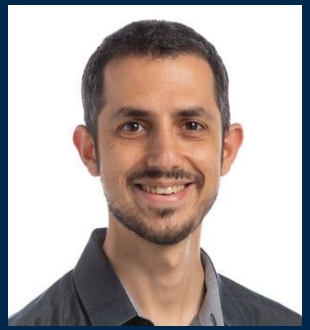
Decomposing QK Space with Contrastive Covariances

Andrew Lee
Harvard



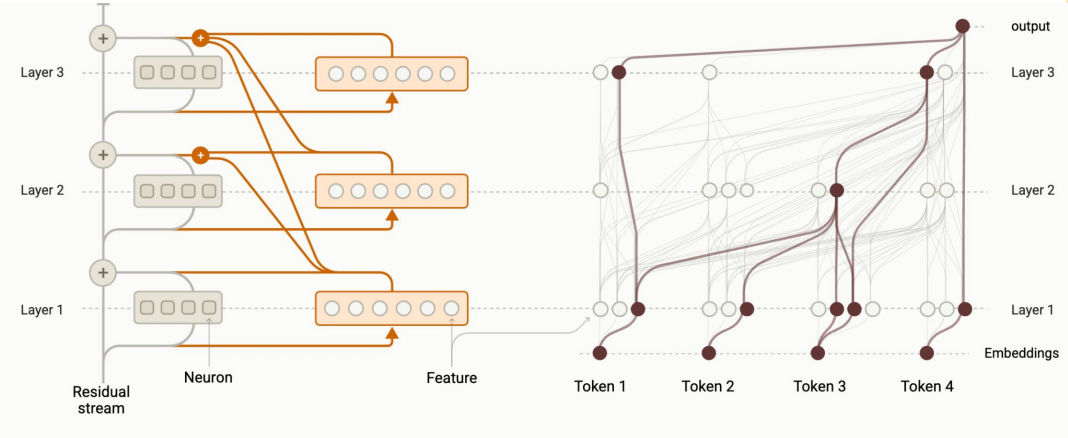
Yonatan Belinkov
Technion - IIT,
Kempner Institute

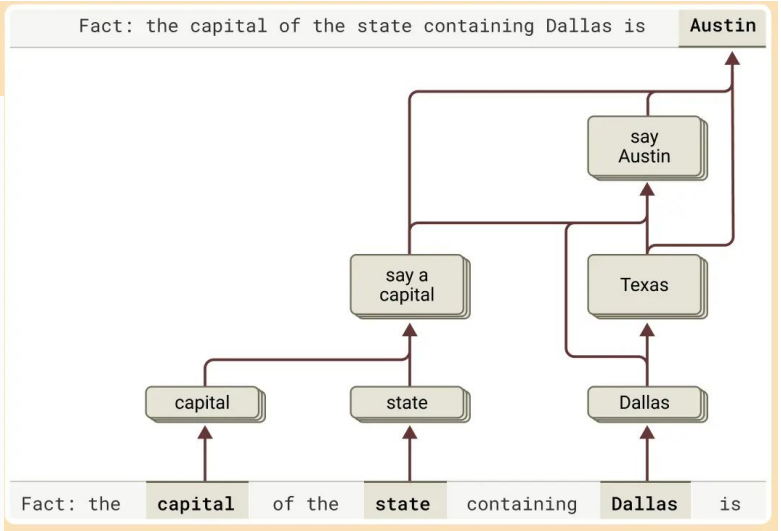
Fernanda Viégas
Harvard,
DeepMind

Martin
Wattenberg
Harvard,
DeepMind

Recent progress on interpretability

- Recent paradigm: Decompose model into sparse, interpretable components
  - E.g., activations, MLP blocks
- Allows researchers to construct “circuits”

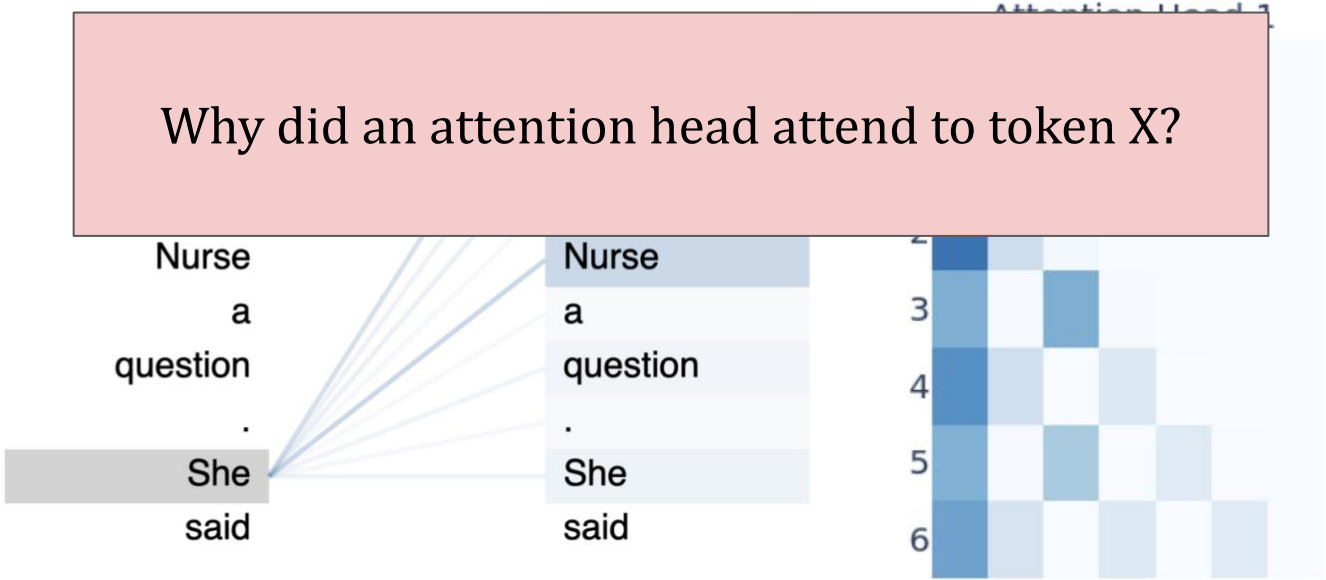




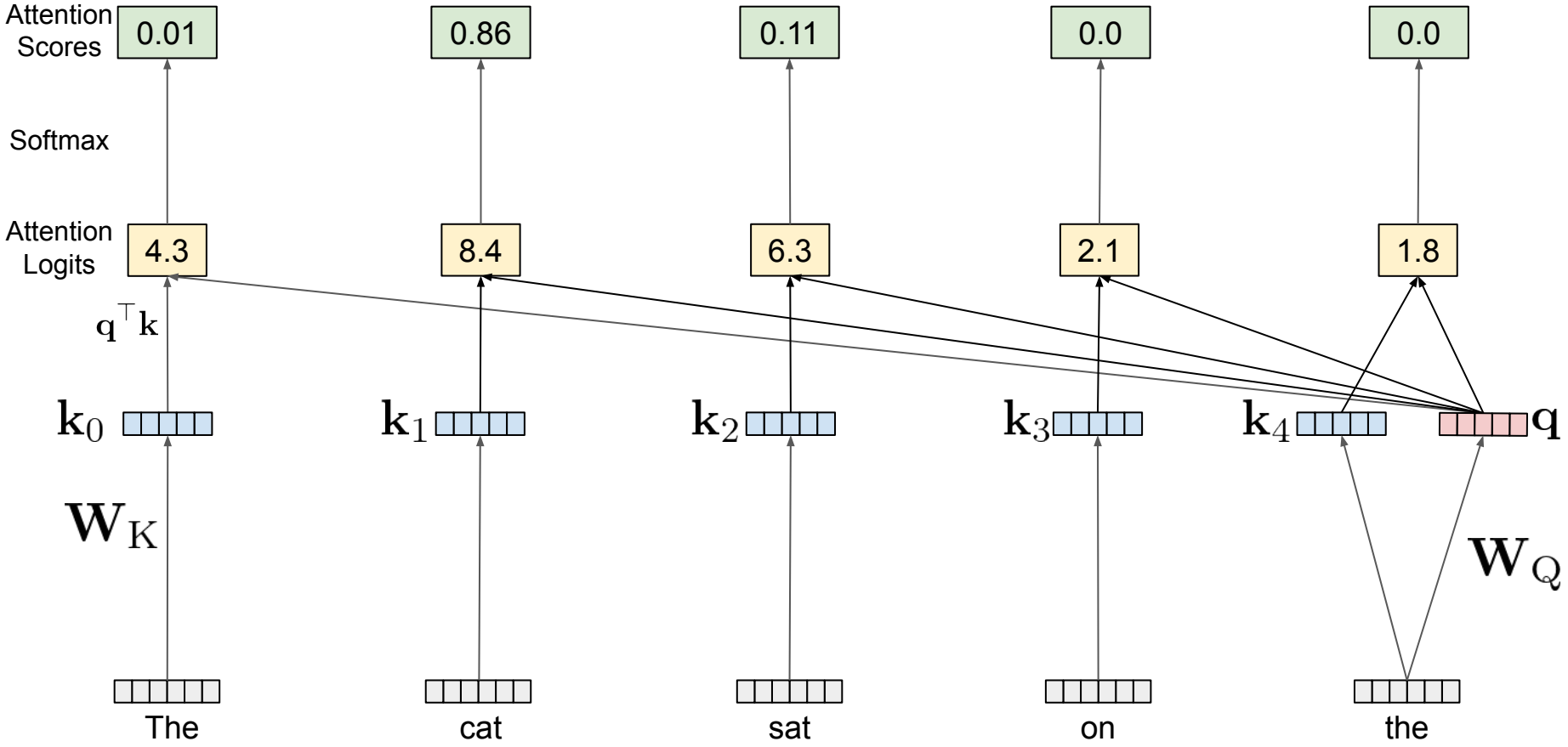
Images from: Tracing the thoughts of a large language model: Anthropic

