Interpreting Black-Box Large Language Models with Sentence-Level Energy Landscapes

Maryam Rezaee **Pooriya Safaei** **Maryam Asgarinezhad** **S. Fatemeh Seyyedsalehi**
Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran
ms.maryamrezaee@gmail.com, pooriya.safaei@sharif.edu
maryamasgn123@gmail.com, seyyedsalehi@sharif.edu

Abstract

The widespread adoption of proprietary Large Language Models (LLMs) accessed through closed APIs has created a critical challenge for responsible deployment: a fundamental lack of interpretability. To address this, we propose a model-agnostic, post-hoc attribution interpreter operating at the sentence level. Our approach trains an Energy-Based Model (EBM) as a surrogate to capture the LLM's internal conceptual relationships between prompts and responses. This energy landscape guides the training of a lightweight interpreter network. Uniquely, our interpreter operates as a standalone tool; once trained, it quantifies the influence of prompt sentences on a user-specified target from the output without requiring further API queries to the LLM. By globally training a local interpreter across diverse inputs, our framework captures broader generation patterns. Experiments demonstrate that our EBM accurately simulates the target LLM, allowing the interpreter to effectively identify the prompt sentences most influential in generating specific target outputs.

1 Introduction

Large Language Models (LLMs) have demonstrated extraordinary performance across complex tasks. Consequently, researchers and developers are rapidly adopting them for diverse applications. However, the critical challenge facing this adoption is a fundamental lack of interpretability. Most powerful LLMs are proprietary and accessed strictly through closed-access APIs. Even when architectures and pre-training datasets are available, their complexity obscures exactly how outputs are generated. In high-stakes domains like medicine and law, this opacity is unacceptable. This prevents meeting the application-grounded standards for responsible deployment (Doshi-Velez and Kim, 2017).

Post-hoc attribution is a primary approach to addressing this opacity. These methods explain model behavior by performing instance-wise feature selection—an interpretability paradigm for identifying an importance vector that measures how much each specific input feature influences the prediction for a given instance (Chen et al., 2018). However, standard attribution techniques, including white-box and model-agnostic methods, struggle in the context of LLMs. White-box methods, which rely on gradients or activations, are incompatible with closed APIs. Furthermore, the faithfulness of popular proxies like attention weights has been challenged (Jain and Wallace, 2019). Model-agnostic methods exist (Ribeiro et al., 2016; Lundberg and Lee, 2017; Seyyedsalehi et al., 2024), but typically target discriminative models with well-defined outputs. Interpreting generative models is significantly harder as the problem is fundamentally ill-posed. These models utilize complex representations to produce high-dimensional outputs like text. Therefore, effective explanation is hindered by the output's interactivity and sheer volume (Schneider, 2024).

Alternatives like prompt-based self-explanation (Wei et al., 2022) are similarly problematic; they rely on the same process we seek to verify, leading to circular logic and motivated reasoning. Consequently, models often produce plausible-sounding yet unfaithful confabulations (Turpin et al., 2023). While automated prompt engineering can steer model behavior to mitigate biases, it is unsuitable for interpretation. These methods optimize instructions for pre-defined targets (Zhou et al., 2023; Clemmer et al., 2024), rendering them unusable for ambiguous interpretation tasks.

To address these limitations, we propose a model-agnostic, post-hoc attribution interpreter. We diverge from standard approaches by shifting the resolution from noisy tokens to coherent sentences, which we define as "concepts," with the goal of relating elements of the output directly to the user prompt at this concept level. While re-

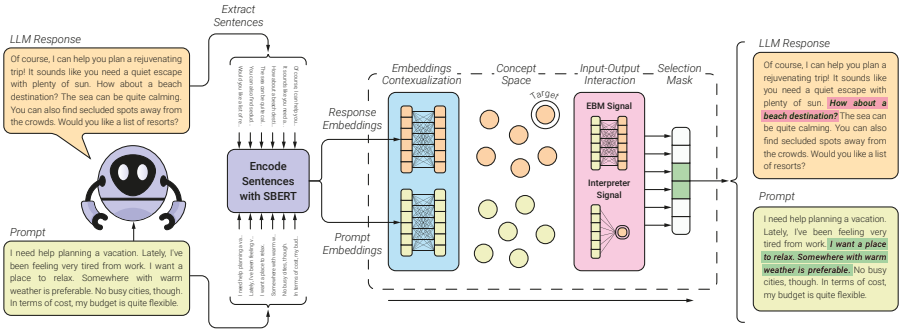


Figure 1: **Overview of the Proposed Framework.** The prompt and response of a black-box LLM are split into sentences and embedded via a pre-trained model. An encoding module maps these embeddings to a concept space where proximity reflects relevance. Subsequently, an interaction module evaluates the consistency between input and output concepts to identify the most influential prompt sentences. These modules are trained using signals from an energy network that simulates the generation process of the target LLM.

cent mechanistic work defines concepts as latent activation vectors (Gao et al., 2025) or distinct architectural bottlenecks (Sun et al., 2025), we adopt a *propositional* definition suitable for black-box analysis. Following the Context Principle (Frege, 1991), individual tokens remain semantically ambiguous without a propositional structure; a sentence, representing a complete thought (Kintsch, 1998), serves as a robust operational concept that captures the causal interactions necessary for generation. Formally, given a prompt $\mathbf{x}$ and an LLM response $\mathbf{y}$, we target a specific subset of the output $\mathbf{y}_T \subset \mathbf{y}$. We then produce an importance vector to identify the subset of prompt sentences $\mathbf{x}_S \subset \mathbf{x}$ that were most influential in generating $\mathbf{y}_T$.

We employ a unique paradigm to globally train a local interpreter. Unlike local interpreters (e.g. LIME (Ribeiro et al., 2016)) which observe only immediate neighborhoods—often causing interpretability illusions (Friedman et al., 2024)—we train across a wider distribution. This enables our model to capture global generation patterns.

Figure 1 outlines our approach. We first train a transformer-based Energy-Based Model (EBM) as a surrogate for the black-box LLM. This EBM maps sentences to a latent "concept space," which simulates the concept-level relationships embedded in the target LLM, and learns a scalar energy function to measure prompt-response compatibility. We chose an EBM because interpretation requires evaluating *global* consistency—which avoids the intractable normalization of probabilities over all possible outputs—unlike the local generation of standard autoregressive models. This energy landscape then guides the training of an interpreter network. Given a prompt and a target subset of the output, the interpreter produces an importance vector that isolates the prompt sentences most influential in generating the given subset.

In summary, our framework makes three core contributions: (1) shifting the unit of analysis from noisy tokens to sentences to enable human-intelligible, concept-level attribution; (2) introducing a transformer-based EBM with novel sampling methods capable of learning a surrogate random field over prompts and responses, with the aim of robustly modeling authentic input-output dynamics; and (3) proposing a post-hoc, model-agnostic framework for interpreting black-box LLMs that finds the specific prompt sentences responsible for triggering a target subset of the LLM's output.

2 Related Work

2.1 Post-hoc Interpretation Methods

Without access to internal circuits for mechanistic interpretability, post-hoc attribution remains the primary tool for explaining black-box behavior by scoring input feature importance for a specific model output. White-box approaches to attribution utilize gradients (Sundararajan et al., 2017; Shrikumar et al., 2017; Chefer et al., 2021) or attention weights (Li et al., 2017; Xie et al., 2017; Hao et al., 2021) to map output signals back to input tokens. However, both are inaccessible via proprietary APIs, and attention is frequently unfaithful to the generation process (Jain and Wallace, 2019).

Perturbation-based methods (Ribeiro et al., 2016; Lundberg and Lee, 2017; Yin and Neubig, 2022) provide a model-agnostic alternative by measuring sensitivity to input alterations, a strategy Hackmann et al. (2024) apply to identify influential

words in prompts. Chen et al. (2018) frames this as instance-wise feature selection by learning an explainer network to maximize mutual information. However, these methods incur prohibitive computational costs for generative tasks (Enouen et al., 2024; Zhao and Shan, 2024). Consequently, recent work explores training language models to produce answer decompositions as intermediate steps for attribution (Balasubramanian et al., 2025), or proposes surrogate frameworks to simulate LLM thinking (Chen et al., 2026). Unlike these approaches, which necessitate model fine-tuning or internal access, our method adopts simulation and decomposition within black-box constraints.

Finally, prompt-based self-explanation (Wei et al., 2022) prompts the LLM to generate rationales. This diverges from attribution, as it justifies outputs rather than isolating influential inputs. Furthermore, these rationales lack faithfulness guarantees, often serving as plausible post-hoc confabulations (Turpin et al., 2023).

2.2 Energy-Based Models in NLP

Energy-Based Models (EBMs) efficiently model high-dimensional structured outputs like text without requiring a normalized probability distribution (LeCun et al., 2006). While standard probabilistic models must intractably sum to one over all possible outputs, EBMs circumvent this by learning a scalar compatibility score where low energy indicates high data density.

In Natural Language Processing (NLP), EBMs directly refine generation. Residual EBMs add a corrective energy term to autoregressive log-probabilities to capture high-level coherence (Deng et al., 2020; Bakhtin et al., 2021), while other approaches use EBMs to define target landscapes for student network knowledge distillation (Tu et al., 2020). Recently, Xu et al. (2025) extended these principles to diffusion models, validating that energy landscapes effectively capture the complex distributions of modern text generation.

Beyond generation, EBMs serve as holistic evaluators and post-hoc rankers. Transformer-based discriminators can act as EBMs to distinguish human from machine text by scoring global sequence structure (Bakhtin et al., 2019). This evaluative capacity naturally extends to post-hoc refinement, where EBMs are utilized to rerank candidate outputs and improve generation quality (Bhattacharyya et al., 2021). Most recently, EBMs have enabled robust language model alignment by explicitly modeling reward distributions—instead of standard scalar points—to capture the uncertainty in human preferences (Lochab and Zhang, 2025).

2.3 Concept-Based Explanations

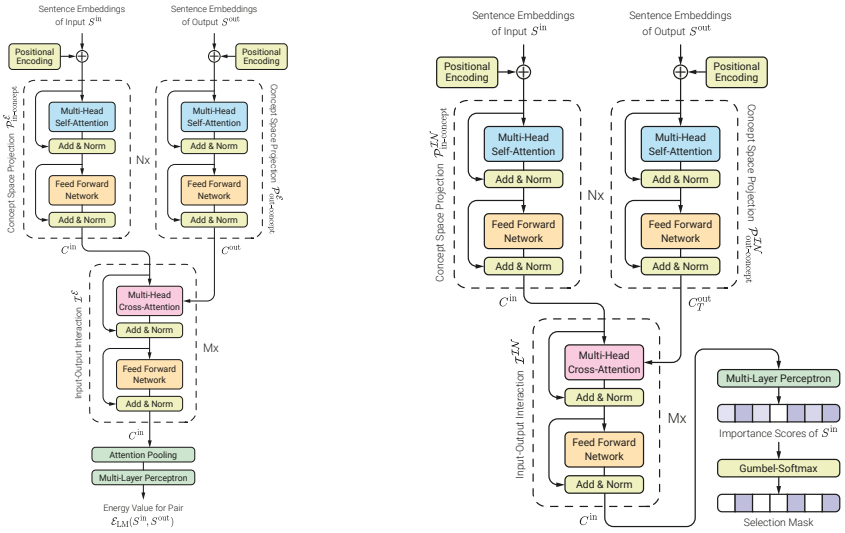
Interpretability increasingly relies on concept-based explanations to map decisions to human-intelligible ideas (Kim et al., 2018). While mechanistic methods dominate white-box settings via Sparse Autoencoders (Gao et al., 2025) or Concept Bottleneck layers (Sun et al., 2025), they require internal access. We address the black-box regime by mapping decisions to propositions rather than latent features. This aligns with rationale evaluation standards, which demand propositional evidence over isolated tokens to ensure faithfulness (DeYoung et al., 2020). We operationalize this by defining the "sentence" as our conceptual unit, relying on its ability as a robust thought for interpretation.

