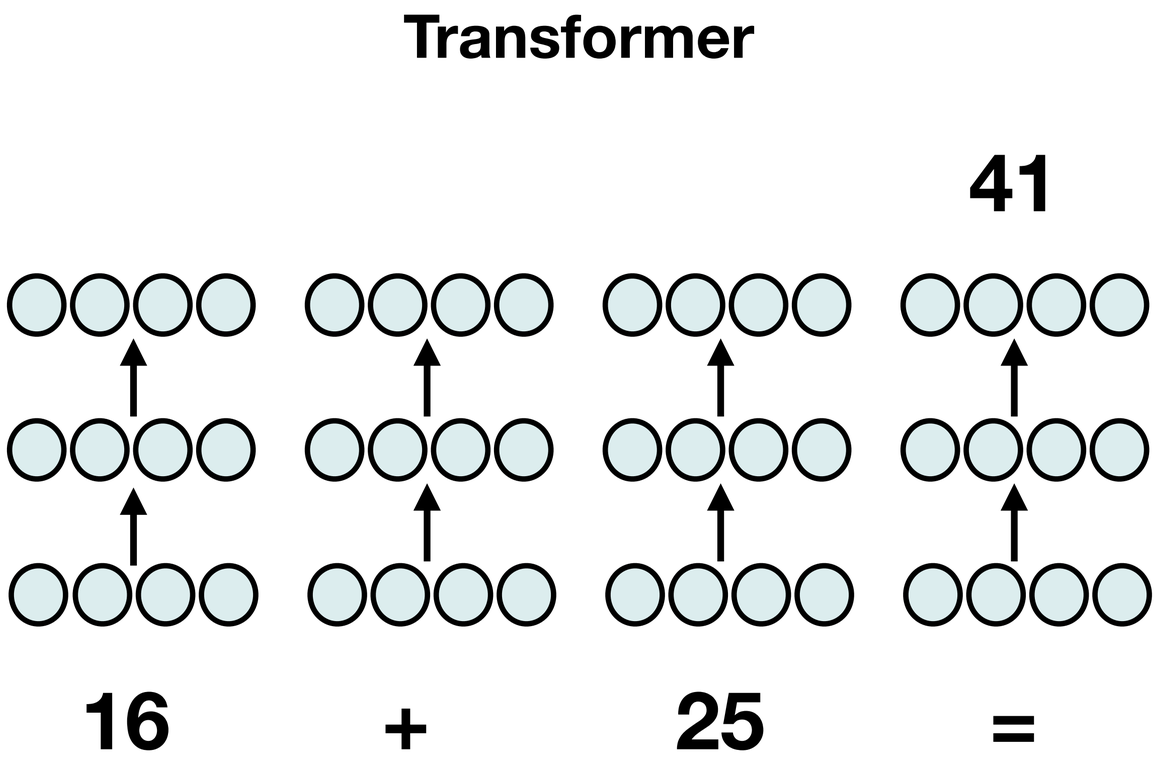
# Reliably steering and interpreting LLMs with *causal representation learning*

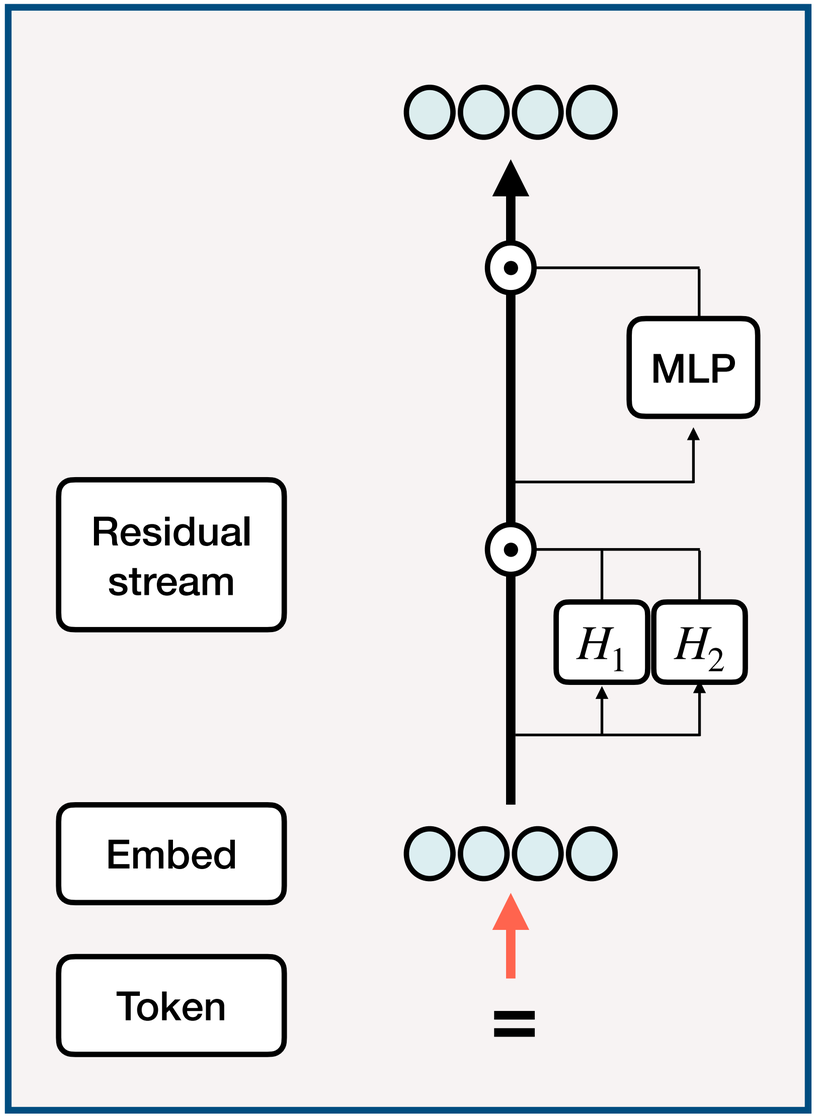
Dhanya Sridhar

April 16, 2025

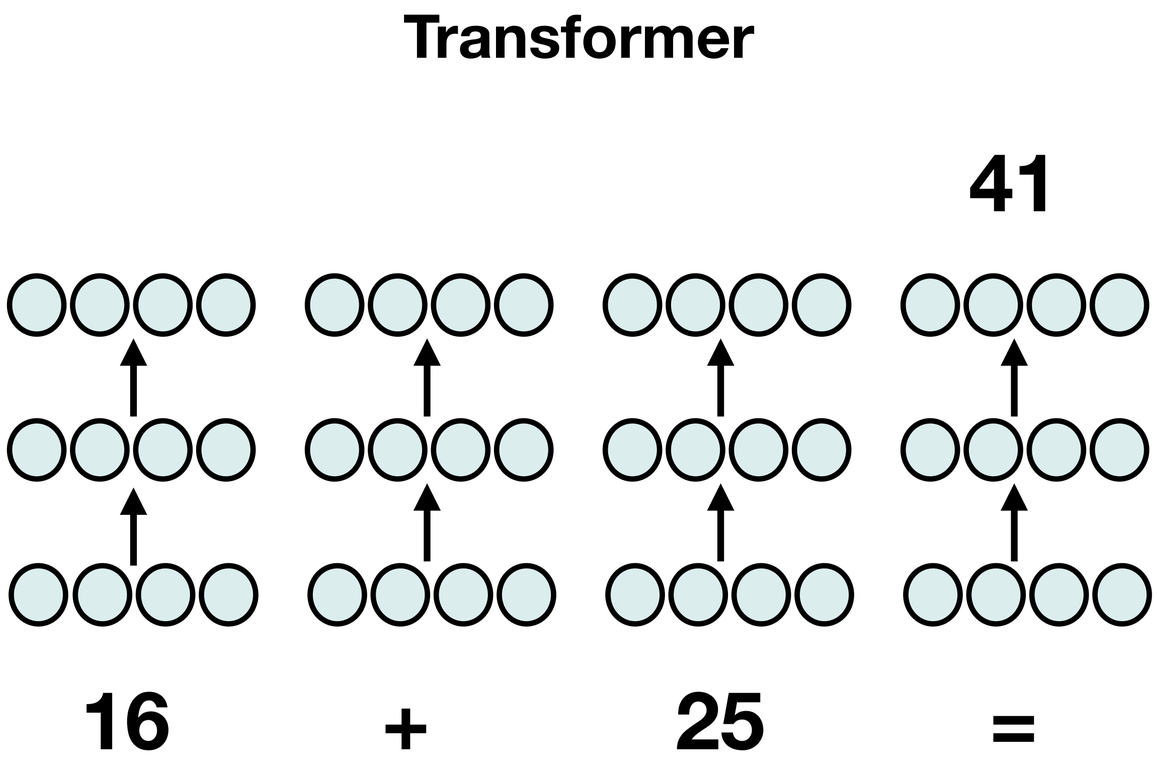


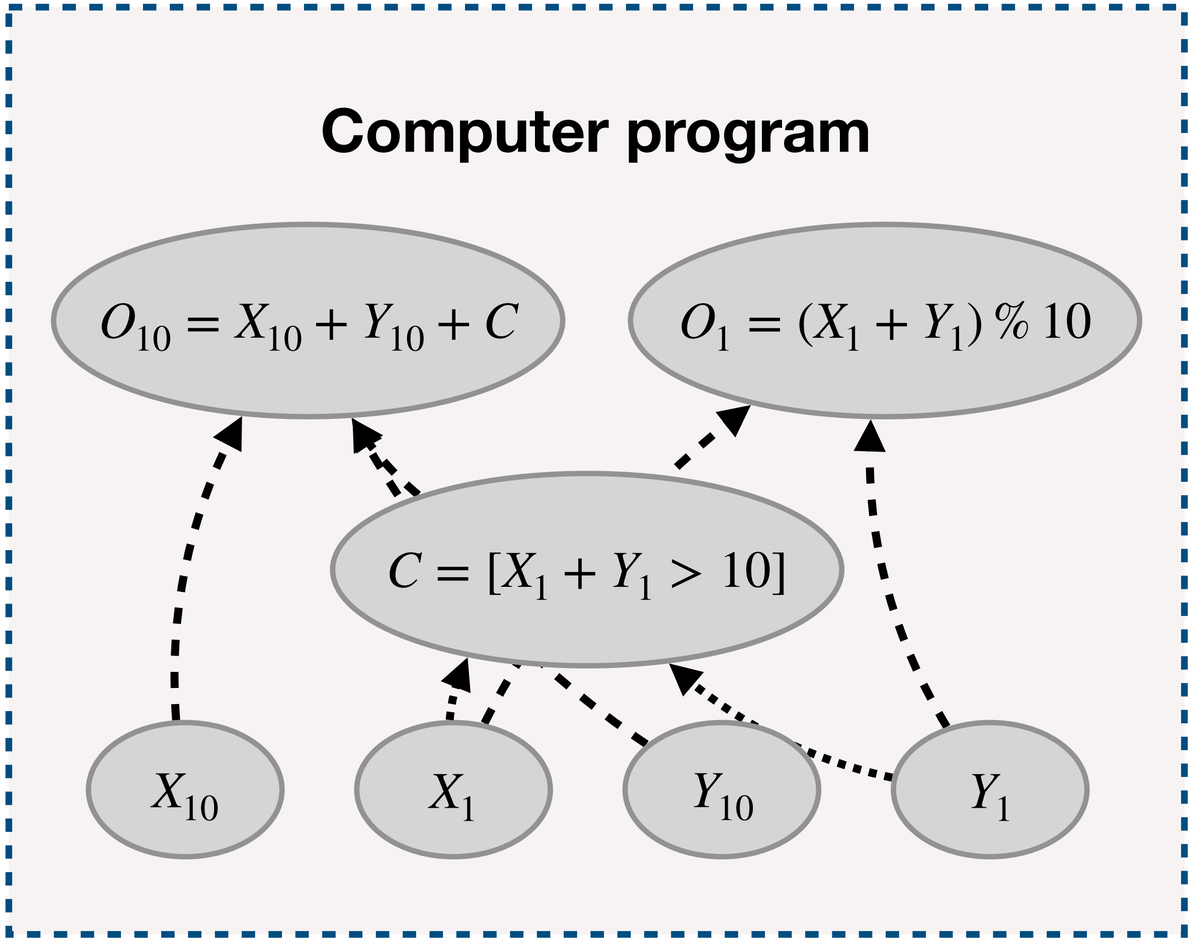

# Simple example: Addition of two digit numbers



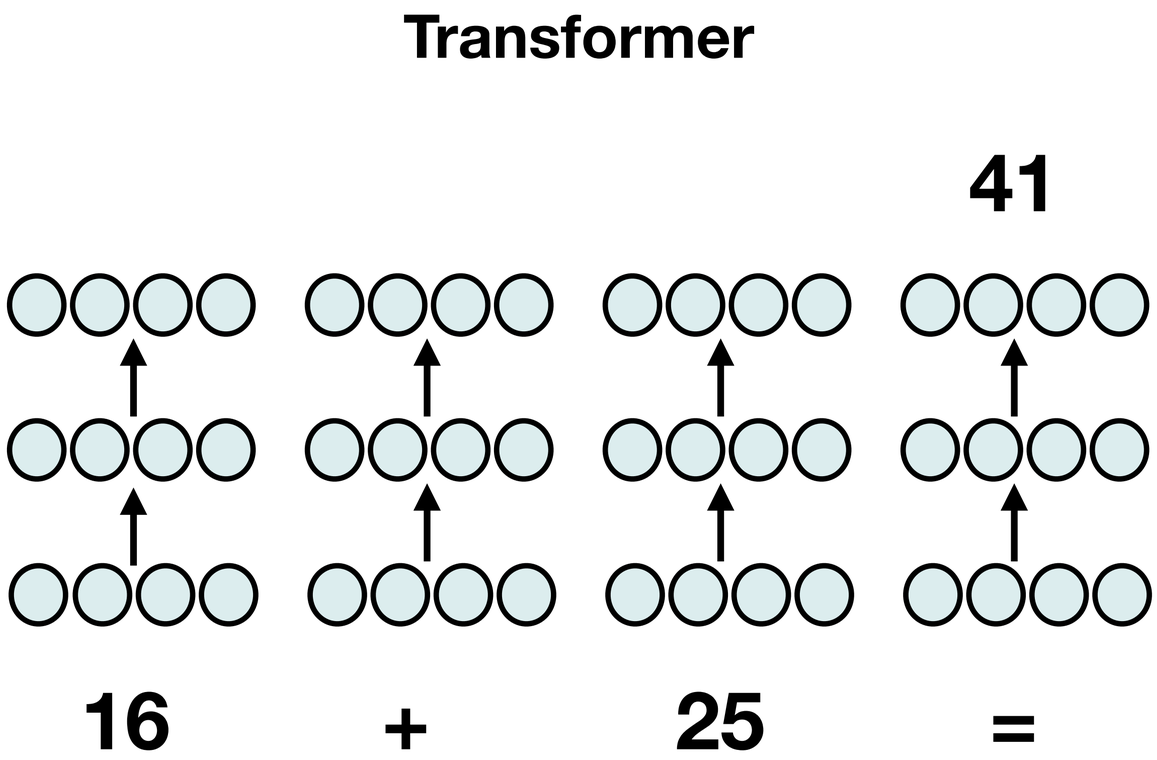



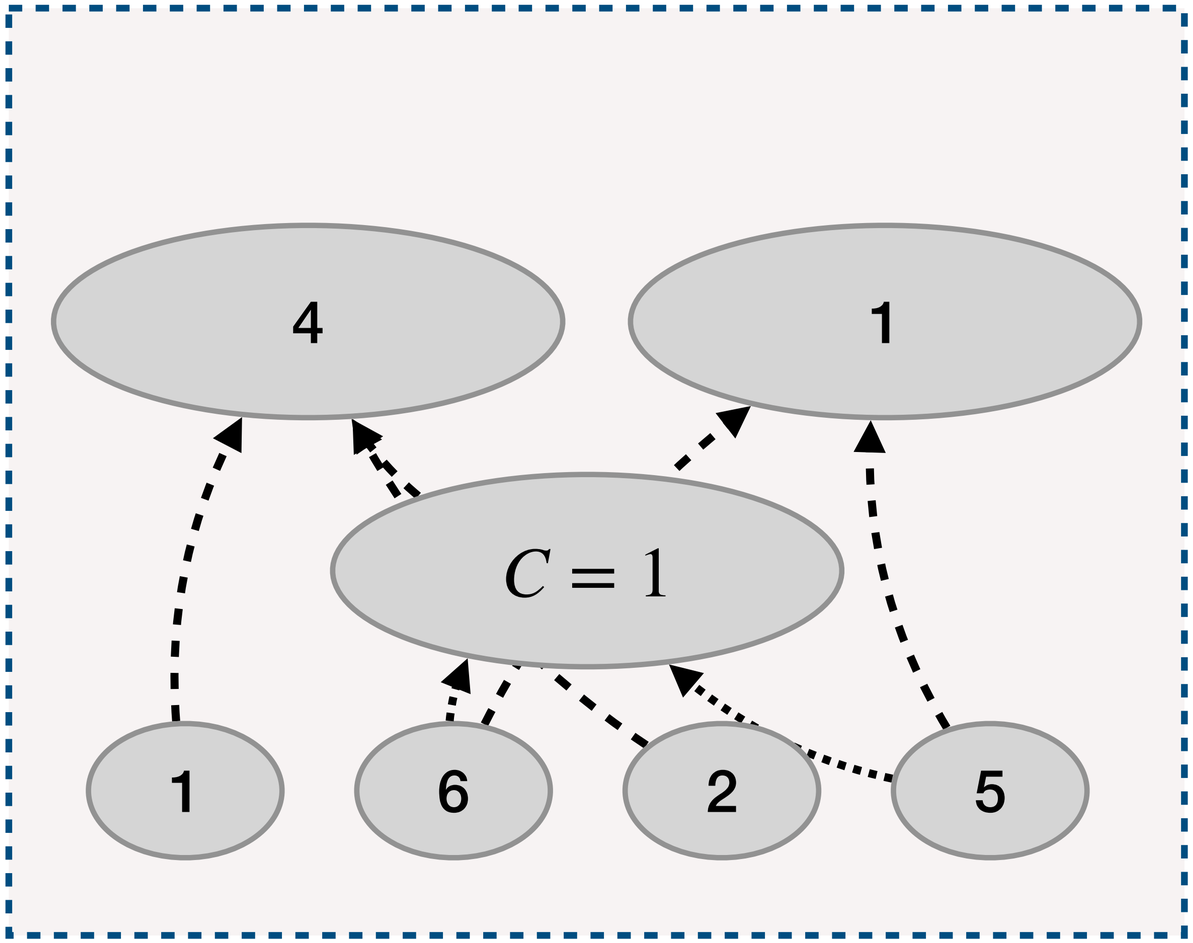
# Simple example: Addition of two digit numbers



