Barely Legal:
AI Introduction for Law Students

Yanai Elazar, 12/01/2025

About Myself

- Yanai Elazar
- Assistant Professor at Bar-Ilan University, Computer Science Department
- Research Interests: Understanding how Generative Models Work

The Journey - Brief History

Pre-2017: Task Specific models, supervised datasets

SQuAD2.0
The Stanford Question Answering Dataset

GLUE

The TAC Relation Extraction Dataset

A large-scale relation extraction dataset with 106k+ examples over 42 TAC KBP relation types.

The Journey - Brief History

Pre-2027: Task Specific models, supervised datasets

2017: Attention is All You Need - The Transformer Revolution

SQuAD2.0
The Stanford Question Answering Dataset

GLUE

The TAC Relation Extraction Dataset

A large-scale relation extraction dataset with 106k+ examples over 42 TAC KBP relation types.

The Journey - Brief History

Pre-2027: Task Specific models, supervised datasets

2017: Attention is All You Need - The Transformer Revolution

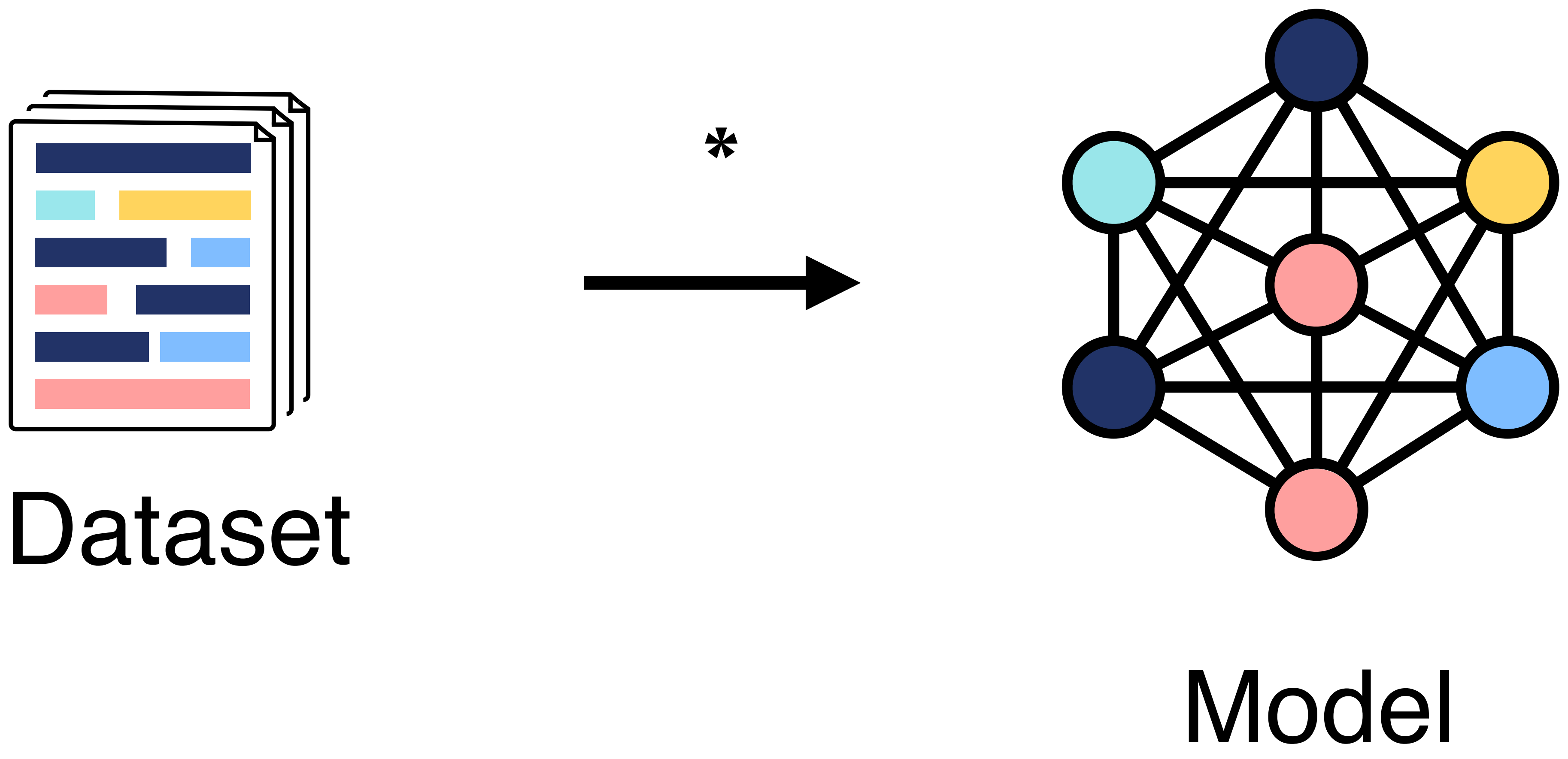
2018-2022: From GPT-1 to ChatGPT: Scaling Works

SQuAD2.0
The Stanford Question Answering Dataset

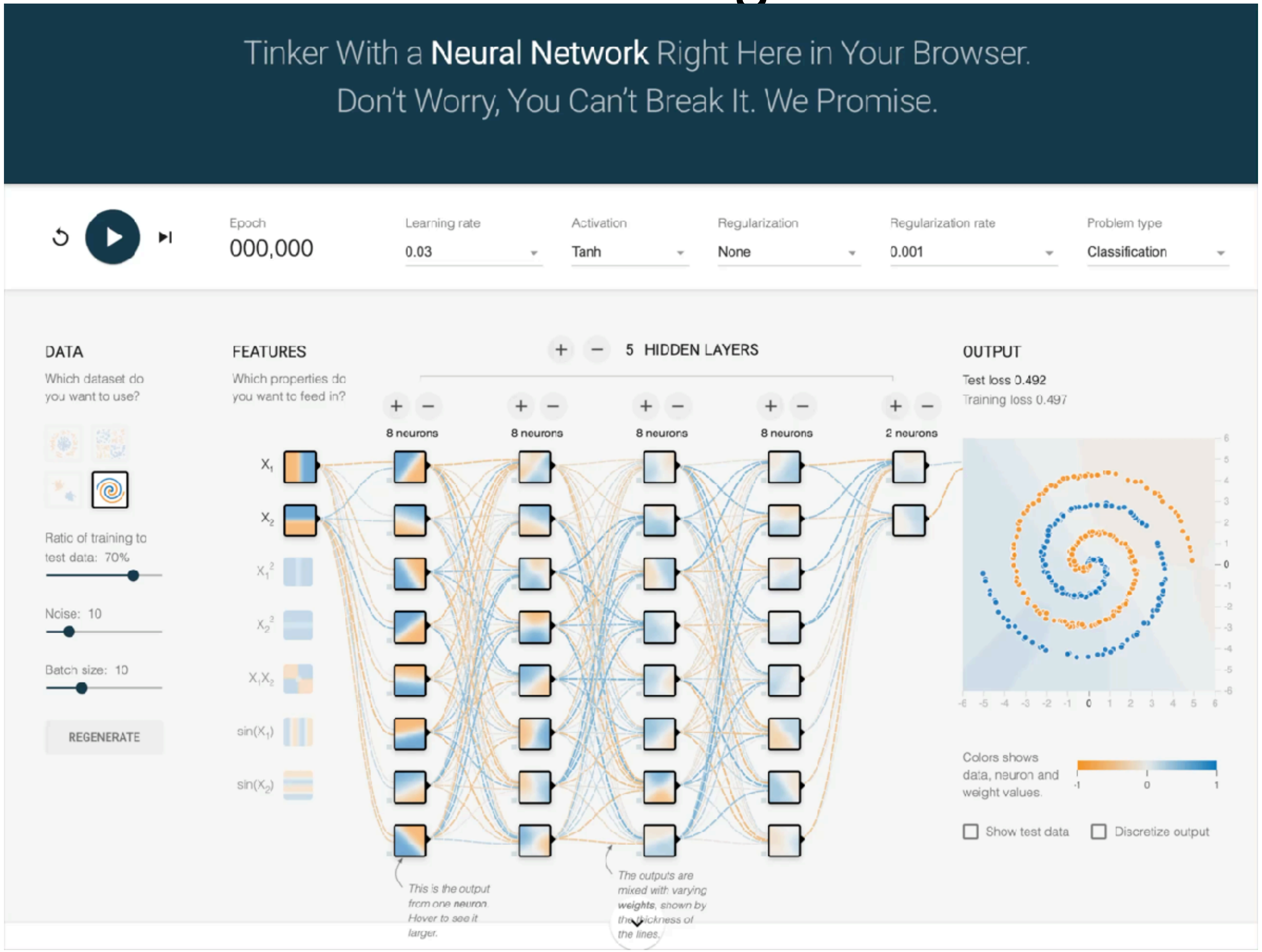
GLUE

The TAC Relation Extraction Dataset
A large-scale relation extraction dataset with 106k+ examples over 42 TAC KBP relation types.

The Backbone - Machine Learning



The Backbone - Machine Learning

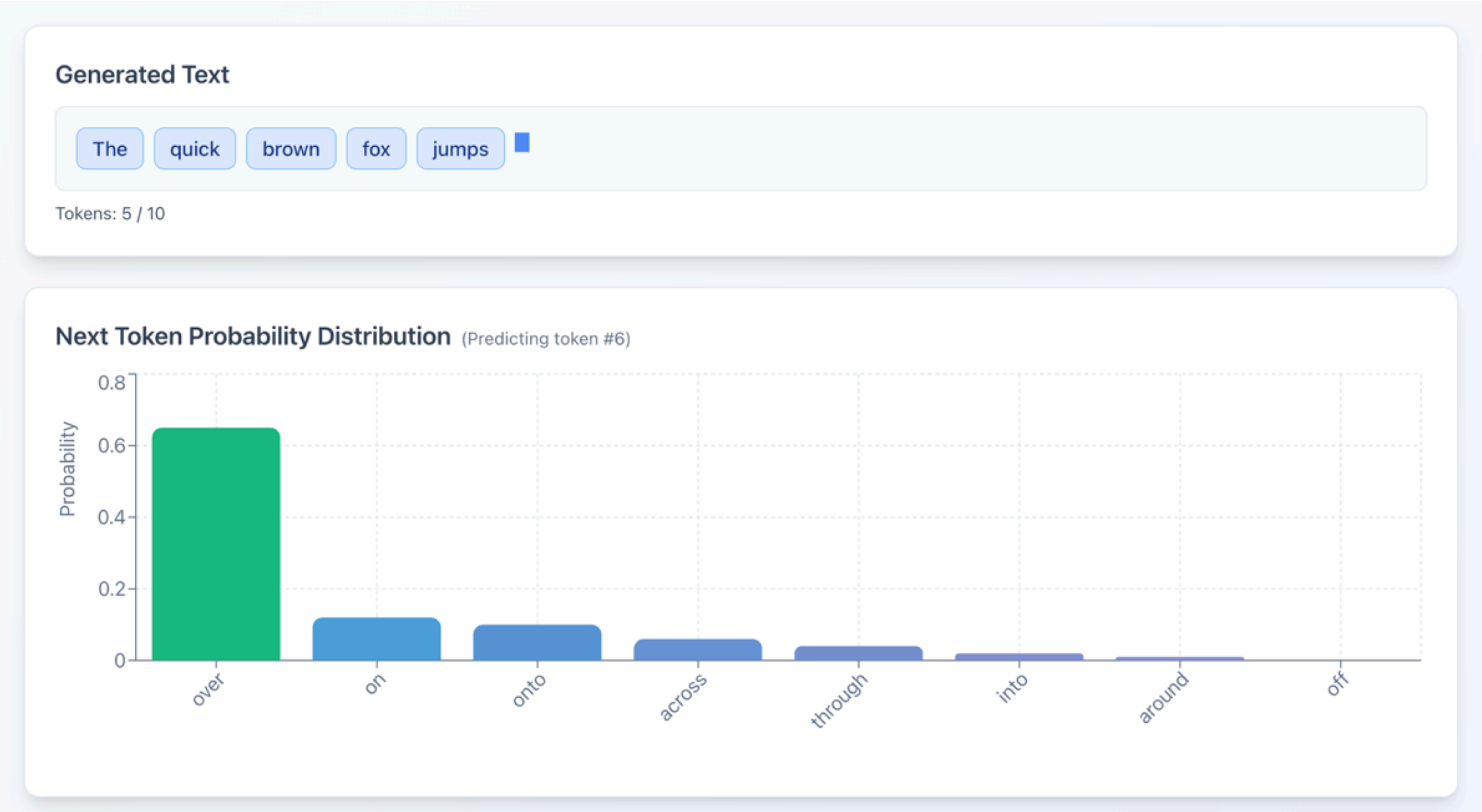


playground.tensorflow.org

The Backbone - Language Models

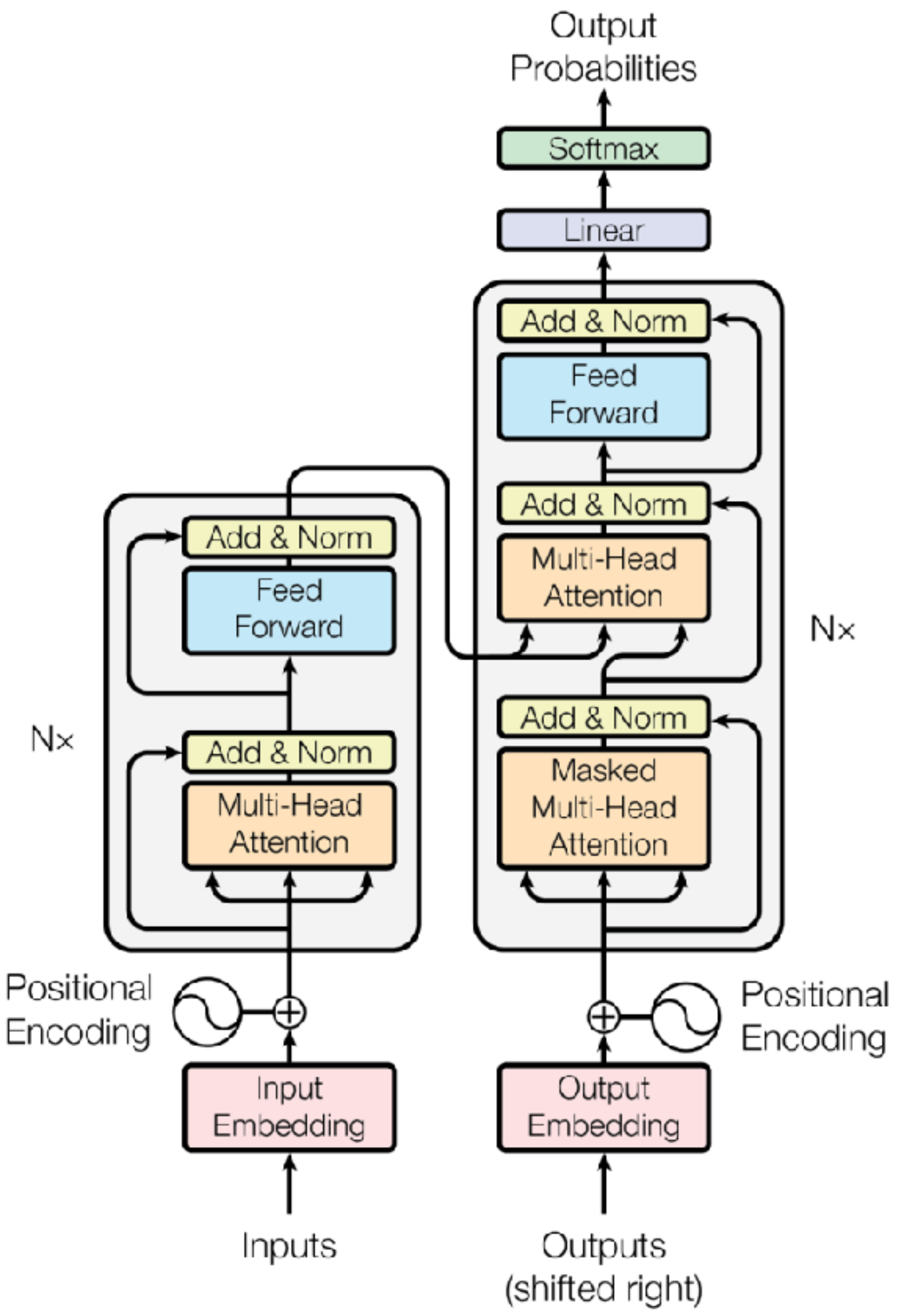
Input: *n* “words”

Output: a distribution over *k* “words”