Treating sentences as semantic objects has historical precedence. BERT utilized Next Sentence Prediction to learn logical relationships (Devlin et al., 2019), while Sentence-BERT confirmed that fine-tuned representations map sentences with similar meanings to proximate points in a semantic space (Reimers and Gurevych, 2019). By leveraging sentences as concepts, our approach parallels recent architectural innovations like Large Concept Models, which shift from tokens directly to sentence-level representations (Barrault et al., 2024).

3 Methodology

We propose a post-hoc, model-agnostic interpreter for black-box LLMs that identifies the input sentences driving part of a response. Departing from standard token-level attribution, we establish the sentence as our fundamental analytical unit. Representing the smallest complete proposition, sentences serve as "concepts," framing generation as an interplay of coherent ideas rather than ambiguous tokens. Formally, let $\mathbf{x}$ be the prompt and $\mathbf{y}$ be the LLM response. We target a subset of output sentences, $\mathbf{y}_T \subseteq \mathbf{y}$, and quantify the influence of each concept in $\mathbf{x}$ on the generation of $\mathbf{y}_T$.

We propose a two-stage framework to achieve this. First, we pre-train an Energy-Based Model (EBM), $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$, to serve as a differentiable surrogate for the black-box LLM. As a function of θ, this model assigns scalar values representing the consistency of a prompt-response pair $(\mathbf{x}, \mathbf{y})$ with the target LLM's generation patterns. Second, we



(a) (b)

Figure 2: **Architectural Overview.** Schematics for (**a**) the energy function $\mathcal{E}_{\text{LM}}$ and (**b**) the interpreter network $\mathcal{IN}$.

leverage this energy landscape to guide the training of a lightweight interpreter, $\mathcal{IN}(\mathbf{x}, \mathbf{y}_T; \alpha)$, parameterized by α. Taking the prompt, response, and a user-specified target output $\mathbf{y}_T$ as inputs, the interpreter generates a sparse, binary vector matching the number of prompt sentences. In this vector, values of 1 identify the subset of prompt sentences $\mathbf{x}_S \subseteq \mathbf{x}$ strictly necessary for generating the target.

3.1 Sentence Extraction and Embedding

The pipeline begins by transforming the input text $\mathbf{x}$ and output text $\mathbf{y}$ into sequences of sentences. We first perform sentence segmentation using the spaCy library (Honnibal and Montani, 2017). Subsequently, we employ a frozen, pre-trained Sentence-BERT module (Reimers and Gurevych, 2019) to map each sentence to a fixed-dimensional vector. This yields embedding sequences S^{in} and S^{out}, which function analogously to token embeddings within our architecture. For further preprocessing and implementation details, including padding strategies, visit Appendix A.

3.2 The Energy-Based Surrogate Model

To approximate the black-box LLM's behavior, we design a globally-aware EBM. Unlike the target LLM, which predicts the next token $P(w_t|w_{<t})$, an EBM acts as a non-normalized compatibility function that evaluates an entire sequence holistically (LeCun et al., 2006). In our approach, the EBM assigns a scalar score to the joint configuration of a prompt and response, constructing an energy landscape over the input and output sentences to learn the shape of the LLM's generation manifold. Specifically, the model learns to distinguish variations of authentic prompt-response pairs from corrupted ones; the lower the assigned energy, the more likely the pair is consistent with concept interactions within the LLM.

As shown in Figure 2a, the architecture processes sentence embeddings in three stages:

Concept Space Projection. Static embeddings $(S^{\text{in}}, S^{\text{out}})$ capture meaning in isolation, but lack the context of the prompt and response. To remedy this, we pass these embeddings through separate, trainable self-attention modules $(\mathcal{P}^{\mathcal{E}}_{\text{in}}, \mathcal{P}^{\mathcal{E}}_{\text{out}})$ to project them into a dynamic concept space $(C^{\text{in}}, C^{\text{out}})$. Here, spatial proximity between vectors reflects the LLM's internal causal dependencies rather than semantic similarity. This target-specific geometry is shaped entirely by the LLM's underlying architecture and training dynamics.

Input-Output Interaction. A cross-attention block allows input concepts C^{in} to attend to output concepts C^{out}, weighing the causal influence of the prompt on the response.

Energy Calculation. The interacting representations are aggregated via attention pooling and passed through a Multi-Layer Perceptron (MLP) to output a scalar energy $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$.

The EBM is trained in two phases. First, it is pre-trained using a custom contrastive sampling strategy. Then, it is fine-tuned alongside the interpreter.

For pre-training, we generate a dataset of prompt-output pairs $(\mathbf{x}, \mathbf{y})$ from the target black-box LLM. To constrain the EBM to the target LLM's input-output dynamics, we employ two complementary contrastive objectives to balance our model's focus.

We define the *fidelity* objective ($\mathcal{L}_{\text{fidelity}}$) as an InfoNCE loss to capture the global generation signature. By treating responses from humans or other LMs as negative samples, we force the EBM to distinguish the target's authentic style.

Conversely, we define the *local dependency* objective ($\mathcal{L}_{\text{dep}}$) to target local conceptual links and causal interactions. We use two specific batch-wise samplers to achieve this. The first sampler, $(x_{\text{part}}, y'_{\text{part}})$, targets response dependency for the InfoNCE loss $\mathcal{L}_{\text{resp-dep}}$. It pairs a partially masked prompt x_{part} with its true partially masked response y_{part}. Keeping this partial prompt fixed, it then contrasts the true partial response against off-topic, partially masked responses y'_{part} from the same batch. This forces the model to verify that the output logically follows that specific input. The second sampler, $(x'_{\text{part}}, y_{\text{part}})$, mirrors this for prompt dependency in the InfoNCE loss $\mathcal{L}_{\text{pmt-dep}}$. It pairs a partial response y_{part} with its true partial prompt x_{part}. It then contrasts this against mismatched partial prompts x'_{part}. This ensures the model traces the output back to its causal antecedent.

The piece-by-piece evaluation of input-output dependencies in these samplers provides a critical advantage—it prevents the model from overfitting to surface heuristics, such as global topic matching or conversational fillers and dataset artifacts. By isolating partial sequences, the surrogate is forced to learn the actual sentence-level causal antecedents driving the target LLM's autoregressive generation.

Thus, we minimize the combined adaptive loss:

$$\mathcal{L}_{\text{total}} = (1-\lambda)\mathcal{L}_{\text{fidelity}} + \lambda\mathcal{L}_{\text{dep}} \tag{1}$$

where λ is a configurable weight and $\mathcal{L}_{\text{dep}}$ is the sum of the two sampler losses, $\mathcal{L}_{\text{resp-dep}}$ and $\mathcal{L}_{\text{pmt-dep}}$. We define the energy scoring term as $h(\mathbf{u}, \mathbf{v}) = \exp(-\mathcal{E}_{\text{LM}}(\mathbf{u}, \mathbf{v}; \theta)/\tau)$. Accordingly, the individual InfoNCE losses are formulated as:

$$\mathcal{L} = -\log\left(\frac{h(\mathbf{x}_i, \mathbf{y}_i)}{h(\mathbf{x}_i, \mathbf{y}_i) + \sum_{(\mathbf{x}', \mathbf{y}') \in \mathcal{N}_i} h(\mathbf{x}', \mathbf{y}')}\right) \tag{2}$$

Here, $\mathcal{N}_i$ constitutes the set of negative samples. Because the quality of the energy landscape hinges on these contrasts, we detail the specific sampling protocols and the full training hyperparameters in Appendix B. This pre-training phase (Fig. 4) yields a globally-aware energy function, which models how the target LLM maps dependencies between input and output sentences, and provides the supervision signal required to train the interpreter.

3.3 The Interpreter Model

Given prompt $\mathbf{x}$, response $\mathbf{y}$, and target $\mathbf{y}_T \subseteq \mathbf{y}$, the interpreter identifies prompt sentences influential on $\mathbf{y}_T$. It outputs a binary vector where 1 indicates a necessary precursor sentence.

Figure 2b illustrates the architecture, which mirrors the EBM's three stages with targeted modifications. As before, embeddings S^{in} and S^{out} are projected into the concept space via self-attention. In the interaction phase, however, we retain only the target concepts C_T^{out} and mask the remainder of the output. The input concepts C^{in} then attend to these targets via cross-attention. Finally, an MLP and Gumbel-Softmax unit (Jang et al., 2017) (see App. C) process the results to yield a binary importance vector for the input sentences.

Let $\tilde{\mathbf{x}} = \mathbf{x} \odot \mathcal{IN}(\mathbf{x}; \mathbf{y}_T, \alpha)$ be the selected subset, where $\odot$ denotes element-wise multiplication that applies the binary mask directly to the input sentences. A successful selection isolates the exact causal antecedents of the target. Therefore, the interpreter must minimize the energy of the selected pair $(\tilde{\mathbf{x}}, \mathbf{y}_T)$ to ensure the subset contains compatible information. Simultaneously, it must maximize the energy of the unselected remainder $(\mathbf{x} - \tilde{\mathbf{x}}, \mathbf{y}_T)$ to guarantee that no causally relevant information is left behind in the discarded text.

Using the pre-trained EBM as a critic, the interpreter optimization is thus:

$$\hat{\alpha} = \underset{\alpha}{\text{argmax}}\, \mathbb{E}_{(x,y)}\Big[\mathcal{E}_{\text{LM}}(\mathbf{x} - \tilde{\mathbf{x}}, \mathbf{y}_T; \theta) - \mathcal{E}_{\text{LM}}(\tilde{\mathbf{x}}, \mathbf{y}_T; \theta)\Big] \tag{3}$$

However, masking inputs inherently causes distribution shifts (Hsia et al., 2024). To prevent this, we fine-tune the EBM alongside the interpreter via a periodic alternating optimization strategy, summarized in Algorithm 1 (further details in App. D.1 and Fig. 5). By intermittently querying the target LLM with dynamically masked prompts to generate fresh responses at fixed epoch intervals, we realign the EBM to the interpreter's evolving distribution. While this joint training incurs API costs, our experiments suggest it is optional for standard benchmarks yet highly beneficial for robust, large-scale deployments.

