The Missing Piece (of Data) in Interpretability Research

Yanai Elazar, June 24nd, 2026

About Myself

- Yanai Elazar
- Assistant Professor at Bar-Ilan University, Computer Science and AI Department
- Leading the DICE lab

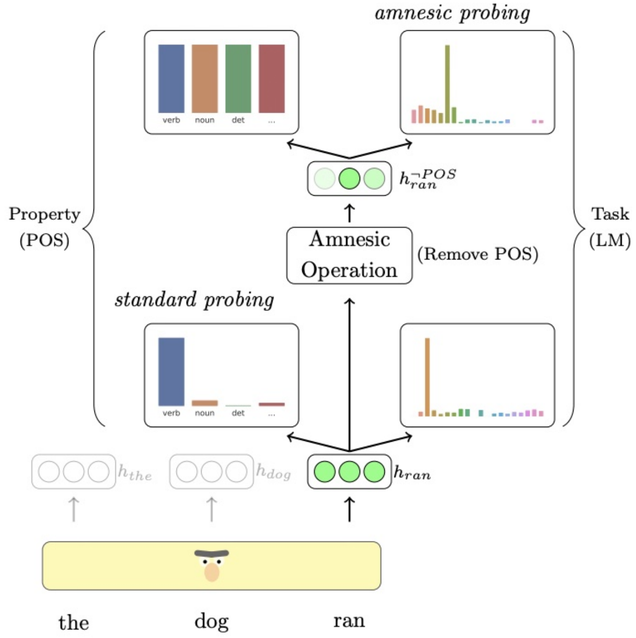
About Myself

- Interested in how neural networks work (“interpretability”)
- This talk won’t be about mechanistic interpretability
  - Although I will address it
  - I did work on it (before it was cool)

Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar[1,2] Shauli Ravfogel[1,2] Alon Jacovi[1] Yoav Goldberg[1,2]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence

TACL 2021

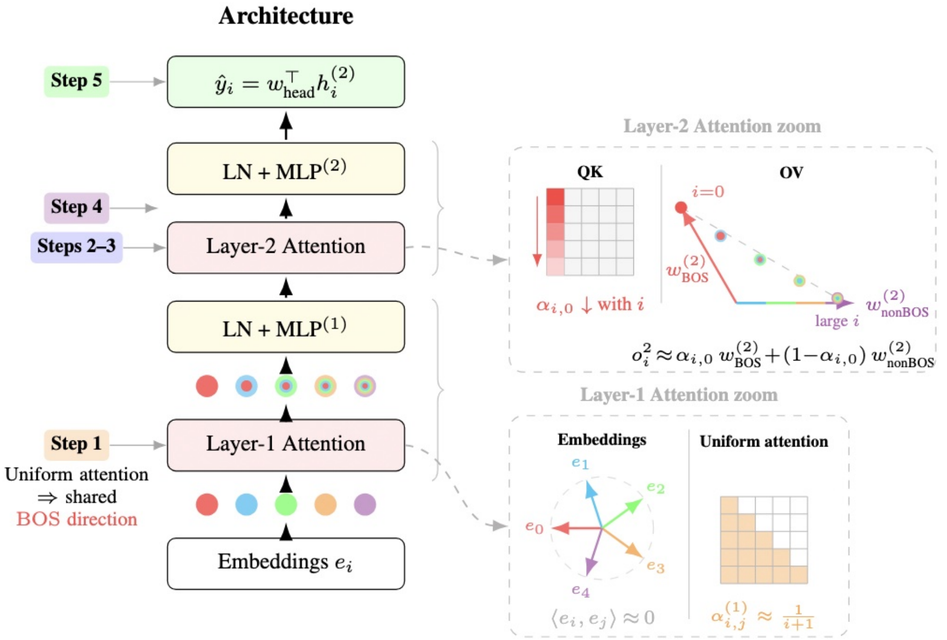


About Myself

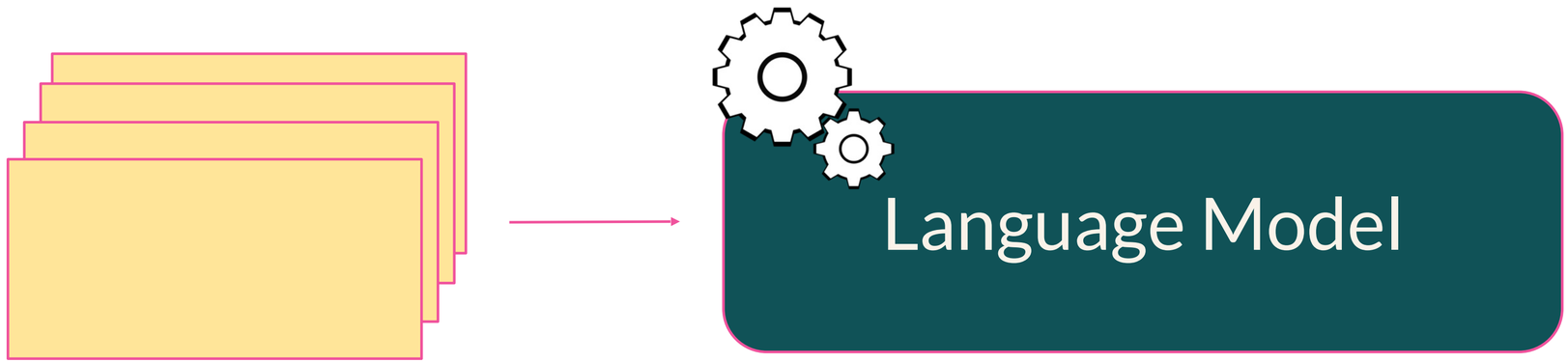
- Interested in how neural networks work (“interpretability”)
- This talk won’t be about mechanistic interpretability
  - Although I will address it
  - I did work on it (before it was cool)
  - I still do (sometimes)

Position Without Positional Embeddings:
A Directional Mechanism in NoPE Transformers

NeurIPS 2026 (under submission)



Language Models Learn All Sorts of Useful Features



Pretraining documents

Learning to predict the next word

Language Models Learn All Sorts of Useful Features

Features learned from pretraining:

Facts:

France → Paris
China → Beijing
Rhode Island → Providence

Language Model

Positive

happy,
joy,
awesome

Negative

bad,
stupid,
upset

Language Models learn ‘linear’ features of these concepts

Meaning the the presence/intensity of the concept can be modulated through a linear or affine transformation

Facts:

France → Paris
China → Beijing
Rhode Island → Providence

Language Model

Positive	Negative
happy, joy, awesome	bad, stupid, upset

Language Models learn ‘linear’ features of these concepts

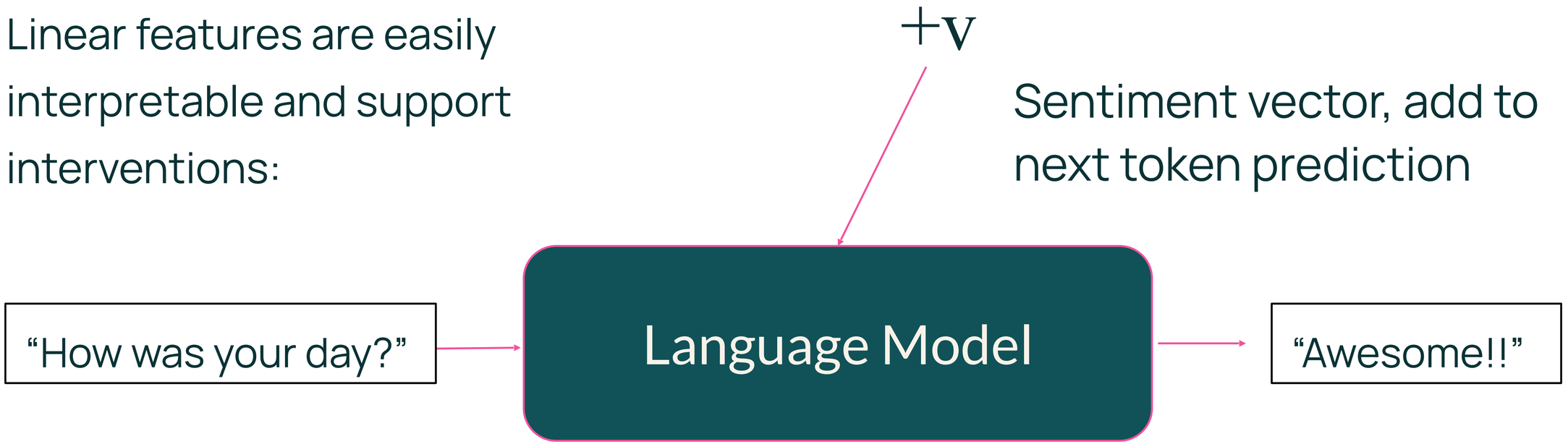
Linear features are easily interpretable and support interventions:

Positive	Negative
happy, joy, awesome	bad, stupid, upset

Tigges et al., 2023

Language Models learn ‘linear’ features of these concepts

Linear features are easily interpretable and support interventions:



Positive	Negative
happy, joy, awesome	bad, stupid, upset

Tigges et al., 2023

Language Models learn ‘linear’ features of these concepts

Linear features are easily interpretable and support interventions:

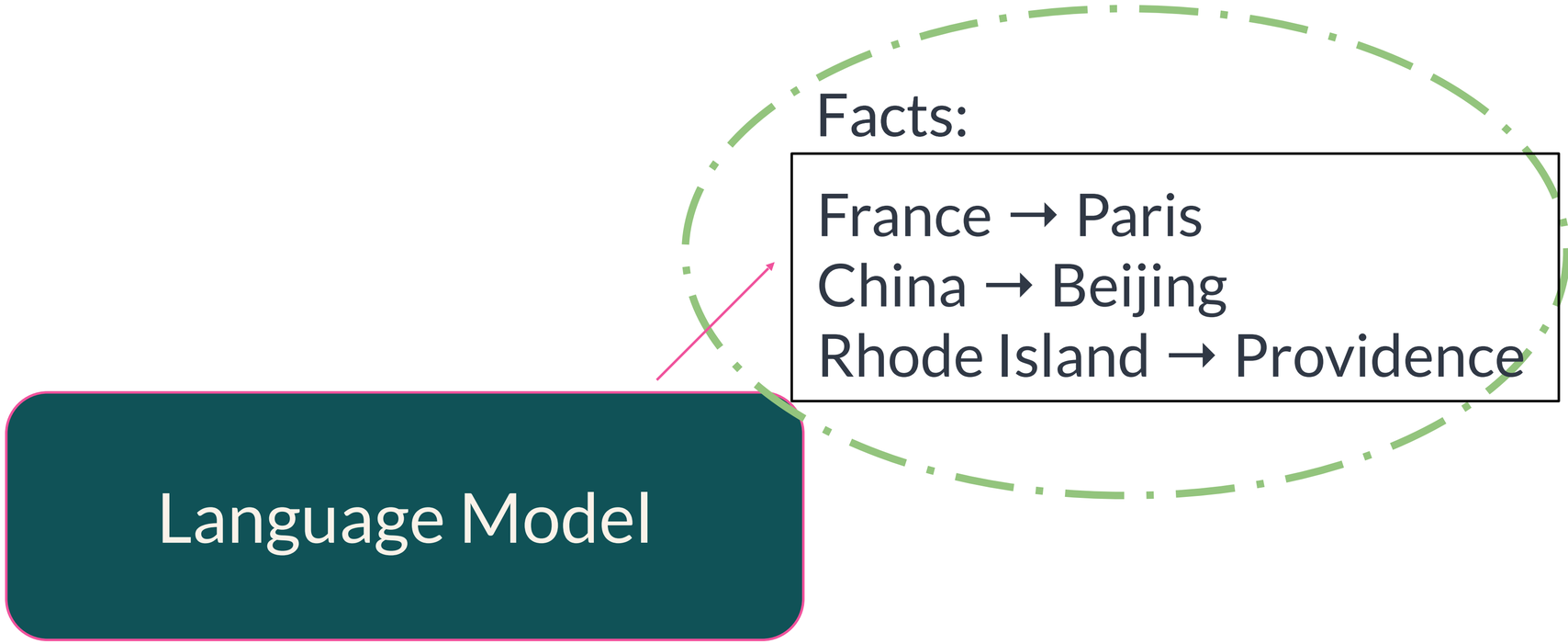
$-\mathbf{V}$

Sentiment vector, subtract from next token prediction

Positive	Negative
happy, joy, awesome	bad, stupid, upset

Tigges et al., 2023

Language Models learn ‘linear’ features of these concepts



Subject-Relation-Object triplets

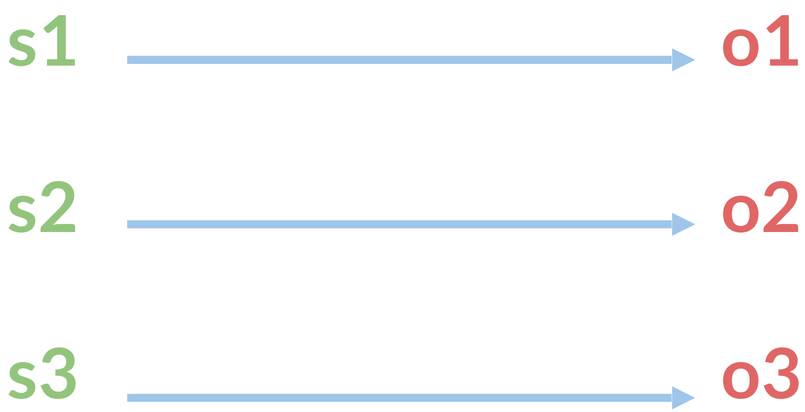
“Oracle’s headquarters is in Texas”

$$\text{headquarters\_in}(\text{Oracle}) = \text{Texas}$$

$$\text{relation}(\text{subject}) = \text{object}$$

Subject-Relation-Object triplets

$$\text{relation}(\text{subject}) = \text{object}$$



This relation is robustly linear when the same transform maps from every subject to every object

Linear Relational Embeddings (LREs)

- Paccanaro and Hinton, 2001

For a binary relation, represent the **subject** as a vector **s**, the **object** as vector **o**, and the relation as an nxn matrix **W**, such that:

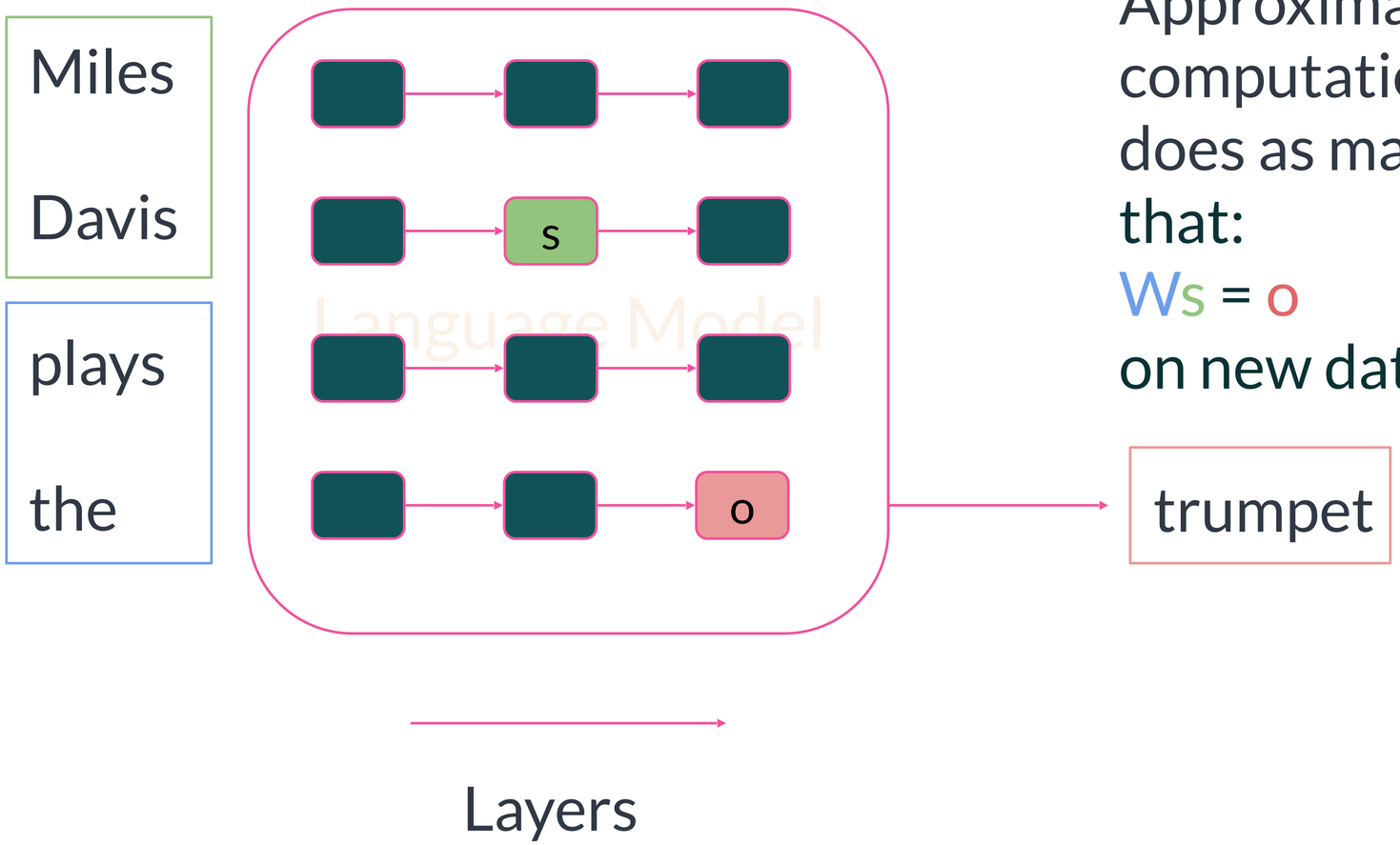
$$Ws = o$$

for all **s** and **o** in the relation.

