Focus-LIME: Surgical Interpretation of Long-Context Large Language Models via Proxy-Based Neighborhood Selection

Junhao Liu , **Haonan Yu** , **Zhenyu Yan** and **Xin Zhang***

Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education
School of Computer Science, Peking University, Beijing 100871, China
{liujunhao, xin}@pku.edu.cn, {haonanyu, zhenyuyan}@stu.pku.edu.cn

Abstract

As Large Language Models (LLMs) scale to handle massive context windows, achieving *surgical* feature-level interpretation is essential for high-stakes tasks like legal auditing and code debugging. However, existing local model-agnostic explanation methods face a critical dilemma in these scenarios: feature-based methods suffer from attribution dilution due to high feature dimensionality, which prevents them from providing faithful explanations. In this paper, we propose **Focus-LIME**, a coarse-to-fine framework designed to restore the tractability of surgical interpretation. Focus-LIME utilizes a proxy model to curate the perturbation neighborhood, allowing the target model to perform fine-grained attribution exclusively within the optimized context. Empirical evaluations on long-context benchmarks demonstrate that our method makes surgical explanations practical and provides faithful explanations to users.

1 Introduction

As Large Language Models (LLMs) scale to handle massive context windows [Ding *et al.*, 2024], they are entrusted with increasingly complex tasks, from auditing legal contracts to debugging extensive codebases. Given the opaque nature of these models, explaining their behavior has emerged as a critical area of research [Valvoda and Cotterell, 2024; Guo *et al.*, 2024]. In particular, practitioners require explainability at varying levels of granularity [Elisha *et al.*, 2025]. Consequently, there is a persistent demand for faithful, fine-grained explanations that can pinpoint the exact locus of model decisions, such as locating a specific important clause or word within a 50-page document. In this paper, we propose to generate surgical local explanations for LLMs by curating the neighborhood of explanations with proxy models.

To illustrate the dilemma of explaining long-context LLMs, consider the motivating example from the Contract Understanding Atticus Dataset (CUAD) [Hendrycks *et al.*, 2021] shown in Figure 1. Here, a user queries a 6,000-word contract to identify clauses designating the specific state or national laws applicable to the agreement. On one hand, **concept-based explanation methods** [Ghorbani *et al.*, 2019; Liu *et al.*, 2024b; Ludan *et al.*, 2024] scale effectively to long contexts by operating at a high level of abstraction. In this scenario, while they might correctly identify the model's reliance on "Jurisdictional Clause" concepts, they inherently lack the granularity to localize specific tokens, leaving the user unable to verify the exact *locus* of the decision. **Conversely, feature-based methods** like LIME [Ribeiro *et al.*, 2016] theoretically offer this surgical precision but collapse under the *curse of dimensionality*. When applied to this extensive context, standard LIME suffers from severe *attribution dilution*: it dissipates the sampling budget across thousands of irrelevant tokens, drowning the critical signal in background noise. To generate faithful explanations, LIME would require an impractically large number of model queries, also resulting in prohibitive computational costs. Our goal is to bridge this gap by strictly curating the neighborhood, enabling the explainer to filter out semantic noise and surgically focus on the decisive evidence. Consequently, our method can precisely identify that specific words like "Governing", "Law", and "Illinois" are the key drivers of the model's decision.

To address the intractability of fine-grained attribution in long-context scenarios, we propose **Focus-LIME**, a framework designed to restore the surgical feature attribution capabilities of LIME for LLMs. Our approach fundamentally reimagines the role of efficiency-oriented techniques. While recent research [Zhao *et al.*, 2025; Jiang *et al.*, 2023; Liu *et al.*, 2026] has explored using proxy models to generate explanations entirely to reduce inference costs, this strategy inherently risks producing unfaithful explanations. Acknowledging this limitation, we draw inspiration from proxy-based efficiency but shift the paradigm from "replacement" to "guidance." We employ the proxy model strictly as a "preliminary scout" for neighborhood selection, tasked with filtering out irrelevant background features. Crucially, we then perform the final perturbation analysis using the target model, but exclusively within this optimized, low-dimensional neighborhood. This "Scout then Scrutinize" strategy aims to preserve fidelity to the target model's decision boundary, while obtaining computational scalability from proxy-based methods.

We extensively evaluate Focus-LIME on long-context benchmarks, including the CUAD and Qasper [Dasigi *et*