Algorithm 1 Interpreter Alternating Optimization

Require: Prompt $\mathbf{x}$, Target Output $\mathbf{y}_T$, Initialized Interpreter $\mathcal{IN}_\alpha$, Pre-trained EBM $\mathcal{E}_\theta$, Grounding Interval N_{ground}

for each training step k **do**
  // Step 1: Interpreter Update
  Generate mask: $M \leftarrow \mathcal{IN}(\mathbf{x}; \mathbf{y}_T, \alpha^{(k-1)})$
  Update $\alpha^{(k)}$ via gradient descent on Eq. 3 to maximize energy gap using frozen $\mathcal{E}_{\theta^{(k-1)}}$
  // Step 2: Periodic EBM Fine-Tuning
  if $k \pmod{N_{\text{ground}}} == 0$ **then**
    Apply hard mask: $\tilde{\mathbf{x}} \leftarrow \mathbf{x} \odot \mathbb{I}(M > 0.5)$
    Query target LLM: $\tilde{\mathbf{y}} \leftarrow \text{LLM}(\tilde{\mathbf{x}})$
    Update $\theta^{(k)}$ to minimize Eq. 1 for new pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$
  else
    Keep EBM frozen: $\theta^{(k)} \leftarrow \theta^{(k-1)}$
  end if
end for

This training process distills the EBM's learned knowledge into the interpreter, thus enabling standalone inference with no reliance on the EBM or need for further LLM API calls.

4 Experiments

Our empirical validation is performed in two stages. First, we conduct an ablation study on the EBM's objective to justify why our tailored contrastive landscape best captures the target LLM's causal dependencies, rather than overfitting to surface heuristics. Second, we train and evaluate the interpreter across two critical axes: *attribution plausibility* and *causal faithfulness*. This ensures that the interpreter is capable of causally valid instance-wise feature selection. Crucially, our framework yields *task-specific* interpreters, where each surrogate is optimized for a targeted domain (e.g., general Q&A). This focused scope allows for effective training even with limited data.

4.1 Ablation Study: Dual-Objective EBM

We base our transformer EBM architecture on the ablations of Bakhtin et al. (2019). However, we introduce critical modifications to the embedding and the objective. The purpose of our concept space projector for embeddings is clear: it contextualizes individual sentence embeddings within the broader text. The objective function (Eq. 1), however, is more complex. We posit that a faithful surrogate must balance two competing requirements: *fidelity*, to capture the target LLM's global distribution, and *local dependency*, to isolate specific sentence-level causal antecedents of the output. While we tested numerous negative samplers and loss formulations to validate this dual objective, we present three representative configurations here: $\mathcal{M}_{\text{Fidelity}}$ ($\lambda = 0$, mimicking standard likelihood), $\mathcal{M}_{\text{Dep}}$ ($\lambda = 1$, isolating semantic links without overfitting to artifacts), and our proposed $\mathcal{M}_{\text{Hybrid}}$ ($\lambda = 0.9$, a dual objective with the optimally observed balance).

Experimental Setup. To construct our corpus, we sampled prompts from the HC3 multi-domain Q&A dataset (Guo et al., 2023) to query our target LLM, `GPT-4o-Mini` (OpenAI et al., 2024), for interpretation. Original HC3 human answers and `GPT-2-Medium` (Radford et al., 2019) generations served as contrastive baselines in the *fidelity* objective. For efficiency, we trained compact 181M-parameter EBMs (∼71M trainable) on $20,000$ pairs; preliminary tests indicate framework scalability, with larger models yielding wider energy gaps and faster convergence, which correlated with higher accuracy in distinguishing authentic pairs. Full configurations are detailed in Appendix B.

Evaluation Framework. To quantify sentence importance across these tests, we employ an ablation-based methodology (Li et al., 2017). For a given prompt and response pair, we establish a baseline energy E_{base}. We ablate each sentence s_i to compute its *Relative Energy Impact*: $\Delta E_i = \mathcal{E}_{\text{ablated}} - E_{\text{base}}$. A higher ΔE_i implies the sentence was necessary for low-energy alignment.

Using this metric, we derived four testing dimensions by analyzing the failure modes observed across various configurations and aligning each with established challenges. We ground the first two in human intuition to test whether the EBM notices input signals essential for the LLM's generation while ignoring output artifacts: (1) **Input Directive Dependency** evaluates the isolation of the main query from background noise, and (2) **Output Semantic Robustness** verifies ambivalence towards semantic fillers. Evaluating the substance of the output, however, is more difficult. Because full generations are too complex for manual heuristics, we use foundation models for (3) **Output Information Ranking** to scalably assess the semantic hierarchy of the response. Finally, we test the exact bidirectional mechanics required by our downstream interpreter through (4) **Causal Disentanglement**, ensuring the EBM separates true causal antecedents from topically related distractors.

Table 1 summarizes the results across these four dimensions, which are further analyzed below and detailed in Appendix E.

Table 1: **EBM Ablation Study Results.** We evaluate the models across four proposed dimensions. $\mathcal{M}_{\textbf{Fidelity}}$ suffers from reliance on artifacts. $\mathcal{M}_{\textbf{Dep}}$ achieves high margins but creates an abstract latent space that hinders downstream interpretation. $\mathcal{M}_{\textbf{Hybrid}}$ balances robustness with causal precision. *Metrics:* **IR@1**: Interrogative Recall@1; **SNR**: Signal-to-Noise Ratio; **Art.** ΔE: Artifact Energy Impact; **R1E**: Rank-1 Error; **nDCG@3**: Normalized Discounted Cumulative Gain; **S-Prec@1**: Soft Precision@1; **Acc**: Counterfactual Accuracy; **ESM**: Energy Separation Margin.

Model	I. Directive Dep.		II. Semantic Robust.		III. Information Rank.		IV. Causal Disentanglement	
	IR@1 ↑	SNR ↑	Art. ΔE ↓	R1E ↓	nDCG@3 ↑	S-Prec@1 ↑	Acc ↑	ESM ↑
$\mathcal{M}_{\text{Fidelity}}$	67.62%	1.45	+0.446	85.1%	0.553	26.0%	62.18%	0.145
$\mathcal{M}_{\text{Dep}}$	80.21%	6.89	**-0.014**	**6.9%**	0.699	52.0%	84.16%	**0.499**
$\mathcal{M}_{\text{Hybrid}}$	**84.77%**	**7.15**	+0.104	15.17%	**0.796**	**81.0%**	**92.40%**	0.382

Dimension I: Input Directive Dependency. In the ERASER benchmark (DeYoung et al., 2020), rationale extraction evaluates whether a model's selected evidence matches human logic. Adapting this framework to a black-box Q&A, we assess the model's ability to isolate the core directive (the question, S_Q) from the remaining non-interrogative context (S_R) as the main causal antecedent of the full model response. We define S_Q strictly as sentences containing explicit interrogative structures, such as ending in a question mark or starting with standard interrogative pronouns.

To quantify this, we first define *Interrogative Recall@1* (IR@1). Functioning as *Top-1 Recall* (Manning et al., 2008), it measures the frequency with which an interrogative sentence produces the maximum energy impact: $\text{IR@1} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[\text{argmax}_i(\Delta E_{j,i}) \in S_{Q,j}]$. We note that perfect recall is not expected, as human prompts frequently rely on ambiguous directives. Additionally, we compute the *Signal-to-Noise Ratio* (SNR) as the average energy impact of directives divided by the average impact of context sentences.

As shown in Table 1, $\mathcal{M}_{\text{Fidelity}}$ exhibits diffuse attention sensitive to background noise. Conversely, $\mathcal{M}_{\text{Hybrid}}$ achieves a $5\times$ SNR improvement, confirming that our *local dependency* objective compels the surrogate to prioritize directives.

Dimension II: Output Semantic Robustness. A prevalent pathology in neural models is reliance on spurious "annotation artifacts" (Gururangan et al., 2018). Drawing from input reduction protocols (Feng et al., 2018)—where models maintain high confidence on meaningless text—we measure susceptibility to fillers. To this end, we filtered the dataset for responses containing isolated fillers (e.g., "Okay!", defined as set S_{Art}).

To evaluate this, we define *Artifact Energy Impact* (Art. ΔE) as the average energy drop when these artifacts are removed: $\text{Art. } \Delta E = \frac{1}{|S_{Art}|}\sum_{i\in S_{Art}}\Delta E_i$. Furthermore, we define *Rank-1 Error* (R1E) as the *False Discovery Rate* (Manning et al., 2008) at cutoff $k = 1$ (FDR@1) to measure the proportion of samples where an artifact erroneously receives the highest importance score: $\text{R1E} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[\text{argmax}_i(\Delta E_{j,i}) \in S_{Art,j}]$.

Results show that $\mathcal{M}_{\text{Fidelity}}$ suffers from this pathology, minimizing energy by over-prioritizing fillers. Conversely, $\mathcal{M}_{\text{Dep}}$ successfully ignores artifacts, yet its abstract latent space led to downstream interpreter collapse in our tests. $\mathcal{M}_{\text{Hybrid}}$ balances this trade-off, retaining the latent space regularization necessary for training while resisting spurious correlations.

Dimension III: Output Information Ranking. To evaluate the EBM's grasp of the informational hierarchy in long generations, we test its ability to rank response sentences by their importance to the overall answer. Specifically, given a full prompt, we sorted its response sentences by their ΔE_i in descending order to form the EBM's predicted ranking. For ground truth, we prompted a powerful yet cost-effective foundation model, `Gemini-2.5-Flash` (Comanici et al., 2025), to assign a relevance score $rel \in \{0, \ldots, 5\}$ to each sentence based on its informational density—ranging from 0 (conversational fluff) to 5 (critical thesis statements).

Evaluation of ranking quality is done through the *Normalized Discounted Cumulative Gain* at cutoff $k = 3$ (nDCG@3) (Järvelin and Kekäläinen, 2002). Let $\pi(i)$ be the index of the sentence ranked i-th by the EBM. The discounted gain is calculated as:

$$\text{DCG@}k = \sum_{i=1}^{k}\frac{2^{rel_{\pi(i)}} - 1}{\log_2(i+1)} \tag{4}$$

This value is then normalized by the ideal DCG (IDCG), which is derived from a perfect sorting of

the oracle's scores. This metric heavily penalizes the model if it fails to place high-value sentences in the top ranks. Additionally, we report *Soft Precision@1* (S-Prec@1). Functioning as a form of *Top-1 Precision* (Manning et al., 2008), it measures the proportion of samples where the EBM's highest-ranked sentence is genuinely important ($rel \geq 4$).

As we see, $\mathcal{M}_{\text{Hybrid}}$ achieves the highest alignment, indicating that the hybrid objective produces sentence rankings most consistent with the generation patterns of state-of-the-art foundation models.

Dimension IV: Causal Disentanglement. This test evaluates the model's ability to identify specific causal links between input and output concepts, rather than relying on mere topical association. This aligns with the counterfactual evaluation principles of Kaushik et al. (2020), who established that robust models must learn "the difference that makes a difference" by distinguishing true causal antecedents from topically related context. To evaluate this disentanglement, we require an oracle to identify sentence relevance for our test.

Because isolating causal antecedents from mere topical overlap is linguistically sensitive, we prompted a state-of-the-art, high-capacity foundation model, `Gemini-3-Pro` (Comanici et al., 2025), to generate counterfactual triplets from the validation data. Each triplet consists of a target response sentence (y_{target}), its main causal antecedent among prompt sentences (x_{cause}), and a topically related but causally irrelevant distractor sentence from the same prompt ($x_{\text{distractor}}$). For a successful disentanglement, the surrogate must assign strictly lower energy to the true causal pair.

