Memorizing Distributions:

Generative Models Recall More Than Verbatim Instances

Yanai Elazar, L2M2, August 1st, 2025

Memorization

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Counterfactual Memorization in Neural Language Models

Chiyuan Zhang Google Research chiyuan@google.com
Daphne Ippolito Carnegie Mellon University daphnei@cmu.edu
Katherine Lee Google DeepMind katherinelee@google.com
Matthew Jagielski Google DeepMind jagielski@google.com
Florian Tramèr ETH Zürich florian.tramer@inf.ethz.ch
Nicholas Carlini Google DeepMind ncarlini@google.com

QUANTIFYING MEMORIZATION ACROSS NEURAL LANGUAGE MODELS

Nicholas Carlini[*] **Daphne Ippolito**[1,2] **Matthew Jagielski**[1]
Katherine Lee[1,3] **Florian Tramèr**[1] **Chiyuan Zhang**[1]

[1]*Google Research*
[2]*University of Pennsylvania*
[3]*Cornell University*

Extracting Training Data from Diffusion Models

Nicholas Carlini[*1] *Jamie Hayes*[*2] *Milad Nasr*[*1]
Matthew Jagielski[+1] *Vikash Sehwag*[+4] *Florian Tramèr*[+3]
Borja Balle[†2] *Daphne Ippolito*[†1] *Eric Wallace*[†5]
[1]Google [2]DeepMind [3]ETHZ [4]Princeton [5]UC Berkeley
[*]Equal contribution [+]Equal contribution [†]Equal contribution

Memorization

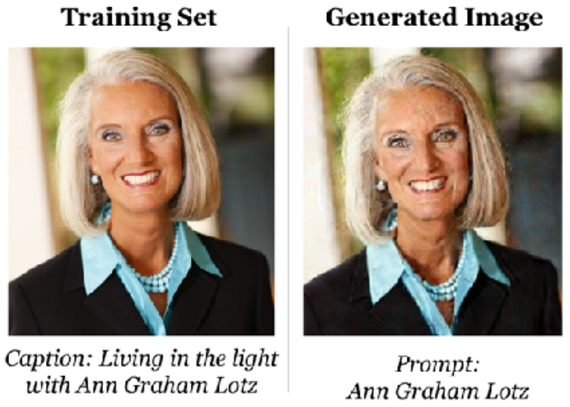
- Memorization has been heavily studied
- Focusing on verbatim memorization

Definition 3.1. A string s is *extractable with k tokens of context* from a model f if there exists a (length-k) string p, such that the concatenation $[p \,||\, s]$ is contained in the training data for f, and f produces s when prompted with p using greedy decoding.

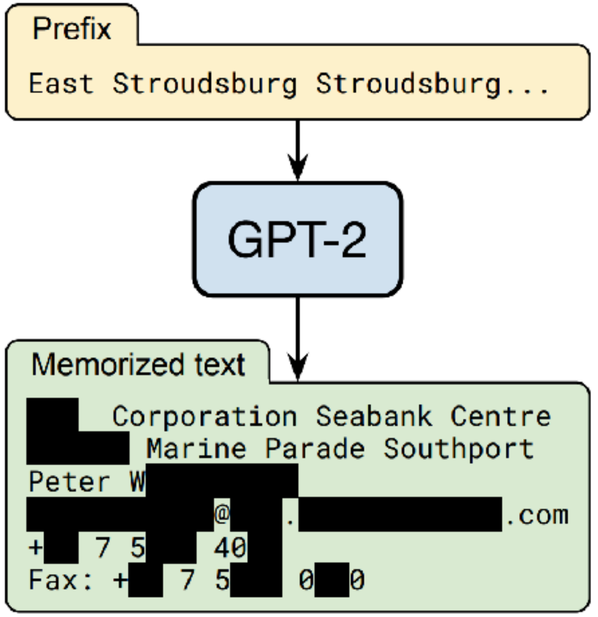
Carlini et al., 2023

Verbatim Memorization

- Memorization has been heavily studied
- Focusing on verbatim memorization



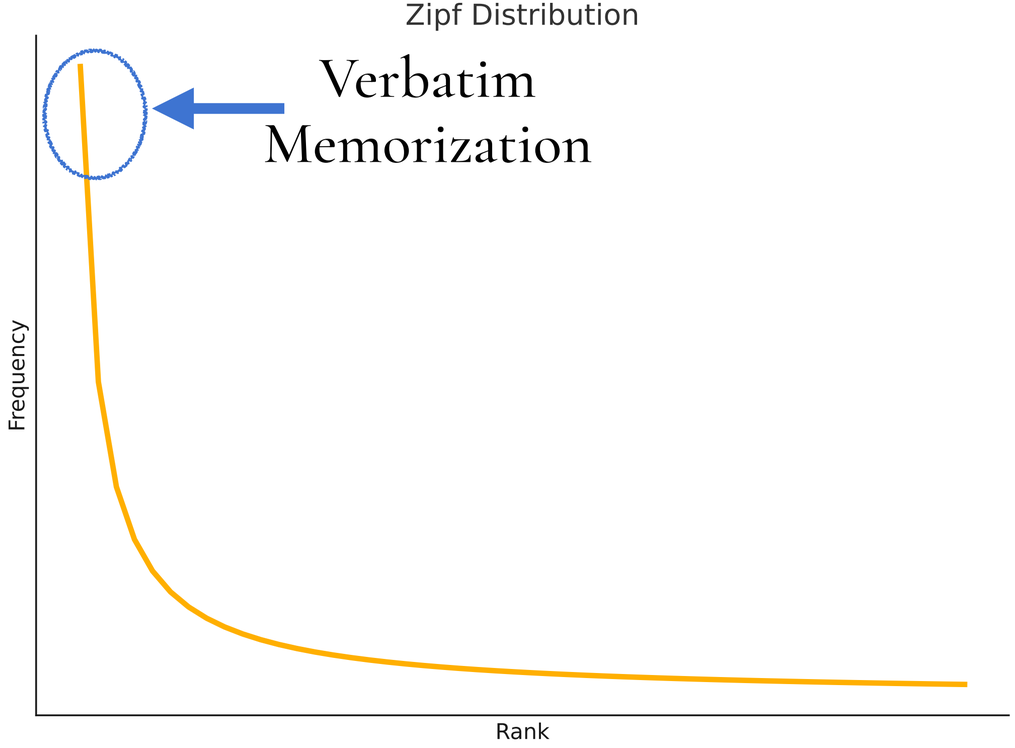
Carlini, Hayes & Nasr et al., 2023



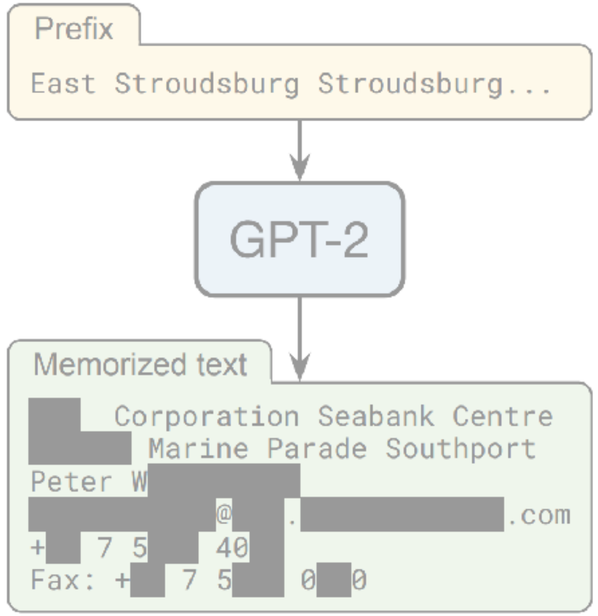
Carlini et al., 2021

Verbatim Memorization

Estimated Memorization (empirical lower) Bounds: 0.1%, 1%



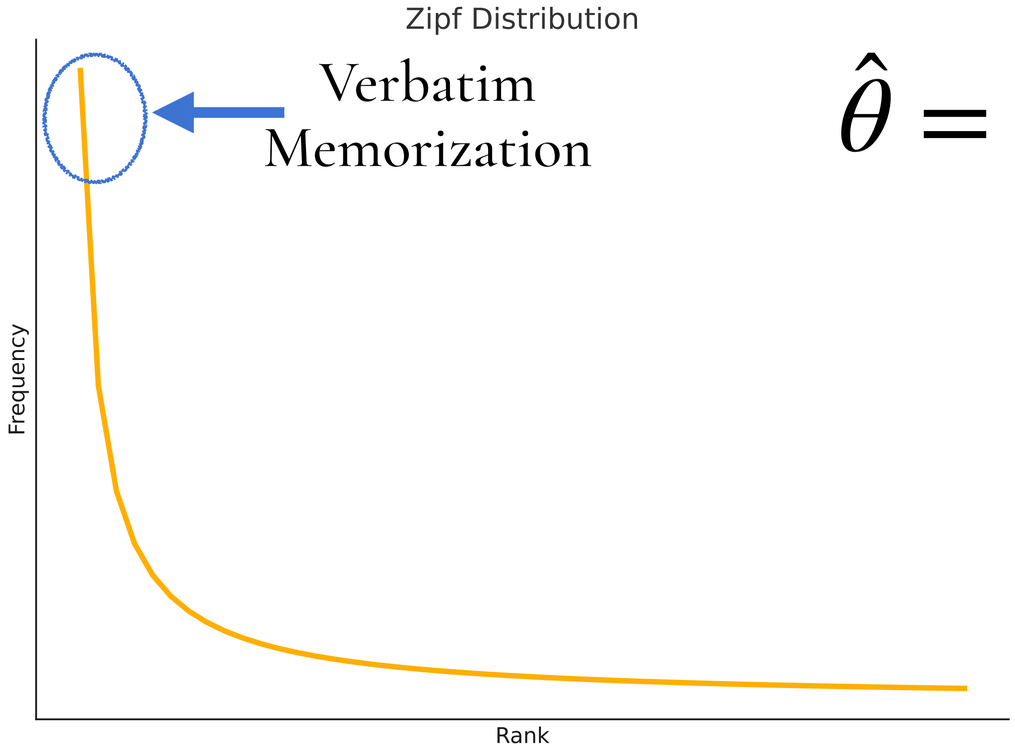
Carlini, Hayes & Nasr et al., 2023



Carlini et al., 2021

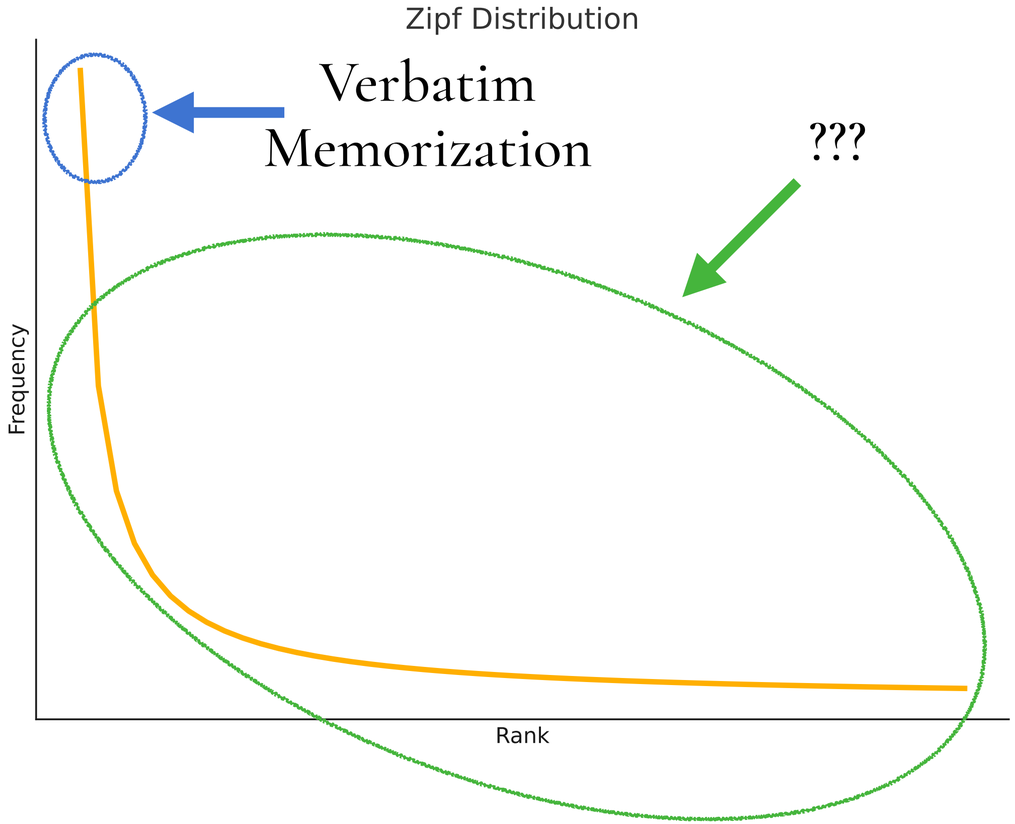
Verbatim Memorization?

Estimated Memorization (empirical lower) Bounds: 0.1%, 1%



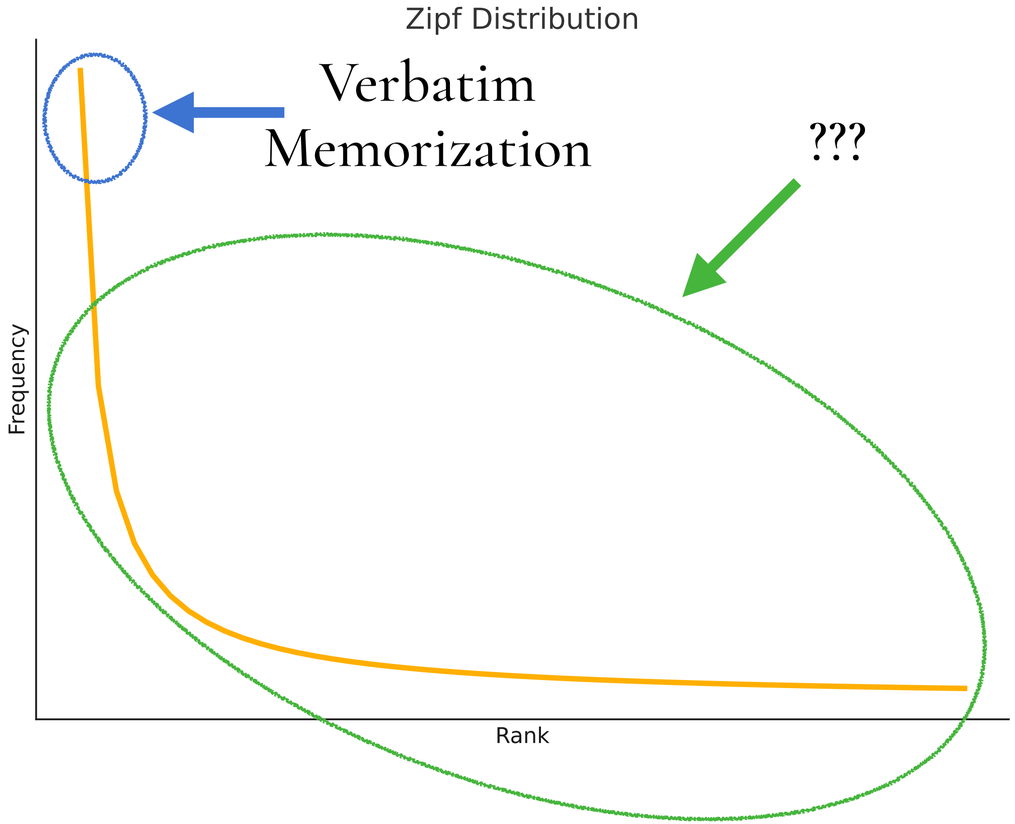
$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Verbatim Memorization?



Verbatim Memorization?

