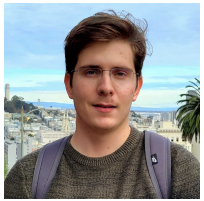
Faithful and Scalable Explanations for LLMs

Justin Singh Kang
PhD Candidate
UC Berkeley, Department of EECS

ITA 2026

Collaborators

SPEX: Scaling Feature Interaction Explanations for LLMs - *ICML 2025*
ProxySPEX: Inference-Efficient Interpretability via Sparse Feature Interactions in LLMs - *NeurIPS 2025 Spotlight*



STATISTICS
UCSB

A Positive Case for Faithfulness: LLM Self-Explanations Help Predict Model Behavior



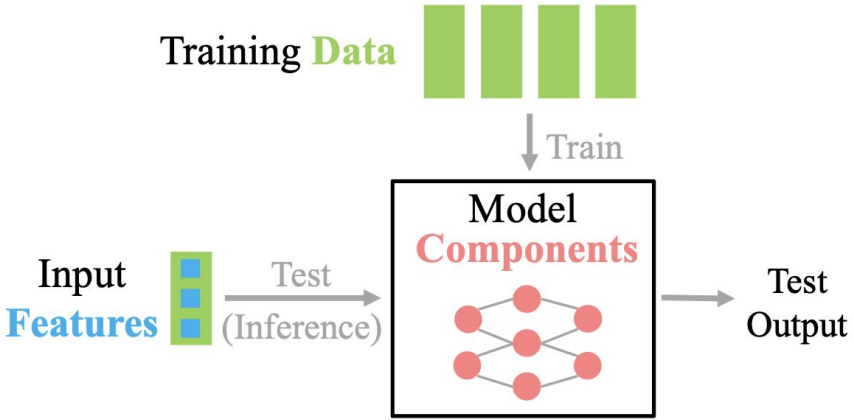

UNIVERSITY OF OXFORD
UCL

Understanding Machine Learning Systems

Feature Attribution: Which features in an input (e.g., images, words) are the most important for the task at hand?

Data Attribution: Which training data points are the most important for the task?

Model Component Attribution: What part of the model is most important for solving this problem?

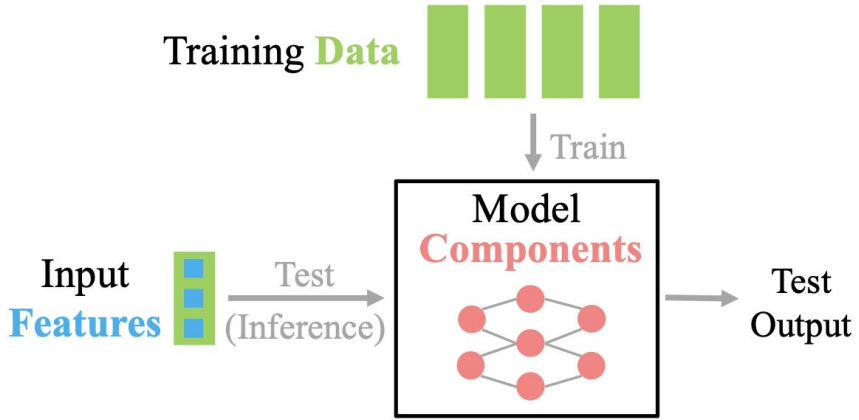


Understanding Machine Learning Systems

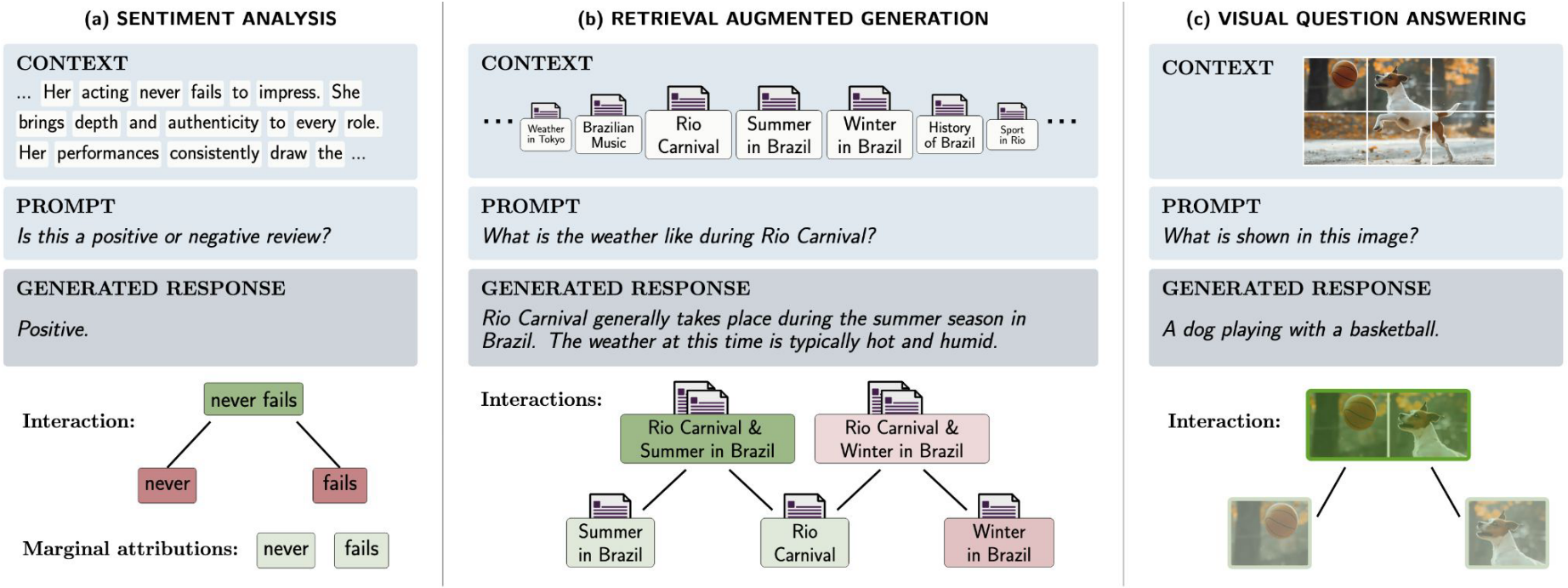
Feature Attribution: **Which features in an input (e.g., images, words) are the most important for the task at hand? With focus on feature interactions.**

Data Attribution: Which training data points are the most important for the task?

Model Component Attribution: What part of the model is most important for solving this problem?