*Corresponding Author

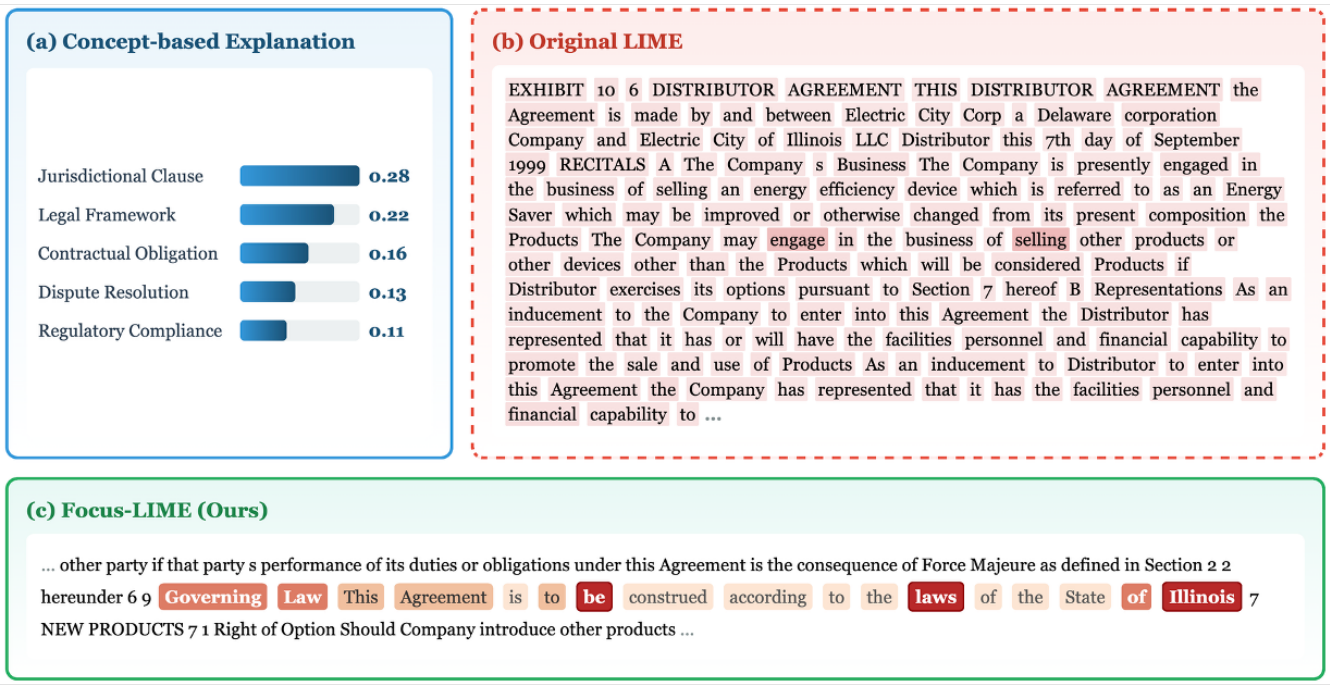

Figure 1: Comparison of explanations generated by different methods when querying LLMs for a "Governing Law" clause in a lengthy contract. **(A) Concept-based Explanation:** Attributes importance to high-level concepts. **(B) Original LIME:** Suffers from *attribution dilution*, where significance scores are distributed noisily across all tokens. **(C) Ours:** Generates a sparse and precise explanation.

al., 2021] datasets. Empirical results demonstrate that our method significantly outperforms both original LIME and pure-proxy baselines in terms of explanation fidelity. Compared with curating neighborhoods using target models, Focus-LIME achieves comparable fidelity while requiring notably fewer model queries, showcasing its efficiency.

In summary, our contributions are as follows:

- We propose Focus-LIME, a model-agnostic framework that enables *surgical* feature-attribution interpretation for long-context LLMs by decoupling neighborhood selection (via proxy) from attribution (via target).
- We instantiate this framework by designing a coarse-to-fine neighborhood contraction algorithm that effectively leverages proxy models to filter semantic noise without sacrificing target fidelity.
- We evaluate the effectiveness of our approach on long-context benchmarks, showing that it makes surgical explanations practical and provides faithful explanations to users.

2 Background and Related Work

In this section, we introduce the necessary background and survey prior studies relevant to our work.

2.1 Large Language Models

A large language model can be regarded as a probabilistic function f that, given a sequence of tokens $\boldsymbol{x} = [x_1, x_2, \ldots, x_t]$, produces a probability distribution over the set of possible next tokens, written as $f(\boldsymbol{x})$. Formally, the mapping is $f : \mathcal{V}^* \to \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ denotes the vocabulary, and the output corresponds to a distribution over all vocabulary elements. Since most LLMs operate with an upper bound on input length n, we typically assume the sequence $\boldsymbol{x} \in \mathbb{X}^n$. Today, LLMs are used in a wide range of applications, including medical [Guo *et al.*, 2024] and legal [Valvoda and Cotterell, 2024], which often involve processing long document inputs and require reliable explanations to ensure their trustworthiness.

2.2 Local Model-Agnostic Explanation Techniques

A model-agnostic local explanation method t takes as input a predictive model f together with a specific instance $\boldsymbol{x}$, and produces an explanation $g_{f,\boldsymbol{x}}$ that characterizes how the model behaves around $\boldsymbol{x}$.

In this work, our attention is mainly on attribution-style approaches, which are the most widely used explanation form. These methods assign numerical relevance scores to the input features of $\boldsymbol{x} \in \mathbb{X}$, aiming to capture how much each feature contributes to the output $f(\boldsymbol{x})$. The explanation $g_{f,\boldsymbol{x}}$ can thus be represented as a vector of attributions $\boldsymbol{a} = [a_1, a_2, \ldots, a_n]$, where each coefficient a_i quantifies the effect of feature x_i on the model's decision. Formally, we can describe an attribution-based explanation method t as a function

$$t : \mathbb{F} \times \mathbb{X}^n \to \mathbb{R}^n,$$

where $\mathbb{F}$ denotes the set of predictive models, $\mathbb{X}^n$ is the input domain, and n is the dimensionality of the instance $\boldsymbol{x}$.