We quantify this using two metrics. First, *Counterfactual Accuracy* (Acc) measures the percentage of triplets where the EBM correctly identifies the true cause. Second, we define the *Energy Separation Margin* (ESM) to capture the model's confidence in its distinction of the pairs, calculated as the average magnitude of their energy difference: $\text{ESM} = \frac{1}{N}\sum_{j=1}^{N}(\mathcal{E}(x_{\text{distractor}}, y_{\text{target}}) - \mathcal{E}(x_{\text{cause}}, y_{\text{target}}))$.

Results show that while $\mathcal{M}_{\text{Dep}}$ achieves the highest ESM—meaning a highly discriminative, sharp energy landscape—this hyper-discrimination reduces its overall accuracy. Meanwhile, $\mathcal{M}_{\text{Hybrid}}$ achieves the highest accuracy, successfully balancing discriminative margin confidence with the robustness required to filter spurious correlations.

4.2 Interpreter Attribution Plausibility

Evaluating explanations for black-box models is fundamentally challenging as no dataset captures the target LLM's internal computation. We must therefore first assess attribution plausibility. Existing attribution datasets are too simplistic to apply to open-ended generation tasks, and human annotation itself reflects human cognitive priors rather than actual model mechanics. In contrast, high-capacity LLMs share the target model's underlying inductive biases (e.g. an autoregressive architecture), allowing them to approximate its internal mechanisms better than human annotators.

We thus assess semantic plausibility quantitatively against LLM "oracles" and qualitatively via human analysis of the results. While lacking absolute ground truth, high alignment with these proxies indicates sensible surrogate attributions.

(a) Scenario A: All Targets (Soft Top-1 Accuracy)

Interpreter	Ours	Gemini	GPT-4o	4o-Mini	GPT-J	GPT-2
ESCI (Ours)	**1.00**	0.75	0.75	0.64	0.79	0.70
Gemini-2.5-Flash	0.67	**1.00**	**0.87**	0.75	0.64	0.58
GPT-4o	0.70	**0.91**	**1.00**	**0.77**	0.70	0.68
GPT-4o-Mini	0.60	0.85	0.81	**1.00**	0.66	0.59
GPT-J-6B	**0.75**	0.69	0.63	0.47	**1.00**	**0.82**
GPT-2-XL	0.67	0.71	0.67	0.51	**0.87**	**1.00**

(b) Scenario A: All Targets (nDCG Score)

Interpreter	Ours	Gemini	GPT-4o	4o-Mini	GPT-J	GPT-2
ESCI (Ours)	**1.00**	0.83	0.82	0.79	0.83	0.81
Gemini-2.5-Flash	**0.85**	**1.00**	**0.89**	0.88	0.79	0.75
GPT-4o	0.81	**0.91**	**1.00**	**0.90**	0.84	0.79
GPT-4o-Mini	0.77	0.89	0.88	**1.00**	0.80	0.75
GPT-J-6B	**0.85**	0.76	0.75	0.74	**1.00**	**0.85**
GPT-2-XL	0.81	0.78	0.80	0.77	**0.90**	**1.00**

(c) Scenario B: Last Target (Soft Top-1 Accuracy)

Interpreter	Ours	Gemini	GPT-4o	4o-Mini	GPT-J	GPT-2
ESCI (Ours)	**1.00**	0.83	0.82	0.71	**0.97**	**0.92**
Gemini-2.5-Flash	0.77	**1.00**	**0.87**	0.75	0.64	0.58
GPT-4o	0.78	**0.91**	**1.00**	**0.82**	0.71	0.59
GPT-4o-Mini	0.66	0.85	0.83	**1.00**	0.68	0.56
GPT-J-6B	**0.98**	0.69	0.70	0.49	**1.00**	0.79
GPT-2-XL	0.96	0.71	0.71	0.50	0.85	**1.00**

(d) Scenario B: Last Target (nDCG Score)

Interpreter	Ours	Gemini	GPT-4o	4o-Mini	GPT-J	GPT-2
ESCI (Ours)	**1.00**	0.86	0.85	0.82	**0.94**	**0.96**
Gemini-2.5-Flash	0.83	**1.00**	**0.93**	0.92	0.82	0.78
GPT-4o	0.82	**0.95**	**1.00**	**0.95**	0.86	0.82
GPT-4o-Mini	0.81	0.93	**0.93**	**1.00**	0.80	0.75
GPT-J-6B	**0.92**	0.79	0.78	0.77	**1.00**	0.84
GPT-2-XL	0.90	0.81	0.79	0.77	0.88	**1.00**

Table 2: **Confusion Matrices Evaluating Interpretation Plausibility.** Each cell (i, j) represents how well the Interpreter in row i matches the scores of the Oracle in column j for interpreting our one target LLM. **Soft Top-1** (left) measures if the Interpreter's top-1 appears in the Oracle's top-2. **nDCG** (right) measures ranking correlation.

ID	Prompt Sentences	ESCI	GPT-4o	Gemini	Response Context & Target Sentence
6	[0] Why do people say "half a dozen" instead of "six"? [1] It seems like such a common occurrence. [2] Why take the time and effort to say the extra . . . [3] Explain like I'm five.	0.00 0.00 0.00 **1.00**	**0.70** 0.20 0.10 0.00	**0.80** 0.00 0.00 0.20	Okay! Imagine you have a box of cookies. If you have six cookies, you can just say "six." But if you say "half a dozen," it's like saying "half of a bigger group" of cookies. People like to use "half a dozen" because it sounds a little fancier. . .
21	[0] RGB lines when you take a picture of your monitor . . . [1] Please explain like I'm five.	**0.98** 0.02	**0.80** 0.20	**1.00** 0.00	. . . the camera gets a little mixed up and shows the colors in a funny way. That's why you see those RGB lines! . . .
33	[0] fuel octane. [1] What happens if I feed my Nissan Versa . . . [3] Please explain like I'm five.	**1.00** 0.00 0.00	0.00 **0.70** 0.30	0.00 **0.70** 0.30	. . . not work anymore! So, it's best to stick with the regular gas (87 octane) that your car is designed to use. That way, it will run smoothly and be happy!
48	[0] What those black lines on the road are. [1] EDIT: Sorry about the confusion, I meant . . . [3] Explain like I'm five.	0.00 **0.99** 0.01	0.40 0.10 **0.50**	**0.55** 0.00 0.45	. . . lines show where the lanes are, while others can tell you if you can park or if you need to stop. They are like guides that help everyone follow the rules of the road!
94	[0] What's the point of finding planets light years . . . [2] Why can't we spend money on improving . . . [3] Please explain like I'm five.	**1.00** 0.00 0.00	0.00 **0.70** 0.30	0.30 0.30 **0.40**	. . . while also exploring space, because both are important for our future. It's like making sure your toys are clean and also dreaming about getting new ones!
142	[0] The most prominent members of the current . . . [1] Explain like I'm five.	**1.00** 0.00	**0.80** 0.20	**0.90** 0.10	. . . their own thing and keep the country safe. People are talking a lot about these ideas as they get ready to vote! . . .

Table 3: **Qualitative Comparison of Attribution Scores. Left:** Prompt snippets. **Middle:** Attribution scores from ESCI and oracles. **Right:** Target sentence from the LLM response.

Experimental Setup. We utilize our best EBM configuration to train an interpreter for the target LLM, `GPT-4o-Mini` (see App. D for architecture). We then use this model to interpret a subset of 200 prompt-response pairs from our HC3-derived dataset (introduced in Sec. 4.1). This yields ∼2,000 distinct attribution tasks, where each task analyzes the causal influence of all sentences within a prompt on a single target response sentence. Following our rationale for proxy annotators, we selected five diverse LLMs (Table 4). `GPT-2-XL` (Radford et al., 2019) and `GPT-J-6B` (Wang and Komatsuzaki, 2021) serve as standard, open-weights causal language models of varying scales. Conversely, `GPT-4o`, `GPT-4o-Mini` (OpenAI et al., 2024), and `Gemini-2.5-Flash` (Comanici et al., 2025) represent state-of-the-art, heavily instruction-tuned proprietary models. Taking each distinct task, we prompted these oracles with an instruction to assign a normalized importance score (0.0 to 1.0) to every prompt sentence to quantify its causal contribution to the target response sentence. We then compare our Energy-Based Concept-Level Surrogate Interpreter (ESCI) against these oracles to measure attribution plausibility.

Table 4: **Model Scale Comparison.** Parameter counts of our framework alongside the five high-capacity LLMs used as oracles for attribution plausibility.

Model	Parameters (Approx.)
ESCI (Ours)	181M (∼71M Trainable)
GPT-2-XL	1.5B
GPT-J-6B	6B
GPT-4o-Mini	High-Capacity Proprietary
Gemini-2.5-Flash	High-Capacity Proprietary
GPT-4o	Massive Proprietary

Quantitative Assessment. Table 2 presents a cross-evaluation matrix detailing pairwise alignment, where each of the six models is evaluated both as the predicting interpreter and the ground-truth oracle. We use *nDCG* (computed exactly as in Sec. 4.1) to measure the full ranking correlation between each interpreter and oracle. Additionally, we define *Soft Top-1 Accuracy* to evaluate the top-ranked choice while accounting for attribution ambiguity; this measures whether the interpreter's highest-ranked prediction successfully falls within the oracle's top-k selections (where $k = 2$; see App. F for complete formulations).

Despite using only ∼71M trainable parameters, ESCI achieves competitive plausibility. In Scenario B (Tab. 2c/d), its close alignment with both standard causal (`GPT-J-6B`) and instruction-tuned (`Gemini-2.5-Flash`) models implies our energy landscape effectively internalizes autoregressive causal mechanics. Crucially, we observe that ESCI yields sparse, highly confident importance scores that sharply isolate *necessary* dependencies. In contrast, generative oracles tend to hedge their attributions and produce diffuse distributions with many non-zero probabilities across sentences. Additionally, white-box methods also require post-processing to even extract valid distributions. Against these challenges, ESCI establishes itself as a robust, standalone tool with comparable attribution strength.

Qualitative Case Studies. Table 3 details attribution behaviors reflecting trends from a manual analysis of all $2,000$ combinations by the authors. ESCI aligns with oracles on clear semantic mappings, but disagreements are revealing. In Sample 33, ESCI over-prioritizes the global topic over the specific question. Conversely, Sample 6

demonstrates ESCI's ability to disentangle stylistic from semantic drivers—crucial for instruction-tuned models—by correctly attributing a stylistic target to the "*Explain like I'm five*" instruction, whereas oracles fixate on semantics. Finally, Sample 94 highlights divergent interpretations when the target is a broadly synthesized analogy.

4.3 Interpreter Causal Faithfulness

While Section 4.2 confirms alignment with architecturally similar oracles, it does not guarantee these selections truly drive the LLM's computation. To validate feature selection under black-box constraints, we measure the impact of prompt interventions. Specifically, in generative scenarios, causal attributions can be logically verified by altering the input prompt and observing the corresponding semantic shift in the generated output. Thus, we adapt the standard metrics of *Sufficiency* and *Comprehensiveness* (DeYoung et al., 2020) to open-ended generation, using semantic similarity instead of discrete probabilities (Atanasova et al., 2023). **Generative Sufficiency** measures target regeneration using *only* selected prompt sentences, while **Generative Comprehensiveness** measures target loss when those sentences are *removed*. A faithful interpreter maximizes the former and minimizes the latter to yield a positive **Gap**.