E.g., capital_of(France) = Paris where **subject=France, object=Paris**

Linear Features of Relations
(LREs: Hernandez et al., 2024)

Language models implement LREs:

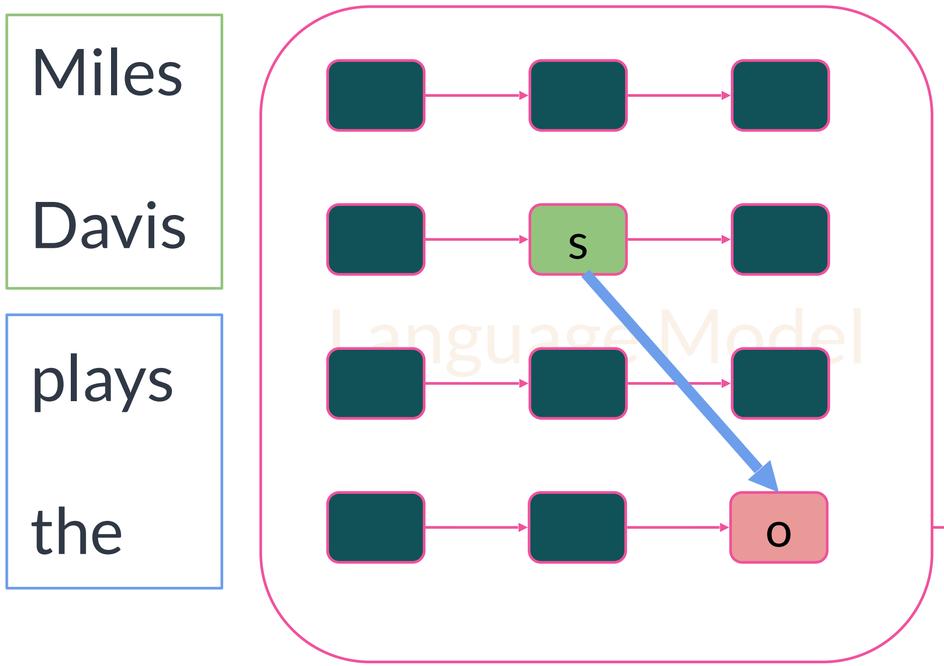


Approximate the computation “plays the” does as matrix W such that:
Ws = o
on new data

trumpet

Linear Features of Relations
(LREs: Hernandez et al., 2024)

Language models implement LREs:

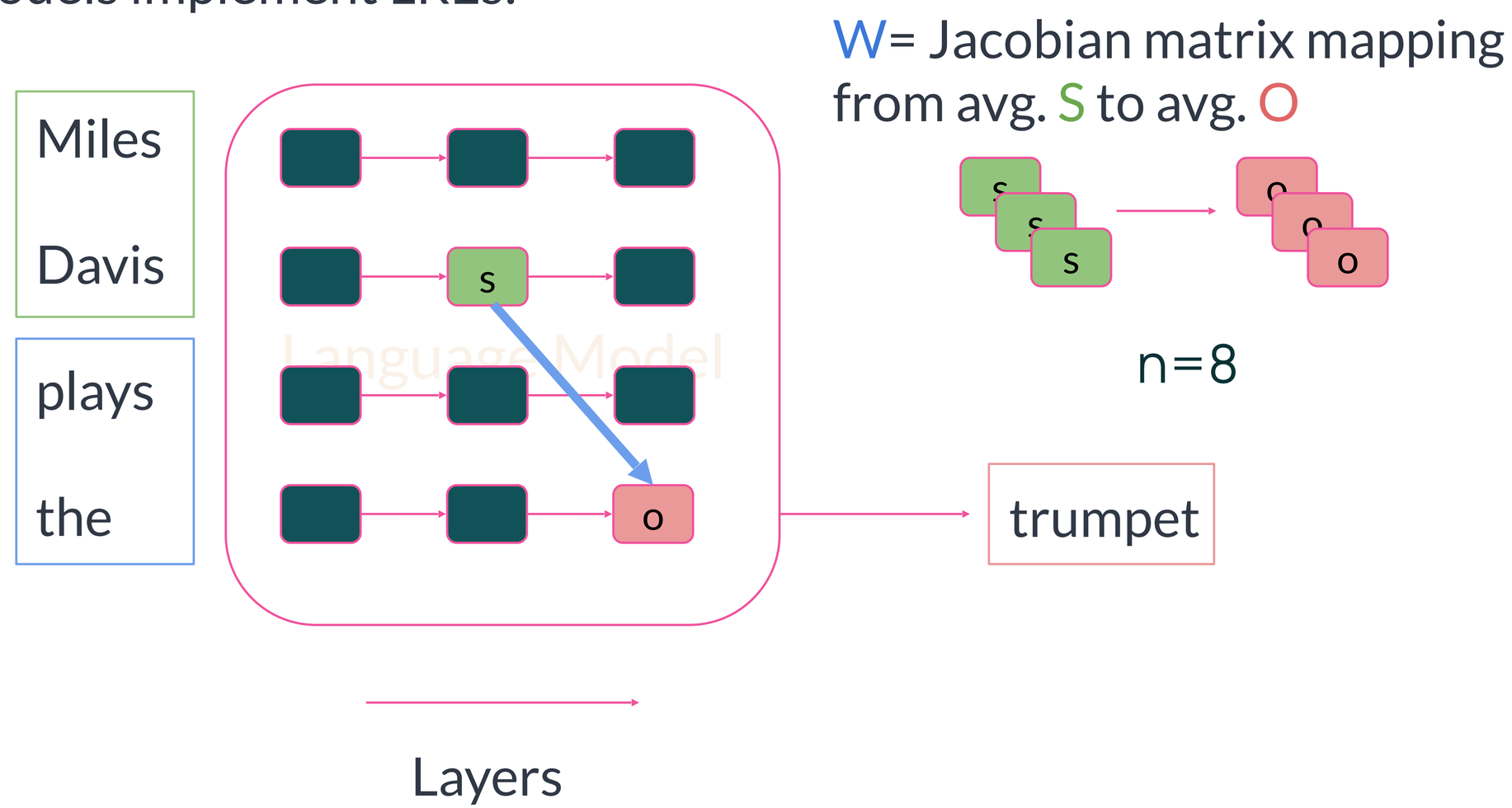


Approximate the computation “plays the” does as matrix W such that:
Ws = o
on new data

trumpet

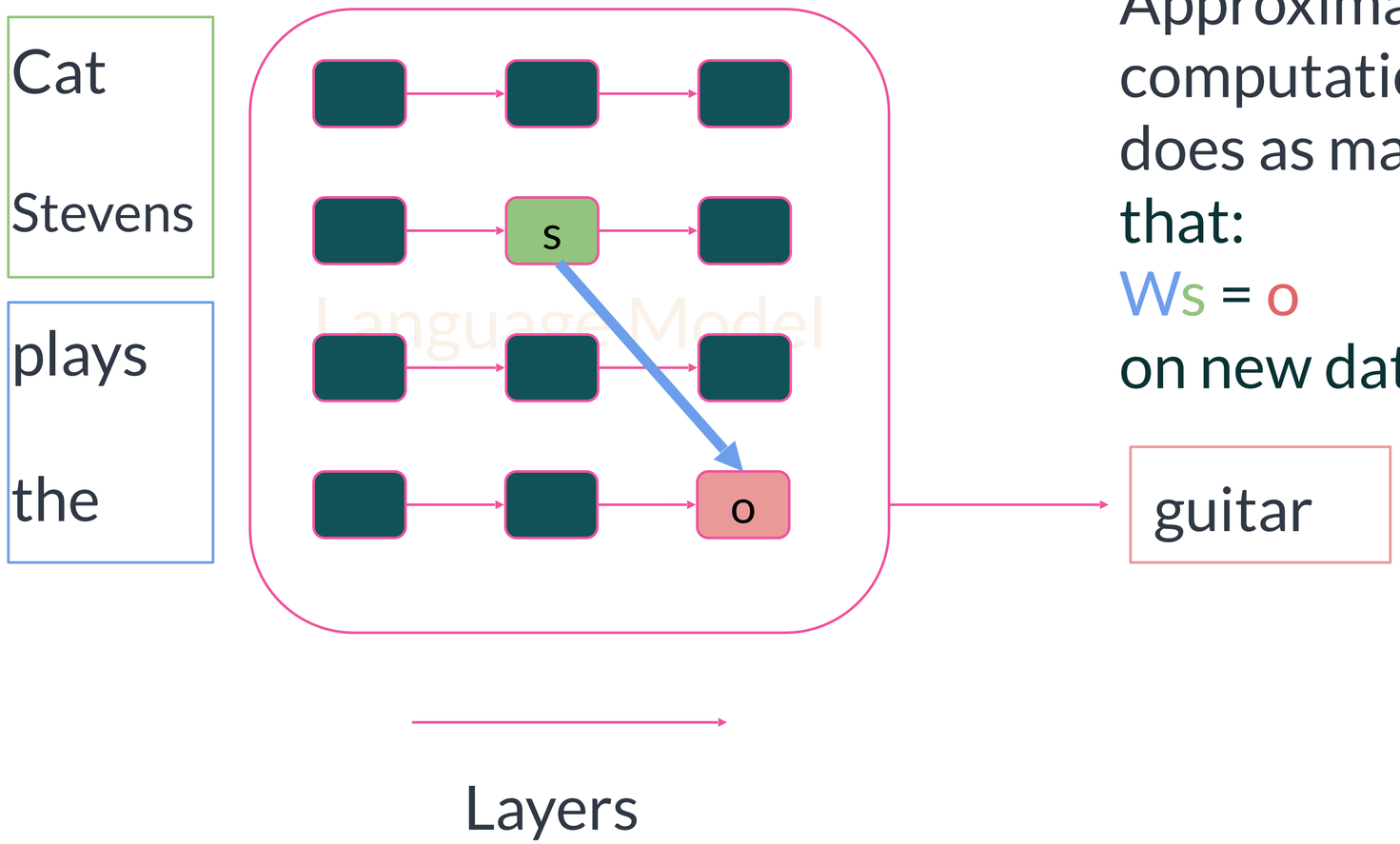
Linear Features of Relations
(LREs: Hernandez et al., 2024)

Language models implement LREs:



Linear Features of Relations
(LREs: Hernandez et al., 2024)

Language models implement LREs:



Approximate the computation “plays the” does as matrix W such that:
Ws = o
on new data

Linear Features of Relations
(LREs: Hernandez et al., 2024)

Metrics: How do we measure “linearity” of a relation?

Linear Features of Relations
(LREs: Hernandez et al., 2024)

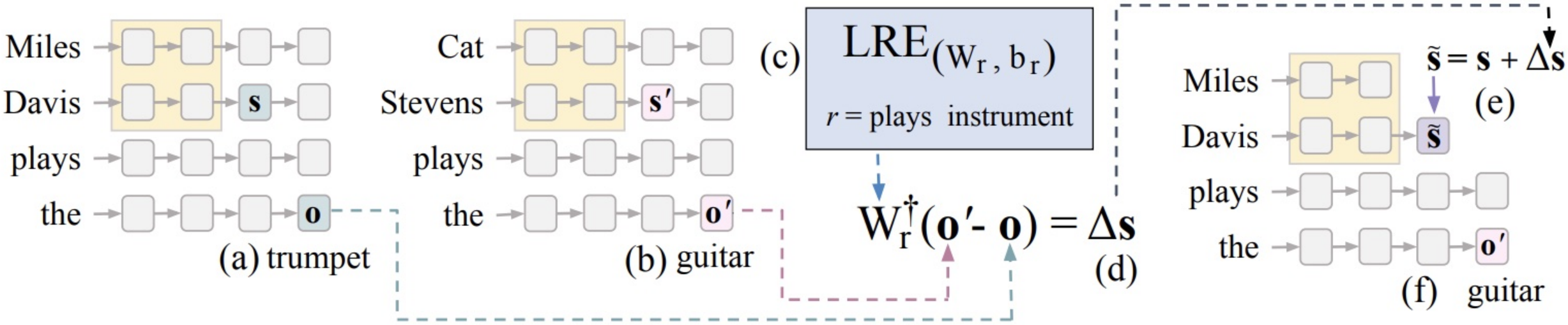
Metrics: How do we measure "linearity" of a relation?

Causality: If the representation space is structured, we should be able to edit LM predictions with W

Goal: get LM to predict "Miles Davis plays the Guitar" by editing the Miles Davis hidden state

Linear Features of Relations
(LREs: Hernandez et al., 2024)

Metrics: How do we measure "linearity" of a relation?



Linear Features of Relations
(LREs: Hernandez et al., 2024)

Metrics: How do we measure “linearity” of a relation?

Causality is the **proportion** of time the **edited object output** is predicted as more likely than the **original output**

Miles Davis plays the____ p(guitar) > p(trumpet)

OLMo implements LREs:

Many relations well approximated by this transformation.

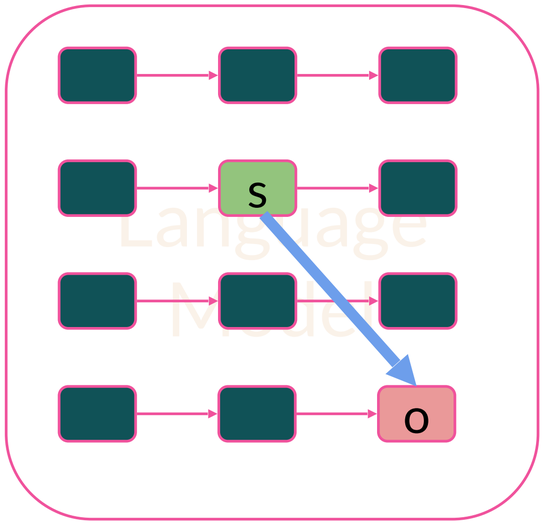
Not true for every relation. Why?

What now?

What now?

Mech
Interp

What causes linear features to form?



Approximate the relation matrix W such that:
Ws = o on new data

1.) What factors of the pretraining data lead to the formation of linear features of some but not all relations in LMs?

2.)When do linear features emerge in pretraining?

What do we look for in the pretraining data?

What factors of the pretraining data lead to the formation of linear features of relations in LMs?

What do we look for in the pretraining data?

What factors of the pretraining data lead to the formation of linear features of relations in LMs?

Hypothesis: More **subjects** and **object** co-occurrences = better linear features (proxy for fact mentions, Elsahar et al., 2018).

What do we look for in the pretraining data?

What factors of the pretraining data lead to the formation of linear features of relations in LMs?

Hypothesis: More **subjects** and **object** co-occurrences = better linear features (proxy for fact mentions, Elsahar et al., 2018).

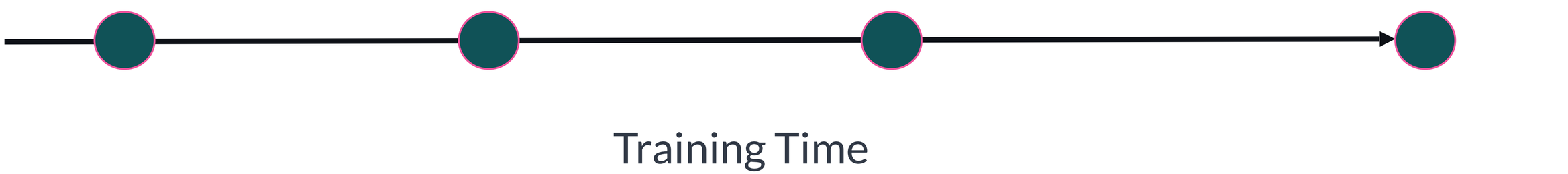
"Paris is the capital of France", ..., "Meanwhile in Paris, France"

"In June, Oracle will move their headquarters to Austin, Texas"

How do we count these Co-Occurrences?