Local explanations describe the target model's behavior in a local neighborhood around a specific input [Dwivedi *et al.*,

2023]. The neighborhood is usually defined as a set of perturbed samples generated by perturbing the input [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Liu and Zhang, 2025; Dwivedi *et al.*, 2023]. Perturbations are typically represented in a binary masking space, where each perturbed instance is encoded by a vector $\boldsymbol{z} \in \{0,1\}^n$. Each entry z_i indicates whether the corresponding feature x_i of the original input $\boldsymbol{x}$ is preserved ($z_i = 1$) or removed/replaced ($z_i = 0$). Formally, there exists a mapping function

$$\phi : \{0,1\}^n \times \mathbb{X}^n \to \mathbb{X}^n,$$

that reconstructs a perturbed sample in the original input space. The neighborhood $\mathbb{N}$ of the input $\boldsymbol{x}$ is then defined as

$$\mathbb{N} = \{\phi(\boldsymbol{z}, \boldsymbol{x}) | \boldsymbol{z} \in \{0,1\}^n\}.$$

When processing long sentence inputs, the number of features n can be very large, leading to a large and high-dimensional neighborhood. Many methods have been proposed to handle this challenge. Some methods [Santhiappan *et al.*, 2024; Achtibat *et al.*, 2024] use attention to narrow the neighborhood, but they require access to the internal model structure, which is not applicable to closed-source LLMs. Some other methods [Shankaranarayana and Runje, 2019; Tan *et al.*, 2023; Yu *et al.*, 2026] try to improve the sampling strategy, but they still provide explanations in a large neighborhood, which can lead to unfaithful explanations.

2.3 Desiderata of Explanations

In this paper, we focus on the fidelity and cost of explanations [Yeh *et al.*, 2019; Burger *et al.*, 2023].

On one hand, explanations should faithfully reflect the model's decision-making process. High fidelity indicates that the explanation accurately captures how the model arrives at its predictions. This ensures that users can trust the explanation and apply it to various practical scenarios, such as model debugging and bias detection.

On the other hand, when explaining large language models, the cost of generating explanations becomes a critical consideration, as querying these models can be computationally expensive. Therefore, it is essential to develop explanation methods that can deliver high-fidelity explanations while adhering to practical computational budgets. In this work, we aim to improve the fidelity of explanations without increasing the overall computational cost.

3 Methodology

In this section, we introduce the Focus-LIME framework for generating surgical explanations for long-context LLMs via proxy-based neighborhood selection.

3.1 Problem Formulation

Following the standard setting of local model-agnostic attribution-based explanation methods [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017], let f_T denote the target model and $g_{\boldsymbol{a}}$ denote a local surrogate parameterized by the attribution vector $\boldsymbol{a}$. The goal is to estimate the optimal attribution vector $\boldsymbol{a}^*$ that minimizes the local approximation error L over the neighborhood:

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}} \mathbb{E}_{\boldsymbol{z} \sim \pi_{\boldsymbol{x}}} \left[L(f_T(\phi(\boldsymbol{z}, \boldsymbol{x})), g_{\boldsymbol{a}}(\boldsymbol{z})) \right] + \Omega(\boldsymbol{a}), \quad (1)$$

where ϕ maps a binary perturbation mask back to the original input space, $\pi_{\boldsymbol{x}}$ is the sampling distribution around $\boldsymbol{x}$, and $\Omega(\boldsymbol{a})$ is a complexity regularizer.

The Cost-Fidelity Dilemma. The fidelity of the local explanation relies heavily on the *sample density* within the neighborhood. To learn a faithful linear surrogate $g_{\boldsymbol{a}}$, the number of perturbed samples K must scale with the feature dimension n to ensure the regression stabilizes.

However, because querying large language models incurs significant computational costs, end users often face a budget B that limits the total number of tokens that can be processed. Under a computational budget B, the maximum allowable samples $K_{max} = \lfloor B/C(f_T) \rfloor$ is fixed, where $C(f_T)$ denotes the average cost of evaluating one perturbed sample with the target model, measured in processed tokens. In long-context scenarios, as n grows, the sample density $\rho = K_{max}/n$ becomes vanishingly small, making the estimation of $\boldsymbol{a}^*$ increasingly unreliable. The limited samples fail to cover the high-dimensional decision boundary, causing the variance of the estimator $\boldsymbol{a}^*$ to explode. We term this phenomenon *attribution dilution*, where the resulting explanation becomes indistinguishable from noise.

Proposed Formulation. To resolve this, we introduce a computationally efficient proxy model f_P for the target model f_T (where $C(f_P)$ is defined analogously and $C(f_P) \ll C(f_T)$) and decompose the problem into two stages. Instead of optimizing over the entire neighborhood $\mathbb{N}$, we seek to identify an *active neighborhood subspace* $\mathbb{S} \subseteq \mathbb{N}$ that contains the most information. Our objective is reformulated as:

$$\boldsymbol{a}^*_{focused} = \arg\min_{\boldsymbol{a}} \mathbb{E}_{\boldsymbol{z} \sim \pi'_{\boldsymbol{x}}} \left[L(f_T(\phi(\boldsymbol{z}, \boldsymbol{x})), g_{\boldsymbol{a}}(\boldsymbol{z})) \right] + \Omega(\boldsymbol{a}), \quad (2)$$

where $\pi'_{\boldsymbol{x}}$ is the sampling distribution restricted to $\mathbb{S}$. The subspace $\mathbb{S}$ is determined by the proxy f_P to filter out irrelevant dimensions (semantic noise), allowing the target model f_T to focus its limited sampling budget on the critical decision boundary. In this case, the sample density $\rho' = K_{max}/|\mathbb{S}|$ can be maintained at a reasonable level, enabling faithful estimation of $\boldsymbol{a}^*_{focused}$.

