A CONCEPT LEVEL ENERGY-BASED FRAMEWORK FOR INTERPRETING BLACK-BOX LARGE LANGUAGE MODEL RESPONSES

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-access APIs has created a critical challenge for their reliable deployment: a fundamental lack of interpretability. In this work, we propose a model-agnostic, post-hoc interpretation framework to address this. Our approach defines an energy model that quantifies the conceptual consistency between prompts and the corresponding LLM-generated responses. We use this energy to guide the training of an interpreter network for a set of target sentences. Once trained, our interpreter operates as an efficient, standalone tool, providing sentence-level importance scores without requiring further queries to the original LLM API or energy model. These scores quantify how much each prompt sentence influences the generation of specific target sentences. A key advantage is that our framework globally trains a local interpreter, which helps mitigate common biases in LLMs. Our experiments demonstrate that the energy network accurately captures the target LLM's generation patterns. Furthermore, we show that our interpreter effectively identifies the most influential prompt sentences for any given output.

1 INTRODUCTION

In recent years, the extraordinary performance of Large Language Models in complex tasks has encouraged machine learning researchers and developers to adopt them in different applications. Powerful LLMs are mostly provided as APIs by developer companies, and the detailed architectures and pre-training datasets are often unavailable. Additionally, when the architecture and dataset are available, their complexity prevents an exact understanding of how outputs are generated. In high-stakes domains such as medicine and law, this opacity prevents human experts from verifying a model's generation against domain knowledge and discovering hidden biases, thereby hindering its ability to satisfy the application-grounded evaluation criteria necessary for responsible deployment (Doshi-Velez & Kim, 2017).

Post-hoc attribution interpretation attempts to explain the behavior of trained machine learning models by finding an importance vector for input features. These methods measure how much each input feature affects the value of the output in each locality of the input space. White-box techniques that rely on gradients or internal activations are immediately disqualified for most real-world scenarios due to the lack of model access. Moreover, the faithfulness of popular proxies like attention weights has been rigorously challenged (Jain & Wallace, 2019). Various techniques have also been introduced for attribution-based interpretation of black-box models (Ribeiro et al., 2016; Lundberg & Lee, 2017; Seyyedsalehi et al., 2022). Most of these methods are developed to explain discriminative models with well-defined vectorized outputs. The problem of interpretation for generative models is fundamentally ill-posed. Generative models learn complex and implicit representations to produce high-dimensional and multifaceted outputs like text. Therefore, explaining them requires grappling with the complexity of their interactive outputs and the sheer volume of information in each generation (Schneider, 2024)

The exact goal of an interpretation method for an LLM is not well-defined and depends on the application. However, an attribution method to interpret an LLM is expected to relate elements of the output text to the user prompt. Here, we consider both input and output texts at a sentence-level resolution. We define the sentence as the fundamental "concept," as it represents the smallest unit of language that expresses a complete thought, proposition, or idea, and use the terms "sentence-level" and "concept-level" interchangeably to distinguish our approach from granular token-level attributions. As the target output, we focus on a subset of output sentences and attempt to find an importance vector for the prompt sentences which shows how much each of the prompt sentences influenced the LLM to generate

[*] Equal contribution.

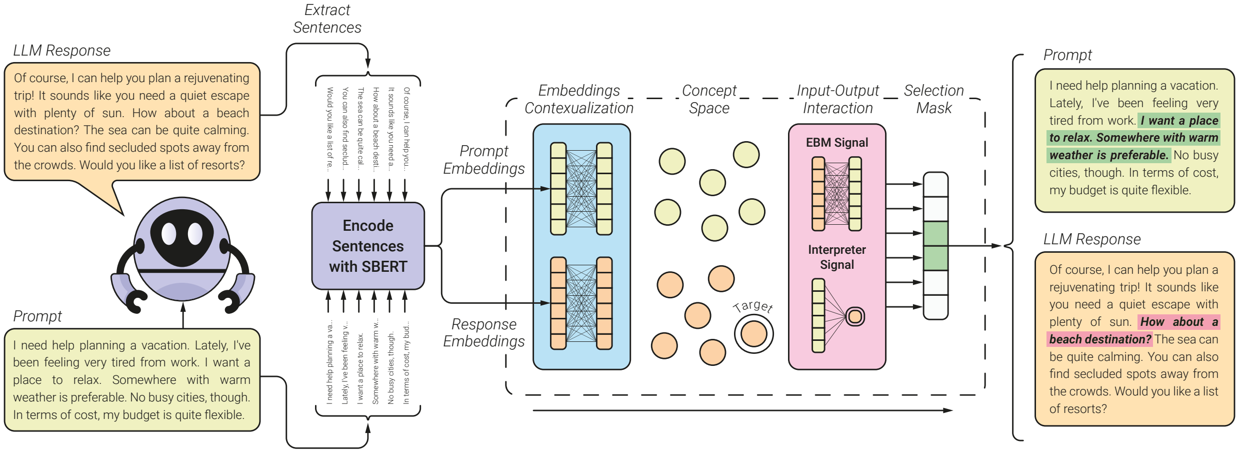

Extract
Sentences
LLM Response
Of course, I can help you plan a rejuvenating trip! It sounds like you need a quiet escape with plenty of sun. How about a beach destination? The sea can be quite calming. You can also find secluded spots away from the crowds. Would you like a list of resorts?
Prompt
I 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.
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...
Prompt Embeddings
Response Embeddings
Embeddings Contexualization
Concept Space
Target
Input-Output Interaction
EBM Signal
Interpreter Signal
Selection Mask
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?