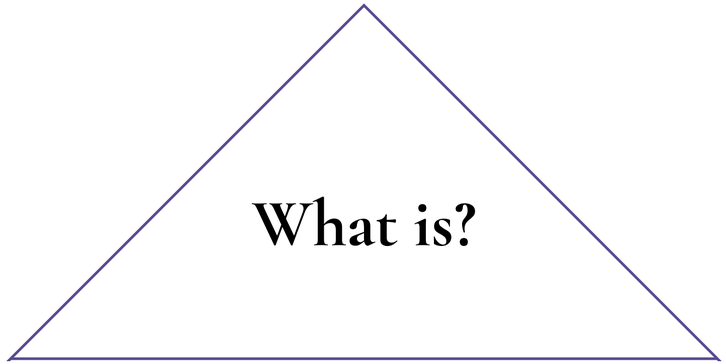
What is learned from the rest of the data (99%)?



Verbatim Memorization?

What is learned from the rest of the data (99%)?

This Talk: Distributional Memorization



This Talk: Distributional Memorization

What is?

How much
it happens?

This Talk: Distributional Memorization

What is?

How much it happens?

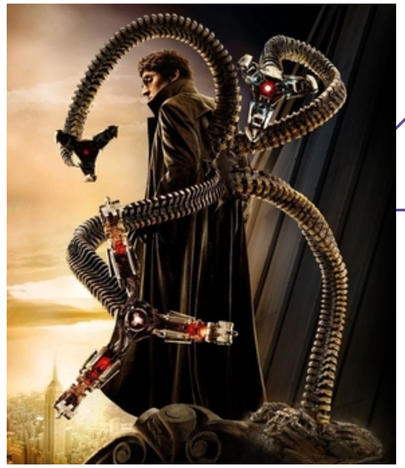
When does it happen?

This Talk: Distributional Memorization

What is?

How much it happens?

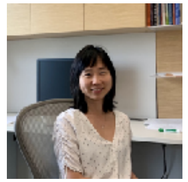
When does it happen?



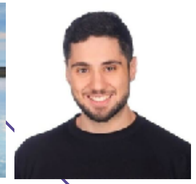
Tools

This Talk: Distributional Memorization

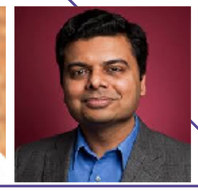

What is?


How much it happens?

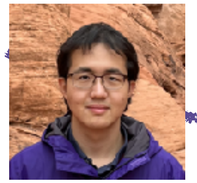







When does it happen?





Tools et el.

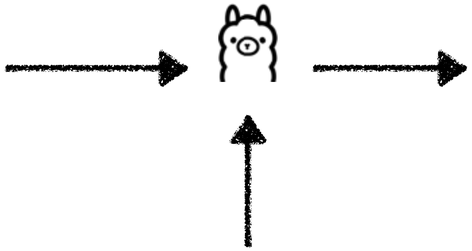
What Is Distributional Memorizing

Memorization Types

Verbatim Memorization

Training Data

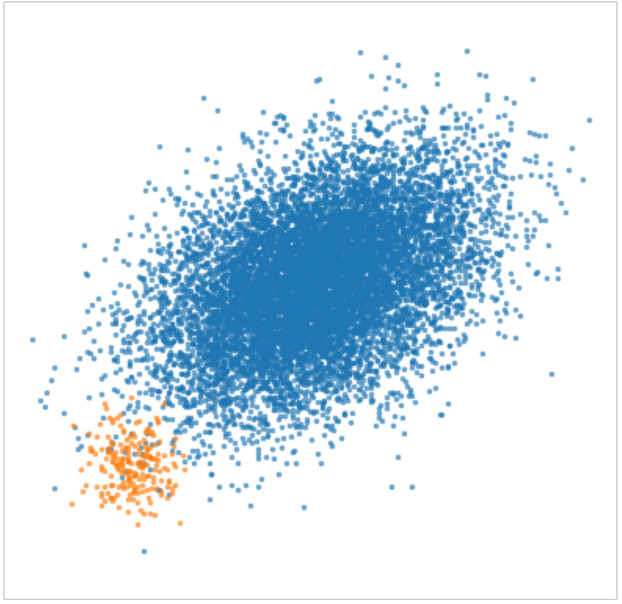
Generation

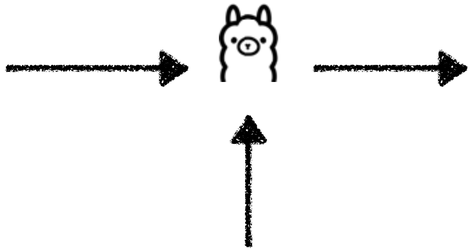


An Image of Leonardo DiCaprio

Prompt

Distribution Memorization





An Image of Leonardo DiCaprio

Prompt

Memorization Types

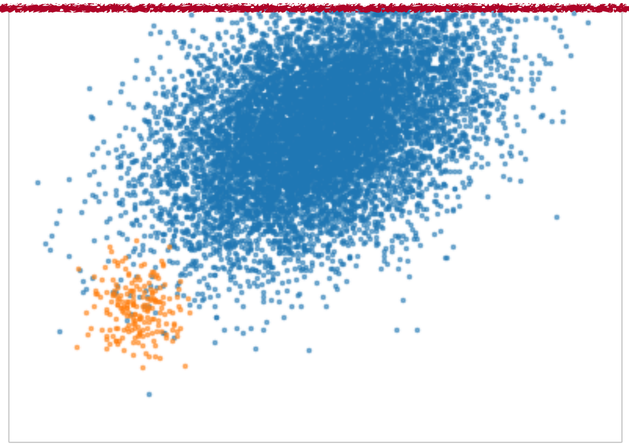
Verbatim Memorization

Training Data

Generation

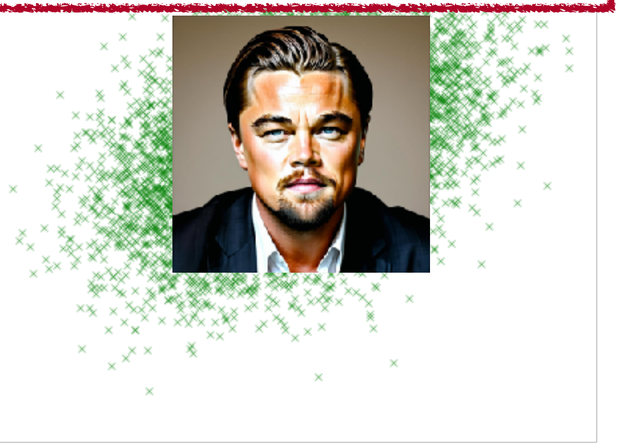
Definition: A model *M* memorize a distribution *D* from the training data if *M* can generate data *D'* that is indistinguishable from *D*

Memorization



An Image of Leonardo DiCaprio

Prompt



The Bias Amplification Paradox in Text-to-Image Generation

Preethi Seshadri, Sameer Singh, Yanai Elazar

NAACL 2024

Models are Biased

- Models encode and exhibit different biases
- This is not a new finding!

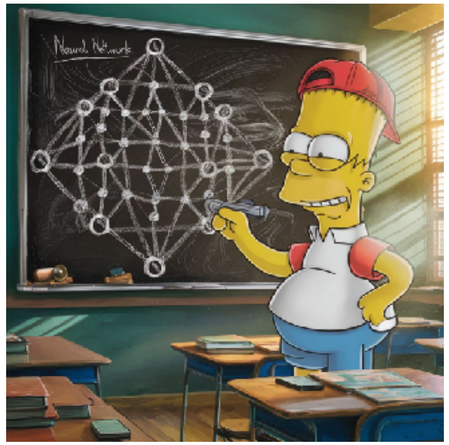
Let's Try It Out!

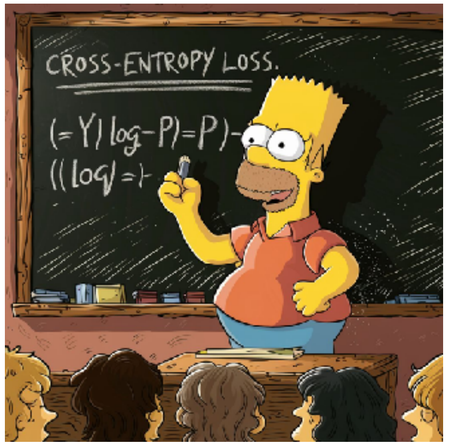
"A photo of a face of an engineer"

1/10 women!

The model is biased!

Where Does The Bias Come From?





Let’s Look At The Data

Where Does The Bias Come From?

5 billion image-caption pairs!

Where Does The Bias Come From?

- Using an index (WIMBD), we have fast access to the data
- … and we can test such associations in the training data

Where Does The Bias Come From?

```
from wimbd.es import get_documents_containing_phrases

# Get documents containing the term:
get_documents_containing_phrases("laion","engineer")
```

ENGI
Engin
Mater

The data is large and noisy, so we need to adjust

Establishing Data Gender Ratios

```python
from wimbd.es import get_documents_containing_phrases

# Get documents containing the term:
get_documents_containing_phrases("laion","engineer")
```

We follow a similar process for the generated images


Filtering

Gender identification

👩




👨

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer
  - Janitor

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer
  - Janitor
  - Lawyer

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer
  - Janitor
  - Lawyer
  - ...

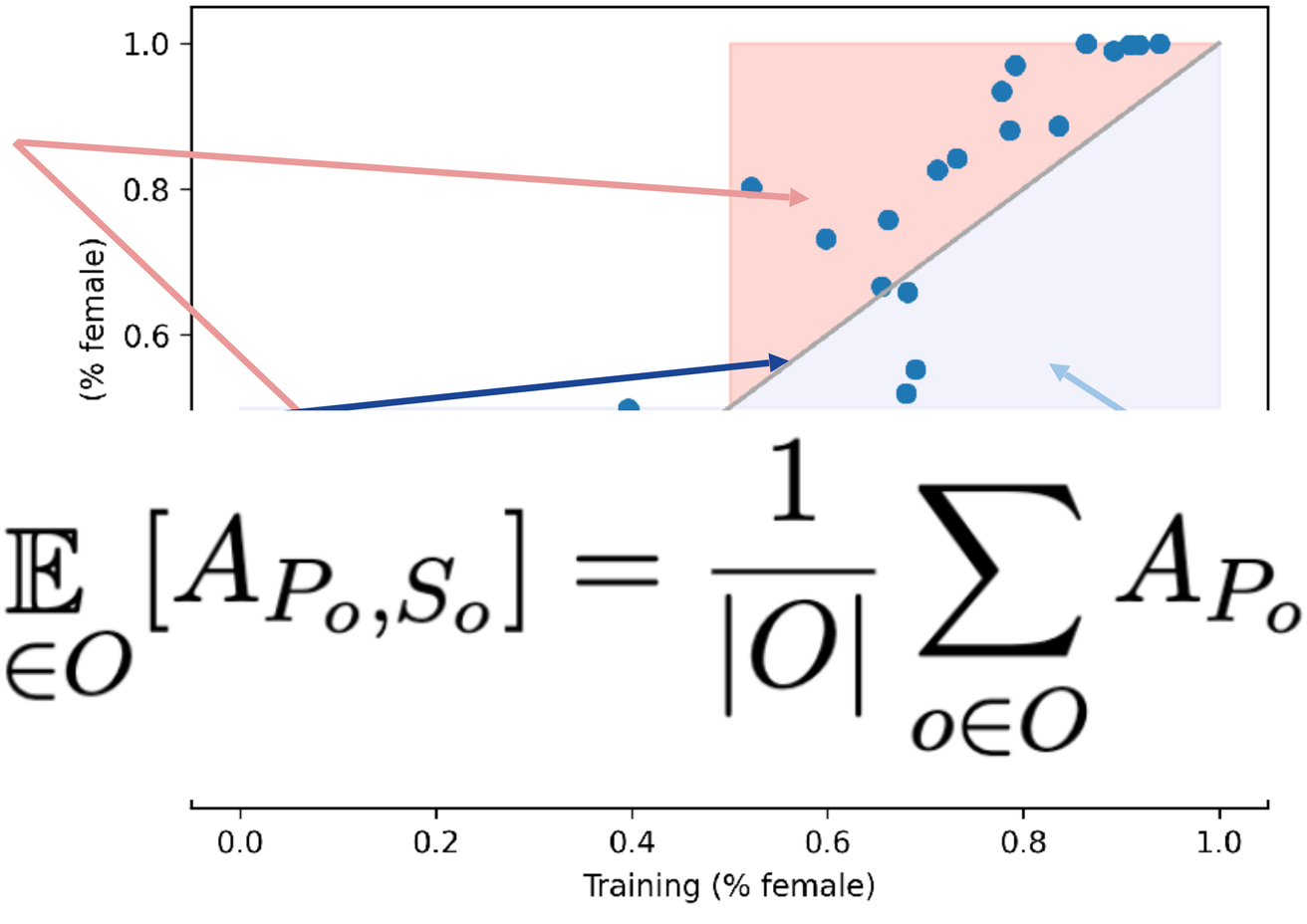
Bias Amplification?

Given the calculated ratios from the data, we can now compare the model's generation to the training data

Bias Amplification?

Peach area:
Bias Amplification

Diagonal:
Bias preservatio



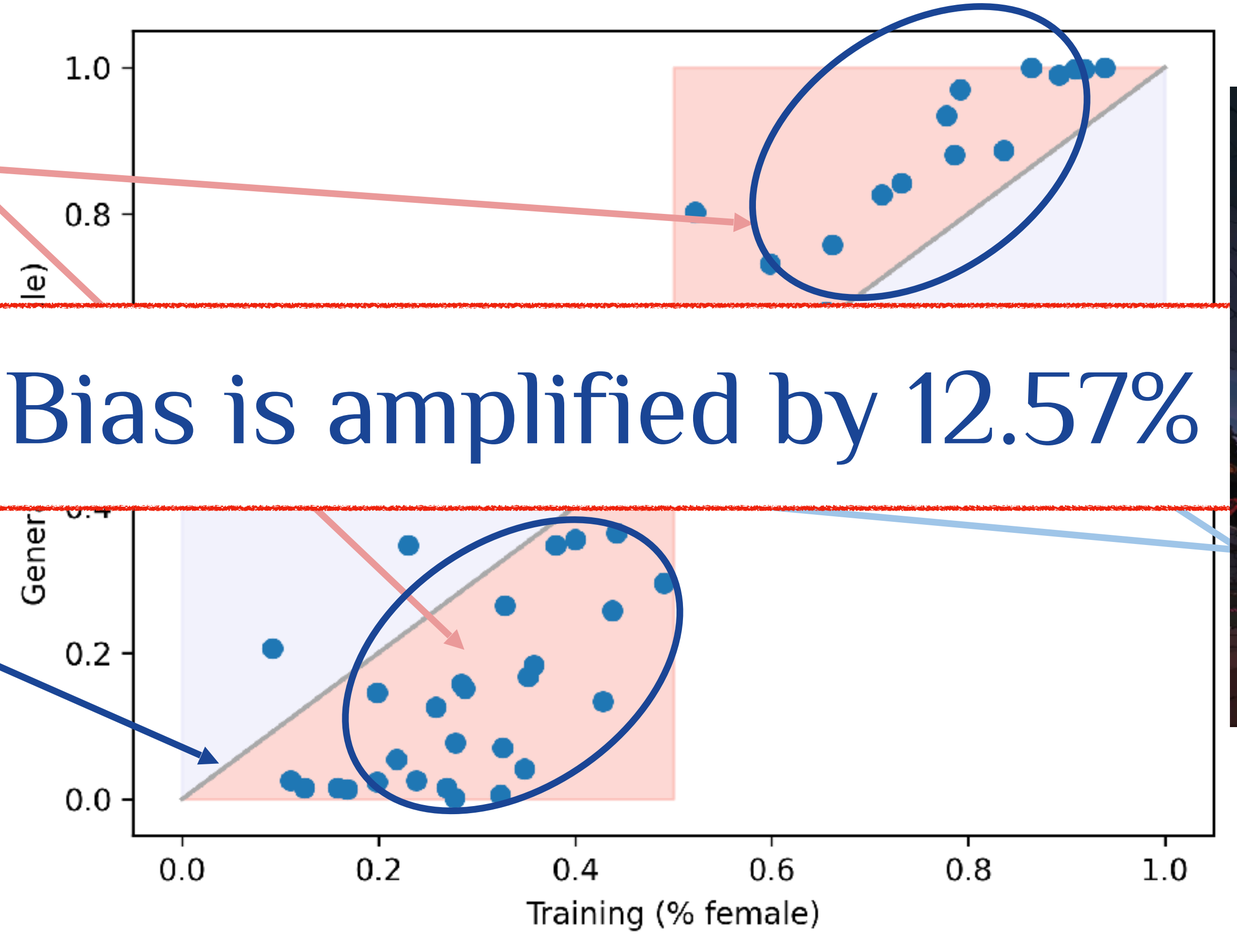
der area:
e-amplification

$$\mathop{\mathbb{E}}_{o \in O}[A_{P_o,S_o}] = \frac{1}{|O|} \sum_{o \in O} A_{P_o,S_o}$$

Bias Amplification!