Recent progress on interpretability

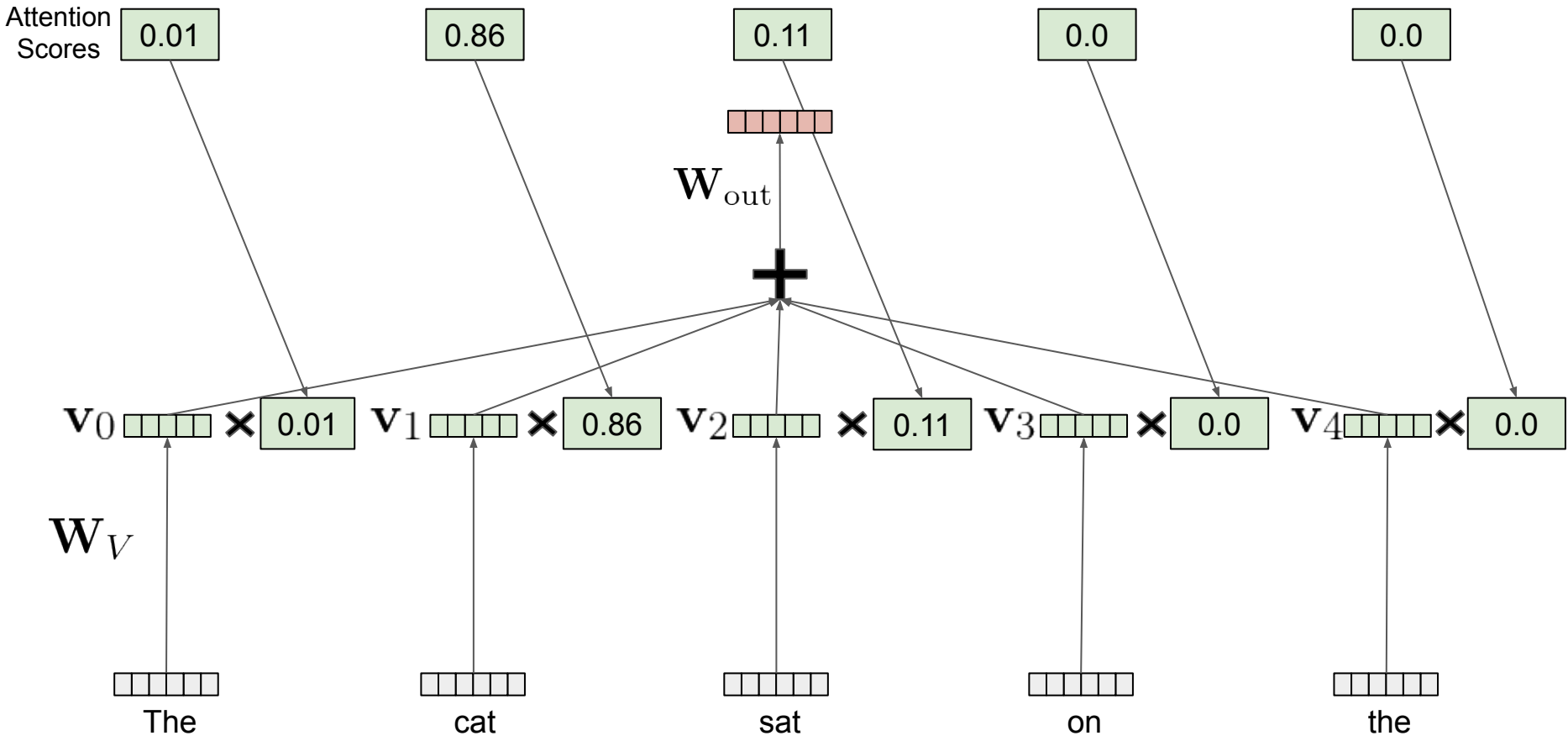
- But we do not know how to interpret attention!
  - In circuit analyses, attention patterns are fixed