Problem:

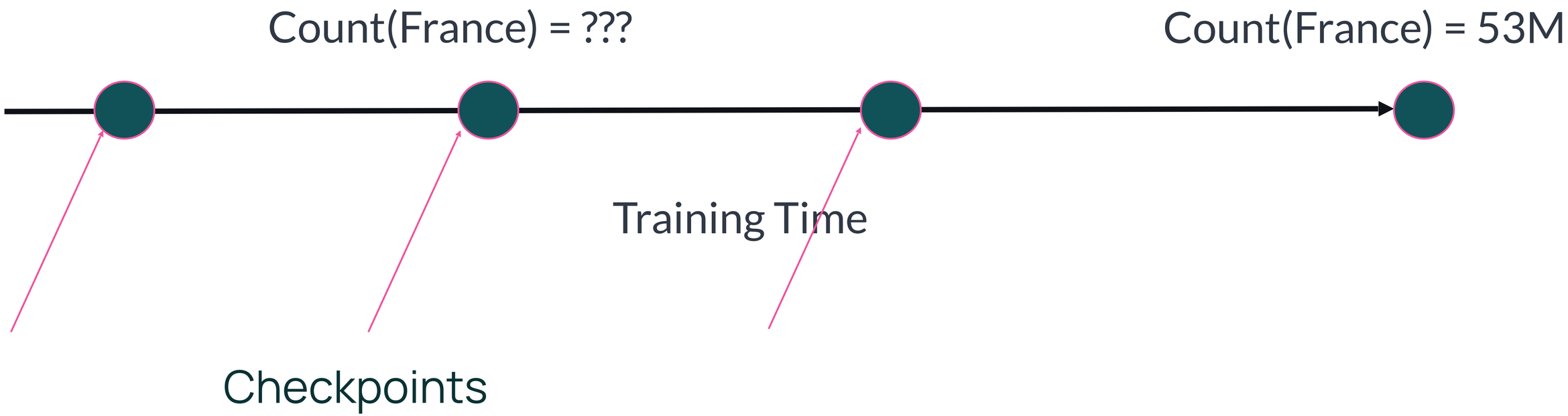
- We have term frequency counts from What's in My Big Data (WIMBD), but only for the final checkpoints (full dataset)



How do we count these Co-Occurrences?

Problem:

- We have term frequency counts from What's in My Big Data (WIMBD), but only for the final checkpoints (full dataset)
- How do we count what a checkpoint has seen?



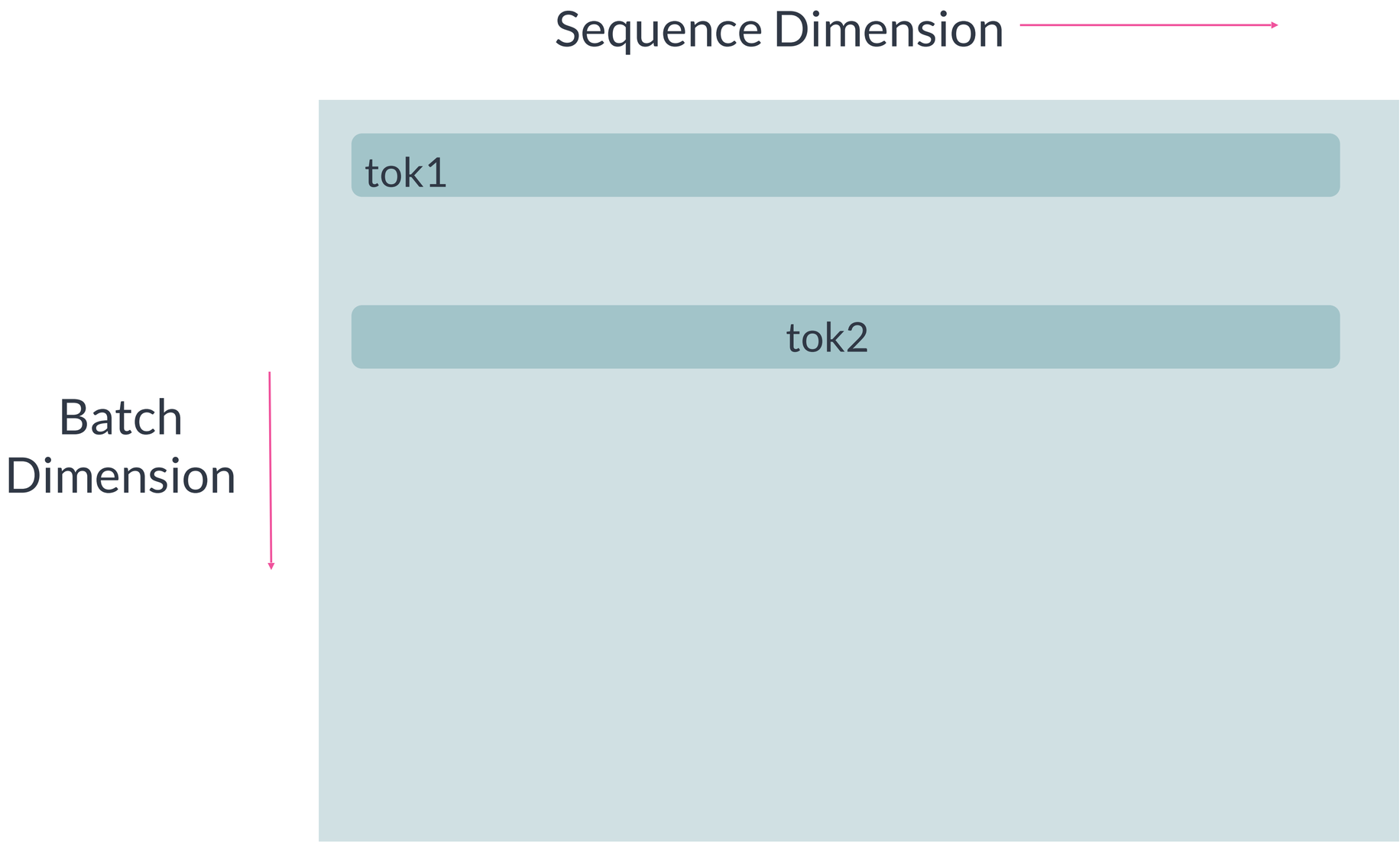
Batch Search

Counting terms in Training batches for any training step

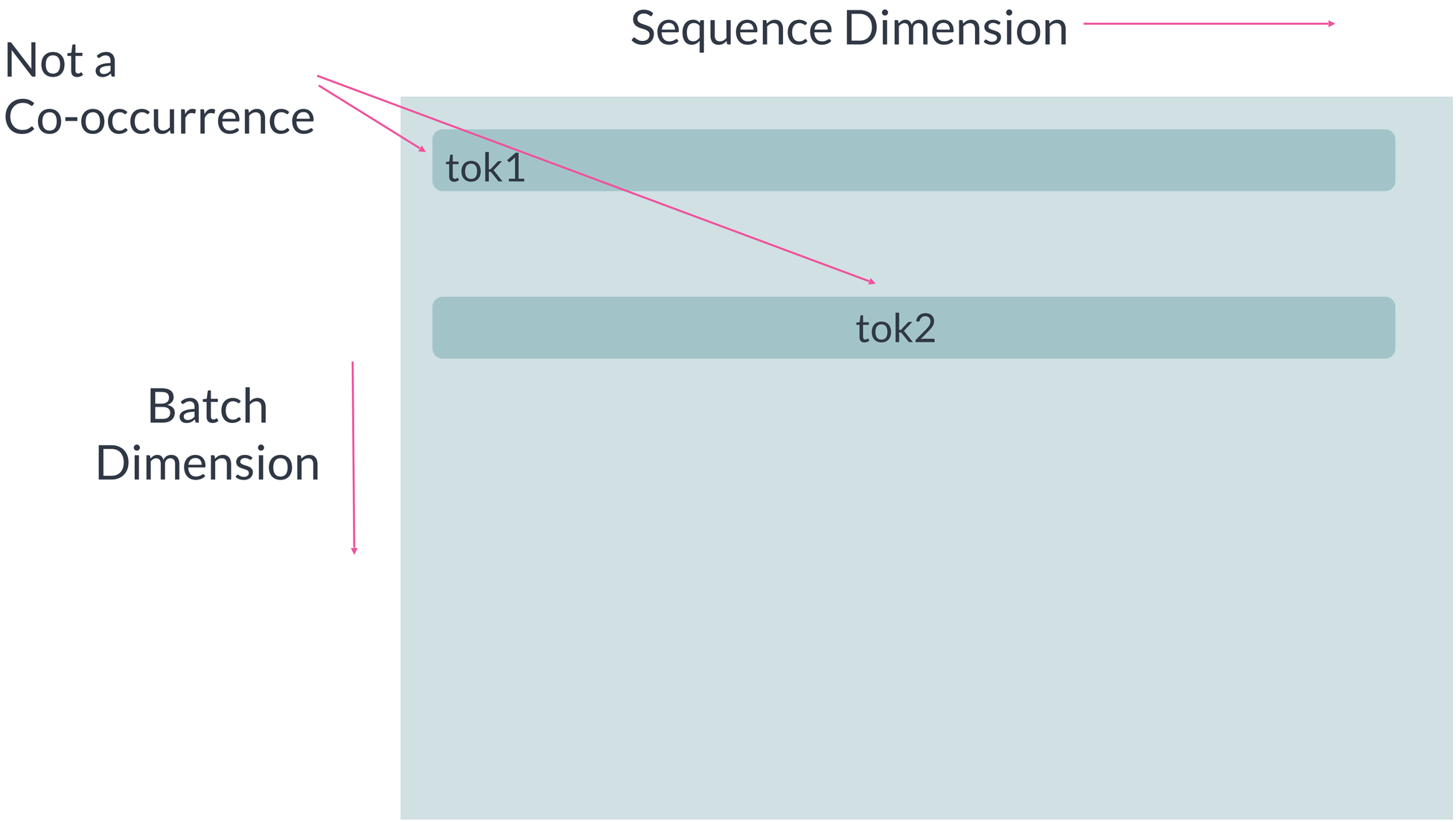
1. Reconstruct all of the training batches for OLMo
2. Count each term in every sequence

10k+ terms, and 2T tokens

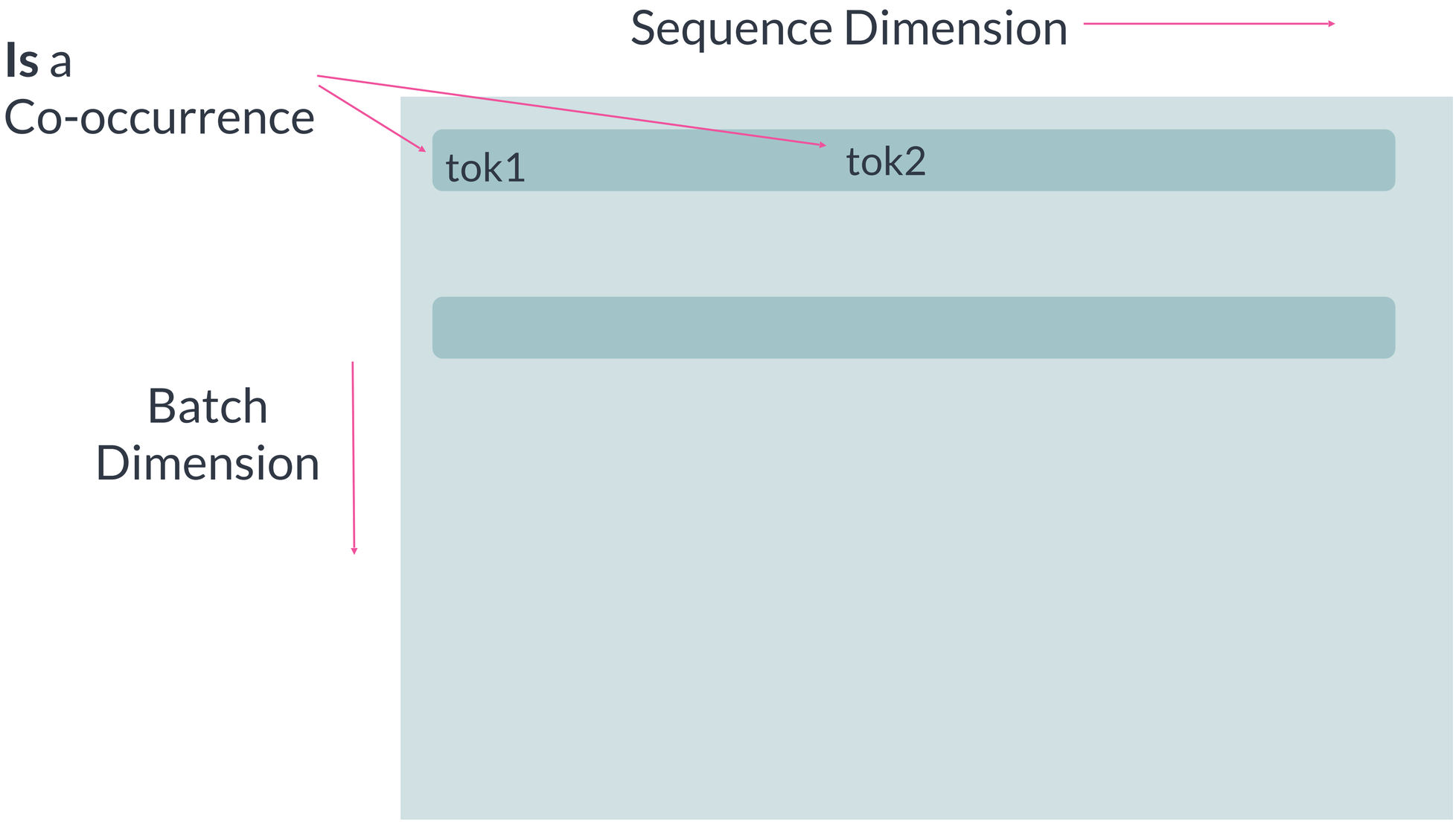
Counting (Co-)Occurrences Throughout Training



Counting (Co-)Occurrences Throughout Training



Counting (Co-)Occurrences Throughout Training



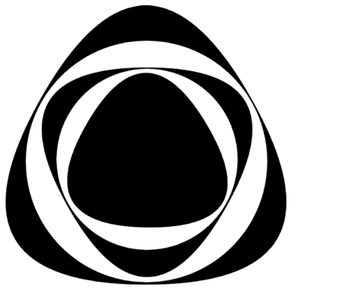
- Count subject-object occs. for the full dataset

- Plot against causality performance

Methods

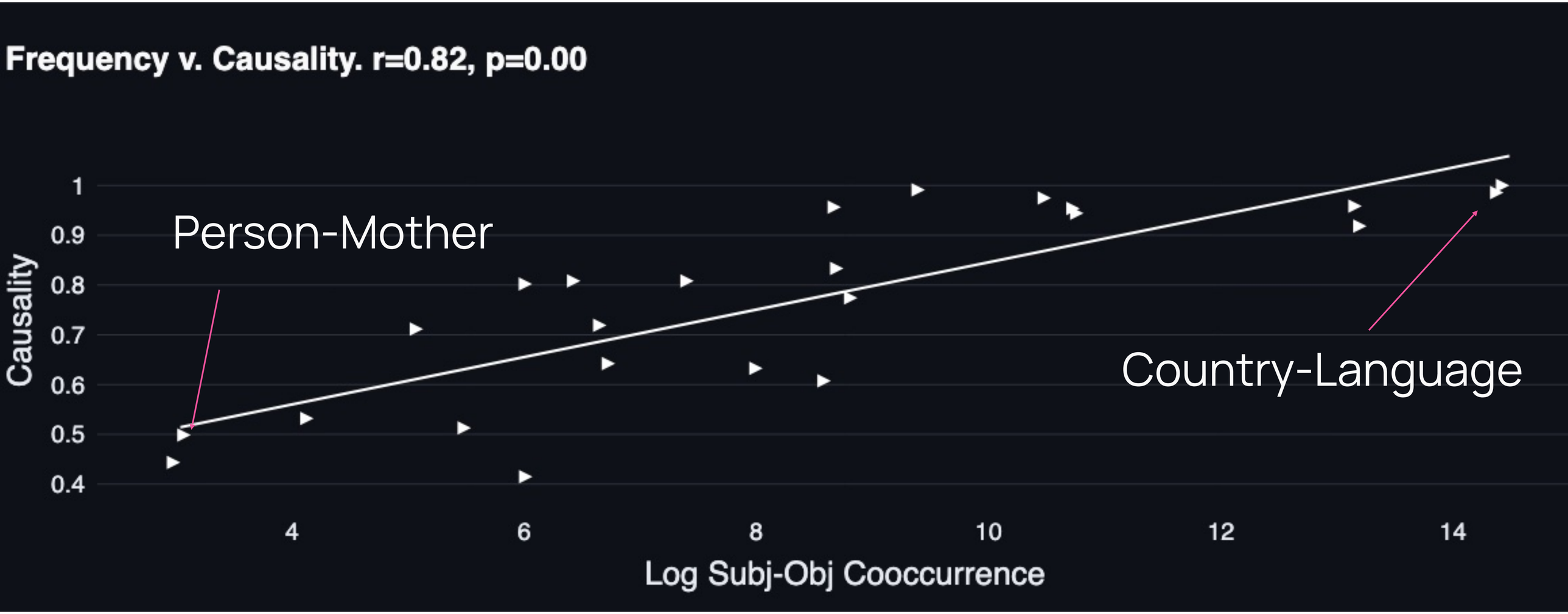
Data: 25 factual recall relations, like “Oracle’s headquarters is in Texas”