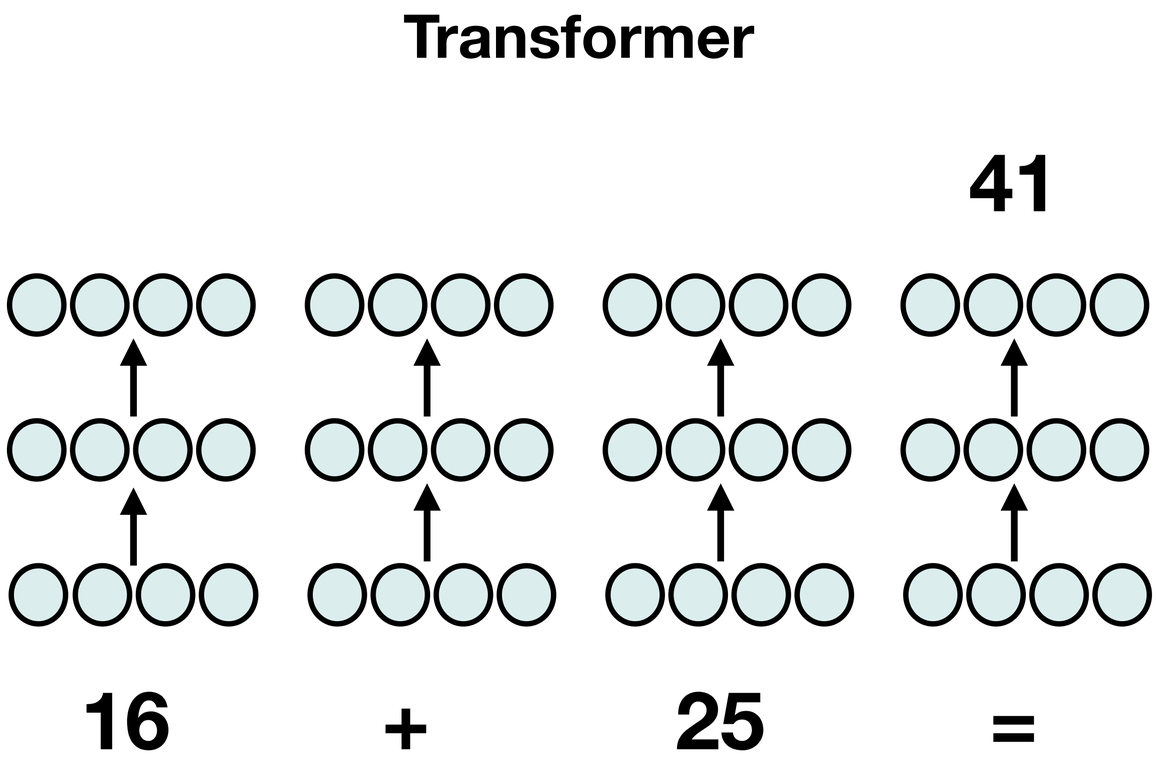



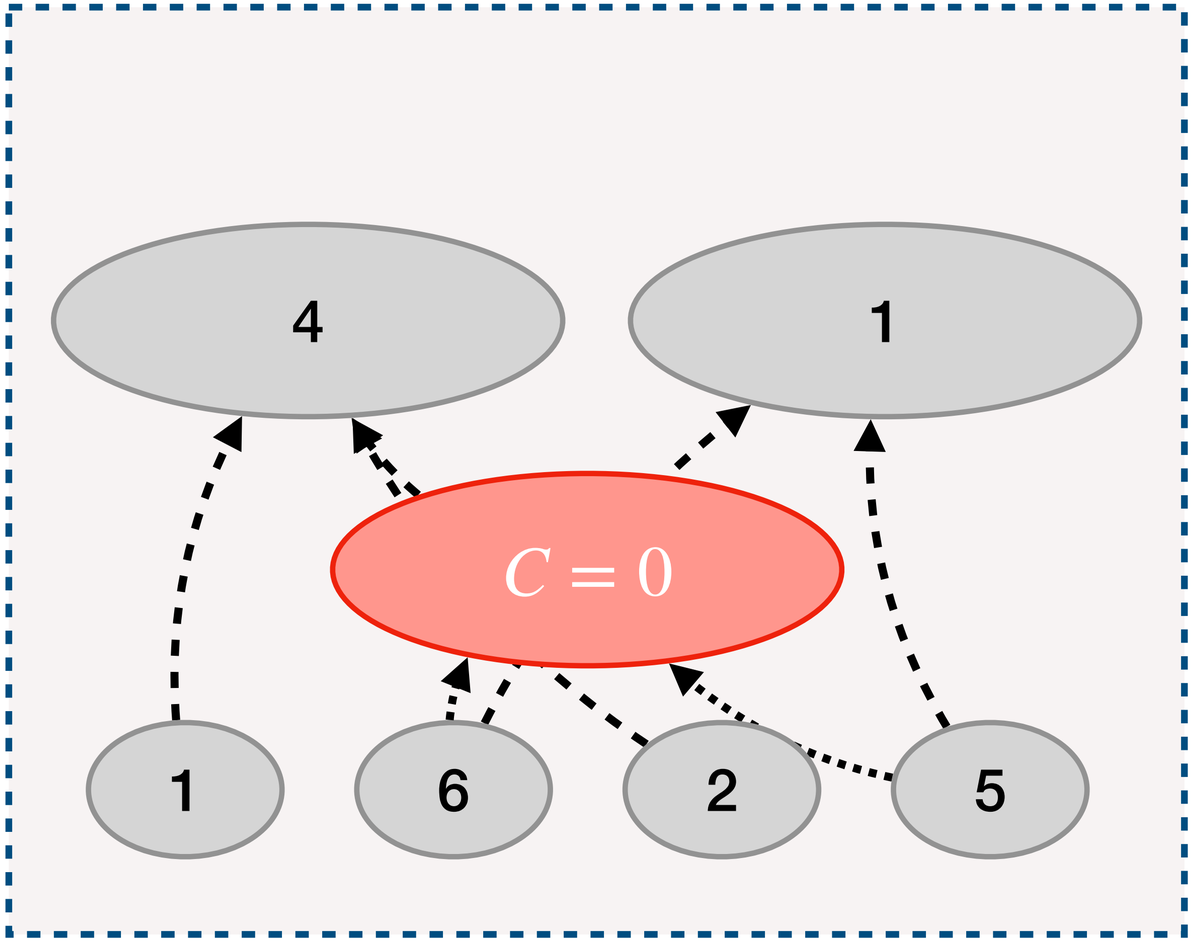
# Simple example: Addition of two digit numbers



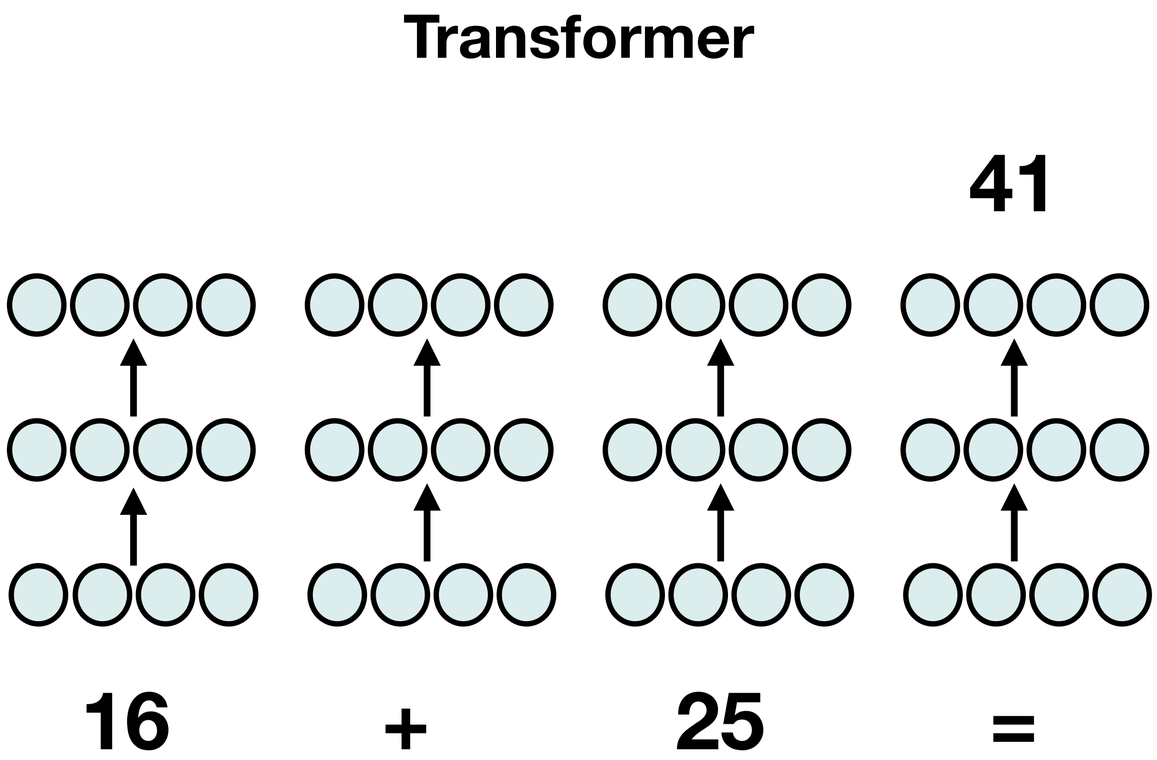



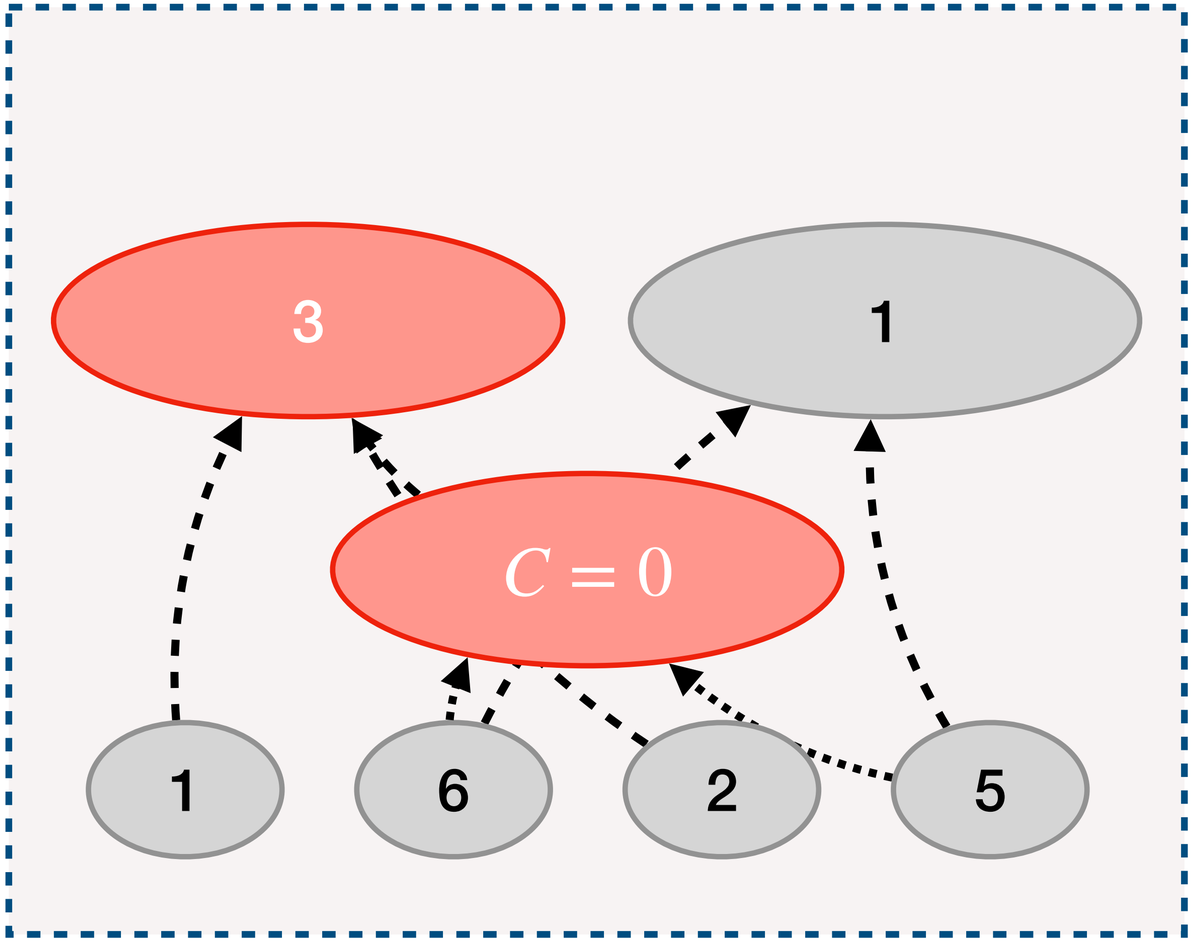
# Simple example: Addition of two digit numbers



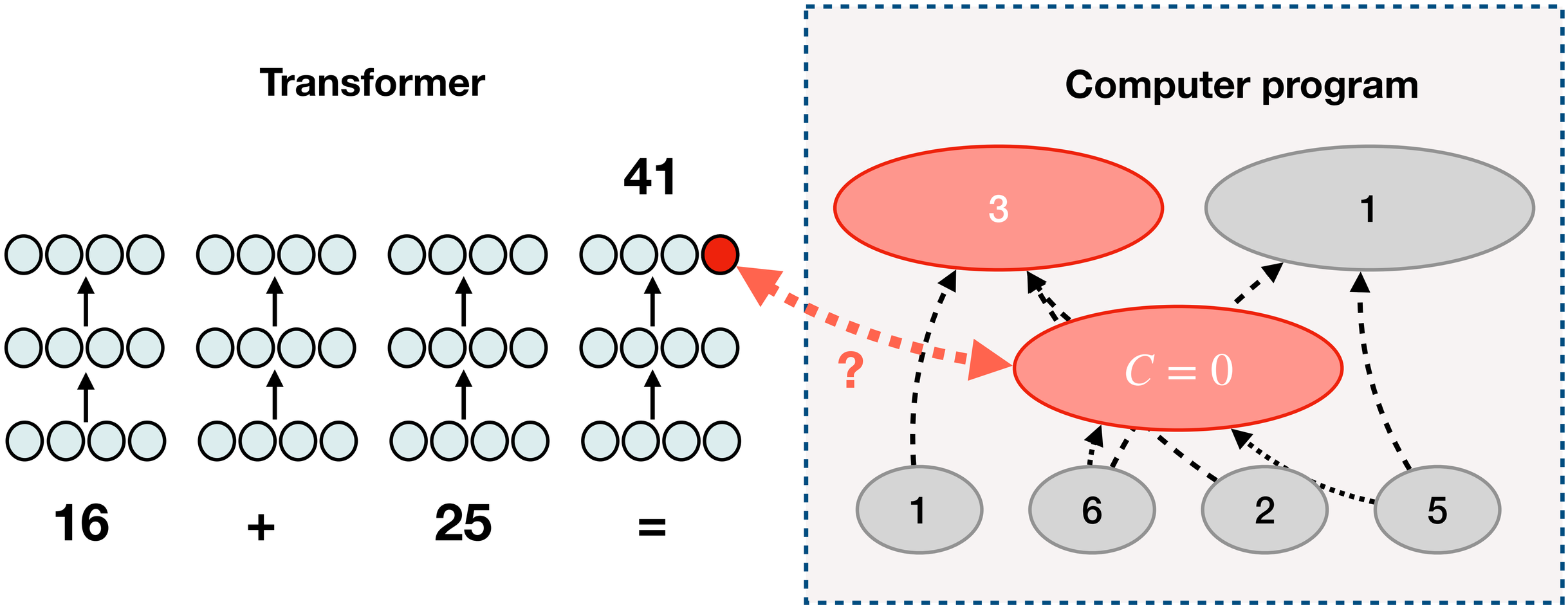



# Simple example: Addition of two digit numbers

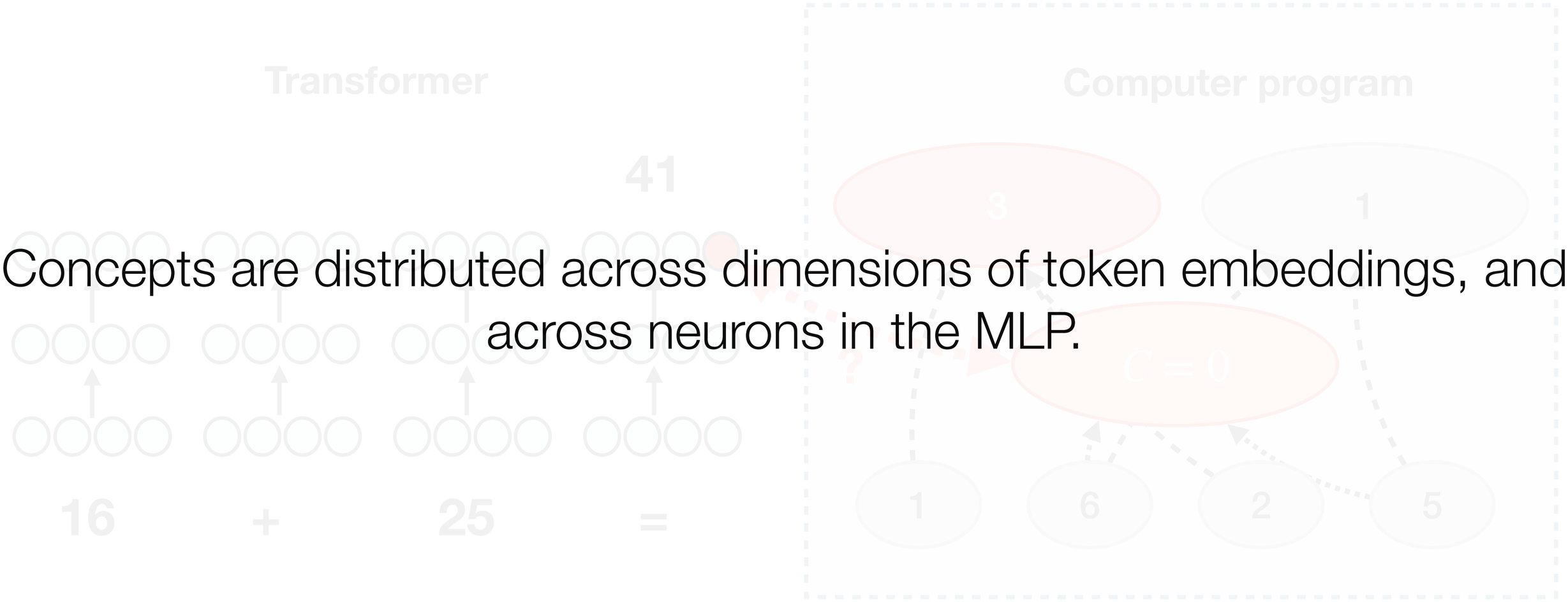


# Simple example: Addition of two digit numbers

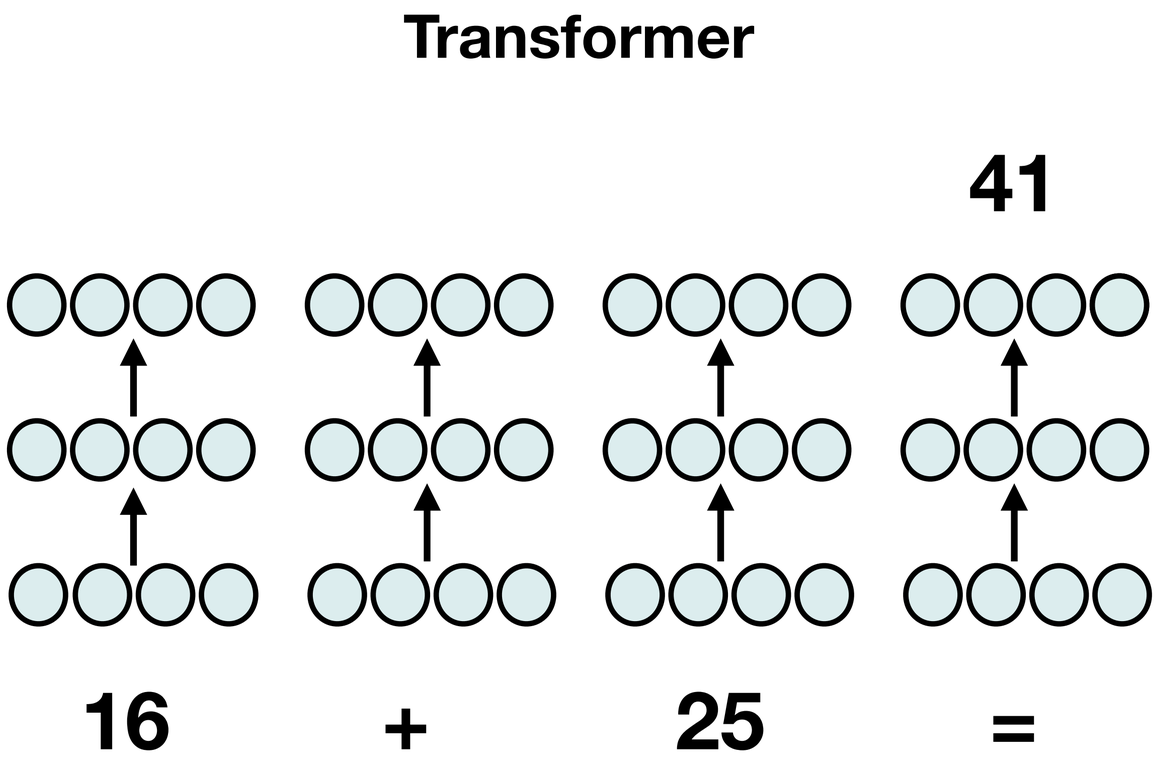


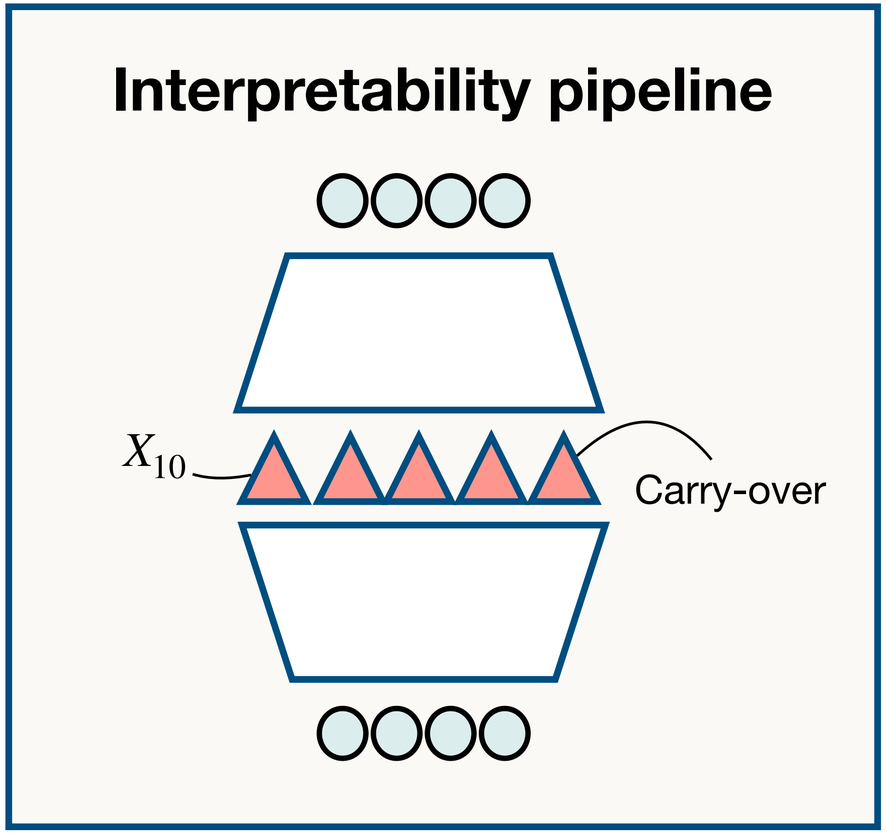
# Simple example: Addition of two digit numbers