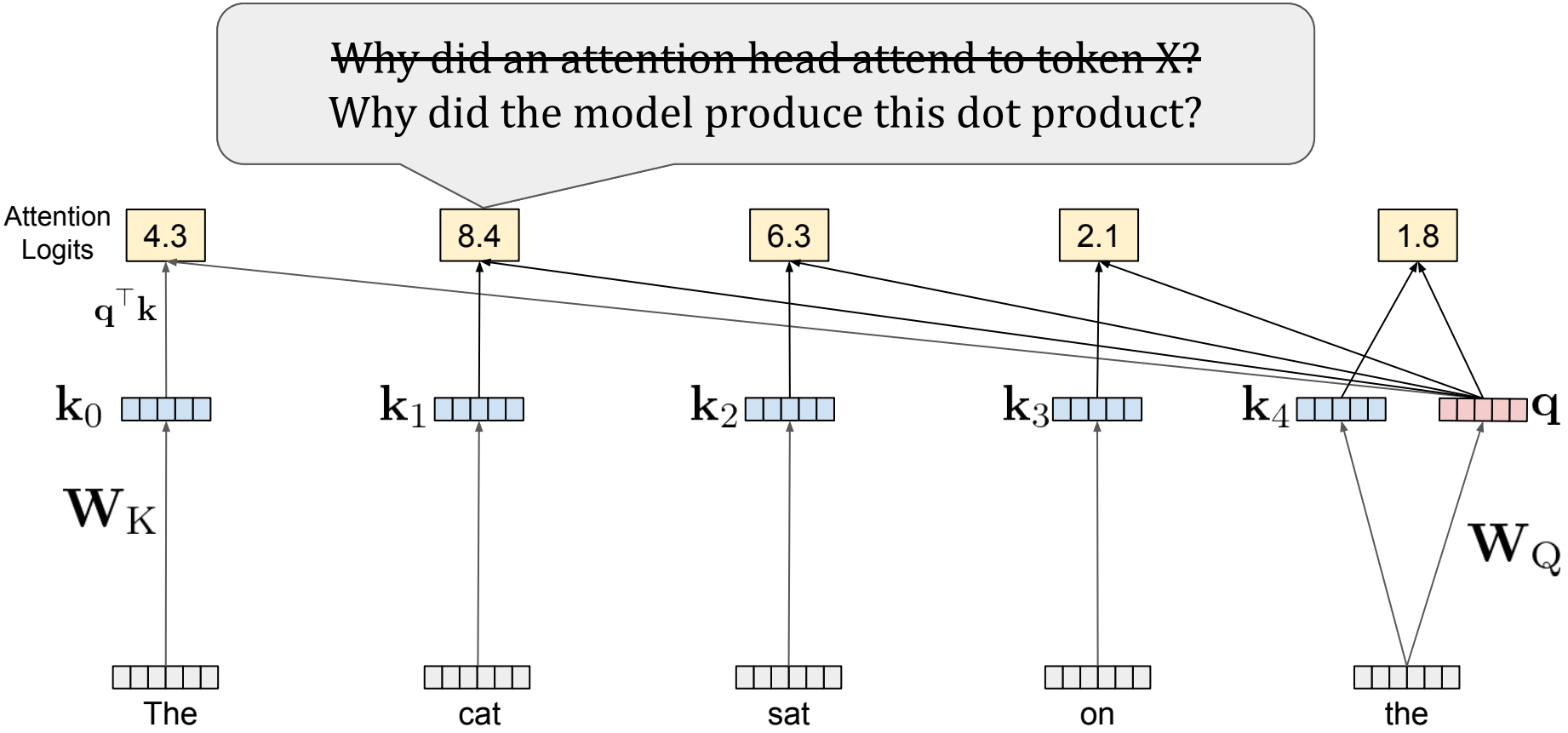
Quick recap: Attention



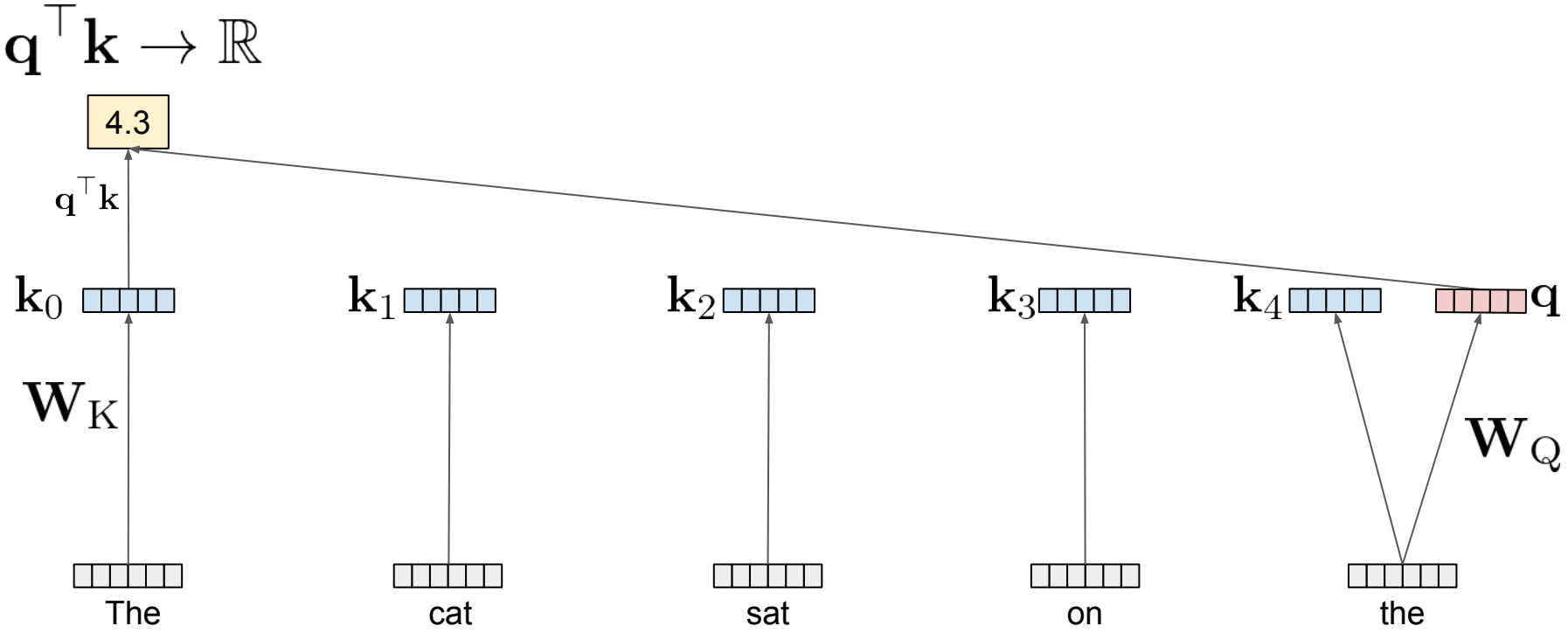
Quick recap: Attention



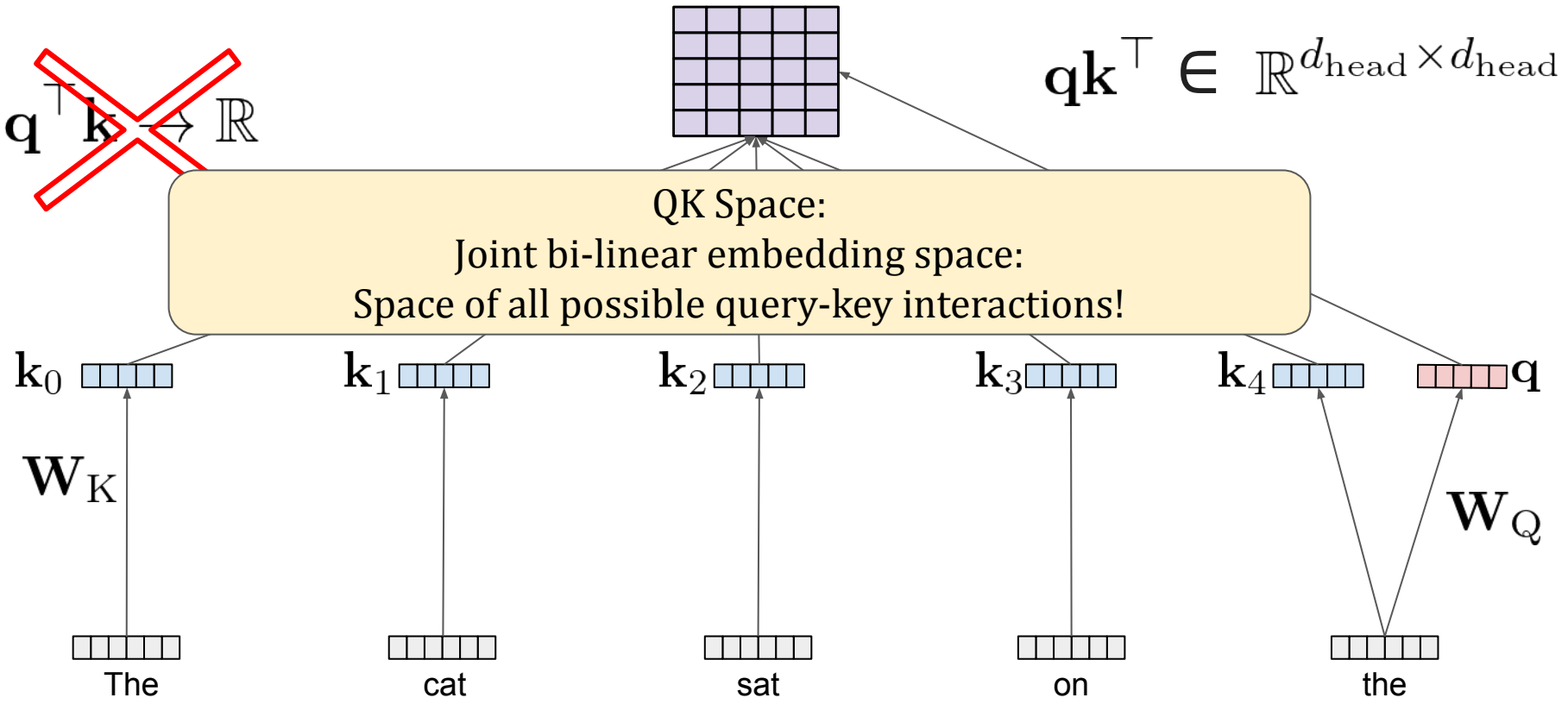
Quick recap: Attention



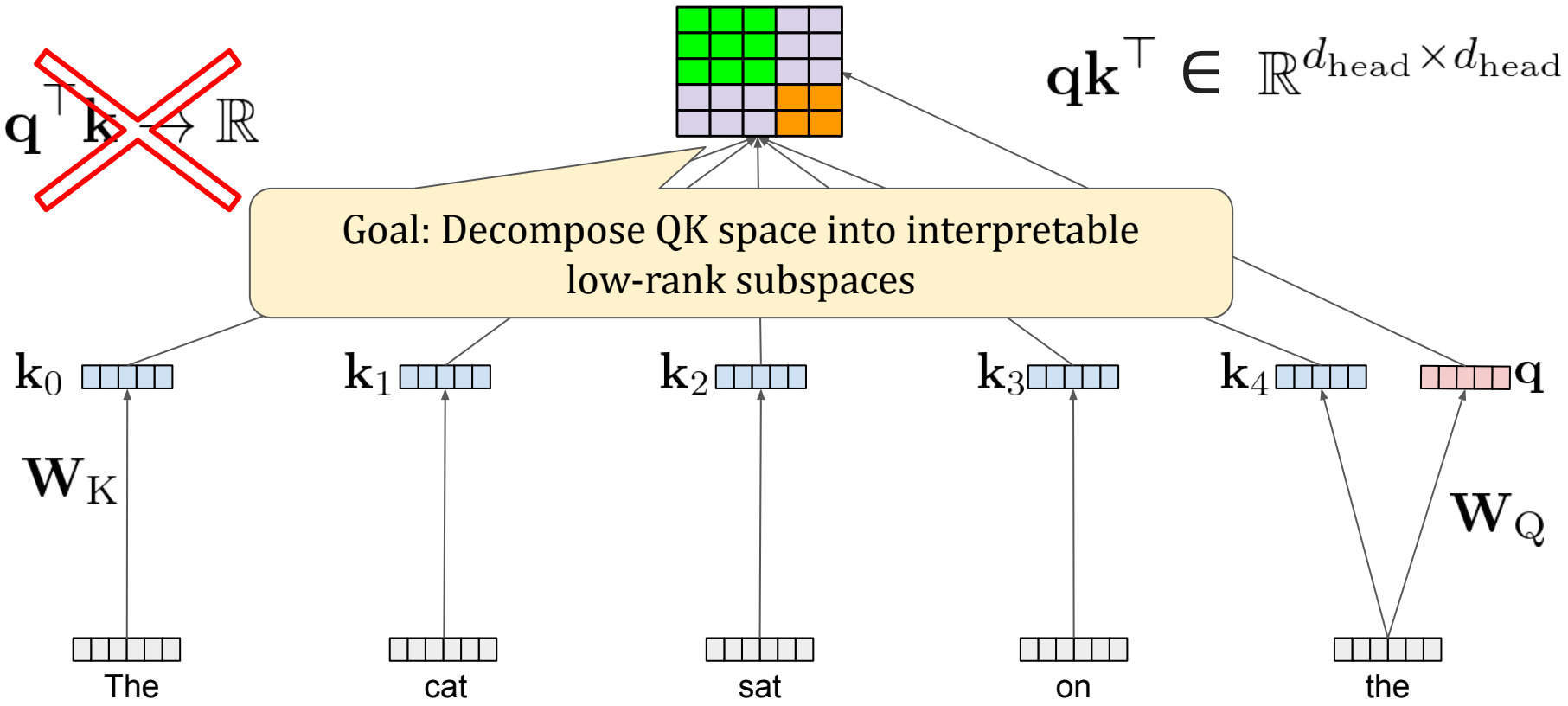
Quick recap: Attention



Query-Key (QK) Space



Query-Key (QK) Space

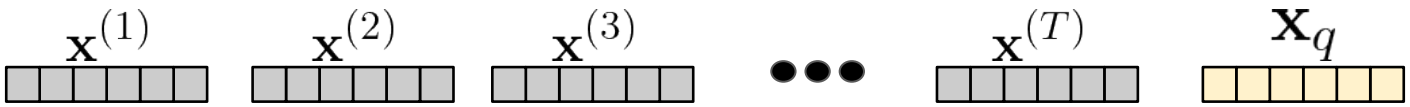


Agenda

1. Toy Task / Model
2. Method: Contrastive Covariance
3. Results on Large Language Models

Toy Task Overview: Payload Retrieval

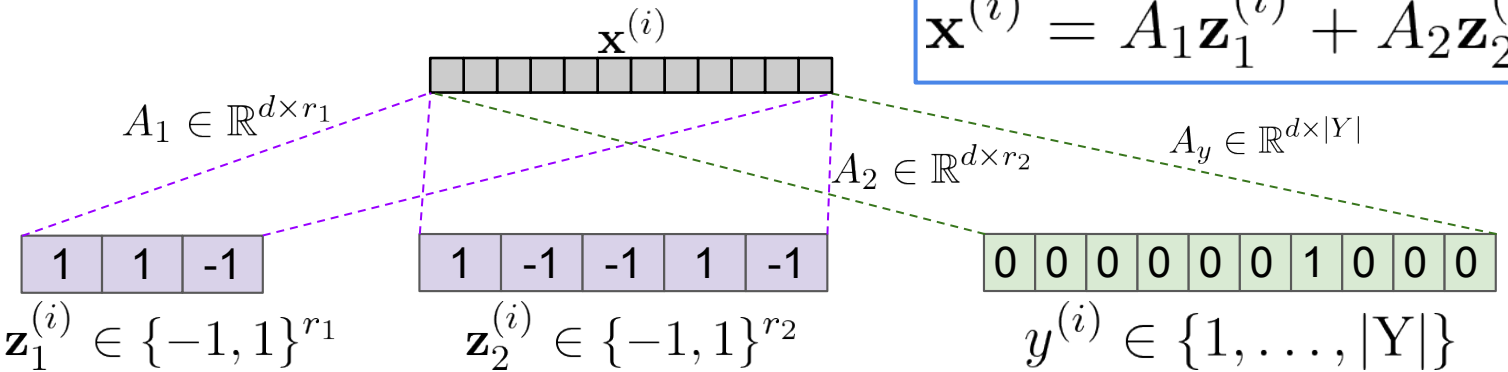
- Assume a set of "payload embeddings" $\{\mathbf{x}^{(i)}\}$...
- and a "query embedding" $\mathbf{x}_q$ that specifies an embedding ($\mathbf{x}^{(i^*)}$) to fetch payload information from
- We will train a *single attention head* to use $\mathbf{x}_q$ to attend to the correct embedding ($\mathbf{x}^{(i^*)}$) and decode the correct payload information
  - No position embeddings, MLPs, residual connections…
  - Just weights $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$, $\mathbf{W}_O$



Toy Task: Payload Retrieval

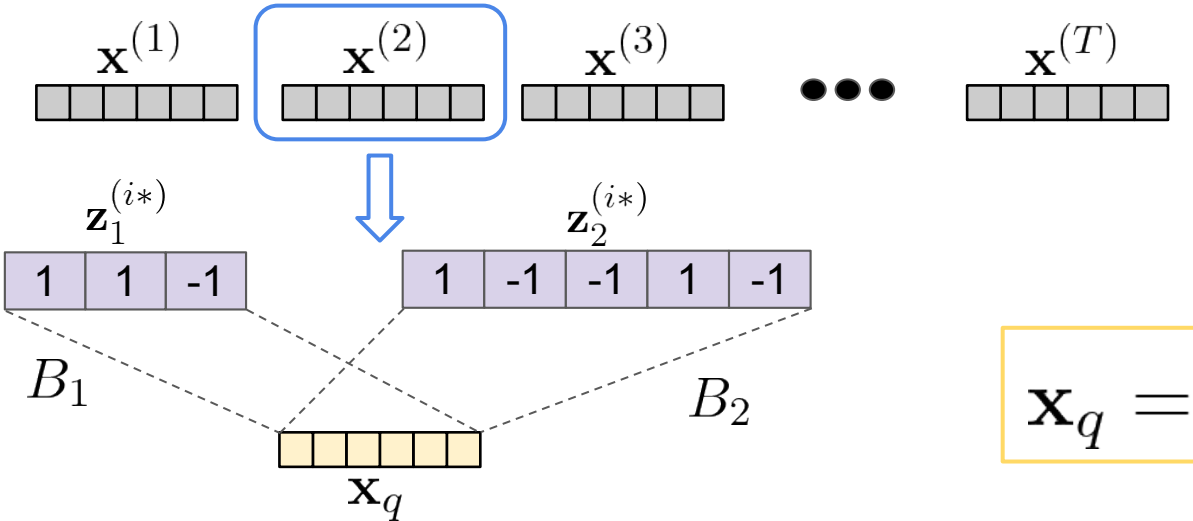
- Each payload embedding $\mathbf{x}^{(i)}$ consists of 3 components:
  - Two latent variables, $\mathbf{z}_1$, $\mathbf{z}_2$
  - Payload (class) $\mathbf{y} \in \{1, \ldots, |Y|\}$
- Each latent variable $\mathbf{z}_1$, $\mathbf{z}_2$ are random sign vectors of length r_1, r_2
- Each are mapped to embedding space via maps A_1, A_2
  - A_1, A_2 are *fixed* random Gaussians
- Payload y is randomly sampled from {1, …, |Y|}
- y is also embedded via map A_y

$$\mathbf{x}^{(i)} = A_1\mathbf{z}_1^{(i)} + A_2\mathbf{z}_2^{(i)} + A_y e_y^{(i)}$$



