STAT 992: Science of Large Language Models

Lecture 5: Linear representation hypothesis, feature superposition

Spring 2026
Yiqiao Zhong

Overview of main findings

- **Linear representation hypothesis:** transformers represent concepts as low-dim linear subspaces (esp. vectors) in the hidden states space

- **Feature superposition**: hidden states are approximately a sparse linear combination of base concepts, e.g.,

$$\text{"apple"} = 0.09 \text{ "dessert"} + 0.11 \text{ "organism"} + 0.16 \text{ "fruit"} + 0.22 \text{ "mobile\&IT"} + 0.42 \text{ "others"}.$$

- Hidden states can encode richer and more contextualized concepts

Table edited by Gemini-3

Isn't that familiar?

- **Sparse coding:** input vectors are a linear representation of basis vectors

$$\boldsymbol{x} = \sum_{j=1}^{K} a_j \boldsymbol{\varphi}_j$$

  - Complete basis: PCA
  - Over-complete basis: dictionary learning

- **What's new: efficient representation learning.**
  - Semantic-rich dictionary through many layers and large context
  - Scalable training: massive dataset and model size

LRH in pre-transformer age

PCA and factor models

- **SVD and spectral decomposition:** Given a data matrix X, calculate the singular value decomposition, or equivalent spectral decomposition

$$X = U\Sigma V^\top, \qquad X^\top X = V\Sigma^\top \Sigma V^\top$$

- **PCA gives best low-rank approximation.** Using top-r singular vectors and singular values, $\hat{X} = U_{\leq r}\Sigma_{\leq r}V_{\leq r}^\top$ [solves](#)

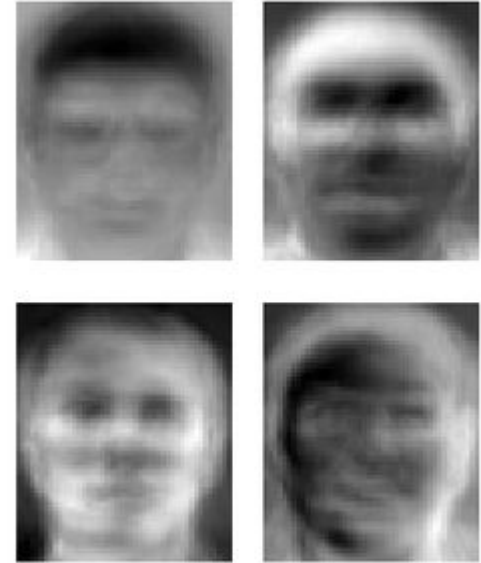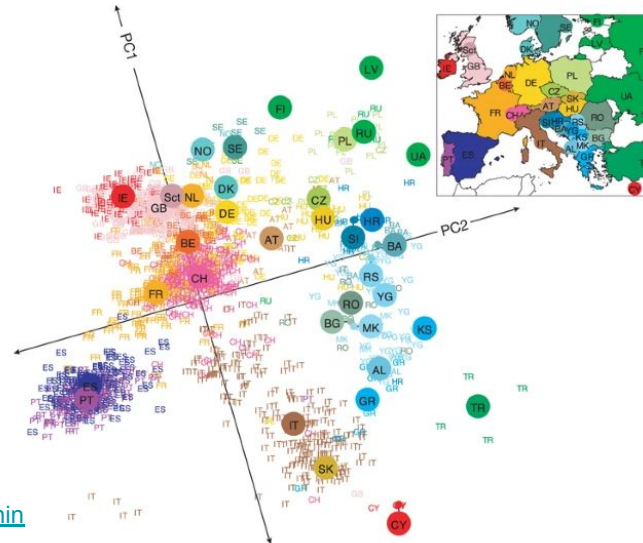$$\left\| X - \hat{X} \right\|_F^2, \quad \text{s.t. } \operatorname{rank}(\hat{X}) \leq r$$

- **Factor model** interpretation: each row is linear combination of a few dominant factor vectors

Classical data analysis

- [Eigenface](#): decompose a face image as a linear combinations

$$\text{Face image}_1 = (23\% \text{ of } E_1) + (2\% \text{ of } E_2) + (51\% \text{ of } E_3) + \dots + (1\% \; E_n).$$

- **Gene expression data analysis**: principal components of gene data mirror geography

