ProxySPEX: Inference-Efficient Interpretability via Sparse Feature Interactions in LLMs

L. Butler* A. Agarwal* J. S. Kang* Y. E. Erginbas B. Yu K. Ramchandran

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley UNIVERSITY OF CALIFORNIA

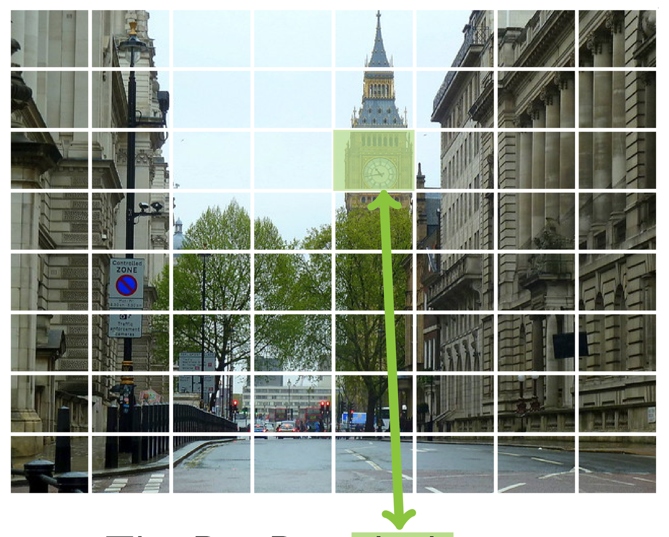
Problem

How can we efficiently identify the influential interactions in large models?

Medical Record:
A 52-year-old woman has a 10-kg weight loss. Her hemoglobin concentration is 7.5 g/dL and leukocyte count is 41,800/mm3. Bone marrow biopsy shows cellular hyperplasia with many immature granulocytic cells.

Model's Decision:
Diagnose with Chronic Myeloid Leukemia

Model Explanation:
10-kg weight loss + cellular hyperplasia
52-year-old + immature granulocytic cells



The Big Ben clock tower peering over the city of London.

1. **(Faithfulness Problem)** Marginal attribution approaches like SHAP/LIME scale, but don't capture important interactions.
2. **(Efficiency Problem)** Prior SOTA still requires tens of thousands of model inferences, which can be prohibitive for complex models such as LLMs.

Formulation: For input $\mathbf{x}$ = "Her acting fails to impress", let $f(\mathbf{x}_S)$ be the output of the LLM under *masking pattern* S. If $S = \{1, 2, 4, 5, 6\}$, then $\mathbf{x}_S$ is "Her acting `[MASK]` fails to impress".

Faithfulness at Scale

We aim to learn an interpretable approximation of f denoted $\hat{f}$. We define faithfulness as the predictive power of $\hat{f}$:

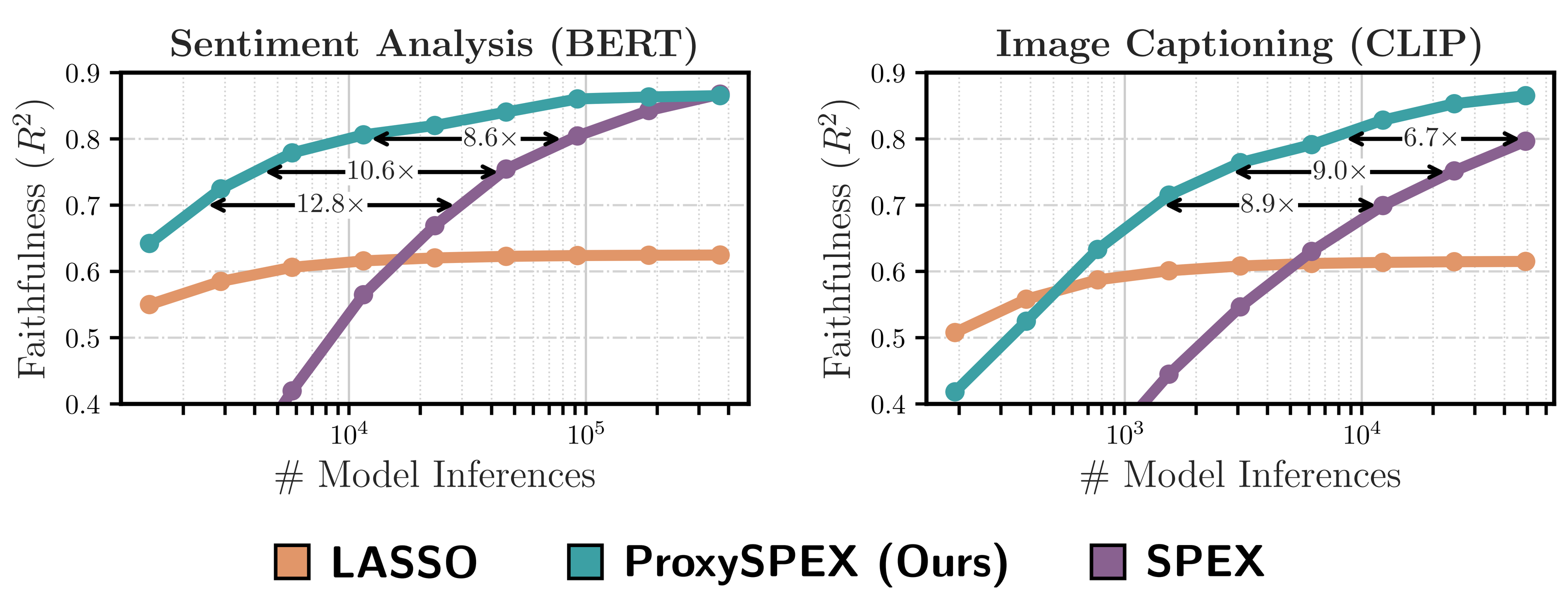
$$R^2 = 1 - \frac{\|\hat{f} - f\|^2}{\|f - \bar{f}\|^2}, \quad \text{where } \|f\|^2 = \sum_{S \subseteq [n]} f(\mathbf{x}_S)^2, \; \bar{f} = \frac{1}{2^n} \sum_{S \subseteq [n]} f(\mathbf{x}_S).$$

$f(\mathbf{x}_S)$ has a unique decomposition under the Fourier transform, expressed as:

$$F(\mathbf{x}_T) = \frac{1}{2^n} \sum_{S \subseteq [n]} (-1)^{|S \cap T|} f(\mathbf{x}_S), \qquad f(\mathbf{x}_S) = \sum_{T \subseteq [n]} (-1)^{|S \cap T|} F(\mathbf{x}_T).$$