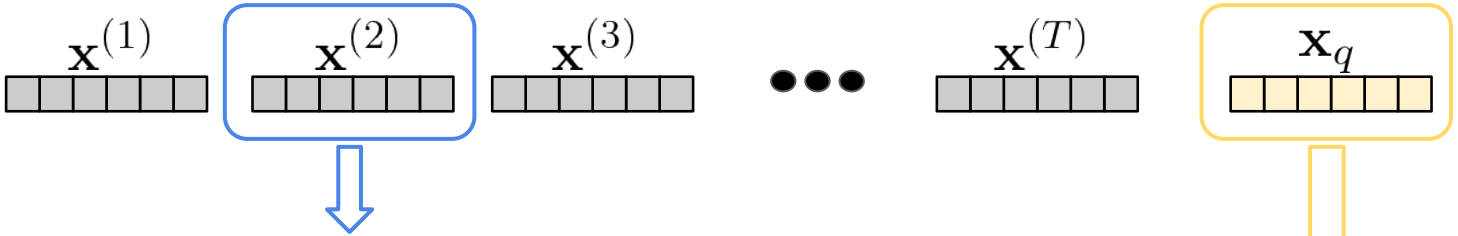
Toy Task: Payload Retrieval

- Given a set of payload embeddings $\mathbf{x}_{1:T}$…
- Randomly select one as "target payload" ($\mathbf{x}^{(i^*)}$)
- Re-use latent variables $\mathbf{z}_1$, $\mathbf{z}_2$ from $\mathbf{x}^{(i^*)}$ to create "query embedding" $\mathbf{x}_q$
- $\mathbf{z}_1^{(i^*)}$, $\mathbf{z}_2^{(i^*)}$ mapped via *different fixed* linear maps B_1, B_2
- Importantly, $\mathbf{x}_q$ *does not* contain any payload information



$$\mathbf{x}_q = B_1\mathbf{z}_1^{(i*)} + B_2\mathbf{z}_2^{(i*)}$$

Toy Task: Payload Retrieval



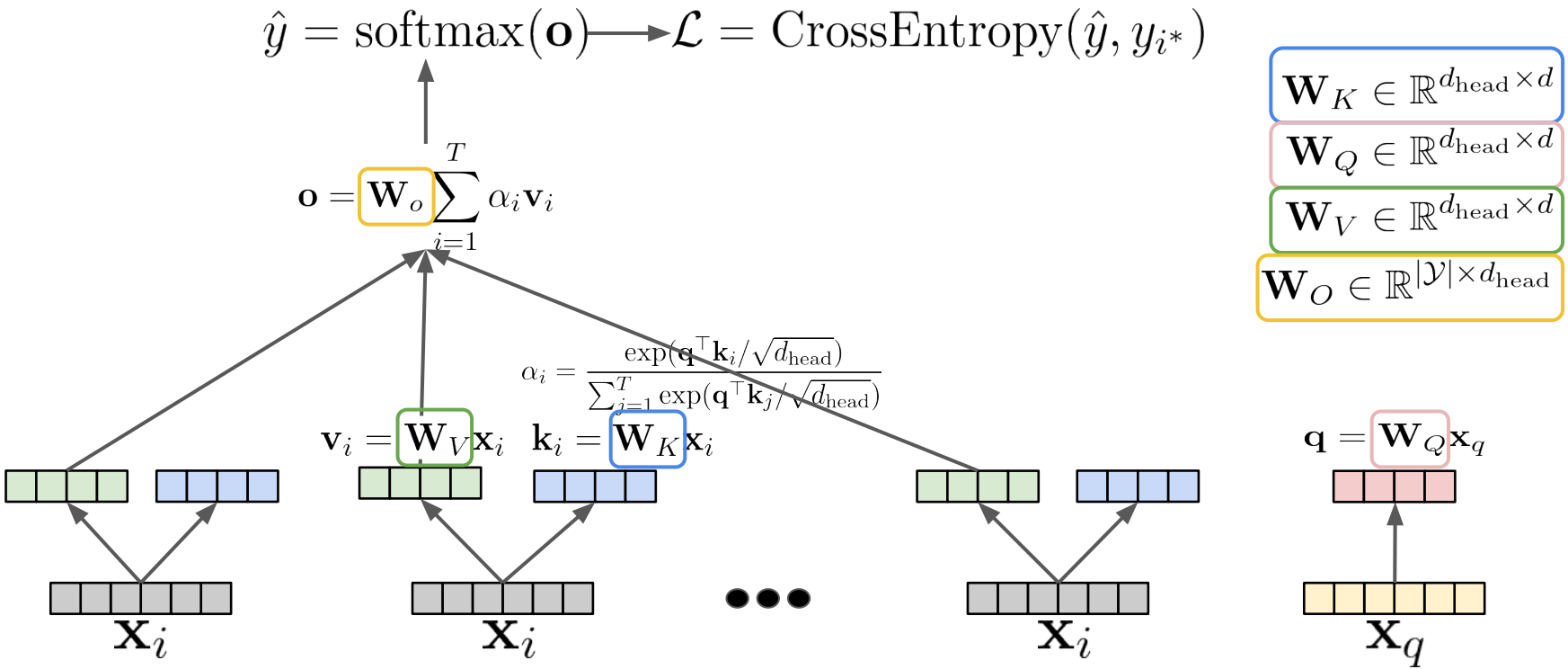
$$\mathbf{x}^{(i)} = A_1 \mathbf{z}_1^{(i)} + A_2 \mathbf{z}_2^{(i)} + A_y e_y^{(i)}$$

$$\mathbf{x}_q = B_1 \mathbf{z}_1^{(i*)} + B_2 \mathbf{z}_2^{(i*)}$$

$$\left\{ \mathbf{x}^{(1:T)}, \mathbf{x}_q, i*, y_{i*} \right\}^N$$

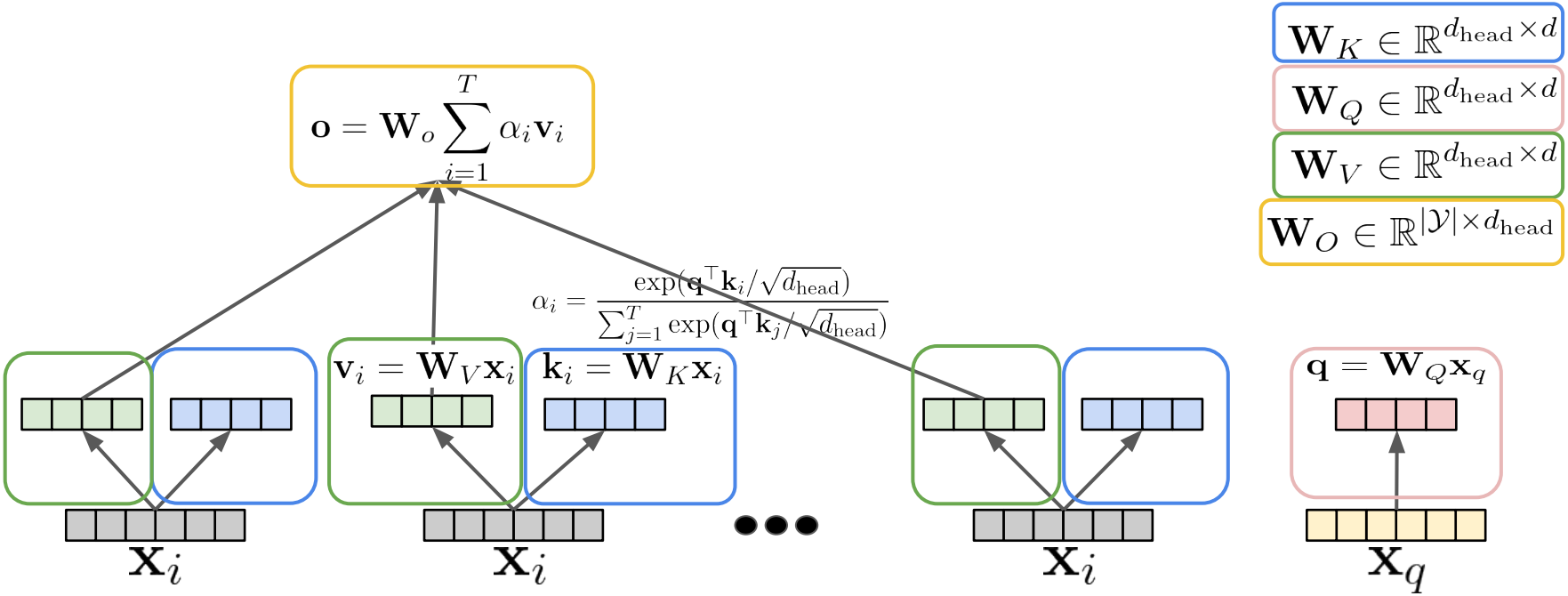
Toy Model

- We will train a *single attention head*



Toy Model (Ideal Solution)

- Ideally, the model learns W_Q, W_K such that **q, k** are aligned when both $\mathbf{z_1}, \mathbf{z_2}$
- Learn W_V, W_O to decode payload information (y)
- As it turns out, the model easily learns this solution and reaches 100% accuracy



Agenda

1. Toy Task / Model
2. **Method: Contrastive Covariance**
3. Results on Large Language Models

Can we recover the subspace in QK-space in which latent variables $\mathbf{z_1}, \mathbf{z_2}$ are embedded?

