Interpreting Black-Box Large Language Models with Concept-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 strictly 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 consistency 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 output without requiring further API queries to the LLM. By globally training a local interpreter across diverse inputs, our framework captures broader generation patterns and mitigates instance-specific biases. 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.

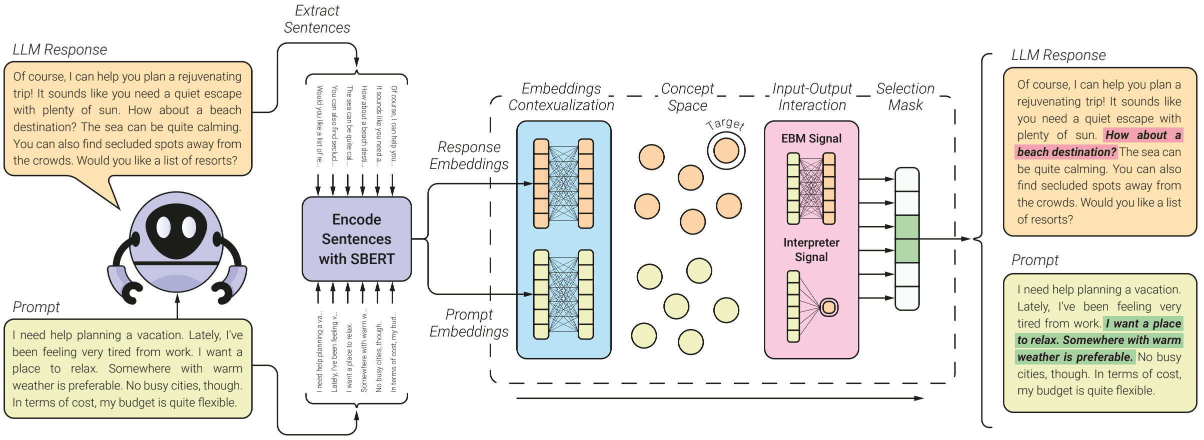
1 Introduction

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, as experts cannot verify the generated output against domain knowledge or detect hidden biases. 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 identifying an importance vector for input features. Essentially, they measure how much each input feature influences the output within a local neighborhood. However, standard attribution techniques, including white-box and model-agnostic, 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 "concepts." We define a concept as a sentence, which is the smallest unit of language that expresses a complete thought. Our goal is to relate elements of



LLM Response

Of course, I can help you plan a rejuvenating trip! It sounds like you need a quiet escape with plenty of sun. How about a beach destination? The sea can be quite calming. You can also find secluded spots away from the crowds. Would you like a list of resorts?

Prompt

I need help planning a vacation. Lately, I've been feeling very tired from work. I want a place to relax. Somewhere with warm weather is preferable. No busy cities, though. In terms of cost, my budget is quite flexible.

LLM Response

Of course, I can help you plan a rejuvenating trip! It sounds like you need a quiet escape with plenty of sun. ***How about a beach destination?*** The sea can be quite calming. You can also find secluded spots away from the crowds. Would you like a list of resorts?

Prompt

I need help planning a vacation. Lately, I've been feeling very tired from work. ***I want a place to relax. Somewhere with warm weather is preferable.*** No busy cities, though. In terms of cost, my budget is quite flexible.

Extract Sentences
Of course, I can help you...
It sounds like you need a...
How about a beach desti...
The sea can be quite cal...
You can also find seclud...
Would you like a list of re...
Encode Sentences with SBERT
I need help planning a va...
Lately, I've been feeling v...
I want a place to relax.
Somewhere with warm w...
No busy cities, though.
In terms of cost, my bud...
Response Embeddings
Prompt Embeddings
Embeddings Contexualization
Concept Space
Target
Input-Output Interaction
EBM Signal
Interpreter Signal
Selection Mask

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.

the output directly to the user prompt at this concept level. 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 and mitigate intrinsic biases.

Figure 1 illustrates an overview of the approach. We first train a transformer-based Energy-Based Model (EBM) to act as a surrogate for the black-box LLM. It maps the prompt and LLM response sentences to a latent "concept space," which simulates the concept-level relationships embedded in the target LLM, and calculates an energy score. This score quantifies the conceptual consistency between the input and output. We then use this energy landscape to guide the training of an interpreter network. Given a prompt and a target output subset, the interpreter produces an importance vector that isolates the prompt sentences most influential in generating that response. In summary, we:

1. Shift the unit of analysis from tokens to sentences. This enables attribution at the level of semantically coherent concepts, making explanations more intelligible to humans.

2. Introduce a transformer-based EBM with novel sampling methods capable of learning a random field over prompts and responses. This model effectively learns authentic input-output dynamics, serving as a robust surrogate for the black-box model.