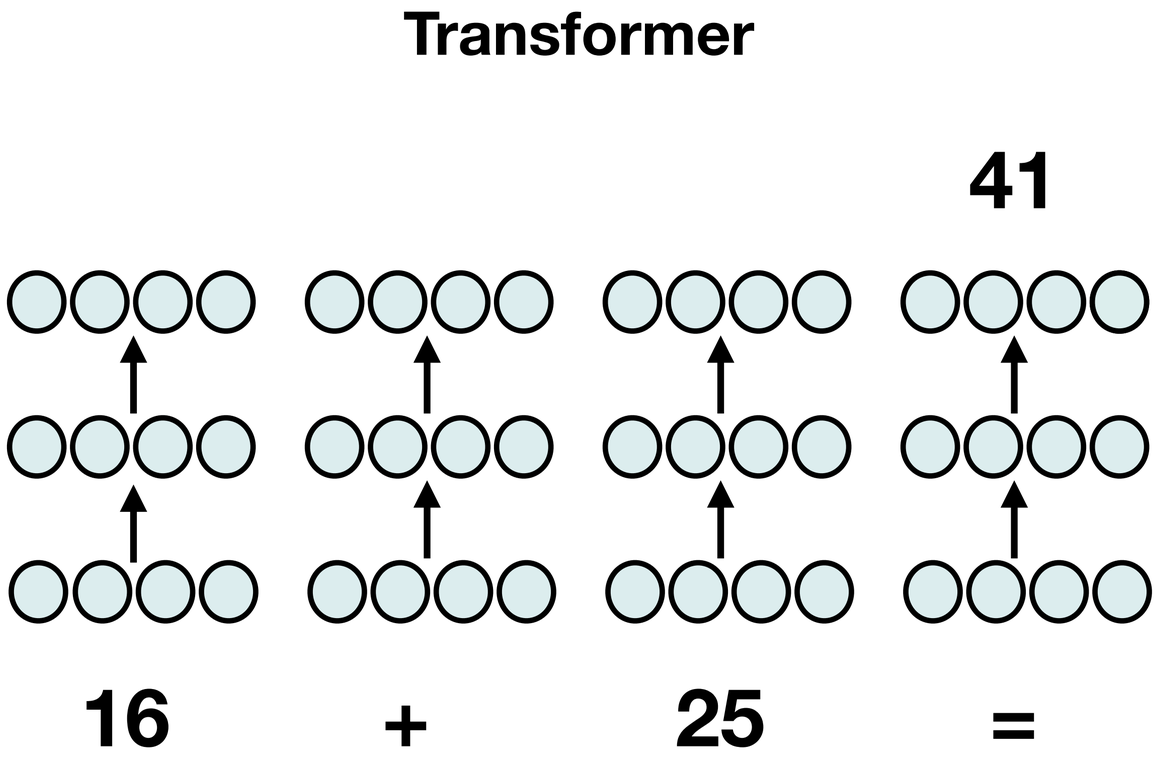

Concepts are distributed across dimensions of token embeddings, and across neurons in the MLP.

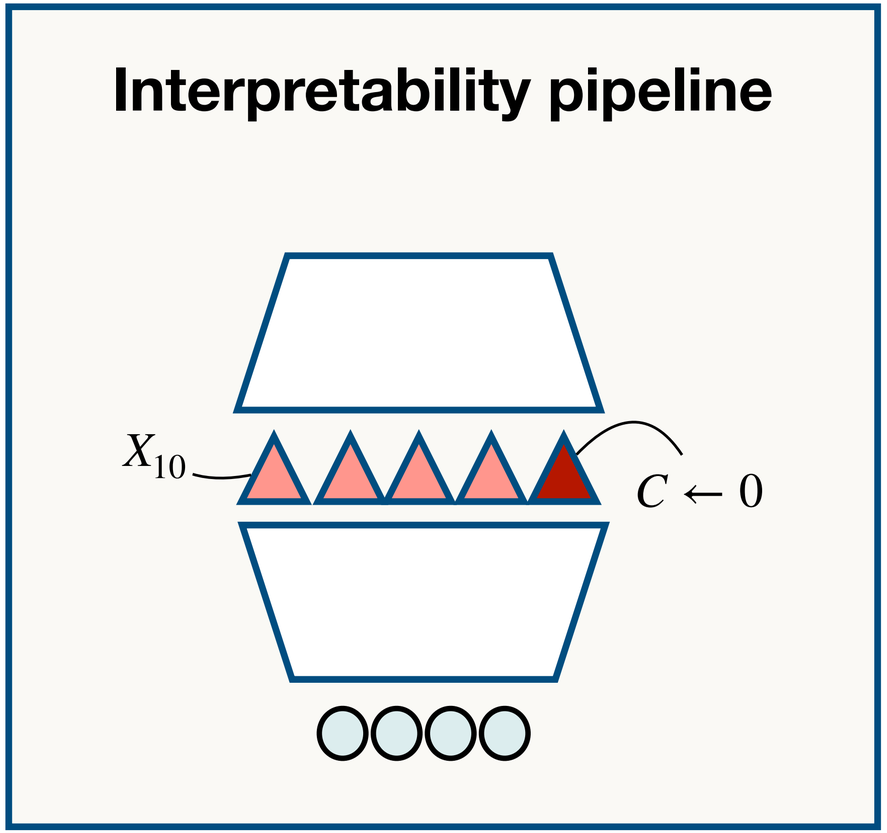
# Unsupervised feature discovery



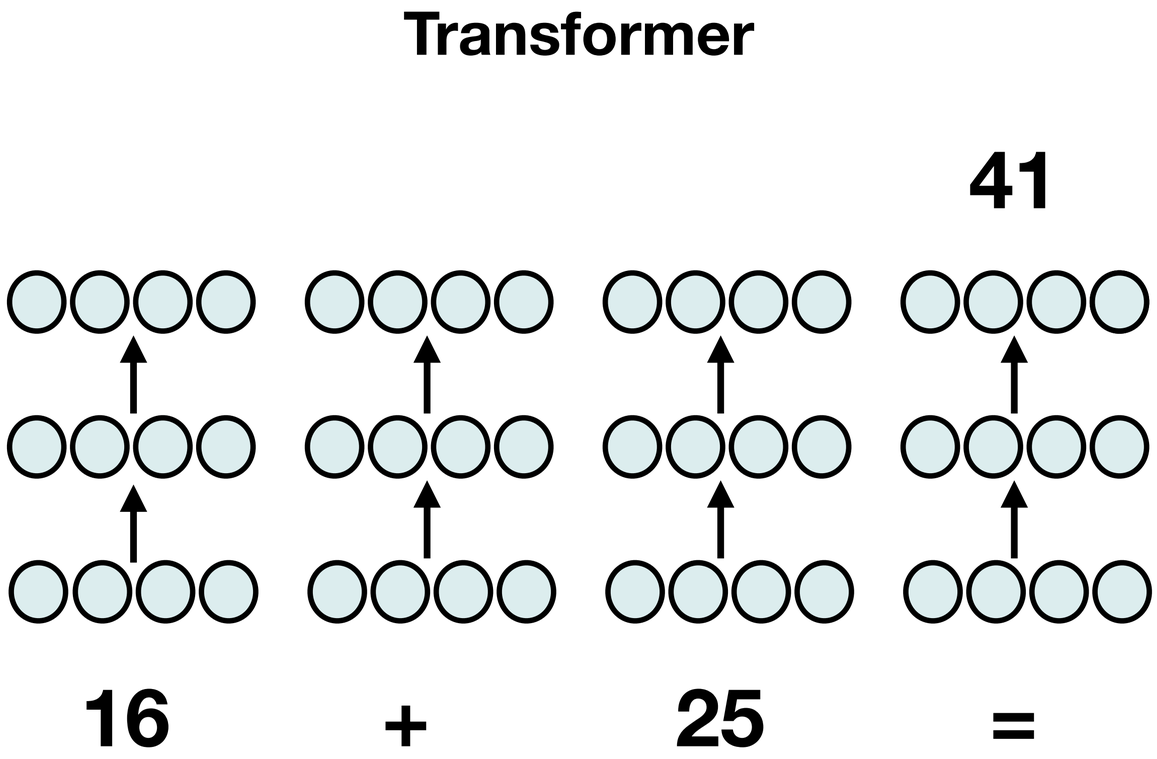



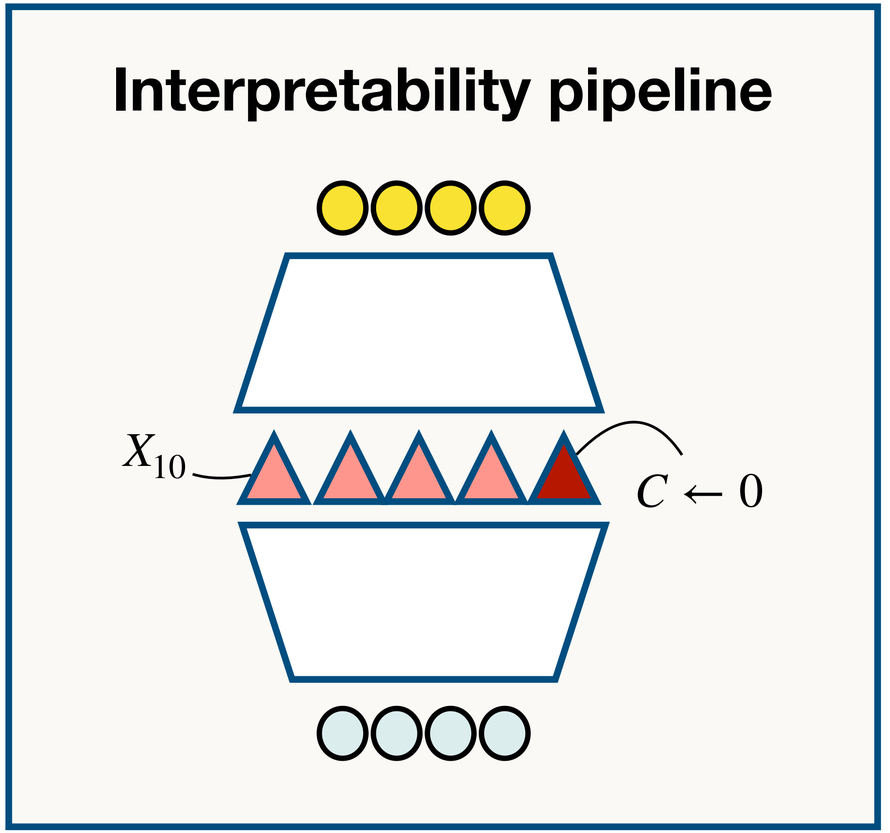
# Unsupervised feature discovery



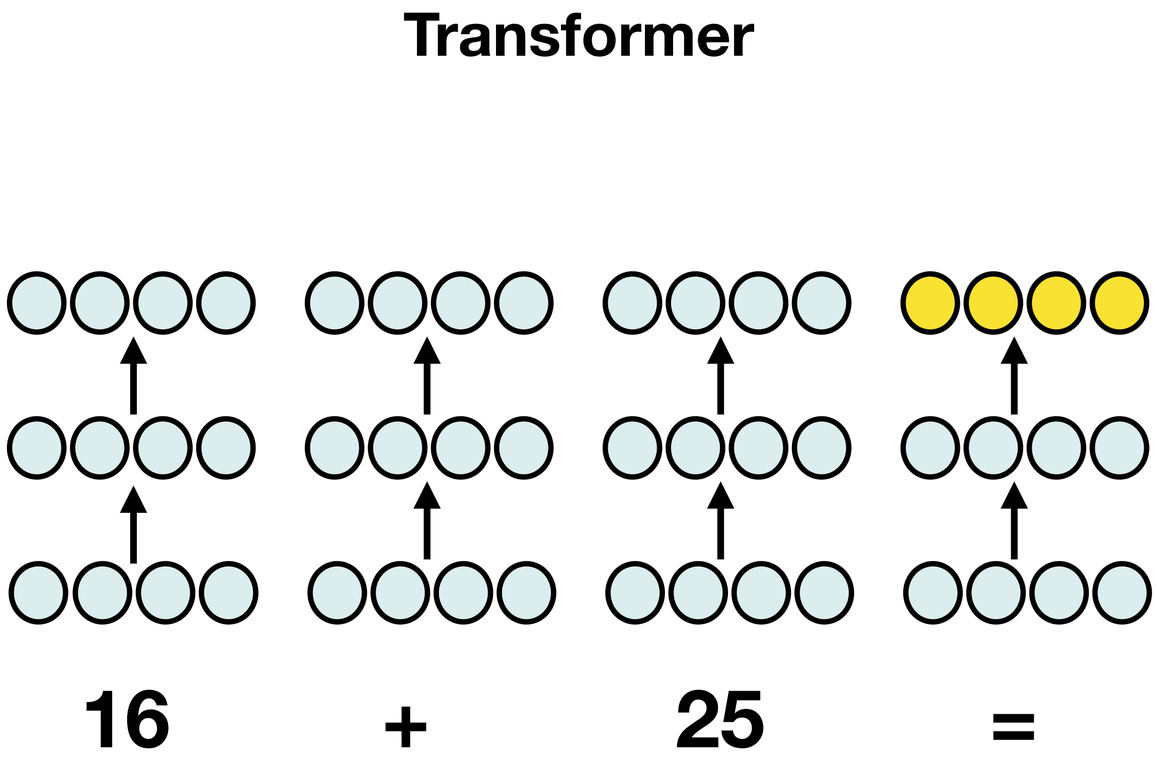



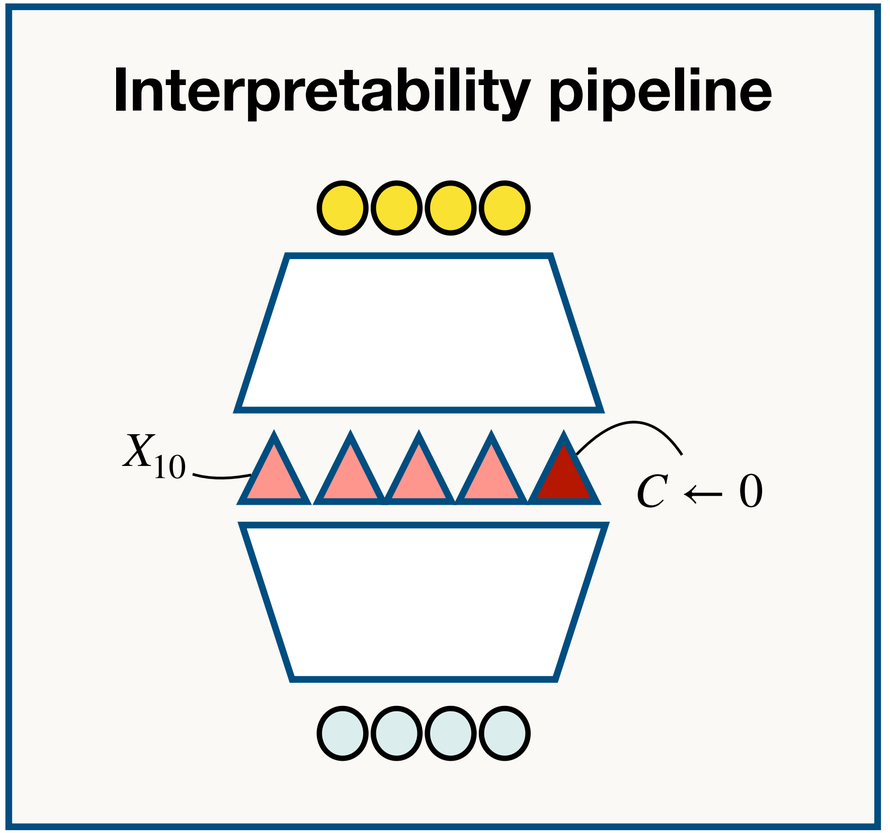
# Unsupervised feature discovery



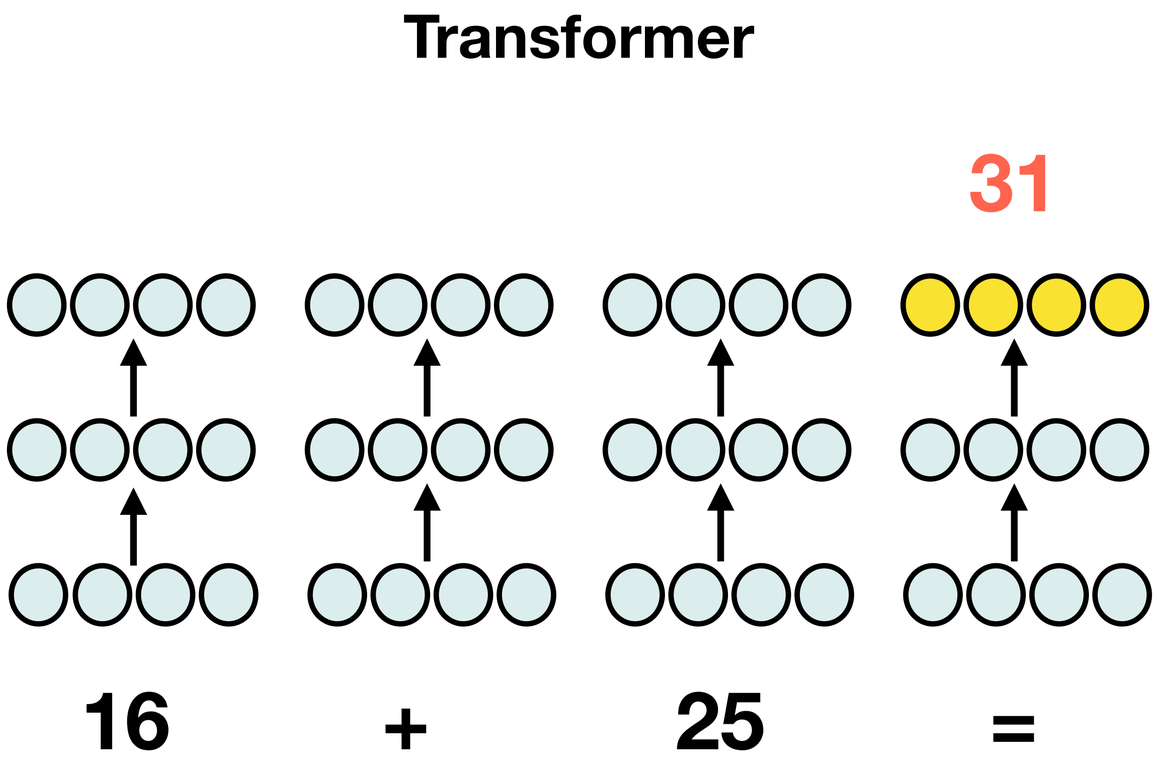



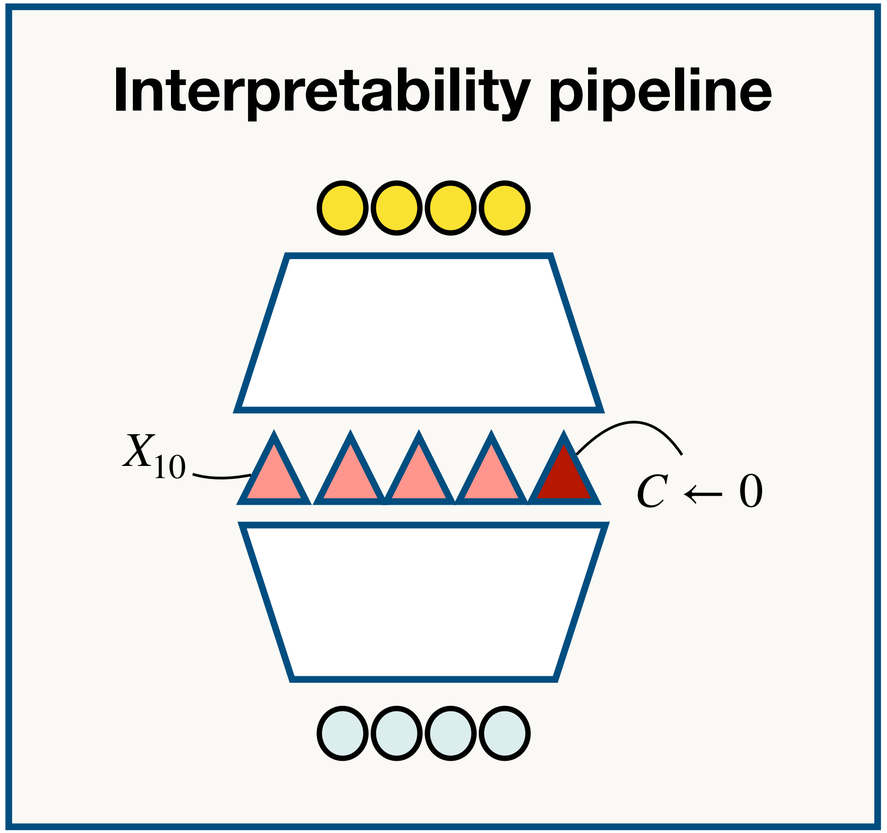
# Unsupervised feature discovery



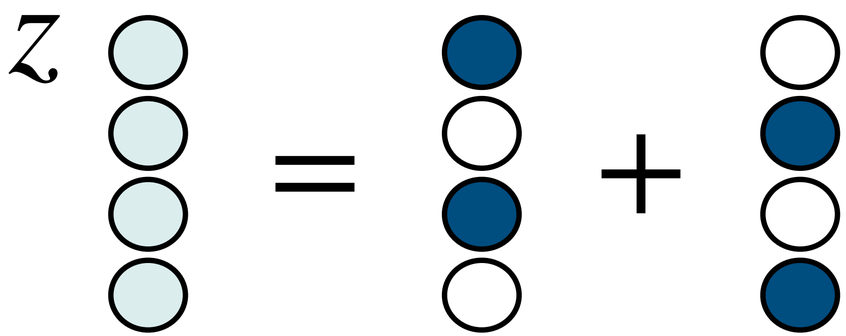
# Unsupervised feature discovery



# Sparse auto-encoders (SAEs)

**Represent $z$ as a sparse sum of concept directions.**




B.A. Olshausen, D.J. Field. 1997. *Sparse coding with an overcomplete basis set: A strategy employed by V1?*