Contrastive Covariance Method

- Per latent variable (i.e., $\mathbf{z}_1$), define "positive" and "negative" covariance
- Take their difference (a.k.a. Contrastive covariance)
- SVD
- Repeat for $\mathbf{z}_2$

$$\mathbf{C}^{+}_{(\mathbf{z}_1)} := \mathbb{E}[\mathbf{q}\mathbf{k}^\top|+] \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}} \qquad \mathbf{C}^{-}_{(\mathbf{z}_1)} := \mathbb{E}[\mathbf{q}\mathbf{k}^\top|-] \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$$

$(\mathbf{q}, \mathbf{k}_i)$ where $\mathbf{z}_1^{(i)} = \mathbf{z}_1^{(i*)}, \mathbf{z}_2^{(i)} = \tilde{\mathbf{z}}_2$ $\qquad$ $(\mathbf{q}, \mathbf{k}_j)$ where $\mathbf{z}_1^{(j)} \neq \mathbf{z}_1^{(i*)}, \mathbf{z}_2^{(j)} = \tilde{\mathbf{z}}_2$

$$\Delta C_{(\mathbf{z}_1)} = \mathbb{E}[\mathbf{q}\mathbf{k}^\top|+] - \mathbb{E}[\mathbf{q}\mathbf{k}^\top|-]$$

$$\Delta \mathbf{C}_{(\mathbf{z}_1)} = \mathbf{U}_{(\mathbf{z}_1)} \mathbf{\Sigma}_{(\mathbf{z}_1)} \mathbf{V}^\top_{(\mathbf{z}_1)}$$

Contrastive Covariance Method

$$\Delta C_{(\mathbf{z}_1)} = \mathbb{E}[\mathbf{q}\mathbf{k}^\top|+] - \mathbb{E}[\mathbf{q}\mathbf{k}^\top|-]$$

$$\Delta \mathbf{C}_{(\mathbf{z}_1)} = \mathbf{U}_{(\mathbf{z}_1)}\mathbf{\Sigma}_{(\mathbf{z}_1)}\mathbf{V}^\top_{(\mathbf{z}_1)} \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$$

- r_1 (Rank of $\mathbf{z_1}$): number of singular values that account for 99% of $|\mathbf{\Delta C}_{(\mathbf{z1})}|_\mathbf{F}$
- $\mathbf{U}_{(\mathbf{z1})}[:r_1]$: basis in query space that encodes $\mathbf{z_1}$
- $\mathbf{V}_{(\mathbf{z1})}[:r_1]$: basis in key space that encodes $\mathbf{z_1}$
- Repeat procedure for $\mathbf{z_2}$

Contrastive Covariance Method

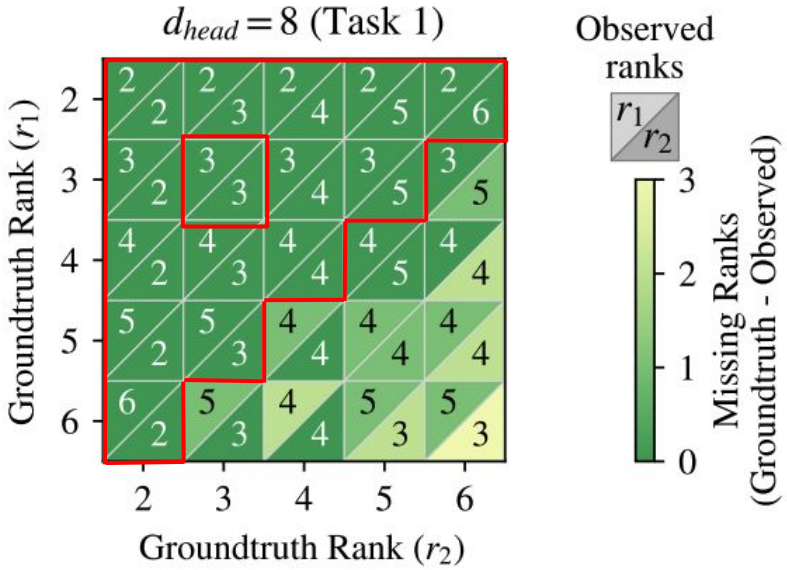
$$\Delta C_{(\mathbf{z}_1)} = \mathbb{E}[\mathbf{q}\mathbf{k}^\top | +] - \mathbb{E}[\mathbf{q}\mathbf{k}^\top | -]$$

$$\Delta \mathbf{C}_{(\mathbf{z}_1)} = \mathbf{U}_{(\mathbf{z}_1)} \mathbf{\Sigma}_{(\mathbf{z}_1)} \mathbf{V}_{(\mathbf{z}_1)}^\top \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$$

- r_1 (Rank of $\mathbf{z}_1$): number of singular values that account for 99% of $|\Delta \mathbf{C}_{(\mathbf{z}1)}|_\mathbf{F}$
- $\mathbf{U}_{(\mathbf{z}1)}[:r_1]$: basis in query space that encodes $\mathbf{z}_1$
- $\mathbf{V}_{(\mathbf{z}1)}[:r_1]$: basis in key space that encodes $\mathbf{z}_1$
- Repeat procedure for $\mathbf{z}_2$

Toy Model Results

- Can we recover the correct ranks of latent features?
- Train attention head on varying task settings (vary r_1, r_2)



Toy Model Results

$$\Delta C_{(\mathbf{z}_1)} = \mathbb{E}[\mathbf{q}\mathbf{k}^\top | +] - \mathbb{E}[\mathbf{q}\mathbf{k}^\top | -]$$

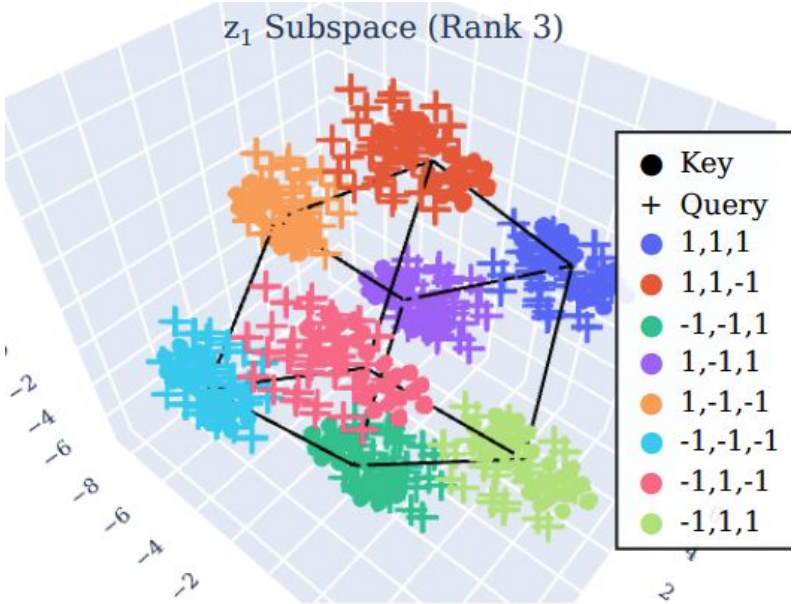
$$\Delta \mathbf{C}_{(\mathbf{z}_1)} = \mathbf{U}_{(\mathbf{z}_1)} \mathbf{\Sigma}_{(\mathbf{z}_1)} \mathbf{V}_{(\mathbf{z}_1)}^\top \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$$

- r_1 (Rank of $\mathbf{z}_1$): number of singular values that account for 99% of $|\Delta \mathbf{C}_{(\mathbf{z}1)}|_F$
- $\mathbf{U}_{(\mathbf{z}1)}[:r_1]$: basis in query space that encodes $\mathbf{z}_1$
- $\mathbf{V}_{(\mathbf{z}1)}[:r_1]$: basis in key space that encodes $\mathbf{z}_1$
- Repeat procedure for $\mathbf{z}_2$

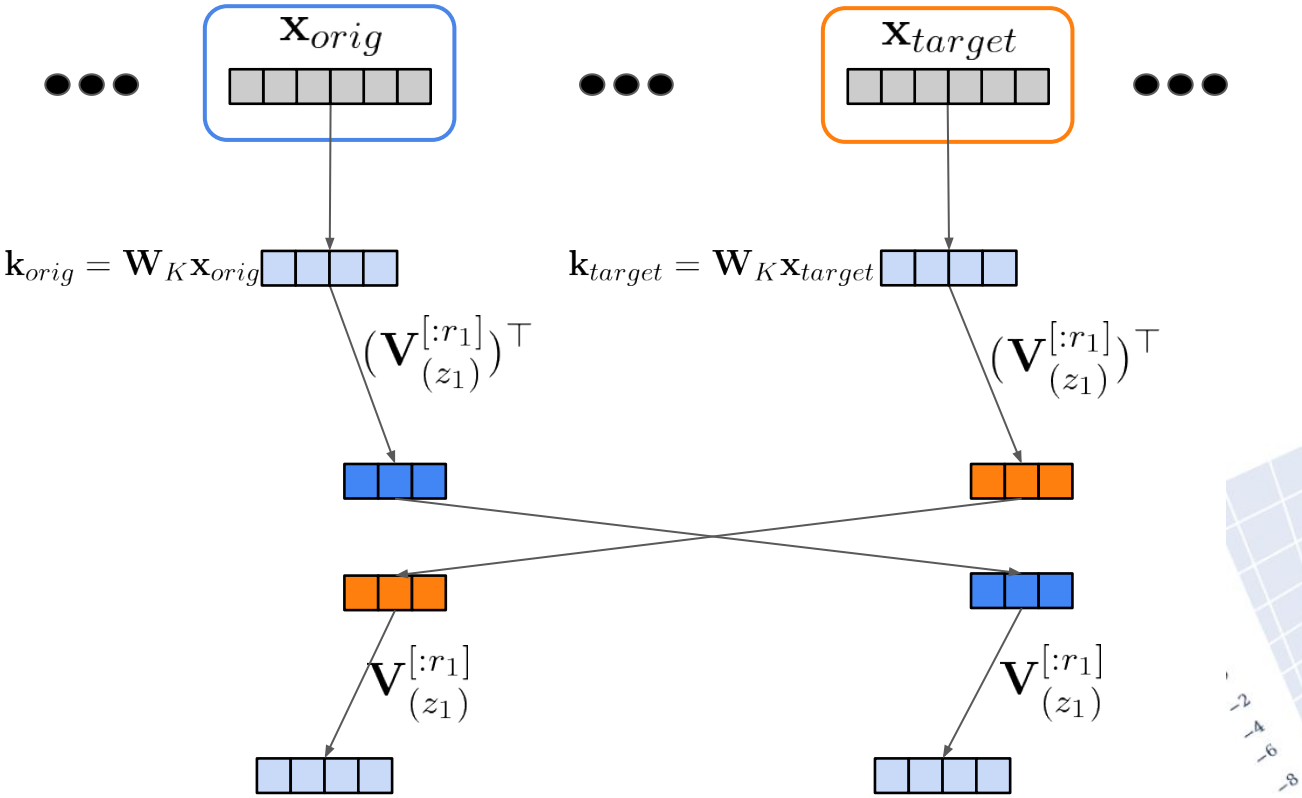
Toy Model Results

- Consider model trained on $r_1 = 3$
  - $\mathbf{z_1} = \{\pm 1, \pm 1, \pm 1\}$ – vertices of a cube
- Given query vectors ($\mathbf{q} = \mathbf{W_Q x_q} \in \mathbb{R}^{\text{d\_head}}$), key vectors ($\mathbf{k} = \mathbf{W_K x_k} \in \mathbb{R}^{\text{d\_head}}$)
- Project them onto $\mathbf{U_{(z1)}} \in \mathbb{R}^3$, $\mathbf{V_{(z1)}} \in \mathbb{R}^3$
- PCA!