3. Propose a post-hoc, model-agnostic framework for interpreting black-box LLMs at a conceptual level. This interpreter finds the specific prompt sentences responsible for triggering a target subset of the LLM's output.

2 Related Work

2.1 Post-hoc Attribution Methods

Attribution methods score the importance of input features for a specific model output. White-box approaches to attribution often utilize gradients, propagating output salient signals back to input tokens (Simonyan et al., 2014; Sundararajan et al., 2017; Shrikumar et al., 2017; Chefer et al., 2021). Others use internal attention weights as proxies for feature importance (Xu et al., 2015; Li et al., 2017; Xie et al., 2017; Hao et al., 2021). However, gradients are inaccessible for proprietary APIs, and attention weights are frequently unfaithful to the reasoning process (Jain and Wallace, 2019). Similarly, influence functions pose data-centric explanations (Koh and Liang, 2017) but remain infeasible without access to the base data or model's Hessian.

Perturbation-based methods offer a model-agnostic alternative; they measure output changes when input segments are removed or altered (Ribeiro et al., 2016; Lundberg and Lee, 2017; Yin and Neubig, 2022). Hackmann et al. (2024) apply this to identify influential words in LLM prompts. However, this approach scales poorly for generative tasks. Validating a single explanation often requires thousands of model queries, incurring high computational costs (Enouen et al., 2024; Zhao and Shan, 2024).

Finally, prompt-based self-explanation leverages the LLM's own generation capabilities. Most notably, techniques like Chain-of-Thought (CoT) ask the model to produce a rationale to justify its output (Wei et al., 2022). While compelling, these explanations lack guarantees of faithfulness; they often represent plausible post-hoc rationalizations rather than the true internal computation path (Turpin et al., 2023).

2.2 Energy-Based Models in NLP

Energy-Based Models (EBMs) have been successfully adapted for generative Natural Language Processing (NLP), primarily through learning global scoring functions. Bakhtin et al. (2019) demonstrated that Transformer-based discriminators can function as EBMs; by distinguishing between human and machine text, these models assign low energy to coherent sequences. This established the potential of EBMs as holistic text evaluators.

Additionally, several paradigms utilize EBMs to refine existing model outputs. The Residual EBM approach adds a corrective energy term to the log-probabilities of a base autoregressive model (Deng et al., 2020; Bakhtin et al., 2021). This allows the system to capture high-level properties, such as coherence, that the base model may miss. Alternatively, Tu et al. (2020) use a powerful autoregressive teacher to define an energy landscape; a student network is then trained via knowledge distillation to generate outputs that minimize this energy. EBMs can also act as post-processing rerankers to select the highest-quality result from a set of candidates (Bhattacharyya et al., 2021).

2.3 Concept-Based Explanations

Interpretability research is increasingly shifting away from granular token-level attributions and toward concept-based explanations. These methods map decisions to human-intelligible ideas rather than individual features (Kim et al., 2018). Our work aligns with this paradigm by defining a "sentence" as the fundamental conceptual unit, as it represents a robust thought for interpretation.

Treating sentences as semantic objects is well-justified by the history of language modeling. Foundational architectures like BERT used Next Sentence Prediction to learn logical relationships (Devlin et al., 2019). Subsequent work on Sentence-BERT confirmed that fine-tuned sentence representations map similar meanings to distinct, nearby points in a vector space (Reimers and Gurevych, 2019). By leveraging sentences as concepts, our work parallels recent architectural innovations such as Large Concept Models, which shift the core computational unit from tokens to sentence-level representations (Barrault et al., 2024).

3 Methodology

Our goal is to develop a post-hoc, model-agnostic method for interpreting black-box LLMs, specifically by identifying which input sentences drive the response. We depart from standard token-level attribution by establishing the sentence as the fundamental unit of analysis. As the smallest linguistic unit expressing a complete proposition, the sentence serves as a robust "concept," enabling us to interpret generation as an interplay of complete 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 seek to 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 energy values representing the consistency of a prompt-response pair $(\mathbf{x}, \mathbf{y})$ with the target LLM's generation patterns. Second, we 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 concepts. We first perform sentence segmentation using the spaCy library (Honnibal and Montani, 2017).