Peach area:
Bias Amplification

Diagonal:
Bias preservatio



Bias Amplification!

Shown in previous work

Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints

Jieyu Zhao[§] **Tianlu Wang**[§] **Mark Yatskar**[‡]
Vicente Ordonez[§] **Kai-Wei Chang**[§]
[§]University of Virginia
{jz4fu, tw8cb, vicente, kc2wc}@virginia.edu
[‡]University of Washington
my89@cs.washington.edu

Zhao et al, 2017, best paper

The Bias Amplification Paradox

But wait!

Why would a model amplify the biases from the training data?

Let's look at the training data again

Training Data Investigation

👩

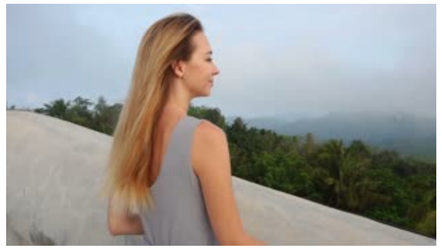
Portrait of young ***woman*** **programmer** working at a computer in the data center filled with display screens

👨

shutterstock · 669546292

programmer configures the... | Shutterstock . vector #669546292

Slow motion **programmer** ***female*** relaxing among nature, young ***woman*** on long-awaited vacation abroad after working year...

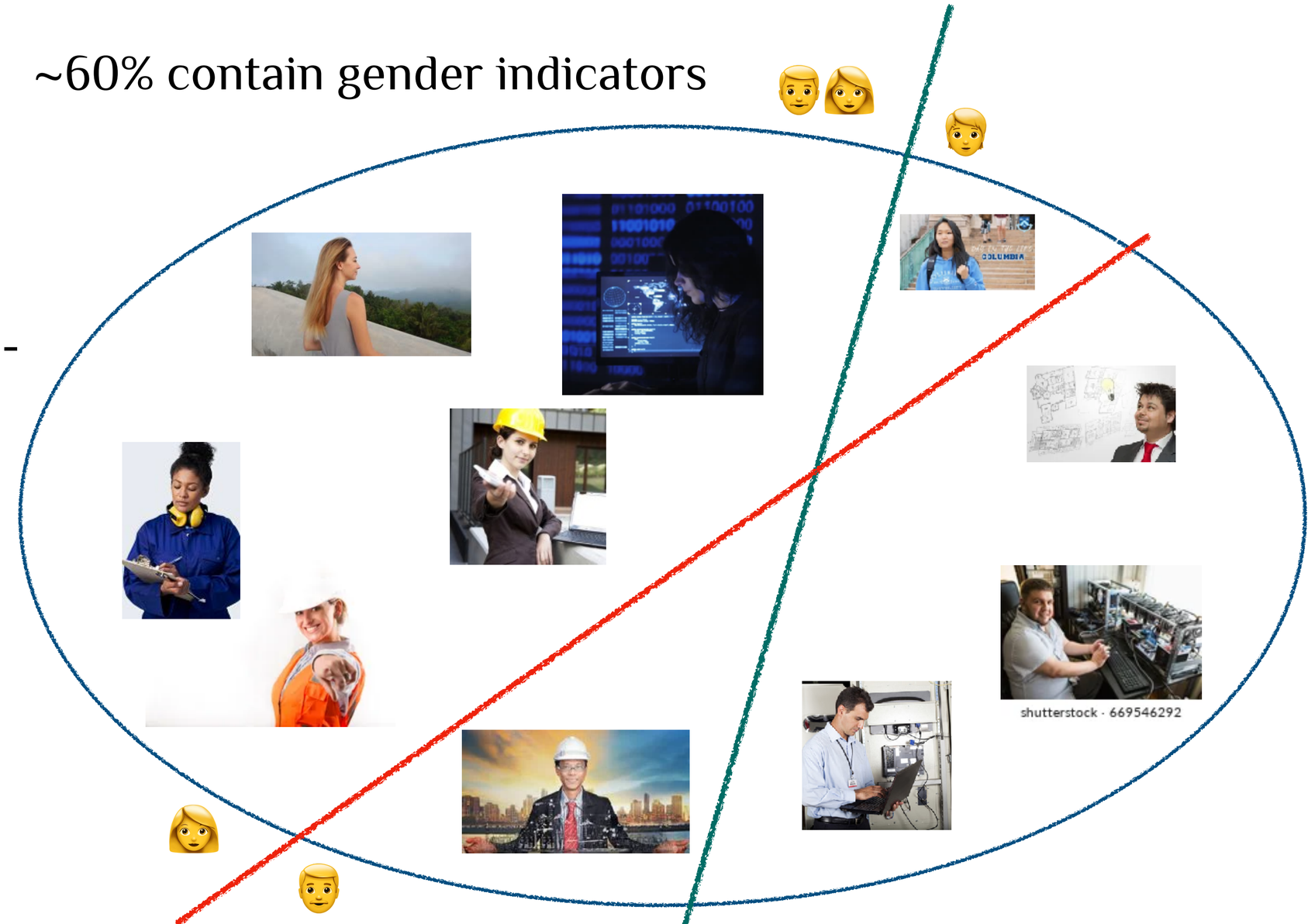


industrial programmer checking computerized machine status

Training Data Investigation

~60% contain gender indicators

Mostly with anti-stereotypical gender (70%)



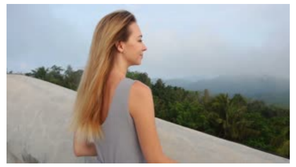
Training Data Investigation

~60% contain gender indicators

Mostly with anti-stereotypical gender (70%)

Test data

“A photo of a face of an engineer”



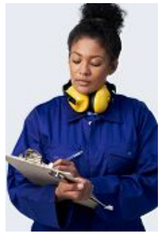
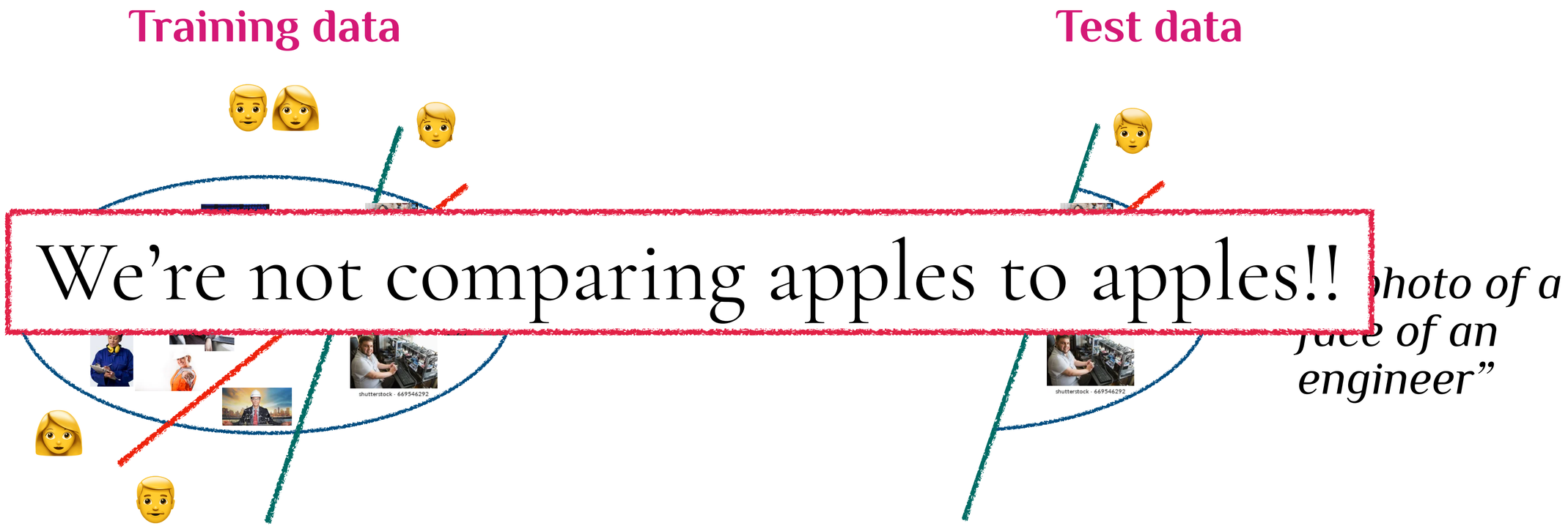





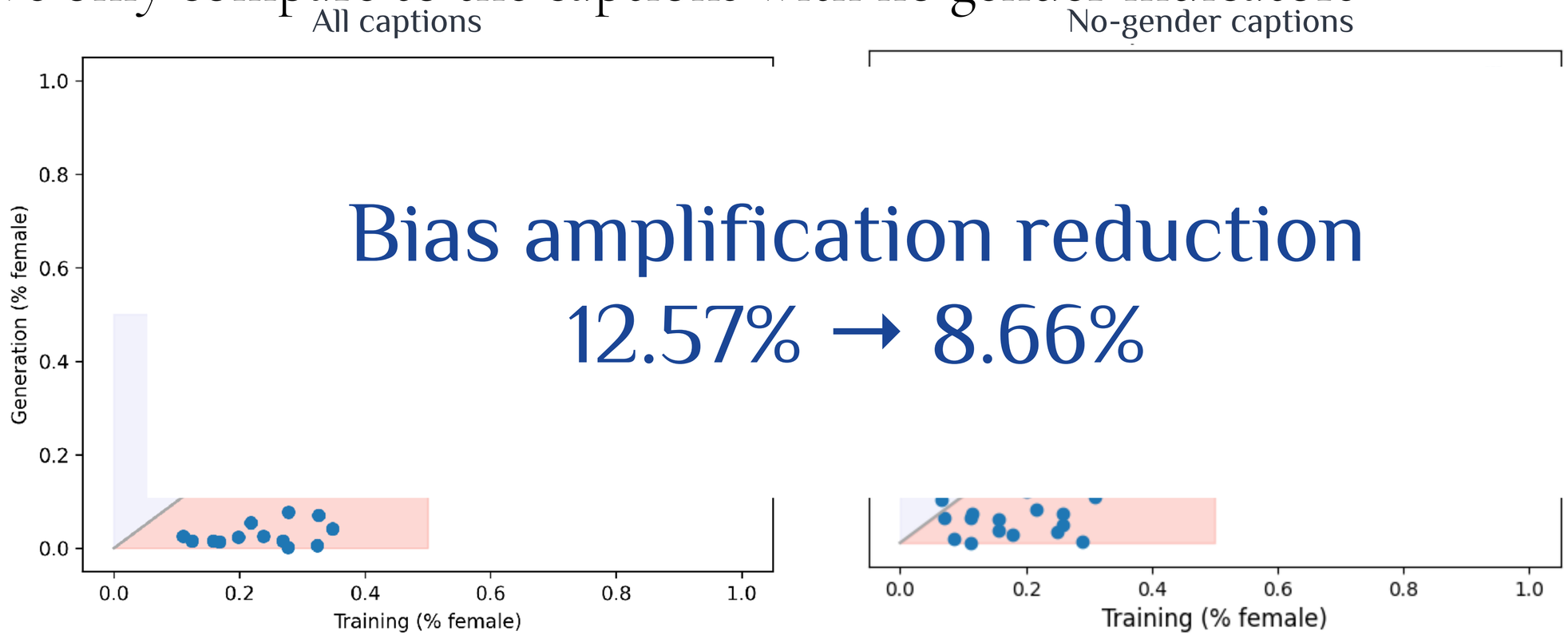
Image Captions & Prompts Mismatch



Matching Distributions

Instead of comparing the generated images to the entire training set:

- We only compare to the captions with no gender indicators



One Mismatch

What about others?

Image Captions & Prompts Mismatch #2

We also found a “

President

Training

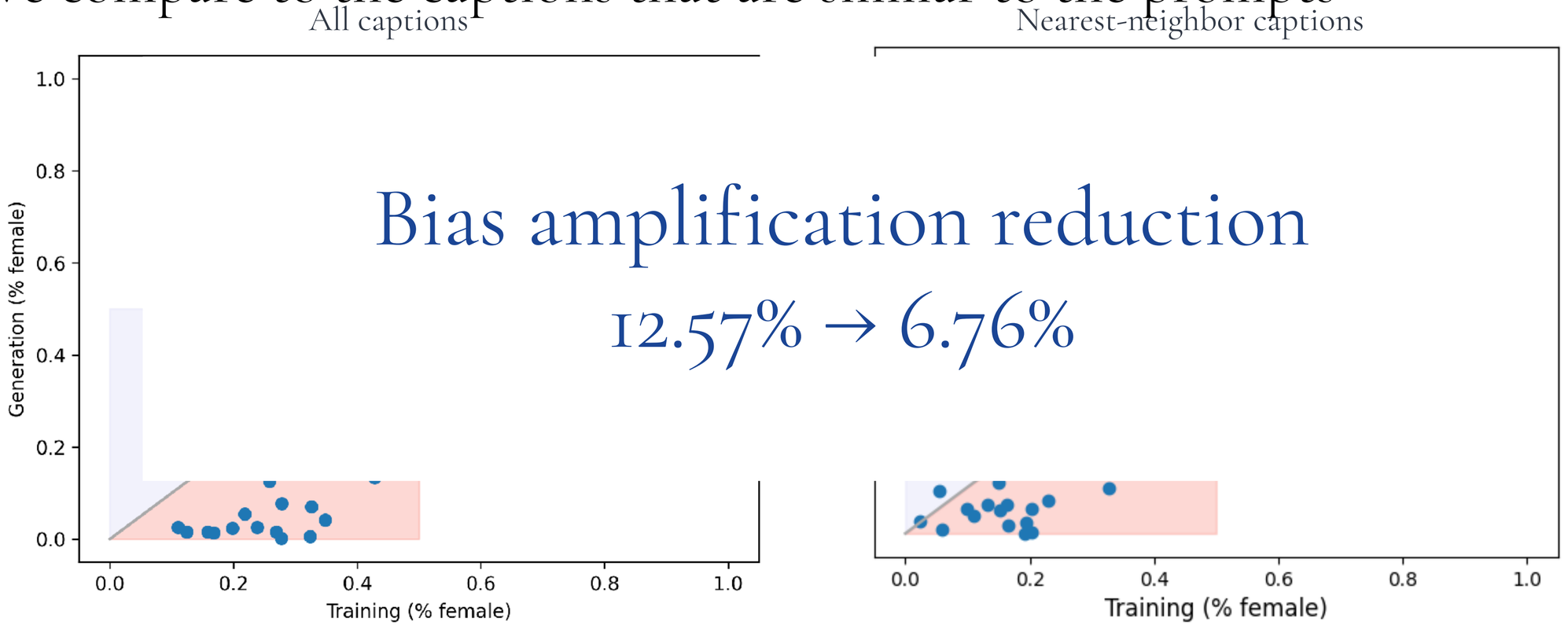
Generation

(a) Training captions for **President**: 1) "Leana Wen, Planned Parenthood president..." 2) "New Schaumburg Business Association President..." 3) "BCCI president N Srinivasan..." 4) "Indiana Pacers president of basketball operations..."