3.2 The Focus-LIME Framework

To address the dilemma of vanishing sample density, we propose **Focus-LIME**, a framework that implements a coarse-to-fine neighborhood curation strategy. The overall pipeline is illustrated in Figure 2. Our core insight is to **decompose** the explanation process into two distinct phases: first using a proxy model to *scout* the relevant neighborhood without additional target model queries, and then using the target model to *scrutinize* the fine-grained attributions within this focused region.

Phase I: Proxy-Guided Scouting. This phase aims to identify an *Active Neighborhood Subspace* $\mathbb{S} \subset \mathbb{N}$ that contains the most relevant features, thereby filtering out background

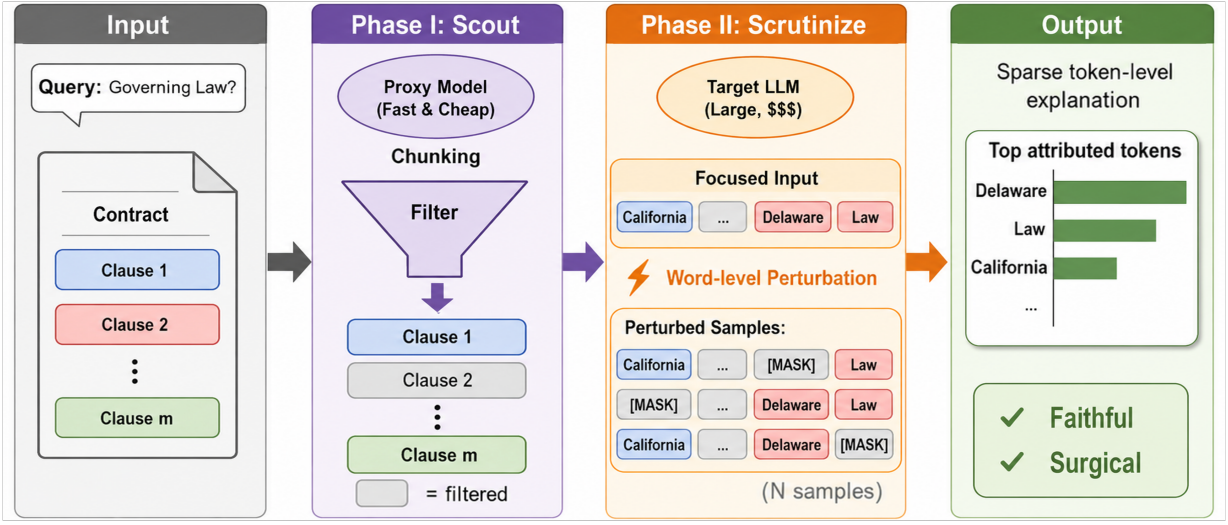

Figure 2: The pipeline of Focus-LIME, which consists of two main steps: 1) neighborhood curation with proxy models, where we first use a smaller proxy model to iteratively narrow the neighborhood by fixing unimportant features and identifying the most faithful neighborhood; and 2) fine-grained attribution with the target model, where we then generate the final explanations using the original LLM, but only in the curated neighborhood.

noise. We formulate Focus-LIME as a flexible framework where the *scouting policy* is a modular component. Depending on the nature of the proxy model and the task, various selection strategies can be employed (e.g., gradient-based saliency [Sundararajan *et al.*, 2017], attention [Santhiappan *et al.*, 2024; Achtibat *et al.*, 2024], and so on). In our experiments, we instantiate this as an **iterative refinement process** in a fully model-agnostic manner:

1. **Initialization:** We initialize the candidate set $\mathbb{C}^{(0)}$ as the entire input document $\boldsymbol{x}$. Let $t = 1$ be the current iteration step.
2. **Hierarchical Decomposition:** At step t, we decompose the surviving candidates from the previous step $\mathbb{C}^{(t-1)}$ into finer-grained units $\mathbb{U}^{(t)} = \{u_1, u_2, \ldots, u_m\}$. For example, in the first iteration, we decompose the document into paragraphs; in the second, we decompose the selected paragraphs into sentences.
3. **Proxy Valuation:** We employ LIME to compute an attribution score for each unit in $\mathbb{U}^{(t)}$ using the proxy model f_P. We follow [Liu *et al.*, 2026] to choose the proxy model.
4. **Top-k Filtering:** We select the top-k_t units with the highest relevance scores to form the refined set $\mathbb{C}^{(t)}$, where k_t is the retention budget at iteration t.
5. **Termination Check:** We check if the total token length of $\mathbb{C}^{(t)}$ satisfies the density constraint or a user-defined stop condition (e.g., maximum number of iterations). If the constraint is met, the iteration terminates, and $\mathbb{C}^{(t)}$ is converted into the final binary *Focus Mask* $\mathbf{m}_{focus}$. Otherwise, we increment t and repeat from Step 2.