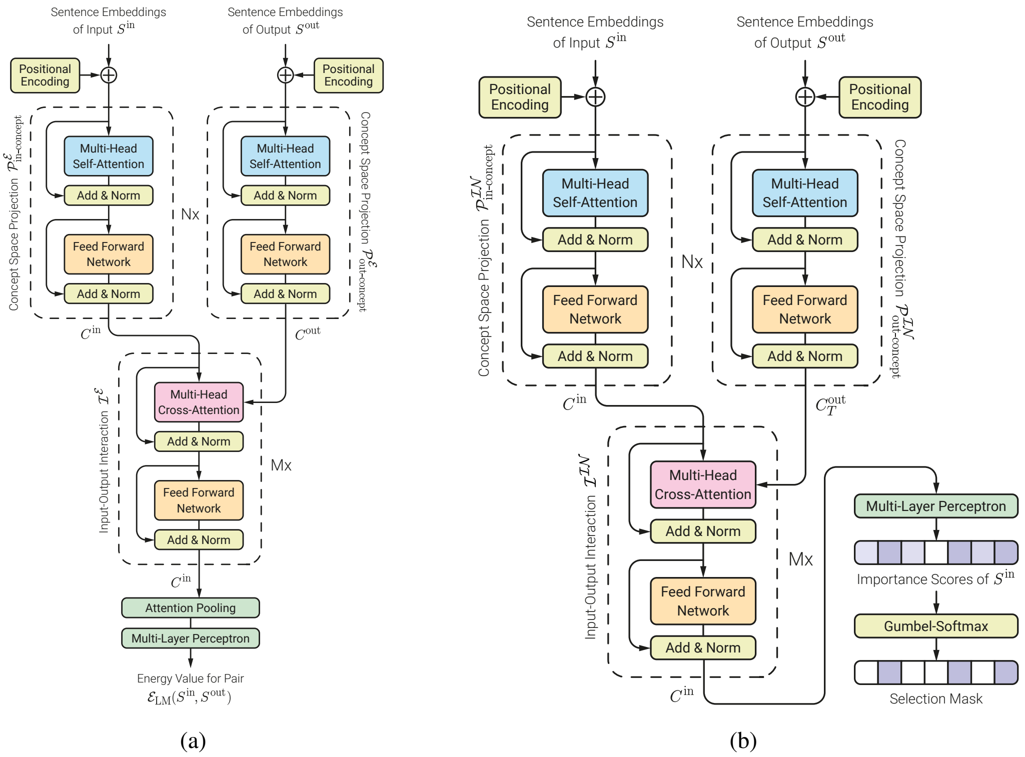

Sentence Embeddings of Input S^{in}
Sentence Embeddings of Output S^{out}
Positional Encoding
Positional Encoding
Concept Space Projection $\mathcal{P}^{\mathcal{E}}_{\text{in-concept}}$
Multi-Head Self-Attention
Add & Norm
Feed Forward Network
Add & Norm
Nx
Multi-Head Self-Attention
Add & Norm
Feed Forward Network
Add & Norm
Concept Space Projection $\mathcal{P}^{\mathcal{E}}_{\text{out-concept}}$
C^{in}
C^{out}
Input-Output Interaction $\mathcal{I}^{\mathcal{E}}$
Multi-Head Cross-Attention
Add & Norm
Feed Forward Network
Add & Norm
Mx
C^{in}
Attention Pooling
Multi-Layer Perceptron
Energy Value for Pair $\mathcal{E}_{\text{LM}}(S^{\text{in}}, S^{\text{out}})$
(a)
Sentence Embeddings of Input S^{in}
Sentence Embeddings of Output S^{out}
Positional Encoding
Positional Encoding
Concept Space Projection $\mathcal{P}^{\mathcal{IN}}_{\text{in-concept}}$
Multi-Head Self-Attention
Add & Norm
Feed Forward Network
Add & Norm
Nx
Multi-Head Self-Attention
Add & Norm
Feed Forward Network
Add & Norm
Concept Space Projection $\mathcal{P}^{\mathcal{IN}}_{\text{out-concept}}$
C^{in}
C^{out}_T
Input-Output Interaction $\mathcal{I}^{\mathcal{IN}}$
Multi-Head Cross-Attention
Add & Norm
Feed Forward Network
Add & Norm
Mx
C^{in}
Multi-Layer Perceptron
Importance Scores of S^{in}
Gumbel-Softmax
Selection Mask
(b)

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

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 that learns to distinguish variation 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:

1. **Concept Space Projection:** Static embeddings (S^{in}, S^{out}) capture meaning in isolation, but lack the specific 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^{in}, C^{out}). Here, distances reflect the LLM's internal dependency structure rather than generic semantic similarity. This is a function of the model's underlying architecture and training dataset.

2. **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.

3. **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 novel set of objectives, then subsequently 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. 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 interactions using two batch-wise samplers. The first, $(x_{\text{part}}, y'_{\text{part}})$, distinguishes the correct partial response from off-topic partial responses for the InfoNCE loss $\mathcal{L}_{\text{resp-dep}}$. The second, $(x'_{\text{part}}, y_{\text{part}})$, mirrors this for partial prompts against negatives in the InfoNCE loss $\mathcal{L}_{\text{pmt-dep}}$. This compels the model to verify input-output dependencies piece-by-piece rather than relying on general heuristics.