Matching Distributions #2

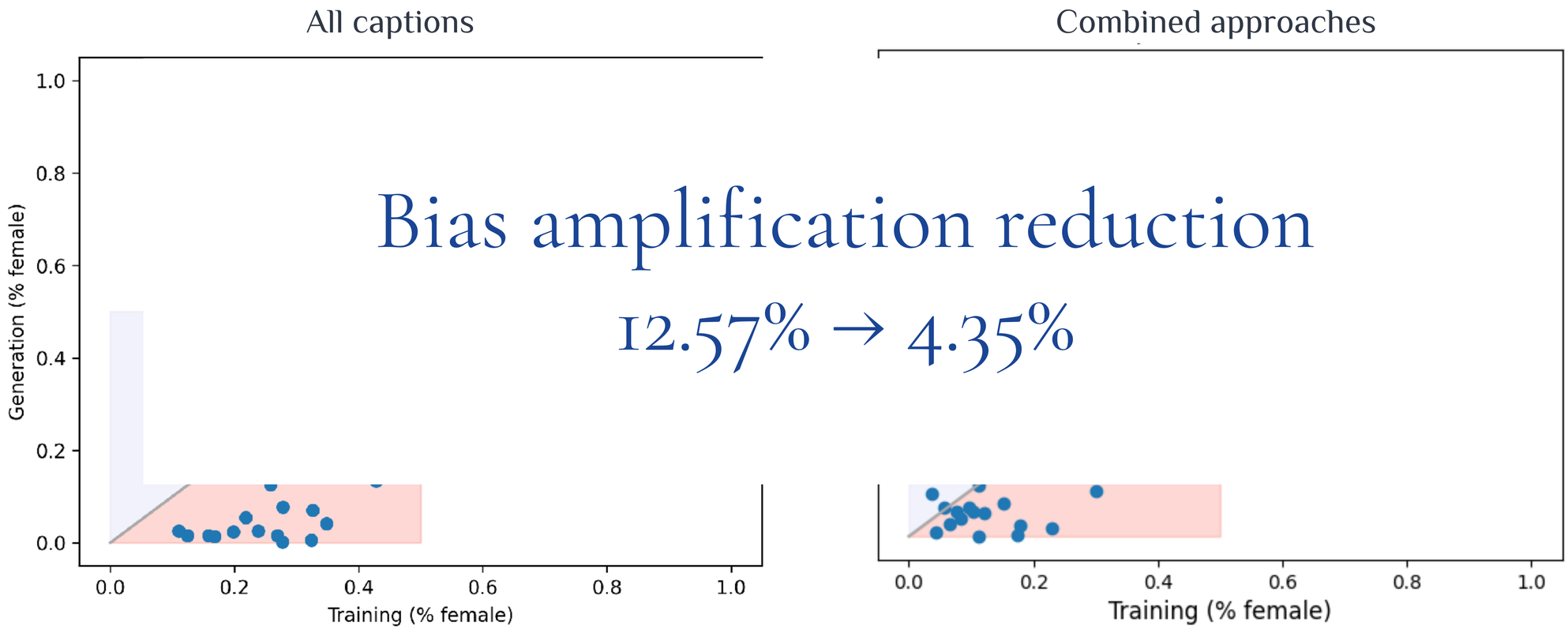
Instead of comparing the generated images to the entire training set:

- We compare to the captions that are similar to the prompts



Matching Distributions: Combined

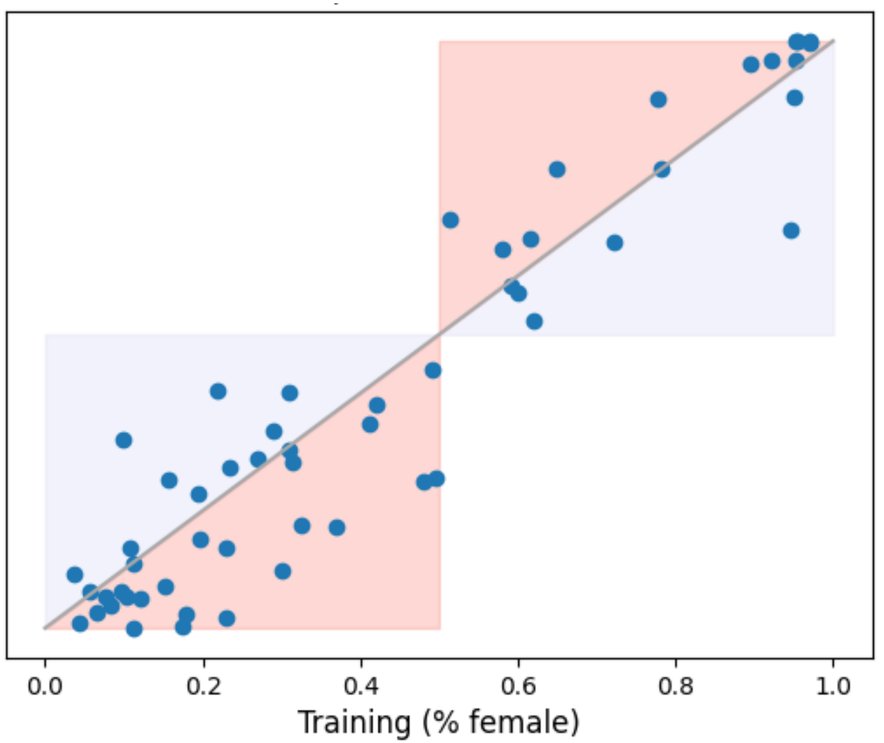
Finally, we combine both approaches



Revisiting the Bias Amplification Claim

While we still observe bias amplification:

- It is significantly reduced
- There may be more confounders
- This problem is more nuanced and involved than originally thought



What is Distributional Memorization?

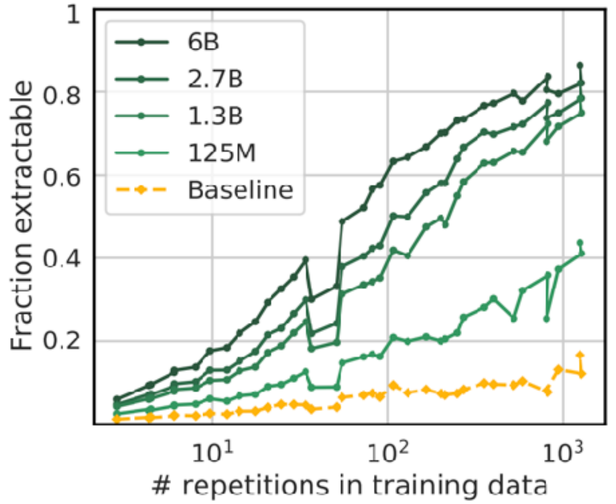
- Models memorize gender ratios from training data
- Evaluation becomes tricky
- We should think about the right abstractions

How Much Distributional Memorization Happen?

Memorized Sequences

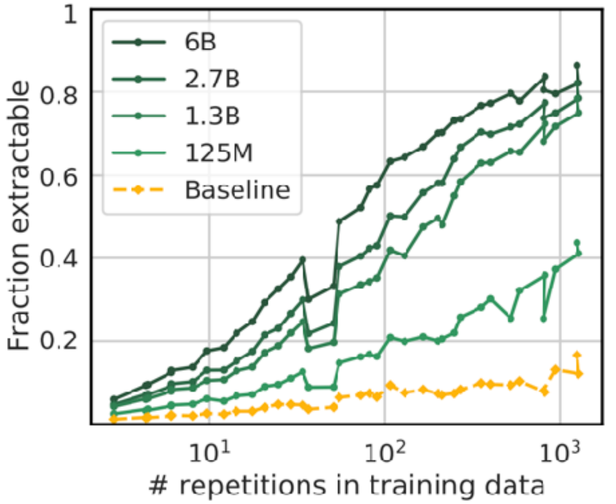
Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*



Memorized Sequences

- This is a strict metric
- Can we relax it?

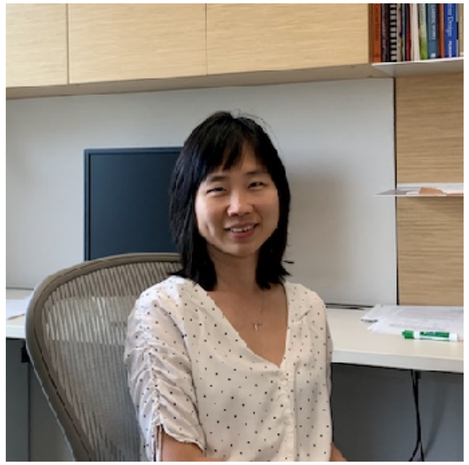


Estimated Memorization (empirical lower) Bounds: 0.1%, 1%

Detection and Measurement of Syntactic Templates in Generated Text

Chantal Shaib, Yanai Elazar, Junyi Jesse Li, Byron C. Wallace

EMNLP 2024



Memorizing Templates

- Instead of looking at raw sequences

The Last Black Man in San Francisco is a poignant, beautifully shot film [...] creates a unique and intense viewing experience.

Memorizing Templates

- Instead of looking at raw sequences
- We inspect their a linguistic abstraction

The Last Black Man in San Francisco is a poignant, beautifully shot film [...] creates a unique and intense viewing experience.

DET ADJ PROPN PROPN ADP PROPN PROPN AUX DET NOUN PUNCT ADV VERB NOUN [...] VERB DET ADJ CCONJ ADJ VERB NOUN

Memorizing Templates

- Instead of looking at raw sequences
- We inspect their a linguistic abstraction

Review the movie “The Last Black Man in San Francisco”

The Last Black Man in San Francisco is a poignant, beautifully shot film [...] creates a unique and intense viewing experience.

DET ADJ PROPN PROPN ADP PROPN PROPN AUX DET NOUN PUNCT ADV VERB NOUN [...] VERB DET ADJ CCONJ ADJ VERB NOUN

Novel sequence

Frequenct POS sequence

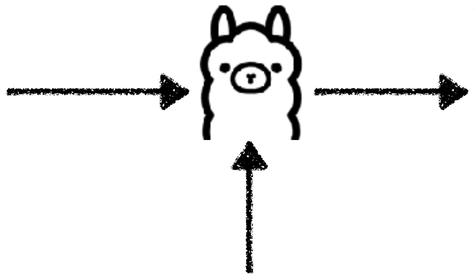
20% of generations

Memorizing Templates (like HMMs)

Training Data

DET ADJ PROPN PROPN ADP PROPN PROPN
AUX DET NOUN PUNCT ADV VERB NOUN [...]
VERB DET ADJ CCONJ ADJ VERB NOUN

Memorizing
Templates



Lorem ipsum dolor sit amet

Prompt

Generation

Memorized
Templates

DET ADJ PROPN PROPN ADP PROPN PROPN
AUX DET NOUN PUNCT ADV VERB NOUN [...]
VERB DET ADJ CCONJ ADJ VERB NOUN

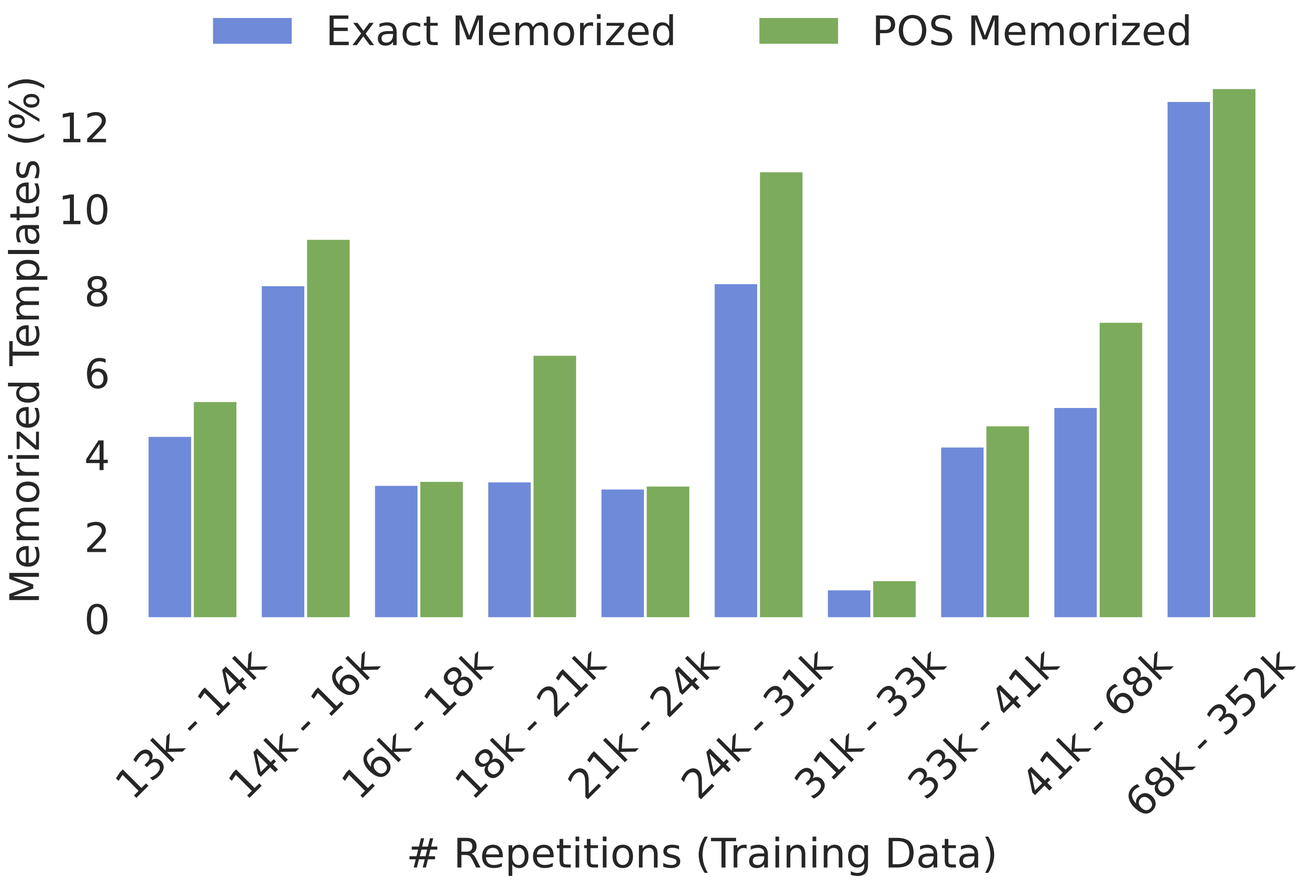
The Last Black Man in San Francisco is a poignant, beautifully shot film [...] creates a unique and intense viewing experience.

Sampling from
the distribution

Results


Exact Memorized
POS Memorized
Memorized Templates (%)
Repetitions (Training Data)

Results


Exact Memorized
POS Memorized
Memorized Templates (%)
0
2
4
6
8
10
12
13k - 14k
14k - 16k
16k - 18k
18k - 21k
21k - 24k
24k - 31k
31k - 33k
33k - 41k
41k - 68k
68k - 352k
Repetitions (Training Data)

Results

6.4% vs. 5.3%
(on average)

Summary

- Generated texts is largely based on “templates”…
- that originate from the training data
- In turn, it allows us to measure memorized templates

How Much it Happens?

- Memorization goes beyond exact match
- It can be seen as template memorization
  that the tokens are sampled from its distribution
  (like HMMs)
- We should extend our definitions of
  memorization beyond exact match

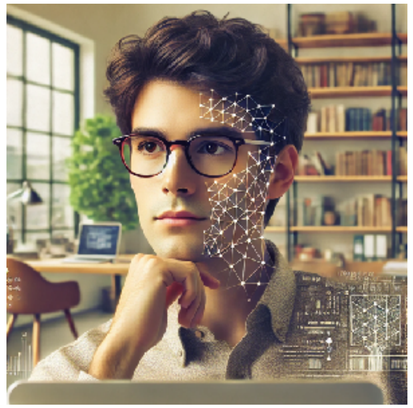
When Does Distributional Memorization Happen?

