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

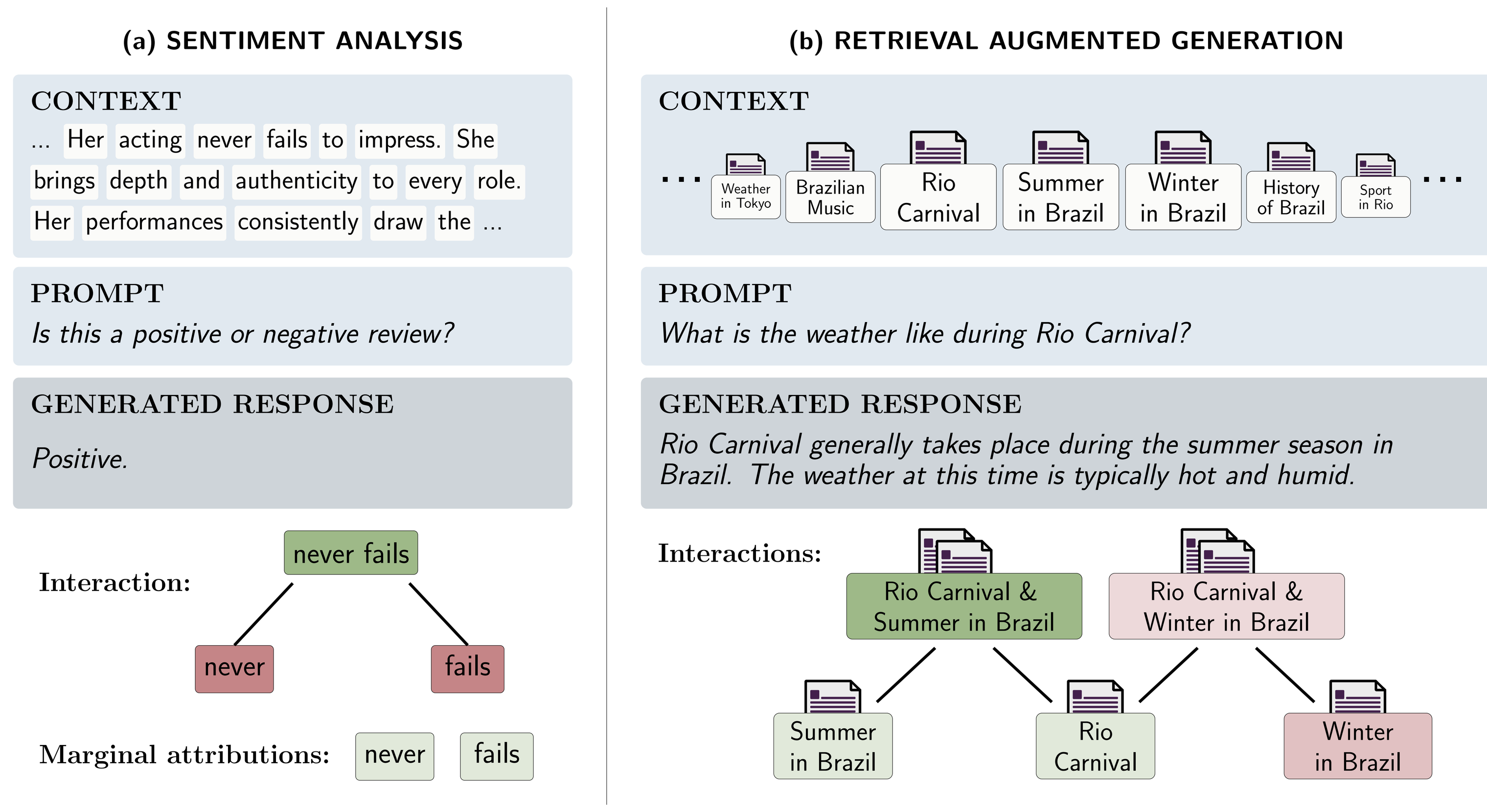
L. Butler* A. Agarwal* J. S. Kang* Y. E. Erginbas B. Yu K. Ramchandran

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley UNIVERSITY OF CALIFORNIA

Problem

How can we **efficiently identify** the **influential feature interactions** in LLMs?