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—ranging from human baselines to the partial-sequence batching strategies—and the full training hyperparameters in Appendix B. This pre-training phase (Fig. 5) yields a globally-aware energy function $\mathcal{E}_{\text{LM}}$, which captures the target LLM's latent structure 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 to $\mathbf{y}_T$. 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^{out}_T 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. A successful selection minimizes the energy of the authentic pair $(\tilde{\mathbf{x}}, \mathbf{y}_T)$ and maximizes the energy of the irrelevant remainder $(\mathbf{x} - \tilde{\mathbf{x}}, \mathbf{y}_T)$. The interpreter optimization is thus:

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

However, masking inputs inherently causes distribution shifts (Hsia et al., 2024). To prevent this, we fine-tune the EBM alongside the interpreter via periodic alternating optimization (Fig. 6, App. D.1). First, we update the interpreter parameters (Eq. 3) given the current EBM. Second, we periodically query the target LLM with the selected prompt subset $\tilde{\mathbf{x}}$ to generate a consistent response $\tilde{\mathbf{y}}$. Third, we update the EBM using this fresh pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ and the loss described in Section 3.2. While this incurs API costs, our experiments suggest it is optional for standard benchmarks yet beneficial for robust, large-scale deployments. This process transfers the EBM's structure to the interpreter, enabling standalone inference with zero additional queries.

4 Experiments

We empirically validate our framework by first establishing the pre-trained EBM's efficacy as a faithful surrogate model. We present a representative ablation study to demonstrate its capacity to capture the target LLM's latent semantic relations. Subsequently, we evaluate the interpreter trained on these energy landscapes across two complementary dimensions: semantic plausibility and causal faithfulness. We emphasize that our framework is designed to train *task-oriented* interpreters where each interpreter is specialized for a specific type of task (e.g. general Q&A). This focused scope allows for effective training even with limited data.

4.1 Validating the Dual-Objective EBM

To justify our dual-objective framework, we compare three EBM designs selected from our broader experiments. We argue that a robust surrogate requires two complementary properties: *fidelity*, to capture the target LLM's global distribution, and *local dependency*, to enforce local causal precision.

Experimental Setup. Models were trained on the HC3 dataset, a multi-domain Q&A corpus containing both human and model-generated text. We employed `GPT-4o-Mini` as the target and `GPT-2-Medium` as the contrastive baseline. We utilized compact 181M parameter models ($\sim$71M trainable) on a subset of $20,000$ samples for efficiency. Detailed configurations are in Appendix B. Preliminary scaling experiments demonstrated that larger models achieved wider energy gaps and faster convergence, which correlated with higher accuracy in distinguishing authentic pairs. This indicates the framework's scalability despite the

Table 1: **EBM Ablation Study Results.** We evaluate the models across four dimensions of interpretability. $\mathcal{M}_{\textbf{Fidelity}}$ 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.

	I. Core Directive		**II. Robustness**		**III. SOTA Alignment**		**IV. Causal Disentanglement**	
	(DeYoung et al., 2020)		(Gururangan et al., 2018)		(`Gemini-2.5-Flash`)		(`Gemini-3-Pro`)	
Model	**IR@1** $\uparrow$	**SNR** $\uparrow$	**Art.** $\Delta E \downarrow$	**R1E** $\downarrow$	**nDCG@3** $\uparrow$	**Prec@1** $\uparrow$	**Acc** $\uparrow$	**ESM** $\uparrow$
$\mathcal{M}_{\text{Fidelity}}$	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**
$\mathcal{M}_{\textbf{Hybrid}}$	**84.77%**	**7.15**	+0.104	15.17%	**0.796**	**81.0%**	**92.40%**	0.382

limited scale of this study. We focus our analysis on the following distinct configurations:

- $\mathcal{M}_{\textbf{Fidelity}}$ **(Baseline):** Trained solely on structural *fidelity* ($\lambda = 0$), mimicking standard likelihood modeling.
- $\mathcal{M}_{\textbf{Dep}}$ **(Ablation):** Trained solely on the *local dependency* objective ($\lambda = 1$) to identify semantic links across partial segments without overfitting to surface artifacts.
- $\mathcal{M}_{\textbf{Hybrid}}$ **(Ours):** A dual-objective model ($\lambda = 0.9$) using dependency samplers for structural dependencies and fidelity signals to regularize the latent space.

Evaluations across four semantic alignments (Tab. 1) follow protocols in Appendix E.