Feature Interactions



Examples of Learned Interactions

A little boy plays with a Nintendo GameCube controller inside a McDonald's

Caption

Examples of Learned Interactions

A little boy plays with a Nintendo GameCube controller inside a McDonald's

Caption

Examples of Learned Interactions

A little boy plays with a Nintendo GameCube controller inside a McDonald's

Caption

Introducing SPEX: A new way to learn feature interactions

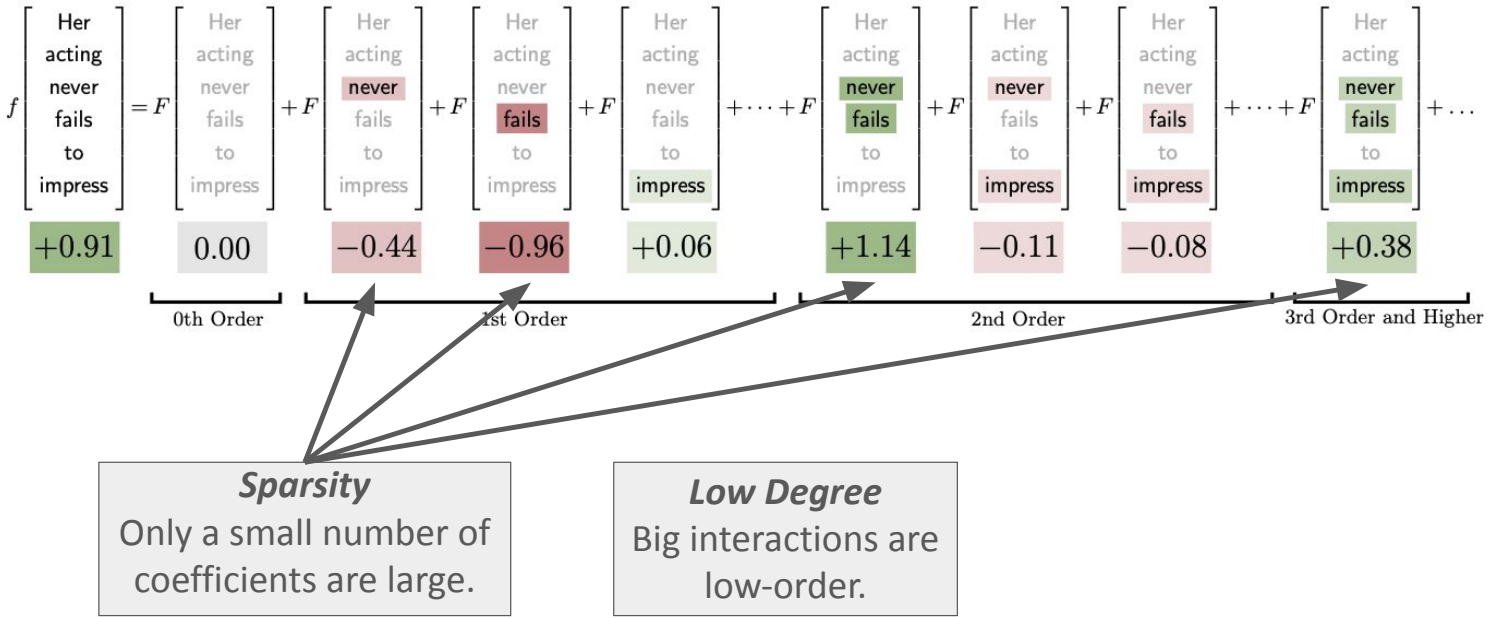
basics-lab.github.io/spex-webapp/

How does SPEX work?

SPEX views feature attribution as equivalent to learning a **Sparse Fourier Transform (Walsh-Hadamard Transform)**.

Decomposing the function (model) in terms of **interactions.**

Learns interactions from masked queries.

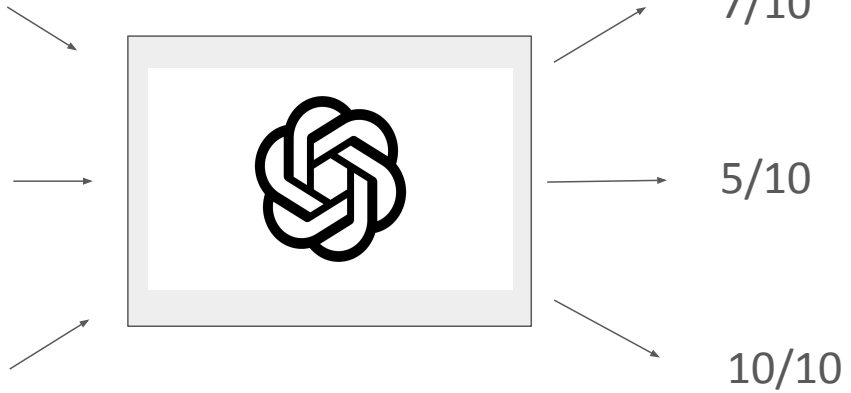


SPEX: Mask and Observe Changes

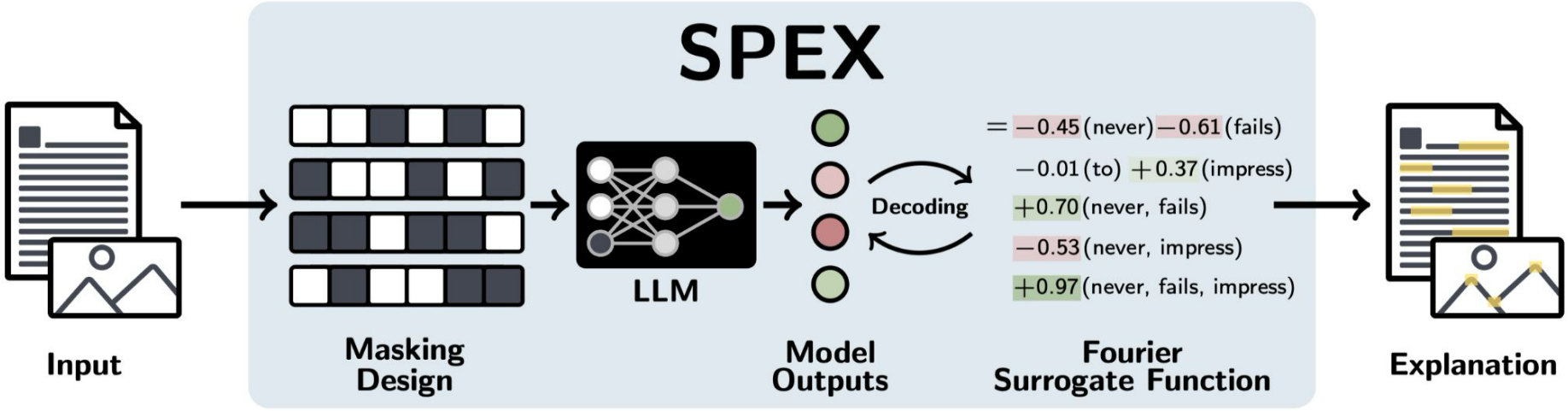
The concept of information [redacted] by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication" and is also referred to as Shannon entropy. [redacted] communication [redacted] source of data, a communication channel, and a receiver. The "fundamental problem of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was [redacted] …