Figure 1: The prompt and response of the target black-box LLM are split in a sentence-wise manner and given to a pre-trained sentence embedding module. Embeddings are contextualized and transferred to a concept space, which simulates the generation process of the LLM. This concept space is trained using signals from an energy network. The proximity of sentences in the concept space demonstrates the influence of prompt sentences in generating output sentences.

the target subset of the output. Then, we can find a subset of prompt sentences that were most influential in generating that target. Formally, given a prompt $\mathbf{x}$, an LLM response $\mathbf{y}$, and a user-selected target subset of the output $\mathbf{y}_T \subset \mathbf{y}$, our objective is to identify the subset of prompt sentences $\mathbf{x}_S \subset \mathbf{x}$ that were most influential in generating $\mathbf{y}_T$.

Prompt-based self-explanation, where an LLM is asked to justify its output using techniques like Chain-of-Thought (Wei et al., 2022), is prone to circular logic; it relies on the same, potentially biased, generative capabilities we seek to understand. Such explanations can produce plausible-sounding confabulations that are unfaithful to the model's true computational process, a form of motivated reasoning (Turpin et al., 2023). While efforts to steer model outputs via automated prompt engineering can overcome certain biases, these methods cast the problem as a search for an instruction that maximizes the probability of a known, pre-defined target answer (Zou et al., 2023; Zhou et al., 2023), which is almost unusable for interpretation tasks.

In this work, we propose a framework to train a post-hoc attribution interpreter for an arbitrary LLM. Here, we globally train a local interpreter. A local interpreter attempts to explain the behavior of the model on a per-instance basis. Local interpreters like LIME (Ribeiro et al., 2016) only observe samples from the neighborhood of the target instance during interpretation. However, this approach may not capture the global behavior of the complex model and can result in interpretability illusions (Friedman et al., 2024). By training globally, our interpreter observes various prompts and LLM responses, which helps it to capture the global behavior of the LLM and mitigate the effects of its intrinsic biases.

Figure 1 illustrates an overview of the proposed approach. We train a transformer-based energy network over the sentences of a prompt and an LLM response. In this network, prompt and response sentences are mapped to a concept space that simulates input-output latent semantic relationships embedded in the target LLM. Then, the energy value is calculated based on the proximity of sentences in this concept space. The energy value can be considered as a metric that indicates how likely two texts are to be a prompt and its corresponding output from the target LLM. Using this metric, we train an interpreter for the target LLM, which takes the prompt and a subset of output sentences as input and returns an importance vector for the prompt sentences. This importance vector shows how much each prompt sentence influenced the LLM to generate the target subset of the output. Using this importance vector, the interpreter selects the subset of the most influential sentences in the prompt for the generation of the output. The key advantages of this work are as follows:

1. We shift the unit of analysis from noisy tokens to semantically coherent concepts, which we define as sentences, enabling a more human-intelligible level of attribution.
2. To operate in a black-box setting, we train a transformer-based energy-based model (EBM) that learns a random field over prompts and their corresponding LLM responses to simulate its thought process. This model successfully distinguishes responses of the target LLM from those written by humans or generated by other language models. As a surrogate for the target LLM, we use this EBM to guide the training procedure of an interpreter.

3. Finally, we propose a post-hoc, model-agnostic framework for interpreting black-box LLMs at a conceptual level. This interpreter finds the most influential sentences of the prompt, which triggered the LLM to generate the target subset of output sentences.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 details our proposed method, and Section 4 presents our experimental results. Section 5 discusses the implications of our findings, and Section 6 concludes with limitations and future directions.

2 Related Work

2.1 Post-hoc Attribution Methods to Explain Language Models

Input-output attribution methods aim to score the importance of input features for a given model output. A major line of work requires white-box access to model internals. Gradient-based methods compute saliency maps by propagating the output gradient back to the input tokens (Simonyan et al., 2014; Sundararajan et al., 2017; Shrikumar et al., 2017; Chefer et al., 2021). Attention-based methods propose using the model's internal attention weights as a direct proxy for feature importance (Xu et al., 2015; Li et al., 2017; Xie et al., 2017; Hao et al., 2021), though their faithfulness has been challenged (Jain & Wallace, 2019). Similarly, influence functions (Koh & Liang, 2017) provide data-centric explanations but are infeasible for proprietary APIs as they require access to the model's Hessian and training data.

To circumvent the need for model access, several paradigms have been developed. Perturbation-based methods offer a model-agnostic alternative by measuring the change in output when parts of the input are removed or altered (Ribeiro et al., 2016; Lundberg & Lee, 2017; Yin & Neubig, 2022). Inspired by this approach, one study finds the influence of individual words in a prompt given to an LLM to generate an output (Hackmann et al., 2024). However, making this approach scalable for generative tasks often incurs a high computational cost, requiring thousands of model queries for a single explanation (Enouen et al., 2024; Zhao & Shan, 2024).

Finally, prompt-based self-explanation uses the LLM's own generative capabilities to produce a rationale, most notably through Chain-of-Thought (CoT) prompting (Wei et al., 2022). This approach is compelling but lacks guarantees of faithfulness, as the generated explanation may not reflect the model's true internal computation path but rather a plausible post-hoc rationalization (Turpin et al., 2023).

2.2 Energy-Based Models in Natural Language Processing

Energy-Based Models (EBMs) have been successfully adapted for generative NLP, primarily through using their ability to learn a global, sequence-level scoring function. One study demonstrated that a Transformer-based discriminator, trained to distinguish between human and machine-generated text, can effectively function as an EBM that assigns low energy to natural, coherent sequences (Bakhtin et al., 2019). This established the potential of EBMs as powerful, holistic text evaluators. One paradigm uses EBMs to refine the output of existing models. The Residual EBM approach adds a corrective energy term to the log-probabilities of a base autoregressive model, allowing the EBM to focus on capturing high-level properties like coherence that the base model may lack (Deng et al., 2020; Bakhtin et al., 2021). In another approach, the EBM is used as a post-processing reranker, which scores results of a base model to select the highest-quality one (Bhattacharyya et al., 2021). The work by (Tu et al., 2020) uses a powerful autoregressive model as a teacher to define an energy landscape; a student network is then trained via knowledge distillation to directly generate outputs that minimize this energy.