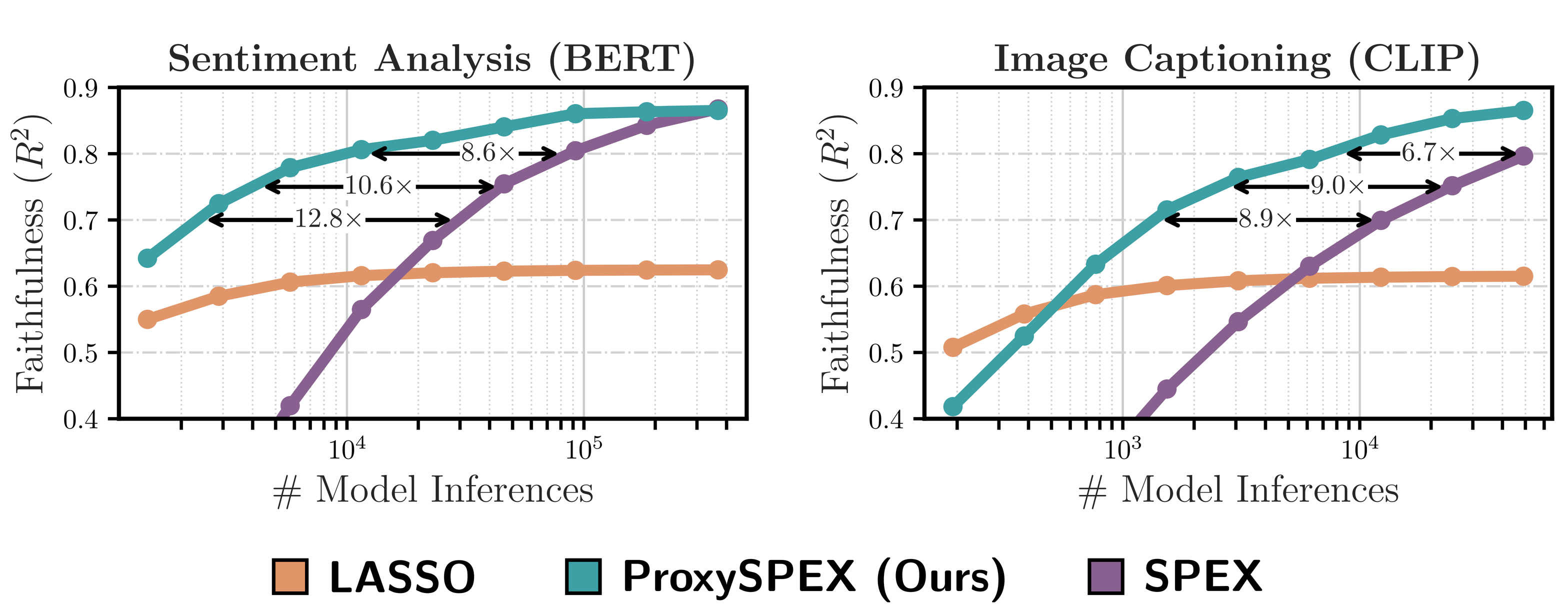
- **Examples:** double negatives in sentiment analysis tasks and multi-document understanding in question answering tasks.
- Marginal attribution approaches like SHAP/LIME scale, but don't capture important interactions.
- A prior approach (SPEX) scales, but still requires tens of thousands of model inferences, which can be prohibitive for complex models such as LLMs.

Faithfulness at Scale

- 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". This masking pattern changes the sentiment score from positive to negative.
- We aim to learn an interpretable approximate function $\hat{f}$ that is faithful to the original function f, measured in terms of R^2:

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

Result: ProxySPEX requires **∼10× fewer inferences** to achieve equally faithful explanations as SPEX.