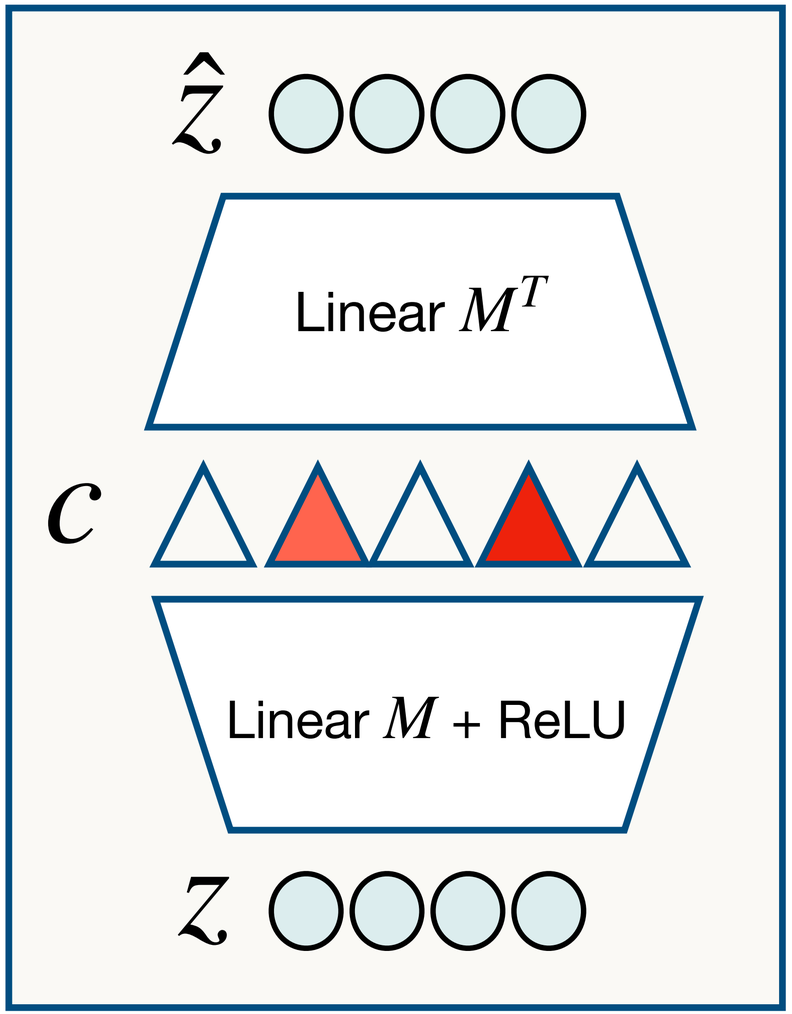
Elad, M. 2010. *Sparse and redundant representations: from theory to applications in signal and image processing.*

# Sparse auto-encoders (SAEs)

**Objective:** $\min_M \ \mathbb{E}_z\left[\|z - \hat{z}\|_2^2\right] + \alpha\|c\|_1$

$$c = \mathrm{ReLU}(Mz + b),\ \hat{z} = M^T c$$

- Adopted in interpretability community around 2023[1,2].
- Assume concepts are represented linearly[3].
- ***If concepts are disentangled,*** each column of the decoder gives us a steering direction.

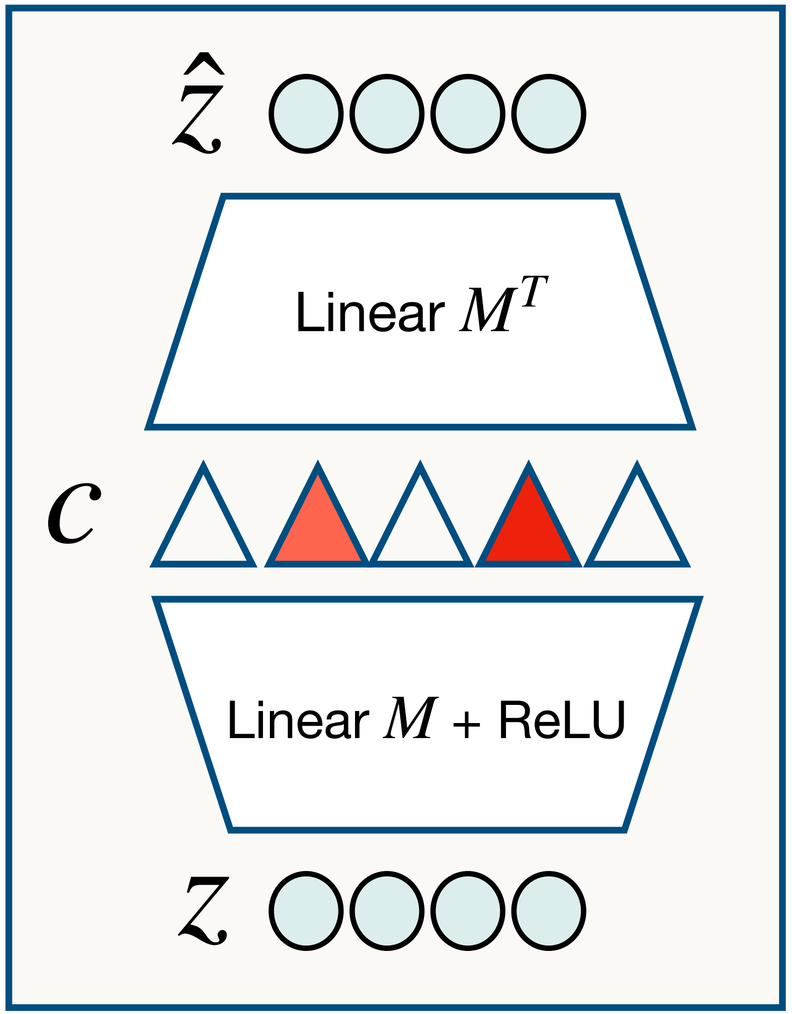


# Sparse auto-encoders (SAEs)

**Objective:** $\min_M \ \mathbb{E}_z\left[\|z - \hat{z}\|_2^2\right] + \alpha\|c\|_1$

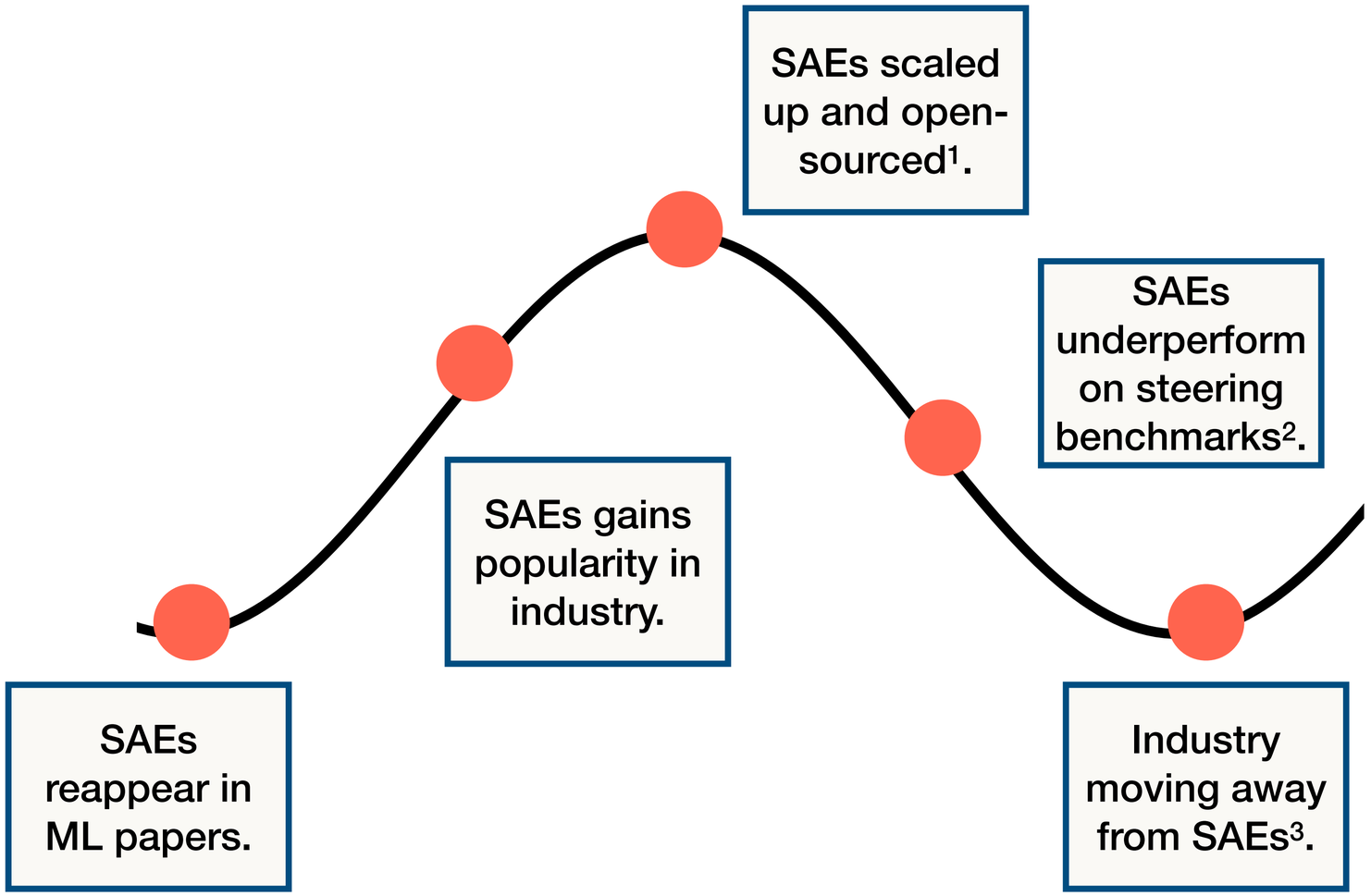
$$c = \mathrm{ReLU}(Mz + b),\ \hat{z} = M^T c$$

- Adopted in interpretability community around 2023[1,2].
- Concepts are represented linearly[3,4].
- ***If concepts are disentangled,*** each column of the decoder gives us a steering direction.




[1]Cunningham et al., 2023. *Sparse Autoencoders Find Highly Interpretable Features in Language Models.* [2]https://transformer-circuits.pub/2023/monosemantic-features.

[3]Mikolov et al., 2013. *Linguistic Regularities in Continuous Space Word Representations.* [4]Jiang et al., 2024. *On the origins of linear representations in large language models.*

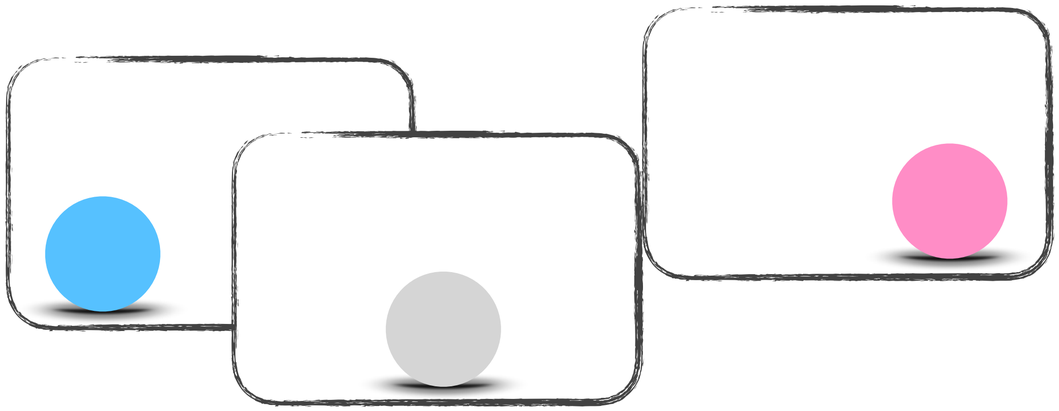



[1]Lieberum et al. 2024. *Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2*

[2]Wu et al. 2025. *AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders.*

[3]https://deepmindsafetyresearch.medium.com/negative-results-for-sparse-autoencoders-on-downstream-tasks-and-deprioritising-sae-research-6cadcfc125b9

We need unsupervised methods that provably disentangle the steerable concepts.