Through this iterative pruning, Focus-LIME effectively "zooms in" on the decision boundary, discarding massive amounts of irrelevant context.

Phase II: Target-Based Scrutinizing. Once the iterative scouting terminates, we obtain a converged *Focus Mask* $\mathbf{m}_{focus} \in \{0,1\}^n$, which delineates the important region from the irrelevant context. In this phase, we engage the target model f_T to perform fine-grained attribution. Unlike standard LIME, which suffers from dilution by sampling uniformly across the entire high-dimensional space, Focus-LIME restricts the perturbation logic as follows:

1. **Constrained Perturbation:** We generate a set of K perturbed samples $\mathcal{Z}' = \{\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_K\}$ using a *Mask-Conditioned Generator*. For each sample $\boldsymbol{z}' \in \{0,1\}^n$, the value of the i-th token z'_i is determined by:

$$z'_i = \begin{cases} 1(Static), & \text{if } \mathbf{m}_{focus}[i] = 0 \\ \sim \text{Bernoulli}(0.5), & \text{if } \mathbf{m}_{focus}[i] = 1 \end{cases} \tag{3}$$

   This strategy ensures that the irrelevant context remains static, while the perturbation is concentrated exclusively within the active region identified using the proxy model.
2. **Target Valuation:** We query the expensive target model f_T with these constrained samples to obtain the prediction vector $\mathbf{y}'$. Crucially, because the dimensionality of the perturbation space has been reduced from n (total tokens) to n_{active}, where $n_{active} := \sum_{i=1}^{n} \mathbf{m}_{focus}[i]$, the sample density $\rho = K/n_{active}$ is restored to a high level, ensuring the stability of the subsequent regression.
3. **Explanation Generation:** Finally, following the standard LIME procedure, we fit a weighted linear regression model to approximate the target model's behavior within the focused neighborhood.

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}} \sum_{\boldsymbol{z}' \in \mathcal{Z}'} \pi_{\mathbb{S}}(\boldsymbol{z}') \cdot L\big(f_T(\phi(\boldsymbol{z}', \boldsymbol{x})), \mathbf{w}^\top \boldsymbol{z}'\big) + \Omega(\mathbf{w}) \tag{4}$$

where $\pi_{\mathbb{S}}(\boldsymbol{z}')$ is the kernel weight within the focused neighborhood, L is the same local fitting loss used in Eq. 1, and only the weights corresponding to the active region ($\mathbf{m}_{focus}[i] = 1$) are optimized. The resulting attribution vector $\mathbf{w}^*$ provides a *surgical explanation*: it identifies the precise tokens that most significantly influenced the target model's prediction within the long-context input.

3.3 Summary

In summary, we presented **Focus-LIME**, a framework designed to overcome the intractability of explaining long-context LLMs. By recognizing that not all tokens in a massive document require equal scrutiny, we reformulated the local explanation problem as a two-stage process:

- **Scout (Phase I):** Utilizing a cost-efficient proxy model to iteratively filter out semantic noise and identify a low-dimensional *Active Neighborhood*.
- **Scrutinize (Phase II):** Leveraging the target model to perform surgical attribution exclusively within this optimized subspace.

This approach resolves the fundamental conflict between *attribution dilution* and *computational budget* in existing model-agnostic feature-attribution methods. This ensures that the final explanation achieves high fidelity to the target model's behavior while keeping the inference cost manageable.

4 Empirical Study

In this chapter, we present a comprehensive empirical evaluation of **Focus-LIME** on challenging long-context benchmarks. Specifically, we aim to answer the following research questions:

- **RQ1 (Faithfulness):** Does Focus-LIME generate more faithful explanations compared to original methods (e.g., LIME) and pure-proxy baselines in long-context scenarios?
- **RQ2 (Mechanism Verification):** Does the neighborhood curation process mitigate attribution dilution, thereby improving explanation fidelity?
- **RQ3 (Human Alignment):** Does Focus-LIME effectively align with human reasoning by accurately localizing the ground-truth evidence spans annotated by experts?

4.1 RQ1: Faithfulness Evaluation

Experimental Setup

Datasets. We use two long-context document QA datasets in our experiments:

- **CUAD** [Hendrycks *et al.*, 2021] is a legal document QA dataset that requires models to answer questions based on long legal contracts.
- **Qasper** [Dasigi *et al.*, 2021] is a long-context QA dataset that contains scientific papers.

We use the yes/no question answering tasks in both datasets for evaluation, retaining examples whose input context exceeds 5,000 tokens. Considering the cost of target-model queries, we evaluate on a subset of 500 examples from each dataset.

Models. We evaluate our method on different combinations of target and proxy models:

- **Target Models (f_T):** We use GPT-4o [Achiam *et al.*, 2023] and DeepSeek V3 [Liu *et al.*, 2024a] as target models.
- **Proxy Models (f_P):** After the screening process, we use Qwen3-30B-A3B [Yang *et al.*, 2025a] and Qwen2.5-32B [Yang *et al.*, 2025b] as proxy models for all target LLMs; these models can run on consumer-grade GPUs.

Baselines. We compare Focus-LIME against the following baselines:

- **Original LIME:** Applies LIME directly to the full document. For each test, we sample 1000 perturbed samples.
- **Proxy Explanation:** Following [Liu *et al.*, 2026], this baseline uses the proxy model to generate explanations. Because the proxy model is much cheaper than the target model, we sample 10,000 perturbed samples for each test.
- **Focus-LIME (w/o Proxy Model):** An ablation of Focus-LIME that skips the proxy-based neighborhood curation step and directly applies constrained perturbation to the full document.