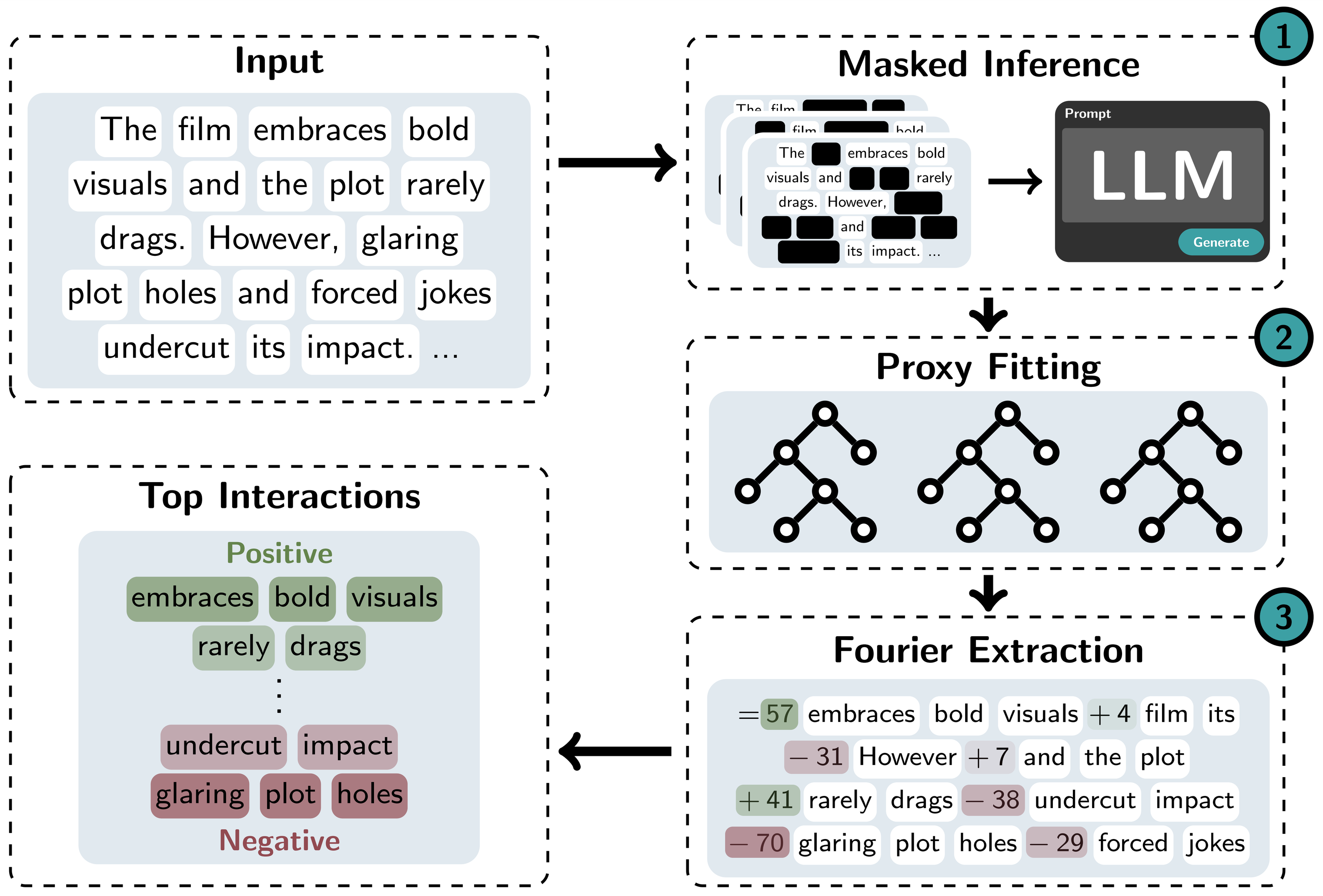
Empirically, we observe that $F(\mathbf{x}_T) \approx 0$ for most T (**sparsity**), large $F(\mathbf{x}_T)$ are **low degree** such that $|T| \leq d$ for some small d, and are correlated (**hierarchy**).

Key Insight: influential interactions are sparse, low-degree, and hierarchical.

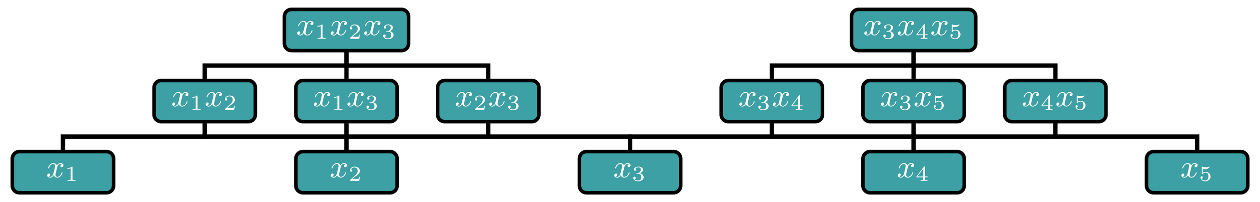


Result: ProxySPEX reduces inference cost by $\sim$10× compared to prior SOTA.

Algorithm



1. ProxySPEX masks words and queries the LLM using this masked input.
2. We fit Gradient Boosted Trees (GBTs) as a proxy model.
3. We extract sparse, hierarchical interactions from the fitted GBTs.



Example: Hierarchical interaction structure extracted from GBTs.