10k+ unique subjects and objects

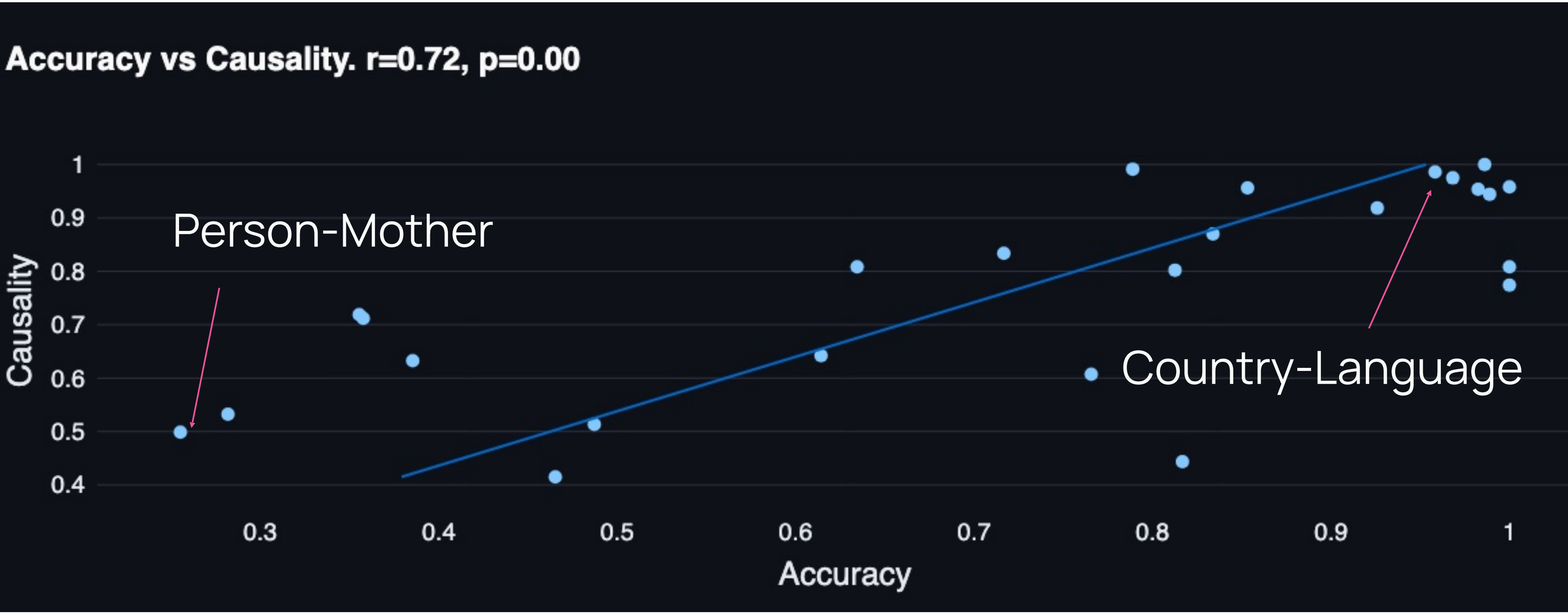


GPT-J

Causality vs Co-Occurrences (Avg. per Relation)



Causality vs Accuracy (Avg. per Relation)



Linear Features and Training Frequency

So far:

- Presence of linear features is highly correlated with high subj/obj co-occurrence statistics
  - Similar to accuracy+frequency

When do these emerge in pretraining?

OLMo Checkpoints

10k steps

25k

50k

100k

150k

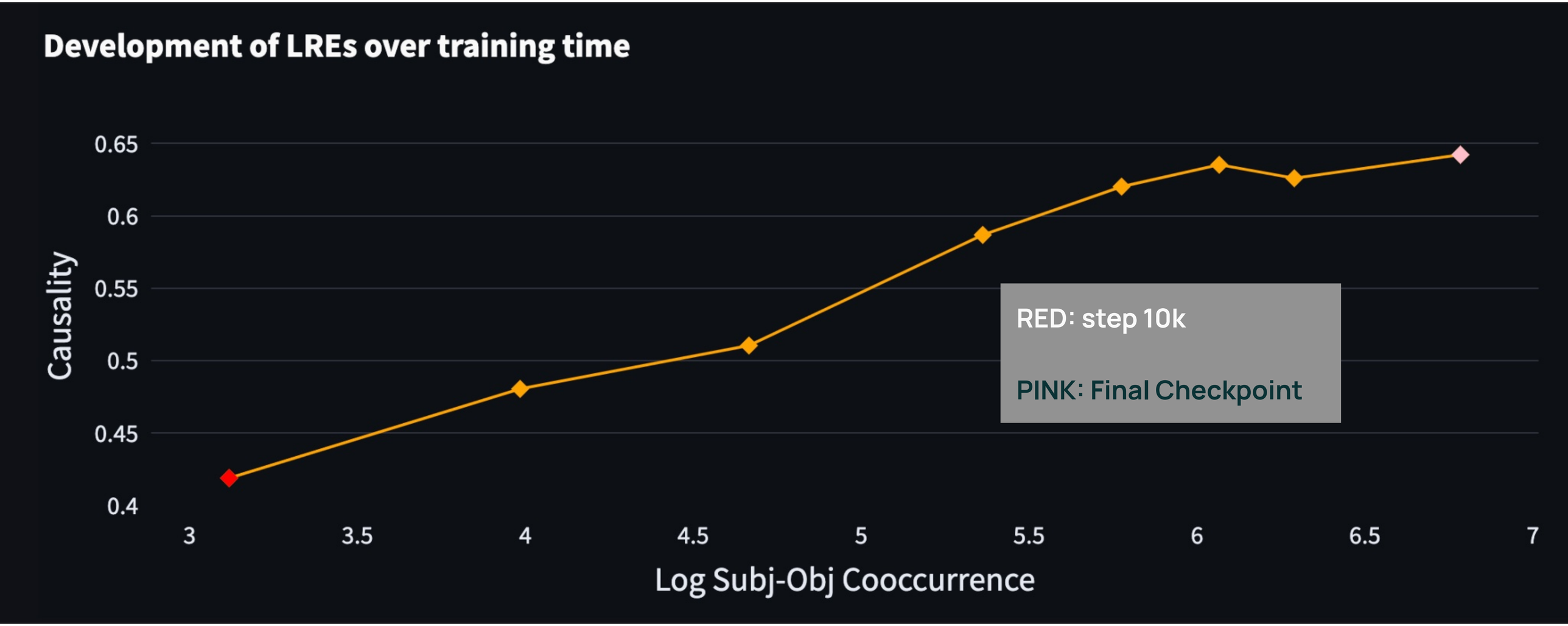
200k

250k

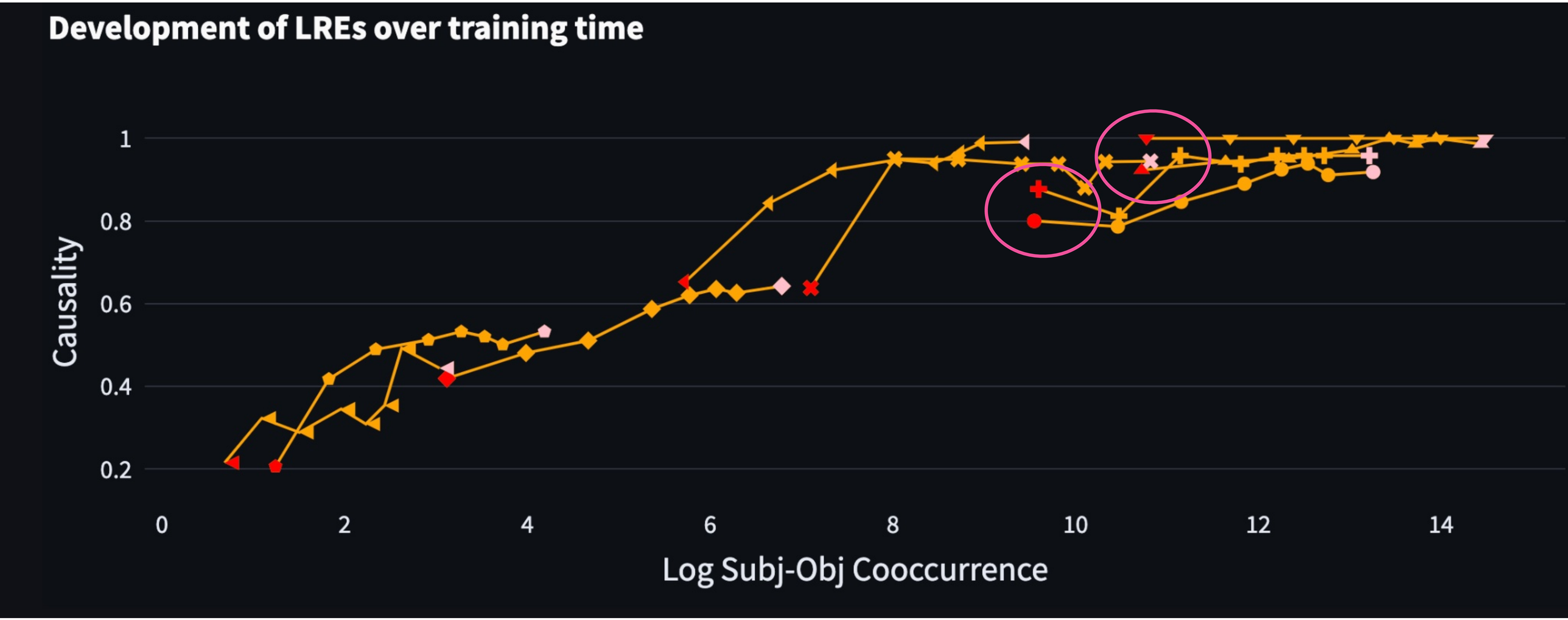
550k (final)

8 Checkpoints

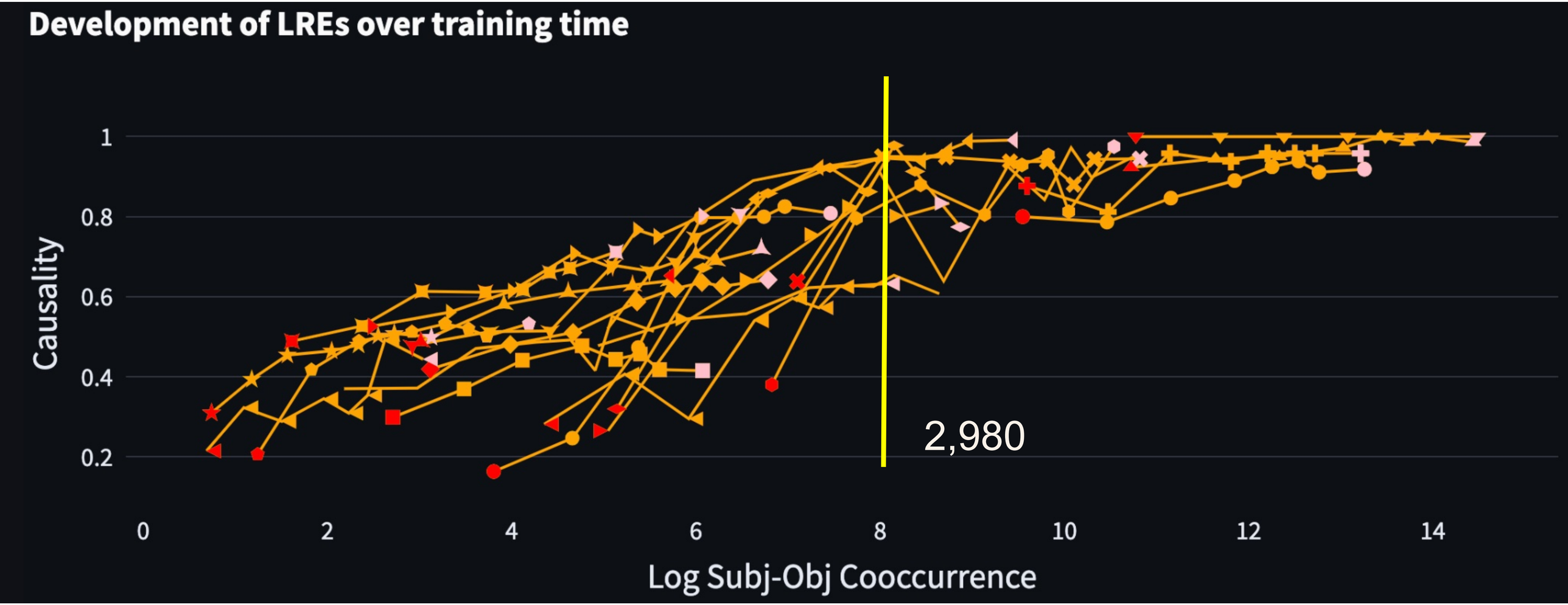
LRE Development: Company HQ relation



LRE Development



Threshold for Linear Features



Threshold for Linear Features

This threshold of ~3k holds for GPT-J as well

Avg Causality above and below threshold	>8	<=8
GPTJ	94	59
OLMo	91	56

Linear features are more correlated with frequency rather than training time

The Frequency←-› LRE Relationship

So far:
Frequency info predicts formation of
linear features after some threshold

The Frequency ← → LRE Relationship

On Linear Representations and Pretraining Data Frequency in Language Models

Jack Merullo[◇] **Noah A. Smith**[♡♣] **Sarah Wiegreffe**[*♡♣] **Yanai Elazar**[*♡♣]

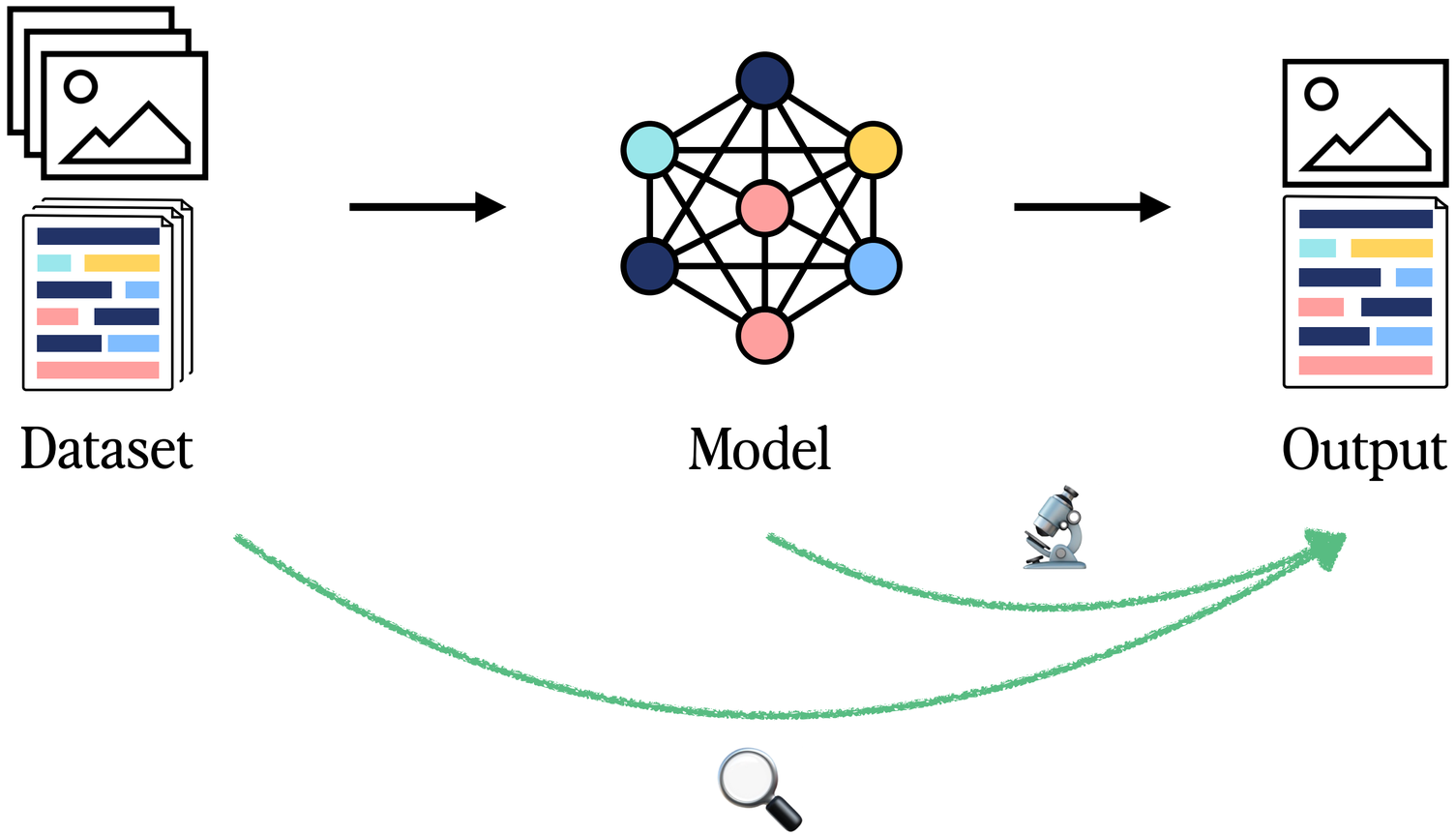
[◇]Brown University, [♡]Allen Institute for AI (Ai2), [♣]University of Washington
[*]Co-senior authors.
jack_merullo@brown.edu, {noah, sarahw, yanaie}@allenai.org

Merullo et al., ICLR 2025

Interpretability Research

- "**Revealing the content of the neural black box**"
  [The 1st BlackboxNLP workshop, EMNLP 2018]
- ***Mechanistic Interpretability*** focuses on "**how?**"
  - "Works at the level of intermediate representations"
    [Saphra and Wiegreffe, BlackboxNLP 2024]
- ***Data Interpretability*** focuses on "**why?**"
  - "Quantifying the effect [of data abstraction] on model behavior"

Interpretability Research



Data Interpretability

- Why models learn or don’t learn?
- Given a model’s behavior, how to explain it??
- Can we develop a general theory???