[redacted] was introduced by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication" [redacted] of three elements: a source of data, a communication channel, and a receiver. The [redacted] of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was generated by the source …

The concept [redacted] "A Mathematical Theory of Communication" and is also referred to as Shannon entropy. [redacted] data communication system composed of three elements: a source of data, a communication channel, and a receiver. The "fundamental problem of communication" – as expressed by S[redacted] – is for the receiver to be able to identify what data was generated by the source …



Overview



Step 1: Design masking pattern*

Step 2: Learn Fourier transform using message passing and BCH decoding for efficiency

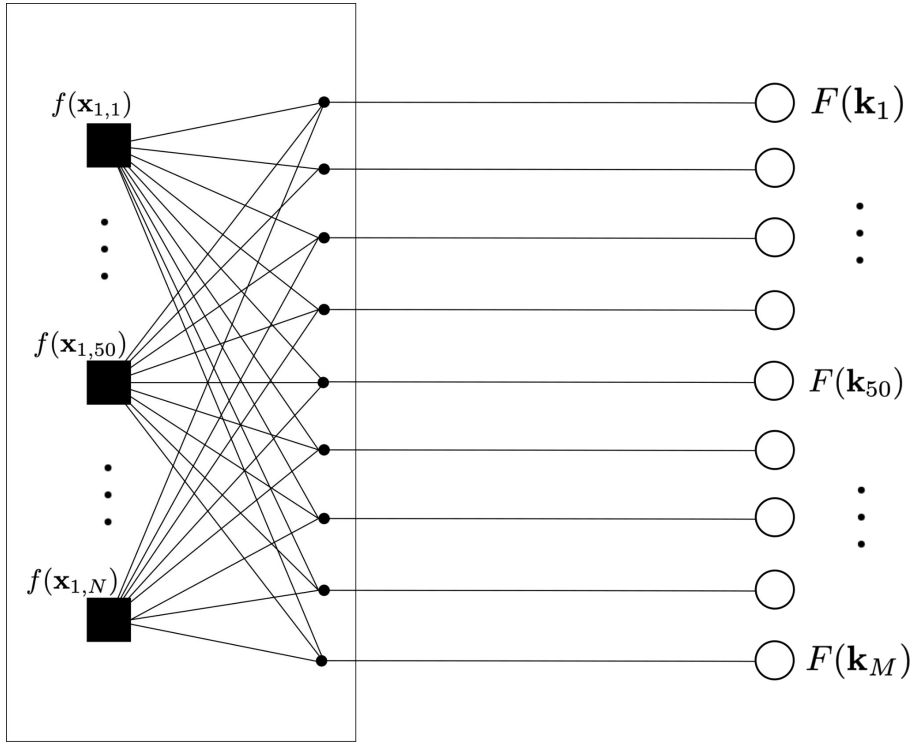
*Masking design based on BCH and LDPC codes

Fourier transform formulation - sampling intuition

Observations from each mask

Sparse unknown interactions, **dense connections**

$$f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{F}_2^n} F(\mathbf{k})(-1)^{\langle \mathbf{x}, \mathbf{k} \rangle}$$



Fourier transform formulation - sampling intuition

Observations from each mask

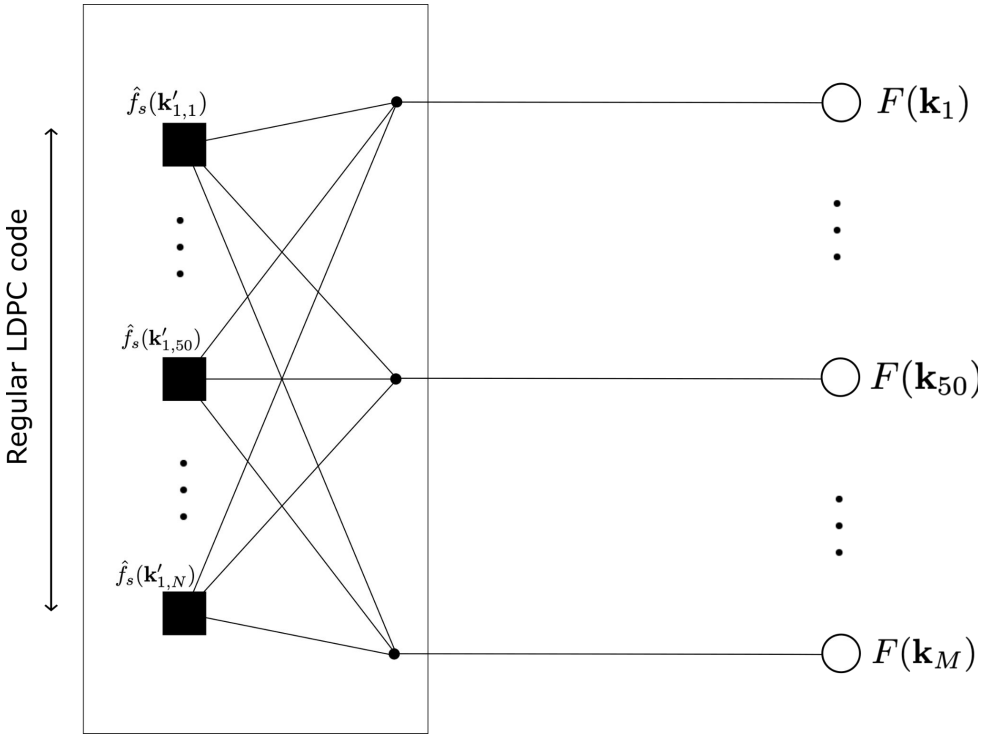
Sparse unknown interactions, **dense connections**

$$f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{F}_2^n} F(\mathbf{k})(-1)^{\langle \mathbf{x}, \mathbf{k} \rangle}$$

By considering **aliasing**, and careful querying we can construct a **sparse graph** [LBP+15].