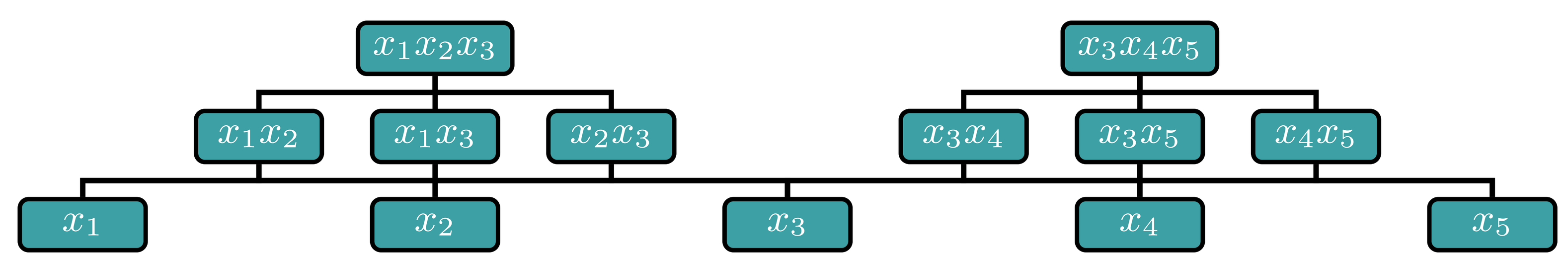
Fourier Sparsity and Spectral Hierarchies

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

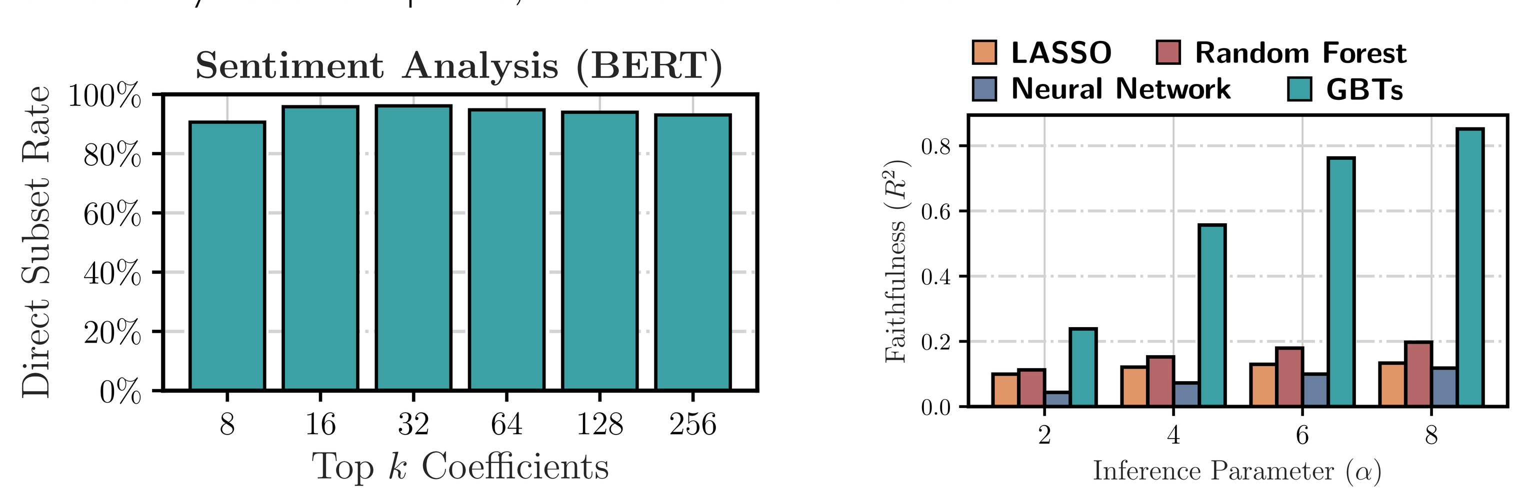
$$F(\mathbf{x}_T) = \frac{1}{2^n}\sum_{S\subseteq[n]} (-1)^{|S\cap T|} f(\mathbf{x}_S), \qquad f(\mathbf{x}_S) = \sum_{T\subseteq[n]} (-1)^{|S\cap T|} F(\mathbf{x}_T).$$

- It has been observed that $F(\mathbf{x}_T) \approx 0$ for most T (**sparsity**), and most large $F(\mathbf{x}_T)$ are **low degree** such that $|T| \leq d$ for some small d.