Data Interpretability

Two Components

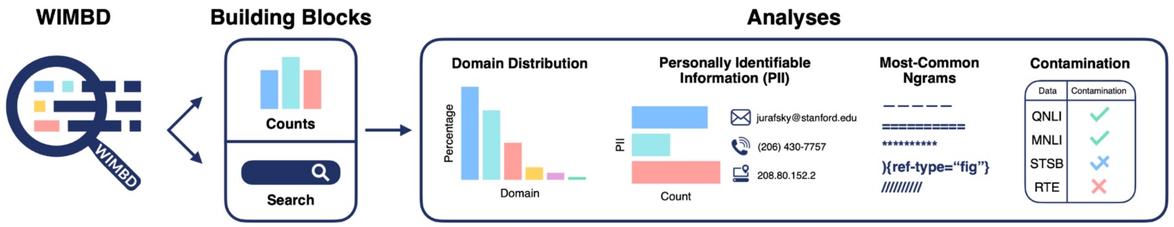
- Convenient & efficienient
  - Abstraction
  - Extraction
- Causal hypothesis testing

Data Interpretability

Two Components

- Convenient & efficienient
  - Abstraction
  - Extraction
- Causal hypothesis testing

What's In My Big Data?



Elazar et al., ICLR 2024

Rusty Dawg

Merrill et al., EMNLP 2024

BatchSearch

Merullo et al., ICLR 2025

Data Interpretability

Two Components

- Convenient & efficienient
  - Abstraction
  - Extraction
- Causal hypothesis testing

Rewriting History: A Recipe for Interventional Analyses to Study Data Effects on Model Behavior

Rahul Nadkarni[1] **Yanai Elazar**[2*] **Hila Gonen**[3*] **Noah A. Smith**[1,4]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Department of Computer Science, Bar-Ilan University
[3]Department of Computer Science, University of British Columbia
[4]Allen Institute for Artificial Intelligence

GENERALIZATION V.S. MEMORIZATION: TRACING LANGUAGE MODELS' CAPABILITIES BACK TO PRETRAINING DATA

Xinyi Wang[1*], **Antonis Antoniades**[1*], **Yanai Elazar**[2,3], **Alfonso Amayuelas**[1], **Alon Albalak**[4], **Kexun Zhang**[5], **William Yang Wang**[1]
[1]University of California, Santa Barbara, [2]Allen Institute for AI,
[3]University of Washington, [4] SynthLabs, [5]Carnegie Mellon University
{xinyi_wang,antonis}@ucsb.edu, william@cs.ucsb.edu

ON LINEAR REPRESENTATIONS AND PRETRAINING DATA FREQUENCY IN LANGUAGE MODELS

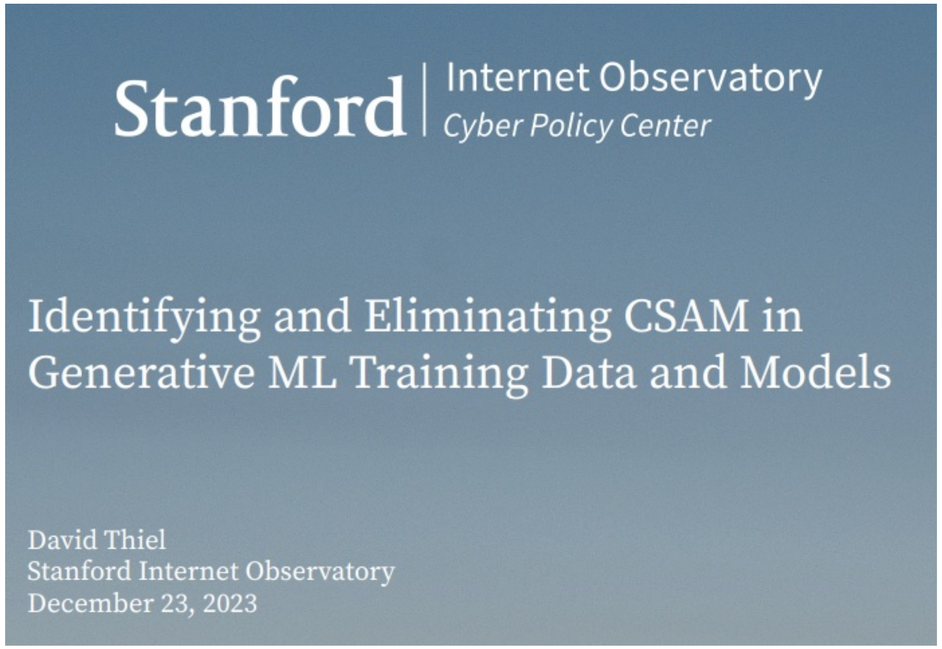
Jack Merullo[◇] **Noah A. Smith**[♡♣] **Sarah Wiegreffe**[*♡♣] **Yanai Elazar**[*♡♣]

[◇]Brown University, [♡]Allen Institute for AI (Ai2), [♣]University of Washington
*Co-senior authors.
jack_merullo@brown.edu, {noah, sarahw, yanaie}@allenai.org

Detection and Measurement of Syntactic Templates in Generated Text

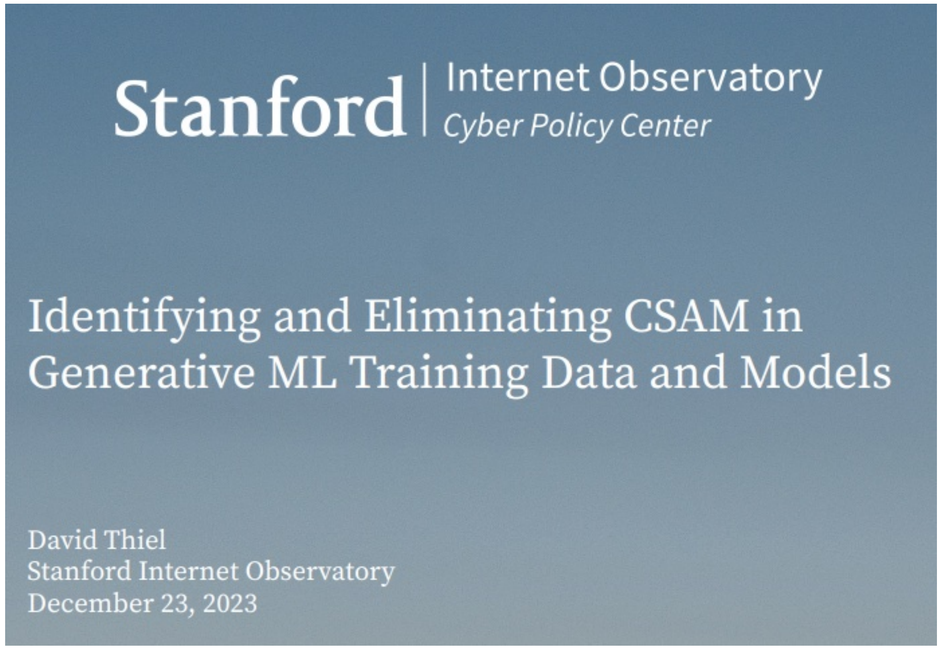
Chantal Shaib[1] **Yanai Elazar**[2,3] **Junyi Jessy Li**[4] **Byron C. Wallace**[1]
[1]Northeastern University, [2]Allen Institute for AI,
[3]University of Washington, [4]The University of Texas at Austin
{shaib.c, b.wallace}@northeastern.edu

Memorization & Generalization



Memorization & Generalization

- Would a single (CSAM) instance be reproduced?
- Would two?
- Would three?
- …
- But also,
  What if we were to remove all of the CSAM?
- Images of children and adult content would still remain



Imitation

✅

Leonardo DiCaprio

❌

Yanai Elazar

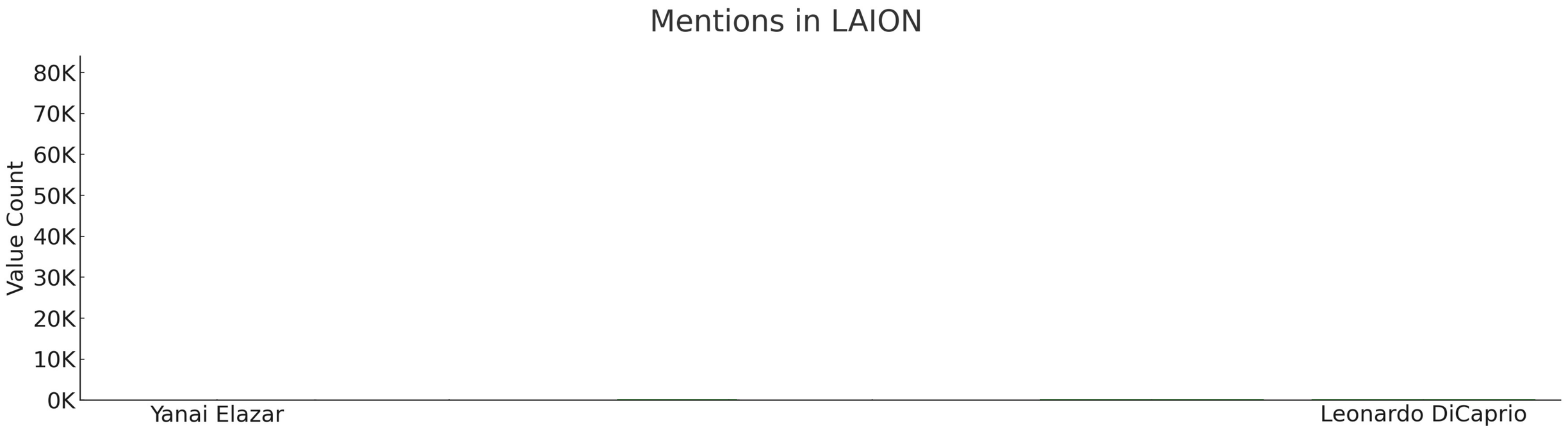
Imitation

Spot the difference

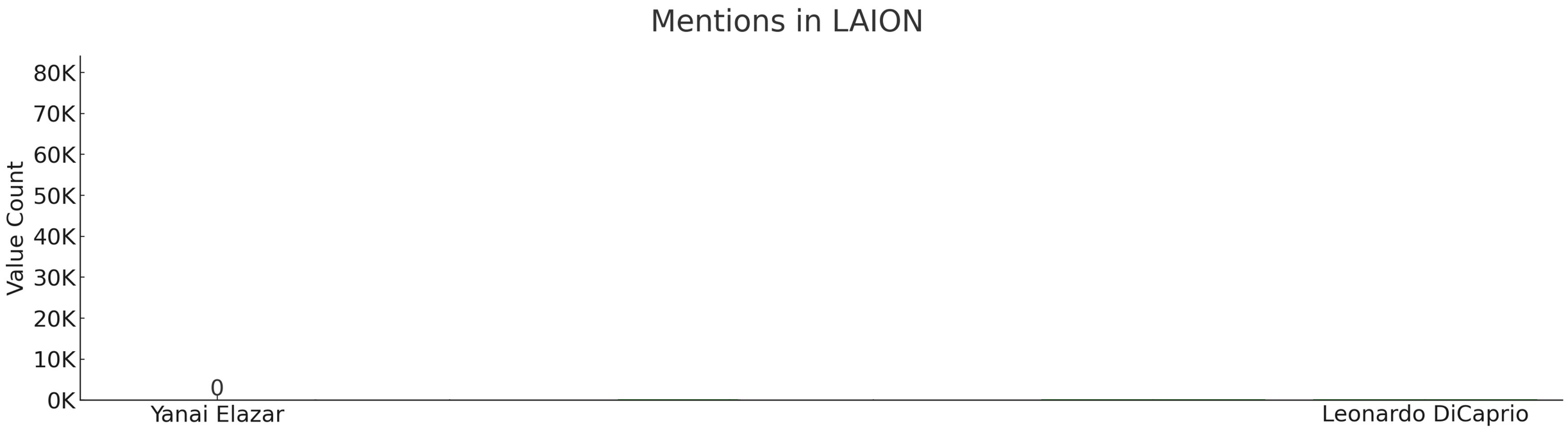
Leonardo DiCaprio

Yanai Elazar

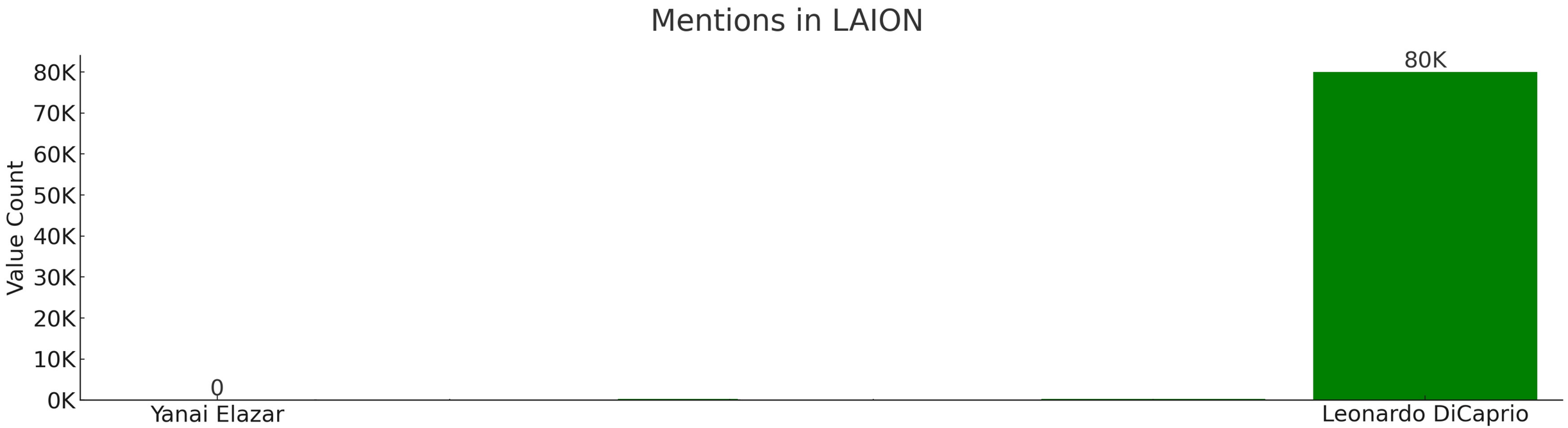
Imitation Threshold?