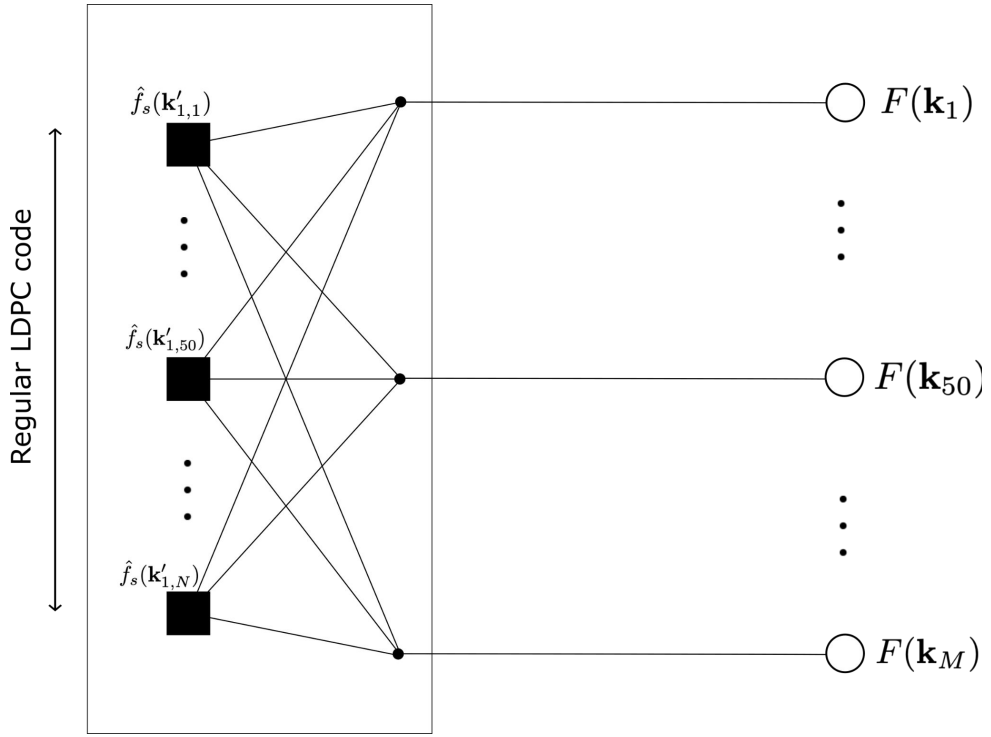
$$f_s(\mathbf{x}') = f(\mathbf{H}^\top \mathbf{x}')$$

$$\hat{f}_s(\mathbf{k}') = \sum_{\mathbf{k} \,:\, \mathbf{H}^\top \mathbf{k} = \mathbf{k}'} F(\mathbf{k})$$



Fourier transform formulation - sampling intuition

Still intractable, number of unknowns exponential in number of features 2^n



Fourier transform formulation - sampling intuition

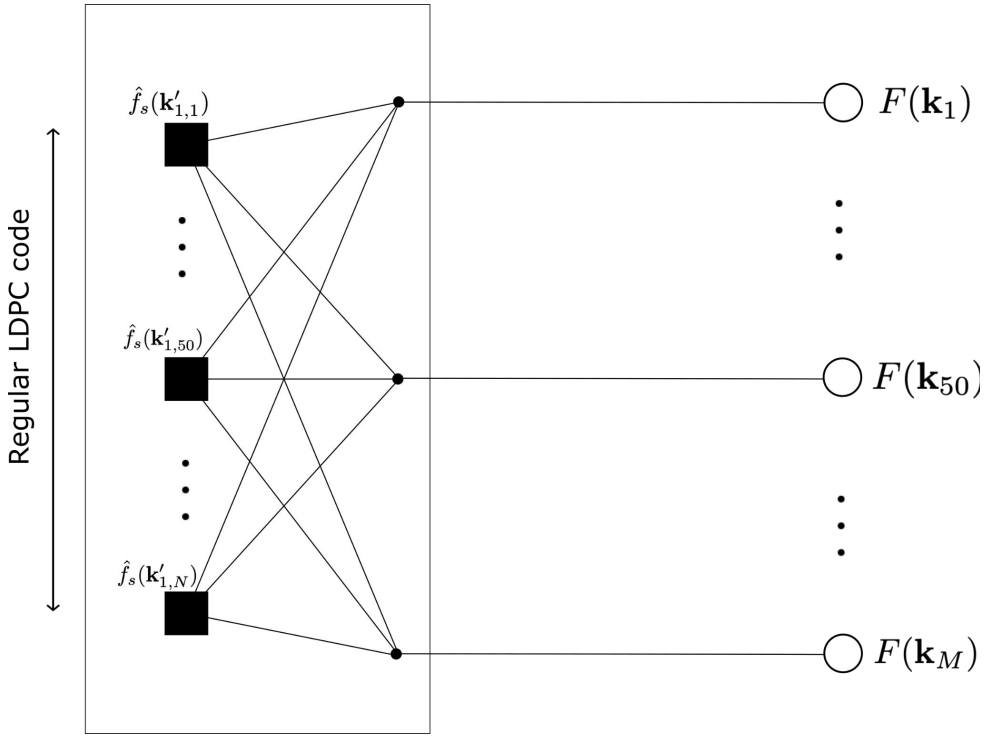
Still intractable, number of unknowns exponential in number of features 2^n

$$f_s(\mathbf{x}') = f(\mathbf{H}^\top \mathbf{x}' + \mathbf{p})$$

$$\hat{f}_s(\mathbf{k}') = \sum_{\mathbf{k} \,:\, \mathbf{H}^\top \mathbf{k} = \mathbf{k}'} F(\mathbf{k})(-1)^{\langle \mathbf{p}, \mathbf{k} \rangle}$$

Same connections

Different modulations



Fourier transform formulation - sampling intuition

Still intractable, number of unknowns exponential in number of features 2^n

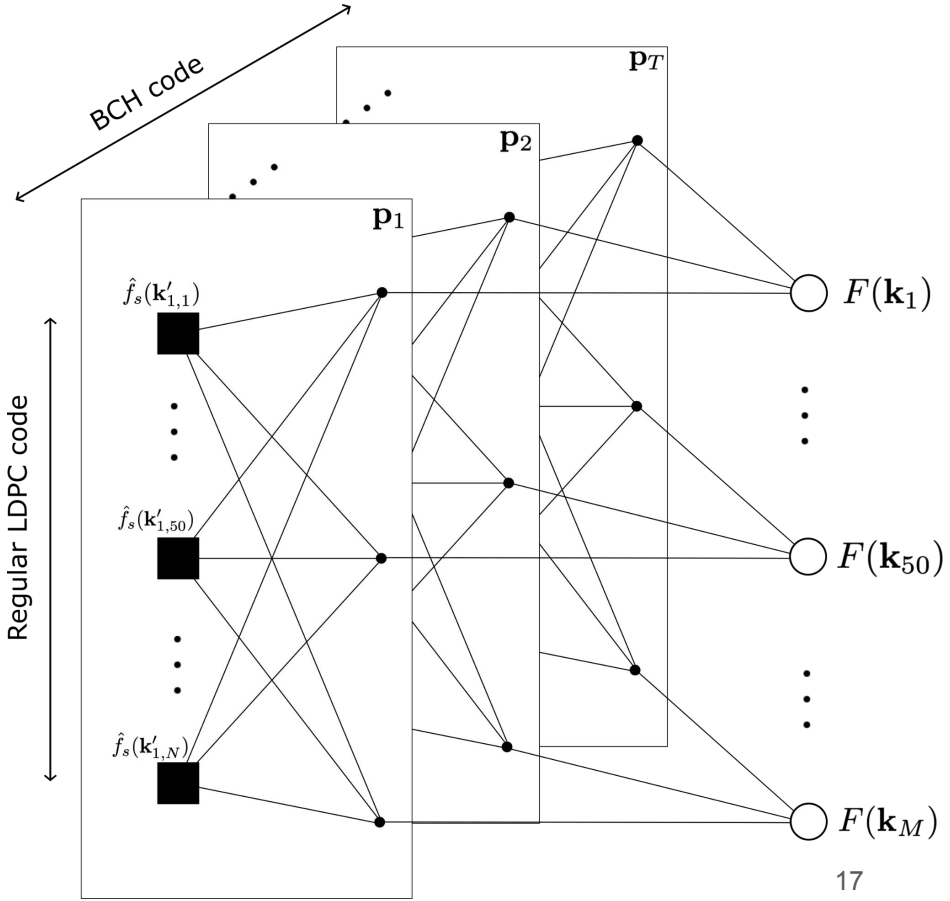
$$f_s(\mathbf{x}') = f(\mathbf{H}^\top \mathbf{x}' + \mathbf{p})$$