2.3 Concept-Based Explanations

Growing consensus in interpretability research suggests that token-level attributions are often too granular for human understanding. This has prompted a shift toward concept-based explanations that map model decisions to higher-level, human-intelligible ideas rather than individual features (Kim et al., 2018). Our work aligns with this paradigm by defining the sentence as the fundamental conceptual unit, as it represents a complete, robust thought for interpretation.

Treating sentences as coherent semantic units is well-justified by the evolution of language models. Foundational architectures like BERT were pre-trained with a Next Sentence Prediction (NSP) task to understand logical sentence relationships (Devlin et al., 2019). Subsequent work, such as Sentence-BERT, confirmed that fine-tuned sentence-level representations map similar meanings to nearby points in a vector space, establishing sentences as distinct semantic objects (Reimers & Gurevych, 2019). By leveraging sentences as our conceptual unit, our work is situated within recent architectural innovations, such as Large Concept Models (LCMs), which propose shifting 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 to interpret the response of any black-box language model to a specific prompt. When analyzing LLMs, working at the level of tokens is often sub-optimal and computationally challenging. While individual words or tokens can be ambiguous, a sentence is the smallest unit of language that expresses a complete thought, proposition, or idea. By treating each sentence as a fundamental concept, we can analyze the model's generation process at a more abstract and human-understandable level, focusing on the interplay of complete ideas rather than fragmented tokens.

Let $\mathbf{x}$ be a prompt to the LLM and $\mathbf{y}$ be the corresponding output. Also, let $\mathbf{y}_T$ denote the target subset of output sentences for which the interpreter should quantify the influence of each prompt sentence. To this end, we propose a two-stage framework: First, we pre-train an Energy-Based Model (EBM), $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$, that evaluates the consistency of a pair $(\mathbf{x}, \mathbf{y})$ with the generation pattern of the target LLM. The energy model is a function with the set of parameters β which outputs a scalar value. Second, we use this EBM to guide the training of a lightweight interpreter, $\mathcal{IN}(\mathbf{x}, \mathbf{y}_T; \alpha)$, with the set of parameters α. The interpreter is a function that takes the prompt, the LLM response, and a user-specified subset of the output sentences, and returns a binary vector whose size equals the number of prompt sentences. In this vector, the values of 1 indicate selected sentences of the prompt as the most influential in generating the target.

3.1 PRE-PROCESSING WITH SENTENCE-BERT

The first step in our pipeline is to pre-process both the input text $\mathbf{x}$ and the output text $\mathbf{y}$ by segmenting them into sequences of sentences. We employ a pre-trained sentence transformer, such as Sentence-BERT (Reimers & Gurevych, 2019), as a frozen embedding module. This module maps each sentence to a fixed-dimensional embedding, producing initial representations as a sequence of embeddings for the prompt, S^{in}, and the output, S^{out}. In our architecture, these sentence embeddings are treated analogously to token embeddings in a standard Transformer; they are fed sequentially into the attention blocks of our EBM and interpreter to capture contextual dependencies between complete thoughts.

To ensure compatibility with the fixed-dimensional attention mechanisms in our architecture, we normalize the sentence counts. We define a task-dependent hyperparameter, N_{max}, representing the fixed sequence length. The sentence sequences are padded with a learnable placeholder or truncated to strictly match this length, resulting in input and output embedding tensors $S^{\text{in}}, S^{\text{out}} \in \mathbb{R}^{N_{\text{max}} \times d}$.

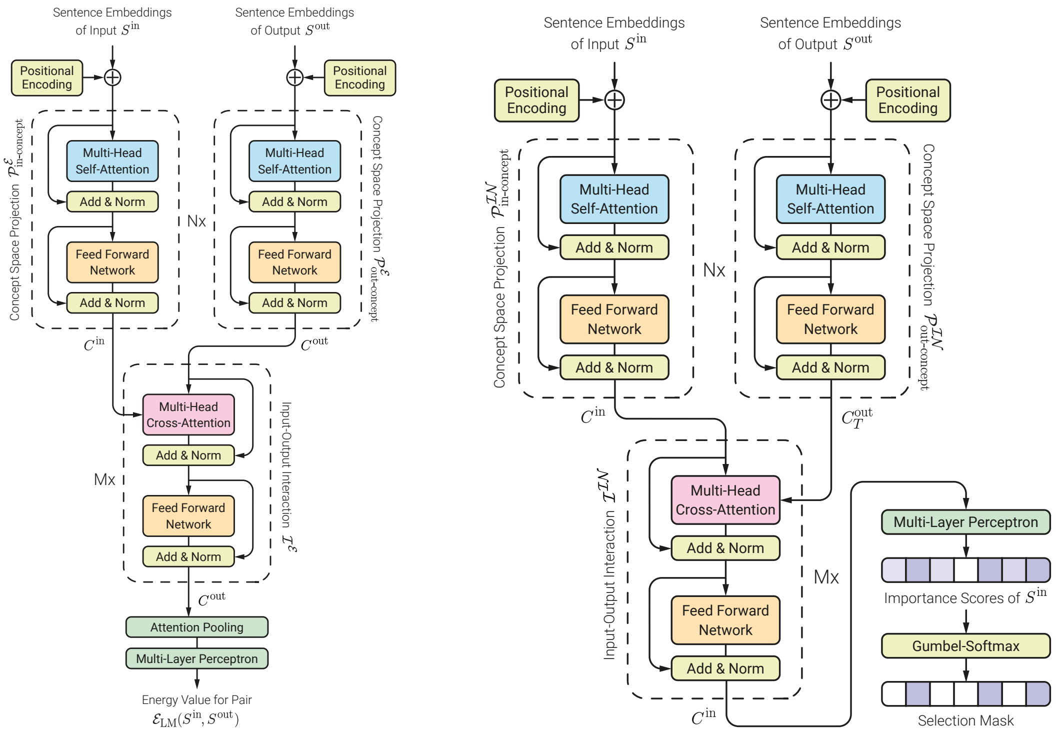
3.2 THE ENERGY-BASED SURROGATE MODEL

To approximate the behavior of the black-box LLM, we design a globally-aware EBM, $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$, that learns to distinguish the most likely target-LLM-generated pairs from inauthentic ones using an energy score. The lower the assigned energy, the more likely the pair is authentic and consistent with what the target LLM would generate. This EBM serves as a differentiable and lightweight surrogate that captures the underlying logic of the LLM.