Imitation

✅

Leonardo DiCaprio

❌



Yanai Elazar

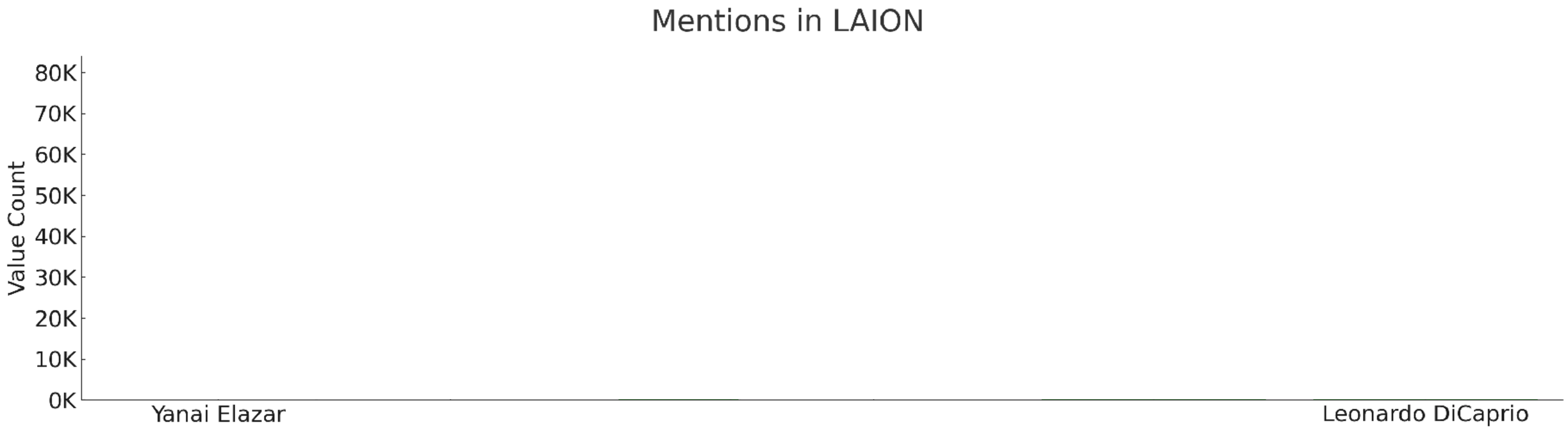
Imitation

Spot the difference

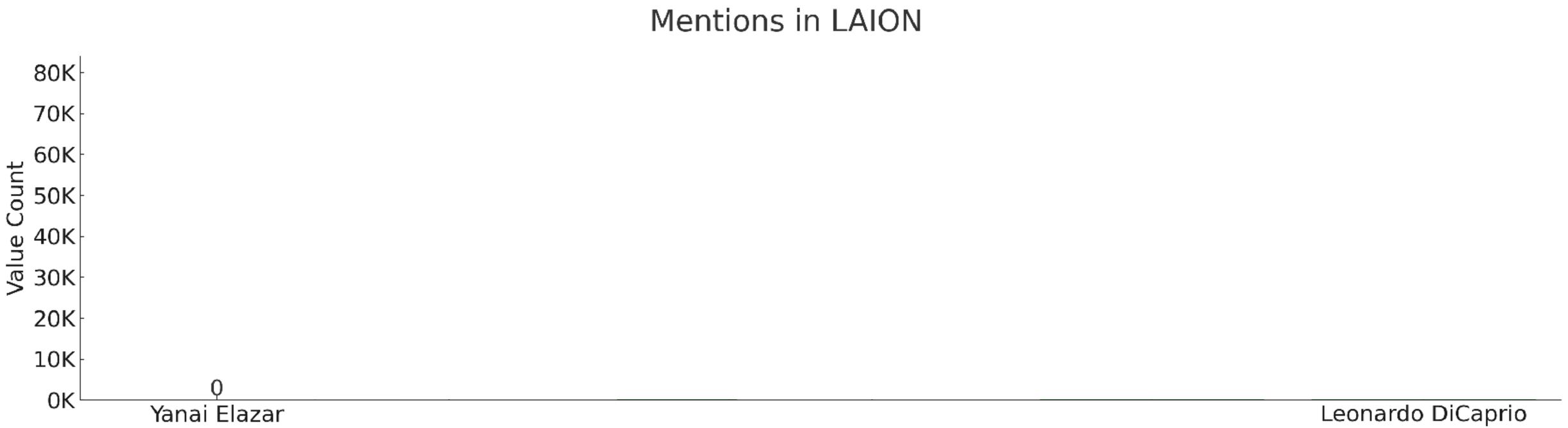
Leonardo DiCaprio

Yanai Elazar

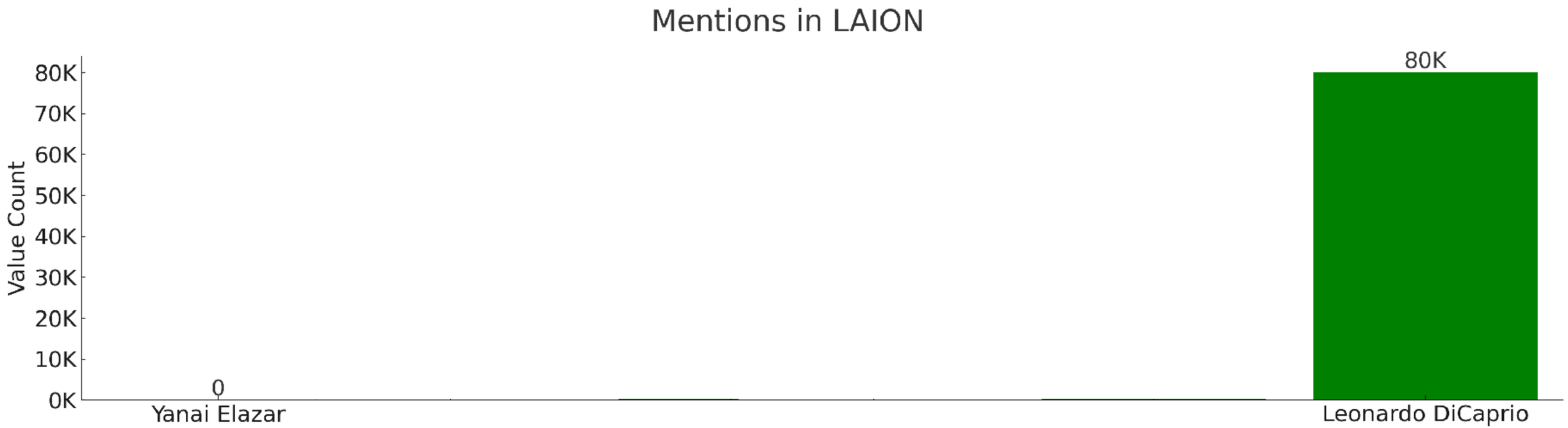
Imitation Threshold?



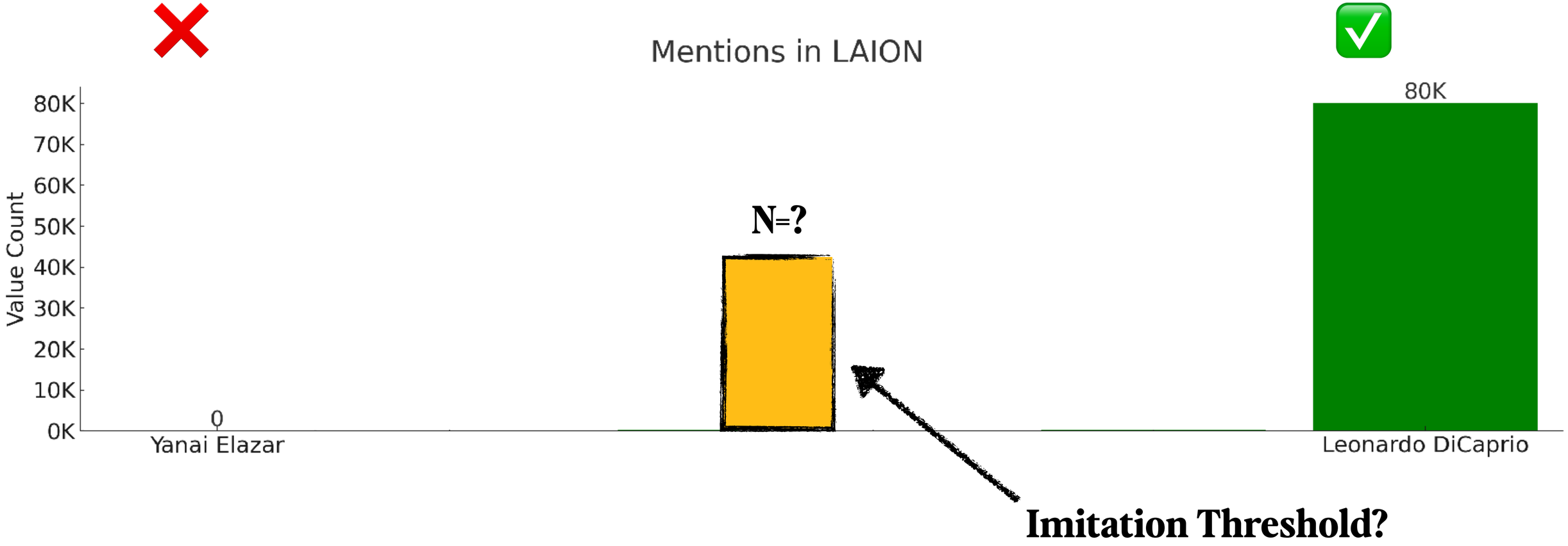
Imitation Threshold?



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*







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*

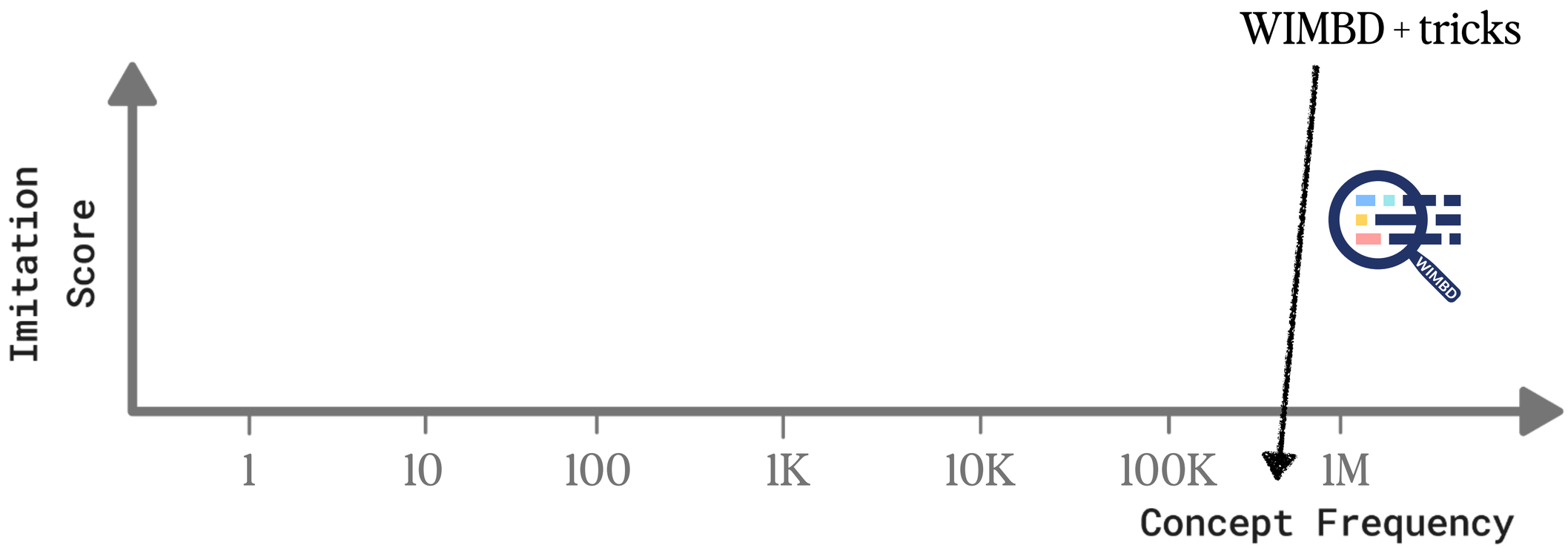
$$P(Imitation \mid do(count(Leo) = x))$$