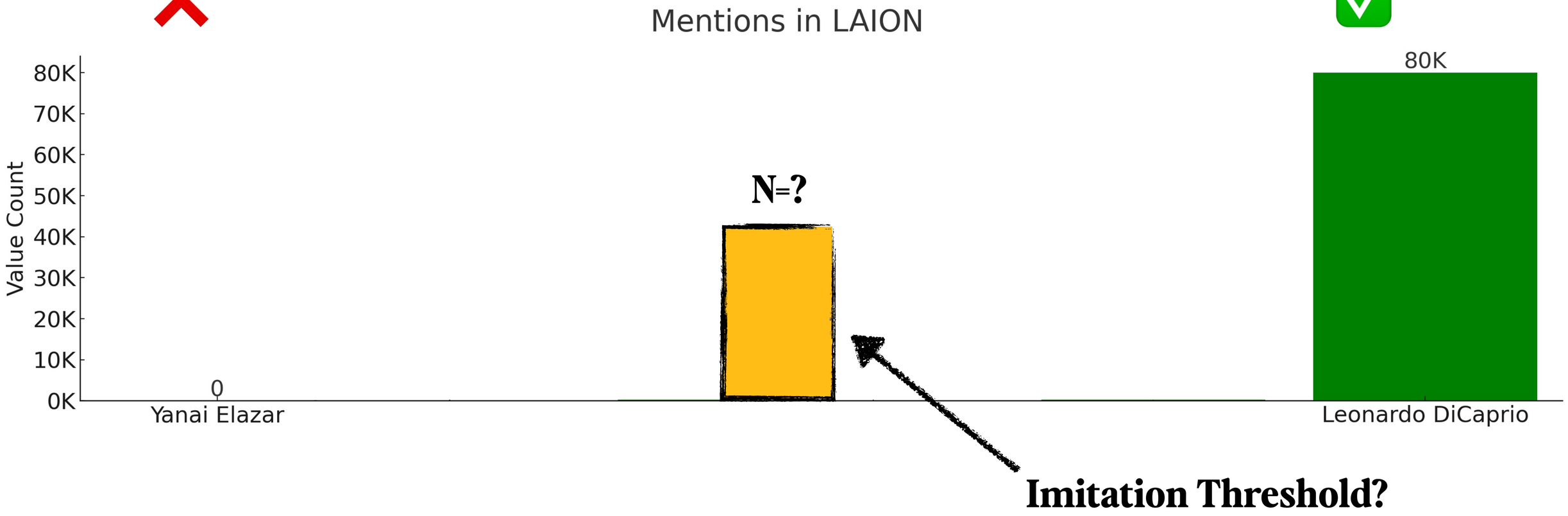
Imitation Threshold?



Imitation Threshold?



Imitation Threshold?



Imitation - Why Should You Care?

- Copyrights

Imitation - Why Should You Care?

- Copyrights

Imitation - Why Should You Care?

- Copyrights
- Privacy

Celebrity → Leonardo DiCaprio

Private individual → Yanai Elazar

Finding the Imitation Threshold

How Many Van Goghs Does It Take to Van Gogh? Finding the Imitation Threshold

Sahil Verma[1] **Royi Rassin**[2] **Arnav Das**[*1] **Gantavya Bhatt**[*1] **Preethi Seshadri**[*3]
Chirag Shah[1] **Jeff Bilmes**[1] **Hannaneh Hajishirzi**[1,4] **Yanai Elazar**[1,4]

[1]*University of Washington, Seattle* [2]*Bar-Ilan University* [3]*University of California, Irvine*
[4]*Allen Institute of AI*







TMLR 2025

Question Formulation

LAION-5B'

Would the model imitate a concept (e.g., *Leo*) if it was trained on *X* of his images instead?

LAION-5B

Count: *100*

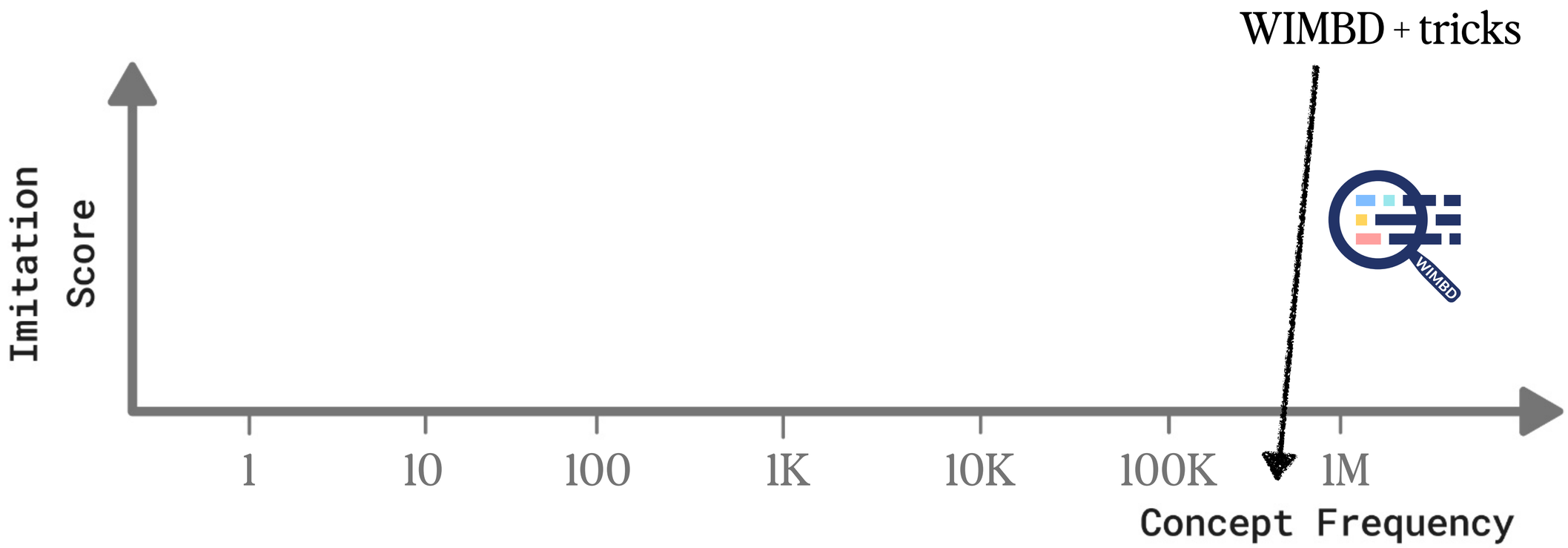
$$P(Imitation \mid do(count(Leo) = x))$$

Count: 80*K*

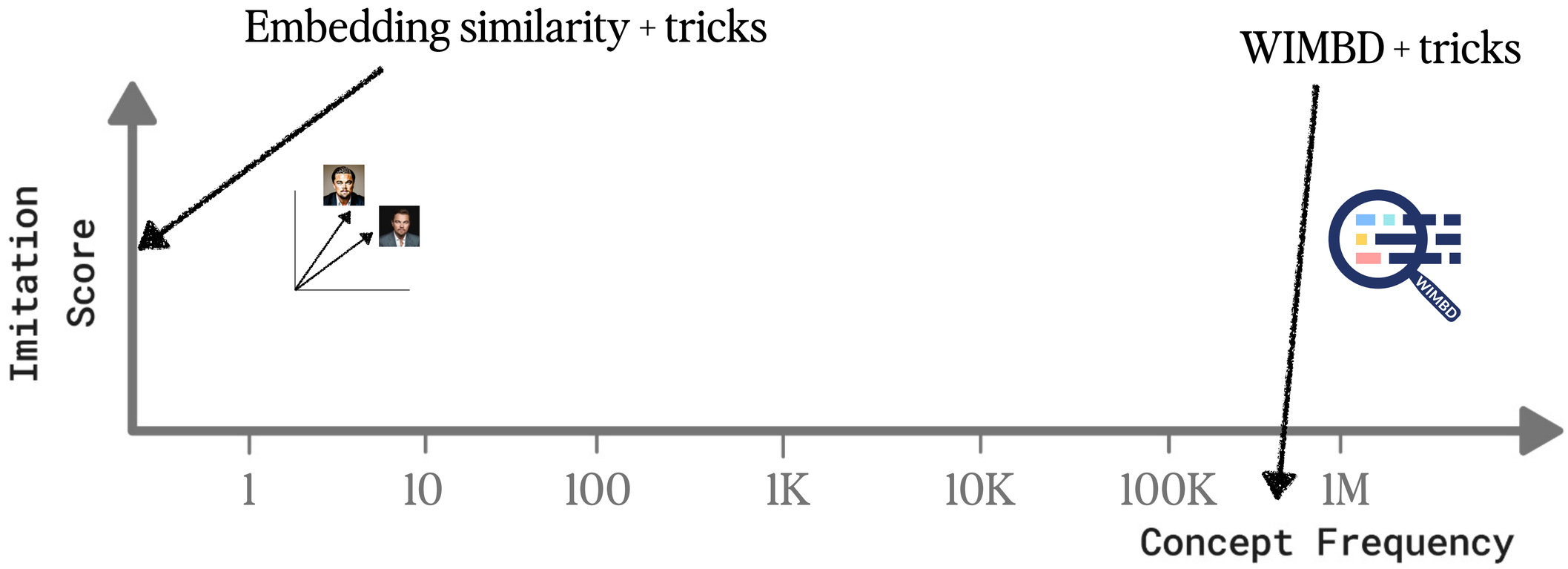
Solutions

1. Counterfactual model

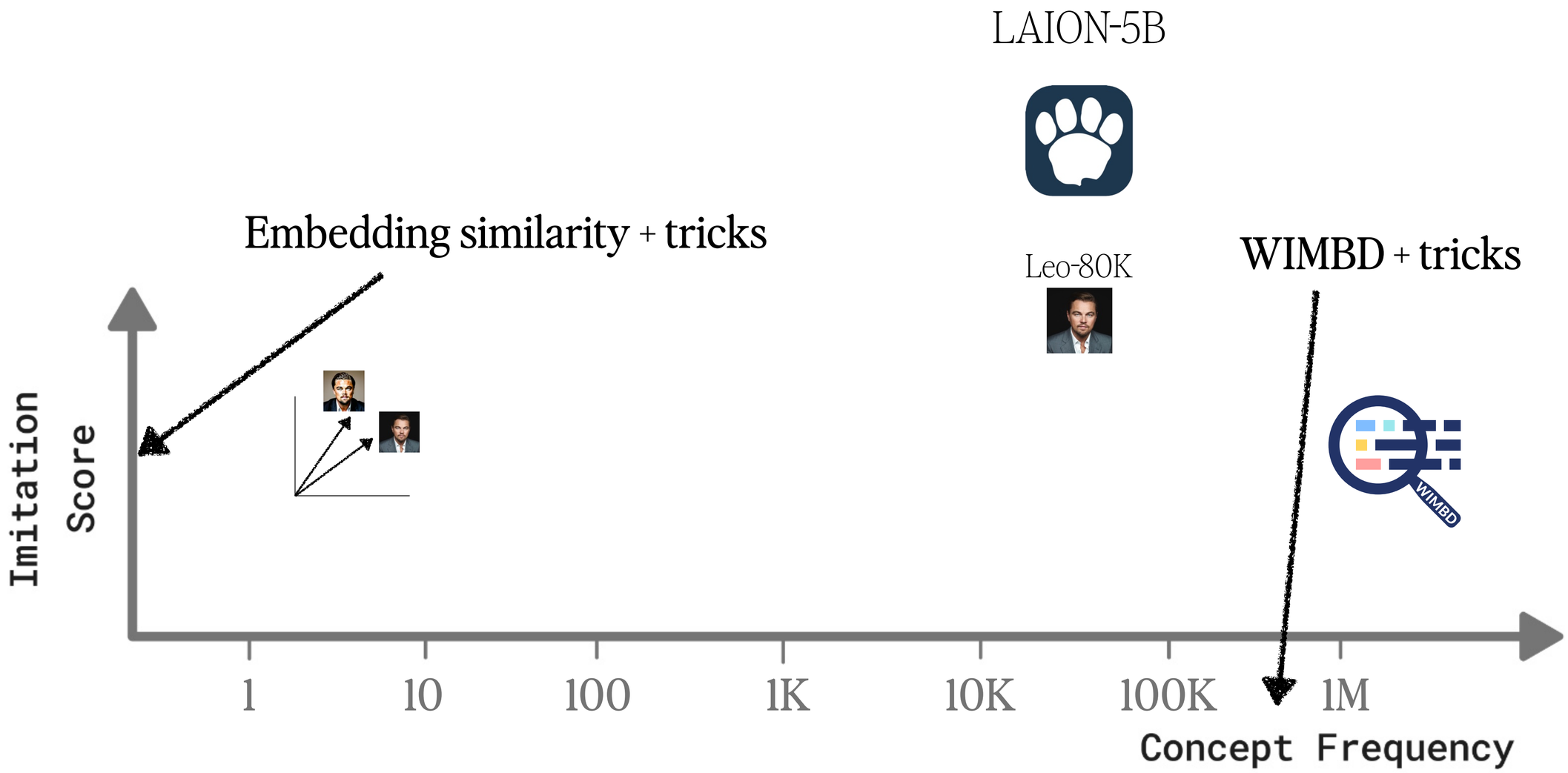
Solutions


WIMBD + tricks
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solutions


Embedding similarity + tricks
WIMBD + tricks
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency
WIMBD

Solutions


LAION-5B
Embedding similarity + tricks
Leo-80K
WIMBD + tricks
WIMBD
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solution #1