Experimental Setup. We benchmark ESCI against four baselines using the same 200 pairs (∼2,000 targets) from Section 4.2. To validate our oracle assumptions, we first evaluate the target's self-assessment (prompting `GPT-4o-Mini` to attribute its own responses) alongside a stronger proxy oracle (`GPT-4o`). As before, these LLMs generate probability distributions over the prompt sentences. To establish a black-box upper bound, we also implement a sentence-level adaptation of LIME (Ribeiro et al., 2016); this treats full sentences, rather than individual tokens, as binary features that are randomly masked to train a linear surrogate and estimate importance scores. Finally, a sparse Random baseline serves as a lower bound to ensure our metrics heavily penalize arbitrary assignments. To convert the scores from all methods into definitive binary selections, we apply a dynamic max-ratio threshold: a sentence is selected if its assigned importance is at least 50% of the maximum score within its prompt.

Evaluation Framework. To quantify each method's faithfulness, we prompt the target LLM to generate *counterfactual responses*—new outputs produced when the model is fed only the prompt subset selected by that specific method as causal or non-causal antecedents of the target sentence. We then measure semantic retention of these new responses by computing the cosine similarity between their embeddings and the original target sentence ($\mathbf{y}_t$), utilizing the `all-mpnet-base-v2` model (Song et al., 2020). Letting $S(\cdot)$ denote the embedding function, we average these scores across all N targets. *Generative Sufficiency* evaluates the counterfactual generated from *only* the selected prompt sentences ($\mathbf{x}_S$): $\mathcal{M}_{\text{suff}} = \frac{1}{N}\sum_{j=1}^{N} \cos(S(\text{LLM}(\mathbf{x}_{S,j})), S(\mathbf{y}_{t,j}))$. Conversely, *Comprehensiveness* evaluates the counterfactual generated from the complement subset: $\mathcal{M}_{\text{comp}} = \frac{1}{N}\sum_{j=1}^{N} \cos(S(\text{LLM}(\mathbf{x}_j \setminus \mathbf{x}_{S,j})), S(\mathbf{y}_{t,j}))$. Detailed formulations and robustness checks justifying cosine similarity over strict logical entailment are provided in Appendix G.

Table 5: **Causal Faithfulness Evaluation.** Semantic similarity to the target under strict interventions. **Sufficiency:** Prompting with *only* selected sentences. **Comprehensiveness:** Selected sentences *removed*. **Gap:** Net causal contribution.

Interpreter	Suff. (↑)	Comp. (↓)	Gap (↑)
LIME (Baseline)	**0.439**	**0.194**	**0.245**
ESCI (Ours)	0.409	0.214	0.195
GPT-4o-Mini	0.407	0.215	0.192
GPT-4o	0.409	0.235	0.175
Random (Baseline)	0.231	0.377	-0.146

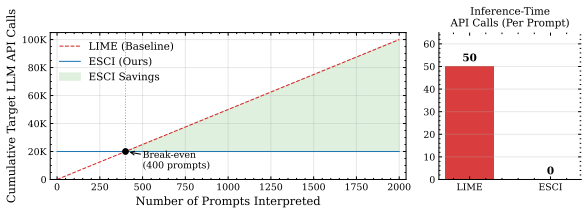


Figure 3: **Computational Scalability.** (Left) LIME API calls scale with prompts. (Right) ESCI eliminates inference API queries for high-volume deployability.

Analysis. Table 5 shows the Random baseline's negative gap confirms arbitrary selection degrades generation. Conversely, sentence-level LIME's exhaustive perturbations establish a faithfulness upper bound, but at prohibitive inference costs (Fig. 3), precluding scalable deployment. Against these extremes, ESCI matches high-capacity LLM

judges. While slightly trailing LIME's precision, ESCI achieves $O(1)$ inference with zero additional API queries. Its one-time pre-training cost ($\sim$20K queries) amortizes rapidly, breaking even after evaluating just 400 prompts. Although LLMs' background knowledge prevents near-zero *Comprehensiveness* across all models, ESCI successfully isolates necessary causal antecedents, providing a Pareto-optimal balance of computational efficiency and attribution fidelity.

5 Conclusion

We introduced a concept-level interpreter for black-box LLMs that shifts post-hoc attribution from tokens to sentences. By modeling generation dynamics as a differentiable energy landscape, we trained a standalone interpreter requiring zero inference API queries. We confirm this surrogate must balance *global fidelity* with *local dependency* to accurately reflect the target. Empirically, ESCI isolates causal sentences with precision approaching exhaustive methods like LIME, yet operates with deployment-ready efficiency, establishing EBMs as scalable tools for diagnosing model behavior.

6 Limitations and Future Work

Computational Trade-offs. A primary limitation of our framework is the upfront pre-training overhead. While ESCI achieves $O(1)$ efficiency at inference time with zero target API queries, it shifts the computational burden entirely to the training phase. Additionally, due to our surrogate's compact size, the framework is highly task-oriented. Although fully training a single model is manageable (e.g., under 30 hours on free-tier GPUs), the manual effort required to discover optimal hyperparameters across diverse tasks presents a significant barrier to out-of-the-box generalizability. Consequently, the tool is less accessible to users lacking the resources to perform this initial setup. Theoretically, a scaled-up EBM trained on multi-task data could yield a universal interpreter; however, within the current scope, a dedicated model remains necessary for each specific application.

Generalizability and Scaling Limits. Due to resource constraints, our validation focused on interpreting `GPT-4o-Mini` on standard Q&A tasks using compact surrogates ($\sim$181M parameters). Future work must expand to diverse domains including complex reasoning and open-ended generation (e.g., *TellMeWhy*, *WikiText*), and distinct, non-GPT architectures. Furthermore, while initial experiments suggest larger EBMs improve performance, their capacity to capture long-range dependencies within massive context windows (e.g., 128k+ tokens) remains unverified. Investigating the scaling laws of the interpreter is crucial to ensure robust attribution in high-complexity regimes.

Lack of Ground-Truth Mechanistic Validation. A fundamental limitation of the black-box setting is the reliance on probabilistic oracles (e.g., `GPT-4o`) rather than deterministic ground truth. While our sufficiency metrics demonstrate causal efficacy, high alignment with an oracle does not guarantee the best fidelity to the target's internal computation. Consequently, our current results confirm *behavioral* simulation rather than *mechanistic* alignment. Validating the latter requires future benchmarking against open-weights architectures (e.g., Llama 3, Pythia), where surrogate attributions can be directly compared with white-box signals like *Integrated Gradients* or attention maps. This would provide deeper theoretical insight to quantify how closely the surrogate energy landscape approximates the target model's true internal computational paths.

Human-Centric Utility. While our metrics confirm causal faithfulness, they do not guarantee that these attributions are actually intuitive or useful to humans. A key limitation of this work is the underlying assumption that mathematical accuracy equates to human intelligibility. Although we provide preliminary qualitative examples, rigorous human-subject studies remain necessary. Future evaluations must confirm whether ESCI practically aids users in downstream auditing tasks, such as detecting LLM hallucinations, identifying biases, or verifying safety compliance.

Scope of Application and Optimization. Our current evaluation is confined to the diagnostic utility of attribution, leaving the framework's broader downstream potential empirically unverified. Theoretically, identifying *necessary* sentences enables automated prompt optimization—pruning irrelevant context to reduce token costs without degrading quality. Similarly, the energy landscape offers a mechanism to audit Chain-of-Thought (CoT) reasoning by filtering unfaithful or confabulated intermediate steps. However, thorough experimentation is necessary to determine if the interpreter maintains robustness when deployed on these high-dimensional generative tasks.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294. Association for Computational Linguistics.

Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. 2021. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *Preprint*, arXiv:1906.03351.

Sriram Balasubramanian, Samyadeep Basu, Koustava Goswami, Ryan Rossi, Varun Manjunatha, Roshan Santhosh, Ruiyi Zhang, Soheil Feizi, and Nedim Lipka. 2025. Decomposition-enhanced training for post-hoc attributions in language models. *Preprint*, arXiv:2510.25766.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. 2024. Large concept models: Language modeling in a sentence representation space. *Preprint*, arXiv:2412.08821.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4528–4537. Association for Computational Linguistics.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791. IEEE.

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892. PMLR.

Lihu Chen, Xiang Yin, and Francesca Toni. 2026. Latent debate: A surrogate framework for interpreting llm thinking. *Preprint*, arXiv:2512.01909.

Colton Clemmer, Junhua Ding, and Yunhe Feng. 2024. Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8581–8590.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. 2024. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13984–14011. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728. Association for Computational Linguistics.

G. Frege. 1991. *The Foundations of Arithmetic*. Wiley.

Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. 2024. Interpretability illusions in the generalization of simplified models. In *Proceedings of the 41st International*

Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pages 14035–14059. PMLR.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112. Association for Computational Linguistics.

Stefan Hackmann, Haniyeh Mahmoudian, Mark Steadman, and Michael Schmidt. 2024. Word importance explains how prompts affect language model outputs. *Preprint*, arXiv:2403.03028.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971. AAAI Press.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary Lipton. 2024. Goodhart's law applies to NLP's explanation benchmarks. In *Findings of the Association for Computational Linguistics*, pages 1322–1335. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR.

W. Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Comprehension: A Paradigm for Cognition. Cambridge University Press.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0):191–246.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. *Preprint*, arXiv:1612.08220.

Anamika Lochab and Ruqi Zhang. 2025. Energy-based reward models for robust language model alignment. In *Second Conference on Language Modeling*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Association for Computing Machinery.

Johannes Schneider. 2024. Explainable generative ai (GenXAI): a survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289.

S. Fatemeh Seyyedsalehi, Mahdieh Soleymani Baghshah, and Hamid R. Rabiee. 2024. SOInter: A novel deep energy-based interpretation method for explaining structured output models. In *The Twelfth International Conference on Learning Representations*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2025. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. 2020. ENGINE: Energy-based inference networks for non-autoregressive machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2826. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 71725–71739. Curran Associates.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–962. Association for Computational Linguistics.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. 2025. Energy-based diffusion language models for text generation. In *The Thirteenth International Conference on Learning Representations*.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198. Association for Computational Linguistics.

Zhixue Zhao and Boxuan Shan. 2024. ReAGent: A model-agnostic feature attribution method for generative language models. In *Proceedings of the AAAI 2024 Workshop on Responsible Language Models (ReLM)*. Association for the Advancement of Artificial Intelligence.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *International Conference on Learning Representations*.

A Preprocessing Details

To ensure compatibility with fixed-dimensional attention mechanisms, we normalize sentence counts during preprocessing. We define a task-dependent hyperparameter, N_{max}, representing the maximum sequence length. After input and output sentences are extracted and embedded, the sequences are padded with a learnable placeholder token or truncated to strictly match this length. This results in dense input tensors $S^{\text{in}}, S^{\text{out}} \in \mathbb{R}^{N_{\text{max}} \times d}$, where d is the embedding dimension of the Sentence-BERT model (768 for `all-mpnet-base-v2`).