The Backbone - Transformers

One (out of many) architectures

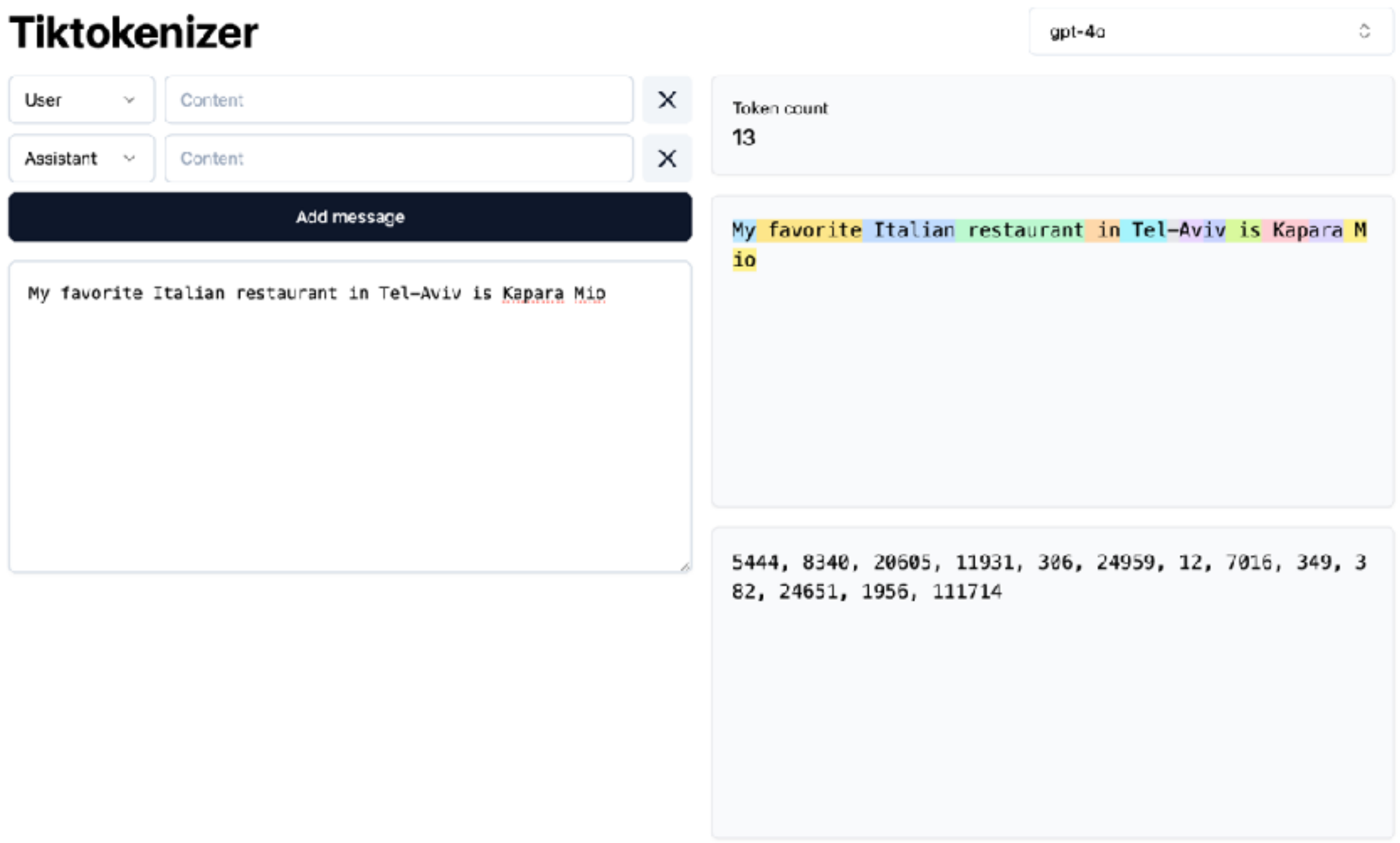


The Backbone - Tokenizers

Transformers don’t know what words are

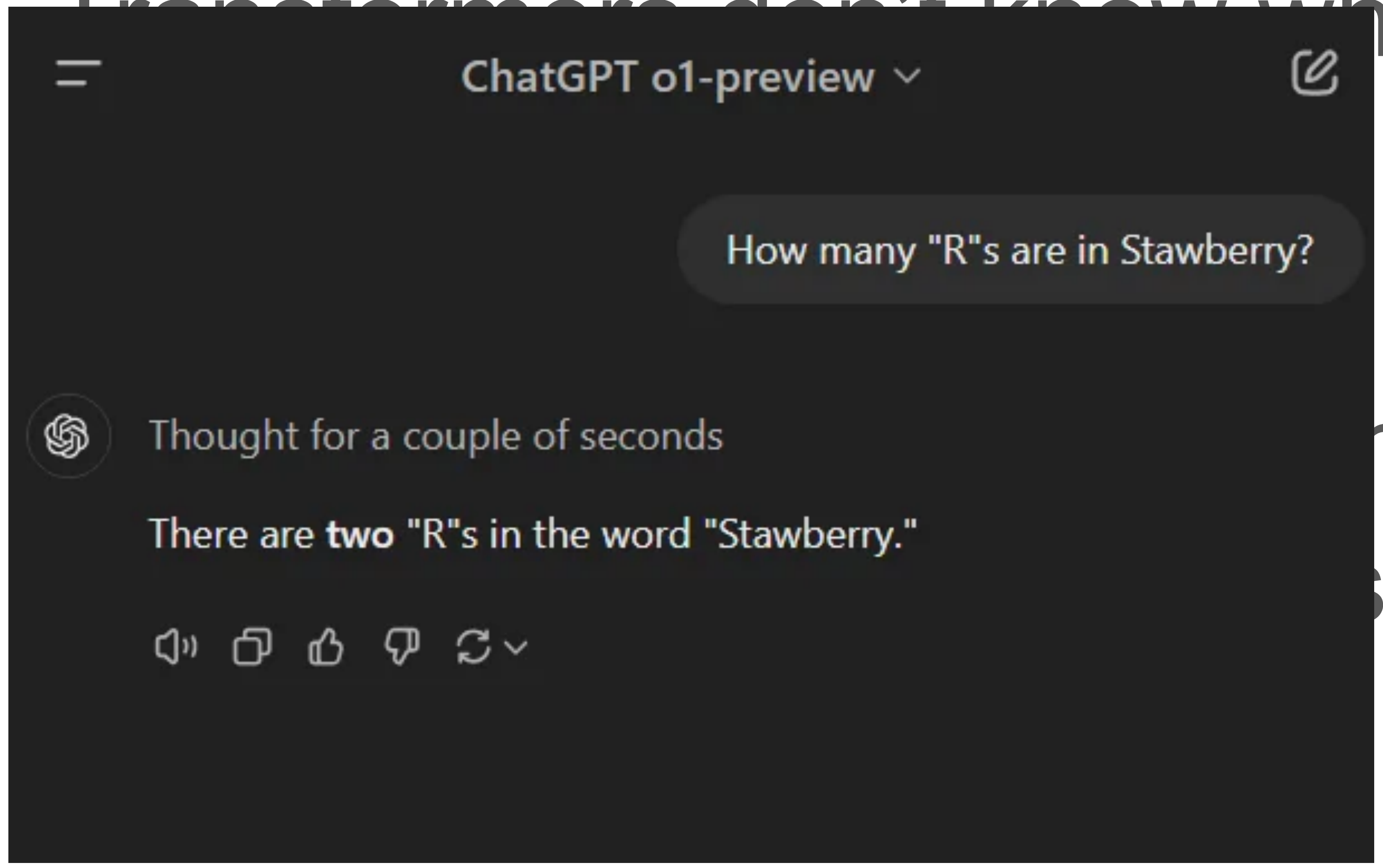
Strings are split into *tokens*

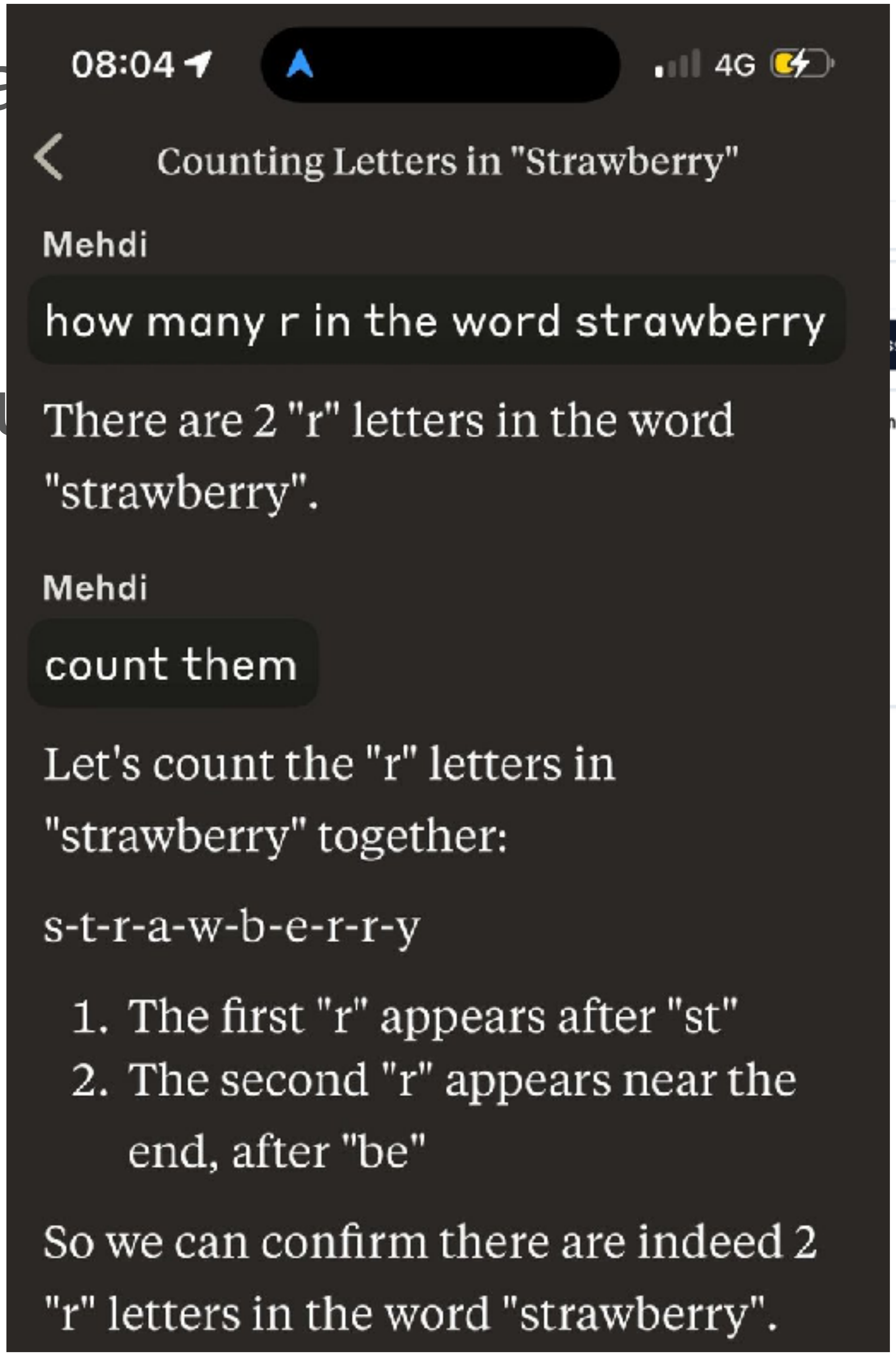
Tokens are represented as numbers that gets converted into a list of numbers (vectors)

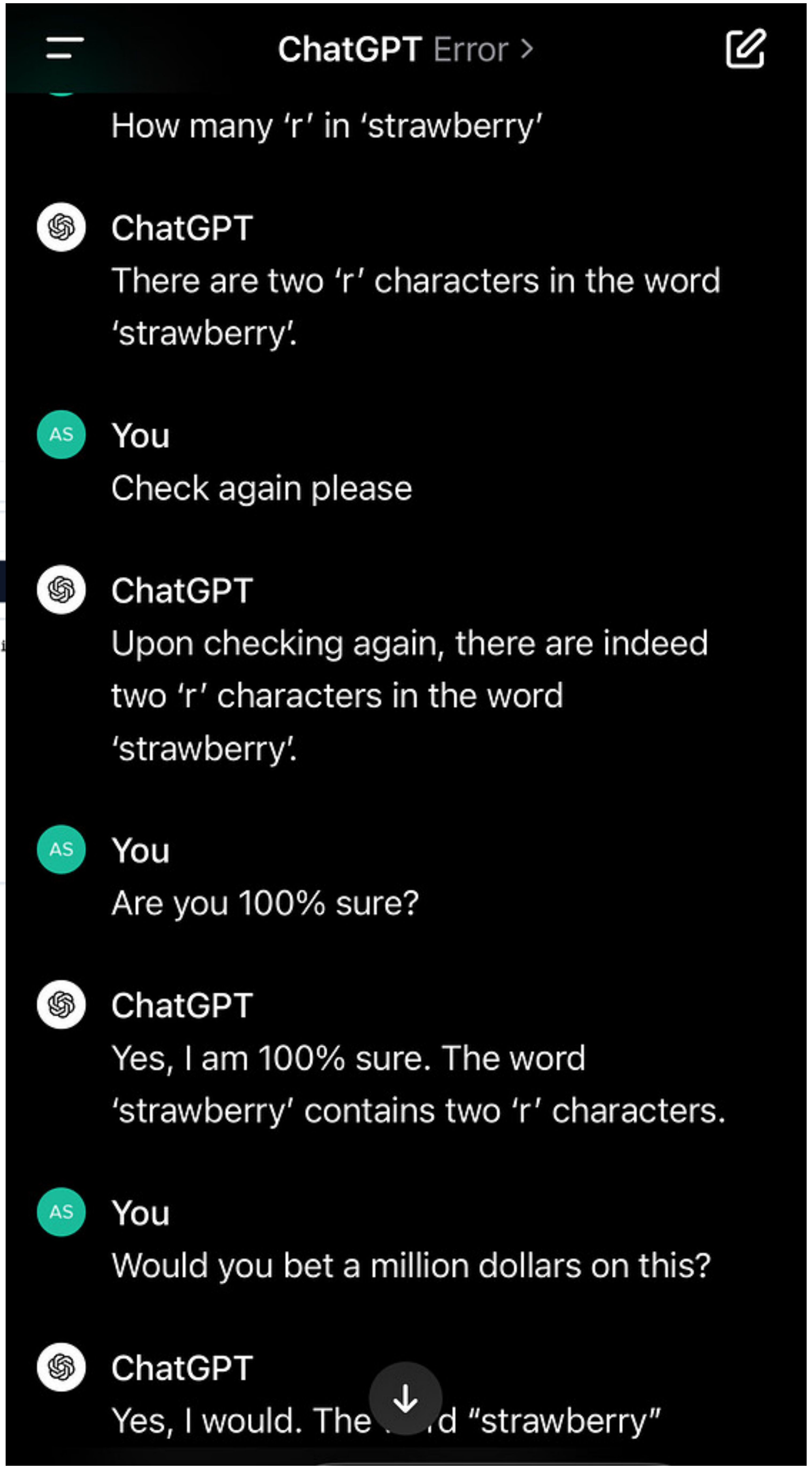


tiktokenizer.vercel.app

The Backbone - Tokenizers







The Backbone - Training Phases

Pre-training: “Reading” the entire internet (aka *self-supervised learning*)

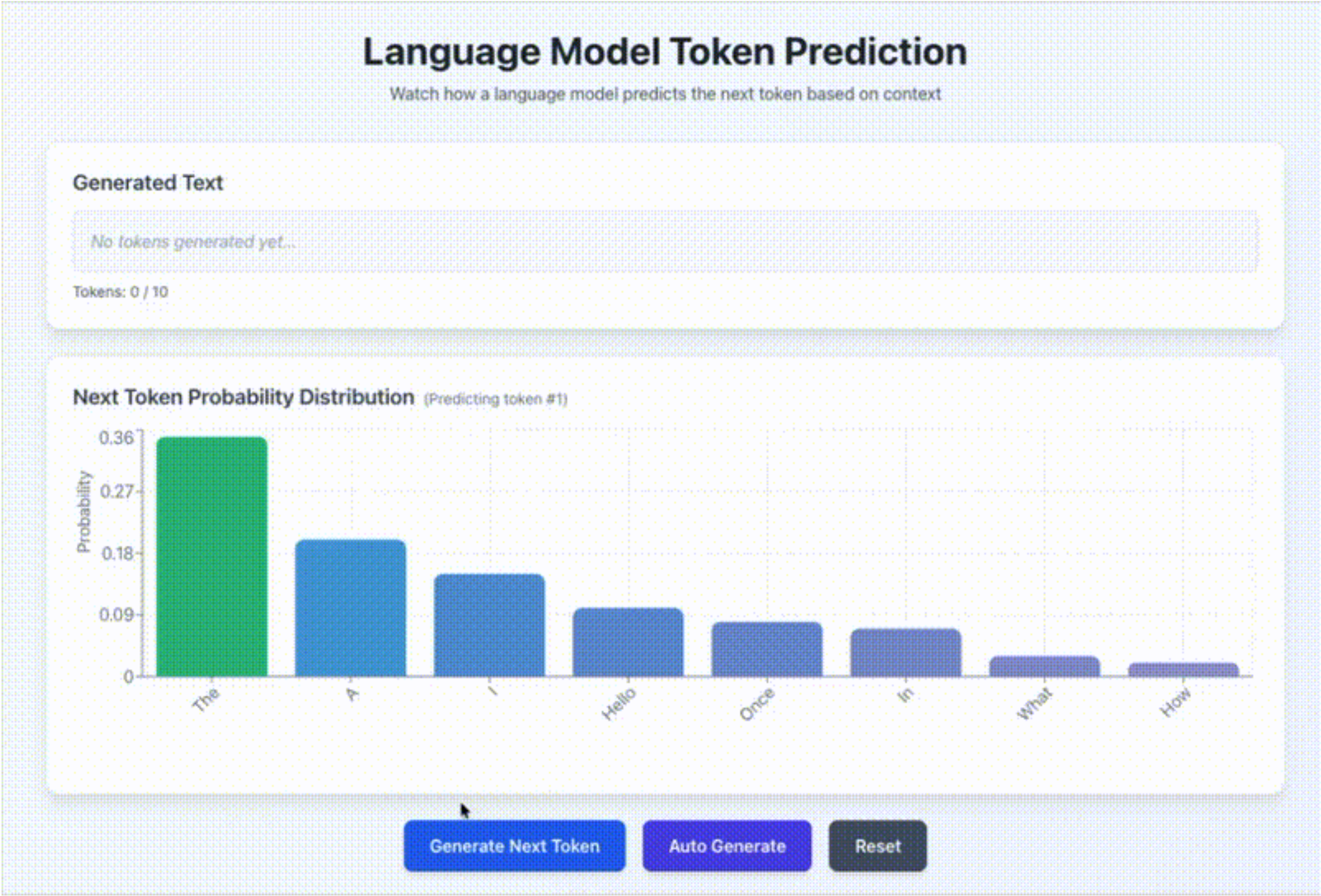
- Reading books, papers, wikipedia, reddit, etc.
- Absorb as much knowledge as possible