[Genes mirror geography within Europe](#), Nature, 2026

From [Wiki](#): Some eigenfaces from AT&T Laboratories Cambridge
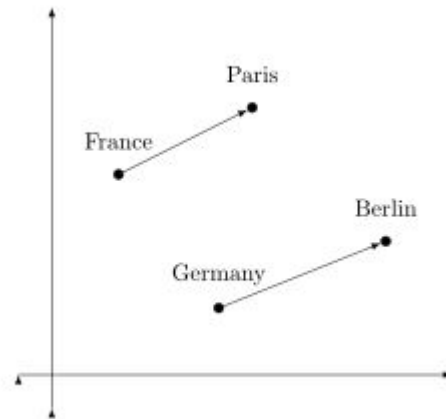
Linear representation and low-rankness

- Low-rank structures underlies interpretable linear features
- Spectral method in broader applications: network data analysis

- Example: stochastic block model (SBM)
  - Connectivity prob within blocks higher
  - In expectation, adjacency matrix is a rank-2 matrix
  - Applying spectral decomposition on one observed matrix
  - Top-2 eigenvectors encode "membership" of nodes

Assortative Case
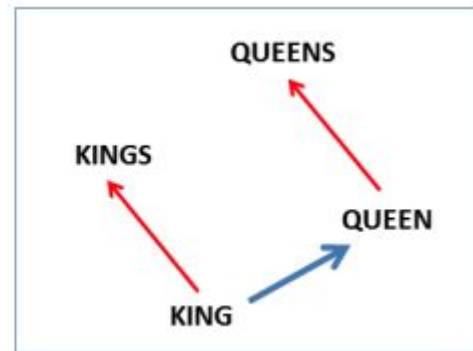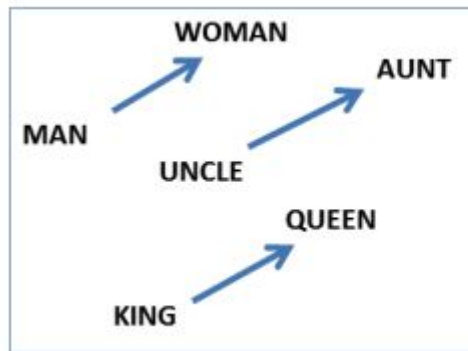
From Wiki: SBM with two blocks

Word embedding

- **Aim**: find embeddings (vector representations) of words / tokens

- Overcoming **prior difficulty**: n-gram models and hidden markov chains not aimed at capturing token semantics

- A transition point in NLP: vector representations are effective for modeling discrete sequence data.
  - Solves polysemy
  - Ideal for neural networks

From Wiki: word embedding

Word embedding

- Similar ideas developed by several groups (2013–2014)
  - Word2vec
  - Glove
- Glove: simple nonlinear matrix factorization finds good word embeddings
  - Co-occurrence matrix: word-word frequency counts within a window
  - Under nonlinear transform, finds low-rank matrix

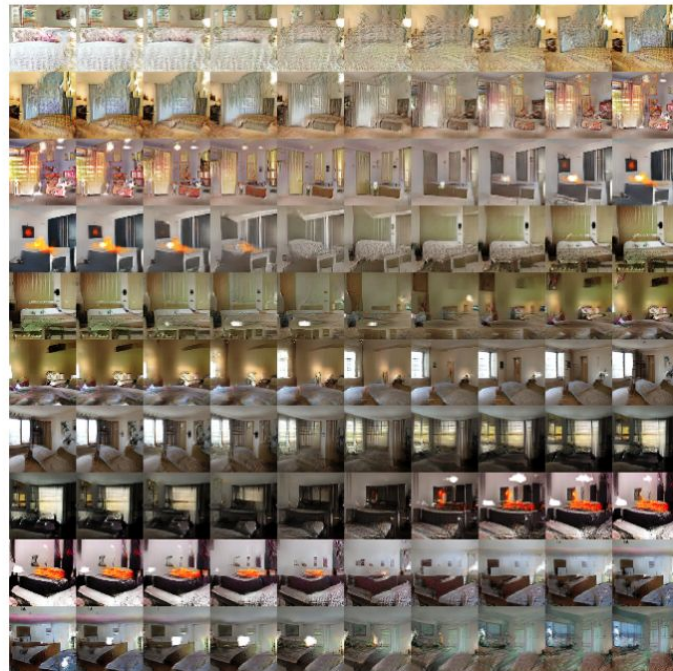- Linear concept representations
  - Composition via additivity