B Energy Network: Pre-training Details

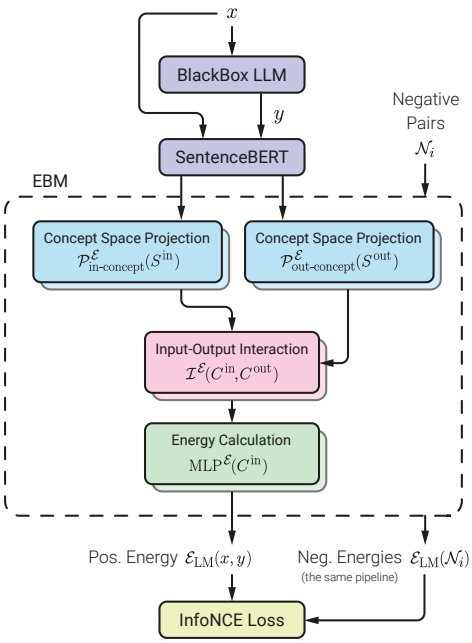


Figure 4: **Pre-training Pipeline of the EBM.** The architecture projects SentenceBERT embeddings into a dynamic concept space via self-attention, followed by a cross-attention mechanism to model input-output interactions. An MLP aggregates these features to compute a scalar energy score $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$. The model is optimized using a dual-objective InfoNCE loss: *fidelity* contrasts authentic pairs $(\mathbf{x}, \mathbf{y})$ against global negatives in $\mathcal{N}_i$ (e.g., human responses) to learn the target distribution, while *local dependency* contrasts partial sequences against batch negatives in $\mathcal{N}_i$ to enforce fine-grained causal precision. Thus, the weighted sum of InfoNCE losses minimizes the energy of authentic pairs while maximizing the energy of corrupted samples.

The Energy-Based Model's training pipeline is illustrated in Figure 4. We trained all EBM variants on dual NVIDIA T4 GPUs (provided by Kaggle's free tier) using the AdamW optimizer. The training process for each EBM required approximately 25 hours. Table 6 details the specific hyperparameters used for the final Hybrid model.

Table 6: **Hyperparameters for the $\mathcal{M}_{\text{Hybrid}}$ EBM.**

Parameter	Value
Architecture	
Encoder Model	`all-mpnet-base-v2`
Frozen Parameters	110M
Trainable Parameters	71M
Total Parameters	181M
Projection Dimension (d_{model})	768
Self-Attention Layers	2
Cross-Attention Layers	6
Attention Heads	8
Dropout Rate	0.1
MLP Layers	2
MLP Hidden Factor	2
Optimization	
Epochs	50
Batch Size	16
Learning Rate	$3e^{-5}$
Scheduler	Linear Warmup
Warmup Steps	200
Loss	
Loss Function	InfoNCE
Local Dependency Weight (λ)	0.9
InfoNCE Temperature (τ)	0.1
Margin	0.5
Negative Candidates (K)	5
Data	
Dataset Size	20,000 samples
Validation Split	10%
Max Sentence Count	16 (Learnable Padding)

Model Configurations. To assess the impact of our dual objectives, we trained three distinct EBM variants. $\mathcal{M}_{\textbf{Fidelity}}$ ($\lambda = 0$) mimics standard likelihood modeling by contrasting positive pairs against only global corruptions. $\mathcal{M}_{\textbf{Dep}}$ ($\lambda = 1$) learns exclusively by contrasting partial segments, forcing the model to identify semantic links without overfitting to the surface artifacts of a single authentic pair. Finally, $\mathcal{M}_{\textbf{Hybrid}}$ ($\lambda = 0.9$) combines these approaches; it relies on dependency samplers to capture structural logic while using the fidelity signal to regularize the latent space.

Negative Sampling Strategies. The training objective relies on a diverse set of negative samples to shape the energy landscape. The specific samplers used for each configuration are:

- $\mathcal{M}_{\textbf{Fidelity}}$ **Samplers:**
  - `response_human`: Swaps the LLM response with a human-written answer from the HC3 dataset.

- `response_other_lm`: Swaps response with a `GPT-2 Medium` output.
- `response_sentence_masking`: Masks a random number of sentences in the LLM's response.
- `prompt_sentence_masking`: Masks a random number of sentences in the data pair's prompt.
- `off_topic`: Swaps response or prompt with one from a different pair in the batch.

- $\mathcal{M}_{\text{Dep}}$ **Samplers:**
  - `partial_response_dep`: Contrasts the authentic partial response (positive) against a mismatched partial response from the batch (negative) given the same partial prompt. This forces the model to verify that the output is a specific logical continuation of the input concepts.
  - `partial_prompt_dep`: Contrasts the authentic partial prompt (positive) against a mismatched partial prompt from the batch (negative) given the same partial response. This ensures that the response is causally attributed to the correct input antecedents rather than generic topics.

- $\mathcal{M}_{\text{Hybrid}}$ **Samplers:**
  - `partial_response_dep`: See $\mathcal{M}_{\text{Dep}}$.
  - `partial_prompt_dep`: See $\mathcal{M}_{\text{Dep}}$.
  - `response_human`: See $\mathcal{M}_{\text{Fidelity}}$.
  - `response_other_lm`: See $\mathcal{M}_{\text{Fidelity}}$.

C Differentiable Top-K Sentence Selection via Gumbel–Softmax

The interpreter network aims to identify the K most important sentences from the input $\mathbf{x}$ influential in generating the target $\mathbf{y}_T$. Since selecting top-K indices is discrete and non-differentiable, we apply a continuous relaxation to the subset sampling via the Gumbel-Softmax trick (Jang et al., 2017; Chen et al., 2018) to enable end-to-end training.

Let the interpreter function produce a vector of unnormalized relevance logits $\mathbf{z} \in \mathbb{R}^n$ for the n input sentences, denoted as $z_i = (\mathcal{IN}(\mathbf{x}, \mathbf{y}_T; \alpha))_i$. To introduce stochasticity, we first generate standard Gumbel noise g_i from i.i.d. uniform samples $u_i \sim \text{Uniform}(0, 1)$ as follows:

$$g_i = -\log(-\log u_i), \quad i = 1, \ldots, n. \tag{C1}$$

Given a temperature $\tau > 0$, a single continuous relaxation of a one-hot vector, denoted as $c \in \Delta^{n-1}$, is computed via the softmax function:

$$c_i = \frac{\exp((z_i + g_i)/\tau)}{\sum_{j=1}^{n} \exp((z_j + g_j)/\tau)}, \quad i = 1, \ldots, n. \tag{C2}$$

As $\tau \to 0$, the vector c approaches a discrete one-hot sample from the categorical distribution defined by $\mathbf{z}$. To approximate a K-hot selection vector (selecting multiple sentences), we draw K independent relaxed samples $\{c^{(j)}\}_{j=1}^{K}$ using Equation C2. We then aggregate these samples by taking their element-wise maximum:

$$m_i = \max_{j=1,\ldots,K} c_i^{(j)}, \quad i = 1, \ldots, n. \tag{C3}$$

The resulting vector $\mathbf{m}$ serves as a continuous proxy for the binary mask. The final output of the interpreter used to gate the input sentences is:

$$\mathcal{IN}(\mathbf{x}, \mathbf{y}_T; \alpha)_i = m_i. \tag{C4}$$

During training, this soft mask allows gradients to backpropagate through the selection process. During inference, we obtain the discrete selection by taking the indices of the top-K logits directly or by hardening the soft mask.

D Interpreter: Training Details

We report the configuration and formulation for the best-performing interpreter, trained utilizing the EBM-guided framework on the Hybrid ($\lambda = 0.9$) energy landscape. The training process required approximately 1 hour on dual NVIDIA T4 GPUs (provided by Kaggle's free tier). Table 7 details the specific hyperparameters.

D.1 Alternating Optimization Details

While Section 3.3 outlines the high-level objective, we detail here the specific gradient updates required for training. To mitigate the distribution shift caused by masking (Fig. 5), we define the joint optimization loop. Let $\theta^{(k)}$ and $\alpha^{(k)}$ denote the parameters at step k.

Step 1: Interpreter Update. We freeze the EBM parameters $\theta^{(k-1)}$ and update the interpreter to improve selection precision. The gradient update is:

$$\alpha^{(k)} \leftarrow \alpha^{(k-1)} - \eta_\alpha \nabla_\alpha \Big(\mathcal{E}_{\text{LM}}(\mathbf{x} \odot M, \mathbf{y}_T; \theta^{(k-1)}) - \mathcal{E}_{\text{LM}}(\mathbf{x} \odot (1 - M), \mathbf{y}_T; \theta^{(k-1)}) \Big) \tag{D1}$$

Table 7: **Hyperparameters for the Interpreter.**

Parameter	Value
Architecture	
Encoder	`all-mpnet-base-v2`
Projection Dim (d_{model})	768
Self-Attention Layers	2
Cross-Attention Layers	6
Attention Heads	8
Dropout Rate	0.1
MLP Layers	1
MLP Hidden Dimension	256
Optimization	
Epochs	50
Batch Size	16
Learning Rate	$1e^{-5}$
Loss Function	InfoNCE ($\tau = 0.1$)
Selection Mechanism	Gumbel-Softmax
Gumbel Temperature	1.0
Data	
Dataset Size	$20,000$ samples
Validation Split	10%
Max Sentence Count	16 (Learnable Padding)

where $M = \mathcal{IN}(\mathbf{x}; \mathbf{y}_T, \alpha^{(k-1)})$ is the generated mask by the interpreter.

Step 2: Periodic Grounding and EBM Fine-tuning. Every N_{ground} steps, we generate a fresh training pair to re-align the EBM. We apply the current hard mask to the prompt and query the black-box LLM:

$$\tilde{\mathbf{x}} = \mathbf{x} \odot \mathbb{I}(M > 0.5) \tag{D2}$$

$$\tilde{\mathbf{y}} = \text{LLM}(\tilde{\mathbf{x}}) \tag{D3}$$

This creates a valid sample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ that represents the model's actual behavior under the current masking policy.

Subsequently, using this new sample at each grounding step, we update the EBM to minimize the energy of the new synthetic pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ while maintaining the structural constraints learned during pre-training. We employ the same dual-objective loss $\mathcal{L}_{\text{total}}$ defined in Equation 1 (Sec. 3.2), consisting of both $\mathcal{L}_{\text{fidelity}}$ and $\mathcal{L}_{\text{dep}}$.