Post-training: Learning fine-grained capabilities, behaviors (aka *supervised learning*)

- Math reasoning
- Coding
- Instruction following

The Backbone - Inference

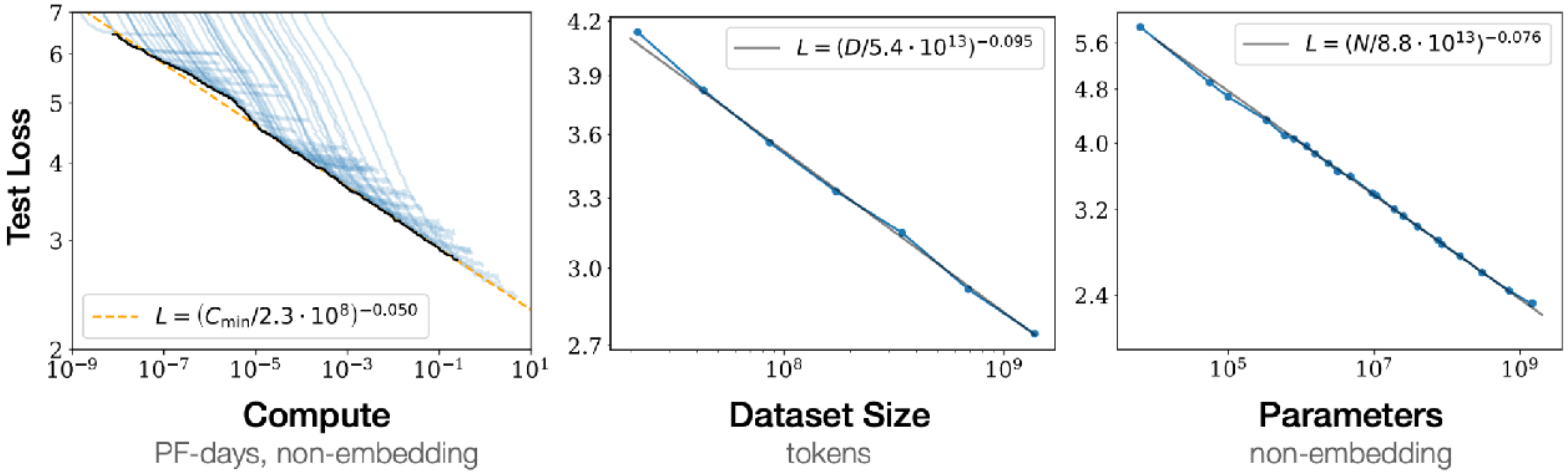
- Model is fixed
- Generate new data (text, images, etc)



The Backbone - Scaling

More = Better

- Model parameters
- Training time
- Data

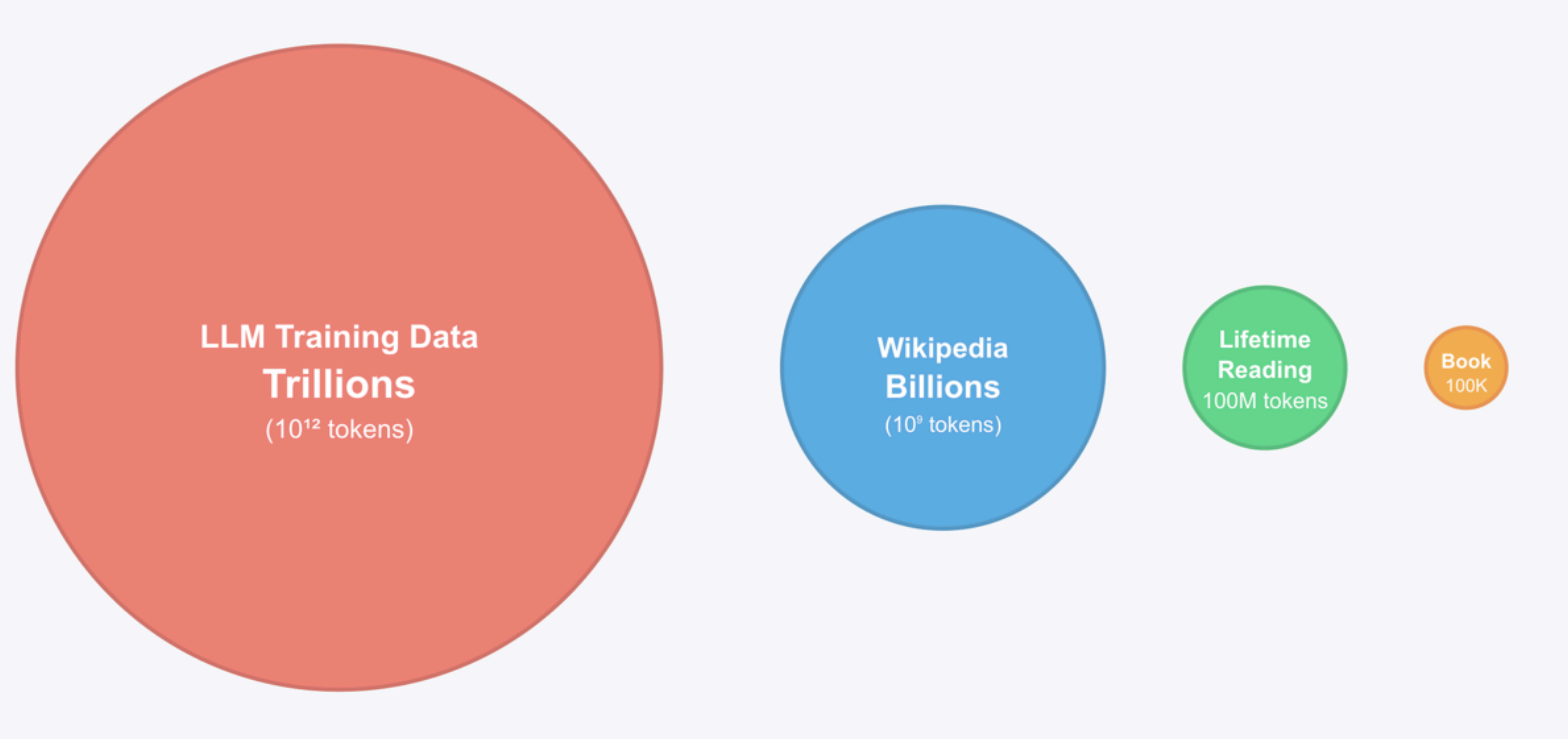


Kaplan et al., 2020

The Backbone - The Data

Models train to “mimic” the data they train on

- LLMs data: Trillions
- Wikipedia: Billions
- A person: 100 millions
- A book: 100 thousands



The Data

Why is the data so important?

Multimodal datasets: misogyny, pornography, and malignant stereotypes

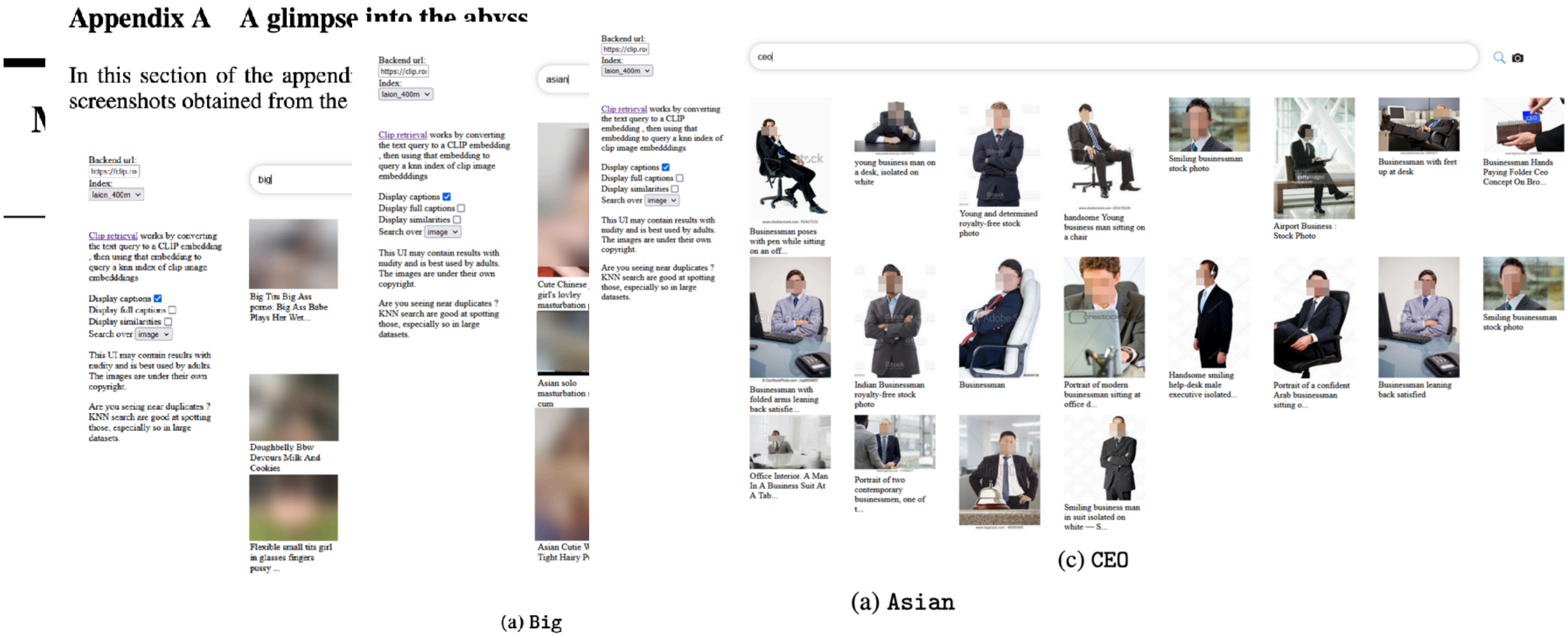
Abeba Birhane*
University College Dublin & Lero
Dublin, Ireland
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu*
Independent Researcher
vinaypra@alumni.cmu.edu

Emmanuel Kahembwe
University of Edinburgh
Edinburgh, UK
e.kahembwe@ed.ac.uk

The Data

Why is the data so important?



The Data

Why is the data so important?

M

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

yanaiela@gmail.com https://github.com/allenai/wimbd wimbd.apps.allenai.org

The Data

Why is the data so important?

Table 3: Most common 10-grams in five of the corpora we consider. n-grams from the top-10 that occur in more than (

n-gram

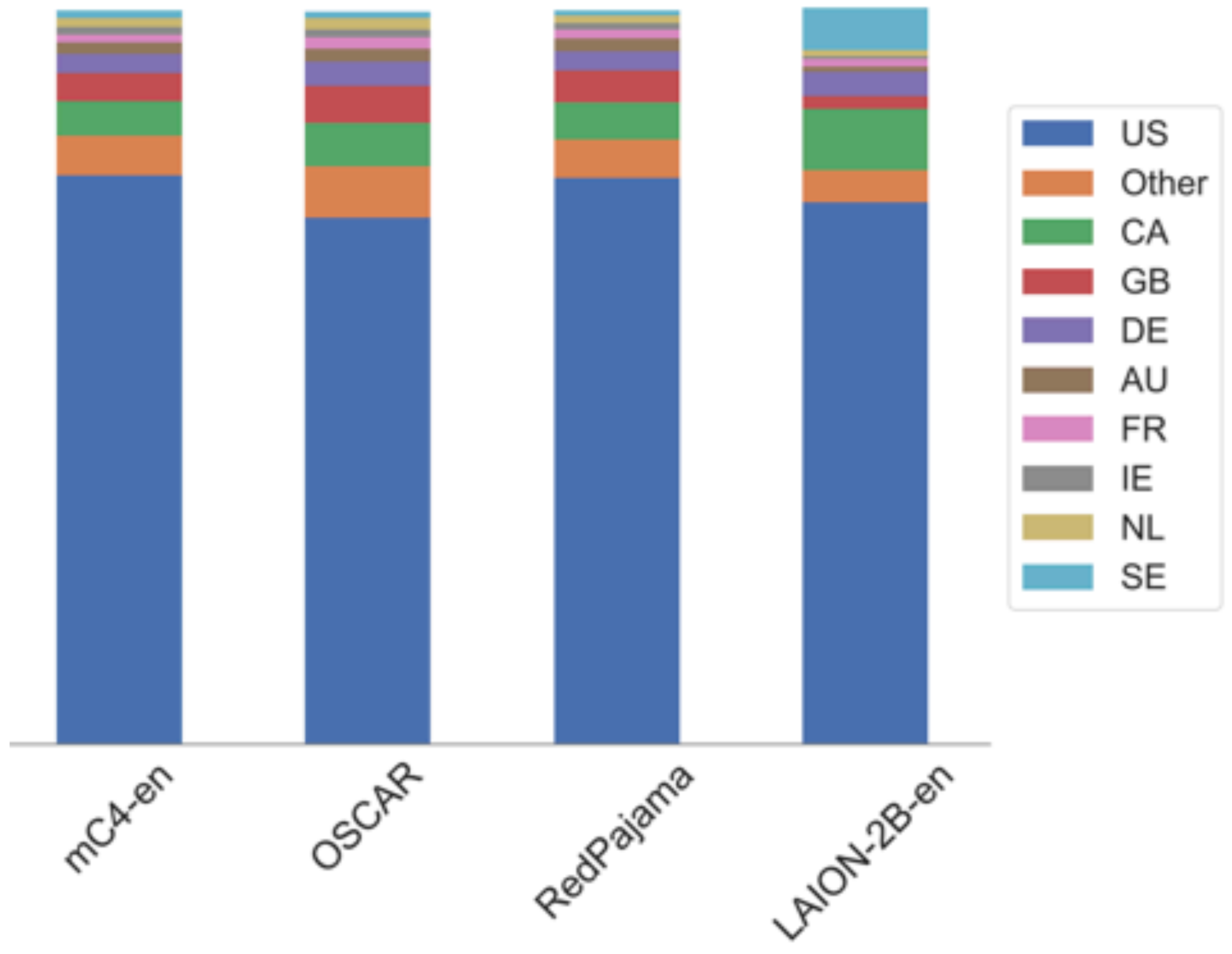
amp ; amp ;

; amp ; amp

n-gram

Corpus	Email Addresses Count	Email Addresses Prec.	Phone Numbers Count	Phone Numbers Prec.	IP Addresses Count	IP Addresses Prec.
OpenWebText	363,789.4	99	532,929.8	87	70,430.0	54
OSCAR	62,802,224.0	100	107,163,132.4	91	3,237,420.6	43
C4	7,614,759.2	99	19,702,198.4	92	796,494.7	56
mC4-en	201,368,945.0	92	4,067,997,426.2	66	97,887,510.2	44
The Pile	19,882,348.2	43	38,019,831.8	65	4,078,794.7	48
RedPajama	35,217,396.0	100	70,264,985.9	94	1,126,129.5	*30
S2ORC	630,130.0	*100	1,465,947.0	*100	0.0	*0
PeS2o	418,136.9	97	226,937.5	*30.8	0.0	*0
LAION-2B-en	636,252.1	*94	1,029,066.6	7	0.0	*0
The Stack	4,329,620.3	53	45,473,381.9	9	4,481,490.7	55

Table 19: Extrapolated ratios of PII frequency (the number of PII matches multiplied by the estimated precision), normalized by number of tokens in a corpus ($\frac{PII * Precision}{\#Tokens}$).



e of URLs (excluding unresolved URLS) given country. Only the nine most common 'other.' We label URLs we were unable to iese documents included.

The Data

Why is the data so important?

GRADE: Quantifying Sample Diversity in Text-to-Image Models

Royi Rassin
Bar-Ilan University

Aviv Slobodkin
Bar-Ilan University

Shauli Ravfogel
Bar-Ilan University
ETH Zürich

Yanai Elazar
Allen Institute for AI
University of Washington

Yoav Goldberg
Bar-Ilan University
Allen Institute for AI

yanaiela@gmail.com https://github.com/allenai/wimbd wimbd.apps.allenai.org

The Data

Why is the data so important?