Metrics. To evaluate the faithfulness of explanations, we conduct **deletion experiments** [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Sun *et al.*, 2023], which measure the change in model output when the most important features (as identified by the explanation) are removed. We use *Area Over most relevant first perturbation curve* (AOPC) as the metric. Specifically, AOPC is defined as:

$$\mathrm{AOPC}_k = \frac{1}{|\mathbb{T}|} \sum_{\boldsymbol{x} \in \mathbb{T}} (p_f(y|\boldsymbol{x}) - p_f(y|\boldsymbol{x}^{(k)})),$$

where $p_f(y|\boldsymbol{x})$ is the probability that f outputs y given the input $\boldsymbol{x}$, $\mathbb{T}$ is the set of all test inputs, and $\boldsymbol{x}^{(k)}$ is the input $\boldsymbol{x}$ with the top k most important features removed. A higher AOPC_k indicates a better explanation. We report $AOPC_{10}$, $AOPC_{50}$, and $AOPC_{100}$ and $AOPC := \frac{\sum_{k=1}^{100} AOPC_k}{100}$ in our experiments.

Results

Tables 1 and 2 present the faithfulness evaluation results on the CUAD and Qasper datasets, respectively. We observe that in the **absence of neighborhood curation**, both Standard LIME and Pure Proxy baselines struggle significantly in long-context scenarios. Specifically, they yield negligible AOPC scores, failing to identify truly decisive features due to severe *attribution dilution*. In contrast, by incorporating the neighborhood curation process, Focus-LIME restores the tractability of the explanation task, making it practical and useful by substantially enhancing fidelity.

Target Model	Method	$AOPC_{10}$	$AOPC_{50}$	$AOPC_{100}$	$AOPC$
GPT-4o	Original LIME	0.006	0.007	0.008	0.008
	Proxy Explanation (Qwen3-30B-A3B)	0.006	0.010	0.022	0.014
	Focus-LIME (w/o Proxy)	0.214	0.433	0.761	0.489
	Focus-LIME (Qwen3-30B-A3B)	0.221	0.452	0.753	0.481
	Focus-LIME (Qwen2.5-32B)	0.207	0.431	0.757	0.473
DeepSeek V3	Original LIME	0.004	0.004	0.008	0.005
	Proxy Explanation (Qwen3-30B-A3B)	0.002	0.006	0.007	0.005
	Focus-LIME (w/o Proxy)	0.143	0.561	0.786	0.535
	Focus-LIME (Qwen3-30B-A3B)	0.164	0.546	0.775	0.524
	Focus-LIME (Qwen2.5-32B)	0.157	0.537	0.773	0.531

Table 1: Faithfulness evaluation on **CUAD** measured by AOPC. Higher is better.

Target Model	Method	$AOPC_{10}$	$AOPC_{50}$	$AOPC_{100}$	$AOPC$
GPT-4o	Original LIME	0.003	0.003	0.004	0.003
	Proxy Explanation (Qwen3-30B-A3B)	0.005	0.010	0.013	0.011
	Focus-LIME (w/o Proxy)	0.083	0.357	0.413	0.323
	Focus-LIME (Qwen3-30B-A3B)	0.102	0.356	0.408	0.318
	Focus-LIME (Qwen2.5-32B)	0.081	0.341	0.391	0.303
DeepSeek V3	Original LIME	0.004	0.006	0.008	0.006
	Proxy Explanation (Qwen3-30B-A3B)	0.002	0.004	0.007	0.004
	Focus-LIME (w/o Proxy)	0.064	0.276	0.384	0.289
	Focus-LIME (Qwen3-30B-A3B)	0.074	0.265	0.375	0.296
	Focus-LIME (Qwen2.5-32B)	0.068	0.258	0.369	0.280

Table 2: Faithfulness evaluation on **Qasper** measured by AOPC. Higher is better.

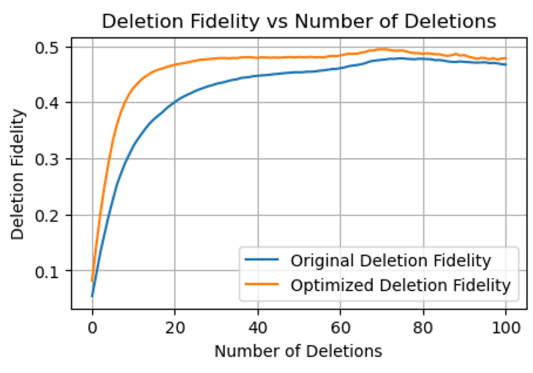

Figure 3: The deletion fidelity (AOPC↑) of explanations on the original and optimal neighborhoods.

Furthermore, Focus-LIME has performance comparable to the ablation without the proxy model, which indicates that using proxy models for neighborhood curation does not sacrifice explanation fidelity. Additionally, Focus-LIME achieves similar performance with different proxy models (Qwen3-30B-A3B and Qwen2.5-32B), indicating that our method is robust to the choice of proxy model.