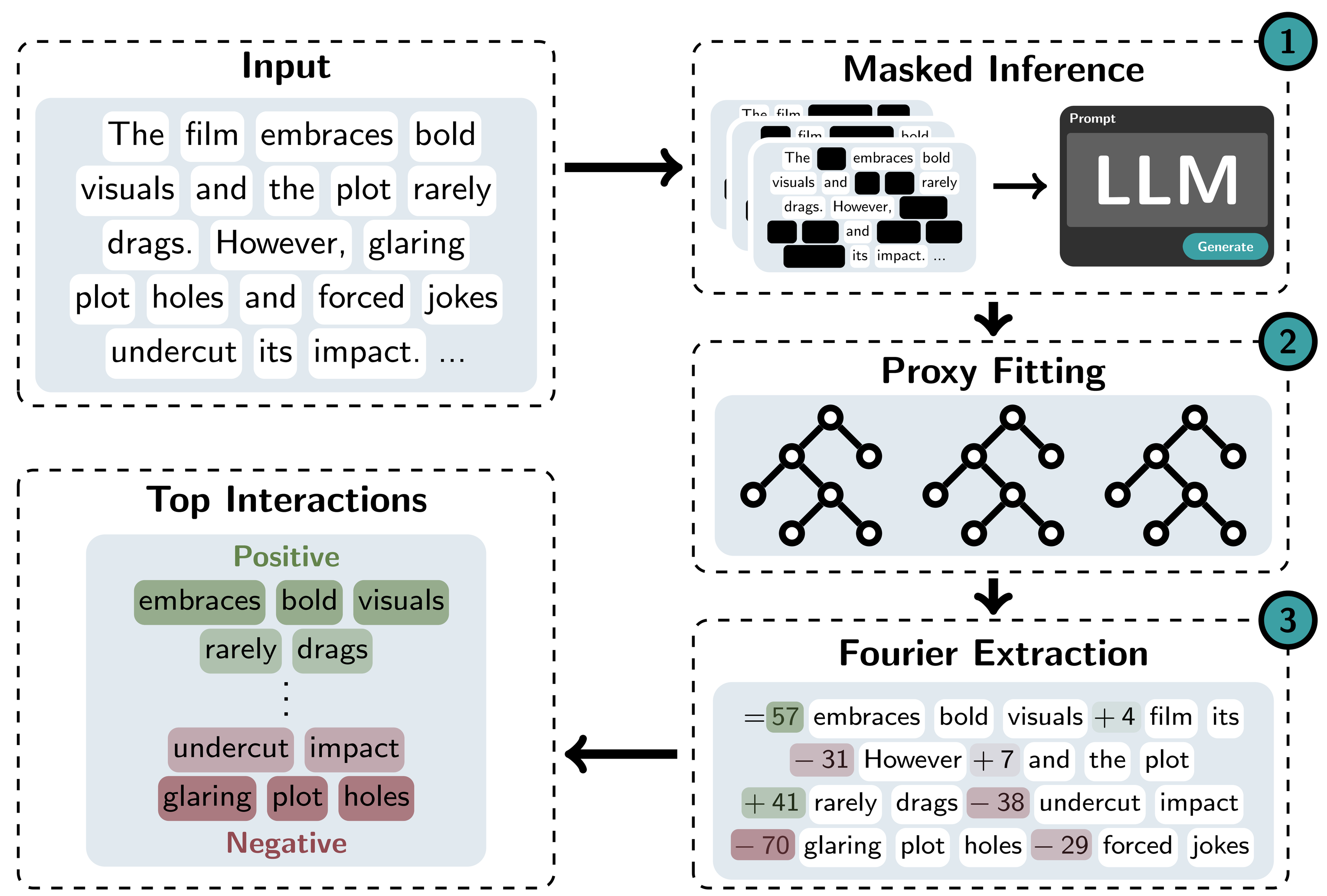
In addition to **sparse** and **low-degree**, influential interactions are **hierarchical**: higher-order interactions are accompanied by their lower-order subsets.