GRADE: Quantifyi

Royi Rassin
Bar-Ilan University

Aviv S
Bar-Ilan

yanaiela@gmail.com http

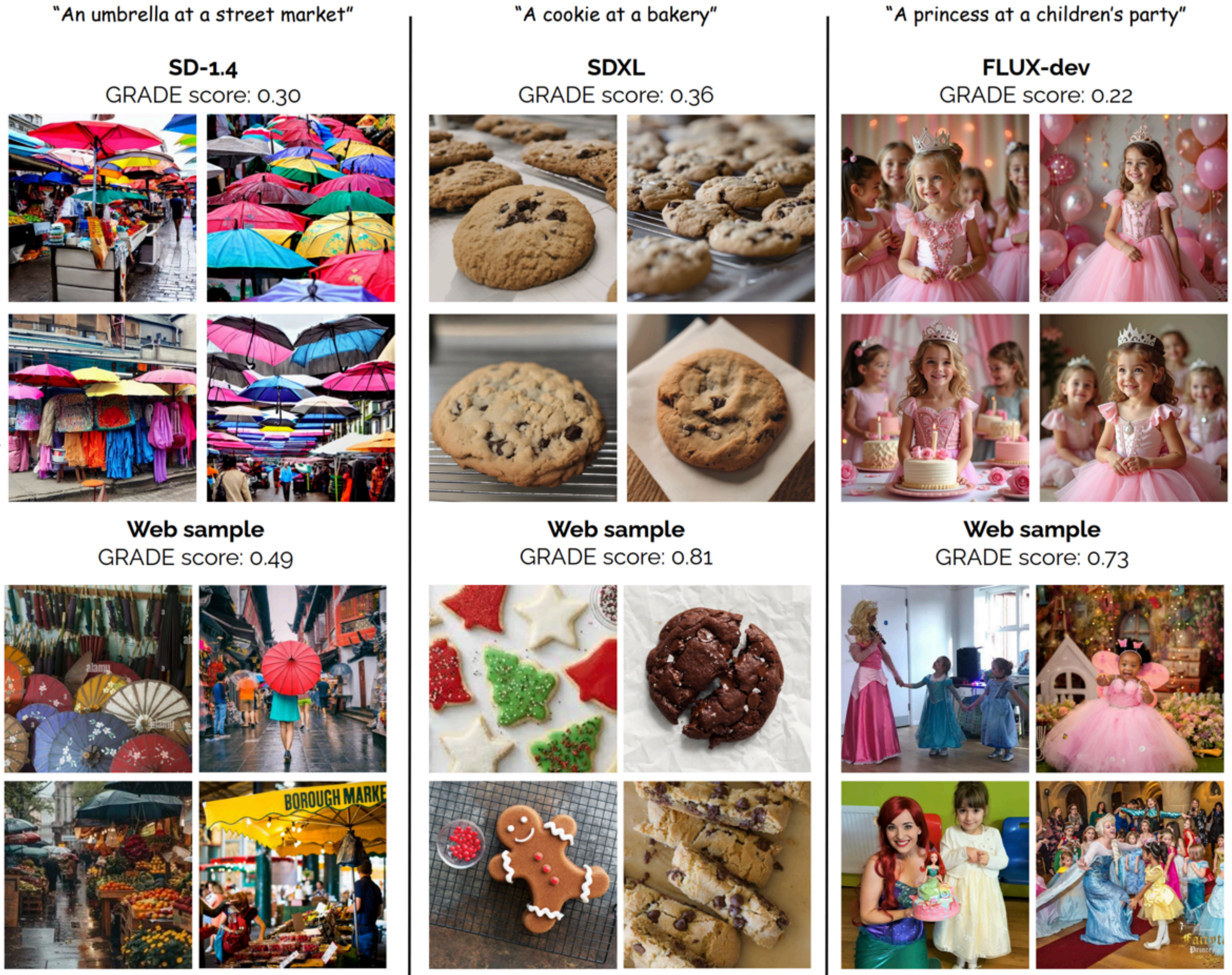
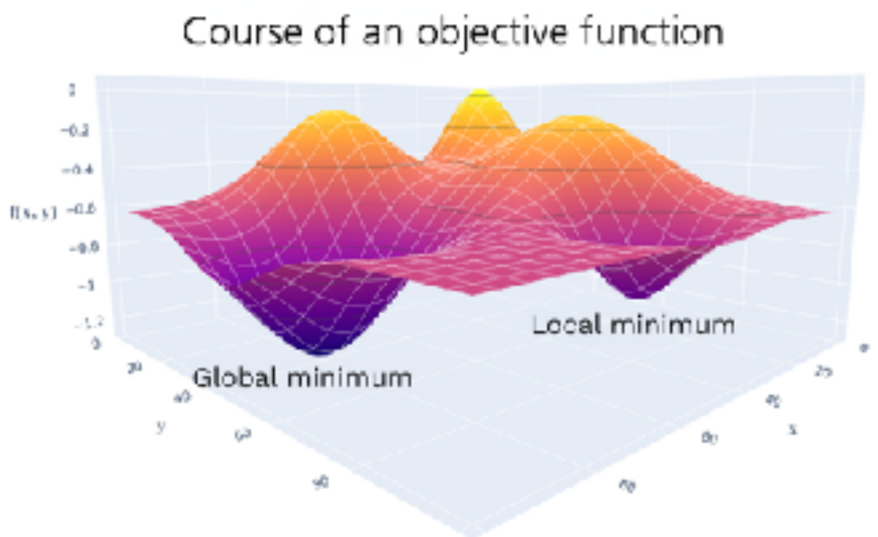


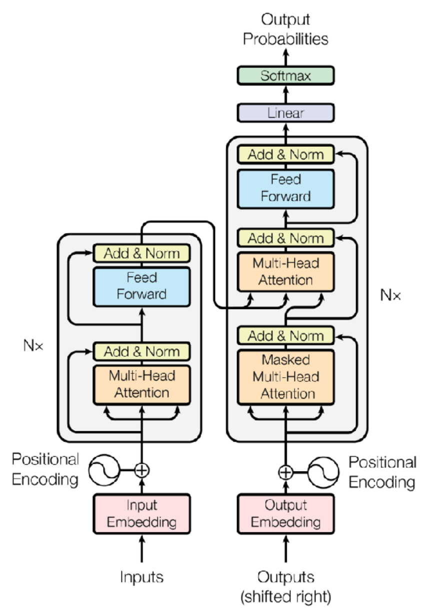
Figure 1. **GRADE scores for T2I generations and corresponding web-search results, for three models and concepts.**

LLMs - Putting It All Together

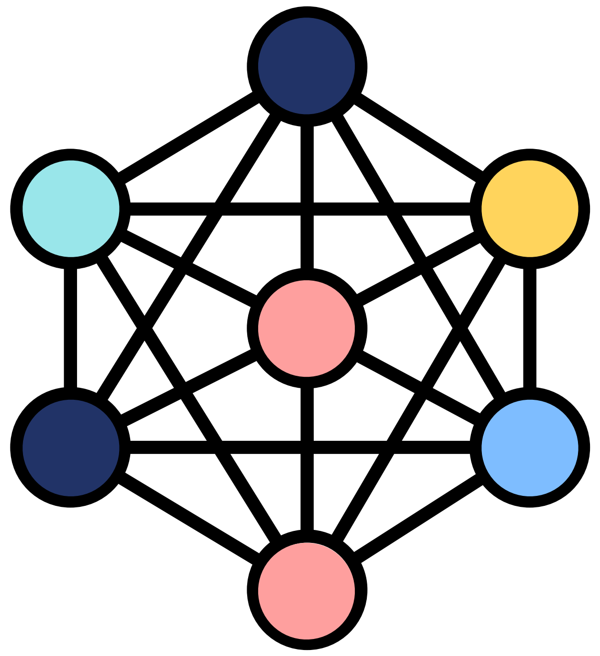


Optimization

Dataset



Architecture

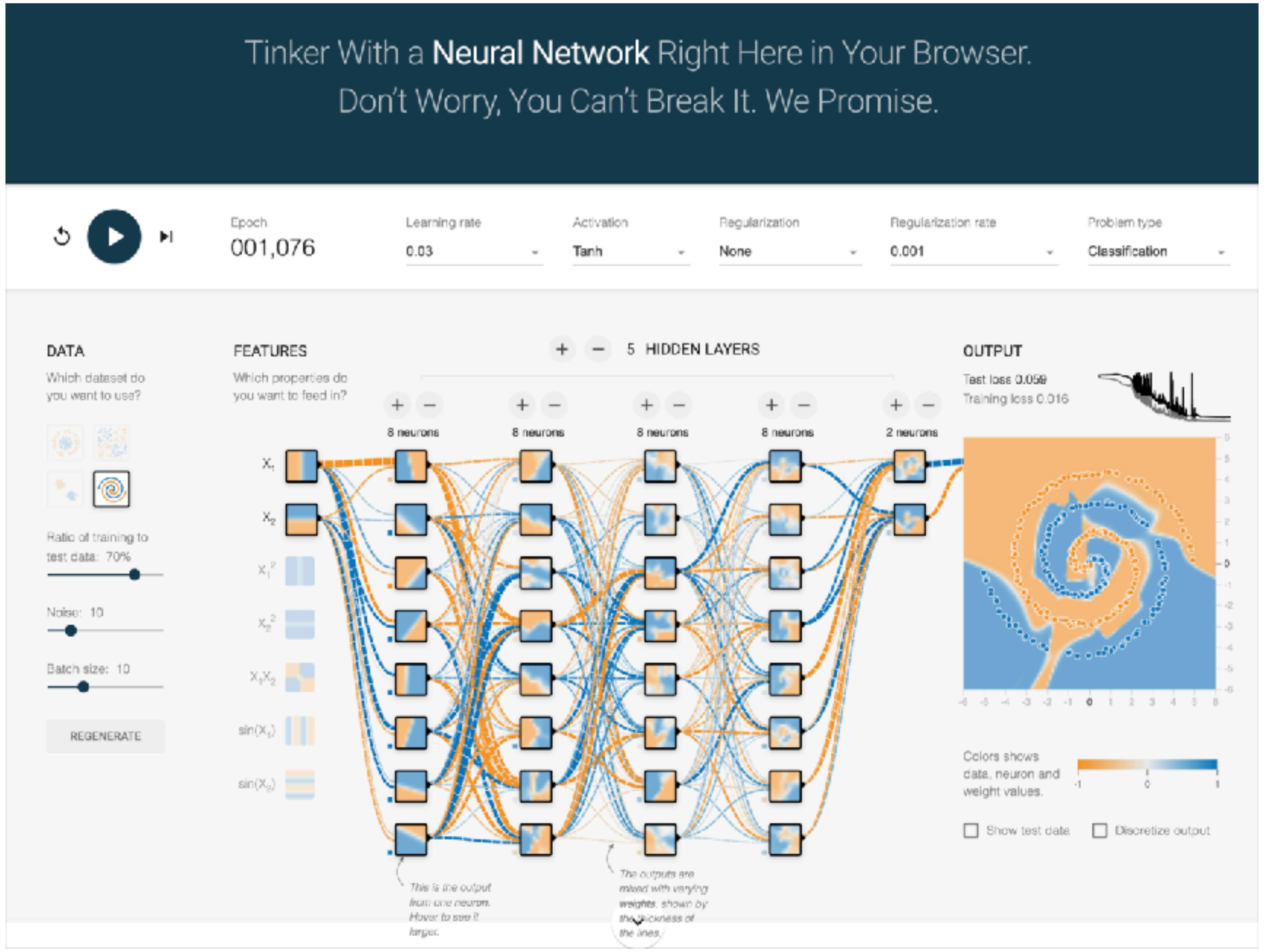


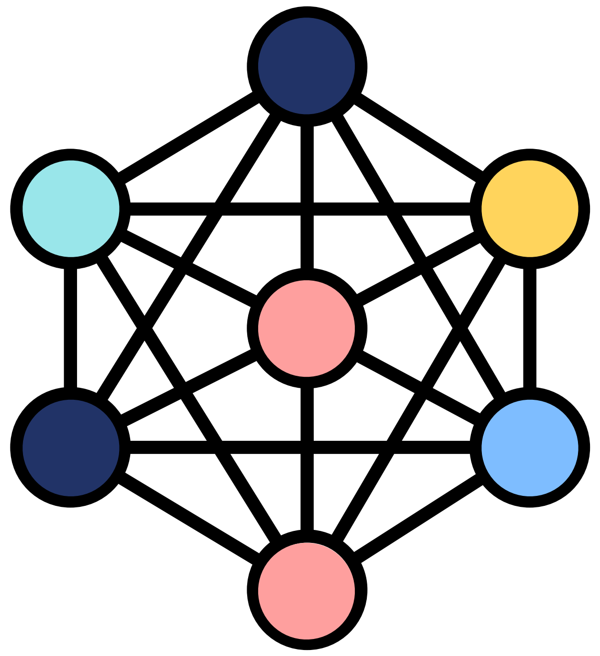
Model

(ChatGPT
Claude
Midjourney)

LLaMA by Meta

LLMs - Putting It All Together





Model

Let's See Some Research

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
- Much documented evidence on biases

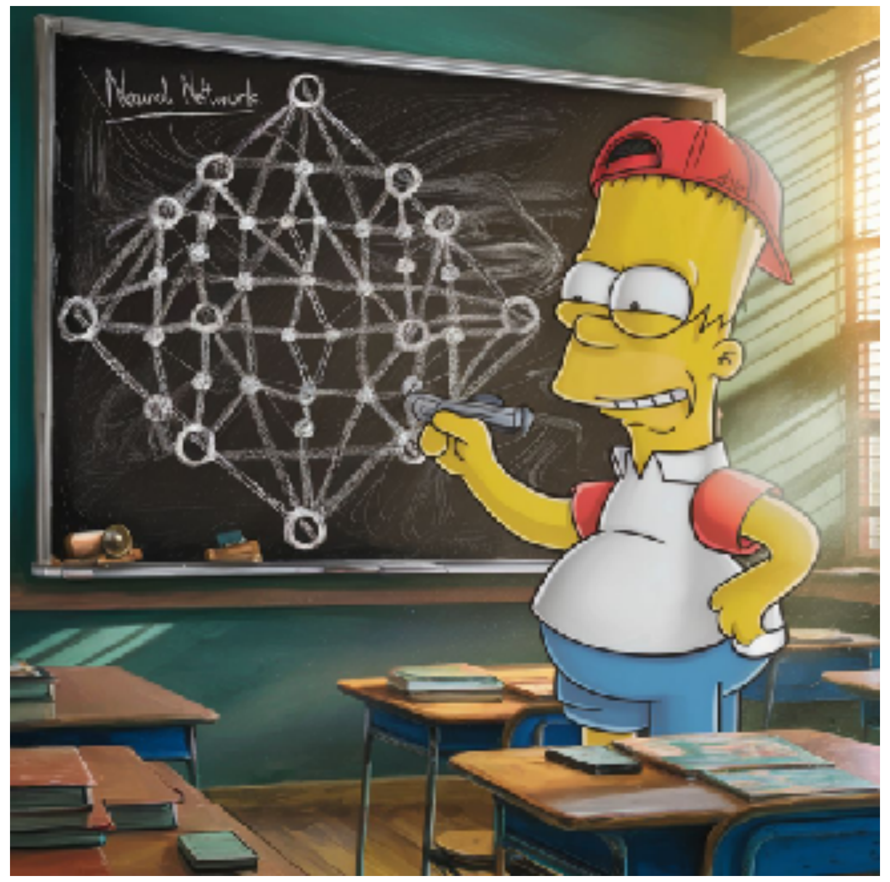
Let's Try It Out!

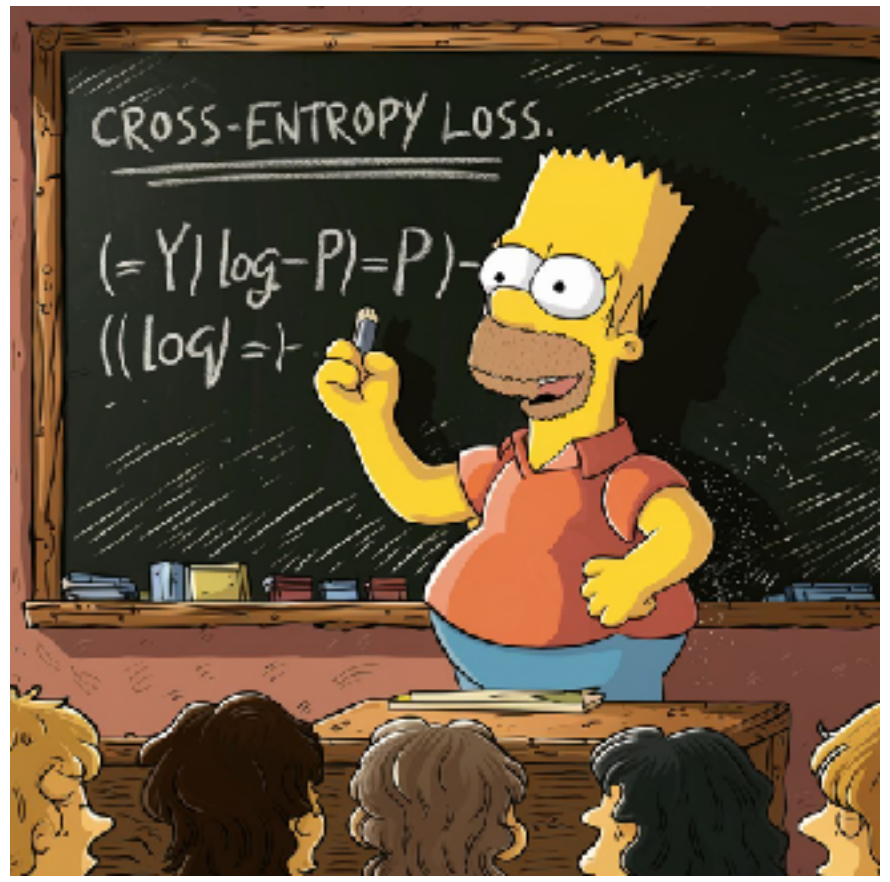
"A photo of a face of an engineer"

1/10 women!

The model is

Where Does The Bias Come From?





Let’s Look At The Data

Where Does The Bias Come From?

5 billion image-caption pairs!

LAION

Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

Where Does The Bias Come From?