Figure 2a shows the architecture of the proposed EBM. In this module, sentence embeddings are processed as follows:

1. **Concept Space Projection:** Static embeddings from Sentence-BERT capture semantic meaning in isolation but lack knowledge of the specific context of the prompt-response pair. To address this, embeddings S^{in} and S^{out} are passed through separate, trainable self-attention modules ($\mathcal{P}^{\mathcal{E}}_{\text{in-concept}}$ and $\mathcal{P}^{\mathcal{E}}_{\text{out-concept}}$). These modules contextualize each sentence with respect to its surrounding sentences, transforming the static representations into a dynamic "concept space" (C^{in} and C^{out}). The concept space captures the LLM's internal dependency structure rather than relying solely on generic semantic similarity. Within this space, the distance between sentences is determined by the LLM's inherent generative patterns, which are a function of its underlying architecture and training dataset.
2. **Input-Output Interaction:** To model the influence of the input on the output, the output concepts C^{out} attend to the input concepts C^{in} through a cross-attention block. This allows the model to weigh the relevance of each input concept to the overall output.
3. **Energy Calculation:** The resulting representations are aggregated via attention pooling and passed through a final multi-layer perceptron (MLP), which outputs a single scalar energy value $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$.

The EBM network is trained in two phases. First, it is pre-trained using an expanded InfoNCE loss. Then, it is fine-tuned during the training of the interpreter. In the pre-training phase, we generate a dataset of a set of prompts and their corresponding outputs $(\mathbf{x}, \mathbf{y})$, by querying the target black-box LLM.


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
Concept Space Projection $\mathcal{P}^{\mathcal{E}}_{\text{out-concept}}$
C^{in}
C^{out}
Multi-Head Cross-Attention
Mx
Input-Output Interaction $\mathcal{I}^{\mathcal{E}}$
C^{out}
Attention Pooling
Multi-Layer Perceptron
Energy Value for Pair $\mathcal{E}_{\text{LM}}(S^{\text{in}}, S^{\text{out}})$
Concept Space Projection $\mathcal{P}^{\mathcal{IN}}_{\text{in-concept}}$
Concept Space Projection $\mathcal{P}^{\mathcal{IN}}_{\text{out-concept}}$
C^{in}
C^{out}_T
Input-Output Interaction $\mathcal{I}^{\mathcal{IN}}$
Importance Scores of S^{in}
Gumbel-Softmax
Selection Mask

(a)

(b)

Figure 2: Proposed architectures of (**a**) the energy network and (**b**) the interpreter network

To apply the InfoNCE loss, we generate a rich set of negative samples for each positive one. In a **structural corruption** setup, for a pair $(\mathbf{x}, \mathbf{y})$, we create two types of negatives: $(\mathbf{x}, \mathbf{y}')$ and $(\mathbf{x}', \mathbf{y})$. The corruptions $\mathbf{x}'$ and $\mathbf{y}'$ are formed by either masking entire sentences (to teach concept continuity) or masking individual tokens (to teach grammatical structure). These negative samples encourage the energy function to assign higher energy to less reasonable sentences. In a **semantic-based** negative sampling, to learn the style and context of pairs, we also construct negatives using outputs from other models or humans, as well as through *off-topic sampling* to switch correct pairs. In the latter, for a positive pair $(\mathbf{x}_i, \mathbf{y}_i)$, we generate negative samples of the form $(\mathbf{x}_i, \mathbf{y}_j)$ and $(\mathbf{x}_j, \mathbf{y}_i)$ for $j \neq i$. This encourages the energy function to capture conceptual inconsistencies and approximate the target LLM's generation distribution by forcing it to distinguish authentic outputs from alternatives, thereby sharpening its discrimination boundary for better guidance of the interpreter.

Finally, the energy model is pre-trained to assign a lower energy score to a positive pair compared to all negative ones by the following loss function:

$$\mathcal{L}_{\mathcal{E}}^{PT} = -\log \left(\frac{\exp(-\mathcal{E}_{\text{LM}}(\mathbf{x}_i, \mathbf{y}_i; \theta)/\tau)}{\exp(-\mathcal{E}_{\text{LM}}(\mathbf{x}_i, \mathbf{y}_i; \theta)/\tau) + \sum_{(\mathbf{x}', \mathbf{y}') \in \mathcal{N}_i} \exp(-\mathcal{E}_{\text{LM}}(\mathbf{x}', \mathbf{y}'; \theta)/\tau)} \right) \tag{1}$$

where $\mathcal{N}_i$ is the set of negative samples for the i-th pair and τ is a temperature hyperparameter. In appendix A, Figure 5 describes the pre-training procedure of the energy block.

3.3 The Interpreter Model

Given a prompt $\mathbf{x}$, the LLM response $\mathbf{y}$, and a user-specified subset of output sentences $\mathbf{y}_T \subseteq \mathbf{y}$ (the target), the interpreter's objective is to quantify the influence of each sentence in $\mathbf{x}$ on the generation of $\mathbf{y}_T$. The interpreter outputs a binary importance vector corresponding to the sentences in $\mathbf{x}$, where a value of 1 indicates that a prompt sentence was a necessary precursor for the target output.