$$\hat{f}_s(\mathbf{k}') = \sum_{\mathbf{k} \,:\, \mathbf{H}^\top \mathbf{k} = \mathbf{k}'} F(\mathbf{k}) (-1)^{\langle \mathbf{p}, \mathbf{k} \rangle}$$

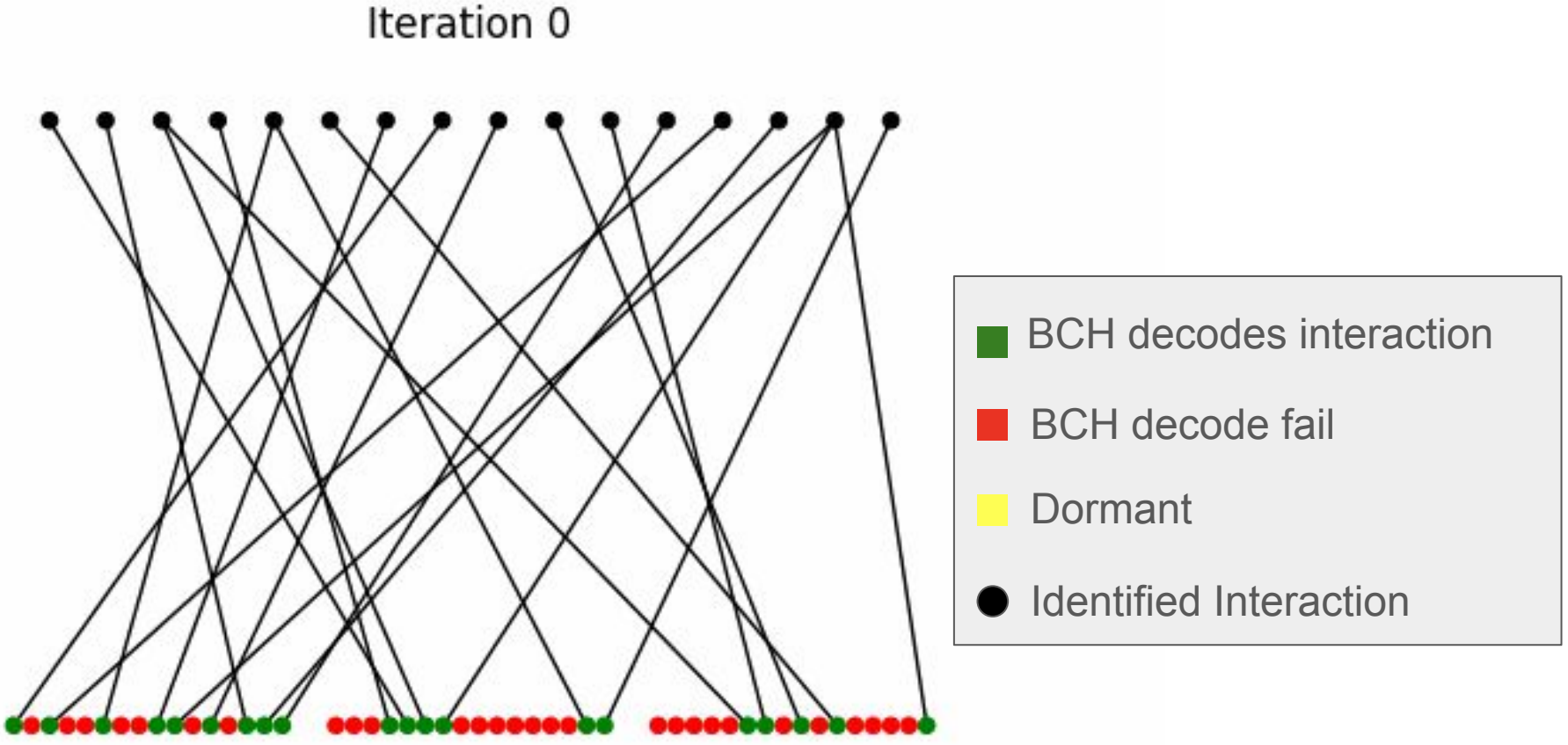
Same connections

Different modulations

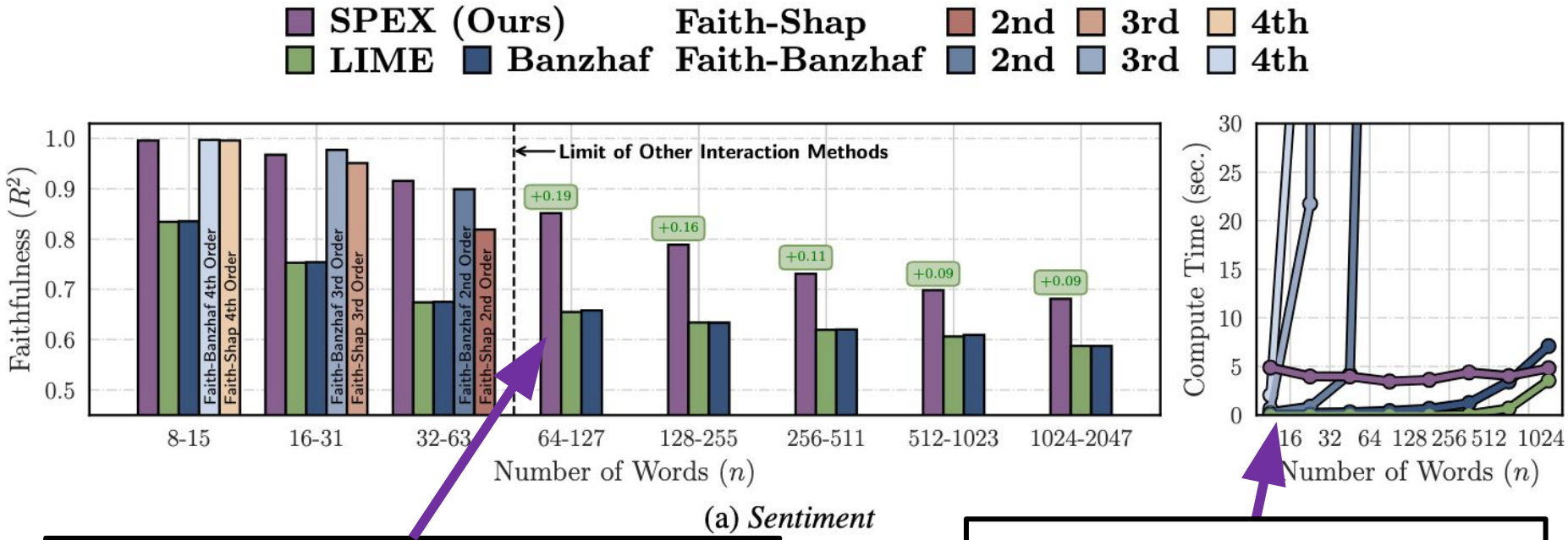
BCH code will tell us which nodes to update (now tractable)