Count: 80*K*

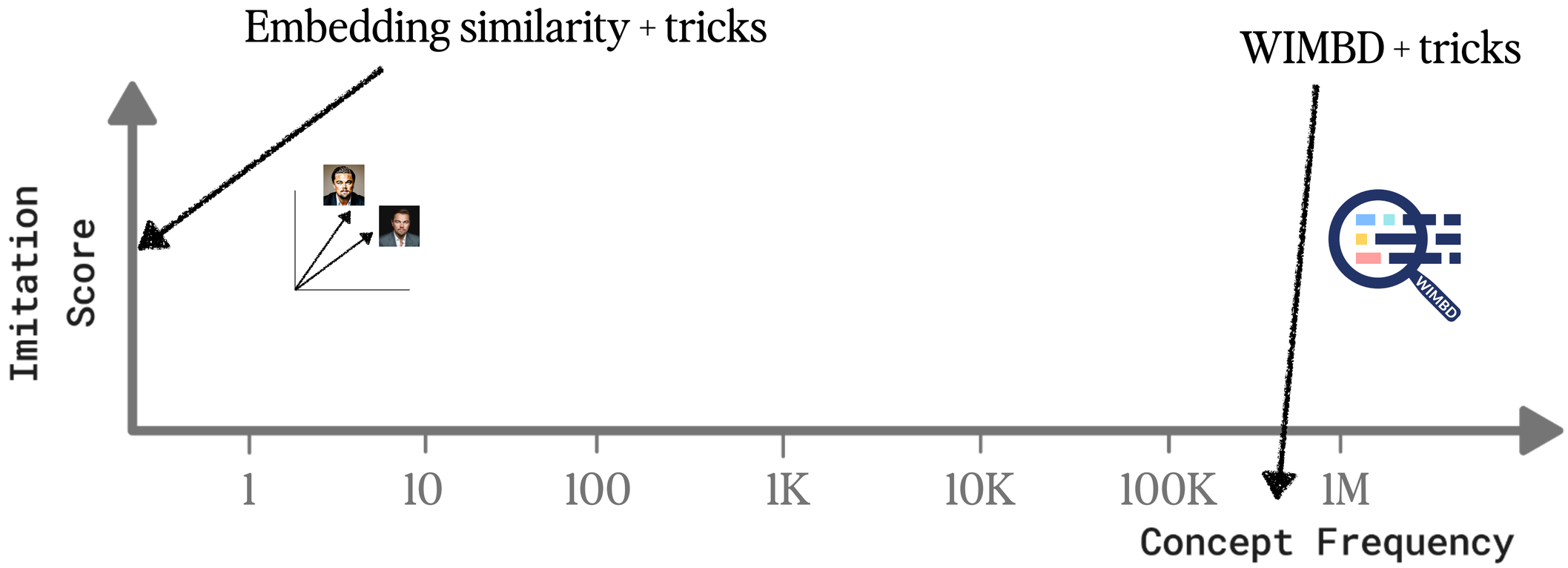
Solutions

1. Counterfactual model

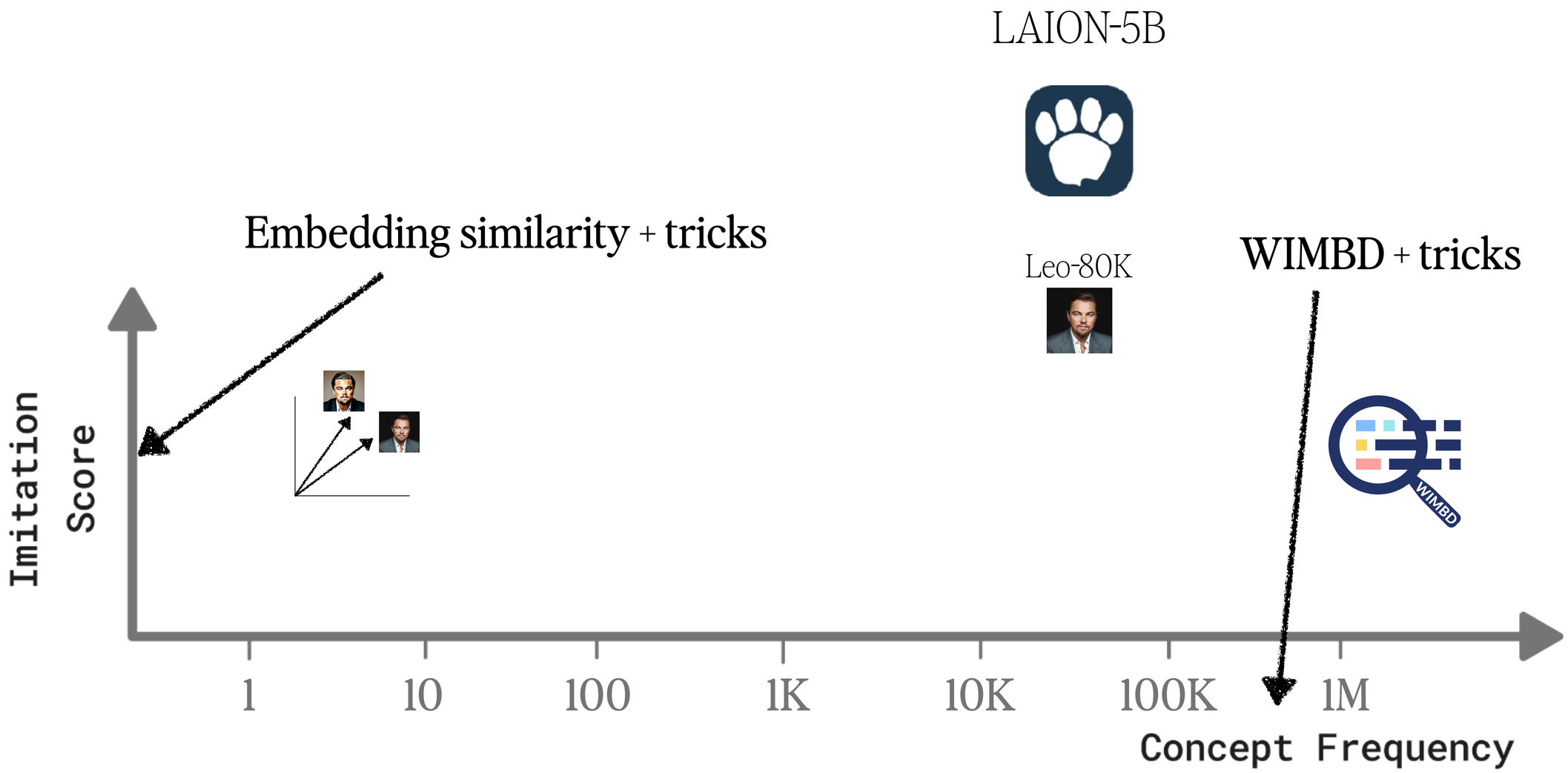
Solutions


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

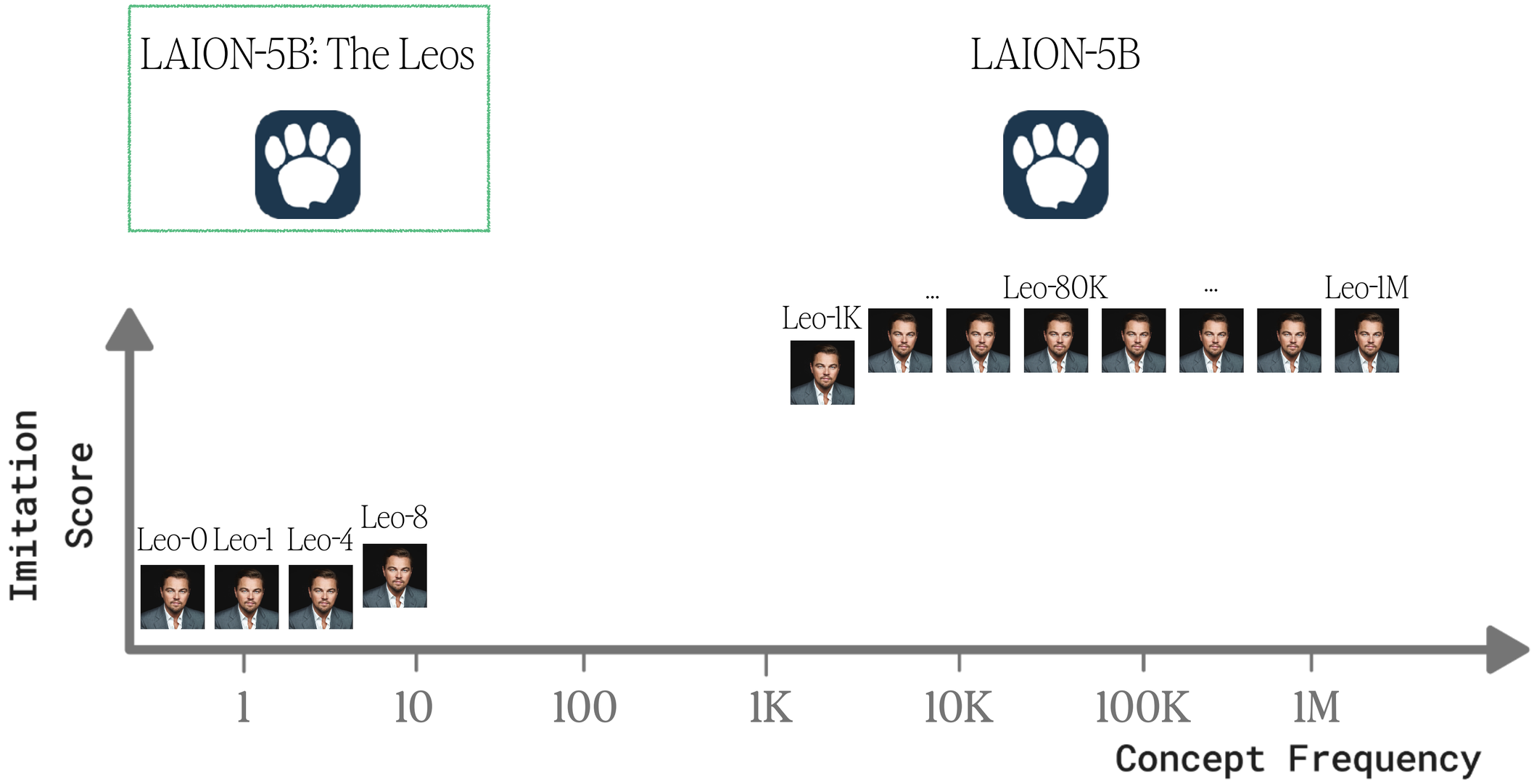
LAION-5B': The Leos

LAION-5B

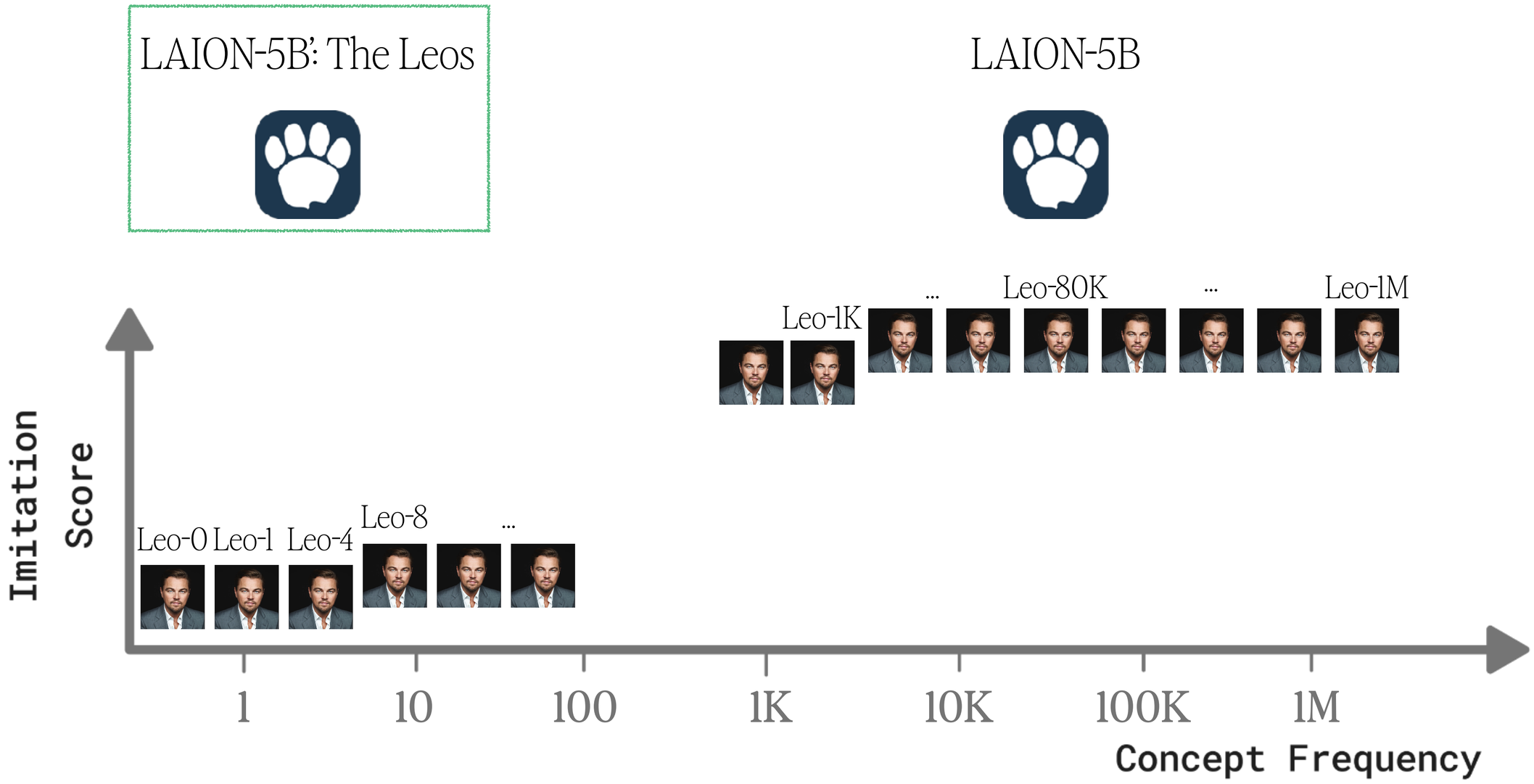
Leo-80K

Leo-0

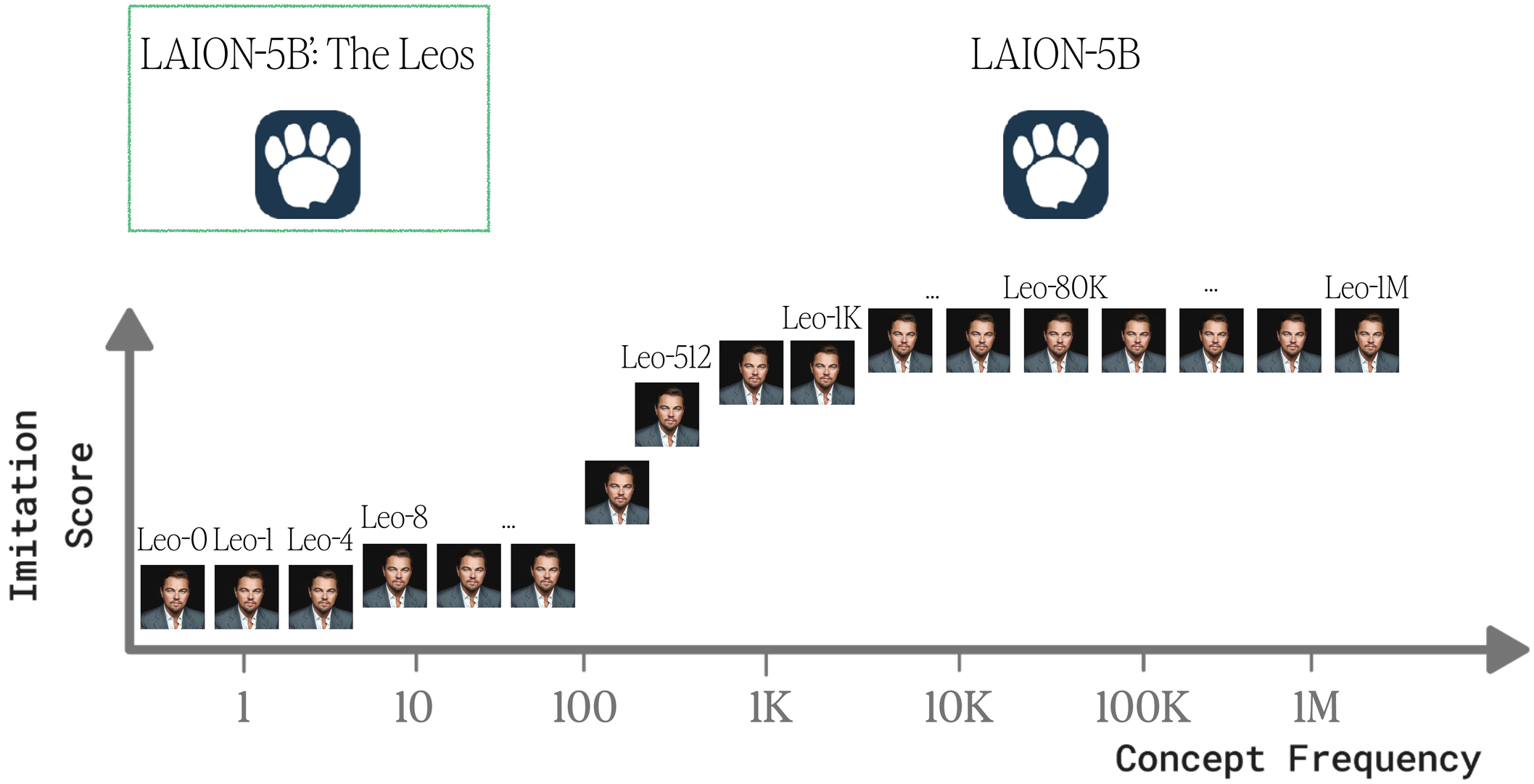
Solution #1


LAION-5B': The Leos
LAION-5B
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