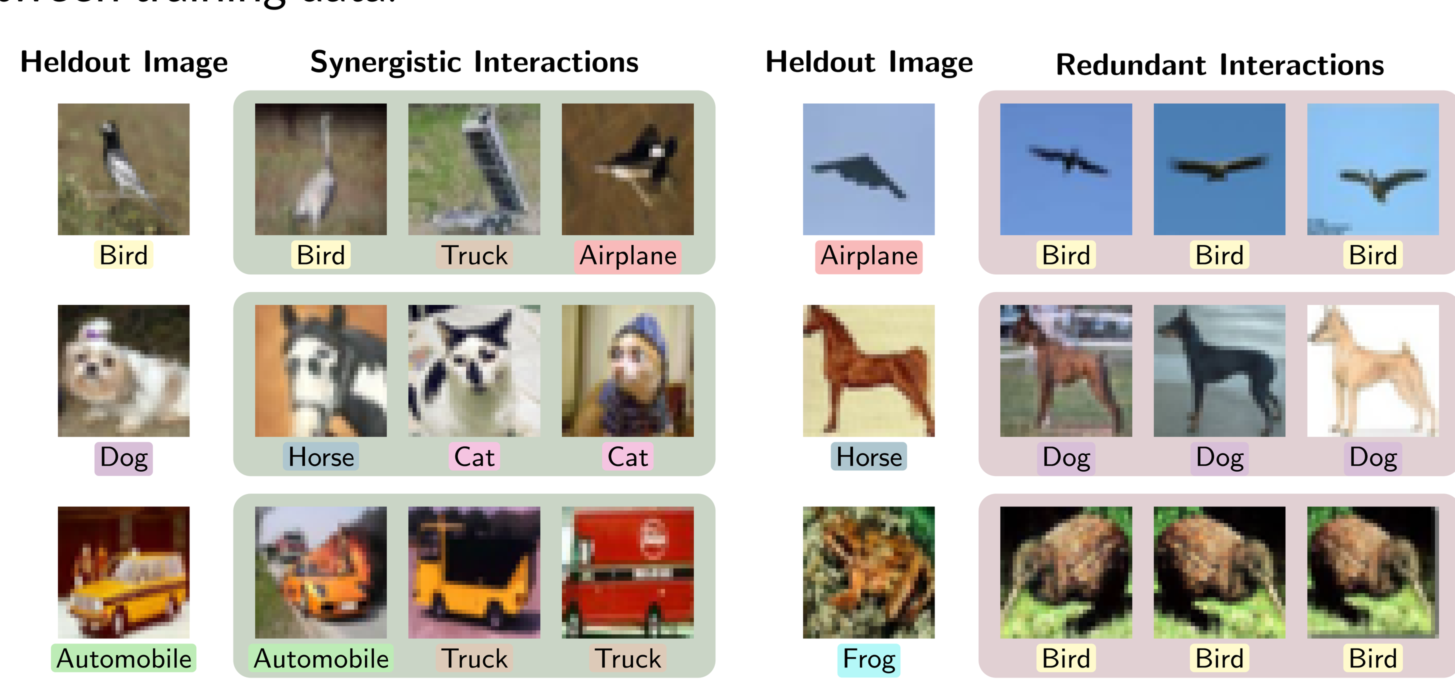
Can we apply ProxySPEX to attribution for model training data, inputs, and internal components?

Experiments: Data Interaction Attribution

Data attribution measures how each training sample influences the prediction of a particular test point $\mathbf{z}$ on class c. We generalize this framework to capture interactions between training samples. For a model trained on S:

$$f(S) \triangleq (\text{logit for } c \text{ on } \mathbf{z}) - (\text{highest incorrect logit on } \mathbf{z}).$$