4.2 RQ2: Mechanism Verification

While RQ1 has demonstrated the superior faithfulness of Focus-LIME on long-context benchmarks, we further investigate the underlying mechanism by verifying our hypothesis that **narrowing the neighborhood can improve the fidelity of local** explanations in general scenarios. This analysis examines whether users can also use this strategy to improve explanation fidelity beyond long-context scenarios.

Experimental Setup. We conducted an experiment on a text classification dataset: the Large Movie Review Dataset (IMDb) [Maas *et al.*, 2011], which contains movie reviews with an average length of 228 words. We use Qwen3-235B-A22B as the target model, and use Qwen3-30B-A3B as the proxy model to optimize the neighborhood. We optimize the neighborhood by greedily fixing the least important feature according to the explanation generated by LIME in each iteration, and evaluate the fidelity of explanations along the path of narrowing the neighborhood. We use AOPC as the fidelity metric, and compare the fidelity of explanations on the original neighborhood and the optimal neighborhood found on the path of narrowing the neighborhood.

Results. Figure 3 shows the deletion fidelity (AOPC) of explanations on the original and optimal neighborhoods when deleting the top 1 to 100 most important features. On average, using the optimal neighborhood improves fidelity by 11.4% relative to the original neighborhood. When deleting the 10 most important features, using the optimal neighborhood improves fidelity by 32.2% relative to the original neighborhood.

Moreover, Figure 4 provides a more fine-grained analysis of the fidelity improvement. Figure 4(a) provides an overview of the relationship between fidelity improvement, the number of deleted features, and the percentage of examples. Specifically, Figure 4(b) shows that over 20% of the examples can

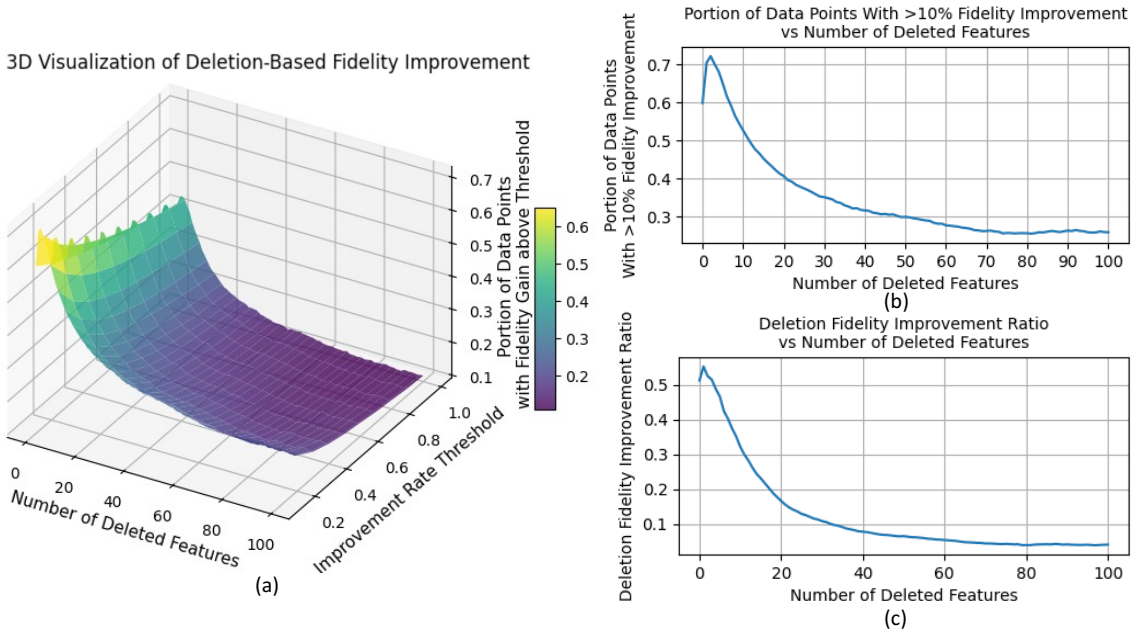

Figure 4: The fidelity of explanations on the original and optimal neighborhoods.

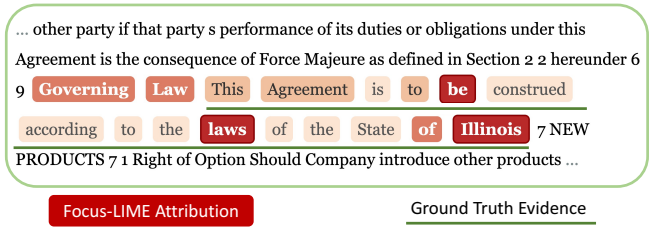

Figure 5: The important words identified by Focus-LIME and the ground-truth evidence highlighted by expert lawyers for the "Governing Law" clause checking task.

achieve more than 10% relative improvement in fidelity by using the optimal neighborhood across different deletion ratios, and when deleting less than 10 features, over 50% of the examples can achieve more than 10% relative improvement in fidelity. Figure 4(c) further shows that using the optimal neighborhood can significantly improve fidelity, especially when deleting a small number of features.

4.3 RQ3: Alignment with Human Evidence

RQ3 investigates the *practical utility* of Focus-LIME: can it serve as a reliable tool for human users to locate critical information? To answer this, we utilize the ground-truth annotations provided in the CUAD dataset. A high-quality explanation for a robust model should align with these human annotations.