$$\Delta \mathbf{C}_{(\mathbf{z}_1)} = \mathbf{U}_{(\mathbf{z}_1)} \mathbf{\Sigma}_{(\mathbf{z}_1)} \mathbf{V}_{(\mathbf{z}_1)}^{\top} \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$$



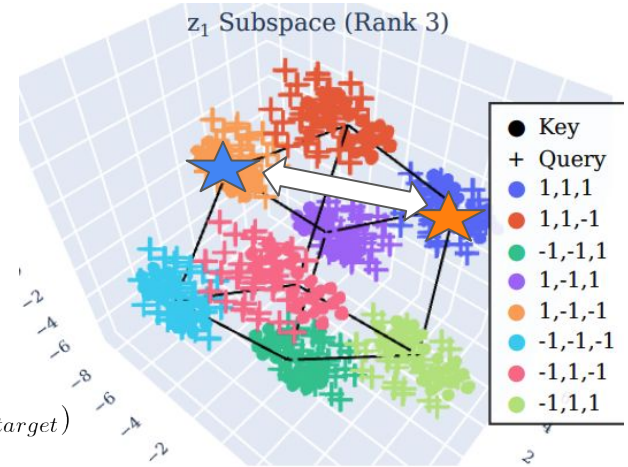
Toy Model Results: Causal Interventions



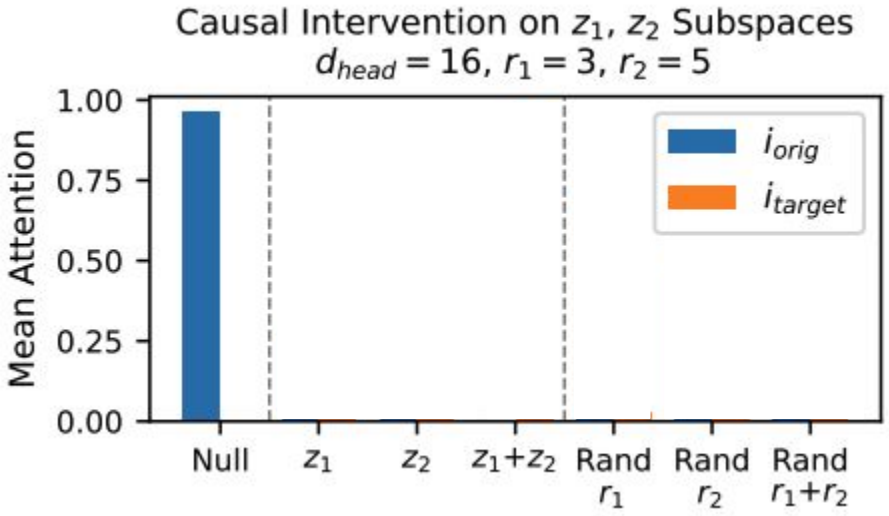
$$\Delta \mathbf{C}_{(\mathbf{z}_1)} = \mathbf{U}_{(\mathbf{z}_1)} \mathbf{\Sigma}_{(\mathbf{z}_1)} \mathbf{V}_{(\mathbf{z}_1)}^\top$$

$\mathbf{U}_{(z_1)}^{[:r_1]}$ encodes $\mathbf{z_1}$ (query space)

$\mathbf{V}_{(z_1)}^{[:r_1]}$ encodes $\mathbf{z_1}$ (key space)



Toy Model Results: Causal Interventions



Agenda

1. Toy Task / Model
2. Method: Contrastive Covariance
3. **Results on Large Language Models**

LLM Results

- “Filter Heads”
- Binding Features

We study Llama 3.1-8B-Instruct and Qwen 3-4B-Instruct

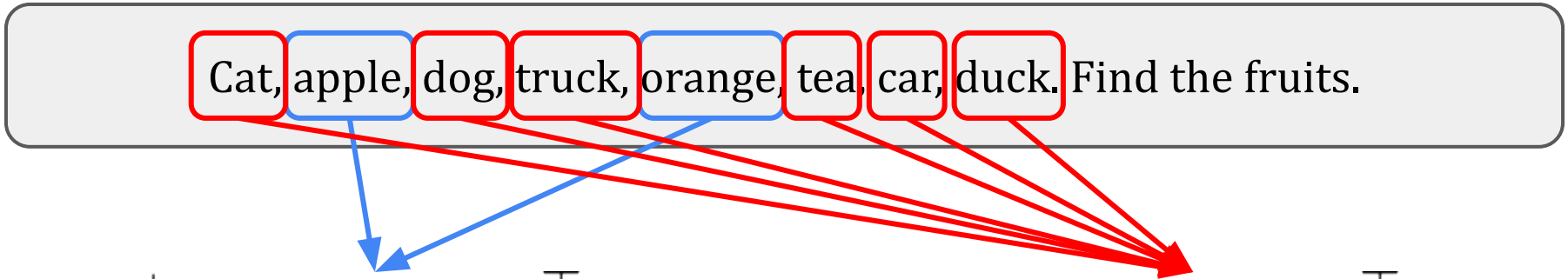
Filter Heads

Cat, apple, dog, truck, orange, tea, car, duck. Find the fruits.

- Attention heads that mirror “filter” functions [1]
- What features are these heads using?

[1] Sharma, A. S., Rogers, G., Shapira, N., and Bau, D. Llms process lists with general filter heads. 2025

Filter Heads

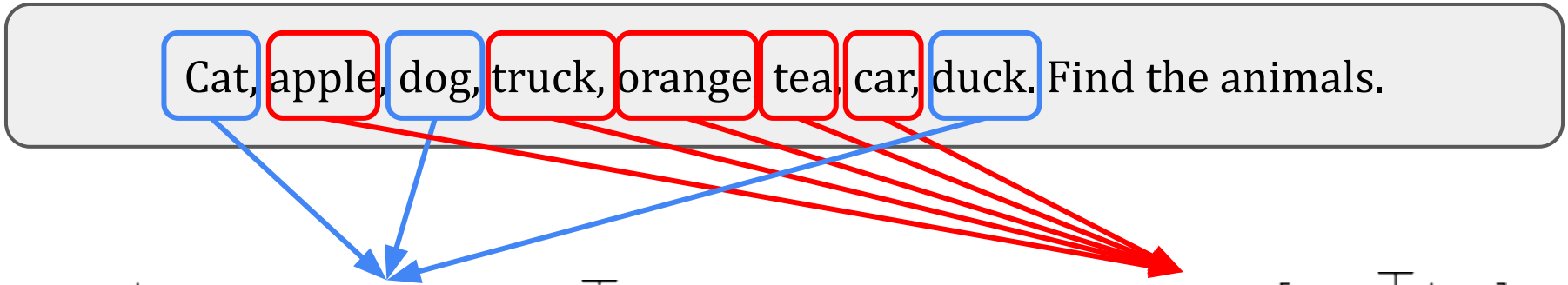


$$\mathbf{C}^{+}_{\text{Fruits}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|+] \qquad \mathbf{C}^{-}_{\text{Fruits}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|-]$$

$$\Delta\mathbf{C}_{\text{Fruits}} := \mathbf{C}^{+}_{\text{Fruits}} - \mathbf{C}^{-}_{\text{Fruits}}$$

$$\Delta\mathbf{C}_{\text{Fruits}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$$

Filter Heads

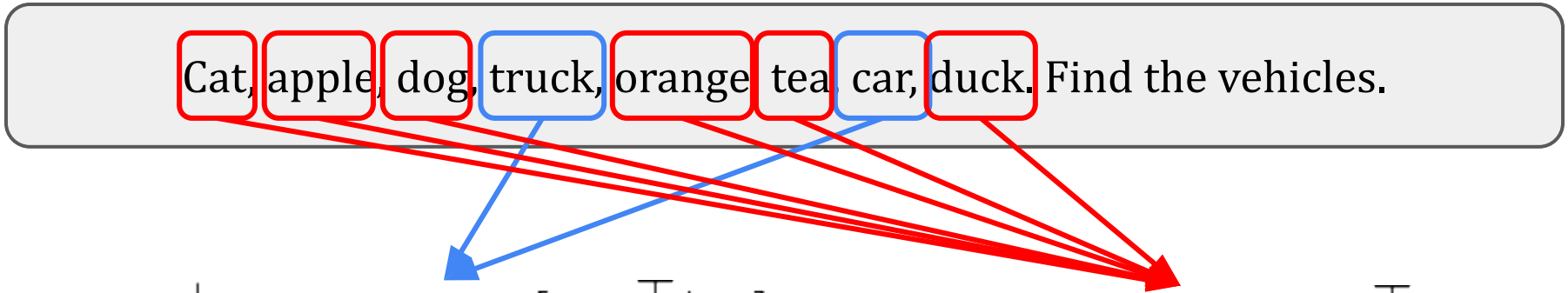


$$\mathbf{C}^{+}_{\text{Animals}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|+] \qquad \mathbf{C}^{-}_{\text{Animals}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|-]$$

$$\Delta\mathbf{C}_{\text{Animals}} := \mathbf{C}^{+}_{\text{Animals}} - \mathbf{C}^{-}_{\text{Animals}}$$

$$\Delta\mathbf{C}_{\text{Animals}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$$

Filter Heads



$$\mathbf{C}^{+}_{\text{Vehicles}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|+] \quad \mathbf{C}^{-}_{\text{Vehicles}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|-]$$

$$\Delta\mathbf{C}_{\text{Vehicles}} := \mathbf{C}^{+}_{\text{Vehicles}} - \mathbf{C}^{-}_{\text{Vehicles}}$$

$$\Delta\mathbf{C}_{\text{Vehicles}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$$

Filter Heads

- Compute $\Delta\mathbf{C}$ for 5 categories
  - Animals, countries, fruits, liquid, vehicles
- Interestingly, every $\Delta\mathbf{C}$ is rank 1
- $\Delta\mathbf{C}_{\text{Category}} = U\Sigma V^{T}$
- Categorical semantic space in query space:
  - $\mathbf{U}_{\text{Categories}} := \text{span}(\mathbf{U}_{\text{Animals}}, \mathbf{U}_{\text{Countries}}, \mathbf{U}_{\text{Fruits}}, \mathbf{U}_{\text{Liquid}}, \mathbf{U}_{\text{Vehicles}})$
- Categorical semantic space in key space:
  - $\mathbf{V}_{\text{Categories}} := \text{span}(\mathbf{V}_{\text{Animals}}, \mathbf{V}_{\text{Countries}}, \mathbf{V}_{\text{Fruits}}, \mathbf{V}_{\text{Liquid}}, \mathbf{V}_{\text{Vehicles}})$