Dimension I: Core Directive Localization. We assess the ability to localize primary intent (e.g., the question) amidst background context using *Interrogative Recall* (IR@1) and *Signal-to-Noise Ratio* (SNR) (DeYoung et al., 2020). We note that perfect recall is not expected, as human prompts often contain implicit or structurally ambiguous directives. As shown in Table 1, $\mathcal{M}_{\text{Fidelity}}$ exhibits diffuse attention sensitive to background noise, while $\mathcal{M}_{\text{Hybrid}}$ achieves a 5× SNR improvement, confirming that the *local dependency* objective compels the model to prioritize semantic directives.

Dimension II: Semantic Robustness. We measure the neural models' prevalent sensitivity to "annotation artifacts" (Gururangan et al., 2018) using *Artifact Energy Impact* and *Rank-1 Error* (R1E). $\mathcal{M}_{\text{Fidelity}}$ suffers from a fidelity trap, overprioritizing conversational 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 regularization necessary for training.

Dimension III: Alignment with SOTA Oracles. Using `Gemini-2.5-Flash` as a reference oracle, we evaluate ranking quality via nDCG@3 (Järvelin and Kekäläinen, 2002) and *Soft Precision@1*. $\mathcal{M}_{\text{Hybrid}}$ achieves the highest alignment, indicating that the hybrid objective produces concept rankings most consistent with the generation patterns of state-of-the-art foundation models.

Dimension IV: Causal Disentanglement. We use counterfactual triplets $(y_{\text{target}}, x_{\text{cause}}, x_{\text{distractor}})$ generated by `Gemini-3-Pro` following protocols from Kaushik et al. (2020). While $\mathcal{M}_{\text{Dep}}$ produces the sharpest energy landscape, its hyper-discrimination reduces overall accuracy. $\mathcal{M}_{\text{Hybrid}}$ achieves the highest accuracy, successfully balancing discriminative confidence with the robustness required to filter spurious correlations.

4.2 Interpreter Semantic Plausibility Analysis

In the absence of ground-truth labels for concept importance in open-ended generation, we evaluate the *plausibility* of our method by measuring its alignment with the attribution judgments of powerful LLMs. We treat these LLMs as "oracles," assuming that consensus among diverse high-capacity models serves as a reliable proxy. For this evaluation, we utilize our most robust EBM configuration (see App. D for detailed architecture).

Experimental Setup. We interpret the outputs of `GPT-4o-Mini` on a subset of 200 prompt-response pairs ($2,000$ combinations) from the HC3-based dataset. For every sentence in the response, we task distinct LLMs (`Gemini-2.5-Flash`, `GPT-4o`, `GPT-4o-Mini`, `GPT-J-6B`, and `GPT-2-XL`) to score the sentences in the original prompt based on their contribution to generating that target. We then compare the rankings produced by our Energy-Based Concept-Level Surrogate Interpreter (ESCI) against these oracles. We report results on two scenarios: *Scenario A* (averaging across all target

(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**

Figure 3: **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 target LLM. **Soft Top-1** (left) measures if the Interpreter's top choice appears in the Oracle's top-2. **nDCG** (right) measures ranking correlation. Our proposed ESCI model shows remarkable alignment with GPT-J and GPT-2, suggesting it captures the probabilistic dependencies of standard causal language modeling effectively despite its small size.

sentences) and *Scenario B* (focusing on the final sentence, which often contains the conclusion).

Quantitative Assessment. Figure 3 presents the pairwise alignment between our interpreter and the five oracles. We employ *nDCG* to measure ranking quality and *Soft Top-1 Accuracy* to measure top choice alignment while accounting for ambiguity in text attribution (see App. F for details).

Despite having orders of magnitude fewer parameters (∼71M trainable, ∼110M loaded), ESCI achieves competitive plausibility against vastly different oracle architectures. In Scenario B (Fig. 3c/d), ESCI aligns closely with both `GPT-J-6B`, a standard causal language model, and `Gemini-2.5-Flash`, a heavily instruction-tuned system. This suggests that our energy landscape effectively internalizes the causal mechanics of autoregressive generation.

We observe that ESCI yields sparse, confident scores, contrasting with the diffuse distributions of instruction-tuned oracles. Additionally, unlike white-box models—which often struggle to produce valid probability distributions and require post-processing intervention—ESCI operates as a robust, standalone tool, while matching them in attribution strength. It sharply isolates *necessary* dependencies, minimizing the ambiguity typical of generative baselines.

