358 [0.3650,0.5067]	0.4745 [0.4309,0.5181]	0.4590 [0.3958,0.5222]
Qwen3-4B	LongRA	MLA++	0.4890 [0.4279,0.5501]	0.4732 [0.3556,0.5907]	0.4119 [0.4024,0.4213]	0.4189 [0.3983,0.4395]	0.4095 [0.3712,0.4478]

Table 8: P-AUC sensitivity to m (lower is better; 95% confidence intervals).

Model	Dataset	Method	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
LLaMA-3.2-3B	IOI	MLA	0.4488	0.4494	0.4523	0.4499	0.4518
LLaMA-3.2-3B	IOI	MLA++	0.4384	0.4379	0.4380	0.4441	0.4476
LLaMA-3.2-3B	CounterFact	MLA	0.3833	0.3860	0.3885	0.3920	0.3955
LLaMA-3.2-3B	CounterFact	MLA++	0.3925	0.3923	0.3940	0.3938	0.3980
LLaMA-3.2-3B	LongRA	MLA	0.4691	0.4740	0.4755	0.4803	0.4774
LLaMA-3.2-3B	LongRA	MLA++	0.4383	0.4385	0.4495	0.4541	0.4518
Qwen3-4B	IOI	MLA	0.3949	0.4322	0.4625	0.4585	0.4834
Qwen3-4B	IOI	MLA++	0.4000	0.4237	0.4737	0.4776	0.4938
Qwen3-4B	CounterFact	MLA	0.4290	0.4258	0.4496	0.4417	0.4560
Qwen3-4B	CounterFact	MLA++	0.4455	0.4395	0.4425	0.4395	0.4530
Qwen3-4B	LongRA	MLA	0.4332	0.4407	0.4617	0.4649	0.4883
Qwen3-4B	LongRA	MLA++	0.4802	0.5048	0.5160	0.5241	0.5352

Table 9: R-AUC sensitivity to m (higher is better; point estimates).

Model	Dataset	Method	P-AUC full [95% CI]	P-AUC top-10 [95% CI]	R-AUC full	R-AUC top-10
LLaMA-3.2-3B	IOI	MLA	0.7825 [0.6946,0.8704]	0.7376 [0.6477,0.8275]	0.4454	0.4583
LLaMA-3.2-3B	IOI	MLA++	0.8448 [0.7736,0.9160]	0.7831 [0.6587,0.9075]	0.4345	0.4188
LLaMA-3.2-3B	CounterFact	MLA	0.7153 [0.4878,0.9428]	0.6616 [0.4262,0.8970]	0.3898	0.4003
LLaMA-3.2-3B	CounterFact	MLA++	0.7004 [0.4685,0.9323]	0.6690 [0.4214,0.9166]	0.3935	0.4005
LLaMA-3.2-3B	LongRA	MLA	0.5794 [0.4123,0.7465]	0.5589 [0.2491,0.8687]	0.4726	0.4245
LLaMA-3.2-3B	LongRA	MLA++	0.6110 [0.3903,0.8317]	0.5198 [0.3234,0.7162]	0.4458	0.4212
Qwen3-4B	IOI	MLA	0.6985 [0.5054,0.8916]	0.6595 [0.4280,0.8910]	0.4768	0.5379
Qwen3-4B	IOI	MLA++	0.6924 [0.5077,0.8771]	0.6566 [0.4265,0.8867]	0.4886	0.5424
Qwen3-4B	CounterFact	MLA	0.7372 [0.5225,0.9519]	0.7508 [0.5604,0.9412]	0.4557	0.4979
Qwen3-4B	CounterFact	MLA++	0.7487 [0.5399,0.9575]	0.7425 [0.5433,0.9417]	0.4550	0.4972
Qwen3-4B	LongRA	MLA	0.4465 [0.2394,0.6536]	0.5081 [0.2944,0.7218]	0.4866	0.5080
Qwen3-4B	LongRA	MLA++	0.4237 [0.2365,0.6109]	0.5505 [0.3310,0.7700]	0.5311	0.5045

Table 10: Full-curve and top-ten protocol comparison on 100 examples per dataset.