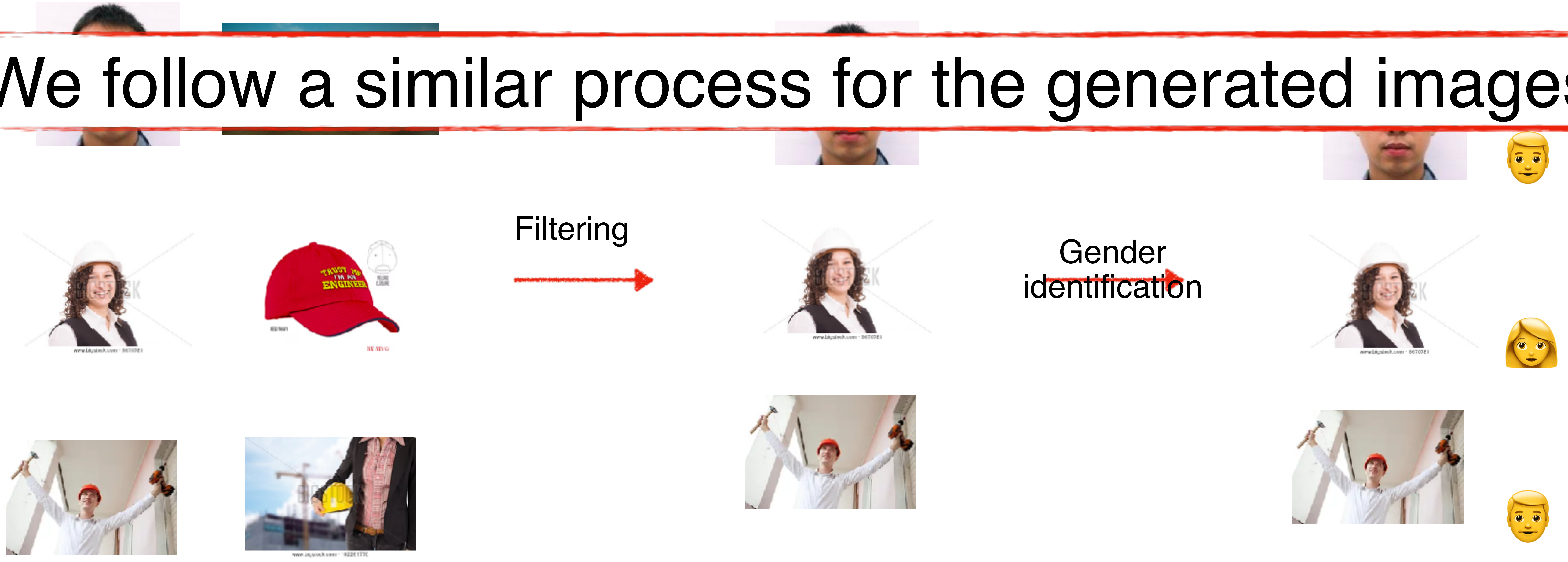
- Using an index (WIMBD), we have fast access to the
- … and we can test such associations in the training data