Qualitative Case Studies. Figure 4 provides a granular look at specific attribution behaviors. In cases of clear semantic mapping (21 and 142), ESCI aligns perfectly with oracles. Disagreements, however, are revealing. In Sample 33, ESCI attributes the simplified output to the topic keyword, whereas oracles point to the specific user question. This suggests ESCI may sometimes over-prioritize the global topic over query nuances. Conversely, Sample 6 highlights ESCI's strength: the target sentence is a pure stylistic simplification. ESCI correctly identifies the instruction "*Explain like I'm five*" as the cause, whereas oracles fixate on the semantic content. This confirms ESCI's ability to disentangle stylistic drivers from semantic ones, a critical capability for interpreting instruction-tuned models. Finally, Sample 94 illustrates total chaos, where the target sentence is an analogy synthesizing multiple parts of the prompt, leading to valid but divergent interpretations across all models.

4.3 Evaluating Interpreter Faithfulness

While plausibility confirms that our interpreter aligns with human and oracle judgments, it does not guarantee that the selected concepts are the true drivers of the LLM's generation. To assess this, we evaluate *Causal Faithfulness* by measuring the generative consequences of intervening on the input prompt. We adapt the standard metrics of *Sufficiency* and *Comprehensiveness* (DeYoung et al., 2020), modifying them for open-ended generation by using semantic similarity rather than classification probability (Atanasova et al., 2023).

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! …

Figure 4: **Qualitative Comparison of Attribution Scores. Left:** Prompt snippets. **Middle:** Attribution scores from ESCI and oracles (Top scores in **bold**, near-zero scores grayed). **Right:** Target sentence from the LLM response. **ID 6**: ESCI correctly attributes the simple tone to the style instruction ("Explain like I'm five"), while oracles focus on the subject. **ID 33**: ESCI over-focuses on the topic keyword ("fuel octane"), while oracles correctly identify the specific question. **ID 94**: A case of chaos where the target response is a broad analogy, leading all models to diverge.

We define **Generative Sufficiency** as the degree to which the target output can be regenerated using *only* the selected prompt sentences. Conversely, **Generative Comprehensiveness** measures the extent to which the target concept is lost when the selected sentences are *removed* from the prompt. A faithful interpreter should maximize the former and minimize the latter, creating a positive *Faithfulness Gap*. We quantify this using cosine similarity between the embeddings of the original target sentence and the generated counterfactuals.

Table 2 compares our ESCI against oracles derived from `GPT-4o` and `GPT-4o-Mini`; for each oracle sentences are selected based on the LLM's importance rankings using a max-ratio thresholding strategy (see App. G for details).

Table 2: **Causal Faithfulness Evaluation.** We measure the semantic similarity of the LLM's response to the target sentence under strict interventions. **Sufficiency:** Prompting with *only* selected sentences. **Comprehensiveness:** Prompting with *everything but* selected sentences. **Gap:** The net causal contribution

Interpreter	Suff. (↑)	Comp. (↓)	Gap (↑)
ESCI (Ours)	**0.409**	**0.214**	**0.195**
GPT-4o (Oracle)	**0.409**	0.235	0.175
GPT-4o-Mini (Oracle)	0.407	**0.215**	0.192

Analysis. Our method achieves *Sufficiency* parity with the `GPT-4o` oracle and slightly outperforms the `GPT-4o-Mini`'s self-explanation. Critically, ESCI also achieves the lowest *Comprehensiveness* score, suggesting that our isolated subsets are not merely relevant, but are necessary causal antecedents for regenerating the target output. This is a significant result given that our interpreter operates as a lightweight surrogate with orders of magnitude fewer parameters than the baselines. We note, however, that no model achieves near-zero *Comprehensiveness*; the consistent residual similarity (~ 0.21) reflects the inherent ambiguity of open-ended generation, where LLMs can often reconstruct semantic content from background knowledge or redundancy. Within this noisy regime, ESCI's ability to match the causal fidelity profile of state-of-the-art oracles confirms its reliability as a computationally efficient, standalone attribution tool.

5 Conclusion

In this work, we introduced a concept-level interpreter for black-box LLMs that shifts post-hoc attribution from noisy tokens to coherent sentences. By distilling latent dynamics into a differentiable energy landscape, we trained a standalone interpreter operating with zero inference-time API costs. Our ablation studies confirm that constraining the surrogate with both of our novel *fidelity* and *local dependency* objectives is necessary for this task. Empirically, ESCI achieves generative sufficiency parity with `GPT-4o` while demonstrating slightly superior causal isolation, disentangling necessary antecedents from stylistic priors. This establishes energy-based surrogates as a scalable pathway for diagnosing model behaviors without full access.