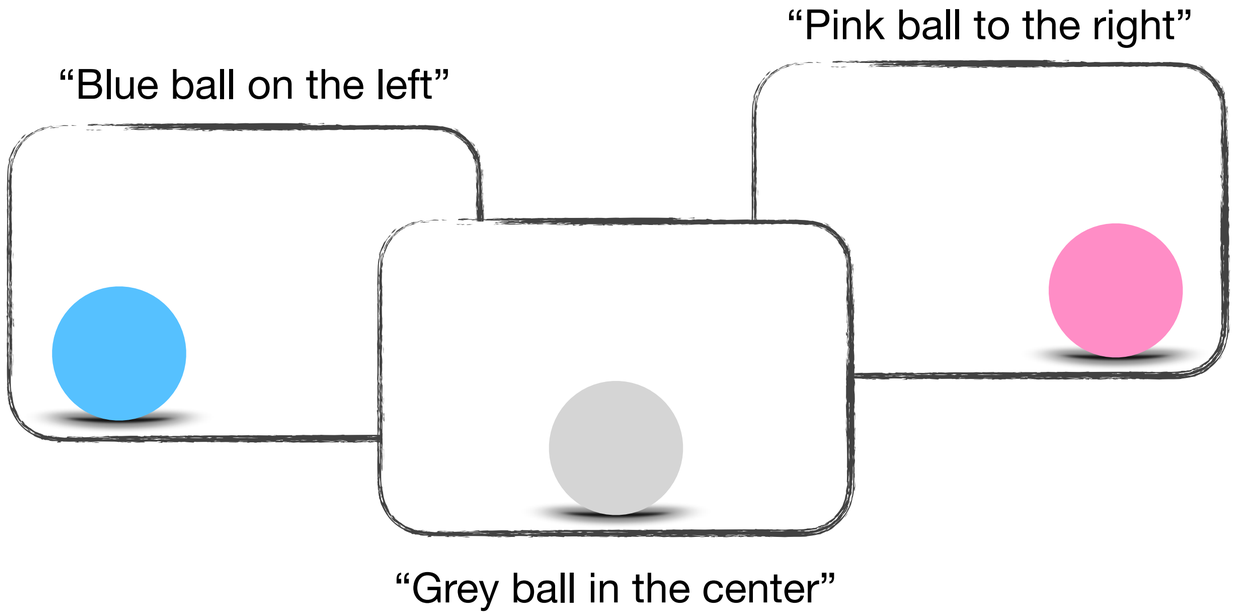


# Causal representation learning

Discovering “intervenable” features

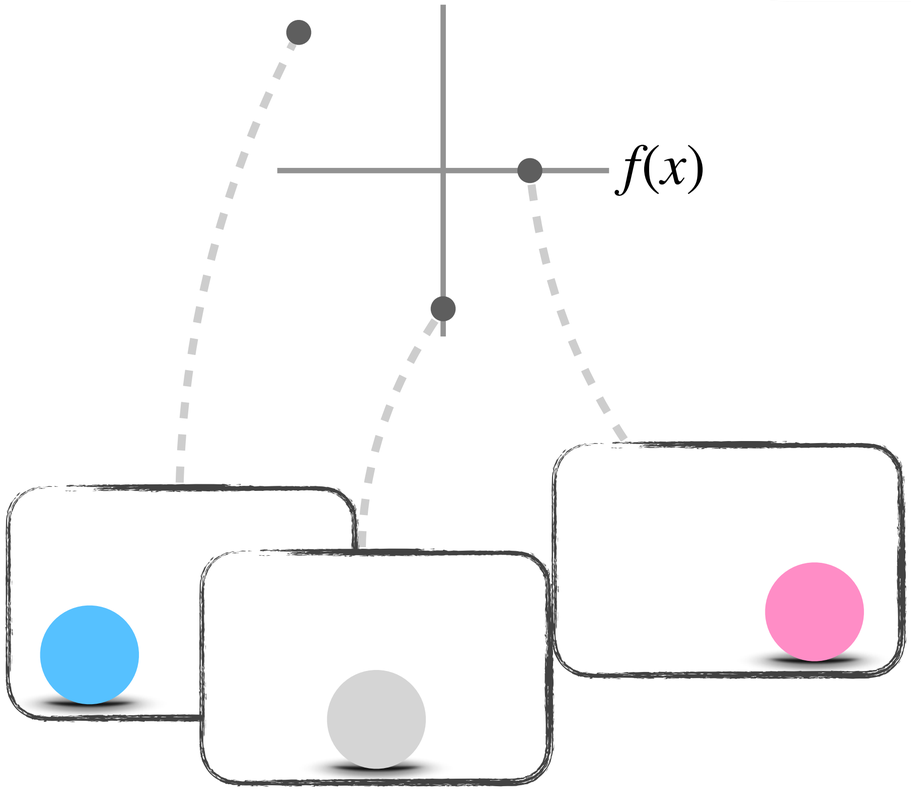
# Learning to encode images

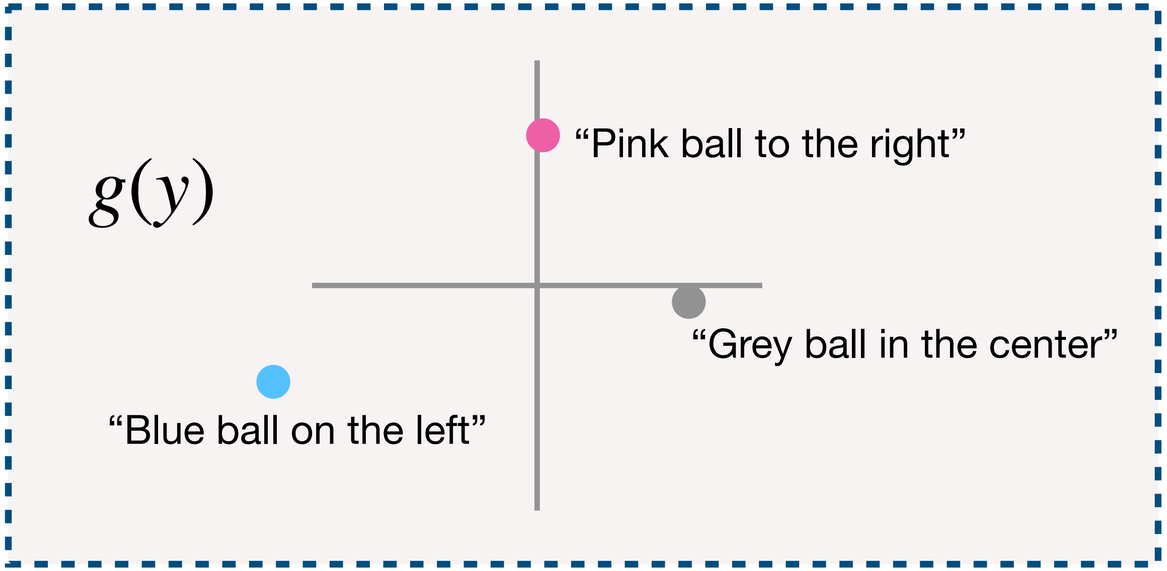
**Supervision from text**



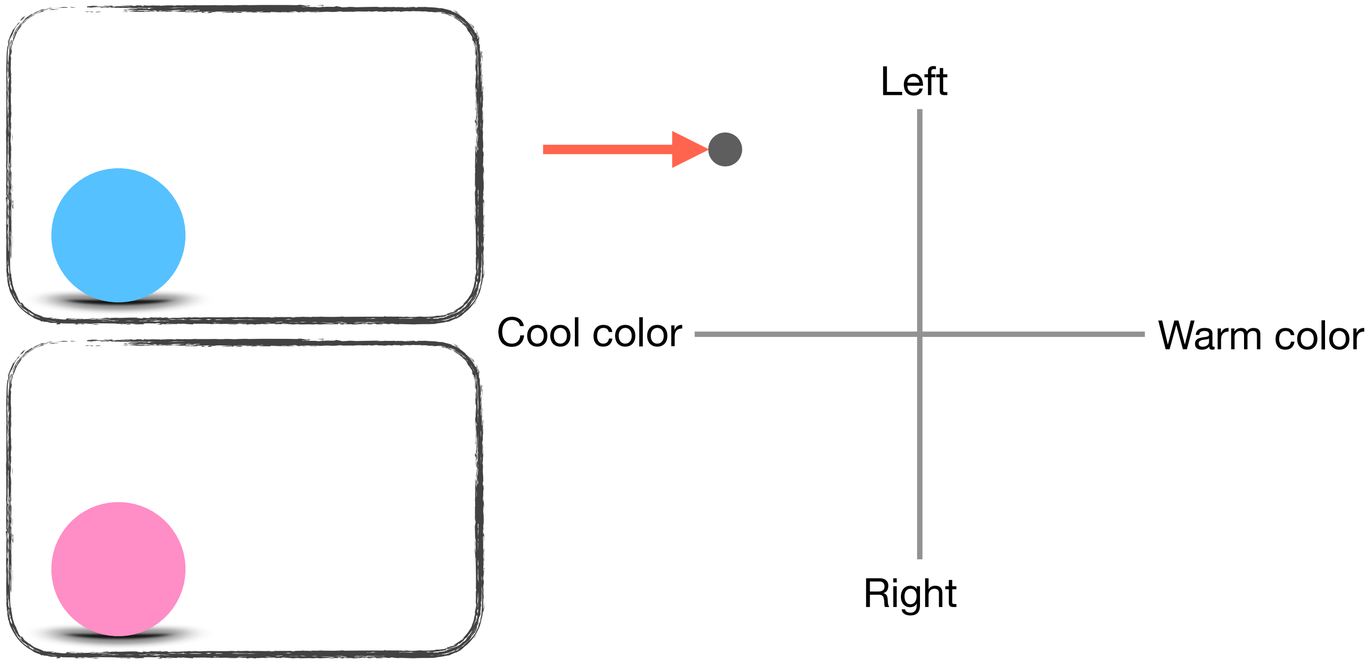
# Learning to encode images

$$\min_{f \in \mathcal{F}} \mathbb{E}_{x,y}\left[(g(y) - W^{\top} f(x))^2\right]$$



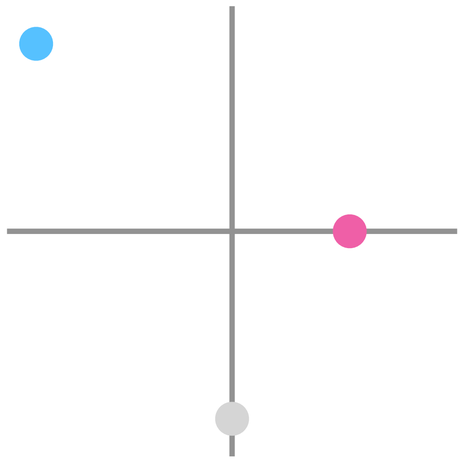
# A “nice” choice of basis

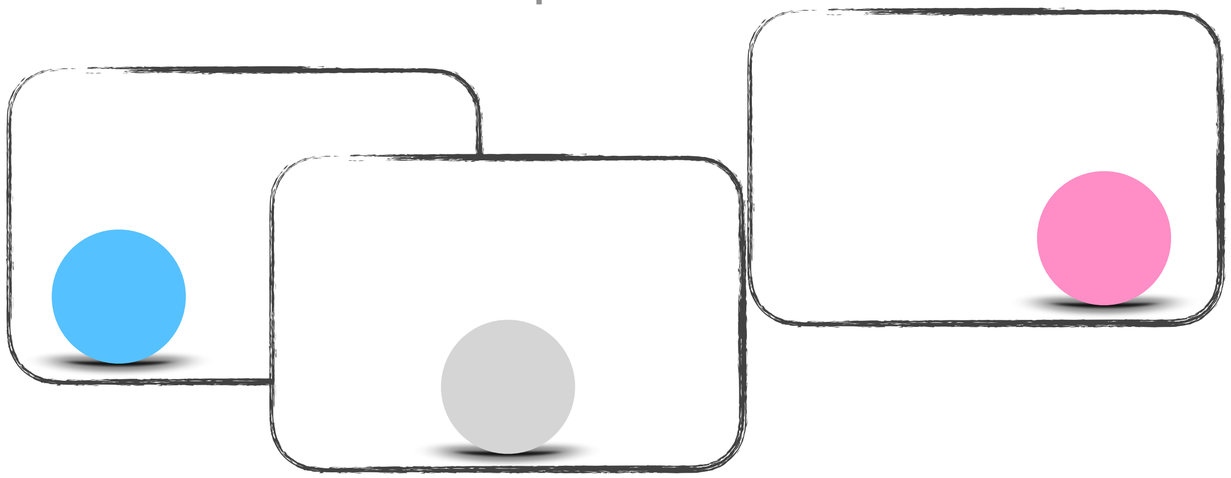



These axes align with concepts we want to intervene upon in isolation.

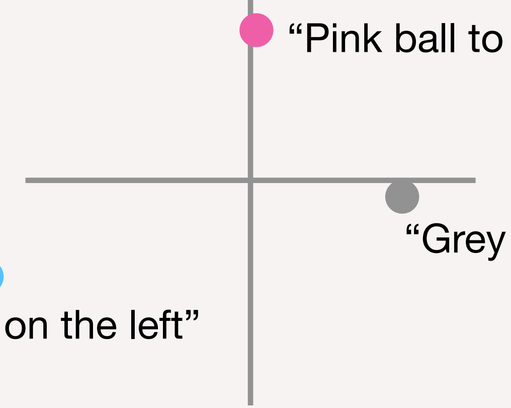
# Are we guaranteed to learn these features?

$f(x)$





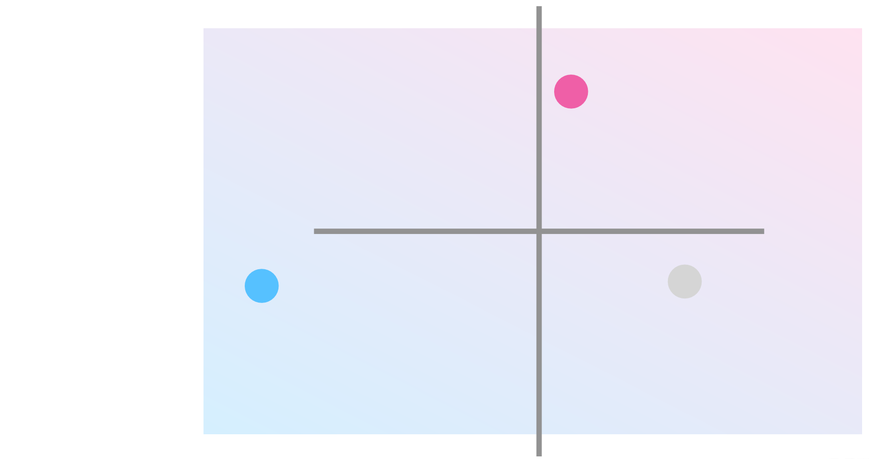
$g(y)$

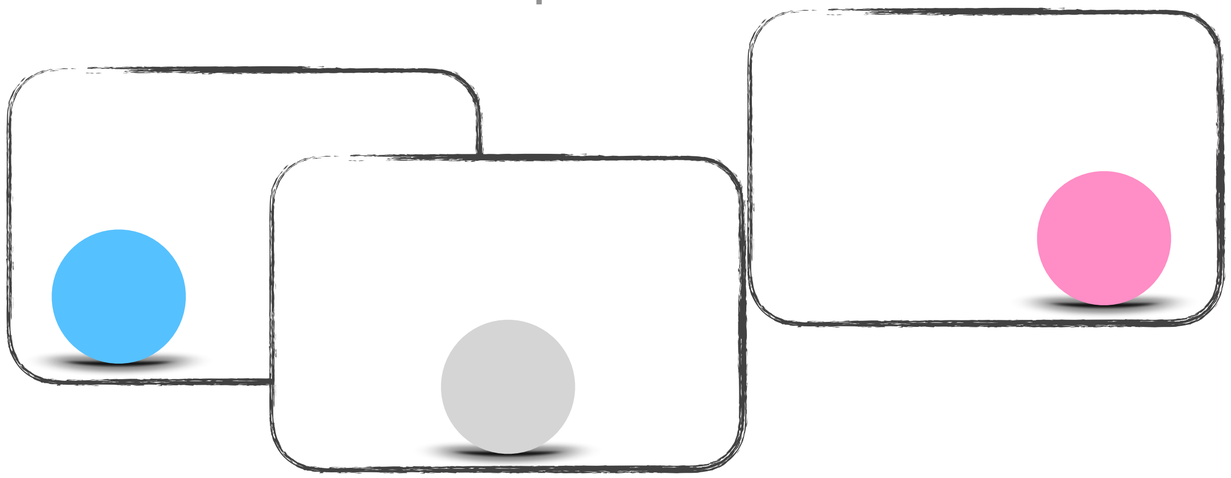



$$\min_{f \in \mathcal{F}} \mathbb{E}_{x,y}\left[(g(y) - W^{\top} f(x))^2\right]$$

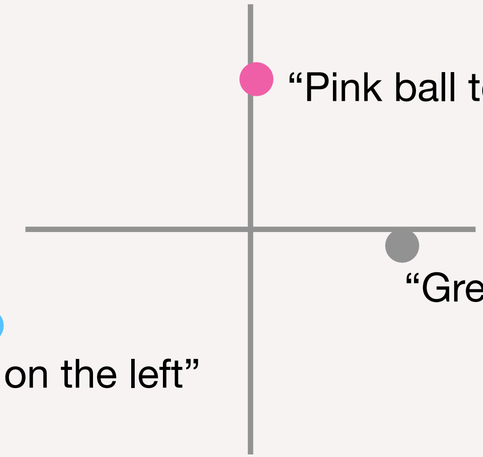
# Learning objective is underspecified!

Rotated by 63 degrees





$g(y)$



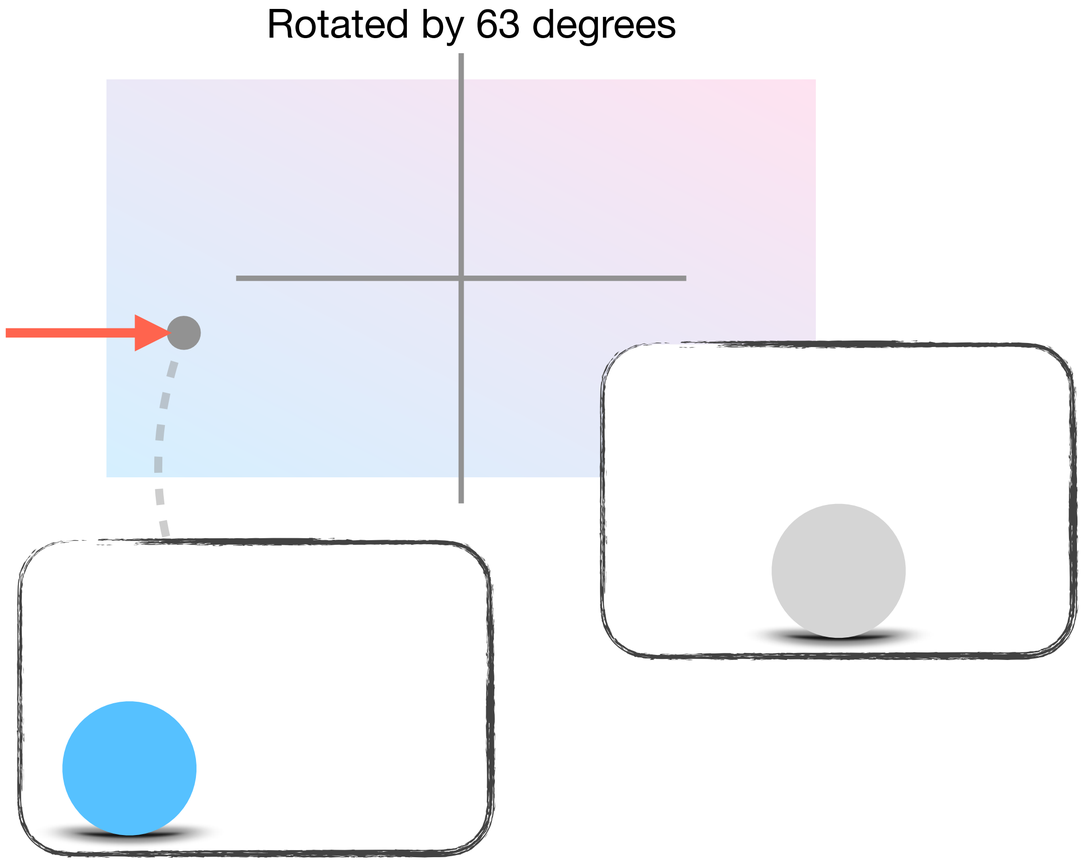
"Pink ball to the right"

"Grey ball in the center"

"Blue ball on the left"

$$\min_{f \in \mathcal{F}} \mathbb{E}_{x,y}\left[(g(y) - W^\top f(x))^2\right]$$

# This indeterminacy affects steering




Axes of the learned basis don’t align with the concepts we want to change!

Learning the vector that changes blue to pink would require labeled data.

Hyvarinen, Aapo, and Hiroshi Morioka. 2016. “Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA.

Roeder, Geoffrey, Luke Metz, and Diederik P. Kingma. 2020. “On Linear Identifiability of Learned Representations.”

***Ahuja, Kartik, Jason Hartford, and Yoshua Bengio. 2022. “Weakly Supervised Representation Learning with Sparse Perturbations.”***

Locatello, Francesco, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. 2020. “Weakly-Supervised Disentanglement Without Compromises.

Brehmer, Johann, Pim de Haan, Phillip Lippe, and Taco Cohen. 2022. “Weakly Supervised Causal Representation Learning.

Ahuja, Kartik, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. 2023. “Interventional Causal Representation Learning.”

Gresele, Luigi, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. 2022. “Independent Mechanism Analysis, a New Concept?”

Kügelgen, Julius von, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. 2023. “Nonparametric Identifiability of Causal Representations from Unknown Interventions.”

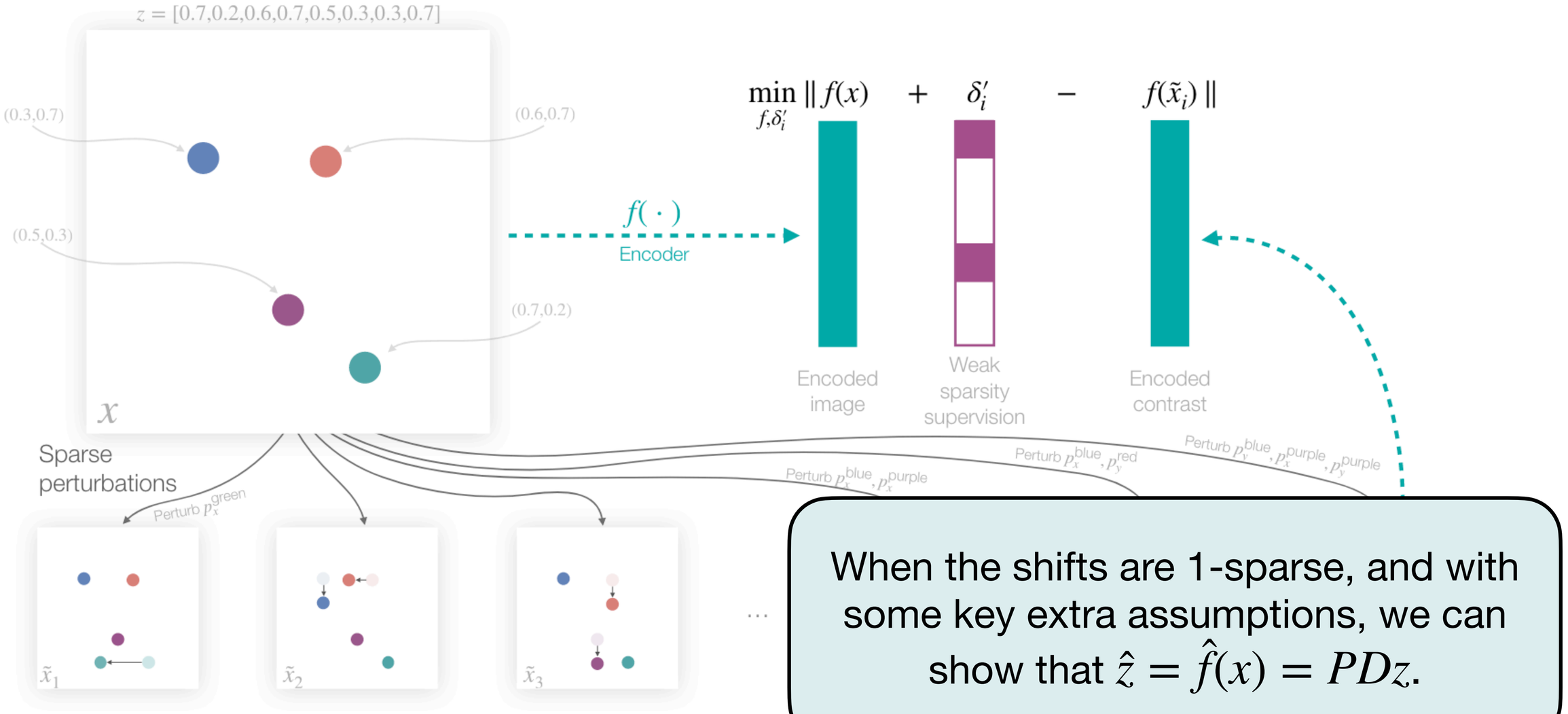
Lachapelle, Sébastien, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. 2022. “Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA.”

Lippe, Phillip, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. 2022. “CITRIS: Causal Identifiability from Temporal Intervened Sequences.”

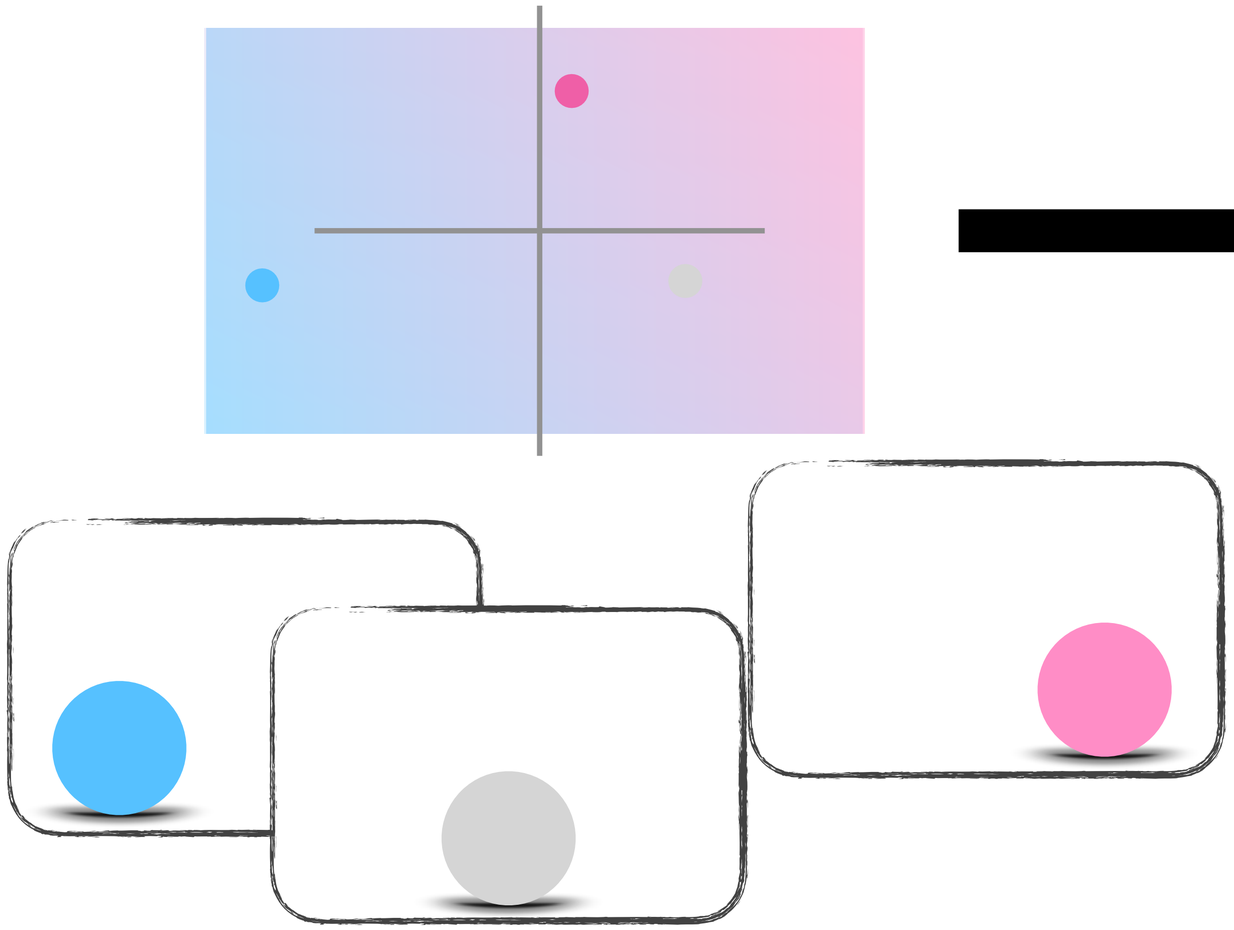
Varıcı, Burak, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. 2025. “Score-Based Causal Representation Learning: Linear and General Transformations.”