SPEX in action



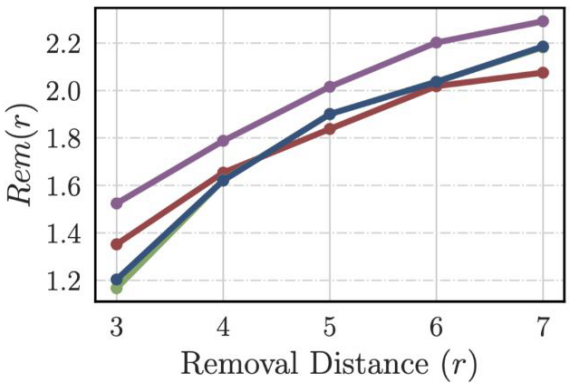
SPEX outperforms other methods - remains tractable



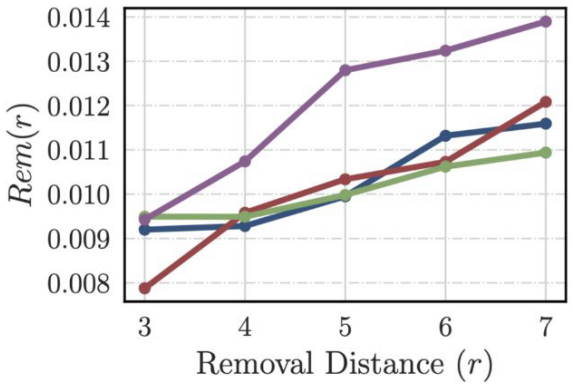
Significant increase in predictive information over first order approaches

Higher order approaches scale poorly

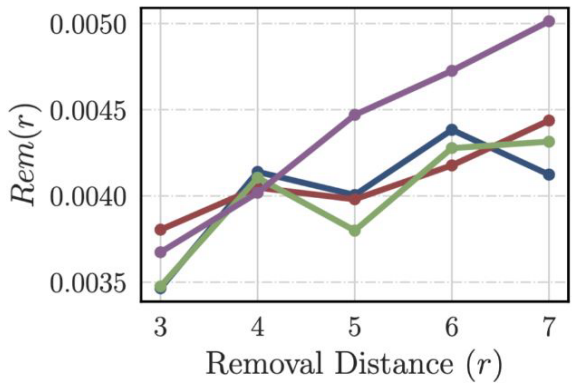
Strong wins in feature selection “Prompt Compression”