6 Limitations and Future Work

Computational Trade-offs. A primary limitation of our framework is the computational overhead of the pre-training phase. Unlike perturbation-based methods (e.g., LIME) which are expensive at *inference* time, our approach shifts this burden to *training*. While this incurs a one-time cost, it is notably modest compared to LLM pre-training; our experiments required less than 30 hours on free-tier GPUs (see App. B and D). Crucially, this investment yields a standalone interpreter capable of $O(1)$ inference with zero additional API queries, making our method suitable for high-volume, real-time analysis, though less accessible for users unable to perform the initial pre-training.

Generalizability and Scaling Limits. Due to resource constraints, our validation focused on standard Q&A tasks using compact surrogate models ($\sim$181M parameters) to interpret `GPT-4o-Mini`. Future work should expand to diverse domains, including complex reasoning and open-ended generation (e.g., *TellMeWhy*, *WikiText*), and target distinct architectures beyond the GPT family. Additionally, while our scaling experiments suggest improved performance with larger EBMs, the efficacy of the energy landscape in capturing long-range dependencies within massive context windows (e.g., $128k+$ tokens) remains to be verified. 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 sufficiency and comprehensiveness quantify causal faithfulness, they remain automated proxies. A key limitation is the assumption that causal accuracy automatically equates to human intelligibility. Although we provide preliminary qualitative analysis, validating the practical utility of ESCI requires rigorous subject studies. Such evaluations are critical to corroborate our plausibility findings against human judgment—providing a grounded check on probabilistic oracles—and to assess downstream usability. Specifically, we aim to measure whether these explanations effectively aid users in high-stakes auditing tasks, such as detecting hallucinations, identifying bias, and verifying safety compliance.

Scope of Application and Optimization. Our current evaluation is confined to the diagnostic utility of attribution, leaving the framework's broader potential for downstream optimization empirically unverified. Theoretically, the identification of *necessary* sentences enables prompt optimization—automatically pruning irrelevant context to reduce token costs without degrading output quality. Similarly, the energy landscape offers a mechanism to audit Chain-of-Thought (CoT) reasoning, potentially filtering unfaithful or confabulated intermediate steps. However, given the complexity of these domains, further experimentation is strictly necessary to determine if the interpreter's performance can be sufficiently optimized to maintain robustness when deployed on such 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.

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.

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.

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.

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.

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: EACL 2024*, 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.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

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

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.

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.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (Workshop Track)*.

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.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057. PMLR.

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.

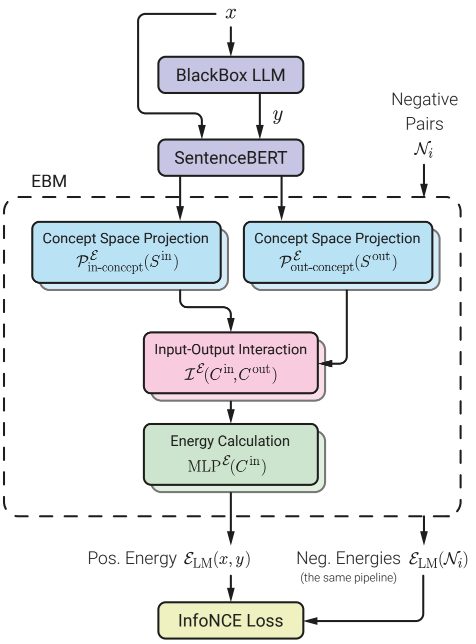
Zhixue Zhao and Boxuan Shan. 2024. ReAGent: A model-agnostic feature attribution method for generative language models. In *Proceedings of the AAAI 2024 Workshop on Responsible Language Models (ReLM)*. Association for the Advancement of Artificial Intelligence.

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 $S^{\text{in}}, S^{\text{out}} \in \mathbb{R}^{N_{\max} \times d}$, where d is the embedding dimension of the Sentence-BERT model (768 for `all-mpnet-base-v2`).

B Energy Network: Pre-training Details



x
BlackBox LLM
y
Negative
Pairs
$\mathcal{N}_i$
SentenceBERT
EBM
Concept Space Projection
$\mathcal{P}^{\mathcal{E}}_{\text{in-concept}}(S^{\text{in}})$
Concept Space Projection
$\mathcal{P}^{\mathcal{E}}_{\text{out-concept}}(S^{\text{out}})$
Input-Output Interaction
$\mathcal{I}^{\mathcal{E}}(C^{\text{in}},C^{\text{out}})$
Energy Calculation
$\text{MLP}^{\mathcal{E}}(C^{\text{in}})$
Pos. Energy $\mathcal{E}_{\text{LM}}(x, y)$
Neg. Energies $\mathcal{E}_{\text{LM}}(\mathcal{N}_i)$
(the same pipeline)
InfoNCE Loss

Figure 5: **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 pipleline is illustrated in Figure 5. 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 3 details the specific hyperparameters used for the final Hybrid model.