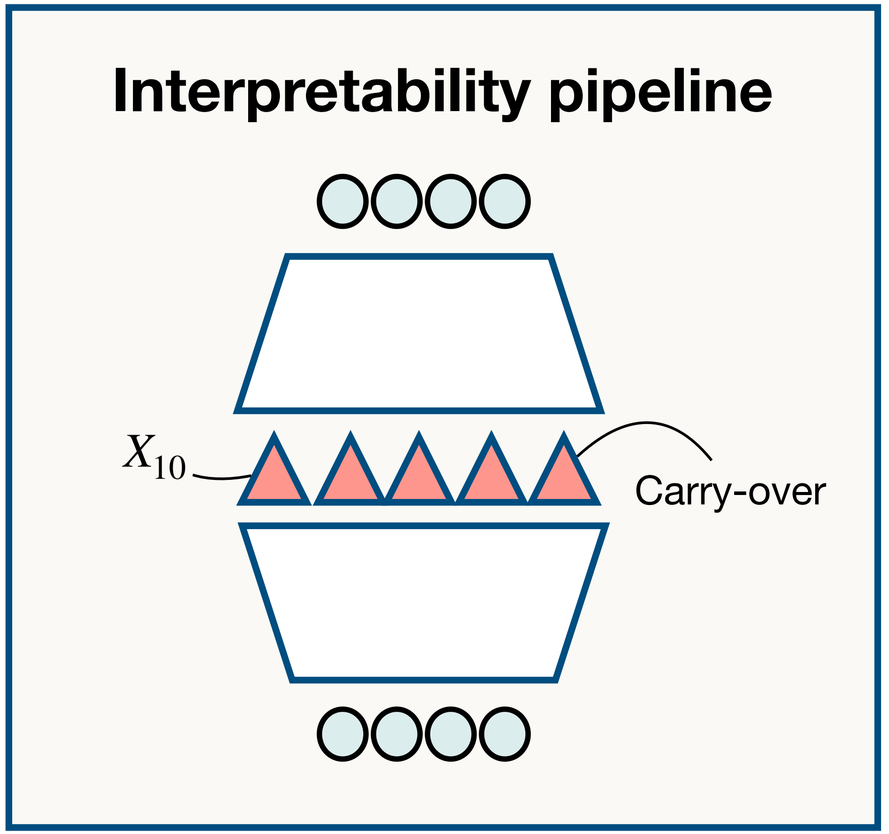
Zhang, Jiaqi, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. 2023. “Identifiability Guarantees for Causal Disentanglement from Soft Interventions.”

# Weak supervision via invariances



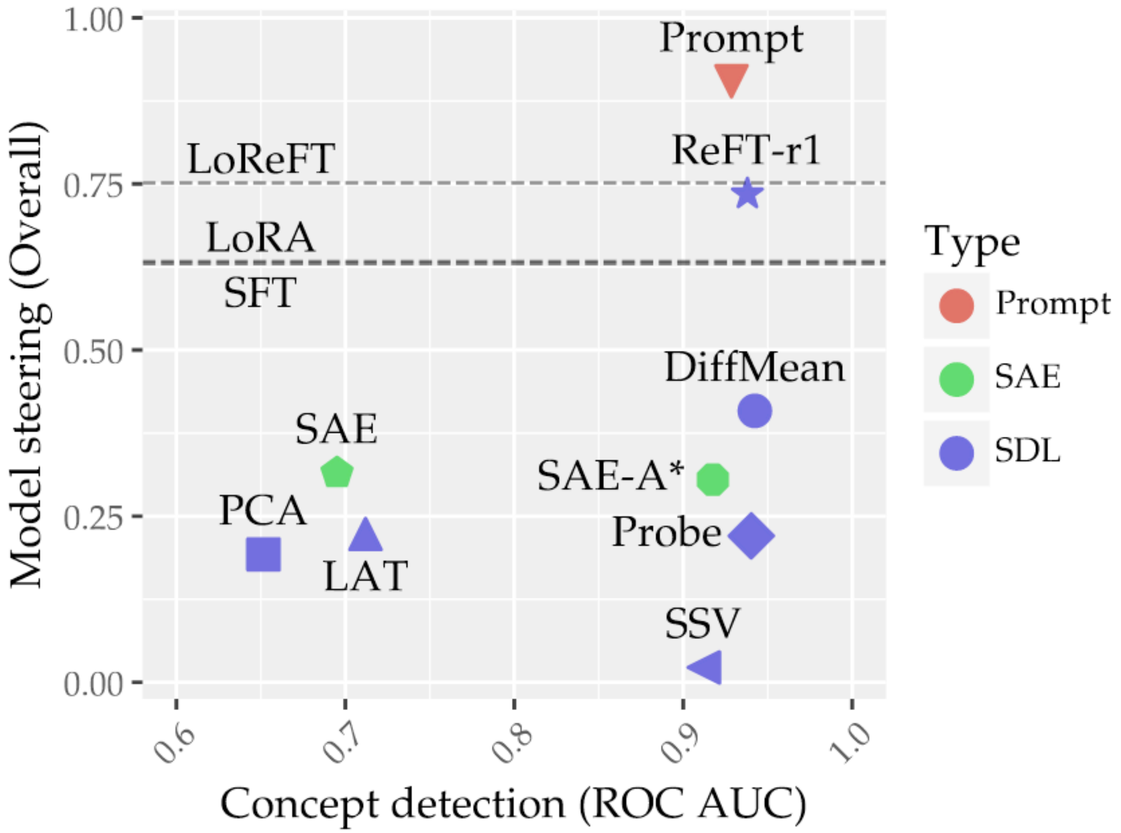

When the shifts are 1-sparse, and with some key extra assumptions, we can show that $\hat{z} = \hat{f}(x) = PDz$.

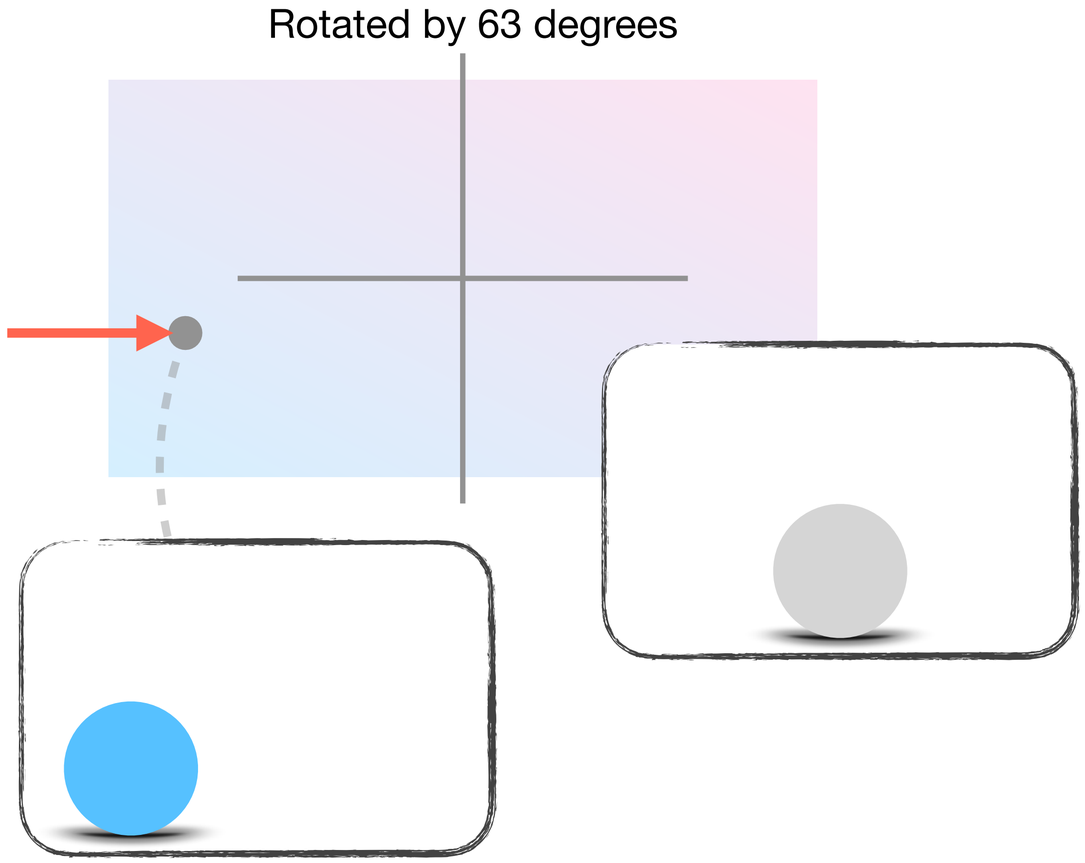


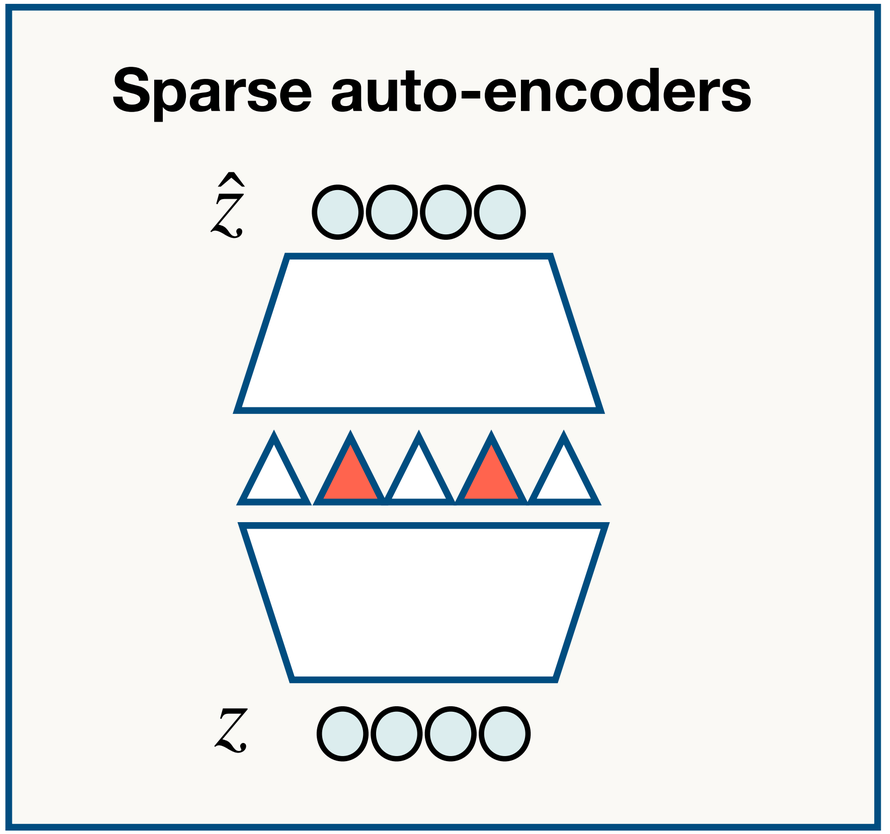

Interpretability pipeline
$X_{10}$
Carry-over

# AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders

Zhengxuan Wu [*1] Aryaman Arora [*1] Atticus Geiger [2] Zheng Wang [1] Jing Huang [1]
Dan Jurafsky [1] Christopher D. Manning [1] Christopher Potts [1]



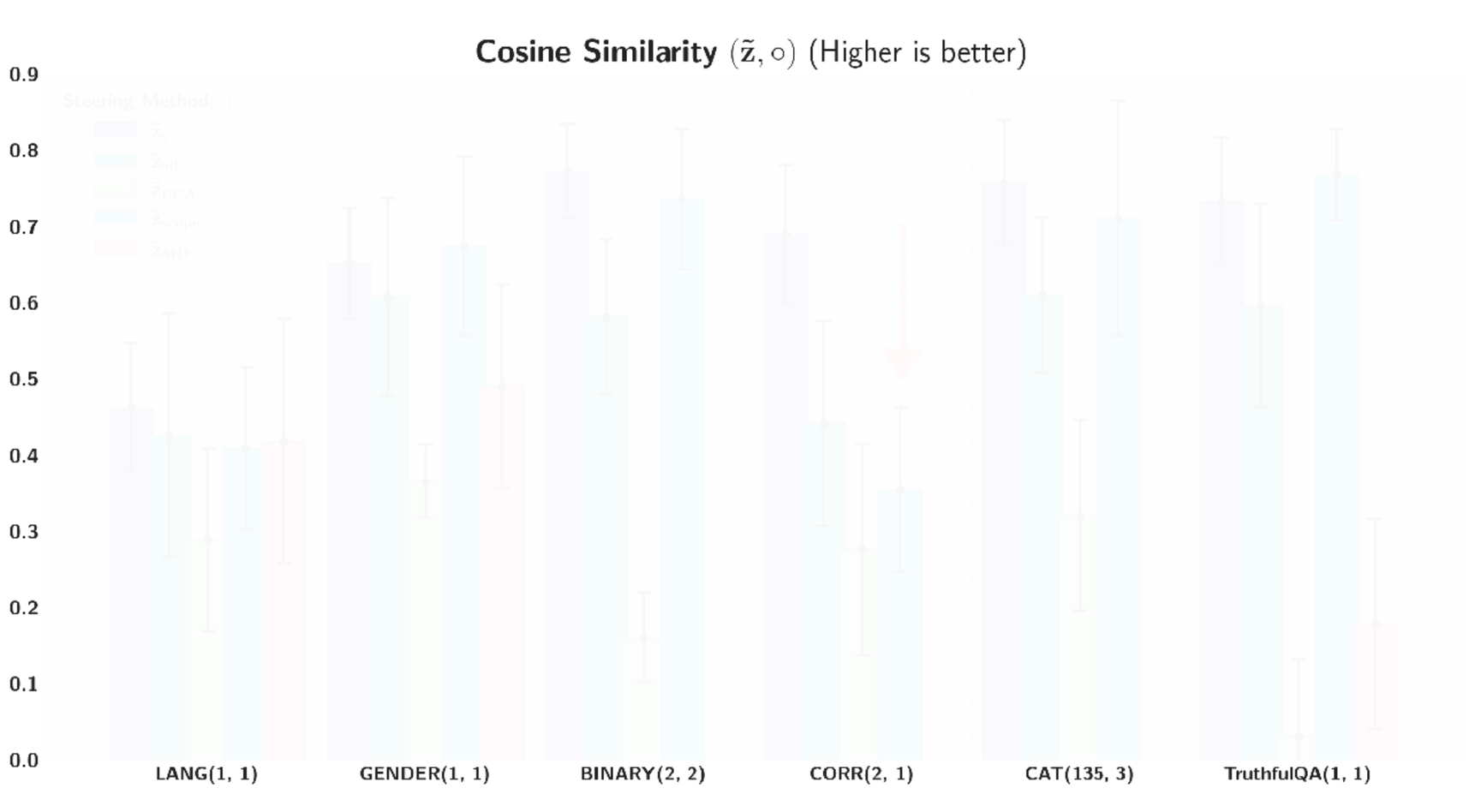



*Do we end up with entangled features like our rotated basis?*

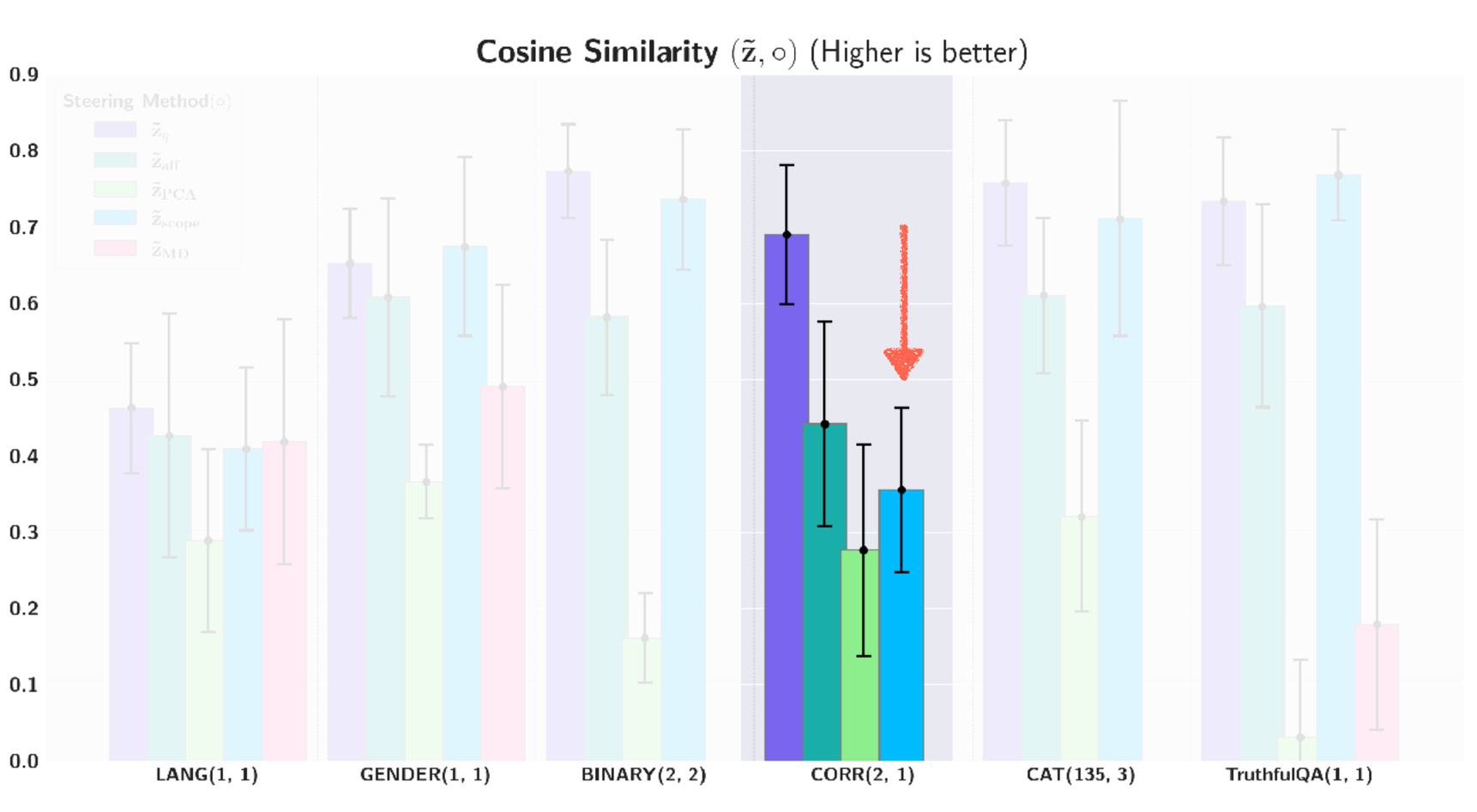
# Evidence from experiments

Single-concept shifts $(z, \tilde{z})$. Use each column of LlamaScope decoder to steer $z$ and take best similarity to true $\tilde{z}$.