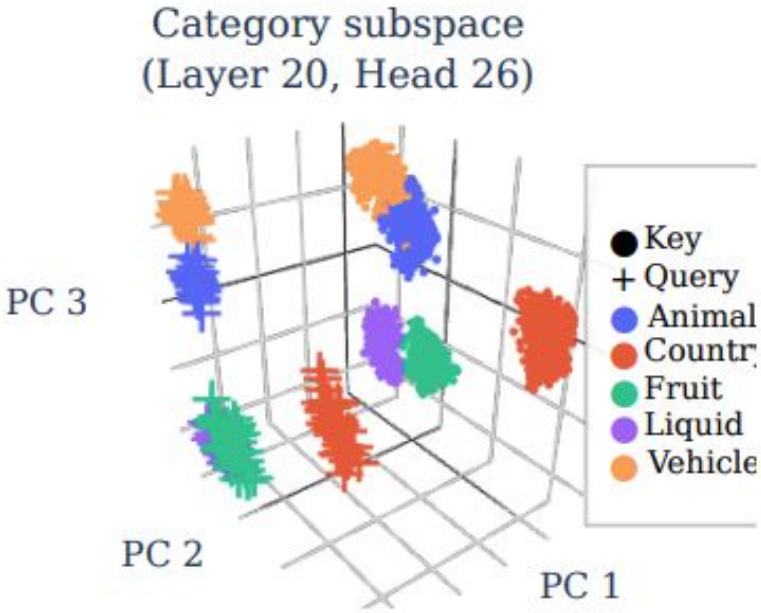
Filter Heads (PCA)

Cat, apple, dog, truck, orange, tea, car, duck. Find the fruits.

- Take key vectors for each entity
- Take query vector (last token in input sentence)
- Project query vectors onto $\mathbf{U}_{\text{Categories}} := \text{span}(\mathbf{U}_{\text{Animals}}, \mathbf{U}_{\text{Countries}}, \mathbf{U}_{\text{Fruits}}, \mathbf{U}_{\text{Liquid}}, \mathbf{U}_{\text{Vehicles}})$
- Project key vectors onto $\mathbf{V}_{\text{Categories}} := \text{span}(\mathbf{V}_{\text{Animals}}, \mathbf{V}_{\text{Countries}}, \mathbf{V}_{\text{Fruits}}, \mathbf{V}_{\text{Liquid}}, \mathbf{V}_{\text{Vehicles}})$
- PCA!

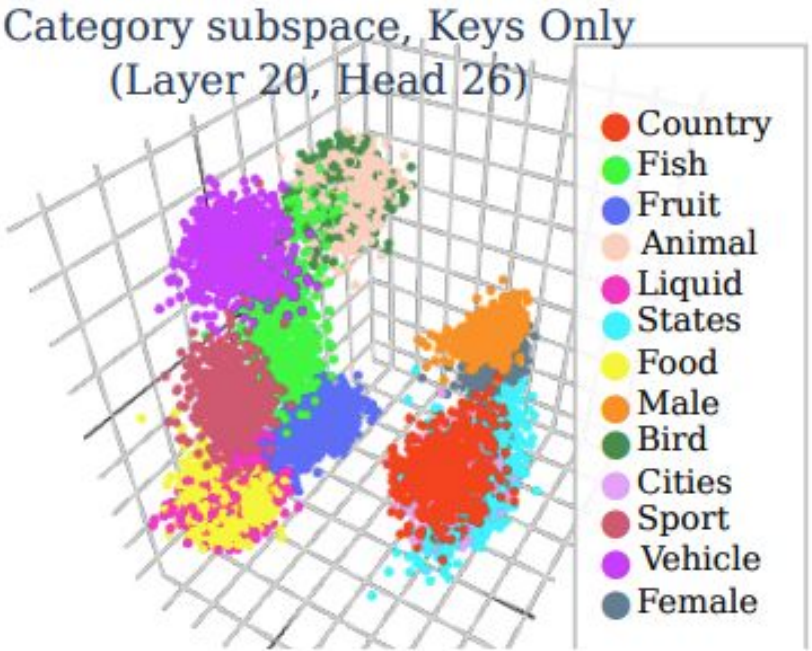
Filter Heads (PCA)

Cat, apple, dog, truck, orange, tea, car, duck. Find the fruits.



Filter Heads (PCA)

- Expanding to 13 categories
- (Visualizing keys only)



Filter Heads (Causal Interventions)

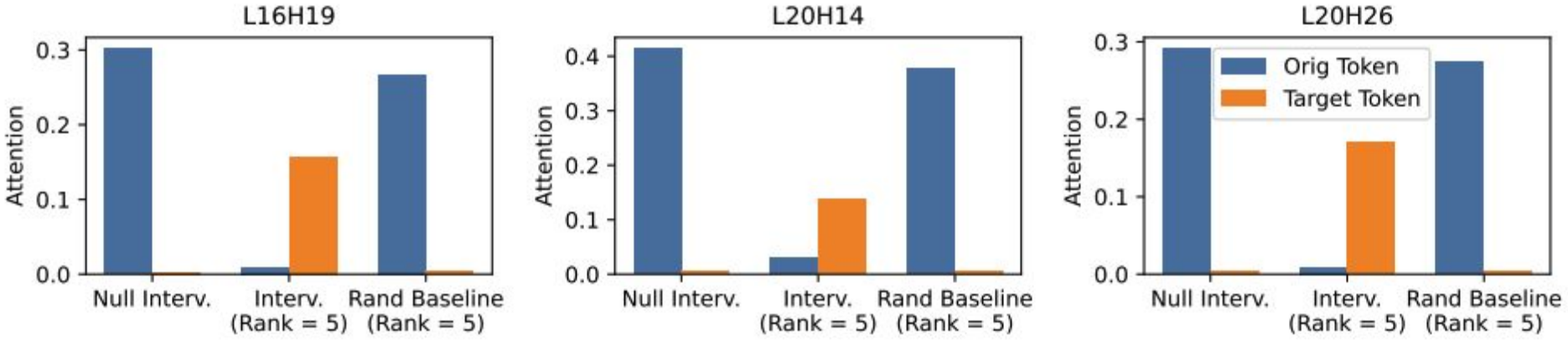
Cat, apple, dog, tea, car, duck. Find the fruits.

- Given queried category c^*, choose a different category c^{target}.
- Project key vectors onto $\mathbf{V}_{\text{Categories}}$, swap coordinates between c^*, c^{target}

Filter Heads (Causal Interventions)

Cat, apple, dog, tea, car, duck. Find the fruits.

- Given queried category c^*, choose a different category c^{target}.
- Project key vectors onto $\mathbf{V}_{\text{Categories}}$, swap coordinates between c^*, c^{target}



Binding Features

The hat is in box O. The jam is in box Z. The cat is in box B. The book is in box J. Which box is the jam in?

- Multiple binding features (“tags”) that the model uses [1]

[1] Yoav Gur-Arieh, Mor Geva, Atticus Geiger. Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context. 2025

Binding Features

The hat[1] is in box O[1]. The jam[2] is in box Z[2]. The cat[3] is in box B[3]. The book[4] is in box J[4].
Which box is the jam[2] in?

- Multiple binding features (“tags”) that the model uses [1]
- Order-IDs
  - Model “tags” the order in which entity-box pairs appear
  - Use “order-ID tag” to retrieve correct box

[1] Yoav Gur-Arieh, Mor Geva, Atticus Geiger. Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context. 2025

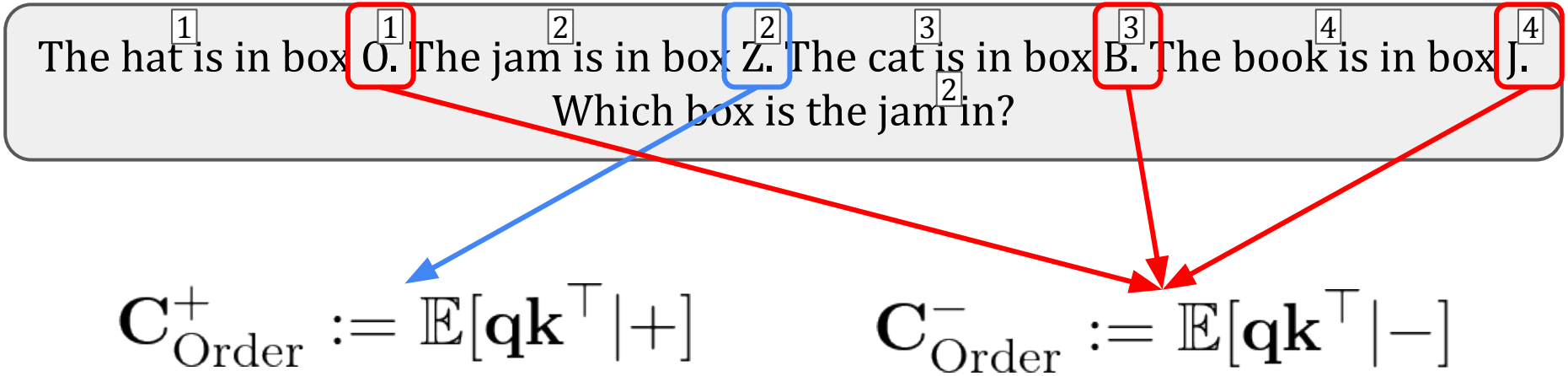
Binding Features

The hat is in box O. The jam is in box Z. The cat is in box B. The book is in box J.
Which box is the jam in?

- Multiple binding features (“tags”) that the model uses [1]
- Order-IDs
  - Model “tags” the order in which entity-box pairs appear
  - Use “order-ID tag” to retrieve correct box
- Lexical-IDs
  - The “intuitive” solution: retrieve box based on queried entity

[1] Yoav Gur-Arieh, Mor Geva, Atticus Geiger. Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context. 2025

Binding Features: Order-ID



$$\mathbf{C}^{+}_{\text{Order}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|+] \qquad \mathbf{C}^{-}_{\text{Order}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|-]$$

Importantly: we use the same entity tokens and box tokens in all of our samples

This allows us to factor out lexical information

Binding Features: Lexical-ID

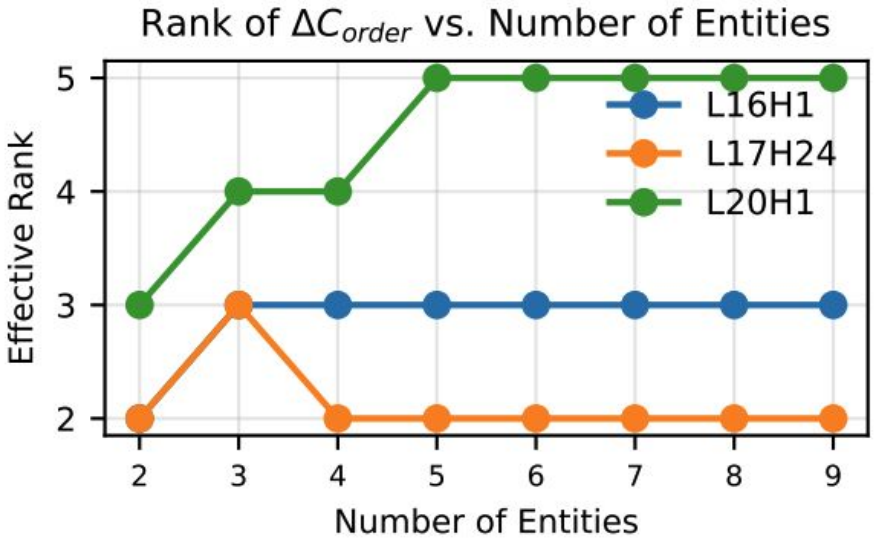
The hat is in box O. The jam is in box Z. The cat is in box B. The book is in box J. Which box is the jam in?