Figure 2b shows the architecture of the interpreter. First, embeddings of prompt sentences S^{in} are projected to a concept space using a self-attention module $\mathcal{P}^{\mathcal{IN}}_{\text{in-concept}}$. Sentence embeddings of the output S^{out} are also projected

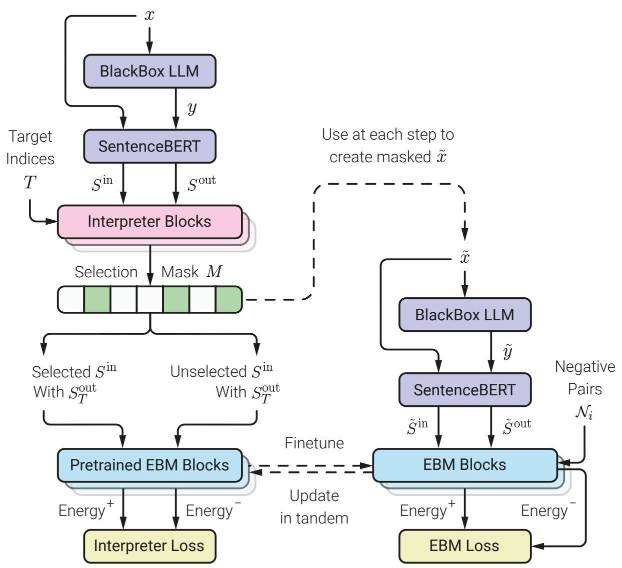


Figure 3: An overview of the EBM-guided training procedure for the interpreter. The framework involves a joint, alternating optimization where the interpreter's parameters are updated using a fixed EBM to create an input mask (left). Subsequently, the EBM is fine-tuned using the LLM's response to the masked prompt to mitigate distribution shift (right).

using another self-attention module $\mathcal{P}^{\mathcal{IN}}_{\text{out-concept}}$. Then, only the target sentences are retained and the others are masked. In the next step, the input concepts C^{in} attend to target concepts C_T^{out} via a cross-attention mechanism. The resulting attention weights are passed through an MLP layer, which results in a vector of importance scores. The last step of the interpreter module is a Gumbel-Softmax unit (Jang et al., 2016) by which we encourage the interpreter to find a specific number of the most important sentences for the target output. Details of this module are provided in appendix B. Considering $\mathbf{x}$ as the prompt, we define,

$$\tilde{\mathbf{x}} = \mathbf{x} \odot \mathcal{IN}(\mathbf{x}; \mathbf{y}_T, \alpha) \tag{2}$$

where $\odot$ denotes the element-wise multiplication of the interpreter output with sentences of $\mathbf{x}$. As the energy function simulates the generation process of the LLM, if the interpreter successfully selects the most important sentences of the input for the target, the value of $\mathcal{E}_{\text{LM}}(\tilde{\mathbf{x}},\ \mathbf{y}_T; \theta)$ should be at a minimum among the energy values for all other possible subsets of input sentences. Equivalently, as $\mathbf{x} - \tilde{\mathbf{x}}$ shows the least relevant parts of the prompt to the target, the value of $\mathcal{E}_{\text{LM}}(\mathbf{x} - \tilde{\mathbf{x}},\ \mathbf{y}_T; \theta)$ should be at a maximum among the energy values for all other possible subsets of input sentences. If $\tilde{\mathbf{x}}$ were the most related subset of the prompt to the target output, the difference between these two values will be maximized. Therefore, we believe the ideal interpreter solves the following optimization problem:

$$\arg\max_{\alpha} \quad \mathbb{E}_{(x,y)} \left[\mathcal{E}_{\text{LM}}(\mathbf{x} - \tilde{\mathbf{x}},\ \mathbf{y}_T; \theta) - \mathcal{E}_{\text{LM}}(\tilde{\mathbf{x}},\ \mathbf{y}_T; \theta) \right] \tag{3}$$

3.3.1 EBM-Guided Training of the Interpreter

In equation 3, the energy value is calculated for the masked prompt and the LLM response. As mentioned in (Hsia et al., 2023), masking samples leads to a distribution shift compared to the original input distribution. Therefore, to mitigate the effect of the out-of-distribution problem, we fine-tune the energy network during the training of the interpreter with an InfoNCE loss.

Figure 3 describes an overview of the proposed procedure to train the interpreter. The interpreter and EBM are optimized jointly in an alternating fashion, allowing them to adapt to each other and improve performance. The final optimization procedure is as follows:

1. **Update Interpreter:** In iteration k, for a fixed EBM $\mathcal{E}(.,.;\theta^{(k-1)})$, we update the interpreter's parameters to maximize equation 3 as follows:

$$\alpha^{(k)} \leftarrow \alpha^{(k-1)} + \beta\ \nabla_\alpha \mathbb{E}_{(x,y)}[\ \mathcal{E}_{\text{LM}}\left(\mathbf{x} - \mathbf{x} \odot \mathcal{IN}(\mathbf{x}; \mathbf{y}_T, \alpha),\ \mathbf{y}_T;\ \theta^{(k-1)}\right) - \mathcal{E}_{\text{LM}}\left(\mathbf{x} \odot \mathcal{IN}(\mathbf{x}; \mathbf{y}_T, \alpha),\ \mathbf{y}_T;\ \theta^{(k-1)}\right)\ ] \tag{4}$$