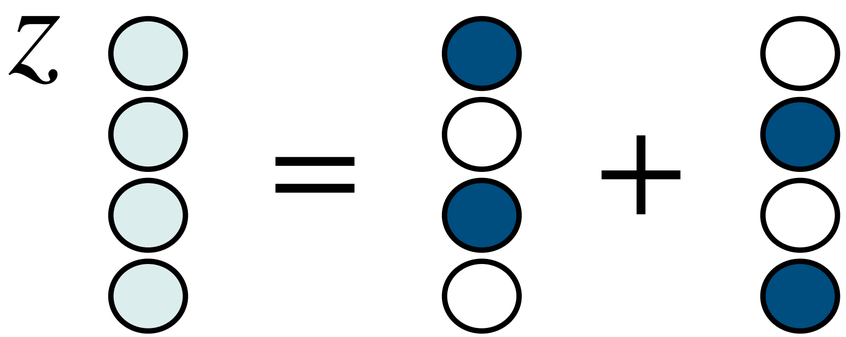
# Evidence from experiments

Single-concept shifts $(z, \tilde{z})$. Use each column of LlamaScope decoder to steer $z$ and take best similarity to true $\tilde{z}$.



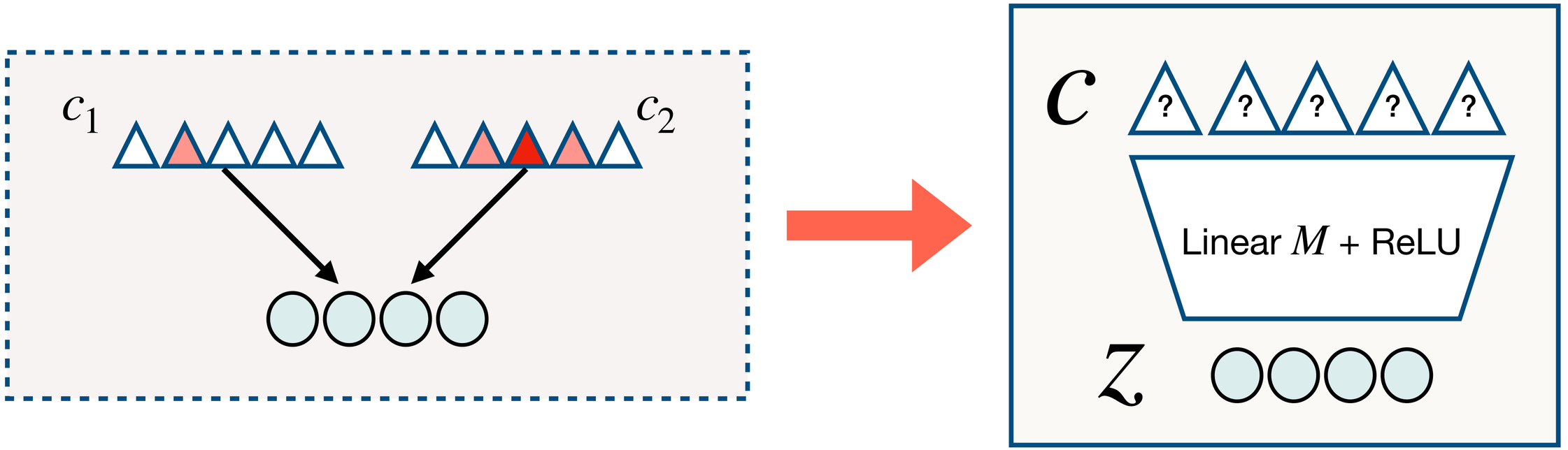
# Sparse auto-encoders (SAEs)

**Represent $z$ as a sparse sum of concept directions.**

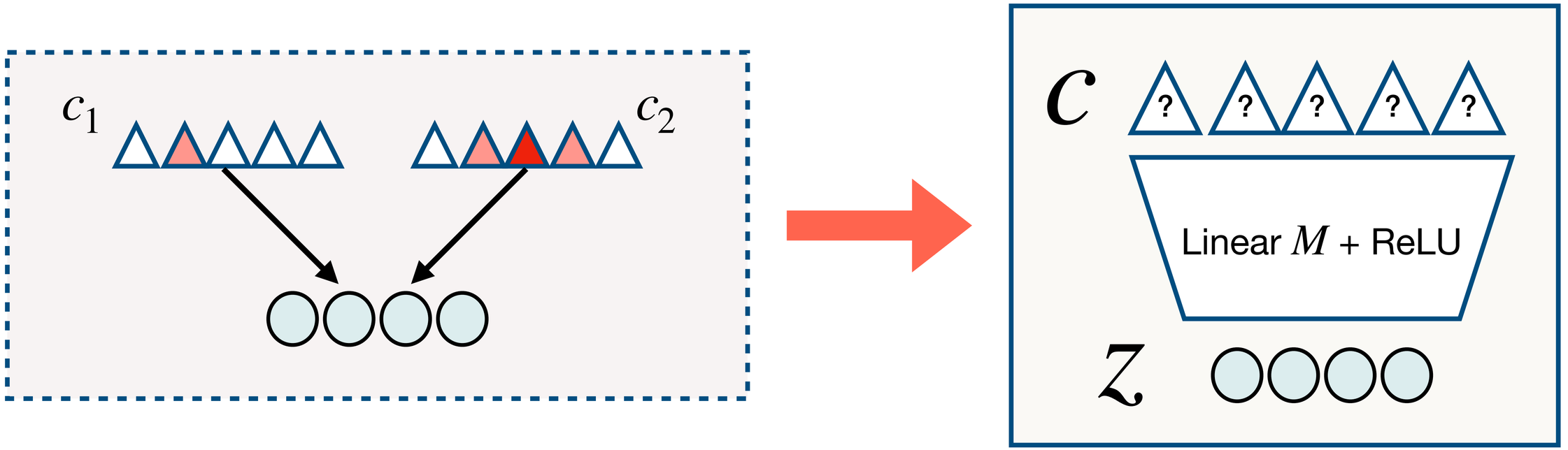



Sparsity regularization should lead to disentangled features — what might be going wrong?

# Issue 1: Some concepts may not learnable

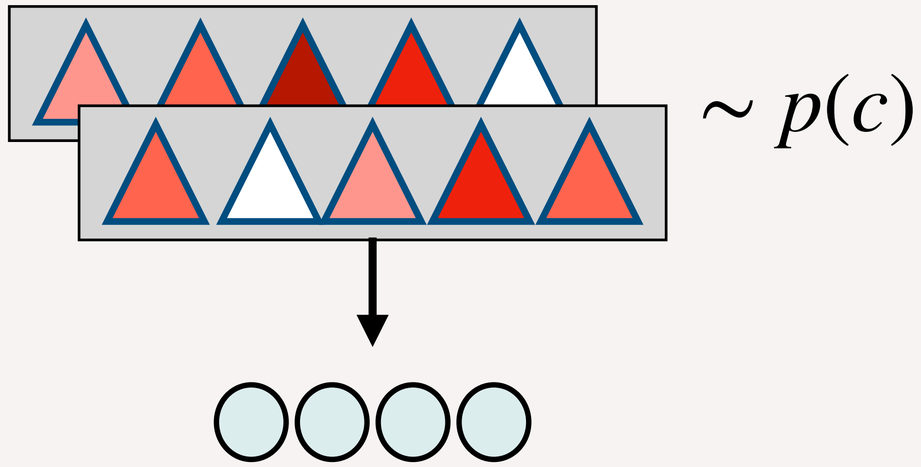


# Issue 1: Some concepts may not learnable




Mapping from concepts to LLM representations is not ***injective.***

# Issue 2: Not enough diversity in underlying concepts



$$\sim p(c)$$

$$\min_{M} \|z - \hat{z}\|_2^2 + \alpha\|c\|_1$$

E.g., if we lack examples $z$ that capture only ***a single concept,*** spurious encoders that entangle concepts also minimize regularizer!

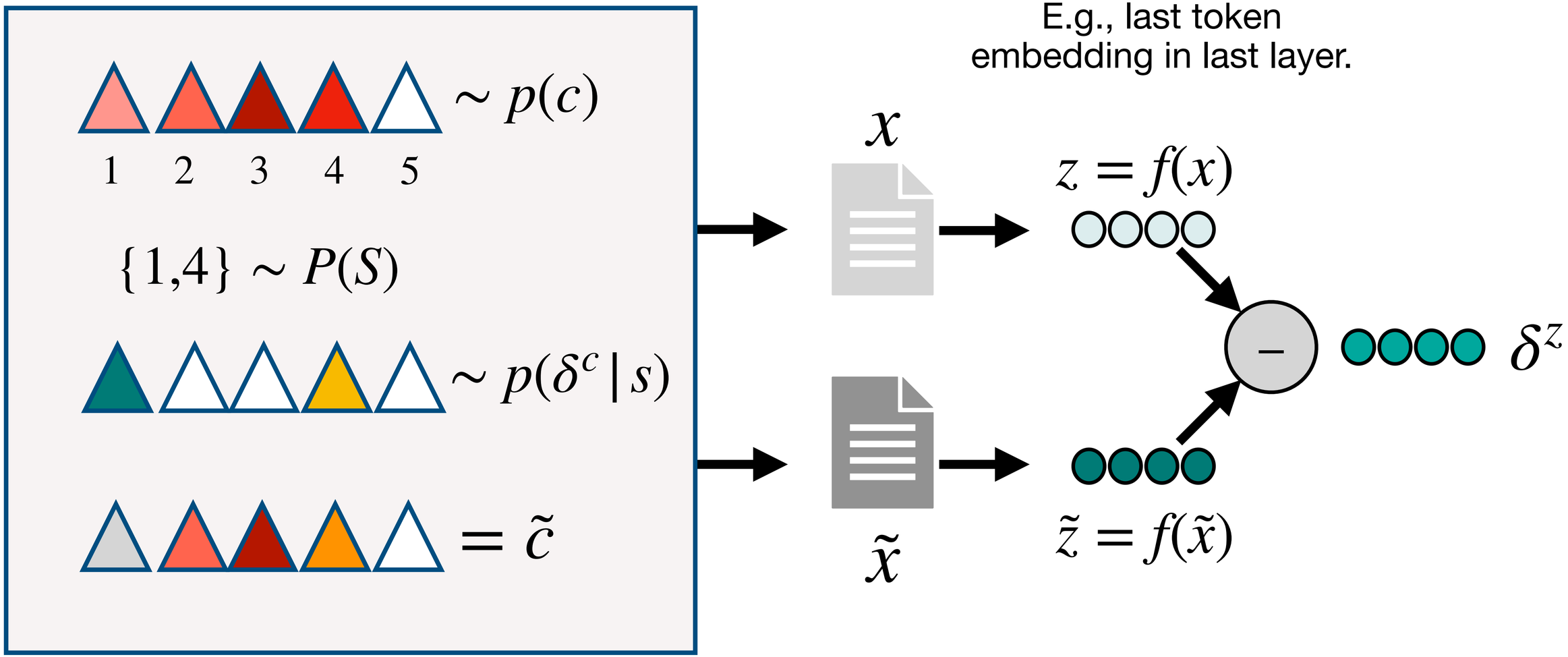

- Moran, Sridhar, Wang and Blei. 2022. *Identifiable deep generative models via sparse decoding.*
- Arora et al. 2013. *A practical algorithm for topic modeling with provable guarantees.*

# Identifiable Steering via Sparse Autoencoding of Multi-Concept Shifts

**Shruti Joshi** [1 2] **Andrea Dittadi** [3 4 5] **Sébastien Lachapelle** [6] **Dhanya Sridhar** [1 2]

arXiv: 2502.12179

# Multi-concept shift datasets

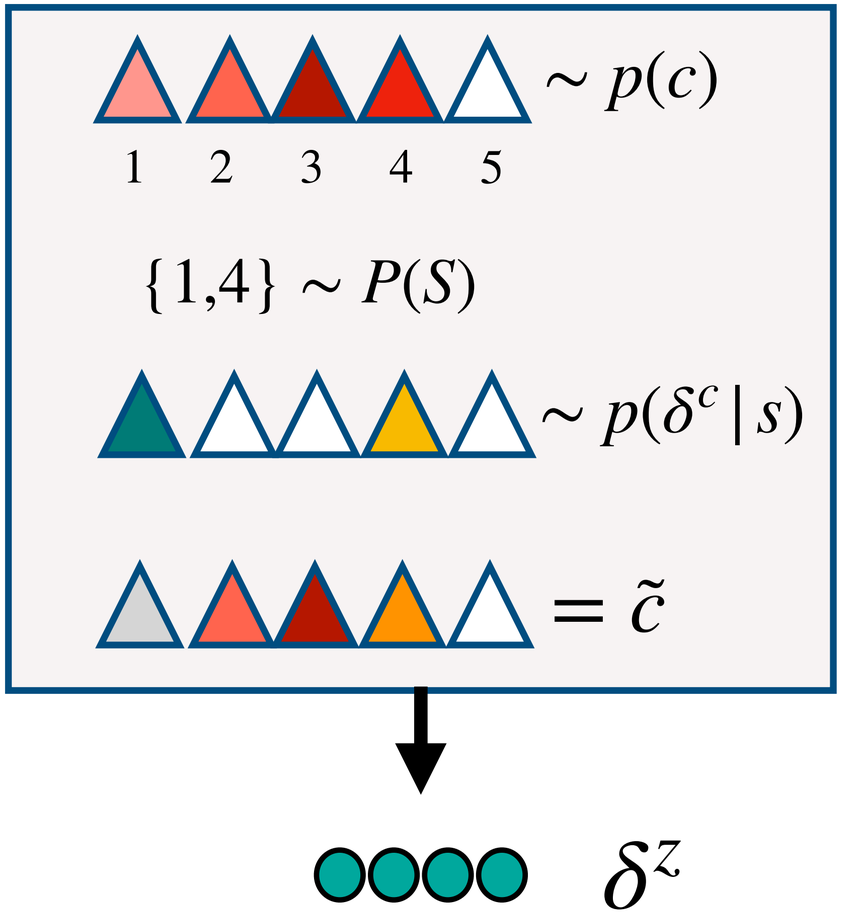


# Multi-concept shift datasets

⚠️ These are more general than single-concept shift contrastive pairs used to learn steering vectors via supervision.

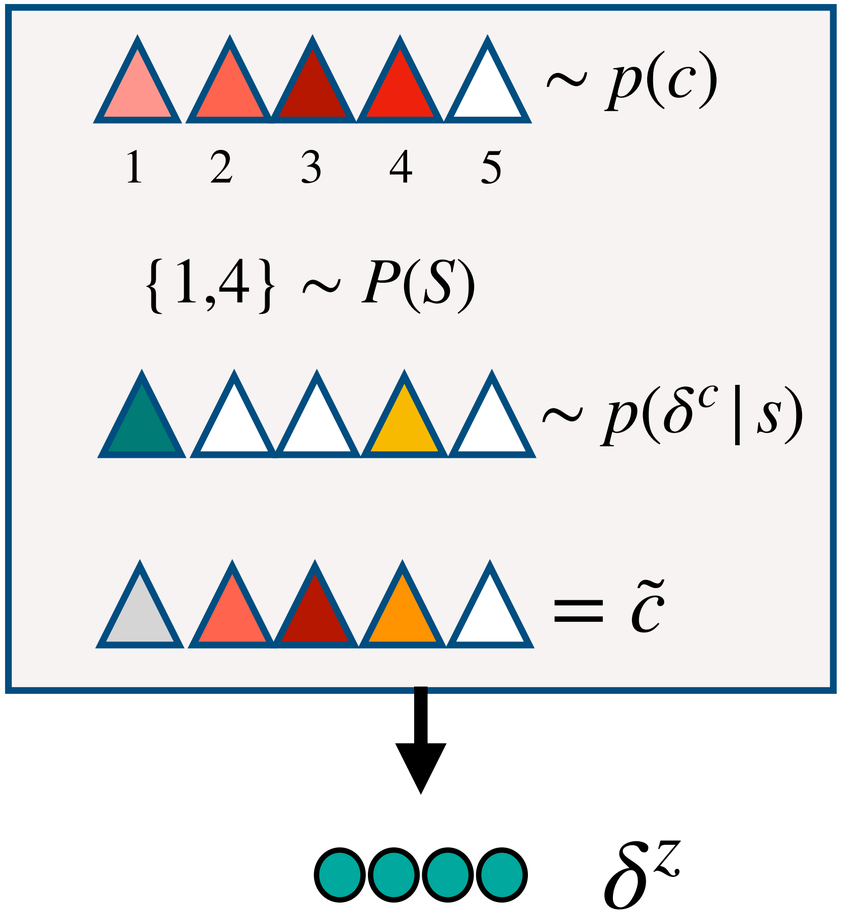
🤖 We could, e.g., use LLMs to generate concept-shifted examples, and learn despite the fact that LLMs don't generate perfect counterfactuals.

# Multi-concept shift datasets




**Tackling lack of diversity:** By formalizing conditions on $P(S, \delta^c)$, we get the diverse shifts we need for sparsity regularization to guarantee concept recovery.

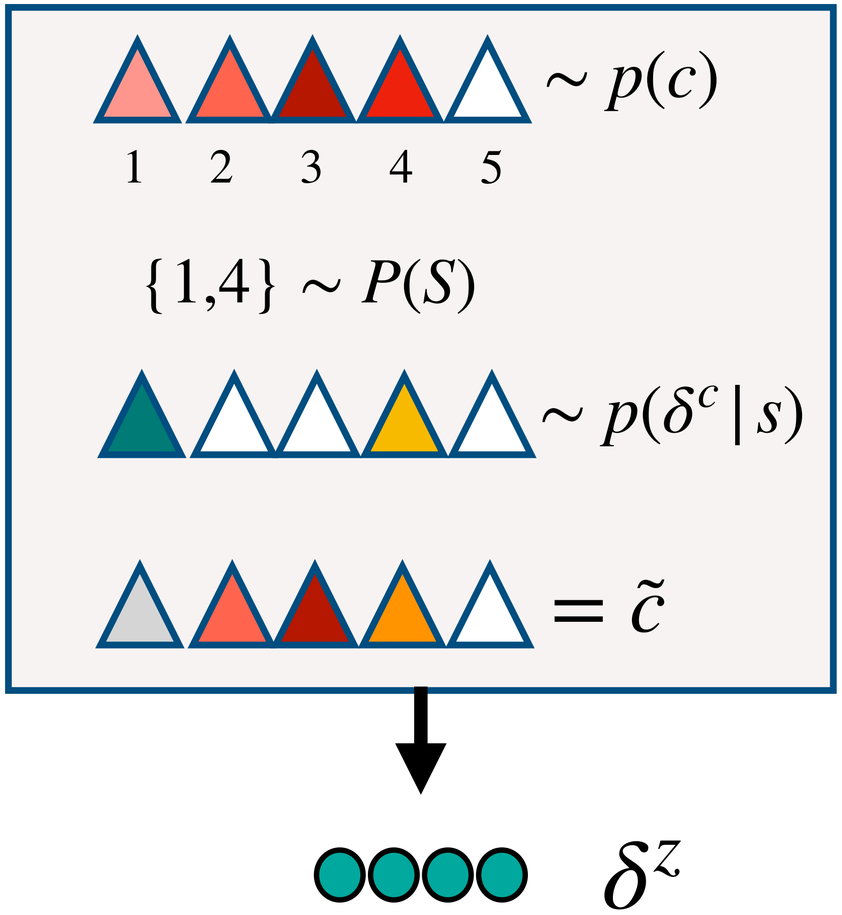
# Multi-concept shift datasets




**Tackling lack of diversity:** By formalizing conditions on $P(S, \delta^c)$, we get the diverse shifts we need for sparsity regularization to guarantee concept recovery.