(Left) Direct subset rate measures the rate at which a top-k interaction has a lower-order subset also contained in the top-k. **(Right)** Gradient Boosted Trees efficiently recover sparse, hierarchical interactions.

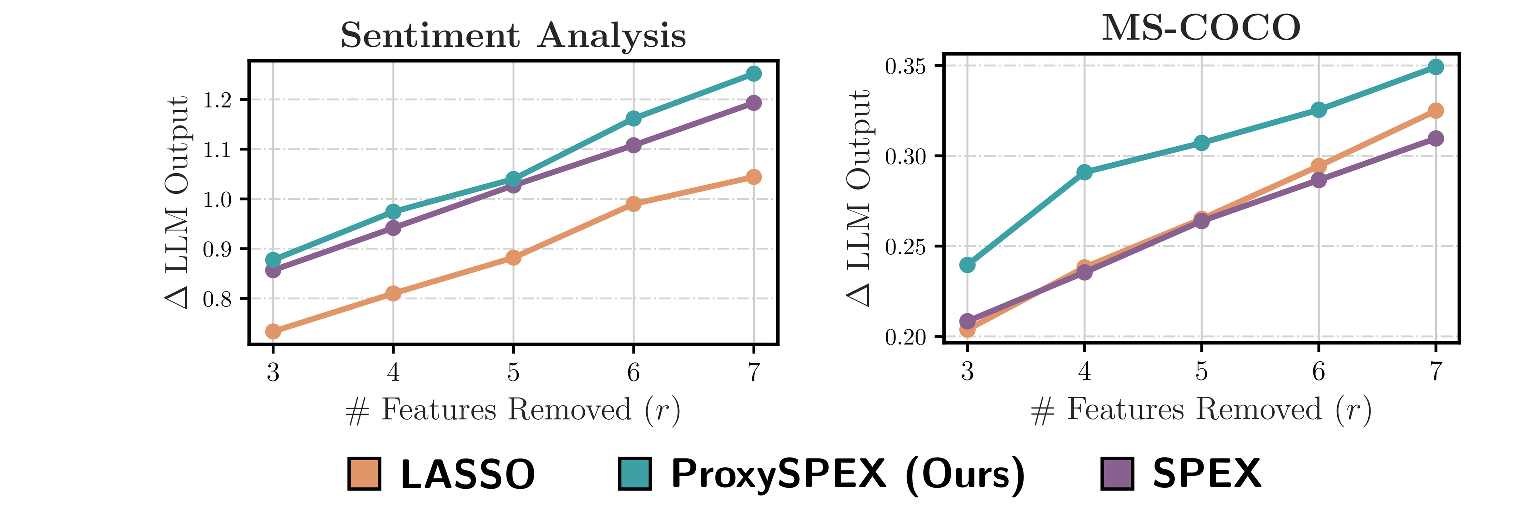


ProxySPEX Algorithm



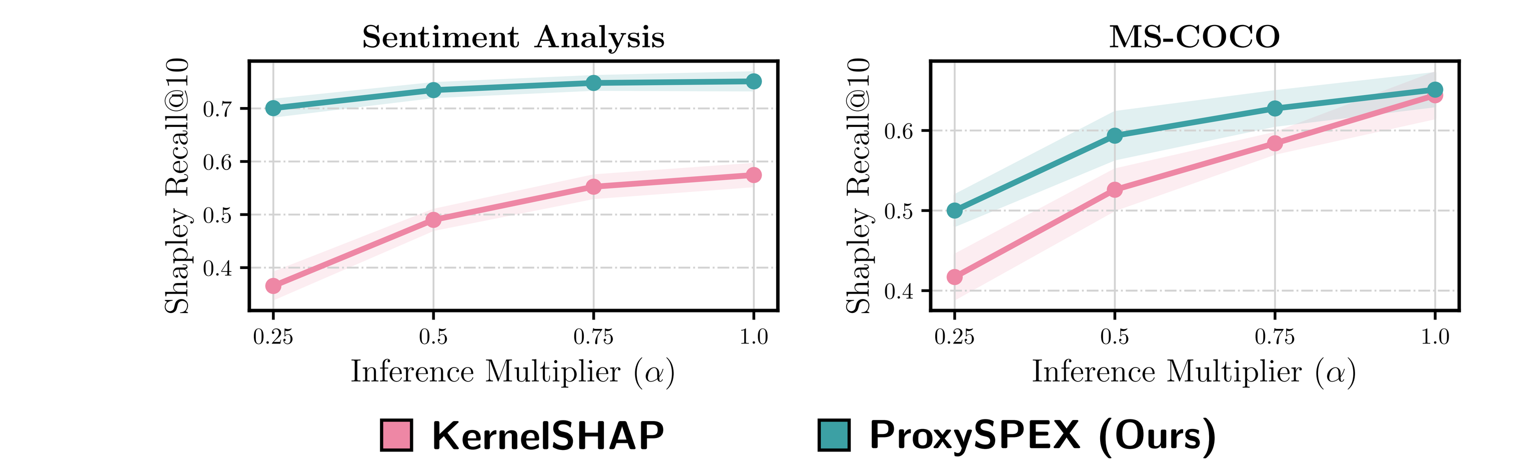
(1) ProxySPEX masks subsets of words and queries the LLM using this masked input. (2) It then fits Gradient Boosted Trees as a proxy model to learn the LLM's hierarchical interactions. (3) A sparse representation is extracted from the fitted GBTs, capturing influential interactions.

Feature Removal



By accounting for interactions, ProxySPEX identifies more influential features across datasets than the LASSO and SPEX.