Establishing Data Gender Ratios

```
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



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 to
- 62 occupations:
  - Accountant
  - Chef

Setup

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

Setup

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

Setup

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





Setup

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

Bias Amplification?

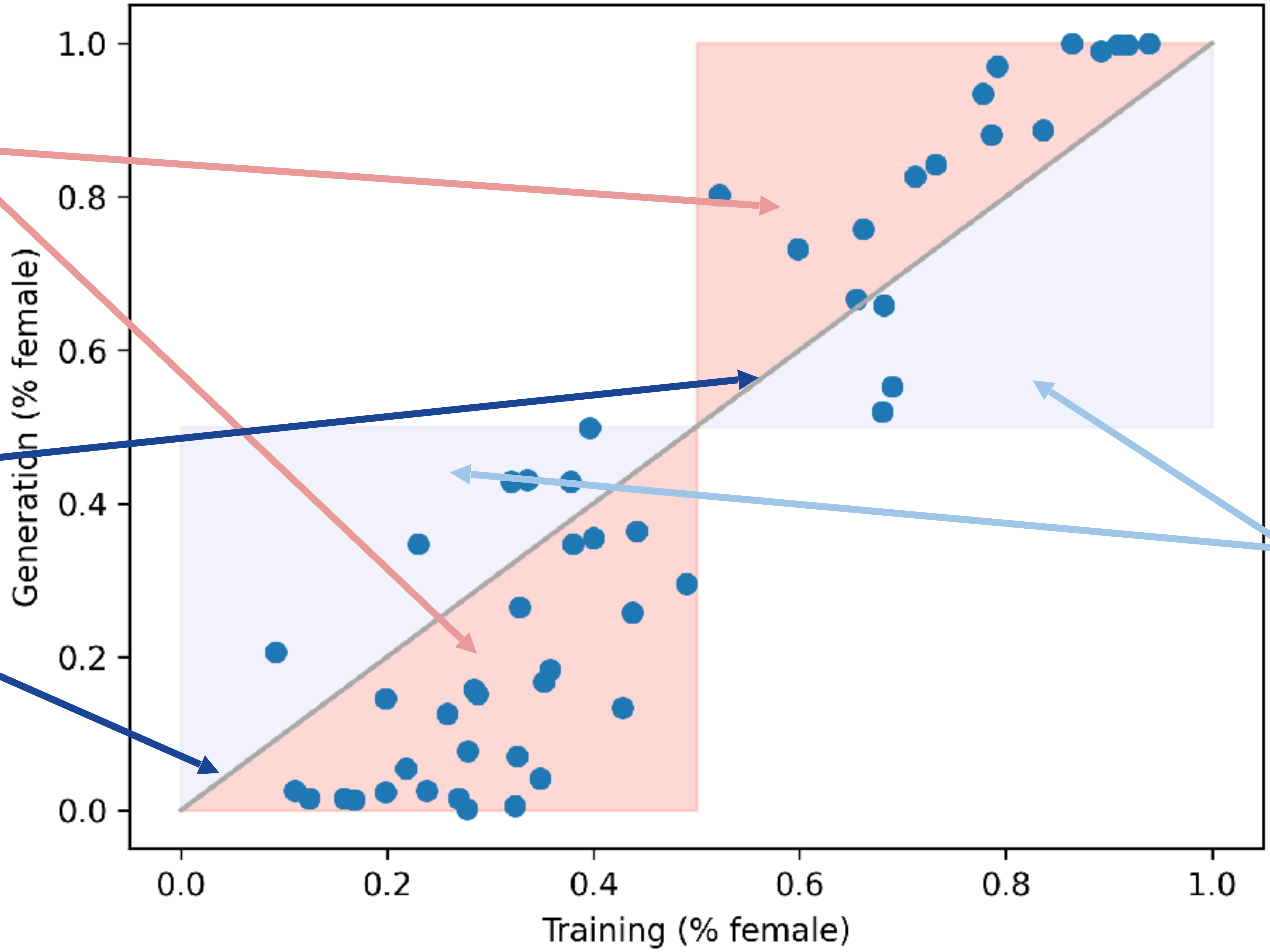
Given the calculated ratios from the data, we can now compare the model's generation to the training data

Bias Amplification?

Peach area:
Bias Amplification

Diagonal:
Bias preservation

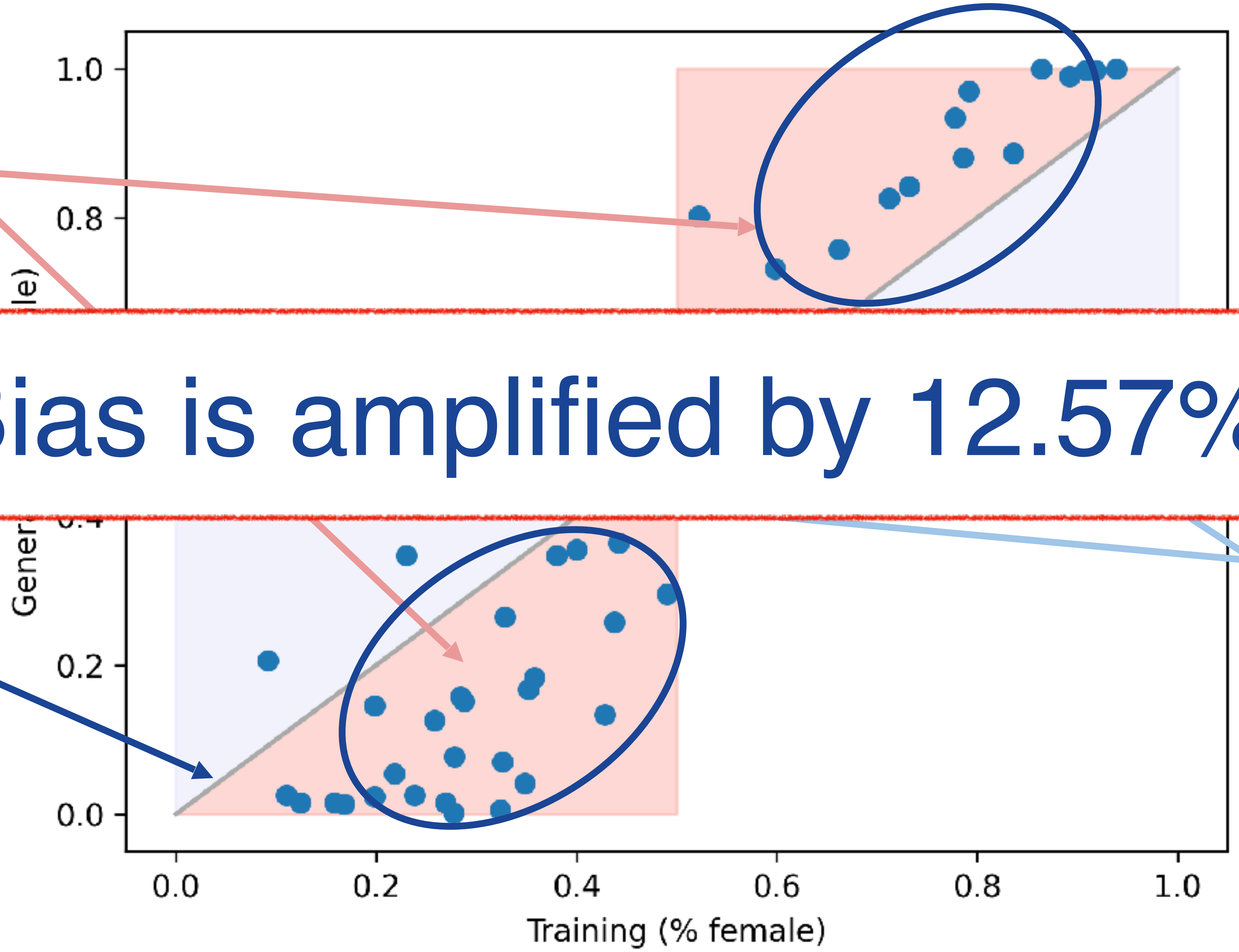
Lavender area:
Bias de-amplification



Bias Amplification!

Peach area:
Bias Amplification

Diagonal:
Bias preservatio



Bias is amplified by 12.57%

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

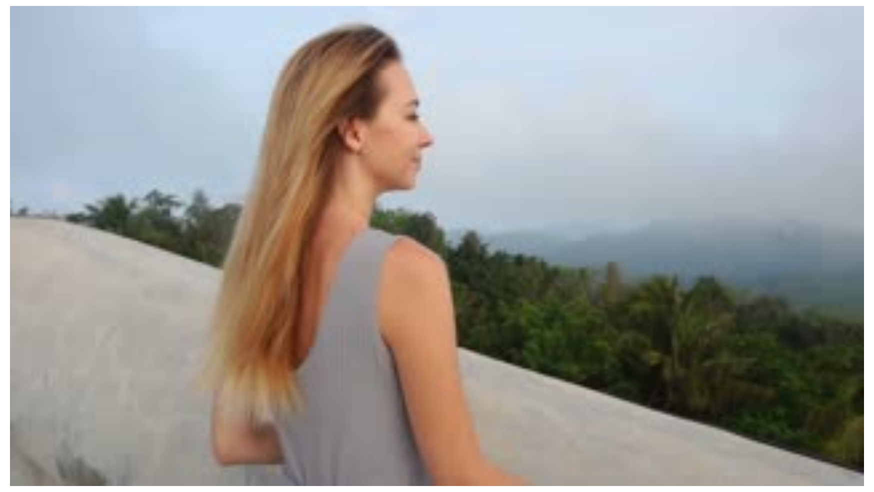
Training Data Investigation

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…

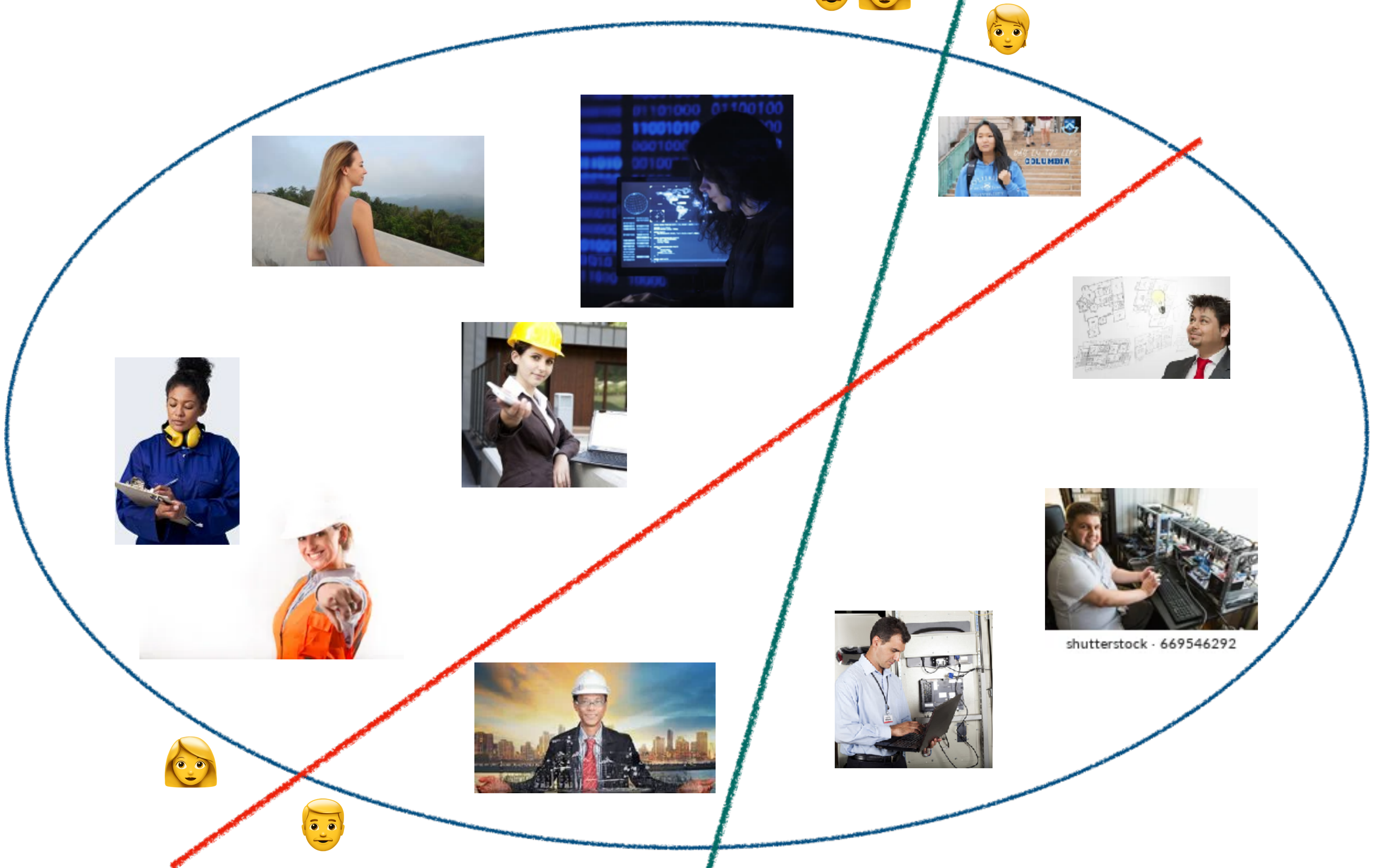


industrial programmer checking computerized machine status

Training Data Investigation

~60% contain gender indicators

Mostly with anti-stereotypical gender (70%)



Training Data Investigation

~60% contain gender indicators

Test data

"A photo of a face of an engineer"

Mostly with anti-stereotypical gender (70%)