Linguistic Regularities in Continuous Space Word Representations, 2013

LRH are also common for non-language data

- Kernels in CNN are well known to extract hierarchical features
- Generative models such as autoencoders and [GAN](GAN)s encode meaning concepts / scenes / objects linearly in latent space

AlexNet paper, 2012: visualizing first-layer conv kernels

DCGAN paper, 2016: linear interpolation in latent space

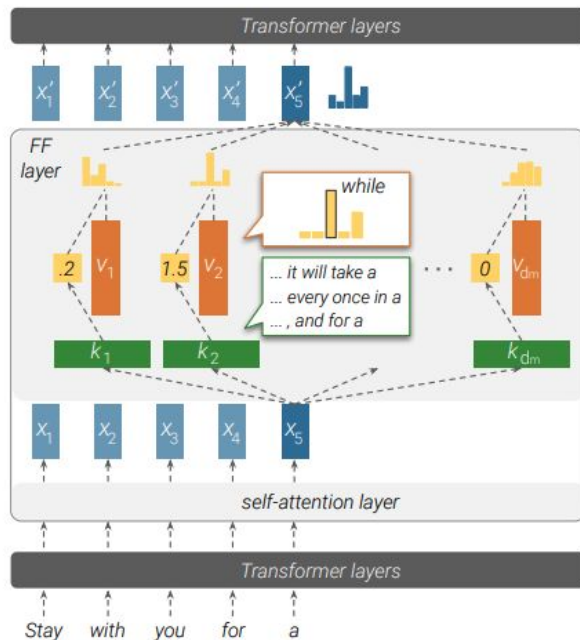# Interpretability of transformers with LRH

Analyzing MLP layers in a transformer

- Transformer Feed-Forward Layers Are Key-Value Memories

- An FFN layer within a transformer is simply a two-layer MLP:

$$\mathrm{FF}(\boldsymbol{x}) = \sum_{j=1}^{D} \sigma(\boldsymbol{x}^\top \boldsymbol{k}_j) \boldsymbol{v}_j$$

- Neural memory interpretation
  - $\boldsymbol{k}_j$ is a key
  - $\boldsymbol{v}_j$ is a value
  - An embedding $\boldsymbol{x}$ matches a key if the inner product is large, then activates the corresponding value

[Transformer Feed-Forward Layers Are Key-Value Memories](#), 2021

Analyzing MLP layers in a transformer

- **Interpreting** value vectors v_i: projection with unembedding matrix
  - **Token embedding** maps tokens to embeddings
  - **Unembedding matrix** (namely final classification weight matrix) maps embeddings to vectors of len of vocab size, which after softmax converted to prob distribution over the vocab
- Top tokens from the prob distribution represents meaning of value vectors
- Interpreting FFN as many "sub-updates" of concepts, additively

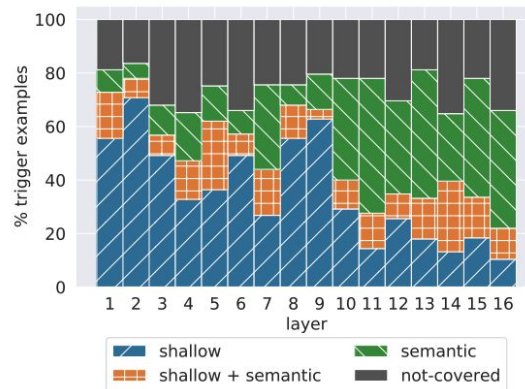$$\mathbf{p}_i^\ell = \text{softmax}(\mathbf{v}_i^\ell \cdot E).$$

		Concept		Sub-update top-scoring tokens
GPT2	$\mathbf{v}_{1018}^3$	Measurement	semantic	kg, percent, spread, total, yards, pounds, hours
	$\mathbf{v}_{1900}^8$	WH-relativizers	syntactic	which, whose, Which, whom, where, who, wherein
	$\mathbf{v}_{2601}^{11}$	Food and drinks	semantic	drinks, coffee, tea, soda, burgers, bar, sushi
WIKILM	$\mathbf{v}_1^1$	Pronouns	syntactic	Her, She, Their, her, she, They, their, they, His
	$\mathbf{v}_{3025}^6$	Adverbs	syntactic	largely, rapidly, effectively, previously, normally
	$\mathbf{v}_{3516}^{13}$	Groups of people	semantic	policymakers, geneticists, ancestries, Ohioans

Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space, 2022

Analyzing MLP layers in a transformer

- Simple strategy: interpreting vectors as top-related tokens
- Often (but not always) useful
- A related strategy: find prompts that activate a key-value pair the most

Key	Pattern	Example trigger prefixes
$\mathbf{k}_{449}^{1}$	Ends with "substitutes" (shallow)	At the meeting, Elton said that "for artistic reasons there could be no substitutes / In German service, they were used as substitutes / Two weeks later, he came off the substitutes
$\mathbf{k}_{2546}^{6}$	Military, ends with "base"/"bases" (shallow + semantic)	On 1 April the SRSG authorised the SADF to leave their bases / Aircraft from all four carriers attacked the Australian base / Bombers flying missions to Rabaul and other Japanese bases
$\mathbf{k}_{2997}^{10}$	a "part of" relation (semantic)	In June 2012 she was named as one of the team that competed / He was also a part of the Indian delegation / Toy Story is also among the top ten in the BFI list of the 50 films you should
$\mathbf{k}_{2989}^{13}$	Ends with a time range (semantic)	Worldwide, most tornadoes occur in the late afternoon, between 3 pm and 7 / Weekend tolls are in effect from 7:00 pm Friday until / The building is open to the public seven days a week, from 11:00 am to
$\mathbf{k}_{1935}^{16}$	TV shows (semantic)	Time shifting viewing added 57 percent to the episode's / The first season set that the episode was included in was as part of the / From the original NBC daytime version, archived

Transformer Feed-Forward Layers Are Key-Value Memories, 2021

How do transformer encode and process semantics?

- **Heuristics of transformers**
  - MLPs / FFNs store <u>static knowledge</u> using key-vector memories, encoding <u>progressively rich</u> semantics in later layers
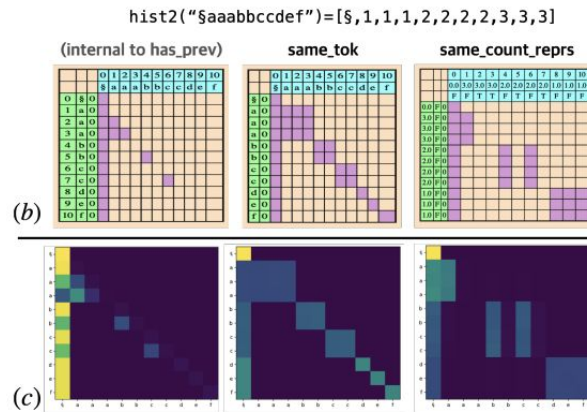 - Self-attentions <u>implements algorithms</u> by mixing and composing token / position information

- **Layer specification**: MLPs and SAs across layers can play different roles
- **Mixture of experts** (MoEs): MLP split into sparsely activated "experts" for knowledge specification

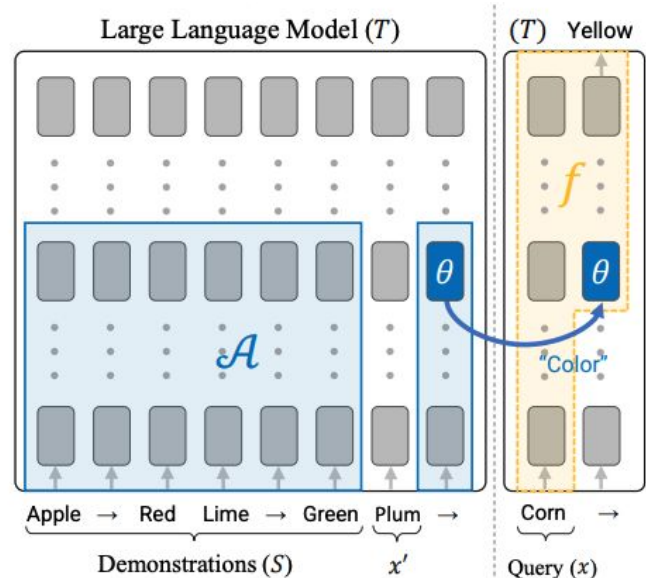Thinking Like Transformers, 2021: implementing programs via attention patterns