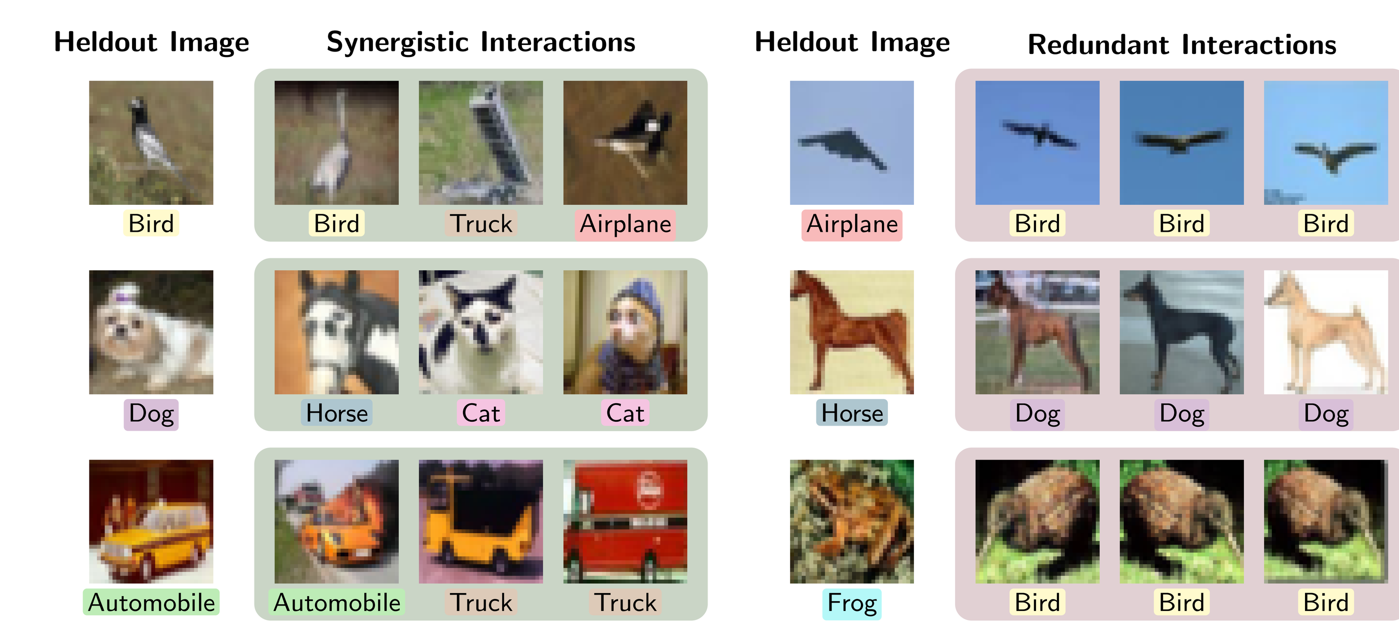
Sample-Efficient Shapley Estimation



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. For small α, ProxySPEX is **superior at recovering the most significant features**, while KernalSHAP outperforms as α increases.

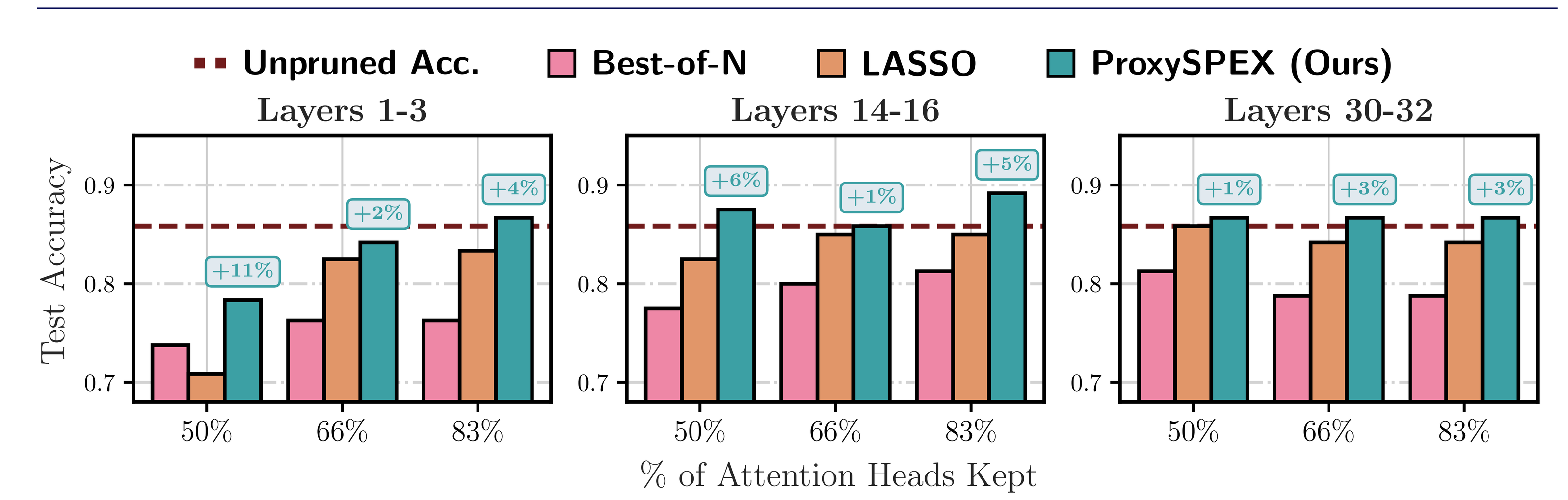
Data Interaction Attribution

Data attribution measures how each training sample influences the prediction of a particular test point. We extend this framework to capture interactions.



Synergistic interactions: data that together are more valuable together than the sum of their parts. **Redundant interactions:** Combined influence is less than the sum of the parts.

Attention Head Interaction Attribution



Attention head pruning for `Llama-3.1-8B-Instruct` for MMLU (high-school-us-history) across different layers. Unpruned accuracy shown by dashed line.

Thirty-Ninth Annual Conference on Neural Information Processing Systems, December 2025