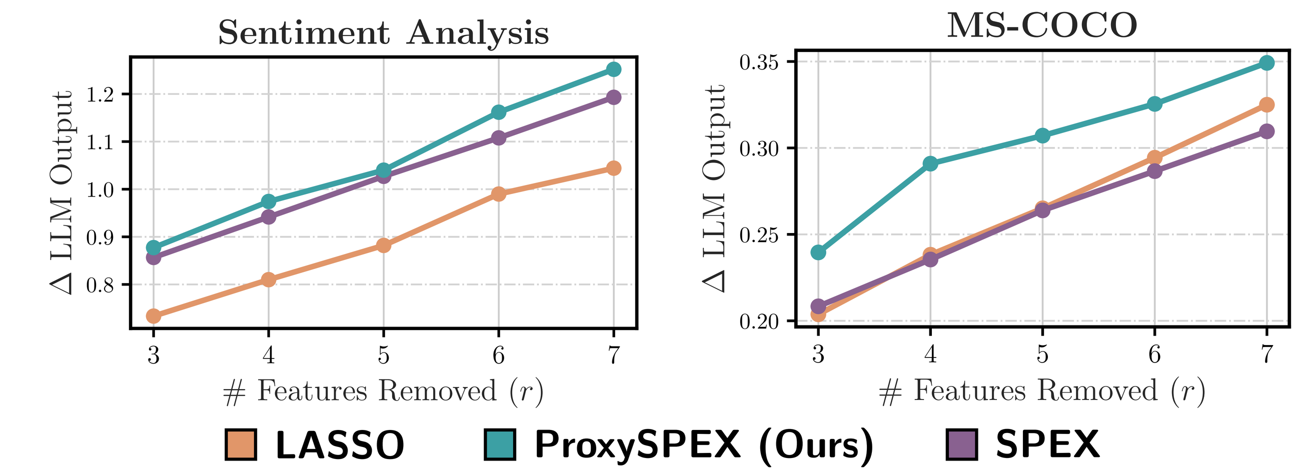
ProxySPEX is the first method efficient enough to learn interactions between training data.



Synergistic Interactions: Combined influence is more than the sum of parts.
Redundant Interactions: Combined influence is less than the sum of parts.

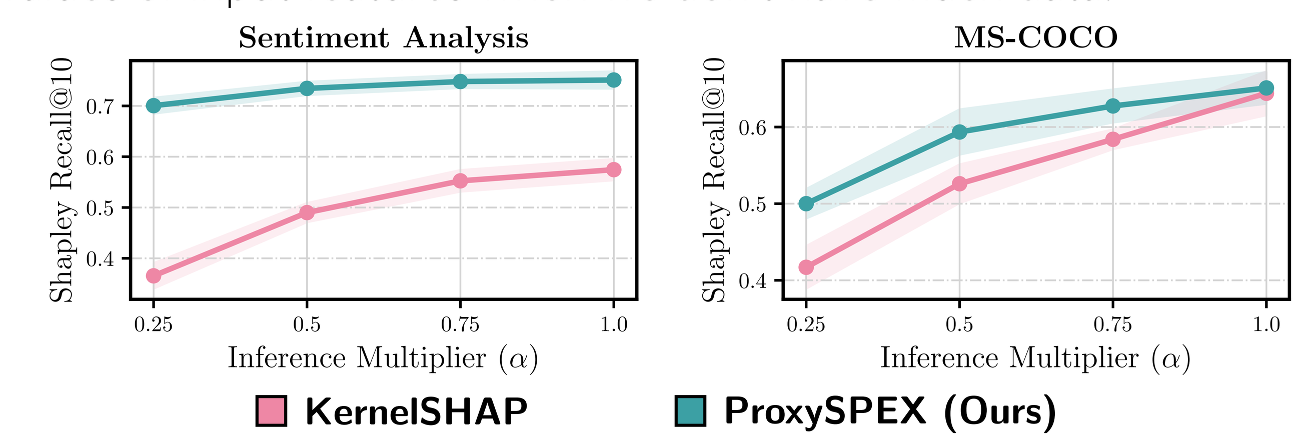
Experiments: Input Feature Interaction Attribution

Feature Removal: By accounting for interactions, ProxySPEX identifies a small set of features to remove that changes the model output.



We use `DistilBERT` for Sentiment Analysis and `CLIP-ViT-B/32` for MS-COCO. Both plots measure a normalized change in logits.

Efficient Shapley Value Estimation: ProxySPEX is SOTA at estimating Shapley values of input features when we don't have much data.