LRH and in-context learning

- An interpretation: in-context (IC) examples activate concepts, which steer the model towards that concepts
- **IC vector**: extract hidden states as concept-encoding vector
- **Inject IC vector** without context: patching this vector or adding this vector to hidden states completes task without context
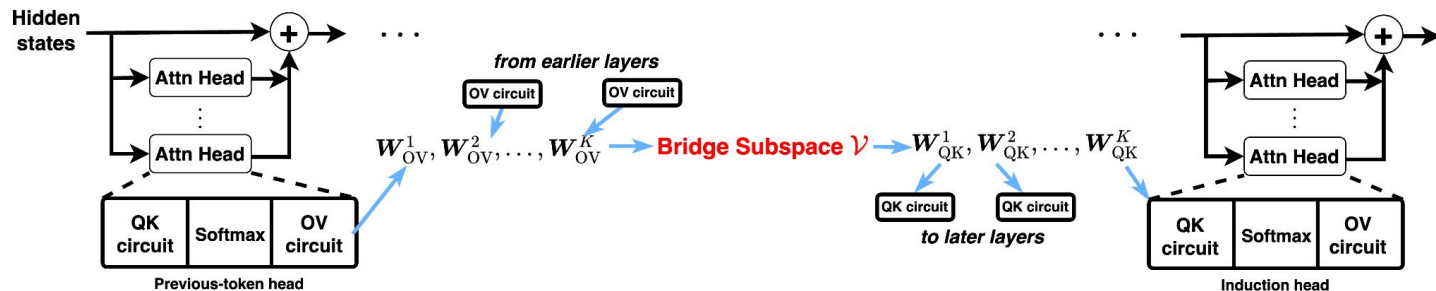

Task	Top tokens in the task vector projection
Previous Letter	e, y, unknown, alphabet, preceding, c Cad, zA, dit, bill
FR-EN	Mason, gram, immer, Santi, latin, utter, Span, Conc, English, equivalent
Present Simple to Gerund	cin, thats, gram, Lorenzo, cian, Isabel, uld, berto, partici, Sah
Country Capital	Paris, its, capital, central, Conc, cities, administrative, Los, Madrid, London

In-Context Learning Creates Task Vectors, 2023

Representation of **compositions** via bridge subspace

- How do transformers compose two layers? How do an earlier layer communicate with a later layer?
- An shared subspace by many pairs of early-layer OV and late-later QK
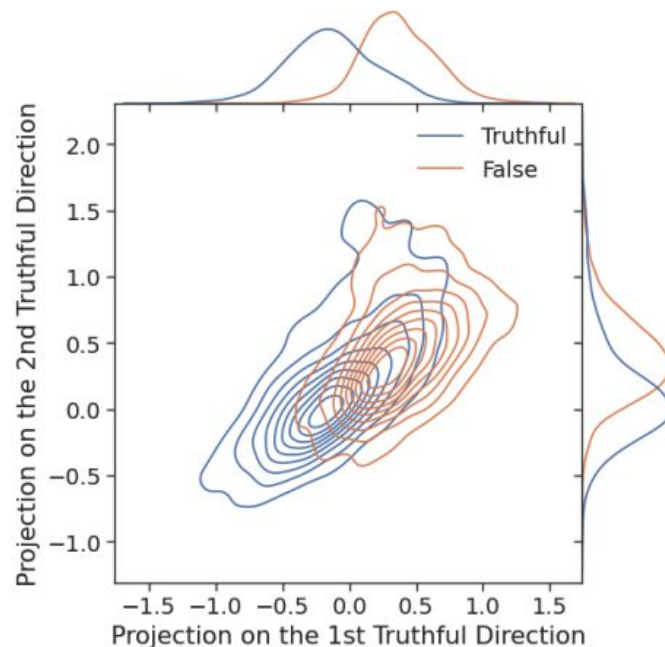- Early layer "writes" in bridge subspace, then "read and processed" by later layers