2. **Getting LLM response to the selected subset of the prompt:** The prompt is masked with the output of the current interpreter. Then, the response of the LLM is obtained for the masked prompt:

$$\tilde{\mathbf{x}} = \mathbf{x} \odot \mathcal{IN}(\mathbf{x}, \mathbf{y}_T; \alpha^{(k)}) \tag{5}$$

$$\tilde{\mathbf{y}} = \text{LM}(\tilde{\mathbf{x}}) \tag{6}$$

3. **Update EBM:** Using the new sample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, the energy function is fine-tuned to capture the LLM behaviour in unexplored areas and avoid the distribution-shift problem:

$$\theta^{(k)} \leftarrow \theta^{(k-1)} + \beta' \nabla_\theta \log \left(\frac{\exp(-\mathcal{E}_{\text{LM}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \theta)/\tau)}{\exp(-\mathcal{E}_{\text{LM}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \theta)/\tau) + \sum_{(\mathbf{x}',\mathbf{y}') \in \mathcal{N}} \exp(-\mathcal{E}_{\text{LM}}(\mathbf{x}', \mathbf{y}'; \theta)/\tau)} \right) \tag{7}$$

where **LM** denotes the target LLM, β and β' are learning rates and τ is the InfoNCE loss temperature. $\mathcal{N}$ shows the set of negative samples generated for the pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. The energy network's parameters are initialized from the pre-trained model. In the interpreter network, parameters of the self-attention modules are initialized with those in the energy network and the cross-attention module is initialized randomly. During this procedure, the information from the energy concept space, which is extracted from the target LLM, is transferred to the interpreter concept space. After training, the proximity of sentences in the concept space of the interpreter depends on the LLM's thought process. Therefore, it acts as a standalone tool for interpreting the target LLM's prompts and responses.

4 Experiments

In this section, we present the empirical validation of our proposed framework. We first demonstrate the effectiveness of the EBM pre-training in creating a faithful surrogate model, and then proceed to evaluate the performance of the final interpreter. We emphasize that our framework is designed to train *task-oriented* interpreters; rather than capturing the universal output space of the LLM, each interpreter is specialized for a specific domain (e.g., sentiment analysis). This focused scope allows for effective training even with limited dataset sizes.

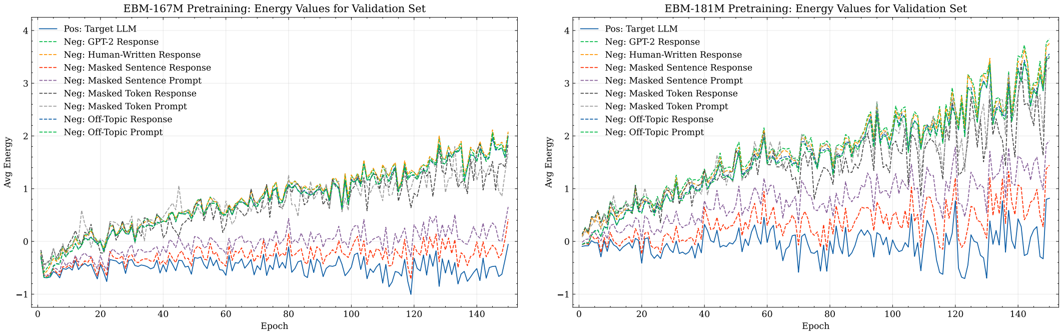
4.1 Validating the EBM as a Faithful Surrogate

The first experiment validates that our EBM can be effectively pre-trained to serve as a surrogate for a target LLM. The goal is to show that the EBM learns a well-defined energy landscape, assigning low energy to authentic prompt-response pairs while assigning high energy to various forms of corrupted or mismatched pairs.

Target LLM and Dataset. Our objective is to model the energy surface of `GPT-4o mini`. We utilized Hello-SimpleAI/HC3 dataset, a corpus widely used for comparing human and LLM-generated text across multiple domains (e.g., science, finance, medicine, and open-domain QA). We used the "question" column from the training split to prompt our target LLM, `GPT-4o mini`, to generate positive samples. For negative sampling, we used the provided human answers and also generated responses from `GPT-2` for the same prompts. We trained two models of different sizes: a smaller EBM-167M on 12,000 samples and a larger EBM-181M on 20,000 samples. For each, 10% of the data was reserved for validation.

Model Architecture. We evaluated two EBM architectures, EBM-167M and EBM-181M, with 167 million and 181 million parameters, respectively. Both models use two concept space projection blocks for the input and output sentences, where each block consists of a self-attention layer with 8 heads and a feed-forward network. The key difference lies in the input-output interaction module; the EBM-167M uses four cross-attention blocks, while the more powerful EBM-181M uses six. Both architectures conclude with a two-layer MLP following the attention pooling stage.

Pre-training and Negative Sampling. The EBMs were pre-trained using the InfoNCE loss with a rich set of negative samples designed to teach different aspects of the target LLM's behavior. These included: **structural corruptions**



(a) EBM-167M Average Validation Energy

(b) EBM-181M Average Validation Energy

Figure 4: Averaged energy value for validation set for EBMs with two different sizes. The blue line shows the value for the target LLM response which successfully is the minimum one. For EBM-167M (**a**), negative samples become successfully distant from the positive one during training. For EBM-181M (**b**), we have a significantly wider energy gap, indicating enhanced discriminative power for larger network.

(masking entire sentences or individual tokens in either the prompt or the response) and **semantic-based negatives** (swapping the response with one from a different model like `GPT-2`, a human-written answer, or an off-topic response from the model).

The EBM pre-training results, summarized in Figure 4 and corroborated by consistent loss convergence (Appendix A, Figure 6), show that both models successfully learned the target LLM's generation patterns. The validation energy plots reveal that the EBMs assign the lowest energy to authentic target samples, effectively distinguishing them from all negatives. As anticipated, semantic-based negatives (e.g., human or `GPT-2` responses) receive the highest energy, while structural corruptions like sentence masking proved to be the most challenging negative class, highlighting the model's sensitivity to nuanced, conceptual changes. Critically, distinguishing difficult off-topic negatives required careful tuning of model capacity and training data balance. Of the two, the larger EBM-181M trained on more data exhibited superior performance, showing a significantly wider energy gap (Figure 4b) than the EBM-167M. We observe that a wider energy gap directly correlates with higher classification accuracy, particularly when distinguishing the target LLM's outputs from hard negatives that are semantically plausible but inauthentic. This superior, well-defined energy surface confirms the framework's scalability and potential for high-fidelity modeling of complex LLMs.