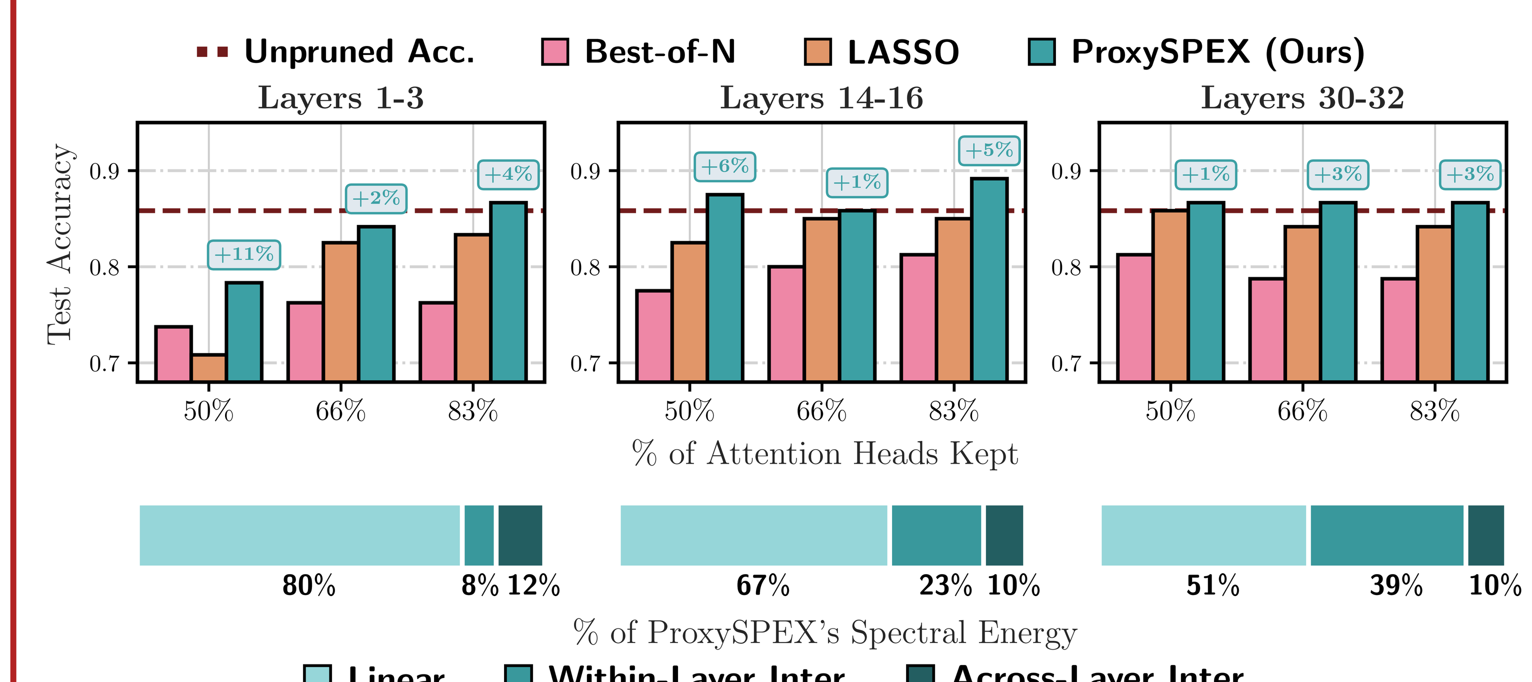


For multipliers $\alpha \in \{0.25, 0.5, 0.75, 1.0\}$, recall of the top ten Shapley values after $\alpha \cdot n \log_2(n)$ inferences.

Experiments: Attention Head Interaction Attribution

To determine which attention heads $\mathcal{H}$ are important for a task, define $\text{LLM}_S(\cdot) \triangleq$ model with only heads $S \in \mathcal{H}$. Then we define the function:

$$f(S) \triangleq \text{Accuracy of LLM}_S \text{ on MMLU training set.}$$



Attention head pruning for `Llama-3.1-8B-Instruct` for MMLU (high-school-us-history) across different layers. Unpruned accuracy shown by dashed line. The faithfulness of ProxySPEX leads to superior performance.

Uniquely, ProxySPEX can be used to measure the amount of interaction energy between heads, both within and across layers, offering exiting new tools for understanding computation in LLMs.