LAION-5B': The Leos

LAION-5B

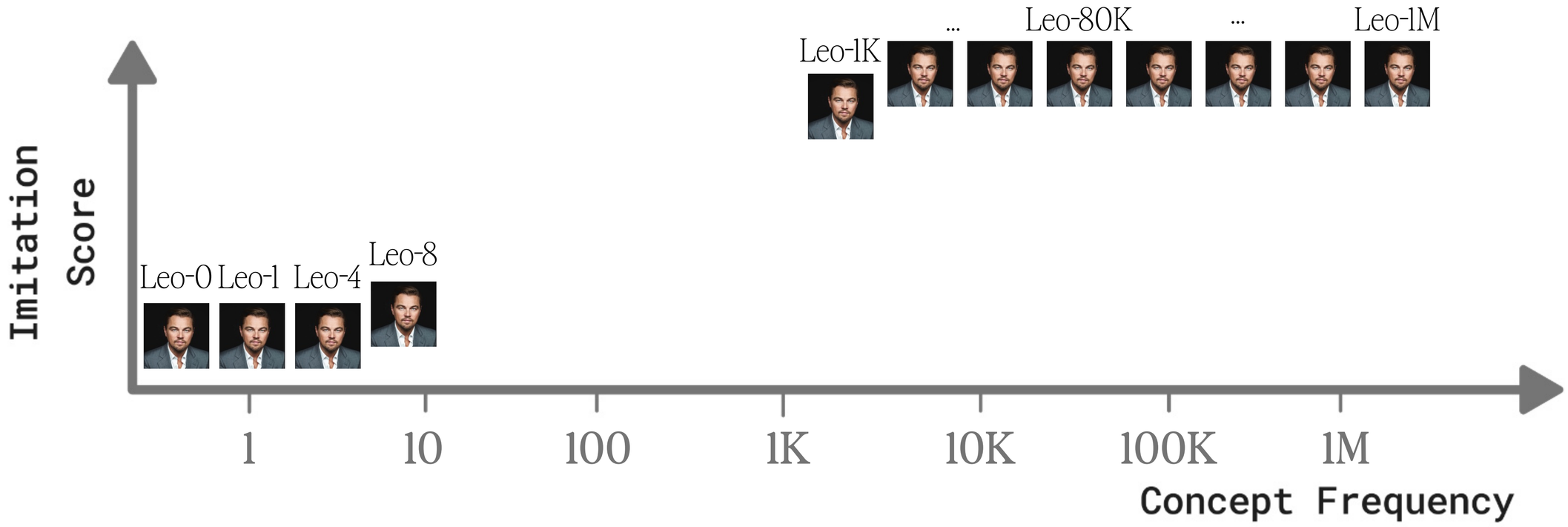
Leo-80K

Leo-0

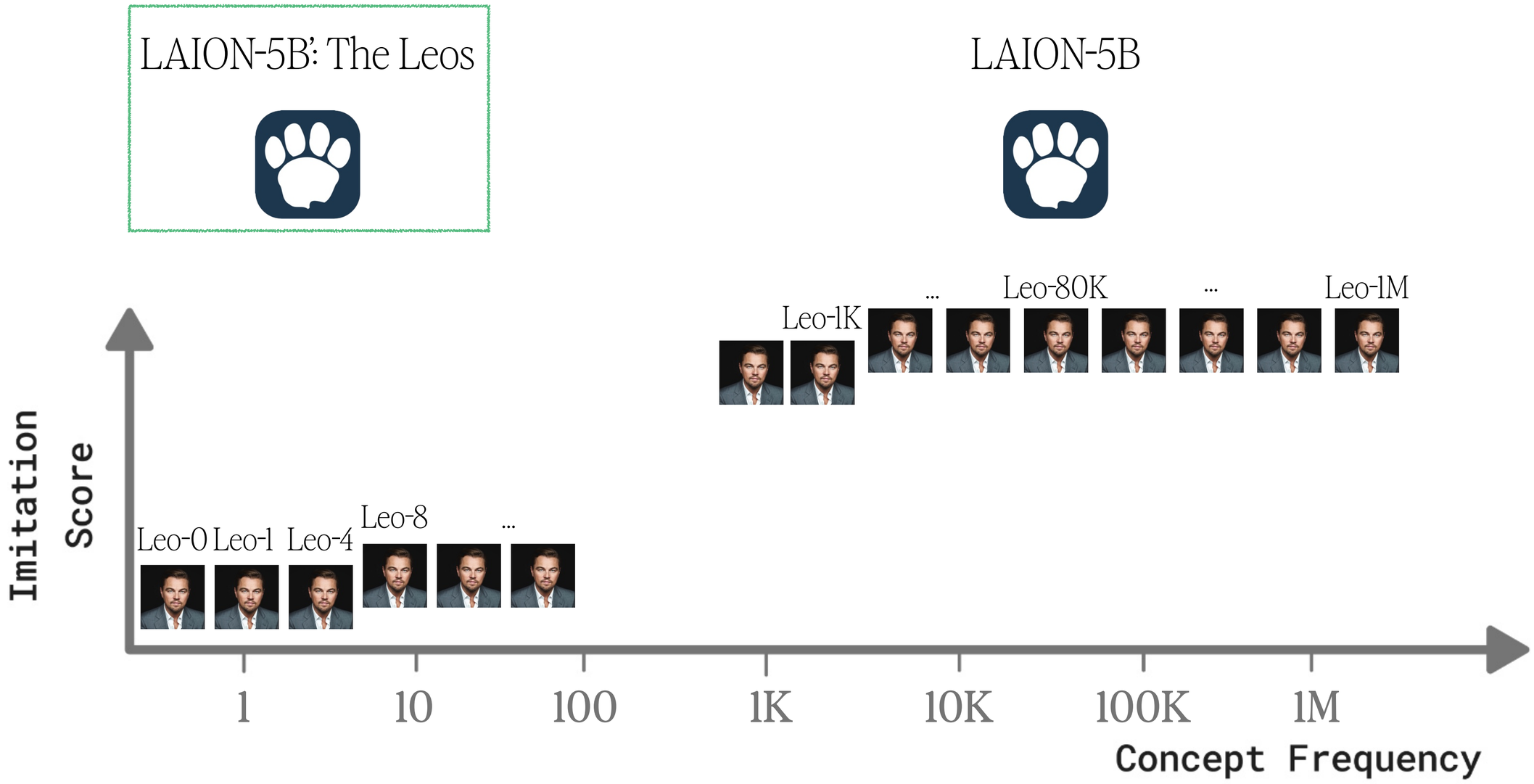
Solution #1

LAION-5B': The Leos

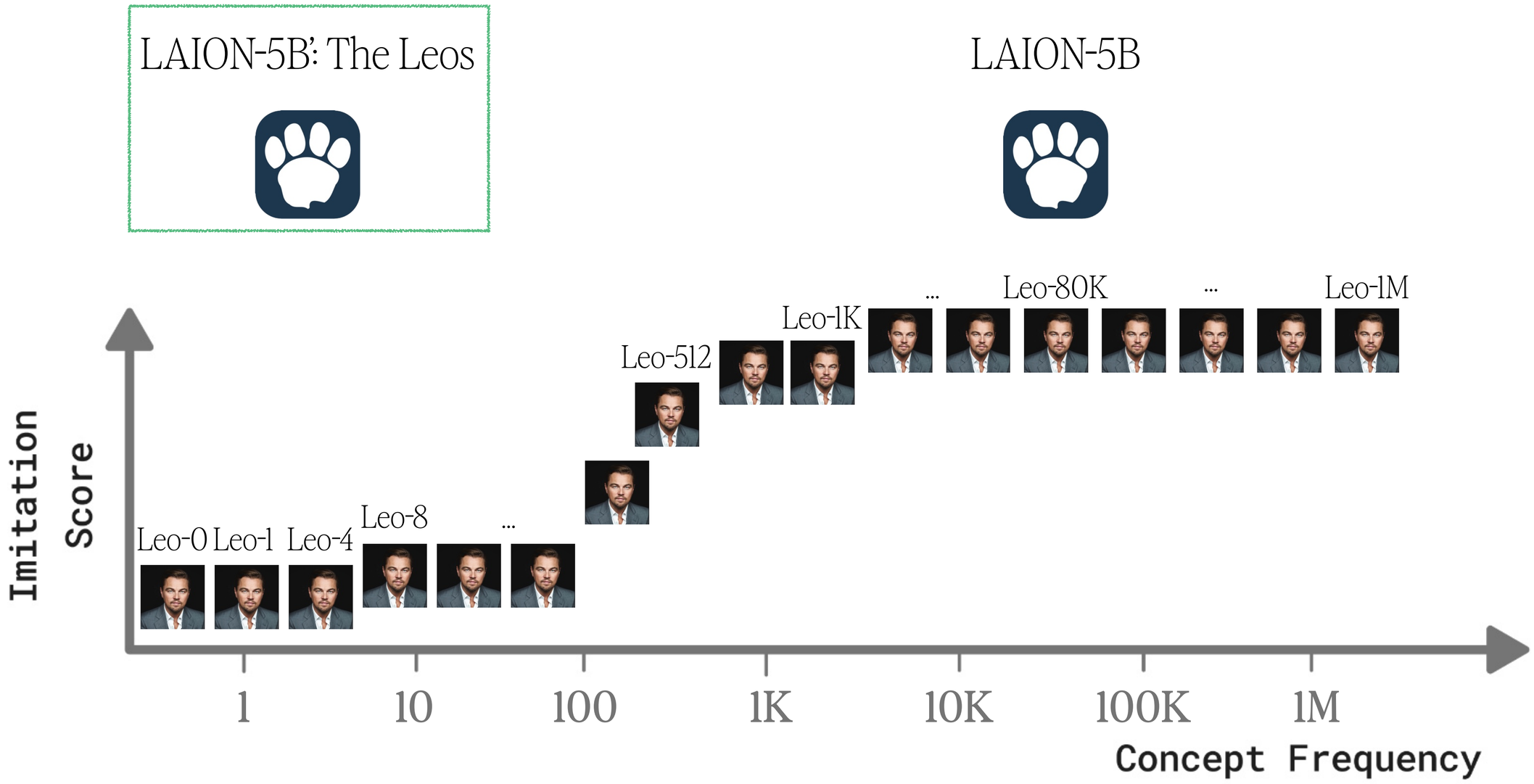
LAION-5B



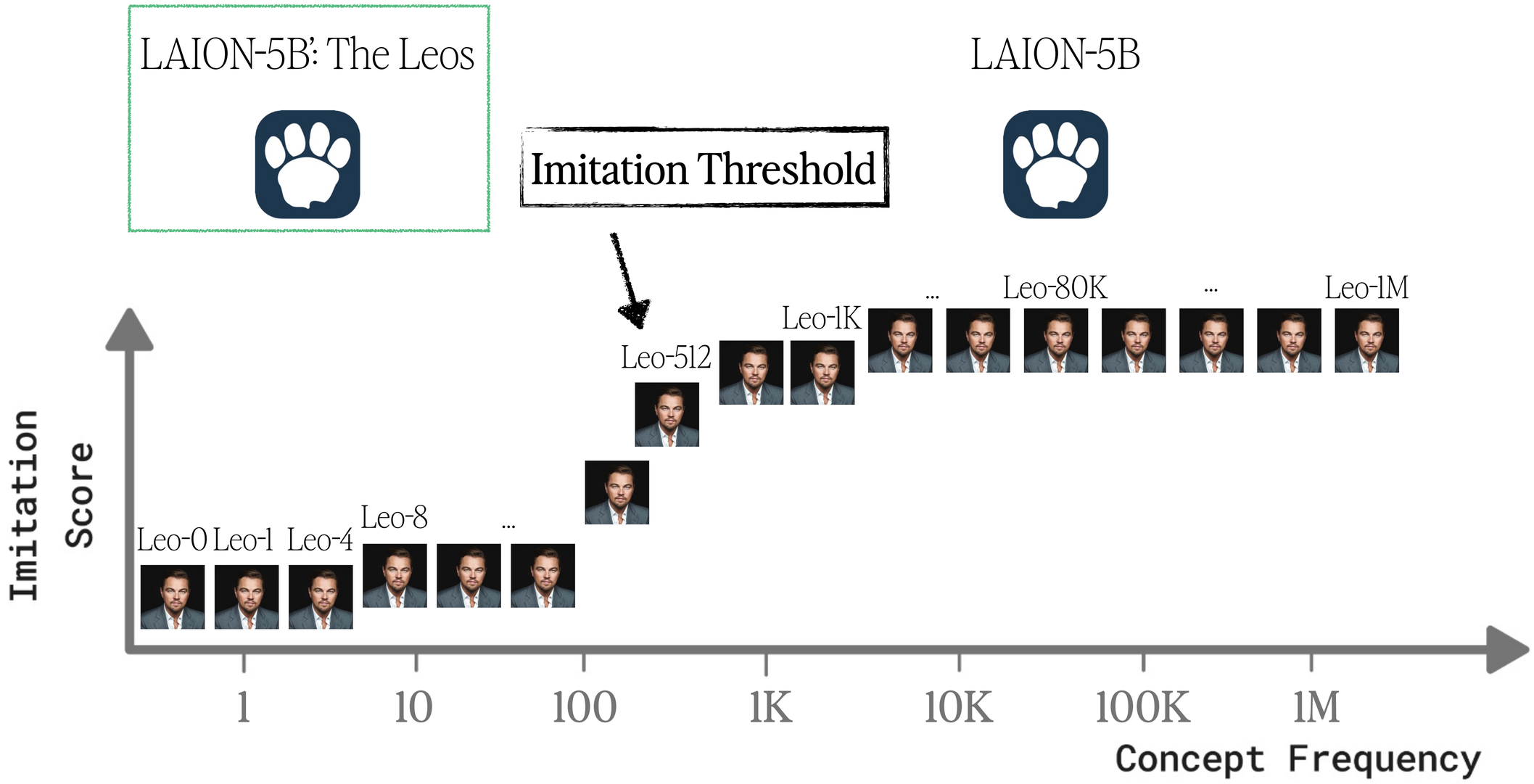
Solution #1


LAION-5B': The Leos
LAION-5B
Leo-1K
...
Leo-80K
...
Leo-1M
Leo-0
Leo-1
Leo-4
Leo-8
...
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solution #1


LAION-5B': The Leos
LAION-5B
Leo-0
Leo-1
Leo-4
Leo-8
...
Leo-512
Leo-1K
...
Leo-80K
...
Leo-1M
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solution #1


LAION-5B': The Leos
LAION-5B
Imitation Threshold
Leo-0
Leo-1
Leo-4
Leo-8
...
Leo-512
Leo-1K
...
Leo-80K
...
Leo-1M
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solution #1