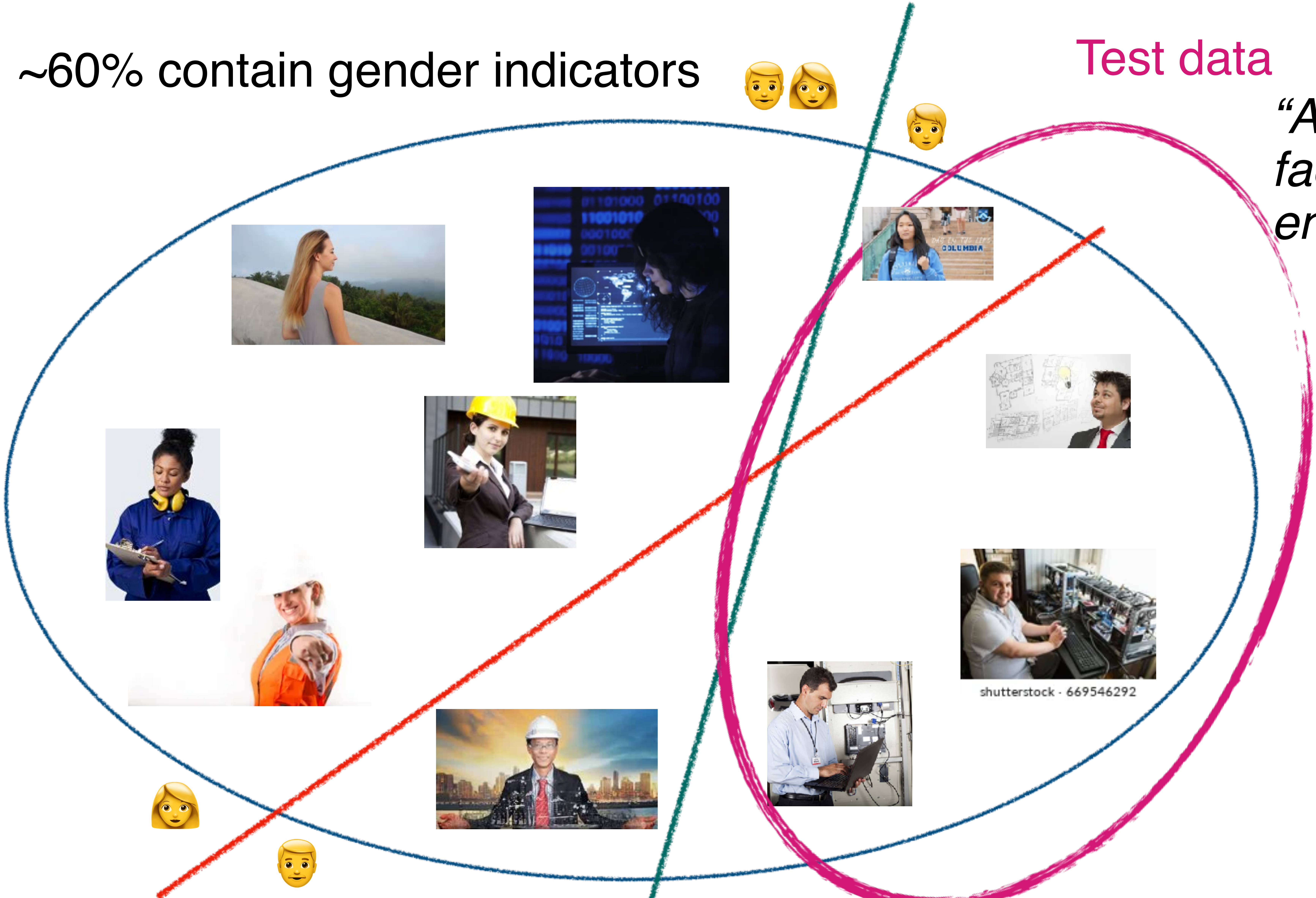
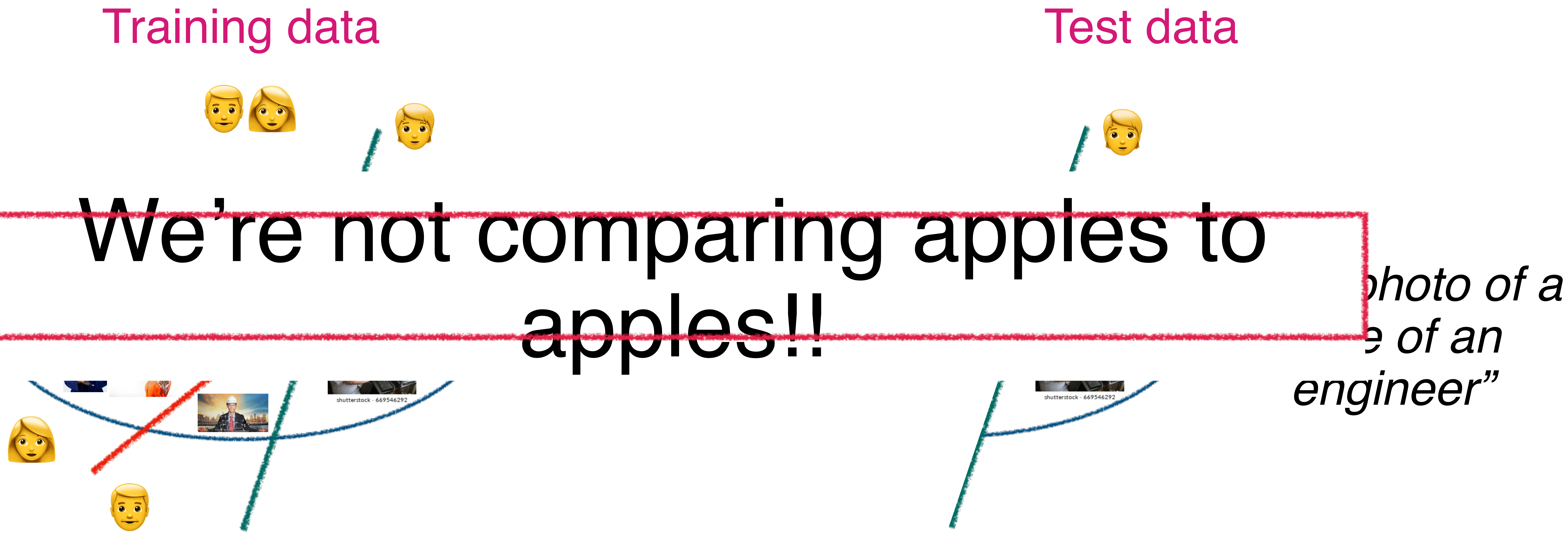


Image Captions & Prompts Mismatch

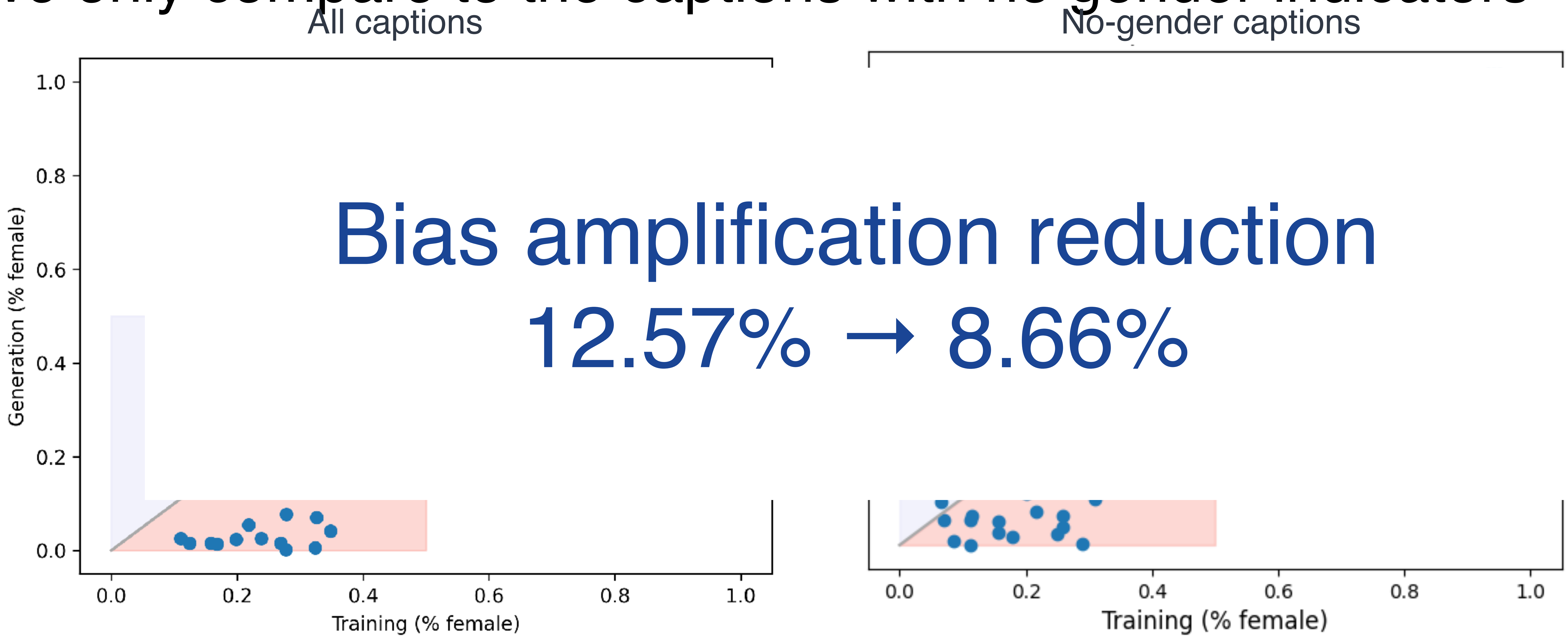


We’re not comparing apples to apples!!

Matching Distributions

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

- We only compare to the captions with no gender indicators



Bias amplification reduction

12.57% → 8.66%

One Mismatch

What about others?

Image Captions & Prompts Mismatch #2

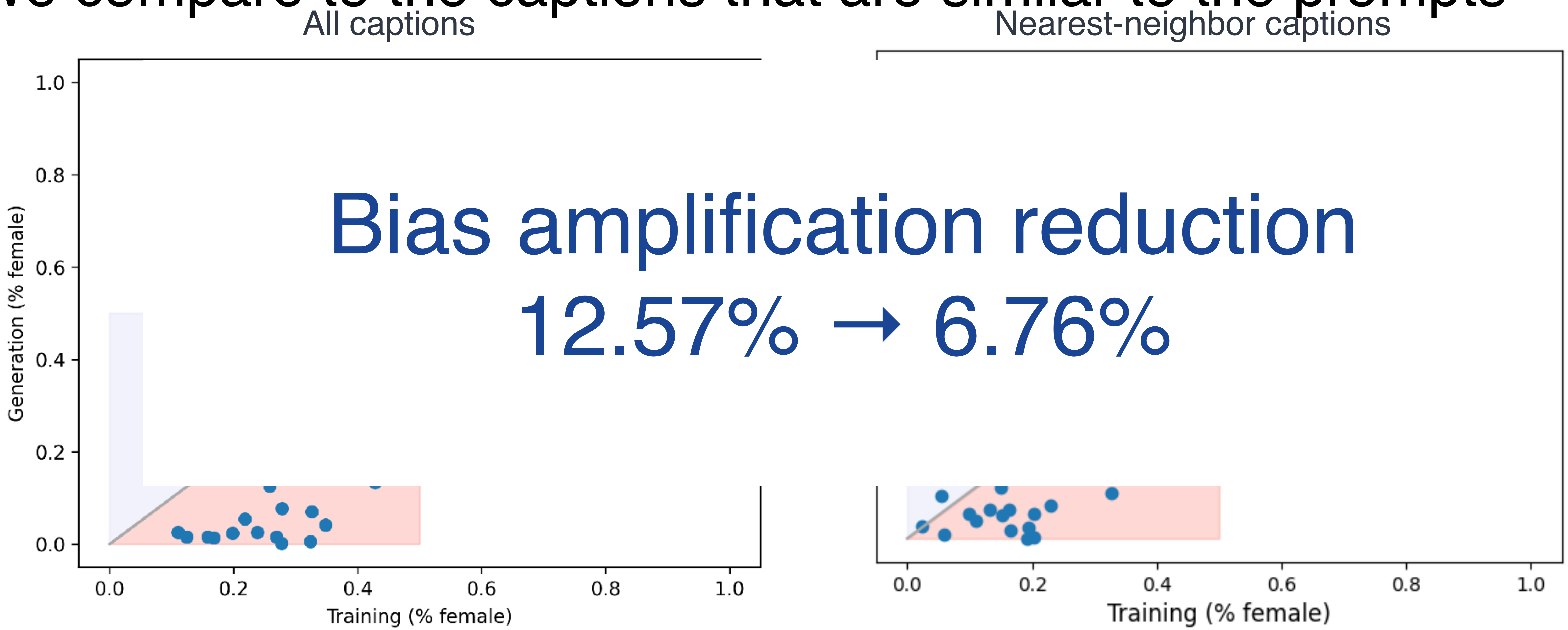
We also found a

(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:

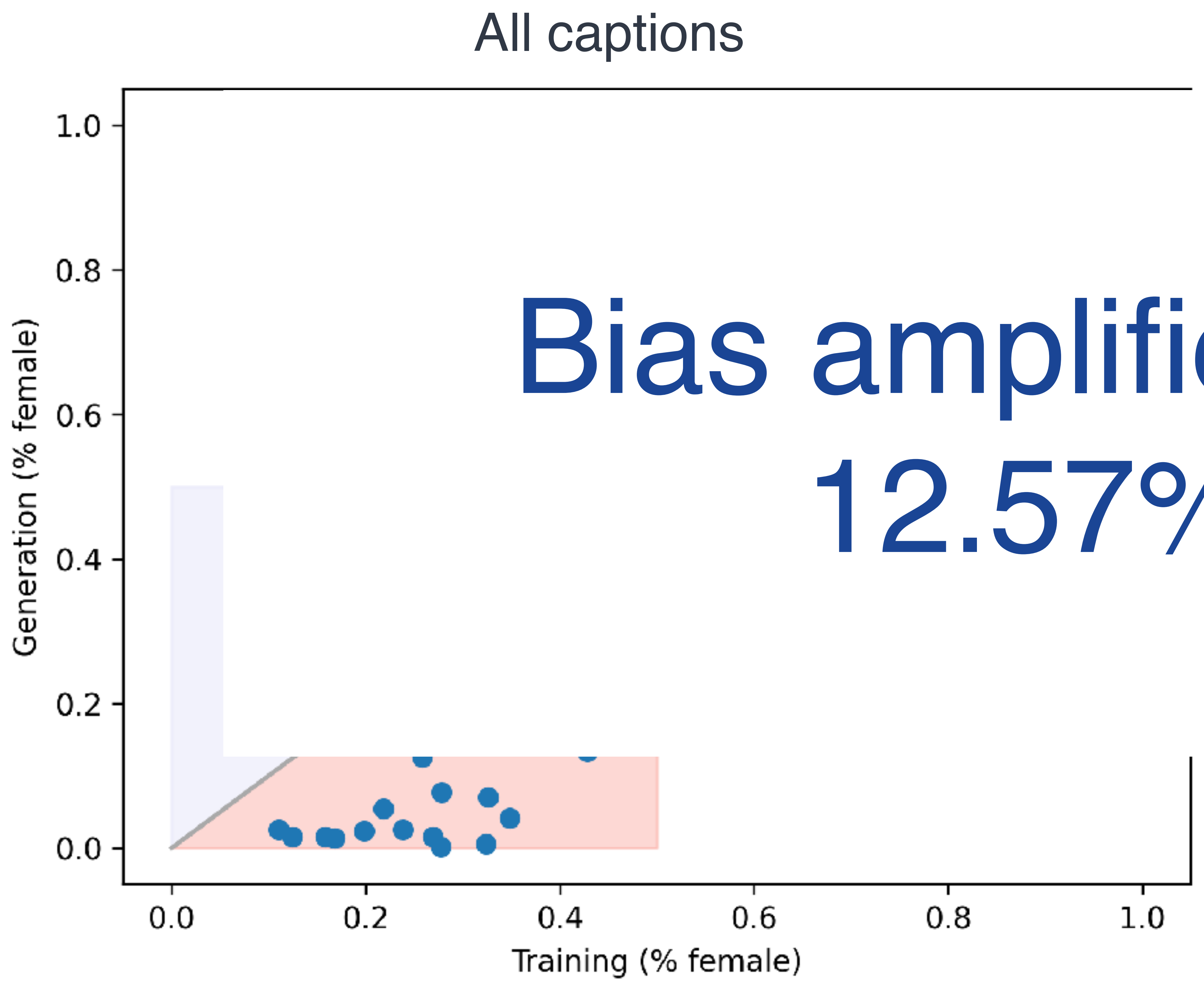
- We compare to the captions that are similar to the prompts

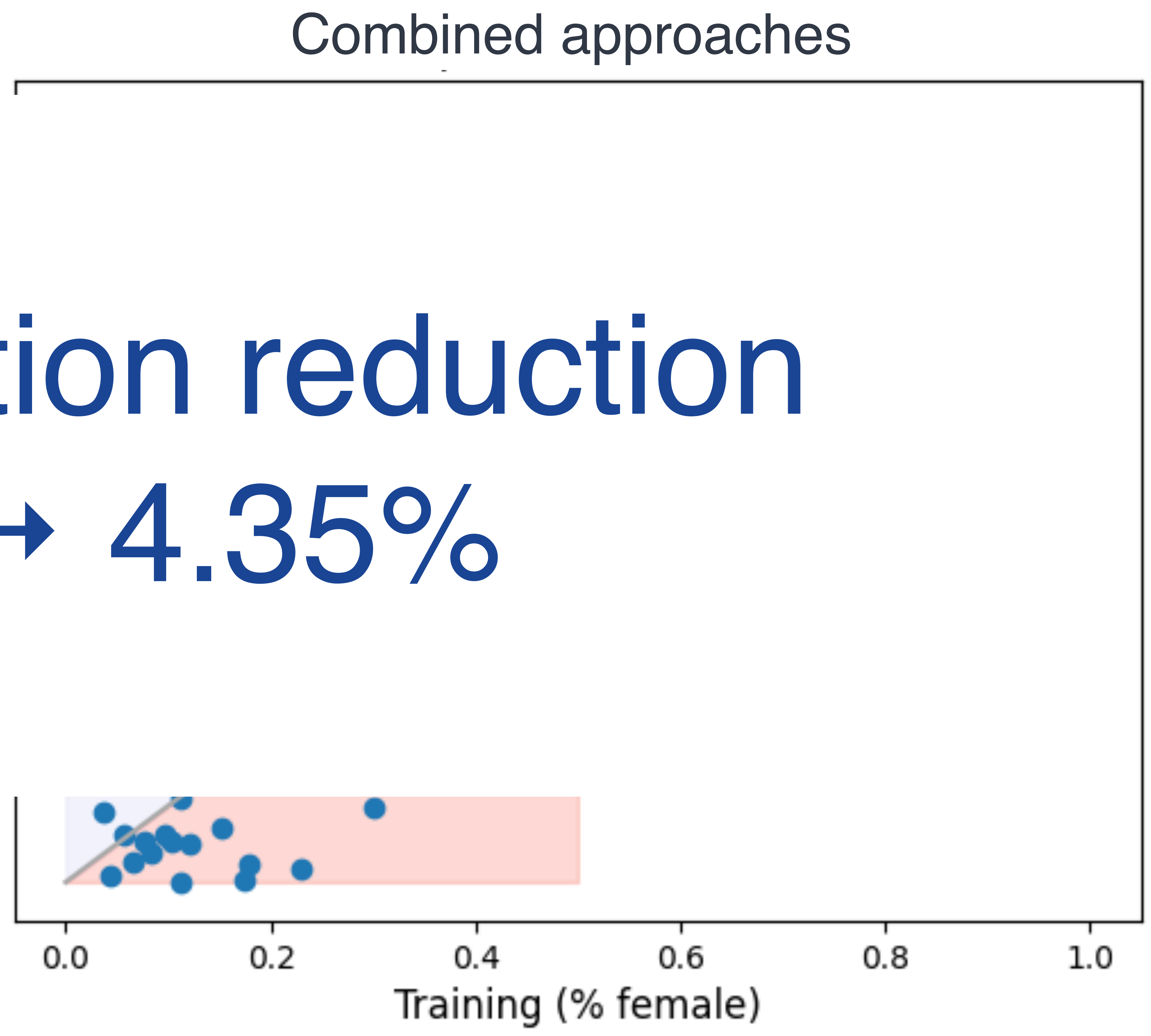


Bias amplification reduction
12.57% → 6.76%

Matching Distributions: Combined

Finally, we combine both approaches





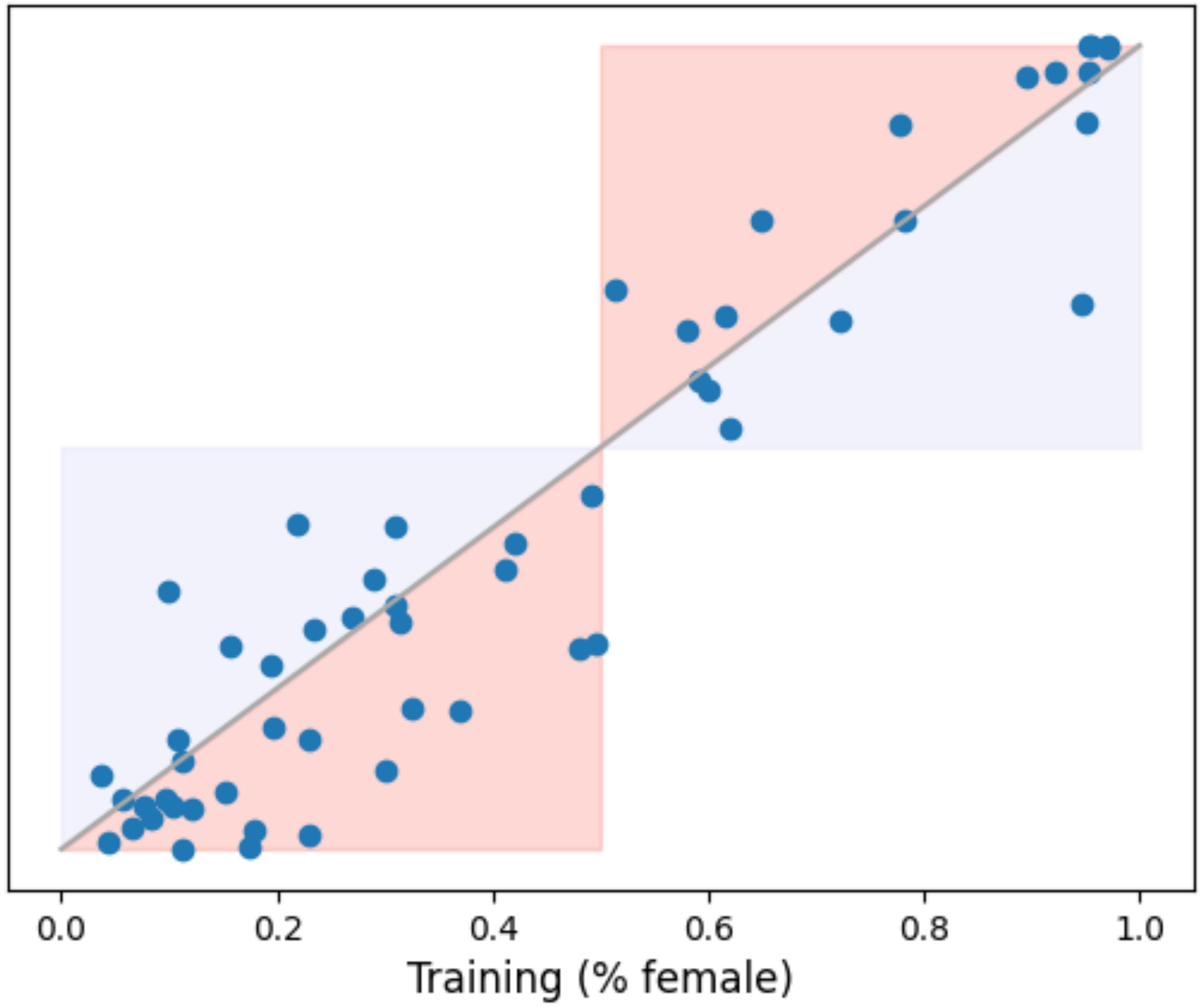
Bias amplification reduction

12.57% → 4.35%

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
- Data dictiates model behavior



Imitation

✅

→ Leonardo DiCaprio

❌

→ Yanai Elazar

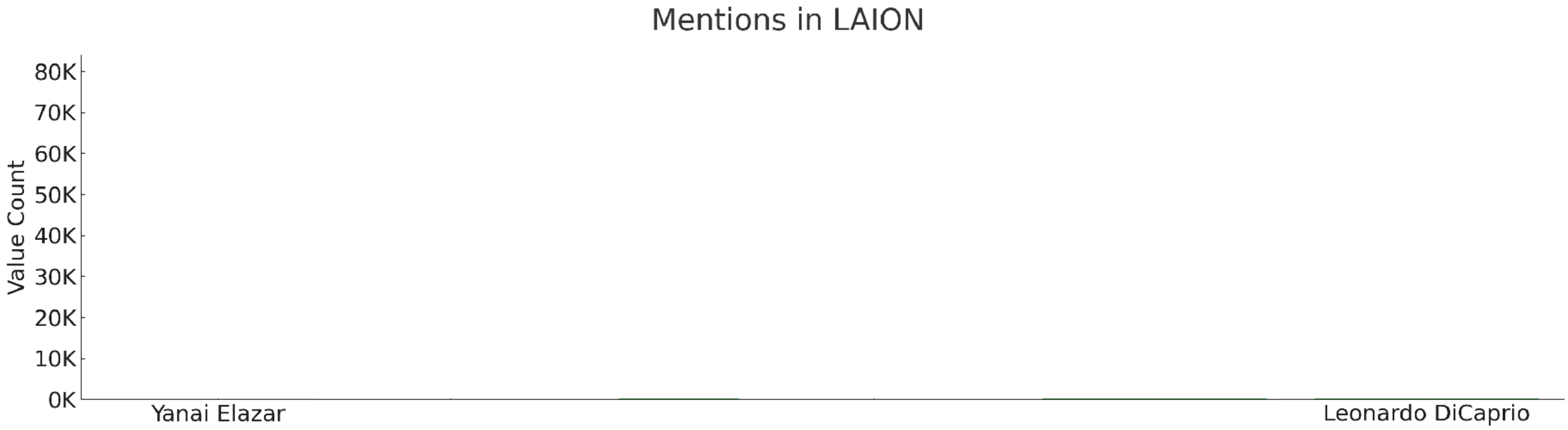
Imitation

Leonardo DiCaprio

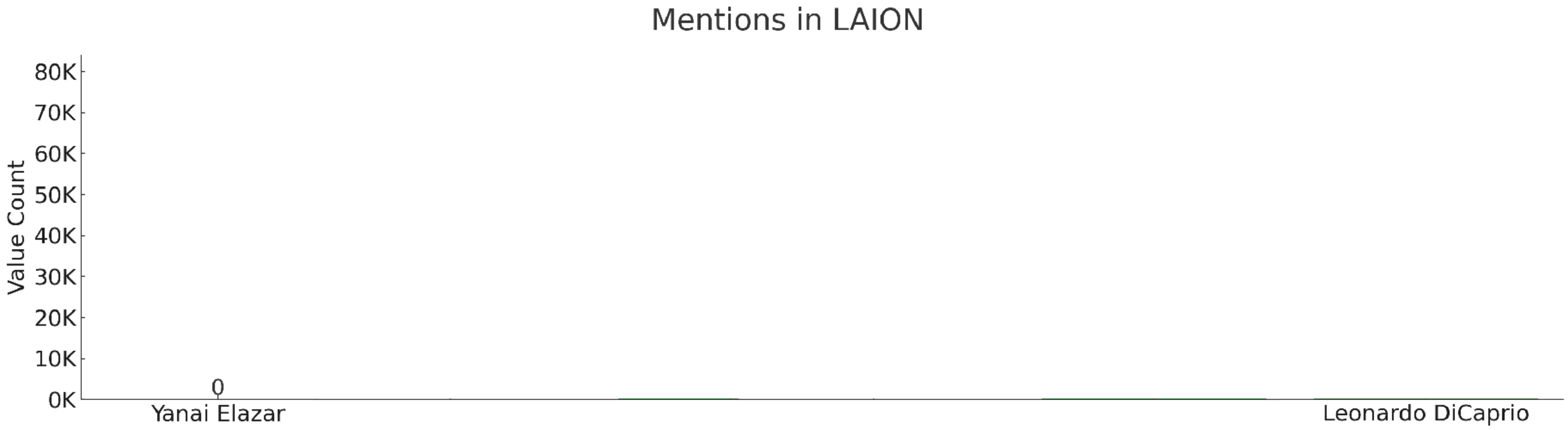
Spot the difference

Yanai Elazar

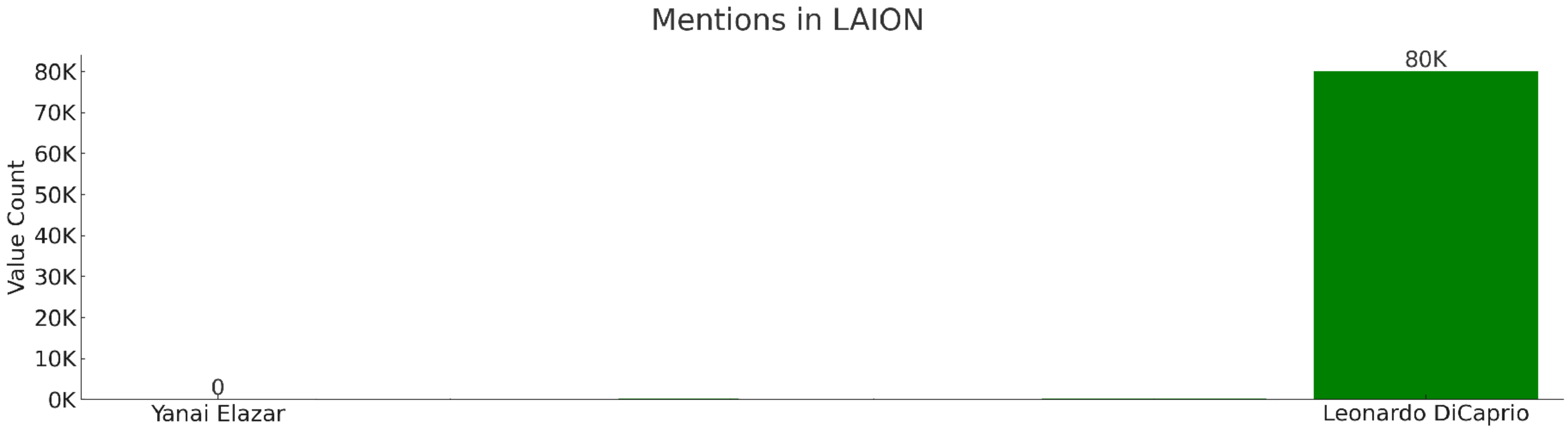
Imitation Threshold?



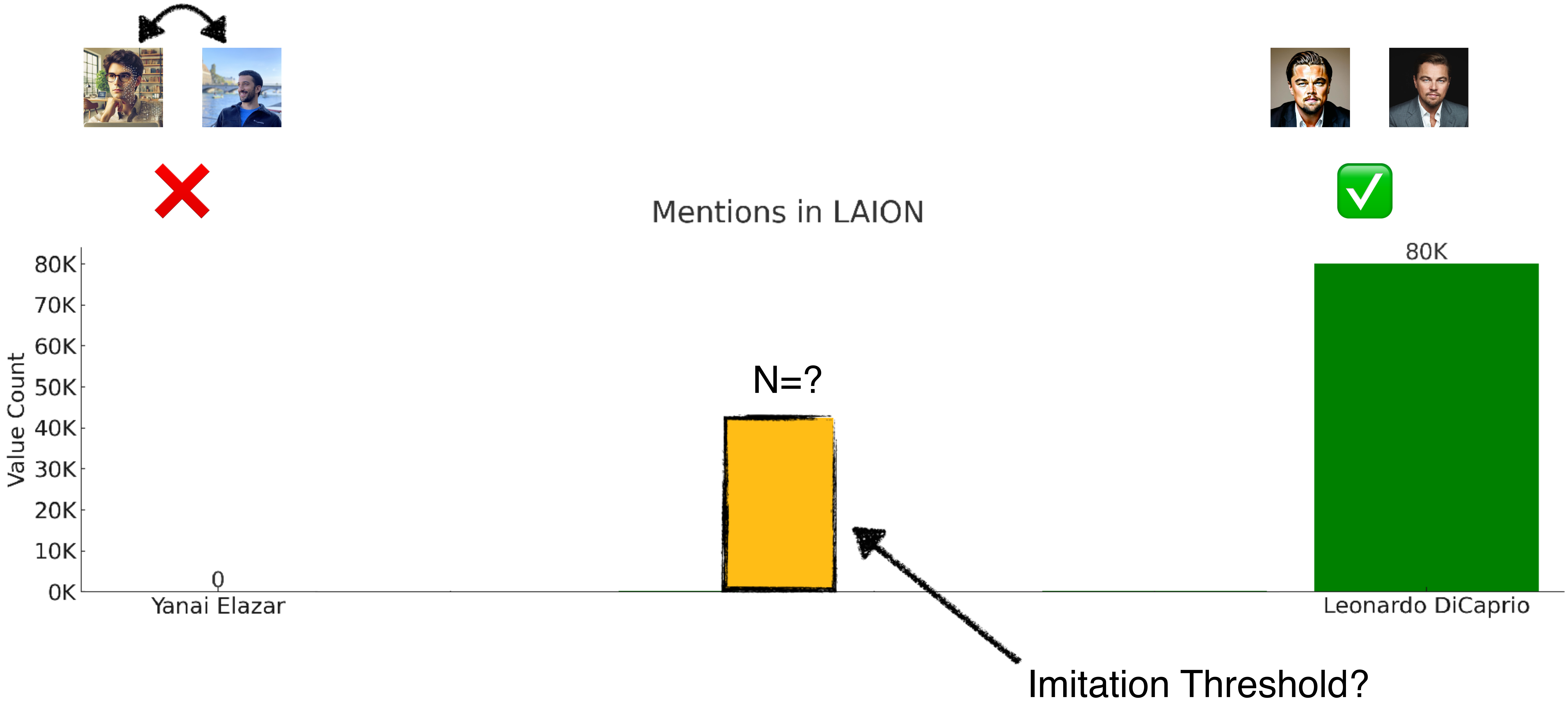
Imitation Threshold?



Imitation Threshold?



Imitation Threshold?