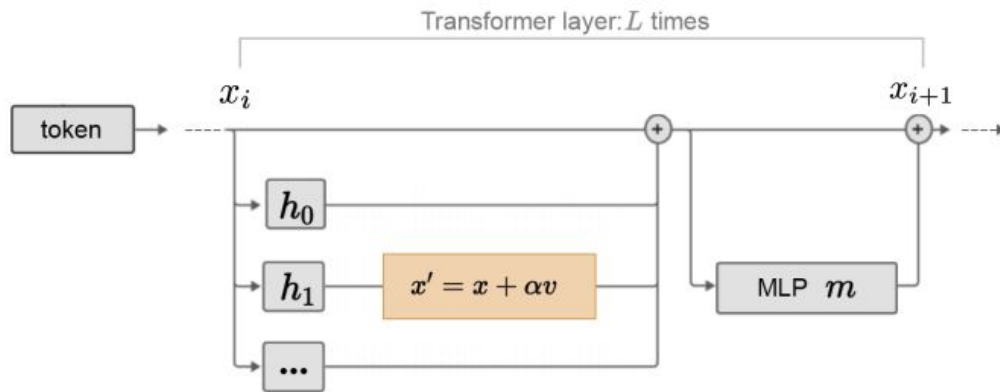$$\mathcal{V} = \mathrm{span}(\boldsymbol{W}_{\mathrm{OV},j}) = \mathrm{span}(\boldsymbol{W}_{\mathrm{QK},k}^{\top})$$

Out-of-distribution generalization via composition: A lens through induction heads in Transformers, 2025
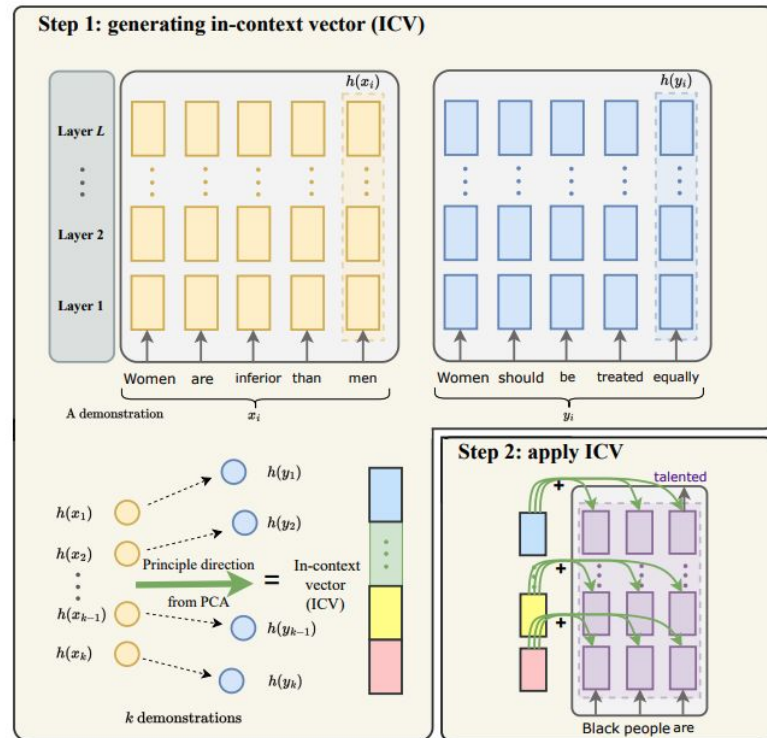
Model adaptation using LRH

Inference-time intervention

- Find harmful (untruthfulness, bias) concept vectors in hidden states space
- Shift the hidden states in the opposite direction

Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, 2024

Steering with IC vectors

- Compute safety-related concept vectors:
  - Prompt with paired examples
  - Calculate top principal component

- Steer the model to reduce harmful-encoding vector

In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering, 2024

Model editing

- Concept / task vectors can be used to edit models (low-cost finetuning)
- Detoxify LLMs: extract hidden states from paired prompts, apply PCA, then project out toxicity-encoding subspaces in MLP value matrices
- Similar variants: task arithmetic, knowledge injection via ROME

	Top Tokens (Layer 14)	Interpretation
μ	, and the - in (" .	Frequent tokens, stopwords
1st svec	s**t f**k ucker b***h slut F**k holes	Toxic tokens
2nd svec	damn really kinda stupid s**t goddamn	Toxic tokens
3rd svec	disclaimer Opinion LĤ Statement Disclaimer Brief	Context dependent topics
4th svec	nation globalization paradigm continent empire ocracy	Context dependent topics

Table 1: Interpreting the top singular vectors of the difference of preference data embeddings. Using GPT-2 and 500 samples from REALTOXICITYPROMPTS, each singular vector of the matrix is interpreted by identifying the top-k tokens it represents. We use the output embedding vector e_j to find top-scoring tokens $j \in \mathcal{V}$ for maximizing $\langle v_i, e_j \rangle$. Tokens have been censored for readability.

Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity, 2024