LAION-5B': The Leos

LAION-5B

Imitation Threshold

Leo-0 Leo-1 Leo-4 ... Leo-1K ... Leo-80K ... Leo-1M

Imitation Score

1 ...0K 1M

Concept Frequency

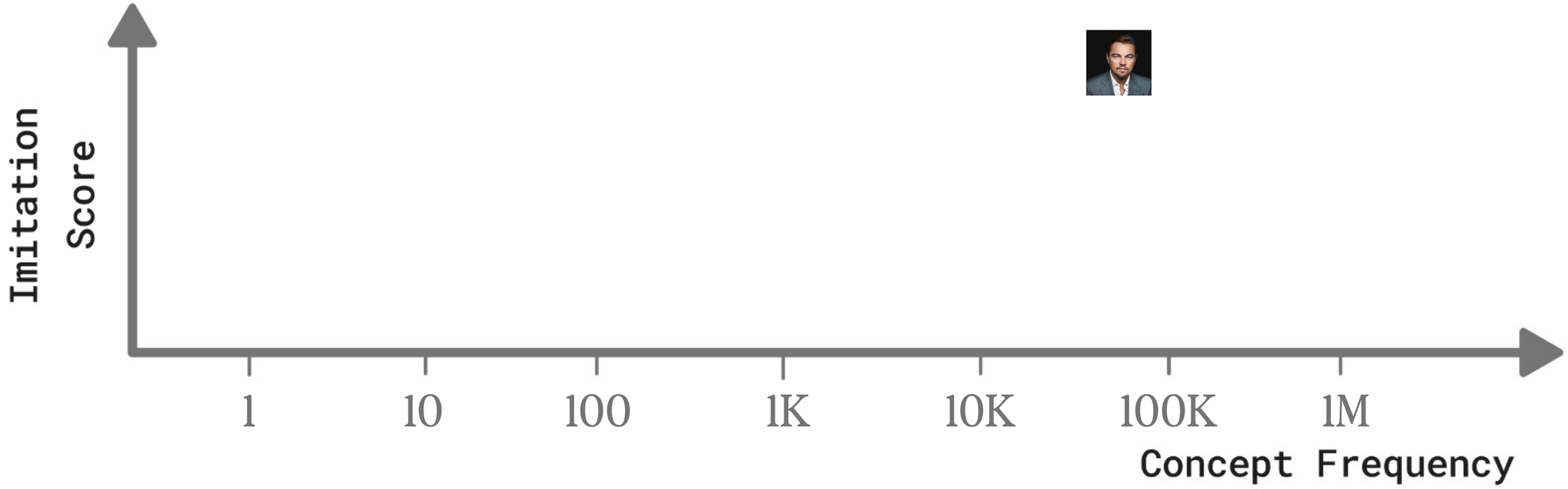
Solutions

1. Counterfactual model ❌

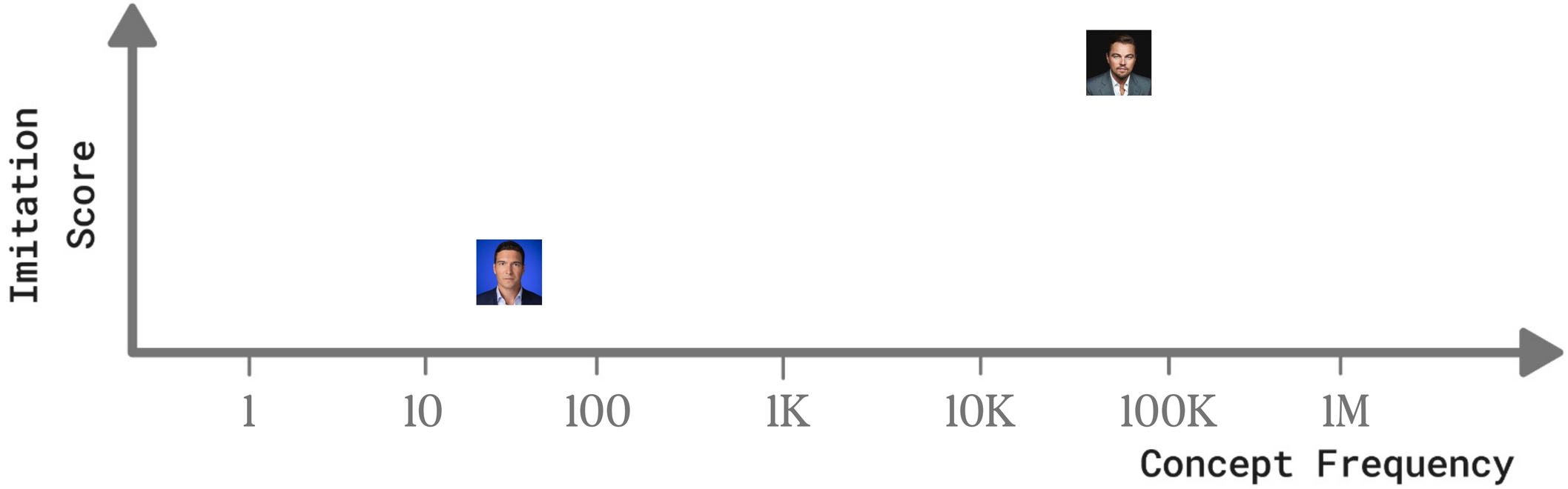
Solutions

1. Counterfactual model ❌
2. Observational approach

Solution #2



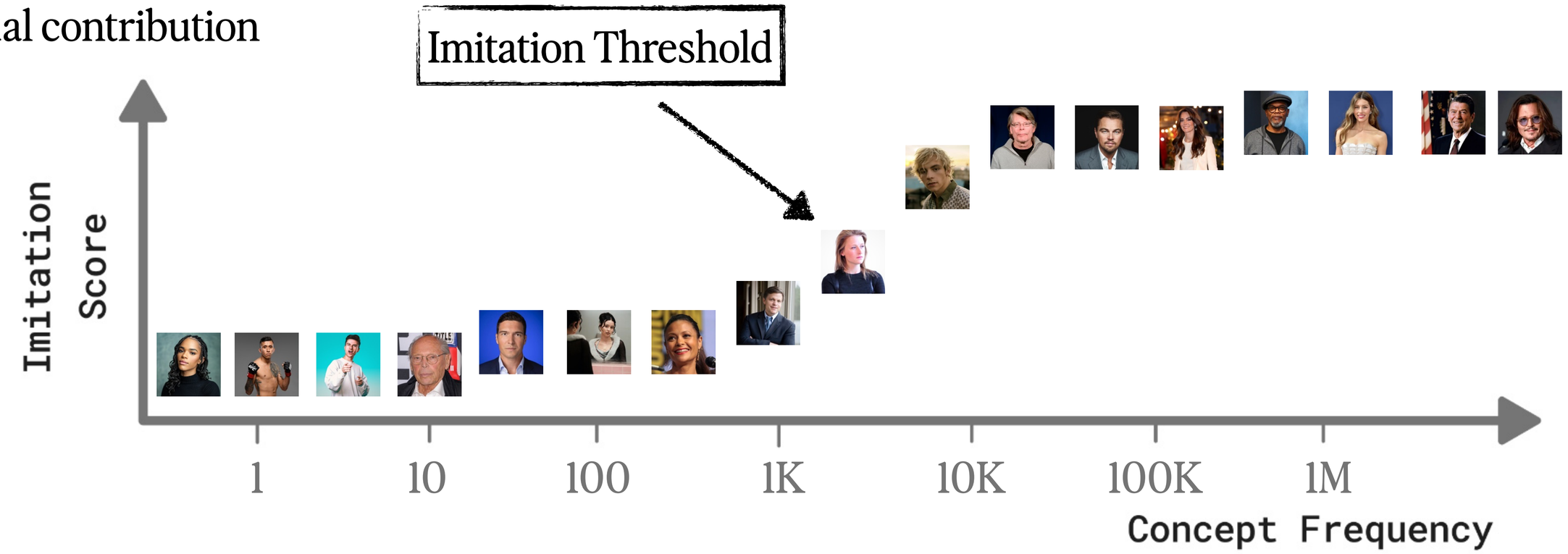
Solution #2



Solution #2

Using some assumptions:

- Distribution invariance
- Lack of confounders
- Equal contribution



Setup

2 domains x 2 datasets

	Human Faces 🧑		Art Style 🖼️	
	Celebrities	**Politicians**	**Classical**	**Modern**

Setup

3 pretraining datasets

	Human Faces 🧑		Art Style 🖼️	
Pretraining Dataset	**Celebrities**	**Politicians**	**Classical**	**Modern**
LAION-400M				
LAION2B				
LAION-5B				

Setup

4 models

		Human Faces 🧑		Art Style 🖼️	
Pretraining Dataset	**Model**	**Celebrities**	**Politicians**	**Classical**	**Modern**
LAION-400M	LD				
LAION2B	SD1.1 SD1.5				
LAION-5B	SD2.1				

Results

		Human Faces 🧑		Art Style 🖼️	
Pretraining Dataset	**Model**	**Celebrities**	**Politicians**	**Classical**	**Modern**
LAION-400M	LD	648	309	219	282
LAION2B	SD1.1	364	234	112	198
	SD1.5	364	234	112	198
LAION-5B	SD2.1	527	369	185	241

Results

		Human Faces 🧑		Art Style 🖼️	
Pretraining Dataset	**Model**	**Celebrities**	**Politicians**	**Classical**	**Modern**
LAION-400M	LD	648	309	219	282
LAION2B	SD1.1	364	234	112	198
	SD1.5	364	234	112	198
LAION-5B	SD2.1	527	369	185	241

Imitation Threshold: 100-650 images

The Imitation Threshold

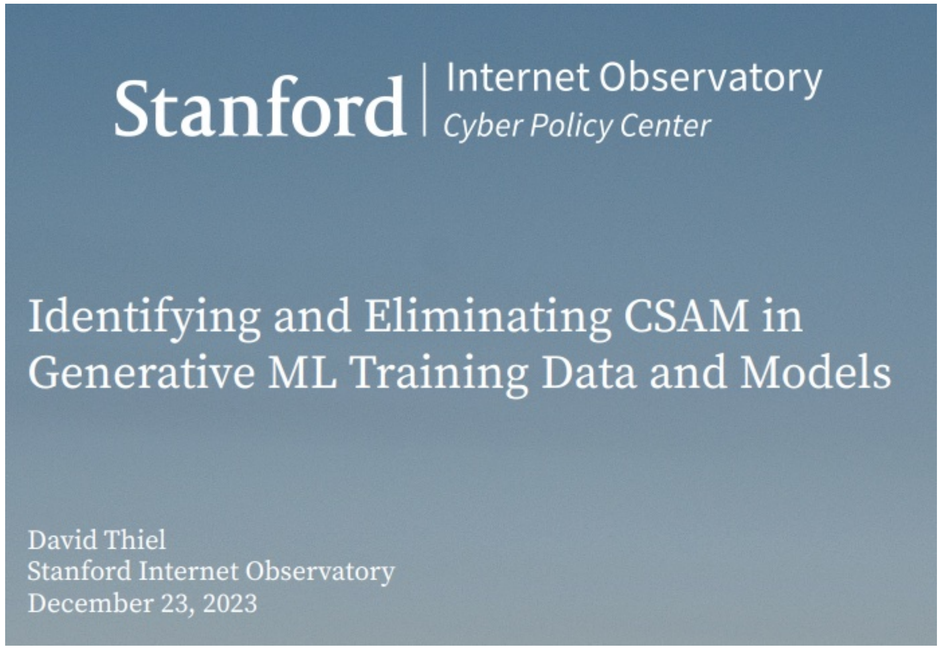
- Memorizing distribution requires to observe enough training instance
- We estimate it to be a few hundreds images
- Implications on privacy, copyrights, etc.

Memorization & Generalization

- Would a single (CSAM) instance be reproduced?
- Would two?
- Would three?
- …

No! We need a few hundreds!

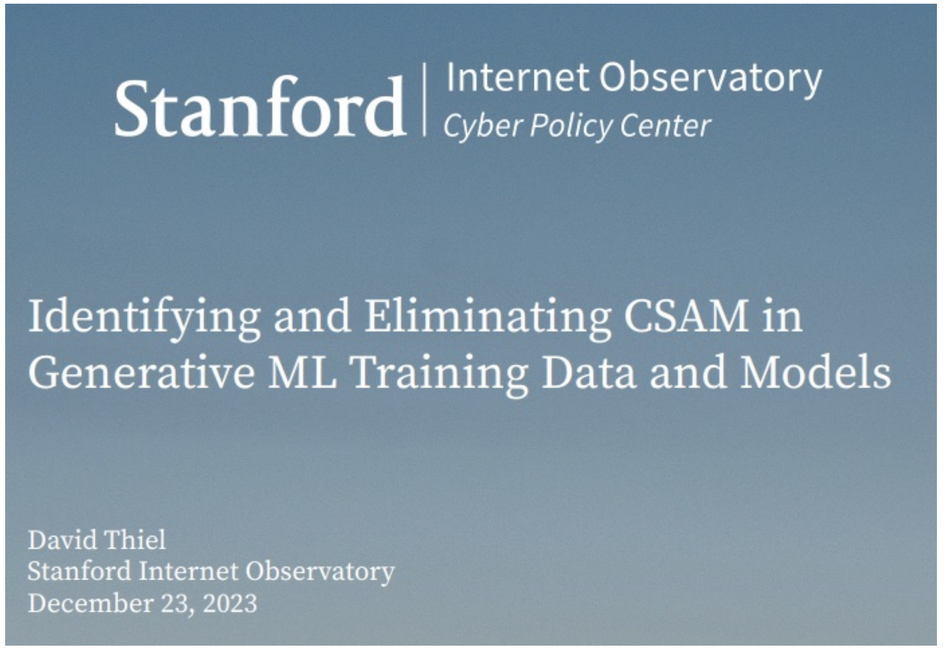
- But also,
  What if we were to remove all of the CSAM?
- Images of children and adult content would still remain



Memorization & Generalization

- Would a single (CSAM) instance be reproduced?
- Would two?
- Would three?
- …

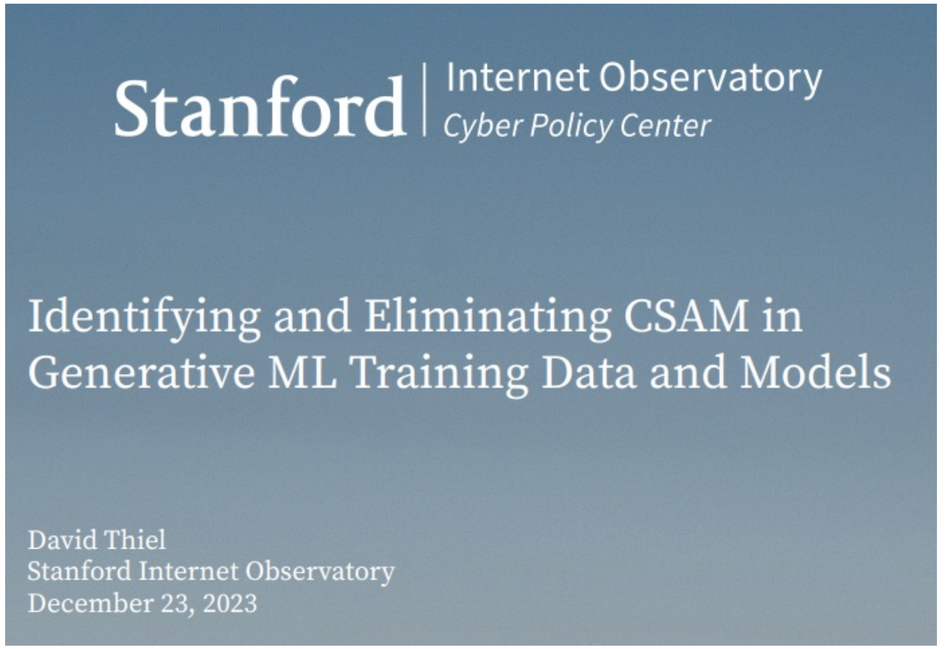
No! We need a few hundreds!



- But also,
  What if we were to remove all of the CSAM?
- Images of children and adult content would still remain

Memorization & Generalization

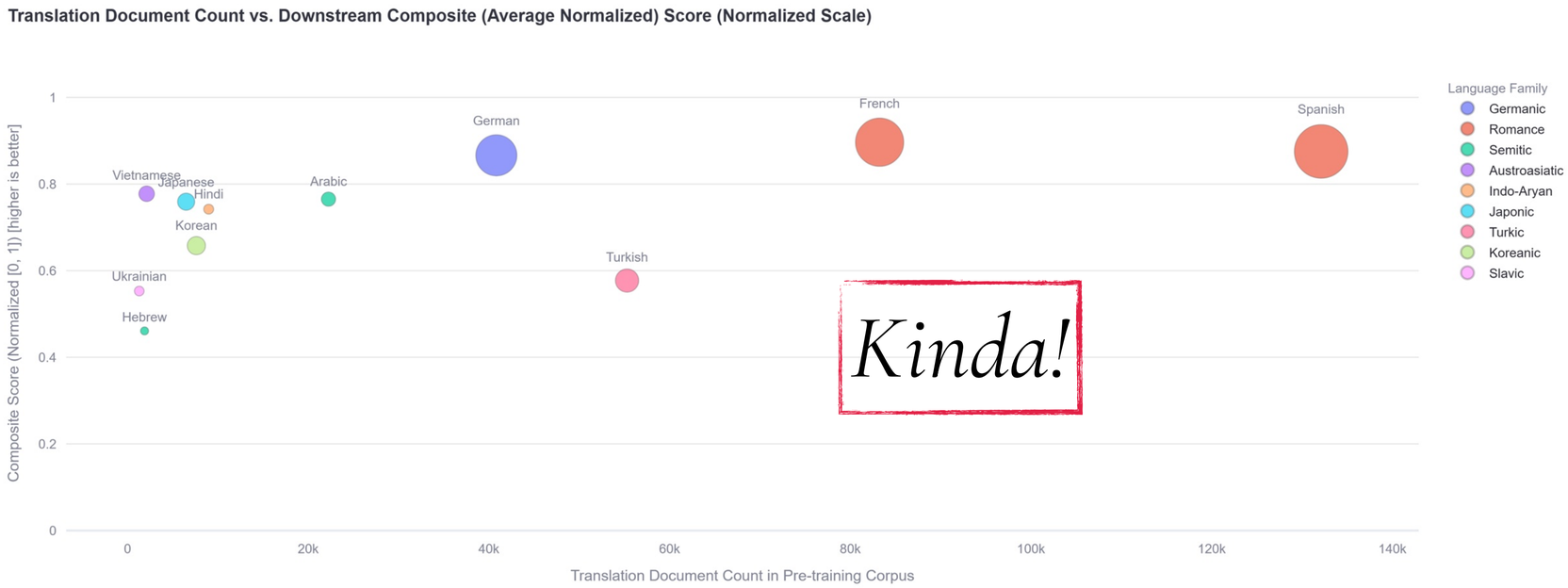
Can we categorize the composition of concepts in the training data?



- But also,
  What if we were to remove all of the CSAM?
- Images of children and adult content would still remain

Performance Predictability Theory

Can we predict the performance of multi-lingual tasks based on the language composition in documents?



Kinda!

Data Inference based on Model Weights

Can we recreate the training data domain composition based on model weights?

Kinda!

unlearn_method	predictor	Train order_score ↑	Train mse ↓	Train spearman_r ↑	Dev order_score ↑	Dev mse ↓	Dev spearman_r ↑	llama order_score ↑	llama mse ↓	llama spearman_r ↑
ga_utility	random	0.499	0.0602	-0.003	0.5	0.0606	0.001	0.568	0.0635	-0.006
ga_utility	lr_raw_self	0.747	0.0362	0.609	0.771	0.0358	0.648	0.571	0.0446	0.111
ga_utility	lr_diff_self	0.787	0.0348	0.704	0.807	0.0341	0.752	0.619	0.0413	0.074
ga_utility	lr_ratio_self	0.781	0.0363	0.692	0.798	0.0355	0.711	**0.667**	0.0426	0.259
ga_utility	lr_curve_self	0.768	0.0333	0.663	0.745	0.0333	0.632	**0.667**	0.0575	**0.371**
ga_utility	lr_full_self	0.715	0.0221	0.597	0.717	0.0242	0.613	0.619	0.0507	0.148
ga_utility	lr_curve_agg	0.787	0.0293	0.672	0.771	0.0309	0.639	0.571	0.0446	0
ga_utility	lr_diff_agg	**0.819**	0.0264	**0.757**	0.824	0.0267	**0.771**	0.619	0.0335	0.074
ga_utility	lr_full_agg	0.778	**0.0209**	0.691	0.781	**0.023**	0.72	0.619	**0.0249**	0.148
ga_utility	lr_ratio_agg	0.801	0.0342	0.715	**0.831**	0.0345	0.764	0.571	0.0561	-0.037
ga_utility	lr_raw_agg	0.807	0.0323	0.735	0.814	0.0334	0.741	0.476	0.0446	-0.371
ga_utility	mlp_curve_self	0.77	0.0318	0.663	0.769	0.0323	0.646	0.571	0.0989	0.111
ga_utility	mlp_full_self	0.748	0.0257	0.645	0.74	0.028	0.639	**0.667**	0.1097	0.259
ga_utility	mlp_curve_agg	0.776	0.0228	0.685	0.79	0.0244	0.696	**0.667**	0.053	0.296
ga_utility	mlp_full_agg	0.719	0.0233	0.574	0.726	0.0254	0.6	0.476	0.0806	-0.259

Data Interpretability

- Why models learn or don't learn?

The data! (mostly)

Data Interpretability

- Why models learn or don't learn? *The data! (mostly)*
- Given a model's behavior, can we explain it?? *The data! (more often than not)*

Data Interpretability

- Why models learn or don't learn? *The data! (mostly)*
- Given a model's behavior, can we explain it?? *The data! (more often than not)*
- Can we develop a general theory??? *Not yet! (tbd)*

Data Interpretability @DICE Lab

yanaiela@gmail.com

@yanaiela