However, because there is no ground-truth human response for the dynamically masked prompt $\tilde{\mathbf{x}}$, we modify the negative sampling set $\mathcal{N}_i$ for the fidelity objective (App. B). We substitute the `response_human` sampler with another distinct, model-based negative, `response_other_lm_stochastic`, to maintain distribution contrast. The gradient update is thus

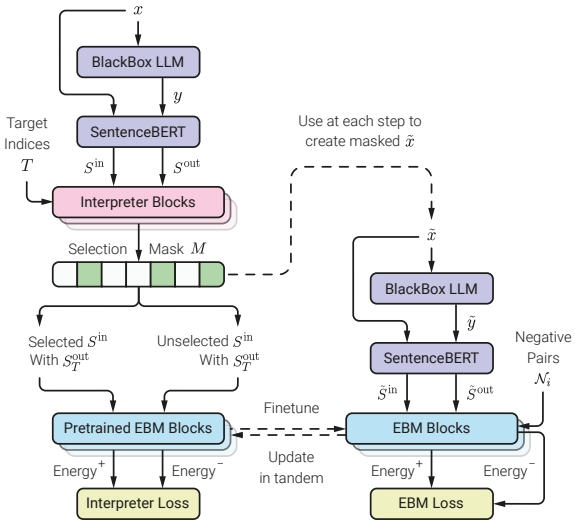


Figure 5: **Overview of the Alternating Optimization Protocol.** The framework employs a joint training strategy to prevent distribution shift. **(Left)** In the standard phase, the interpreter generates a binary mask over the prompt sentences; its parameters are updated to maximize the energy gap using the frozen EBM as a critic. **(Right)** Periodically, the EBM is fine-tuned to adapt to the interpreter's evolving distribution. This involves querying the target LLM with the currently masked prompt to obtain a fresh, ground-truth response, thereby grounding the energy landscape in the model's actual behavior under partial input.

computed using this modified negative set $\tilde{\mathcal{N}}$:

$$\theta^{(k)} \leftarrow \theta^{(k-1)} - \eta_\theta \nabla_\theta \Big((1-\lambda)\mathcal{L}_{\text{fidelity}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\mathcal{N}}) + \lambda \mathcal{L}_{\text{dep}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \Big) \tag{D4}$$

This alternating procedure ensures that as the interpreter's selections evolve, the energy landscape adapts to provide accurate supervision for those specific sparse inputs.

E Energy Network: Evaluation Protocols

To evaluate the EBM's semantic alignment beyond aggregate accuracy, we developed a suite of granular diagnostic tests. This section details the mathematical formulations, dataset filtering criteria, and specific metrics for each testing dimension.

E.1 Dataset Preparation & Filtering

For all diagnostic tests, we utilized specific subsets of the HC3 validation set ($N = 1000$). To generate the ground-truth importance scores used for evaluation, we employed an ablation-based energy drop methodology. For each sample pair $(\mathbf{x}, \mathbf{y})$,

we systematically removed each sentence to create variants. We calculated the energy for two modes:

- **Prompt Ablation:** Pairs $(\mathbf{x}_{\setminus i}, \mathbf{y})$, where $\mathbf{x}_{\setminus i}$ is the prompt with the i-th sentence removed.
- **Response Ablation:** Pairs $(\mathbf{x}, \mathbf{y}_{\setminus j})$, where $\mathbf{y}_{\setminus j}$ is the response with the j-th sentence removed.

The importance of a sentence was quantified by the positive energy drop caused by its removal relative to the baseline energy $\mathcal{E}(\mathbf{x}, \mathbf{y})$. Using these scored samples, we applied specific filters to isolate relevant linguistic phenomena:

- **Interrogative Subset ($N = 769$):** Used for *Dimension I*. We filtered for prompts containing explicit interrogative structures, defined as sentences ending in a question mark or starting with standard interrogative pronouns (e.g., "What", "How", "Why").
- **Artifact Subset ($N = 890$):** Used for *Dimension II*. We filtered for responses containing distinct conversational fillers (e.g., "Okay!", "Sure!", "Here is the answer:") appearing as isolated sentences.
- **Oracle Subset ($N = 500$):** Used for *Dimension III*. A random subset of validation samples was selected for external scoring by `Gemini-2.5-Flash`.
- **Counterfactual Subset ($N = 500$):** Used for *Dimension IV*. `Gemini-3-Pro` was employed to generate specific counterfactual triplets from validation data (see App. E.5 for details).

E.2 Dimension I: Input Directive Dependency

This test assesses the model's ability to distinguish the primary user intent (the directive) from supplementary context or conversational filler.

Ablation Methodology. For a given prompt $\mathbf{x}$ consisting of n sentences $\{s_1, s_2, \ldots, s_n\}$ and a fixed response $\mathbf{y}$, we calculate the baseline energy $E_{\text{base}} = \mathcal{E}(\mathbf{x}, \mathbf{y})$. We then systematically remove each sentence s_i to create an ablated prompt $\mathbf{x}_{\setminus i}$ and compute the *Relative Energy Impact* (ΔE_i):

$$\Delta E_i = \mathcal{E}(\mathbf{x}_{\setminus i}, \mathbf{y}) - E_{\text{base}} \tag{E1}$$

A positive ΔE_i implies that sentence s_i was necessary for the low-energy alignment (i.e., it was semantically important).

Metric Definitions. Let S_Q be the set of indices corresponding to interrogative sentences and S_R be the set of indices for the remaining context.

- **Interrogative Recall@1 (IR@1):** Adapting the *rationale extraction* evaluation protocol from **ERASER** (DeYoung et al., 2020), we define this as the frequency with which the sentence producing the maximum energy impact is an interrogative sentence.

$$\text{IR@1} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left[\text{argmax}_i(\Delta E_{j,i}) \in S_{Q,j}\right] \tag{E2}$$

  The metric yields a value in $[0, 1]$, where an ideal score of 1 indicates that the explicit question is consistently ranked as the primary causal driver. However, we note that perfect recall is not expected, as our dataset analysis revealed that human-written Q&A prompts often contain implicit or structurally ambiguous directives where the semantic core is not the grammatical question.

- **Attribution Signal-to-Noise Ratio (SNR):** Adapting standard signal processing definitions to attribution magnitude, we define this as the ratio of the average energy impact of questions to the average energy impact of non-question context sentences.

$$\text{SNR} = \frac{\frac{1}{|S_Q|} \sum_{i \in S_Q} \Delta E_i}{\frac{1}{|S_R|} \sum_{k \in S_R} \Delta E_k + \epsilon} \tag{E3}$$

  where $\epsilon = 1e^{-9}$ is a constant for stability. A high SNR indicates the model is highly sensitive to the directive and insensitive to noise.

E.3 Dimension II: Output Semantic Robustness

This test measures the model's susceptibility to non-semantic conversational artifacts, addressing a prevalent pathology in neural models known as reliance on spurious correlations or "annotation artifacts" (Gururangan et al., 2018). Grounded in concepts originally established for NLI datasets, this evaluation assesses whether the model has learned to treat high-frequency tokens (e.g., conversational fillers such as "Okay!") as proxies for output validity, independent of their actual semantic content.

Metric Definitions. Let S_{Art} be the set of indices corresponding to artifact sentences.

- **Artifact Energy Impact (Art. ΔE):** The average change in energy when an artifact is removed. This metric is adapted from *Input Reduction* methods (Feng et al., 2018), where we aim to measure the model's sensitivity to the removal of negligible features.

$$\text{Art. } \Delta E = \frac{1}{|S_{Art}|} \sum_{i \in S_{Art}} (\mathcal{E}(\mathbf{x}, \mathbf{y}_{\setminus i}) - E_{\text{base}}) \tag{E4}$$

- **Rank-1 Error (R1E):** The proportion of samples where an artifact sentence is assigned the highest importance rank (Rank 1). This metric quantifies the fidelity trap, where the model overfits to surface-level plausibility markers rather than semantic drivers.

$$\text{R1E} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left[\underset{i}{\text{argmax}}(\Delta E_{j,i}) \in S_{Art,j}\right] \tag{E5}$$

E.4 Dimension III: Output Information Ranking

This test validates the EBM's internal ranking of sentence importance against a gold standard ranking generated by a state-of-the-art LLM to assess ranking alignment and accuracy.

Oracle Setup. For each sample in the *Oracle Subset*, `Gemini-2.5-Flash` was provided with the prompt, response, and list of sentences as derived by the EBM, and instructed to assign an integer *Information Density Score* $y_i \in \{0, \ldots, 5\}$ to each response sentence s_i. The scoring criteria were:

- **0 (Fluff):** Purely conversational filler or phatic expressions (e.g., "Okay!") with zero informational value.
- **1 (Minor Context):** Generic transitions or polite formatting that aids flow but adds no unique content.
- **2 (Useful Background):** Contextual definitions or analogies that facilitate understanding without constituting the direct answer.
- **3 (Supporting Info):** Elaborations or details necessary for a complete explanation; removing these makes the answer feel thin.
- **4 (Important):** Key facts, steps, or reasoning that directly address the user's request.
- **5 (Critical):** The core thesis or direct solution; the response is conceptually incomplete without this sentence.

Metric Definitions. Let $\mathcal{S} = \{s_1, \ldots, s_M\}$ be the set of sentences in a response. Let rel_i be the oracle's score for sentence i, and let π be the permutation of indices induced by sorting the EBM's energy impact scores ΔE in descending order (i.e., $\pi(1)$ is the index of the most important sentence according to the EBM).

- **nDCG@3 (Normalized Discounted Cumulative Gain):** We measure the ranking quality at cutoff $k = 3$. The Discounted Cumulative Gain (DCG) is computed as:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_{\pi(i)}} - 1}{\log_2(i+1)} \tag{E6}$$

  The Ideal DCG (IDCG) is computed similarly using the permutation π^* that sorts the oracle's scores perfectly. The final metric is:

$$\text{nDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k} \tag{E7}$$

  This metric penalizes the model heavily if it fails to place high-value (oracle score 5) sentences in the top ranks.

- **Soft Precision@1:** This metric assesses the utility of the single most important sentence identified by the EBM. It is defined as the proportion of samples where the EBM's top choice received a high relevance score (≥ 4) from the oracle:

$$\text{S-Prec@1} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left[rel_{\pi_j(1)} \geq 4\right] \tag{E8}$$

E.5 Dimension IV: Causal Disentanglement

This test evaluates the model's ability to identify specific causal links between input and output concepts, distinguished from mere topical association.

Counterfactual Setup. For each sample in the *Counterfactual Subset*, the model identified:

1. A specific target response sentence (y_{target}).
2. The high-impact prompt sentence (x_{cause}) that directly necessitated y_{target}.
3. A low-impact distractor sentence ($x_{\text{distractor}}$) from the *same* prompt that was topically related but causally irrelevant to y_{target}.

Metric Definitions. We quantify discriminative performance by comparing the energy assigned to causal versus distractor antecedents. Let $\mathcal{E}(x, y)$ denote the scalar energy score, where lower values indicate higher compatibility. For a successful disentanglement, the model must assign strictly lower energy to the true cause than to the distractor, satisfying the condition $\mathcal{E}(x_{\text{cause}}, y_{\text{target}}) < \mathcal{E}(x_{\text{distractor}}, y_{\text{target}})$. We aggregate this behavior using two metrics:

- **Counterfactual Accuracy:** The percentage of triplets where the EBM correctly assigns lower energy to the causal pair.

$$\text{Acc} = \frac{100}{N} \sum_{j=1}^{N} \mathbb{I}[\mathcal{E}(x_{\text{cause}}, y_{\text{target}}) < \mathcal{E}(x_{\text{distractor}}, y_{\text{target}})] \tag{E9}$$

- **Energy Separation Margin (ESM):** The average magnitude of the energy difference between the distractor and the cause. A larger positive margin indicates higher confidence in the causal distinction.

$$\text{ESM} = \frac{1}{N} \sum_{j=1}^{N} (\mathcal{E}(x_{\text{distractor}}, y_{\text{target}}) - \mathcal{E}(x_{\text{cause}}, y_{\text{target}})) \tag{E10}$$

F Interpreter: Plausibility Evaluation

To construct the plausibility benchmark, we prompted five diverse LLMs (`Gemini-2.5-Flash`, `GPT-4o`, `GPT-4o-Mini`, `GPT-J-6B`, and `GPT-2-XL`) to act as data annotators.