4.2 Interpretation Results

We train our interpreter model using guidance from the pre-trained energy block, EBM-181M (introduced earlier). The interpreter incorporates two concept-space projection blocks—one for input sentences and one for output sentences. Each projection block contains self-attention layers (initialized with the energy block weights to better align with concept spaces) and six cross-attention layers with eight attention heads, which is followed by a single MLP layer. To facilitate a more descriptive evaluation, we selected a movie review sentiment analysis task as the target LLM scenario. The goal is to identify which parts of each review most strongly influence the LLM's generated response. Specifically, we provide IMDB reviews (Lakshmi, 2020) to `GPT-4o mini` and ask it to produce a sentence summarizing the sentiment of the review. This generated sentiment sentence is then passed as the target output to the interpreter network. Table 1 presents some examples of post-hoc interpretation results produced by the interpreter model.

5 Discussion

The growing integration of generative models into critical applications heightens the demand for transparency and interpretability. In this work, we contribute to this effort by introducing a novel method for elucidating the generative mechanics of Large Language Models (LLMs) at a conceptual level. Results of our study provide important insights into the potential of energy-based models (EBMs) as a surrogate for black-box Large Language Models (LLMs) for further analysis. By employing a concept-level approach, we have demonstrated that the energy model successfully guides training of a model-agnostic, post-hoc input-output attribution method to interpret the LLM response at the sentence level.

Table 1: Results of interpreting the `GPT-4o mini` responses to IMDB review prompts

Review Full Text (Top Sentences Highlighted)	**Predicted Sentiment**
Sample #1	
This is an excellent film, with an extraordinary cast and acting. I was very disappointed when it didn't win the Academy Awards for best film and best actress (Whoopi Goldberg). It certainly deserved it. In any case, take a look at it; I am sure you will enjoy it very much.	Based on the given text, the review sentiment is **positive**.
Sample #2	
Some people might consider this movie a piece of artwork—to be able to express your imagination on film in order to create a movie filled with antagonizing pain and death. I personally think that this movie is a disgust, which should have never been released. This movie is repulsive, illogical, and meaningless. Not only is it a complete waste of time but it makes you sick for days to come. The appalling images shown in the film not only make you grasp for air but they set in your mind and it takes days to forget them. Such a shame that people waste their imagination on such inhumane suffering. "Kill Bill" would be another example, but at least "Kill Bill" has its purpose, meaning, climax, and resolution.	Based on the given text, the review sentiment is **negative**.
Sample #3	
Officially the first martial arts movie in USSR cinematography featuring actual martial artists like Tadeush Kas'yanov and Russian Bruce Lee - Talgat Nigmatullin. Bad people hijack a ship on the high seas, but fortunately, just about everybody on board is a trained martial artist. A collectible for martial arts aficionados.	Based on the given text, the review sentiment is **positive**.
Sample #4	
Slow, incomprehensible, boring. Three enthusiastic words that describe the movie of the book. This is surely a case where the movie should never have been made at the expense of the book. The best part of the movie was the scenery, excellent. The worst part was the slow moving interactions of the actors which combined with endless meaningful glances. The editing is abrupt and patchy. However, despite this, the actors worked very hard at least trying to be a little believable with a terrible script. It was startling that although set in Peru there was hardly a person of Peruvian descent wandering about the set—even in the flashback scenes depicting Peru in the 17th century. If you have any sense of history, try to avoid this movie.	Based on the given text, the review sentiment is **negative**.

Our method identifies the salient sentences in a prompt that trigger specific components of an LLM's response. We hypothesize that this attribution capability could serve as a mechanism for diagnosing model failures; for instance, a generated sentence with no strong attribution to the prompt might indicate a hallucination or the activation of internal parametric biases. Unlike local surrogate methods, our globally trained interpreter captures not only instance-specific relationships but also broader semantic patterns embedded in the LLM's generation process across the dataset. Furthermore, we demonstrate that our approach is scalable, with larger EBMs (e.g., EBM-181M) yielding improved performance in distinguishing authentic from corrupted data pairs. Ultimately, this framework offers a scalable pathway toward verifiable and transparent AI deployment in high-stakes environments.

6 Limitations and Future Work

Cost-Benefit Trade-off. A primary limitation of our approach is the computational cost associated with the contrastive pre-training of the EBM. Our framework shifts the computational burden from inference to training, incurring a significant one-time cost. However, unlike perturbation methods (e.g., LIME) that require thousands of queries per explanation, our trained interpreter operates as a standalone $O(1)$ tool with zero additional API queries, making it highly suitable for high-volume analysis once deployed.

Experimental Scope and Validation. We acknowledge that the current work serves as a proof-of-concept. Due to computational constraints, our current experiments rely on smaller datasets. However, our results suggest the framework remains effective at this scale because our interpreters are *task-oriented*—specialized for specific behaviors rather than the entire LLM manifold.

Future Directions. Future work will focus on three key areas: 1) **Validation:** We will incorporate rigorous faithfulness metrics (e.g., Leave-One-Out) and benchmark against baselines like Sentence-LIME to further validate the attribution quality. This includes specific stress-tests on hallucinated outputs to verify if the interpreter correctly identifies a lack of grounding in the prompt. 2) **Breadth:** We will expand to diverse tasks (e.g., QA, reasoning, and generation on datasets like WikiText) and models. 3) **Modality:** We aim to extend the framework to multimodal settings, which will necessitate advances in cross-modal concept alignment.

REFERENCES

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

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

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. Large concept models: Language modeling in a sentence representation space, 2024.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. Energy-based reranking: Improving neural machine translation using energy-based models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 4528–4537, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.349.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13984–14011, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.832.

Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14035–14059. PMLR, 21–27 Jul 2024.

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

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971, May 2021. doi: 10.1609/aaai.v35i14.17533.

Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C Lipton. Goodhart's law applies to nlp's explanation benchmarks. *arXiv preprint arXiv:2308.14272*, 2023.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1885–1894. JMLR.org, 2017.

N. Pathi Lakshmi. Imdb dataset of 50k movie reviews. `https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews`, 2020.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In John DeNero, Mark Finlayson, and Sravana Reddy (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020.

Johannes Schneider. Explainable generative ai (GenXAI): a survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11):289, 9 2024. ISSN 1573-7462. doi: 10.1007/s10462-024-10916-x.

S Fatemeh Seyyedsalehi, Mahdieh Soleymani, and Hamid R Rabiee. Sointer: A novel deep energy based interpretation method for explaining structured output models. *arXiv preprint arXiv:2202.09914*, 2022.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. ENGINE: Energy-based inference networks for non-autoregressive machine translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2819–2826, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.251.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. An interpretable knowledge transfer model for knowledge base completion. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 950–962, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1088.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 184–198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.14.

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

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

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

Appendix

A Energy Network Pre-training

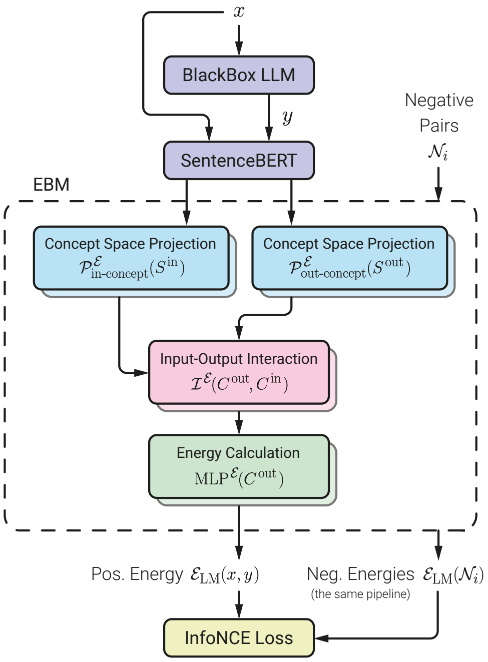
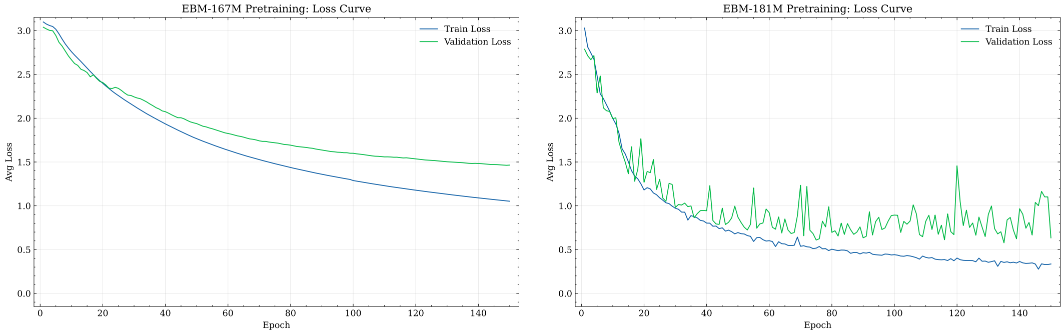

x
BlackBox LLM
y
SentenceBERT
Negative Pairs $\mathcal{N}_i$
EBM
Concept Space Projection $\mathcal{P}^{\mathcal{E}}_{\text{in-concept}}(S^{\text{in}})$
Concept Space Projection $\mathcal{P}^{\mathcal{E}}_{\text{out-concept}}(S^{\text{out}})$
Input-Output Interaction $\mathcal{I}^{\mathcal{E}}(C^{\text{out}}, C^{\text{in}})$
Energy Calculation $\text{MLP}^{\mathcal{E}}(C^{\text{out}})$
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 globally-aware energy-based surrogate model, $\mathcal{E}_{\text{LM}}(\mathbf{x}, \mathbf{y}; \theta)$. Input and output sentences are first embedded via SentenceBERT and then projected into separate concept spaces through trainable self-attention modules. Output concepts cross-attend to input concepts to model input–output dependencies, after which an MLP computes a scalar energy score. Positive pairs $(\mathbf{x}, \mathbf{y})$ from the target LLM and negative pairs $\mathcal{N}_i$ are both scored, and the model is trained with an InfoNCE loss to assign low energy to authentic pairs and high energy to mismatched ones.



(a) EBM-167M Training and Validation Loss

(b) EBM-181M Training and Validation Loss

Figure 6: Pre-training Loss for both EBMs. (**a**) The EBM-167M training and validation loss shows steady convergence. (**b**) The larger EBM-181M loss decreases more rapidly, reflecting the benefits of increased capacity and data volume.

B 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 $u_i \sim \mathrm{Uniform}(0, 1)$ as follows:

$$g_i = -\log(-\log u_i), \quad i = 1, \ldots, n. \tag{A1}$$

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

$$c_i = \frac{\exp((z_i + g_i)/\tau)}{\sum_{j=1}^{n} \exp((z_j + g_j)/\tau)}, \quad i = 1, \ldots, n. \tag{A2}$$

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 Eq. A2. We then aggregate these samples by taking their element-wise maximum:

$$m_i = \max_{j=1,\ldots,K} c_i^{(j)}, \quad i = 1, \ldots, n. \tag{A3}$$

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 defined as:

$$\mathcal{IN}(\mathbf{x}, \mathbf{y}_T; \alpha)_i = m_i. \tag{A4}$$

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.

C THE LLM USAGE

Some 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 ICLR publications.