**What about unmeasurable concepts?** They get cancelled out when we consider embedding differences $\delta^z$!

# Multi-concept shift datasets

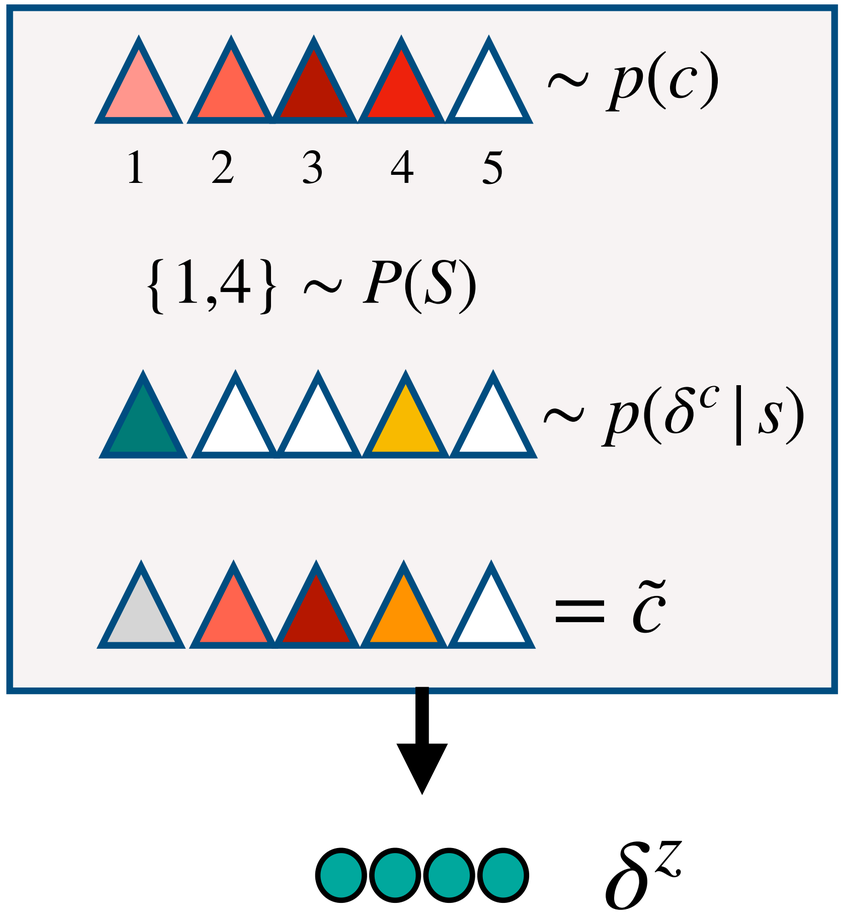


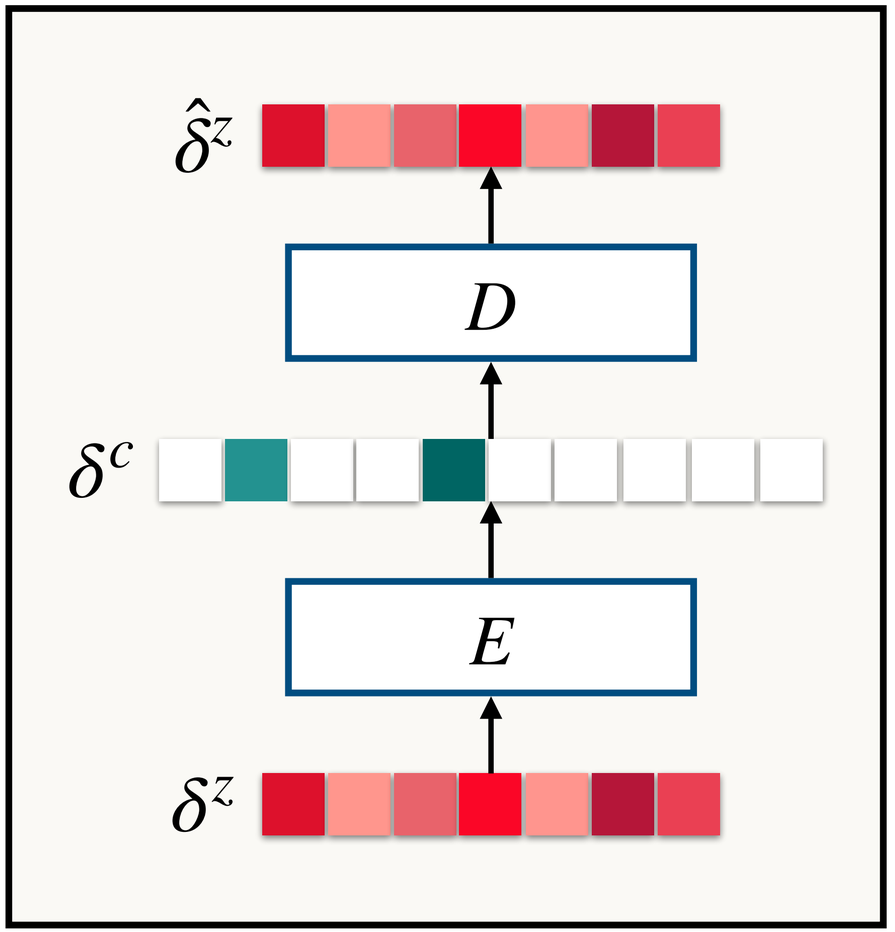

**Tackling lack of diversity:** By formalizing conditions on $P(S, \delta^c)$, we get the diverse shifts we need for sparsity regularization to guarantee concept recovery.

**What about unmeasurable concepts?** They get cancelled out when we consider embedding differences $\delta^z$!

**Key detail here:** We restrict ourselves to recovering only the concepts that vary across the dataset!

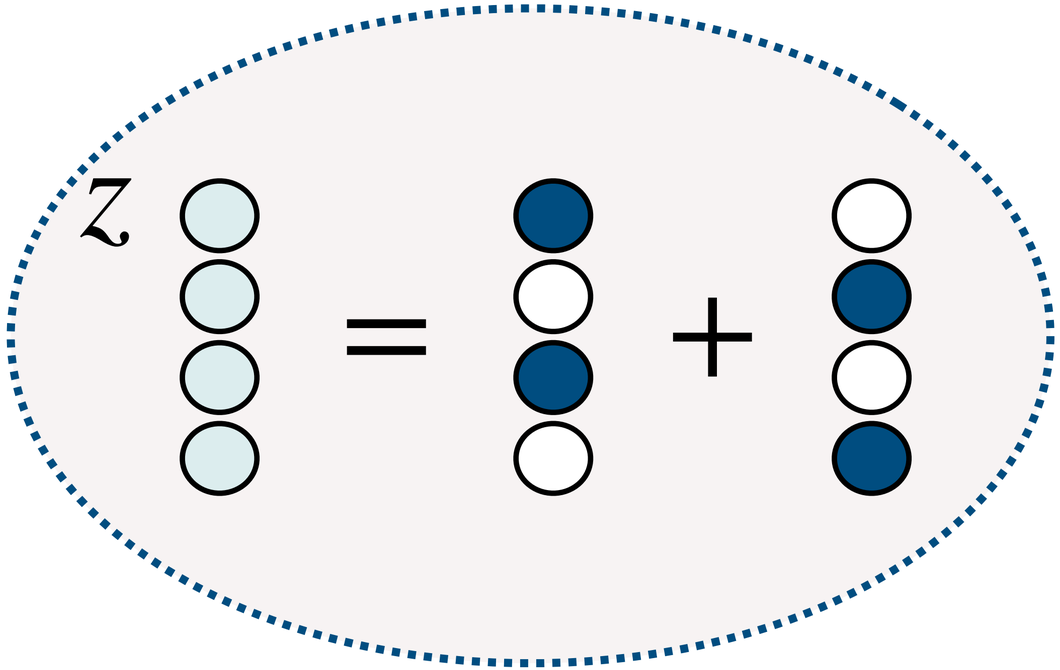
# Sparse *shift* auto-encoders (SSAEs)



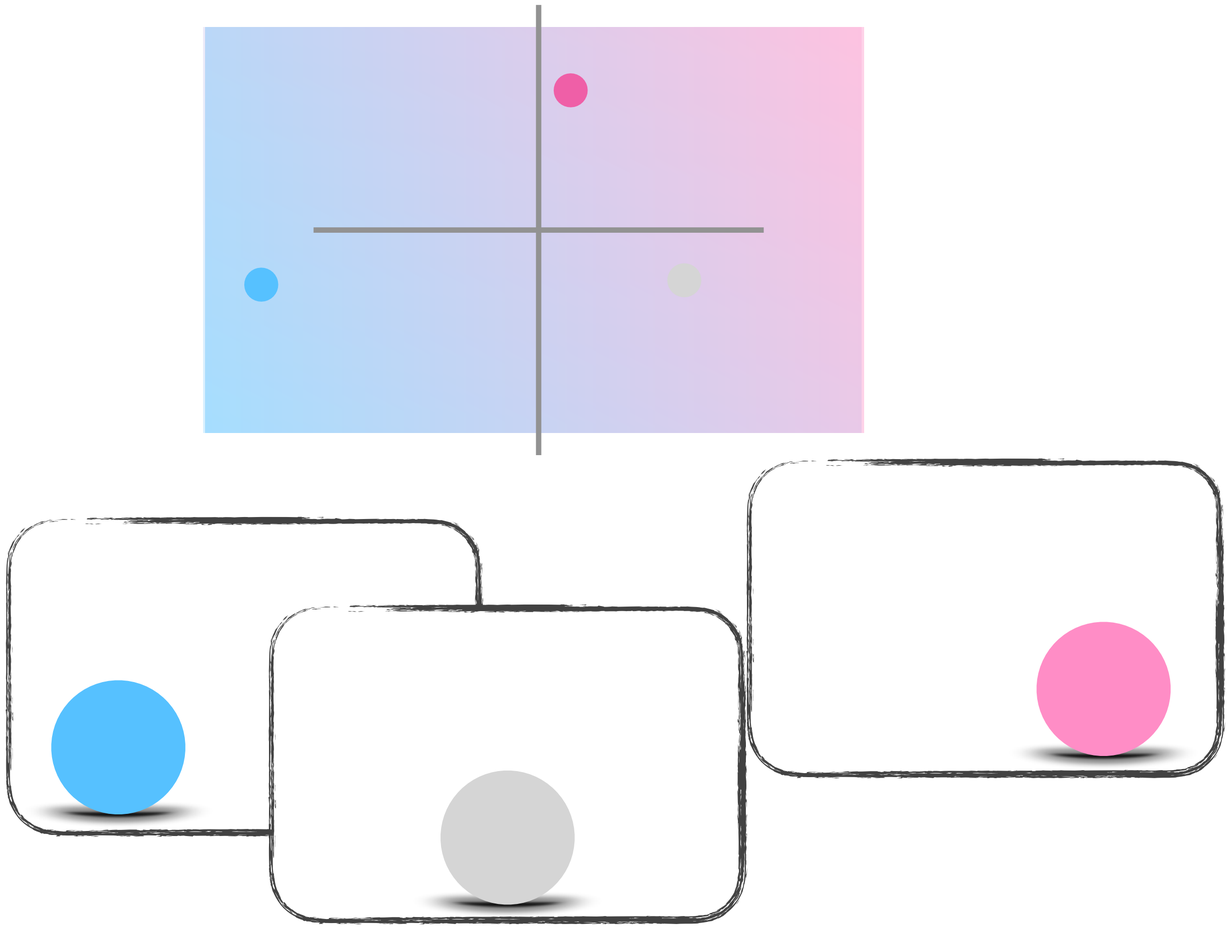



# Sparse *shift* auto-encoders (SSAEs)

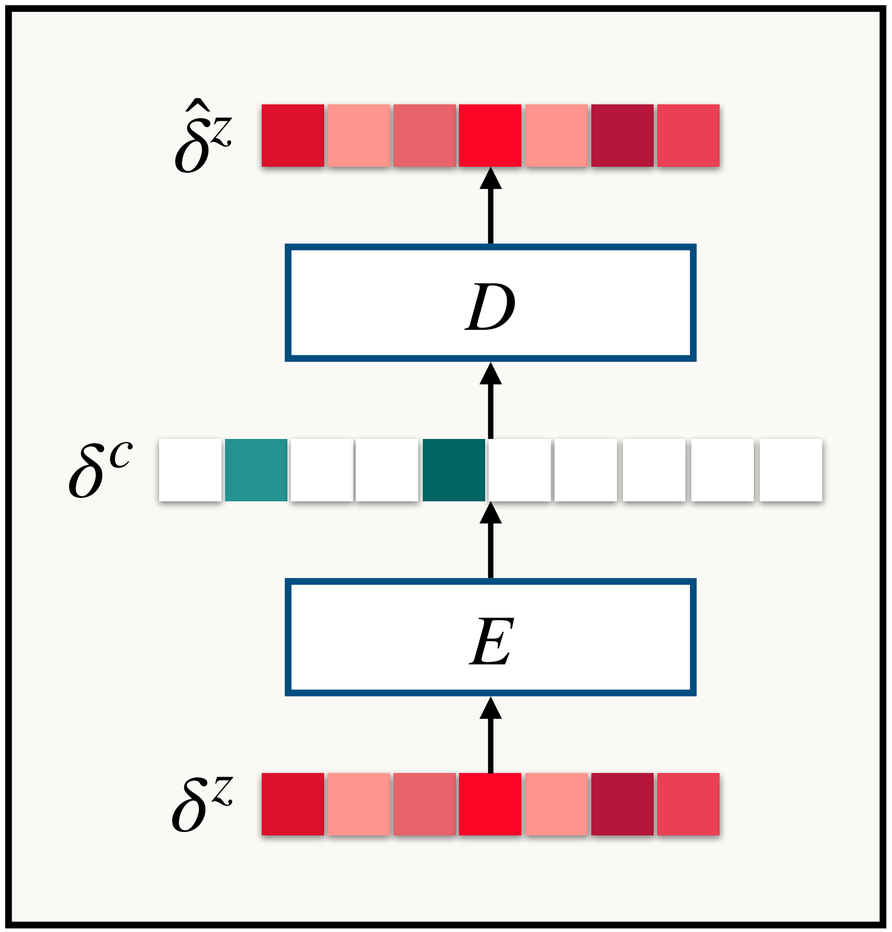
Not basis where concept activations are sparse..




… but where interventions are sparse



# SSAEs provably recover concept shifts



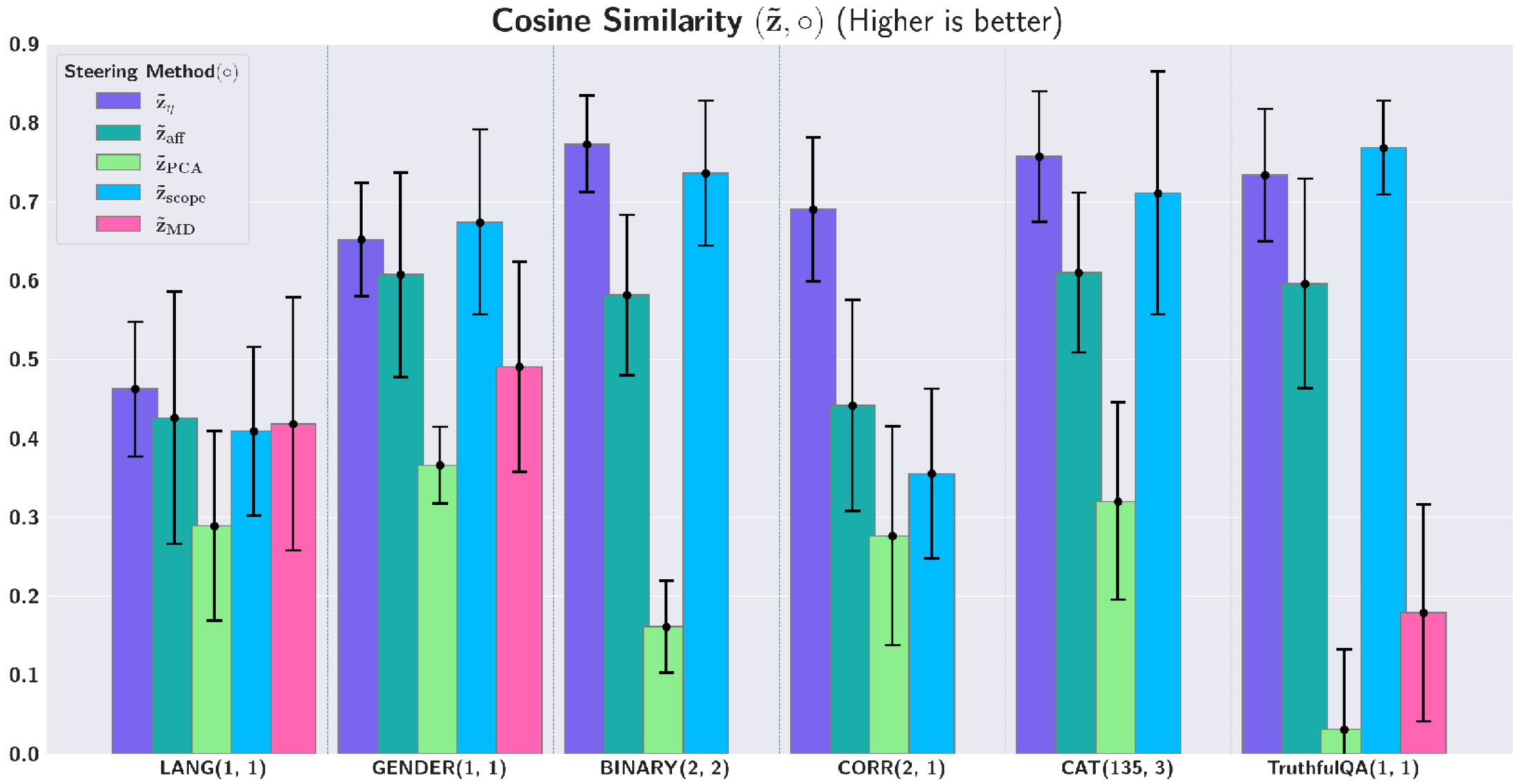

**Objective:**

$$\min_{E,D} \mathbb{E}_{z,\tilde{z}} \left[ \|\delta^z - \hat{\delta}^z\|_2^2 \right] \text{ s.t. } \mathbb{E}\left[ \|\delta^c\|_0 \right] \leq \beta$$

**Informal theorem statement:** We estimate concept shifts $\delta^c$ and linear mixing function $D$ up to permutations and scaling of the target solutions.

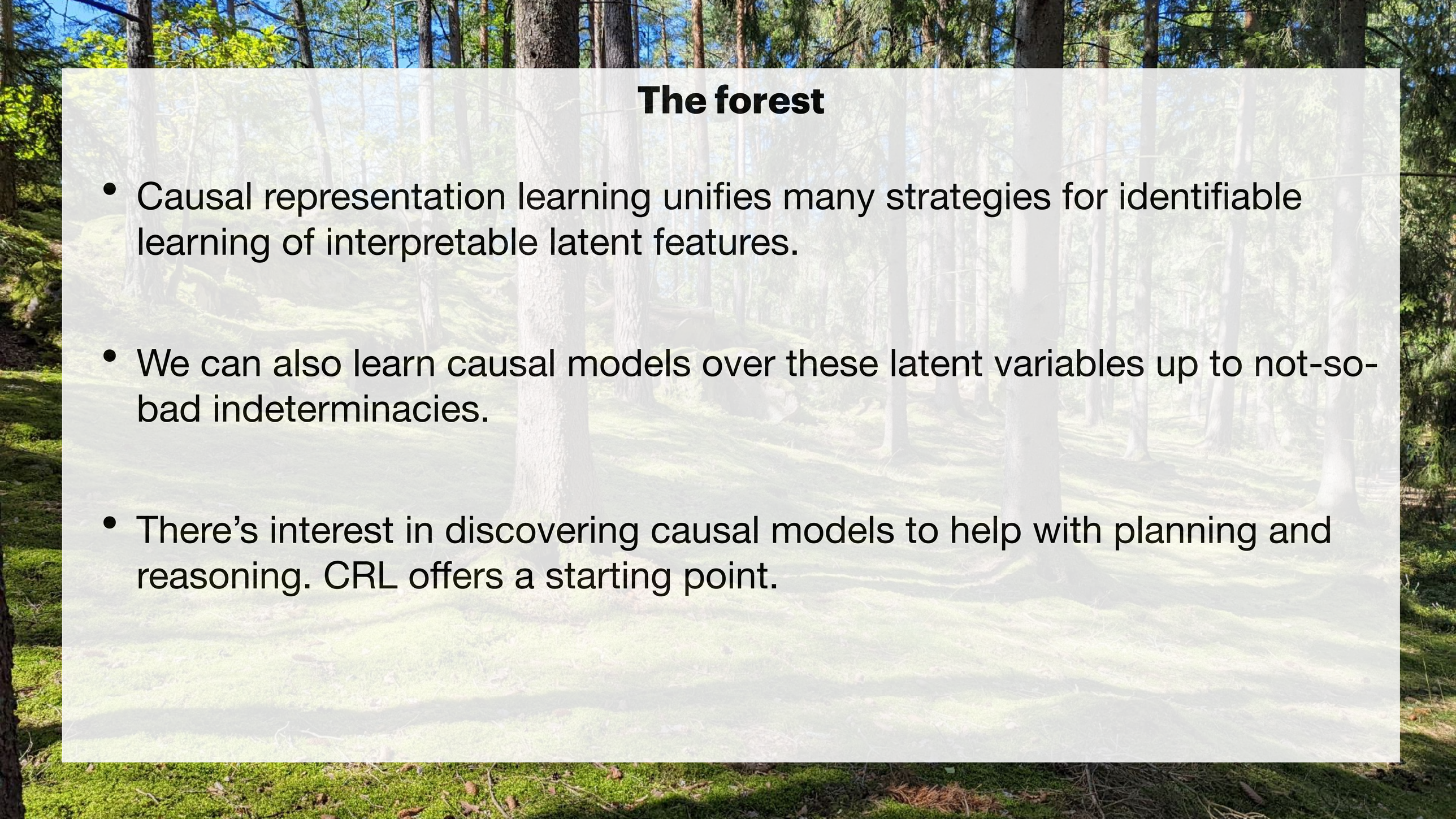
Columns of the decoding matrix are parallel to steering directions.

# How does this work with Llama 3.1 embeddings?



# How does this work with Llama 3.1 embeddings?


Competitive with SAEs or significantly better where SAEs had not worked well.



## The forest

- Causal representation learning unifies many strategies for identifiable learning of interpretable latent features.
- We can also learn causal models over these latent variables up to not-so-bad indeterminacies.
- There’s interest in discovering causal models to help with planning and reasoning. CRL offers a starting point.