(d) *Sentiment* $n \in [64, 127]$



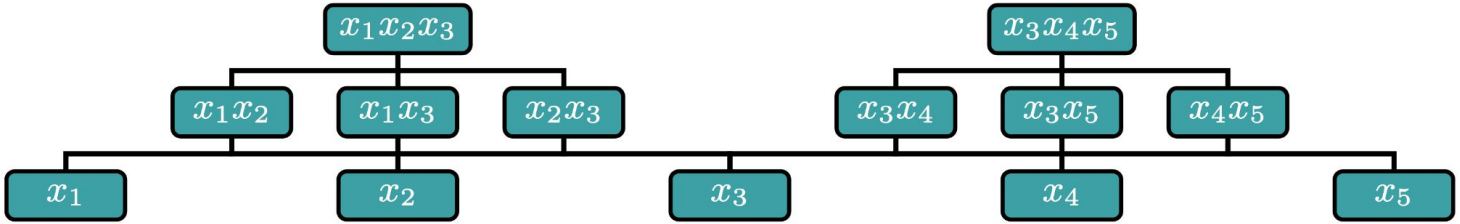
(e) *DROP* $n \in [64, 127]$



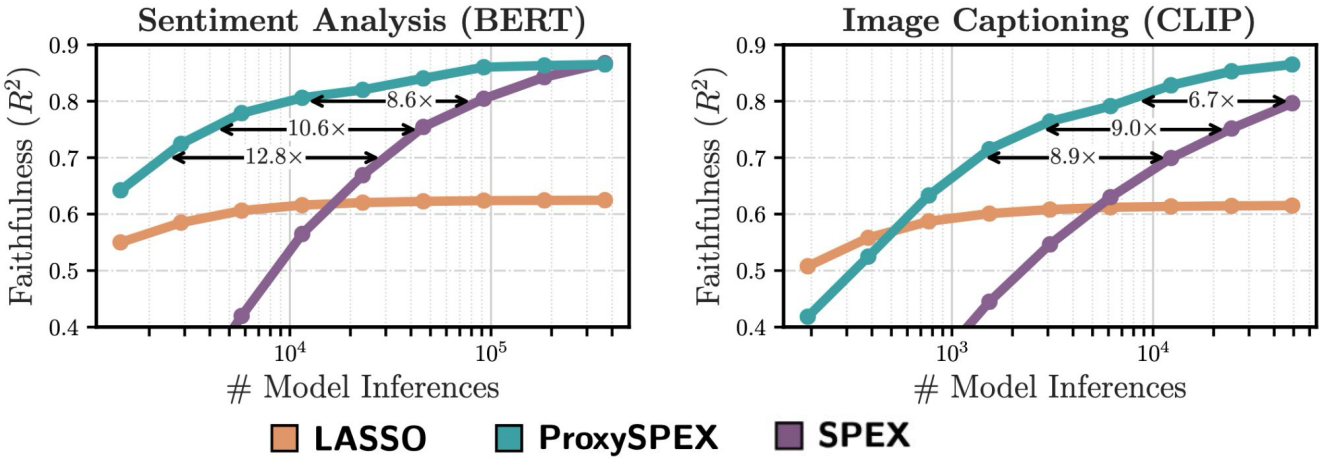
(f) *HotpotQA* $n \in [64, 127]$

ProxySPEX: Interactions are not independent

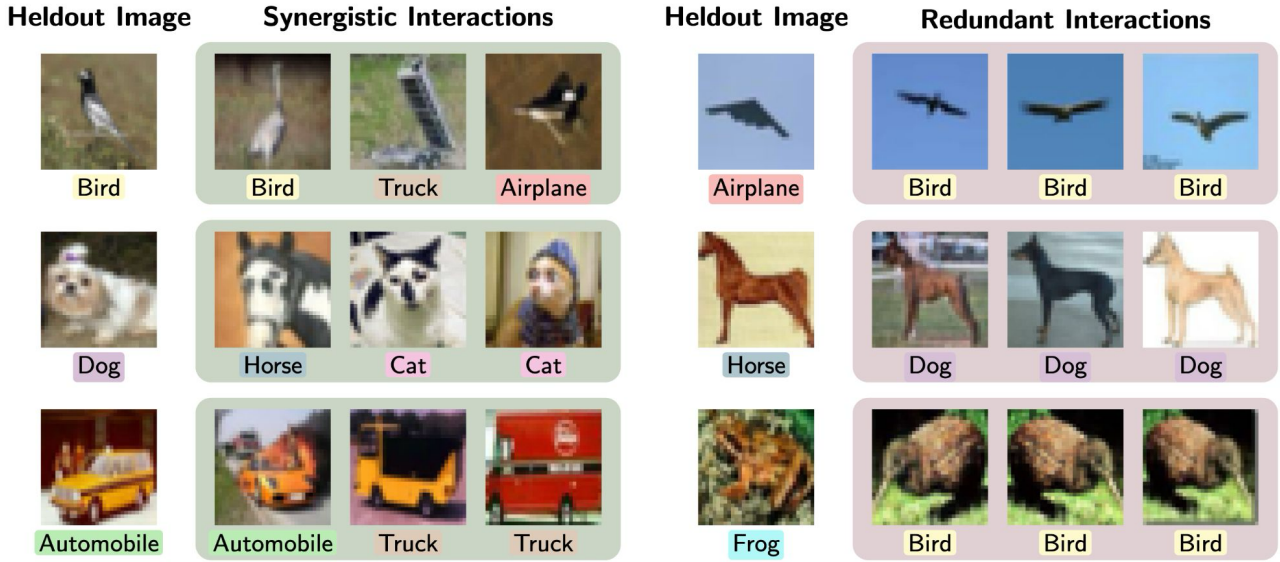
SPEX treats each interaction as independent – not true in many tasks



We leverage a **proxy model** (gradient boosted trees) to capture hierarchy.



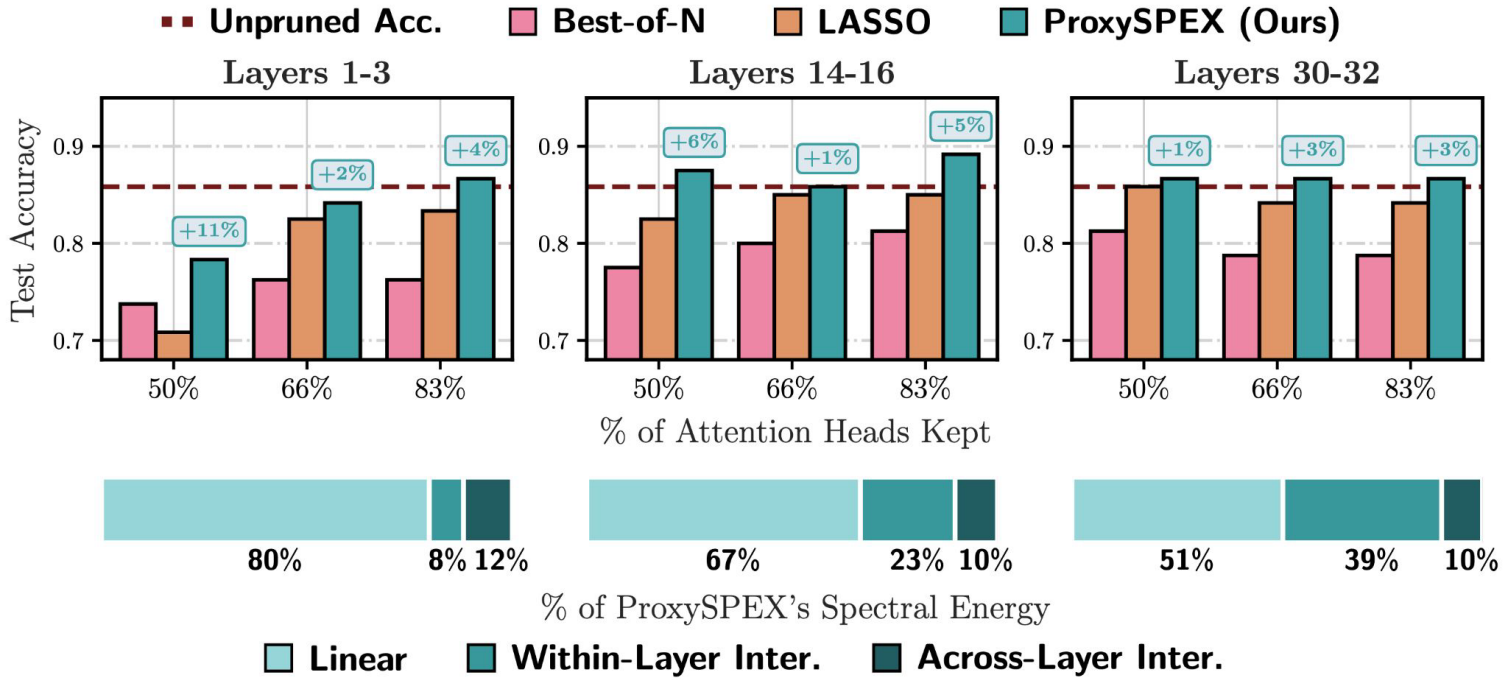
Towards Data Attribution



Synergistic Interactions: Data that shape decision making in a way they couldn't individually. (Whole is more than the sum of parts)

Redundant Interactions: Interchangeable data, and semantically duplicate for the purpose of classification. (Whole is less than the sum of parts)

Mechanistic Interpretability (Attention Head Attribution)



Developing new empirical insights into transformer computation

Access SPEX/ProxySPEX through the SHAP-IQ repository

- Teaming up with researchers at *LMU Munich*, **SPEX is now part the open-source project SHAP-IQ** to enable large scale interaction explanations.
- Quickly generate visualizations that **capture important interactions** for **mission-critical tasks**