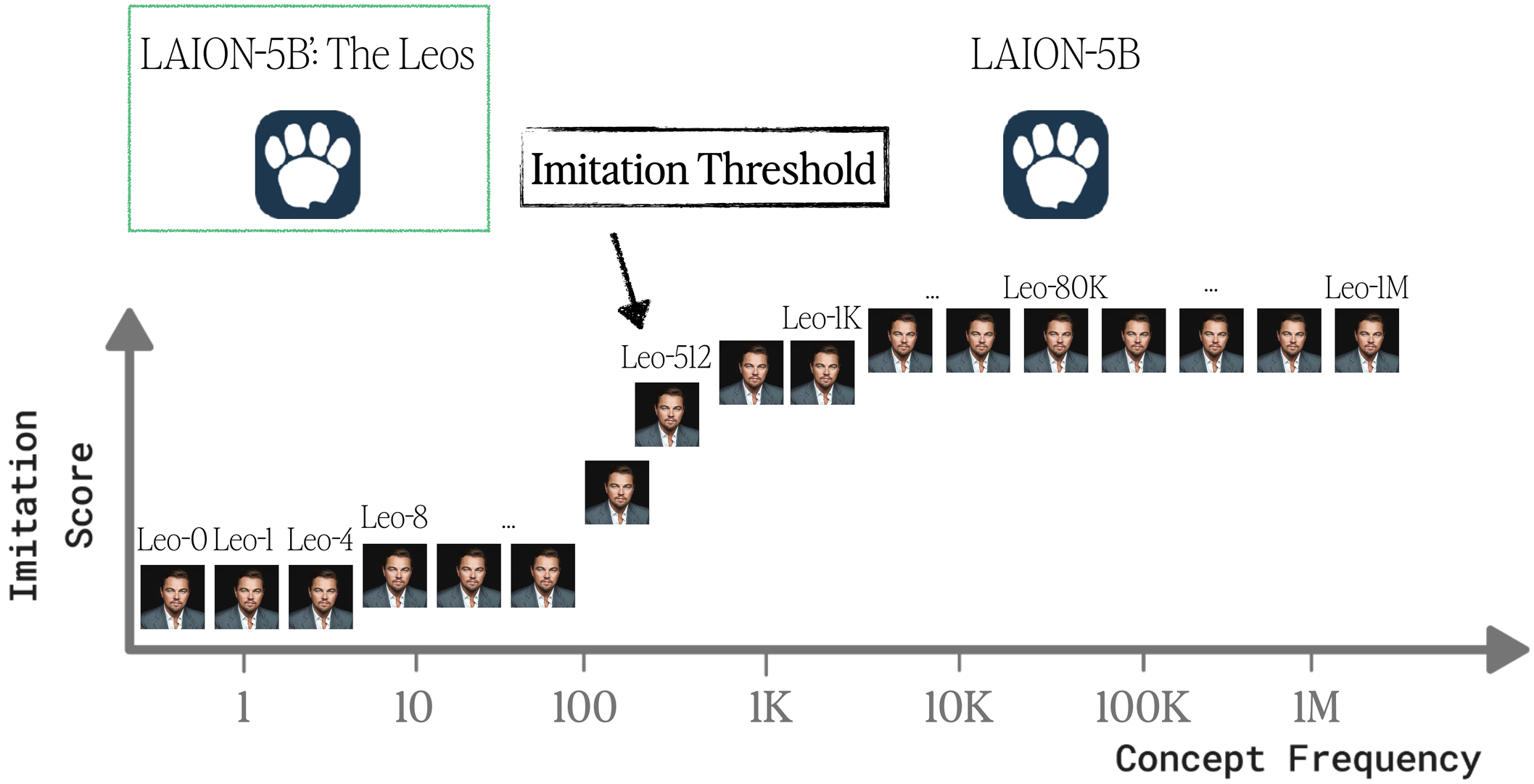
Solution #1


LAION-5B': The Leos
LAION-5B
Leo-0
Leo-1
Leo-4
Leo-8
...
Leo-1K
...
Leo-80K
...
Leo-1M
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 00K 1M

Concept Frequency

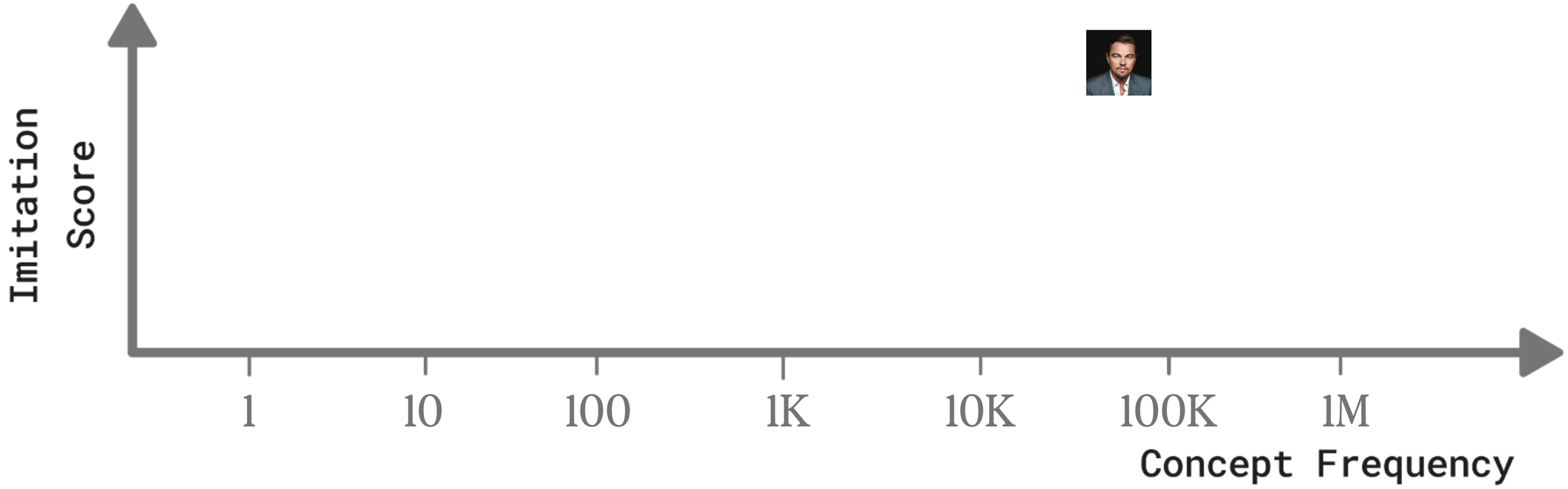
Solutions

1. Counterfactual model ❌

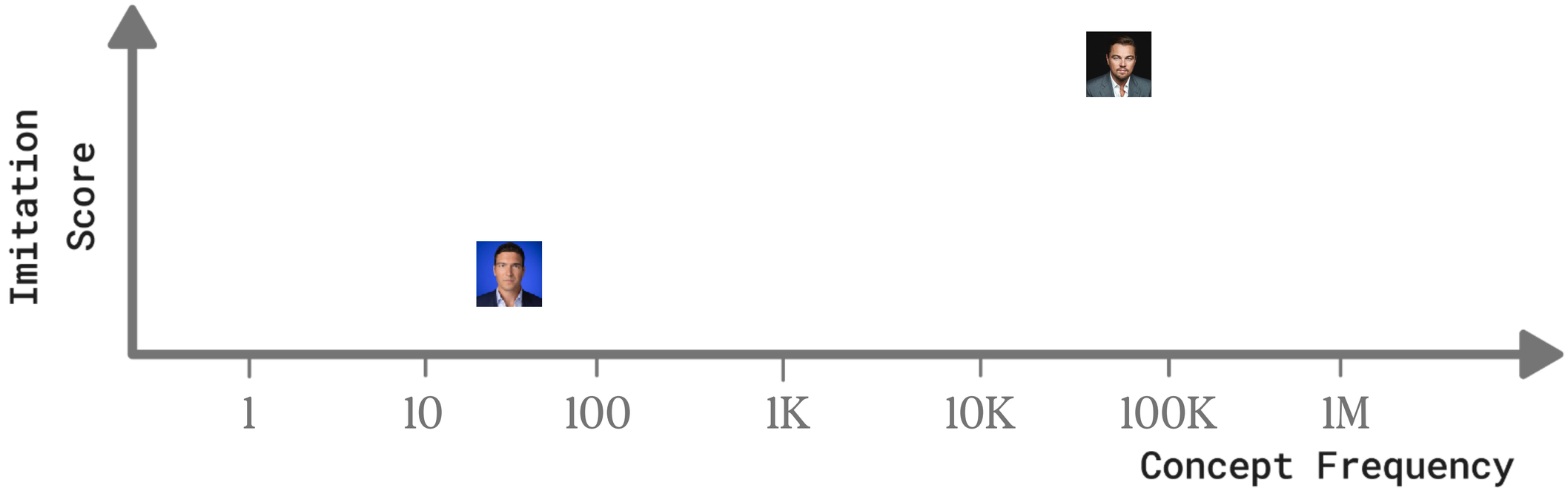
Solutions

1. Counterfactual model ❌
2. Observational approach

Solution #2



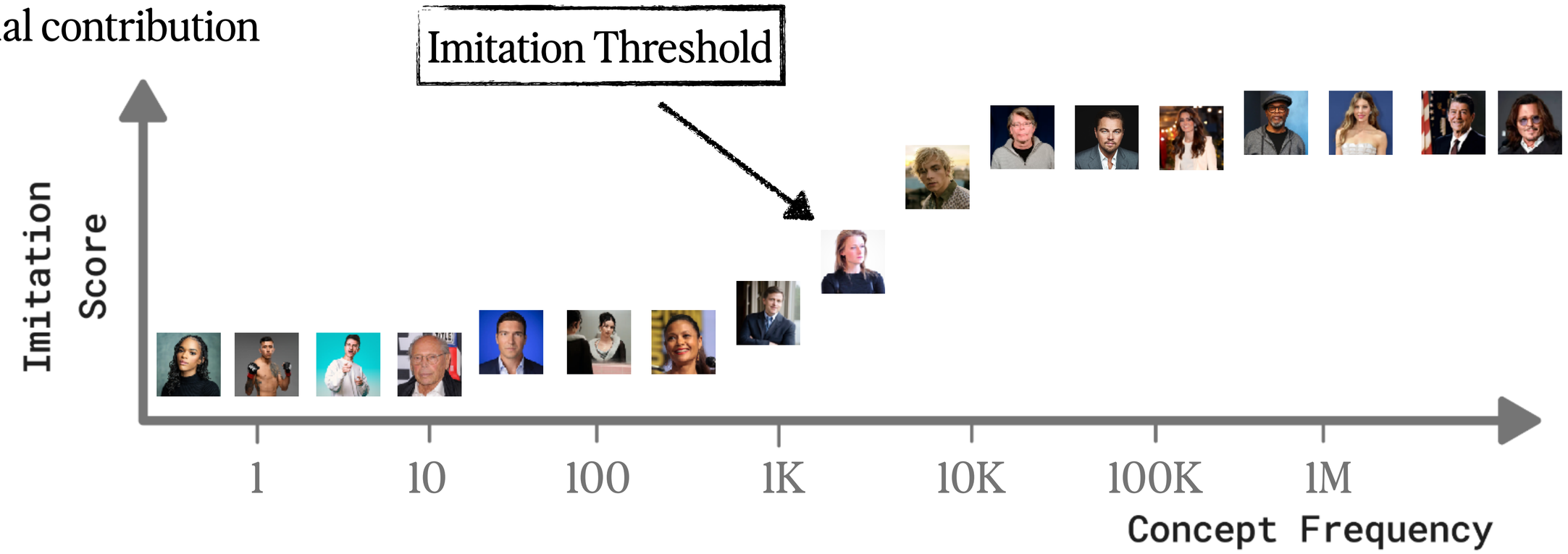
Solution #2



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

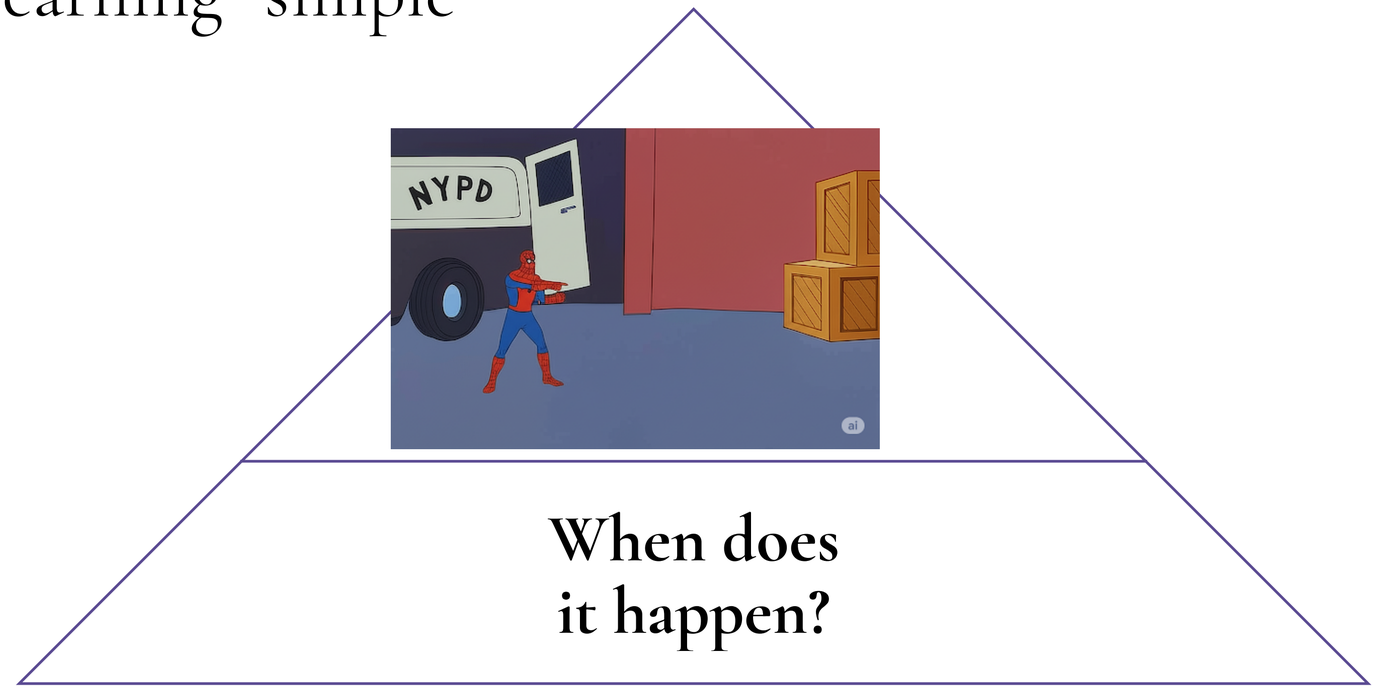
Imitation Threshold: 100-650 images

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.

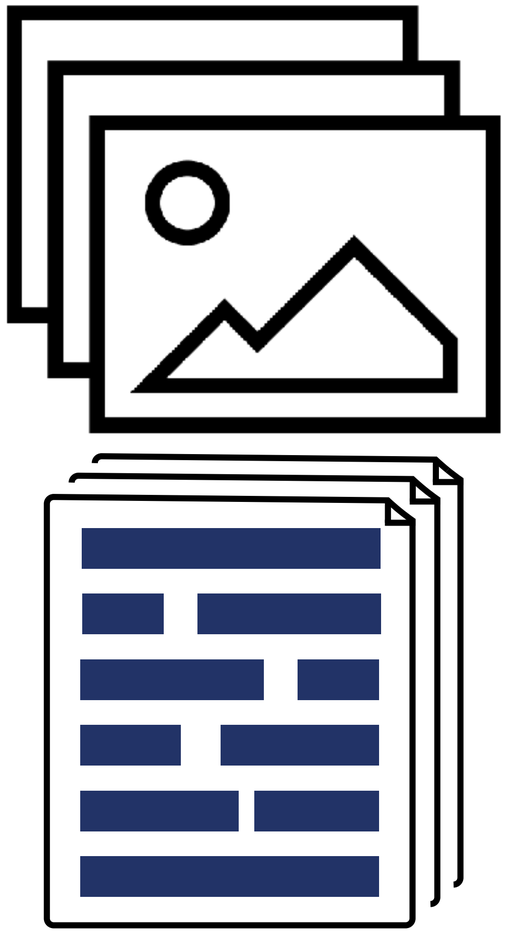
When does it Happen?