Quantitative Analysis: Evidence Localization. We treat the top-k words identified by the explainer as a "retrieval set" and evaluate their overlap with the human ground truth using **Recall**. As the length of human-annotated evidence varies across examples, we choose k as a relative ratio of the evidence length, specifically 100%, 150%, or 200% of the total number of words in the human-annotated evidence. In this experiment, we follow the setting of RQ1, using GPT-4o as the target model and Qwen3-30B-A3B and Qwen2.5-32B as proxy models.

k	100%	150%	200%
GPT-4o + Qwen3-30B-A3B	0.43	0.65	0.91
GPT-4o + Qwen2.5-32B	0.40	0.62	0.87

Table 3: Alignment with human-annotated evidence on **CUAD**, measured by Recall@k. Higher values indicate better evidence localization. GPT-4o is used as the target model, while Qwen3-30B-A3B and Qwen2.5-32B are used as proxy models.

As shown in Table 3, our methods can effectively localize the ground-truth evidence spans annotated by experts. In Figure 5, we present an example where Focus-LIME successfully identifies the critical clauses that determine the model's decision, demonstrating its practical utility in assisting users as they navigate and verify long documents.

5 Conclusion

In this paper, we introduced **Focus-LIME**, a framework designed to resolve the critical challenge of surgical word-level attribution dilution in long-context LLMs. By leveraging a cost-efficient proxy to rigorously curate the search space, our method restores the tractability of surgical interpretation, allowing the target model to focus its sampling budget on the most relevant subspace. Extensive experiments demonstrate that Focus-LIME makes word-level explanations for long-context LLMs practical and has the potential to improve explanation fidelity in more general settings.

Ethical Statement

This work presents an interpretability method that is generally not harmful, though its explanations could potentially be misused to probe model weaknesses or support adversarial attacks.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China (NSFC) under Grant No. W2411051.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Achtibat *et al.*, 2024] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: Attention-aware layer-wise relevance propagation for transformers, 2024.

[Burger *et al.*, 2023] Christopher Burger, Lingwei Chen, and Thai Le. Are your explanations reliable? investigating the stability of lime in explaining text classifiers by marrying xai and adversarial attack. *arXiv preprint arXiv:2305.12351*, 2023.

[Dasigi *et al.*, 2021] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers, 2021.

[Ding *et al.*, 2024] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

[Dwivedi *et al.*, 2023] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), January 2023.

[Elisha *et al.*, 2025] Yehonatan Elisha, Seffi Cohen, Oren Barkan, and Noam Koenigstein. Rethinking saliency maps: A cognitive human aligned taxonomy and evaluation framework for explanations, 2025.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Guo *et al.*, 2024] Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189, 2024.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *NeurIPS*, 2021.

[Jiang *et al.*, 2023] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models, 2023.

[Liu and Zhang, 2025] Junhao Liu and Xin Zhang. Rex: A framework for incorporating temporal information in model-agnostic local explanation techniques. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):18888–18896, Apr. 2025.

[Liu *et al.*, 2024a] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[Liu *et al.*, 2024b] Junhao Liu, Haonan Yu, and Xin Zhang. Conlux: Concept-based local unified explanations. *arXiv preprint arXiv:2410.12439*, 2024.

[Liu *et al.*, 2026] Junhao Liu, Haonan Yu, Zhenyu Yan, and Xin Zhang. Revitalizing black-box interpretability: Actionable interpretability for llms via proxy models, 2026.

[Ludan *et al.*, 2024] Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. Interpretable-by-design text understanding with iteratively generated concept bottleneck, 2024.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[Maas *et al.*, 2011] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[Ribeiro *et al.*, 2016] 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*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Santhiappan *et al.*, 2024] Sudarsun Santhiappan, Mukesh Kumar Reghu, Mahendran Saminathan, and Aakash Veerappan. Speeding up lime using attention weights. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data*

(11th ACM IKDD CODS and 29th COMAD), CODS-COMAD '24, page 444–448, New York, NY, USA, 2024. Association for Computing Machinery.

[Shankaranarayana and Runje, 2019] Sharath M. Shankaranarayana and Davor Runje. Alime: Autoencoder based approach for local interpretability. In Hujun Yin, David Camacho, Peter Tino, Antonio J. Tallón-Ballesteros, Ronaldo Menezes, and Richard Allmendinger, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2019*, pages 454–463, Cham, 2019. Springer International Publishing.

[Sun *et al.*, 2023] Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21826–21840. Curran Associates, Inc., 2023.

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

[Tan *et al.*, 2023] Zeren Tan, Yang Tian, and Jian Li. Glime: General, stable and local lime explanation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36250–36277. Curran Associates, Inc., 2023.

[Valvoda and Cotterell, 2024] Josef Valvoda and Ryan Cotterell. Towards explainability in legal outcome prediction models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7269–7289, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[Yang *et al.*, 2025a] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[Yang *et al.*, 2025b] Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[Yeh *et al.*, 2019] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[Yu *et al.*, 2026] Haonan Yu, Junhao Liu, and Xin Zhang. Manchors: Memorization-based acceleration of anchors via rule reuse and transformation, 2026.

[Zhao *et al.*, 2025] Yunlong Zhao, Haoran Wu, and Bo Xu. Leveraging attention to effectively compress prompts for long-context llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26048–26056, 2025.