SHAP-IQ

682 stars

downloads 128k

Add the SPEX approximator to shapiq (#379)

justinkang221 authored last week · ✓ 8 / 8

Added ProxySPEX to sparse approximators, with additional testing and exampl

landonbutler commit

Tomorrow 10:50AM: Talk by R. Teal Witter

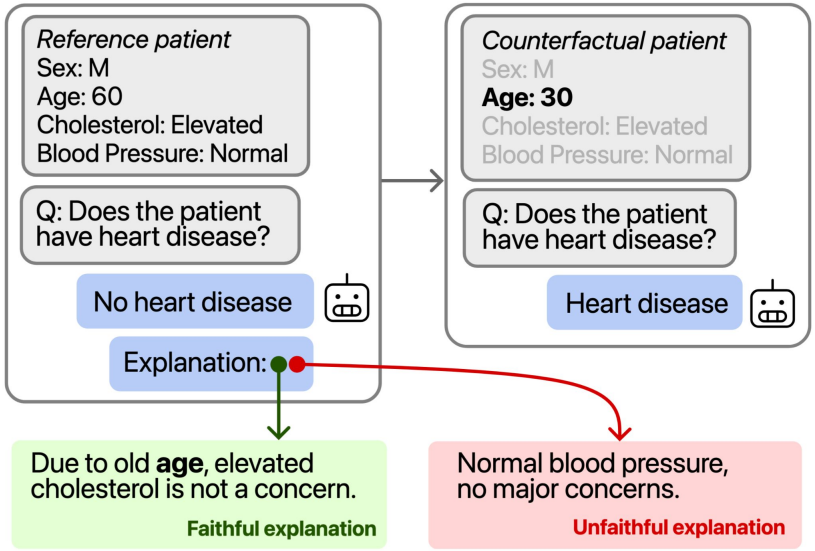
Evaluating the faithfulness of **natural language explanations**

Just ask the model "Why?"

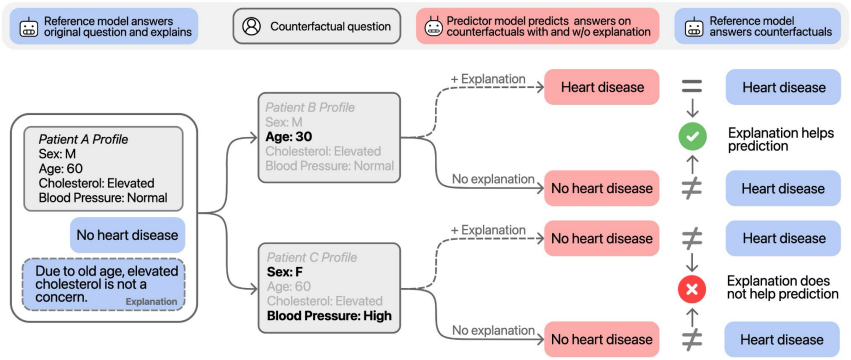
LLMs are trained for ***plausible*** not ***faithful*** explanations**.**

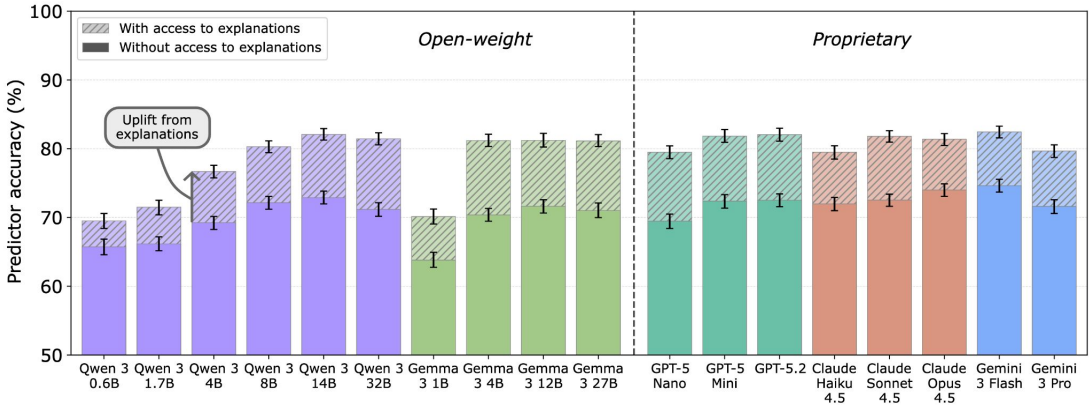
"Unfortunately, we do not currently have viable dedicated evaluations for faithfulness"

-Anthropic, Sonnet 4.5 Model Card



Introducing Simulatability Gain: Agentic Predictive Framework for Faithfulness





Model	Egregious unfaithfulness (%)
Qwen 3 0.6B	15.1
Qwen 3 32B	7.4
Gemma 3 1B	12.9
Gemma 3 27B	6.2
GPT-5 Nano	8.4
GPT-5.2	7.7
Claude Haiku 4.5	8.8
Claude Opus 4.5	6.4
Gemini 3 Flash	5.8
Gemini 3 Pro	7.8

[MKG+26] A Positive Case for Faithfulness: LLM Self-Explanations Help Predict Model Behavior

Concluding Remarks

- SPEX/ProxySPEX scales to large input spaces, and captures interactions.
- Codes can play a central role in efficient learning.
- Efficiency is critical for massive scale like data and model components.
- Natural Language Explanations *should be faithful*, but more work is needed.
- Sparse transforms also very interesting for theoretical results, see [E**K**P+26].

An alternative basis

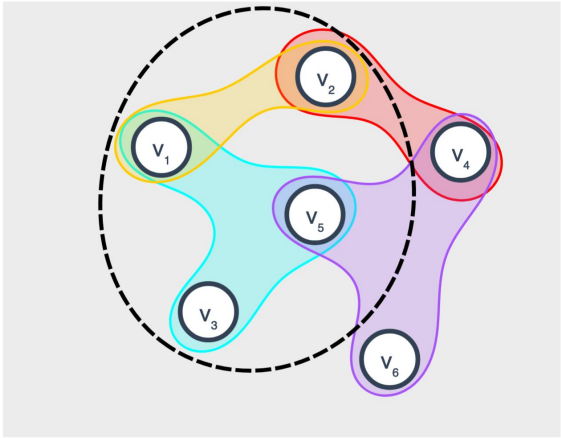
- **Fourier transform** is a basis over **Parity functions.**
- **Mobius transform** is a basis over **AND functions.**

[**K**BE+24]:

Fourier transform ⟷ *codes over the binary fields*
Mobius transform ⟷ *binary group testing designs*.

[E**K**PR26]:

An *adaptive* framework, **solves open question about complexity of *hypergraph reconstruction*** from edge counting oracle.



[EKPR26] *"Adaptive Sparse Möbius Transforms for Learning Polynomials", https://arxiv.org/pdf/2602.06246*