- Memorizing distribution happens
- But when is an under-explored area
- We have shown initial results for learning “simple” concepts
- But there so much more to learn

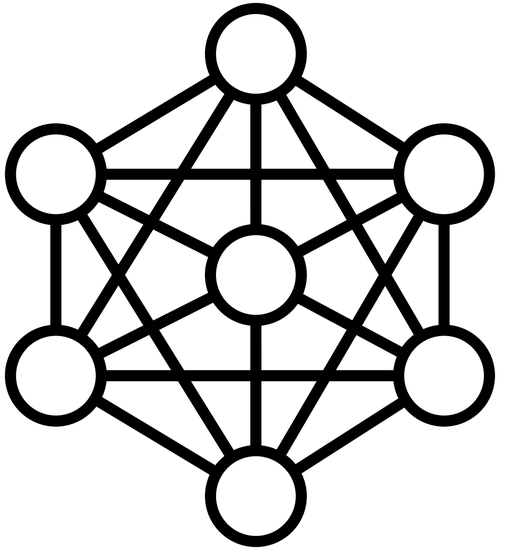


Measuring Distributional Memorization

The Data Behind Memorization

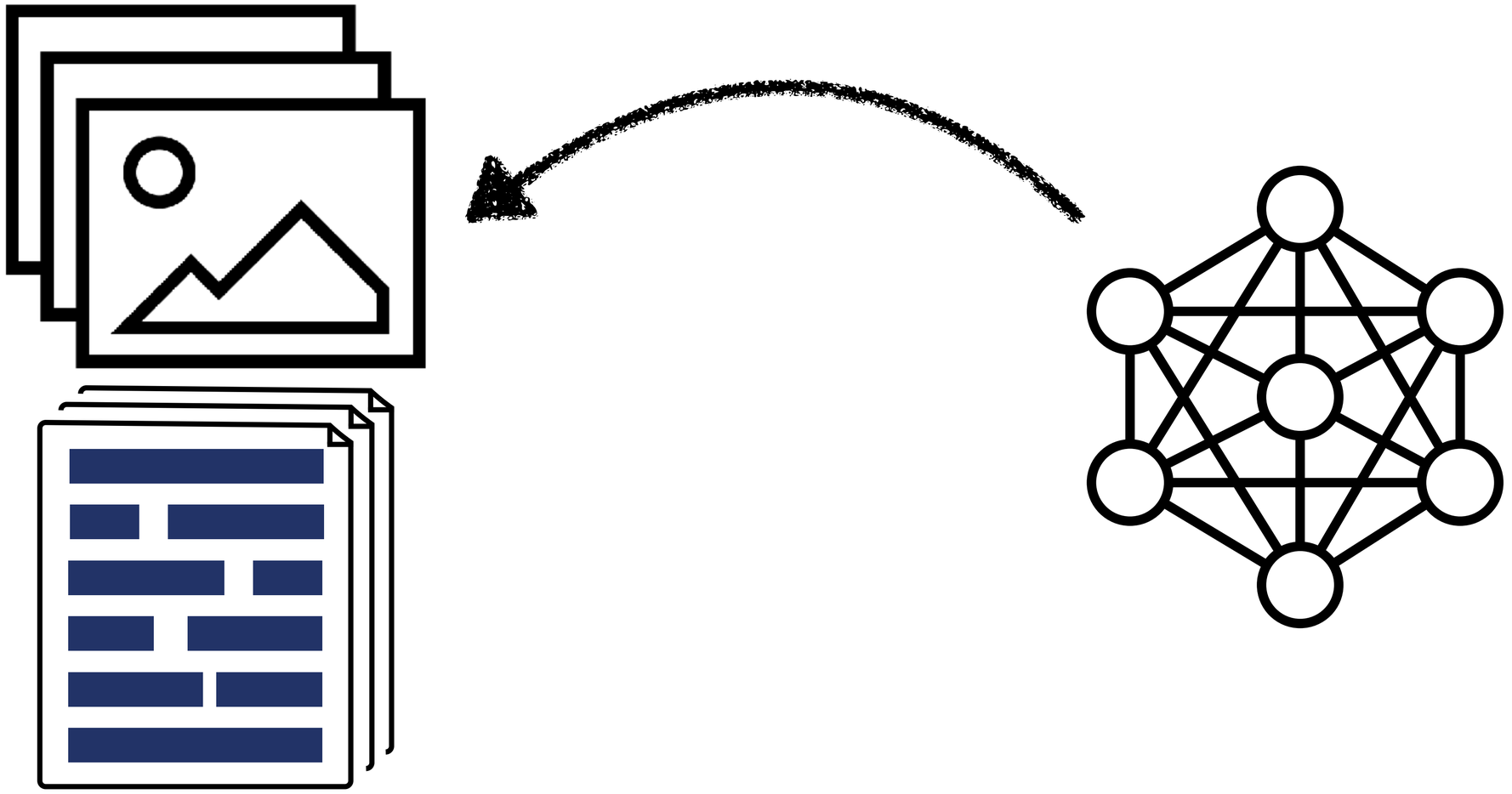


Dataset



Model

The Data Behind Memorization



Dataset

Model

The Tools Behind Memorization

- Modern datasets are huge
- Naively querying is inefficient
- We need something better

The Tools Behind Memorization

Suffix-Array

Deduplicating Training Data Makes Language Models Better

Katherine Lee[*†] **Daphne Ippolito**[*††‡] **Andrew Nystrom**[†] **Chiyuan Zhang**[†]
Douglas Eck[†] **Chris Callison-Burch**[‡] **Nicholas Carlini**[†]

Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens

Jiacheng Liu[♡] **Sewon Min**[♡]
Luke Zettlemoyer[♡] **Yejin Choi**[♡♠] **Hannaneh Hajishirzi**[♡♠]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Allen Institute for Artificial Intelligence liujc@cs.washington.edu

Other Data Structures

Evaluating n-Gram Novelty of Language Models Using RUSTY-DAWG 🐶

William Merrill[α,β] **Noah A. Smith**[γ,β] **Yanai Elazar**[β,γ]
[α]New York University [β]Allen Institute for AI [γ]University of Washington
willm@nyu.edu, noah@allenai.org, yanaiela@gmail.com

INFINI-GRAM MINI: Exact n-gram Search at the Internet Scale with FM-Index

Hao Xu[♡] **Jiacheng Liu**[♡♠] **Yejin Choi**[♢] **Noah A. Smith**[♡♠] **Hannaneh Hajishirzi**[♡♠]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Allen Institute for AI [♢]Stanford University

WHAT'S IN MY BIG DATA?

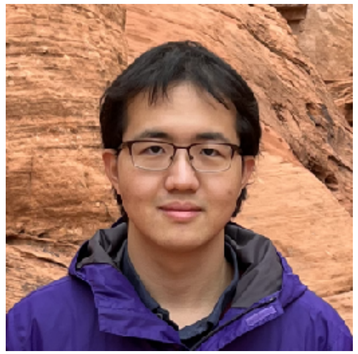
Yanai Elazar[1,2] **Akshita Bhagia**[1] **Ian Magnusson**[1] **Abhilasha Ravichander**[1]
Dustin Schwenk[1] **Alane Suhr**[3] **Pete Walsh**[1] **Dirk Groeneveld**[1] **Luca Soldaini**[1]
Sameer Singh[4] **Hannaneh Hajishirzi**[1,2] **Noah A. Smith**[1,2] **Jesse Dodge**[1]

[1]Allen Institute for AI
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington
[3]University of California, Berkeley [4]University of California, Irvine

Combining the Tools & Generation

OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens

Jiacheng Liu[αω] **Taylor Blanton**[α] **Yanai Elazar**[αω] **Sewon Min**[αβ]
YenSung Chen[α] **Arnavi Chheda-Kothary**[αω] **Huy Tran**[α] **Byron Bischoff**[α] **Eric Marsh**[α]
Michael Schmitz[α] **Cassidy Trier**[α] **Aaron Sarnat**[α] **Jenna James**[α] **Jon Borchardt**[α]
Bailey Kuehl[α] **Evie Cheng**[α] **Karen Farley**[α] **Sruthi Sreeram**[α] **Taira Anderson**[α]
David Albright[α] **Carissa Schoenick**[α] **Luca Soldaini**[α] **Dirk Groeneveld**[α]
Rock Yuren Pang[ω]
Pang Wei Koh[αω] **Noah A. Smith**[αω] **Sophie Lebrecht**[α] **Yejin Choi**[σ]
Hannaneh Hajishirzi[αω] **Ali Farhadi**[αω] **Jesse Dodge**[α]
[α]Allen Institute for AI [ω]University of Washington [β]UC Berkeley [σ]Stanford University



ACL best! 🥳🥳 *demo 2025*

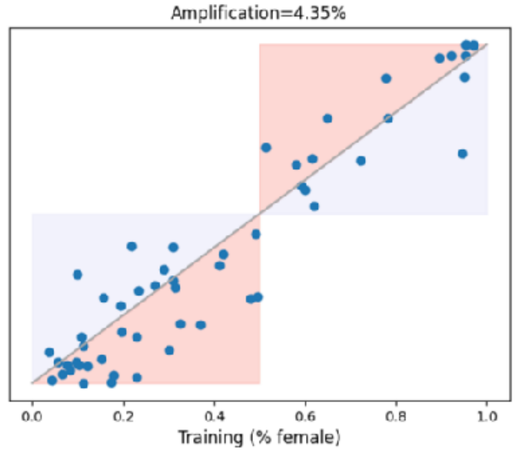
The Tools Behind Memorization

- Most current focus on string matching
- We need to do better

Memorizing Distributions

The **Bias**
Amplification Paradox

What is?



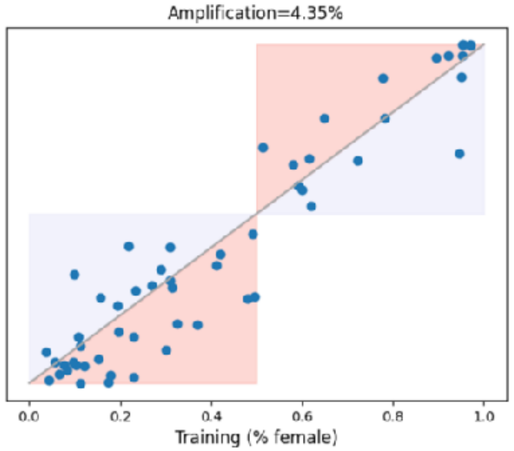
Memorizing Distributions

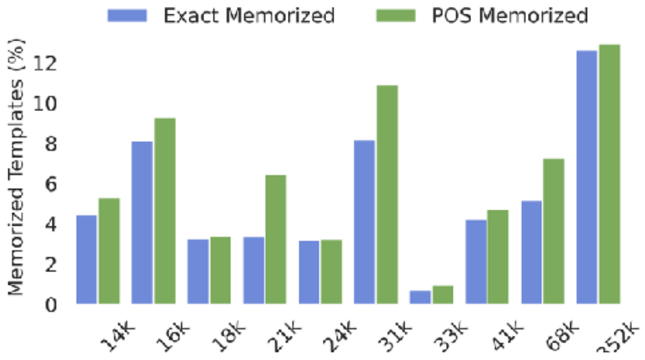
The **Bias Amplification Paradox**

What is?

Syntactic Templates in texts

How much it happens?





Memorizing Distributions

The **Bias Amplification Paradox**

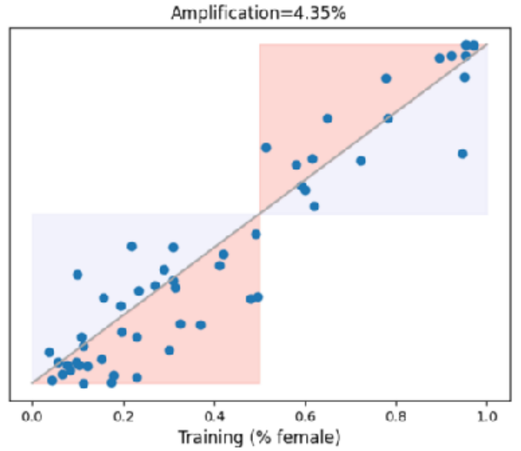
What is?

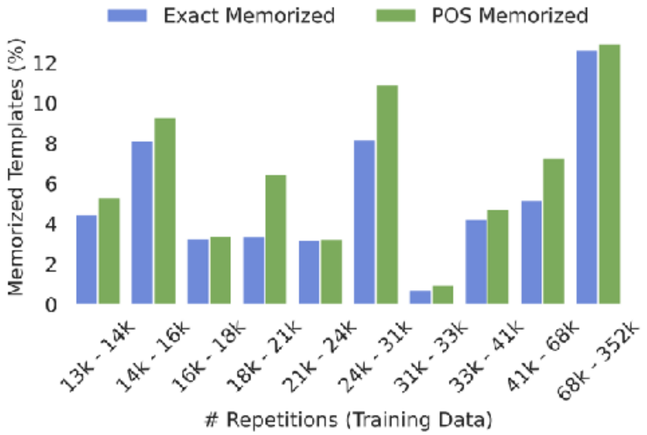
Syntactic Templates in texts

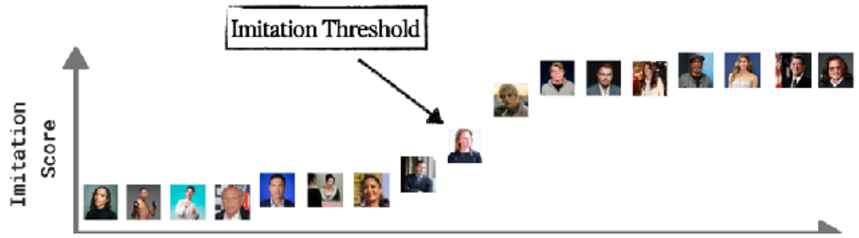
How much it happens?

Imitation Threshold

When does it happen?







Memorizing Distributions

The **Bias Amplification Paradox**

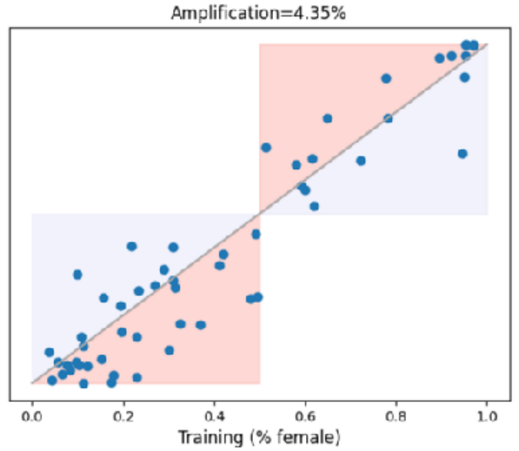
What is?

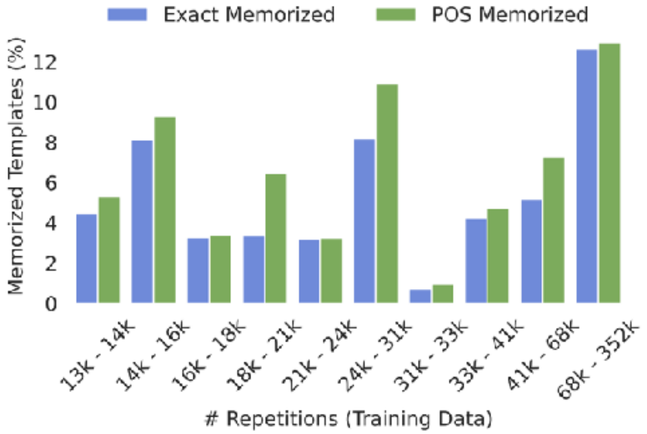
Syntactic Templates in texts

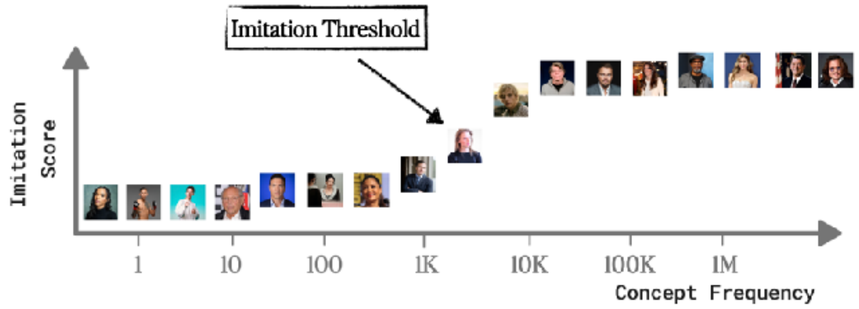
How much it happens?

Imitation Threshold

When does it happen?







WIMBD

∞-gram

OLMoTrace

Thank You!

Questions?

✉ yanaiela@gmail.com

@yanaiela

@yanai.bsky.social