Table 3: **Hyperparameters for the $\mathcal{M}_{\textbf{Hybrid}}$ 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.

- $\mathcal{M}_{\textbf{Dep}}$ **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.

- $\mathcal{M}_{\textbf{Hybrid}}$ **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 a discrete, non-differentiable operation, we employ the Gumbel-Softmax relaxation 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:

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

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:

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

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:

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

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 4 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. 6), 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:

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

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

Table 4: **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)

Step 2: Periodic Grounding. 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.

Step 3: EBM Fine-tuning. 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 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.

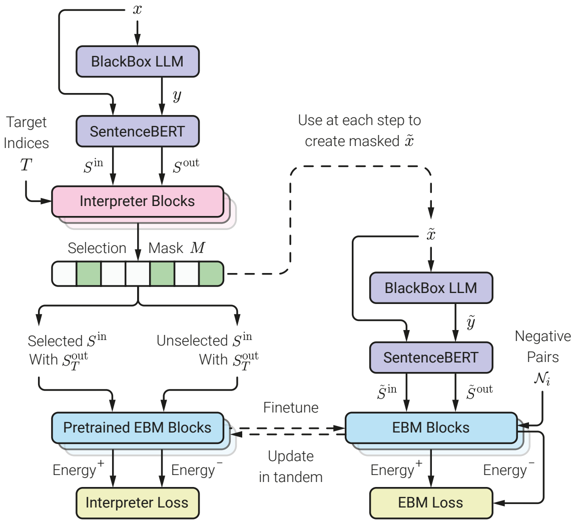


Figure 6: **Overview of the Alternating Optimization Protocol.** The framework employs a joint training strategy to prevent distribution shift. **(Left)** In the standard phase, the interpreter generates a binary mask over the prompt sentences; its parameters are updated to maximize the energy gap using 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.

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: Core Directive Localization

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_{NC} be the set of indices for non-causal 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.

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

  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.

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

  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: 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.

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

- **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: Oracle Alignment

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 oracle 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.

Oracle Baseline Construction. To compare our interpreter's binary selections against the continuous importance scores $s_i \in [0, 1]$ produced by the oracle LLMs, 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 the model's confidence distribution, ensuring we capture the primary drivers of the generation while discarding marginal contributors.

Metric Definitions. For our evaluation (Table 2), we utilize the following definitions. Let $S(\cdot)$ be the embedding function and $\mathbf{y}_t$ be the target.

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

$$\mathcal{M}_{\text{suff}} = \cos(S(\text{LLM}(\mathbf{x}_S)), S(\mathbf{y}_t)) \quad \text{(G2)}$$

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

$$\mathcal{M}_{\text{comp}} = \cos(S(\text{LLM}(\mathbf{x} \setminus \mathbf{x}_S)), S(\mathbf{y}_t)) \quad \text{(G3)}$$

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 (specifically `deberta-v3-large`) to detect logical entailment between the counterfactual generation and the original target. However, as shown in Table 5, NLI metrics proved too rigid for generative tasks.

Table 5: **Robustness Check: NLI Metrics.** Faithfulness scores computed using *DeBERTa-v3 Entailment* probabilities. The extremely low sufficiency scores (< 0.15) indicate that NLI penalizes valid semantic paraphrases, making it unsuitable for evaluating open-ended generation.

Interpreter	Suff. (↑)	Comp. (↓)	Gap (↑)
ESCI (Ours)	**0.146**	0.083	0.063
GPT-4o	0.093	**0.061**	0.033
GPT-4o-Mini	0.145	0.073	**0.072**

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 is because NLI models are trained on premise-hypothesis pairs that require strict logical implication, whereas generative recovery often involves paraphrasing. Consequently, we adopted *Cosine Similarity* (using `all-mpnet-base-v2`) for our primary evaluation. As detailed in the main text, *Cosine Similarity* yielded sufficiency scores in the ~ 0.41 range, 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!", "Sure, here is the answer"). 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. As shown in Table 6, the target LLM frequently regenerates these targets even under strict counterfactual conditions. Notably, while ESCI achieves the highest *Trivial Sufficiency* (indicating strong causal isolation), it yields a *Trivial Comprehensiveness* of 40.8%. This seemingly high hallucination rate is a byproduct of our method's sparsity; because ESCI selects significantly fewer sentences than the baselines, the complement set (used for comprehensiveness) remains larger and more likely to contain residual context that triggers the model's strong RLHF priors.

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

Table 6: **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. (↓)
ESCI (Ours)	**32.5%**	40.8%
GPT-4o	21.6%	**5.2%**
GPT-4o-Mini	15.5%	16.4%

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.