Prompting Strategy. For a given sample tuple consisting of a prompt $P = \{s_1^p, \ldots, s_n^p\}$ and a specific target response sentence s_t^r, each oracle was provided with the full text context and instructed to: "*Assign an importance score (0.0 to 1.0) to every Prompt Sentence. The scores MUST sum to exactly 1.0.*" To maximize determinism, we utilized a temperature of $T = 0$ or close to it.

Metric Definitions. Let $\mathbf{y}_{\text{oracle}} \in \mathbb{R}^n$ be the vector of ground-truth importance scores provided by an oracle for the n sentences in the prompt. Let $\mathbf{y}_{\text{interp}} \in \mathbb{R}^n$ be the predicted importance scores output by the interpreter.

- **Soft Top-1 Accuracy:** This metric addresses the inherent ambiguity in attribution where multiple prompt sentences may be necessary. We define a match if the interpreter's single highest-scored sentence falls within the top-k sentences identified by the oracle.

  Let $i^* = \text{argmax}_{i \in \{1,\ldots,n\}}(\mathbf{y}_{\text{interp}}^{(i)})$ be the index of the sentence chosen by the interpreter. Let $\mathcal{S}_k(\mathbf{y}_{\text{oracle}})$ be the set of indices corresponding to the k largest values in $\mathbf{y}_{\text{oracle}}$. The metric is defined as:

$$\text{SoftAcc@}k = \mathbb{I}\left[i^* \in \mathcal{S}_k(\mathbf{y}_{\text{oracle}})\right] \tag{F1}$$

  In our experiments, we set $k = 2$.

- **nDCG (Normalized Discounted Cumulative Gain):** Similar to EBM experiments, we utilize nDCG to evaluate the quality of the entire ranking order. This metric penalizes the interpreter if it assigns low importance scores to sentences that the Oracle deemed critical.

  Let π be a permutation of indices $\{1, \ldots, n\}$ that sorts the scores $\mathbf{y}_{\text{interp}}$ in descending order, such that $\mathbf{y}_{\text{interp}}^{(\pi(1))} \geq \mathbf{y}_{\text{interp}}^{(\pi(2))} \geq \ldots$. The DCG is computed using the oracle's scores as the true relevance grades:

$$\text{DCG} = \sum_{j=1}^{n} \frac{\mathbf{y}_{\text{oracle}}^{(\pi(j))}}{\log_2(j+1)} \tag{F2}$$

  The Ideal DCG (IDCG) is computed similarly using the permutation π^* that sorts $\mathbf{y}_{\text{oracle}}$ in descending order. The normalized score is:

$$\text{nDCG} = \frac{\text{DCG}}{\text{IDCG}} \tag{F3}$$

G Interpreter: Generative Faithfulness

Quantifying faithfulness in open-ended generation is fundamentally distinct from classification tasks. Unlike classification, where the output is a discrete label, generative outputs are high-dimensional and semantically flexible. A true causal driver may not reproduce the *exact* tokens of the target, but should reproduce its *semantic* core. To validate our interpreter, we devised a three-stage evaluation pipeline: (1) deriving comparable baselines via dynamic max-ratio thresholding, (2) establishing metric definitions robust to generative variance, and (3) filtering non-causal RLHF artifacts to strictly isolate semantic drivers.

Baseline Implementations. To compare our interpreter's binary selections against the continuous importance scores $s_i \in [0, 1]$ produced by the various baselines, we employed a *Max-Ratio Thresholding* strategy. For a given prompt, a sentence i is selected if its importance score is within a factor of the maximum score assigned to any sentence in that prompt:

$$\text{Select } i \iff s_i \geq 0.5 \cdot \max_j(s_j) \tag{G1}$$

This dynamic thresholding adapts to each model's confidence distribution, ensuring we capture the primary drivers of the generation while discarding marginal contributors. The baselines were constructed as follows:

- **LLM Oracles:** `GPT-4o` and `GPT-4o-Mini` were prompted to output a normalized probability distribution of causal importance across the prompt sentences for each specific target.
- **Sentence-Level LIME:** We adapted LIME (Ribeiro et al., 2016) for generative attribution. To mitigate the prohibitive computational cost of standard word-level LIME, we employed a two-dimensional deduplication strategy. For a prompt of n sentences, we generated a strictly unique set of binary masks $M = \{\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(K)}\}$ (using an exhaustive set if $2^n \leq 50$, otherwise randomly sampled). We queried the target LLM exactly once per mask to generate a shared pool of counterfactual responses $\tilde{\mathbf{y}}^{(k)}$. For a given target sentence $\mathbf{y}_t$, we calculated the cosine similarity $sim_k = \cos(S(\tilde{\mathbf{y}}^{(k)}), S(\mathbf{y}_t))$. We then fitted a Ridge regression to predict sim_k from $\mathbf{m}^{(k)}$, weighted by an exponential distance kernel:

  $$w_k = \exp\left(-\frac{d(\mathbf{m}^{(k)}, \mathbf{1})^2}{\sigma^2}\right) \tag{G2}$$

  where d is the cosine distance and $\sigma = 0.25\sqrt{n}$. The resulting regression coefficients were ReLU-clipped to remove negative causal impacts and $L1$-normalized to yield the final probability distribution $\mathbf{s}$.
- **Sparse Random Baseline:** A naive uniform random assignment (e.g., normalizing independent $U(0, 1)$ values) produces flat distributions ($s_i \approx 1/n$) that trivially fail the max-ratio thresholding. To rigorously simulate the confident sparsity of actual trained attribution models, we sampled the probability vectors from a symmetric Dirichlet distribution:

  $$\mathbf{s} \sim \text{Dir}(\boldsymbol{\alpha}), \quad \text{where } \alpha_i = 0.3 \ \forall i \tag{G3}$$

  This ensures the baseline critically tests the effect of arbitrary sentence assignment without failing due to mechanical flatness.

Metric Definitions. For our evaluation (Table 5), we utilize the following definitions. Let $S(\cdot)$ be the sentence embedding function—specifically implemented using the `all-mpnet-base-v2` model to capture dense semantic representations—and $\mathbf{y}_t$ be the target sentence.

- **Generative Sufficiency ($\mathcal{M}_{\text{suff}}$):** The similarity between the target and generated response using *only* the selected sentences $\mathbf{x}_S$:

  $$\mathcal{M}_{\text{suff}} = \cos(S(\text{LLM}(\mathbf{x}_S)), S(\mathbf{y}_t)) \tag{G4}$$

- **Generative Comprehensiveness ($\mathcal{M}_{\text{comp}}$):** The similarity between the target and the response generated using the complement subset $\mathbf{x} \setminus \mathbf{x}_S$:

  $$\mathcal{M}_{\text{comp}} = \cos(S(\text{LLM}(\mathbf{x} \setminus \mathbf{x}_S)), S(\mathbf{y}_t)) \tag{G5}$$

To instantiate these metrics, we select *Cosine Similarity* over sentence embeddings as our comparison function. This choice is grounded in our robustness analysis (see *Selecting Similarity Function* below), which demonstrates that strict logical entailment metrics (NLI) are overly rigid for validating open-ended generation.

Selecting Similarity Function. We initially attempted to evaluate faithfulness using Natural Language Inference (NLI) models to detect logical entailment between the counterfactual generation and the original target. Specifically, we employed the `cross-encoder/nli-deberta-v3-large` model, treating the generated response as the premise and the target sentence as the hypothesis. We extracted the softmax probability of the "entailment" class to quantify sufficiency and comprehensiveness.

However, as shown in Table 8, NLI metrics proved too rigid for generative tasks.

The NLI Sufficiency scores hovered around $0.09 - 0.15$, implying that even the full correct context rarely entailed the target according to the NLI model. This occurs because NLI models are

Table 8: **Robustness Check: NLI Metrics.** Faithfulness scores computed using *DeBERTa-v3* logical entailment probabilities. In this setup, **Sufficiency** and **Comprehensiveness** measure the probability that the counterfactually generated response logically entails the original target sentence. The uniformly low sufficiency scores (< 0.15) indicate that strict NLI models heavily penalize valid generative paraphrasing. This demonstrates that NLI is overly rigid for open-ended generation tasks, justifying our selection of Cosine Similarity for the primary evaluation in Table 5.

Interpreter	Suff. (↑)	Comp. (↓)	Gap (↑)
LIME	0.112	**0.058**	0.054
ESCI (Ours)	**0.146**	0.083	0.063
GPT-4o-Mini	0.145	0.073	**0.072**
GPT-4o	0.093	0.061	0.033
Random	0.098	0.153	-0.055

trained on premise-hypothesis pairs that require strict logical implication, whereas generative recovery in open-ended tasks relies heavily on semantic paraphrasing. Consequently, we adopted *Cosine Similarity* for our primary evaluation, utilizing the `all-mpnet-base-v2` SentenceTransformer model to compute the pairwise cosine similarity between the dense embeddings of the generations and targets. As detailed in the main text, *Cosine Similarity* yielded sufficiency scores in the ~ 0.4 range, effectively capturing the soft semantic retention characteristic of open-ended generation.

Filtering RLHF Priors (Trivial Targets). A major confounder in interpreting instruction-tuned models is the prevalence of conversational fillers (e.g., "Okay!"). These outputs are often driven by Reinforcement Learning from Human Feedback (RLHF) priors rather than specific prompt content. If included, they artificially inflate comprehensiveness scores (lower is better), as the model will often hallucinate these polite preambles even when the causal instruction is removed.

We conducted a trivial target analysis on a subset of conversational fillers ($n = 157$) identified via regex matching to examine how models handle these RLHF priors. As shown in Table 9, the metrics expose that these fillers lack specific causal antecedents.

For example, LIME severely struggles with *Trivial Sufficiency* (performing worse than Random), indicating that it removes too many sentences and fails to find a sufficient prompt subset to reliably trigger the filler. Conversely, while ESCI achieves a much higher *Trivial Sufficiency*, it suffers from an inflated *Trivial Comprehensiveness*. This high hallucination rate in the complement set is a direct byproduct of ESCI's extreme sparsity; because ESCI aggressively isolates only the core logical sentences, the remaining complement set (used for comprehensiveness) remains larger and retains enough generic conversational context to independently trigger the LLM's polite RLHF priors. Ultimately, because these conversational fillers are global stylistic habits rather than logical consequences of specific prompt sentences, no attribution method can meaningfully isolate them.

Table 9: **Trivial Target Analysis.** We measure how often conversational fillers (e.g., "Okay!") persist under intervention. **Triv. Suff:** Percentage of times the filler is generated given *only* the instruction. **Triv. Comp:** Percentage of times the filler is hallucinated when the instruction is *removed*. Values confirm these are RLHF priors, not causally sensitive targets.

Interpreter	Triv. Suff. (↑)	Triv. Comp. (↓)
LIME	15.2%	**1.8%**
ESCI (Ours)	**32.5%**	40.8%
GPT-4o-Mini	15.5%	16.4%
GPT-4o	21.6%	5.2%
Random	15.6%	49.4%

To prevent these non-causal hallucinations from skewing the semantic evaluation, we strictly filtered these targets from the main benchmark.

H The LLM Usage

Parts of the initial drafts of this manuscript were revised with the assistance of a Large Language Model. The model was prompted to improve the fluency, conciseness, and overall academic tone of the text to meet the standards of ACL publications.