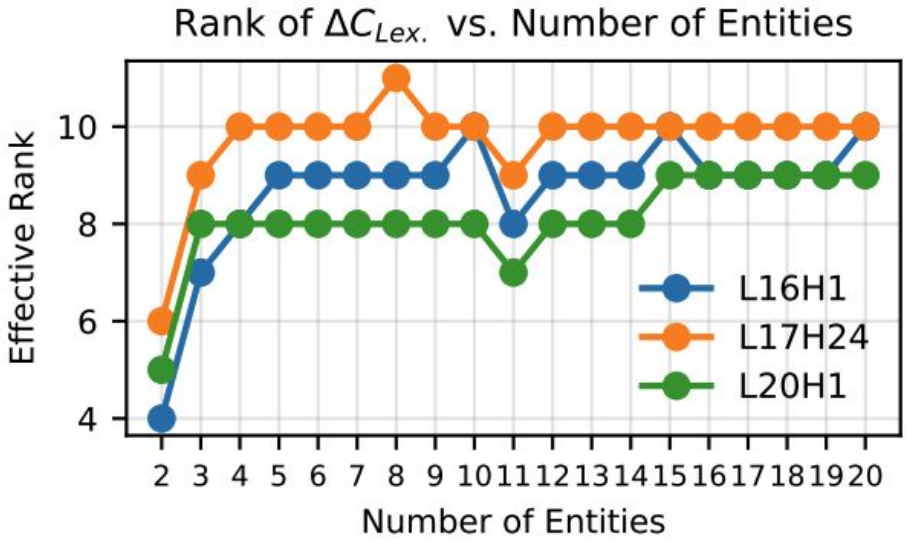
The hat is in box O. **The cup is in box Z** . The cat is in box B. The book is in box J. Which box is the **cup** in?

$$\mathbf{C}^{+}_{\text{Lex}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|+] \qquad \mathbf{C}^{-}_{\text{Lex}} := \mathbb{E}[\mathbf{q}\mathbf{k}^{\top}|-]$$

Low-Rank Binding Features

- ΔC_{Order}, $\Delta C_{Lexical}$ are low-rank
- i.e., binding features live in low-rank subspaces

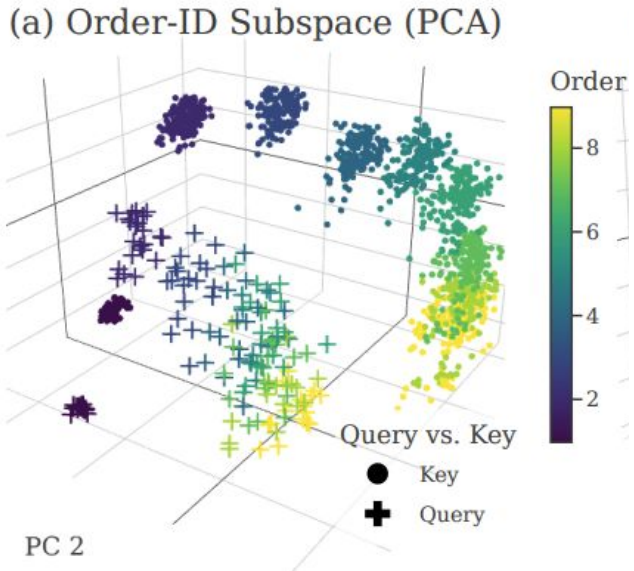




Binding Subspaces (PCA)

The hat is in box O. The jam is in box Z. The cat is in box B. The book is in box J. Which box is the jam in?

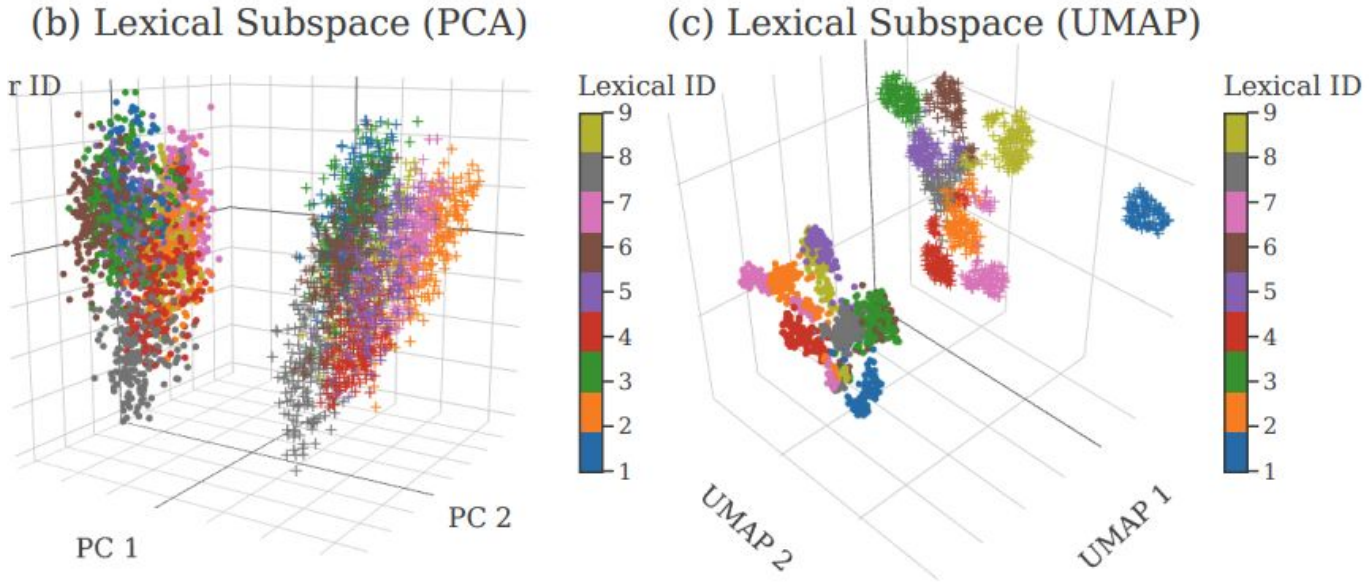
- Project key, query vectors onto $\mathbf{U}_{\text{Order}}$, $\mathbf{V}_{\text{Order}}$



Binding Subspaces (PCA)

The hat is in box O. The jam is in box Z. The cat is in box B. The book is in box J.
Which box is the jam in?

- Project key, query vectors onto $\mathbf{U}_{\text{Lex}}$, $\mathbf{V}_{\text{Lex}}$

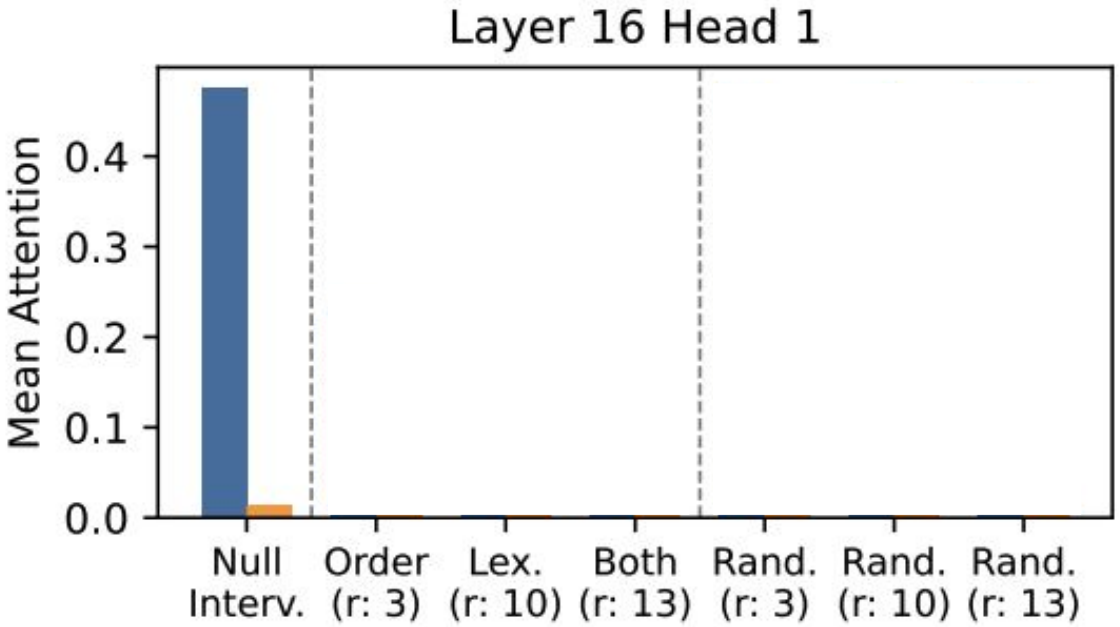


Binding Subspaces (Causal Interventions)

The hat is in box O. The jam is in box Z. The cat is in box B. The book is in box J. Which box is the jam in?

- Given queried box b^*, choose a different box b^{target}.
- Project key vectors onto $\mathbf{V}_{\mathbf{Order}}$, $\mathbf{V}_{\mathbf{Lexical}}$, or both. Swap coordinates between b^*, b^{target}

Binding Subspaces (Causal Interventions)

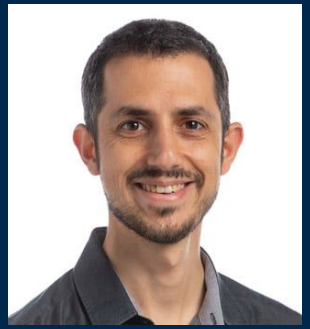


Takeaways

- Decomposing QK space reveals low-rank subspaces encoding features
- Future directions:
  - Contrastive covariance relies on carefully crafted +/- pairs - how to remove this constraint?
  - SAEs assume rank-1 decompositions - should we be designing low-rank decompositions?

Thank you!

Andrew Lee
Harvard



Yonatan Belinkov
Technion - IIT,
Kempner Institute

Fernanda Viégas
Harvard,
DeepMind

Martin
Wattenberg
Harvard,
DeepMind