Imitation - Why Should You Care?

- Copyrights

Imitation - Why Should You Care?

- Copyrights

Issues Jobs **VentureBeat**

ce ⌄ | Security ⌄ | Data Infrastructure ⌄ | Automation ⌄ | Enterprise An

The copyright case against AI art generators just got stronger with more artists and evidence

Credit: VentureBeat made with OpenAI DALL-E 3 via ChatGPT

Imitation - Why Should You Care?

- Copyrights
- Privacy

Leonardo DiCaprio

Celebrity

Yanai Elazar

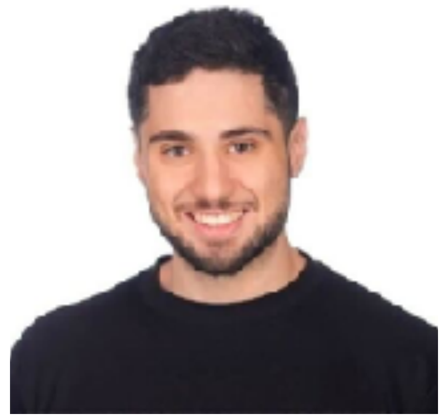
Private individual

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

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

LAION-5B'

Count: *100*

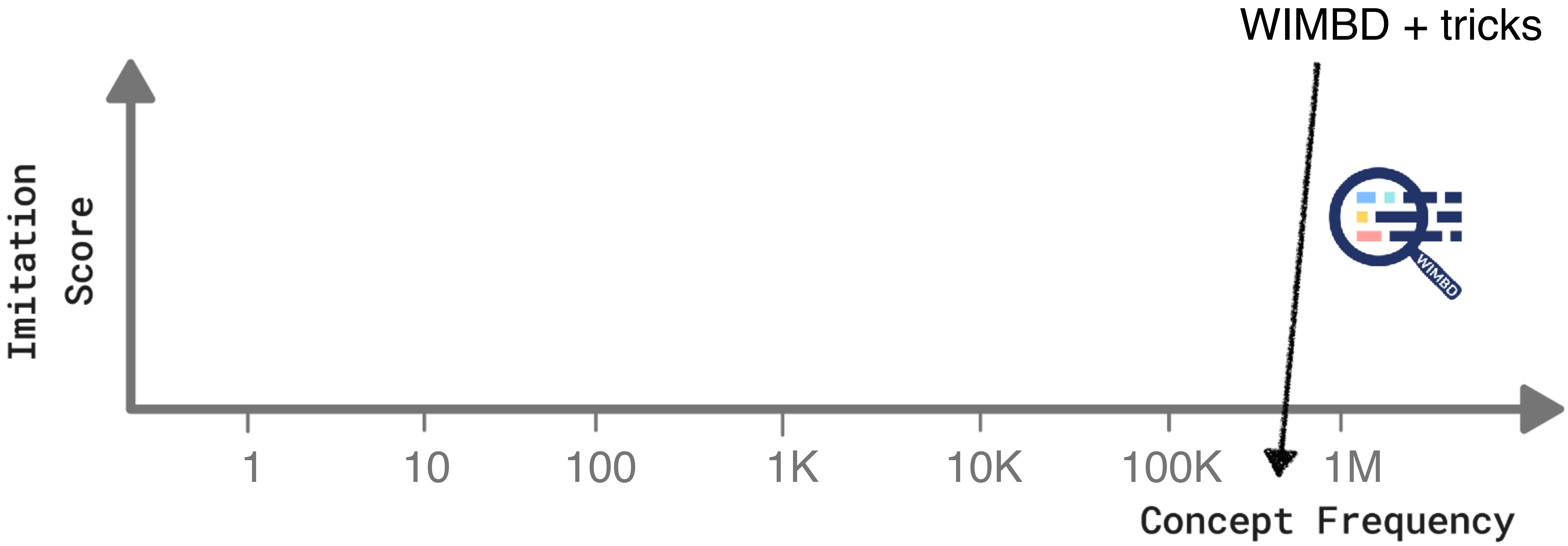
LAION-5B

Count: 80*K*

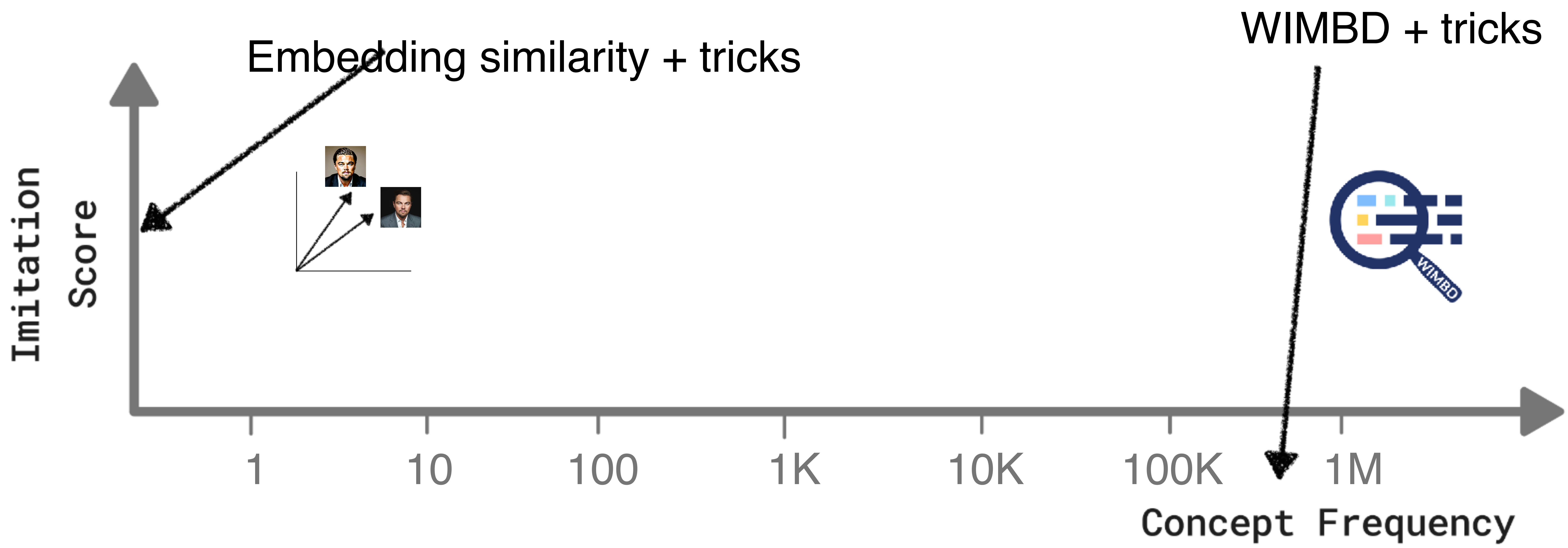
Solutions

1. Counterfactual model

Solutions

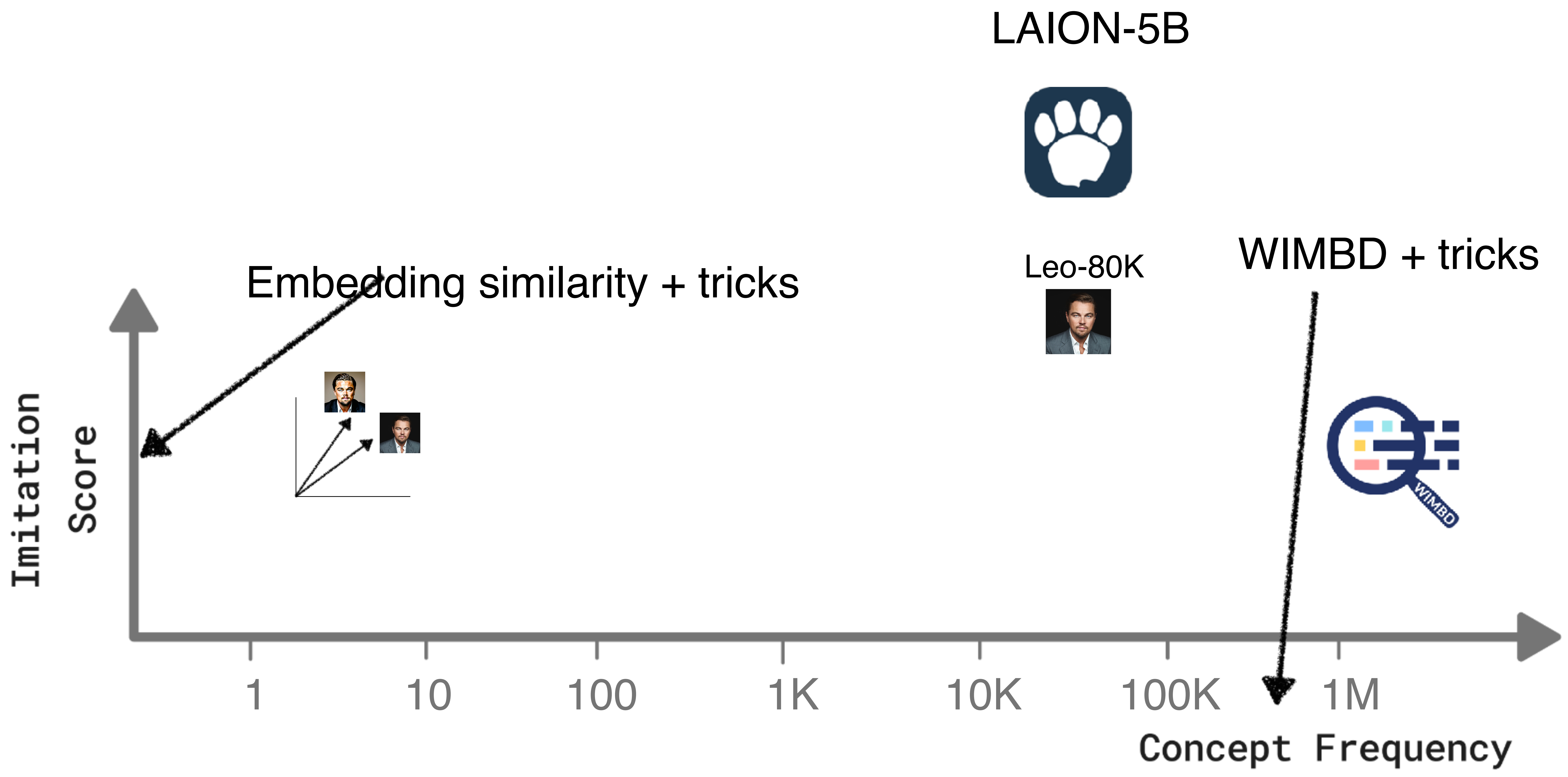

WIMBD + tricks
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency
WIMBD

Solutions



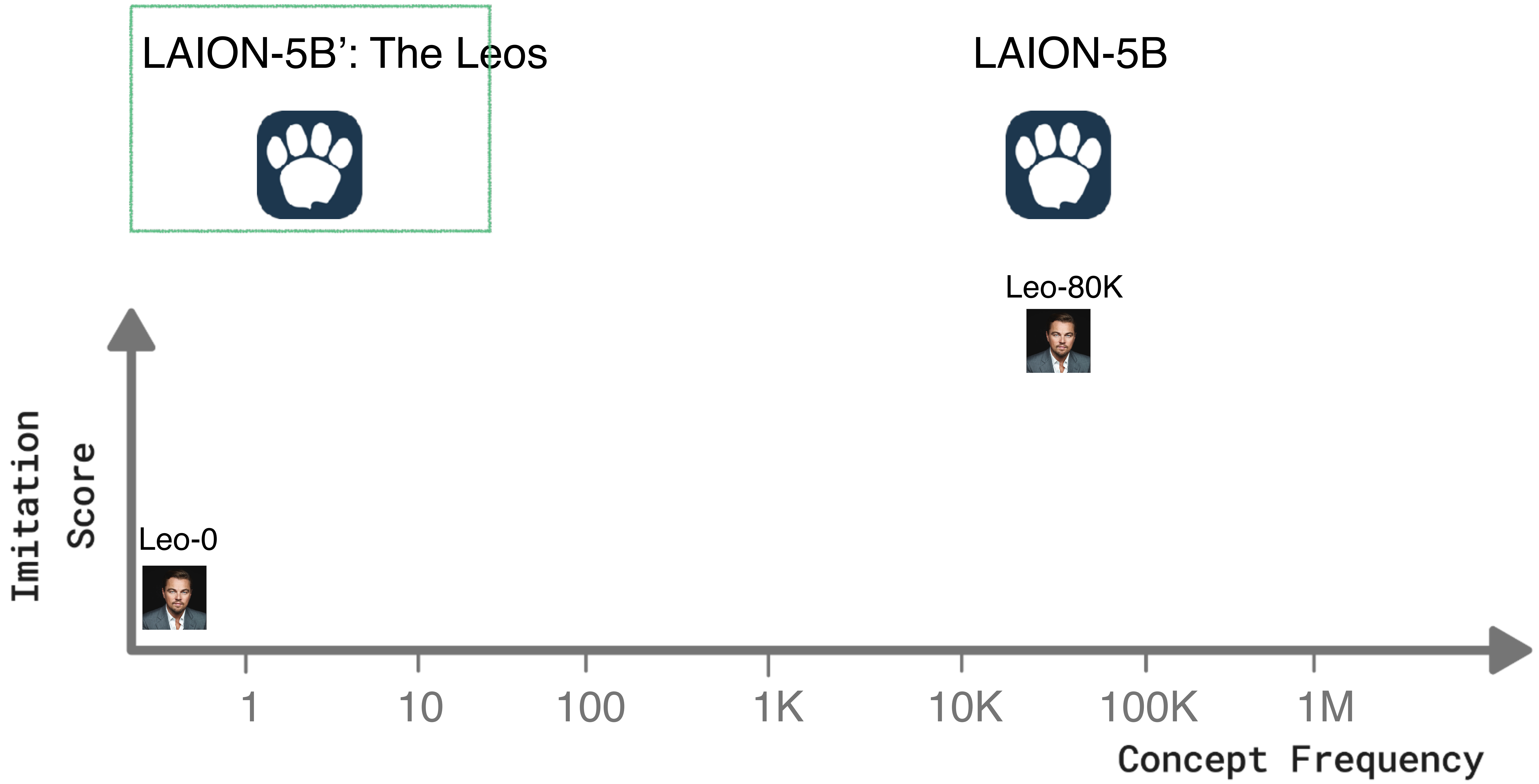
WIMBD + tricks
Embedding similarity + tricks
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solutions



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

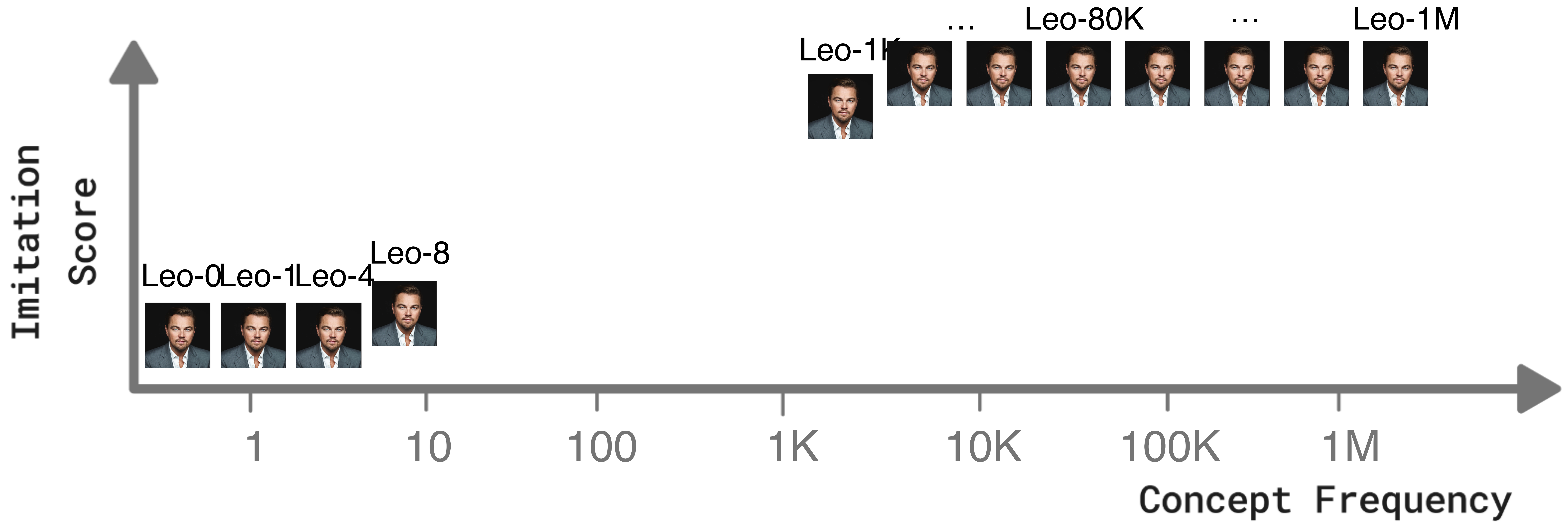
Solution #1


LAION-5B’: The Leos
LAION-5B
Leo-80K
Leo-0
Imitation Score
1
10
100
1K
10K
100K
1M
Concept Frequency

Solution #1

LAION-5B’: The Leos

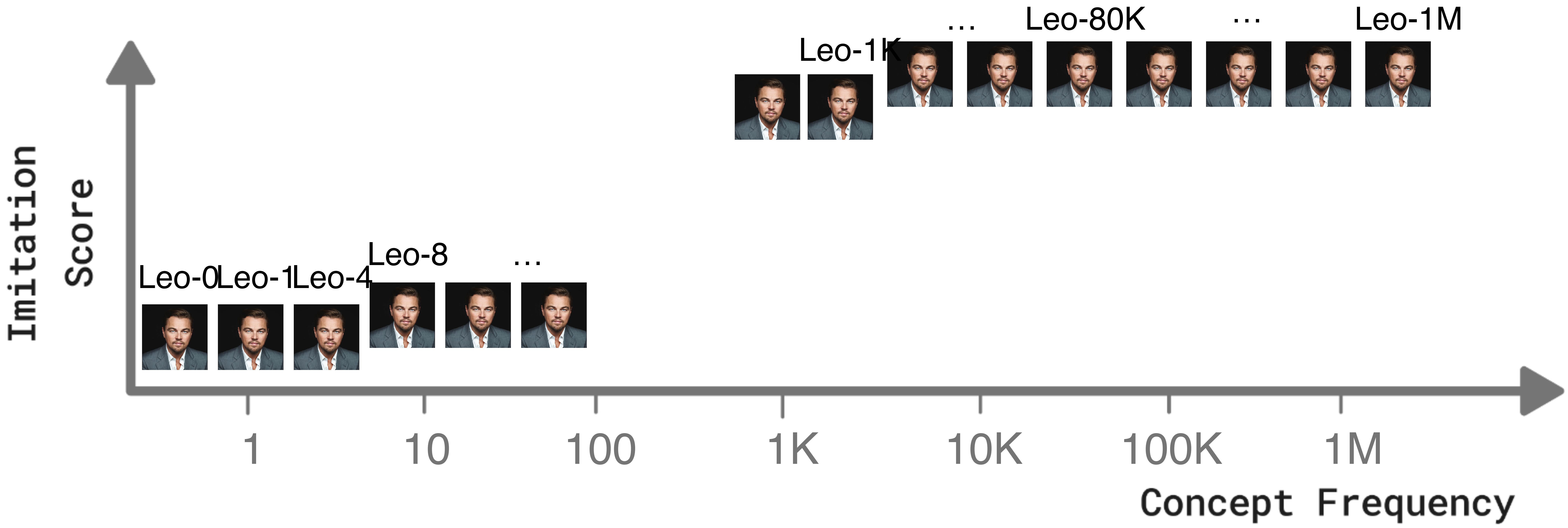
LAION-5B



Solution #1

LAION-5B': The Leos

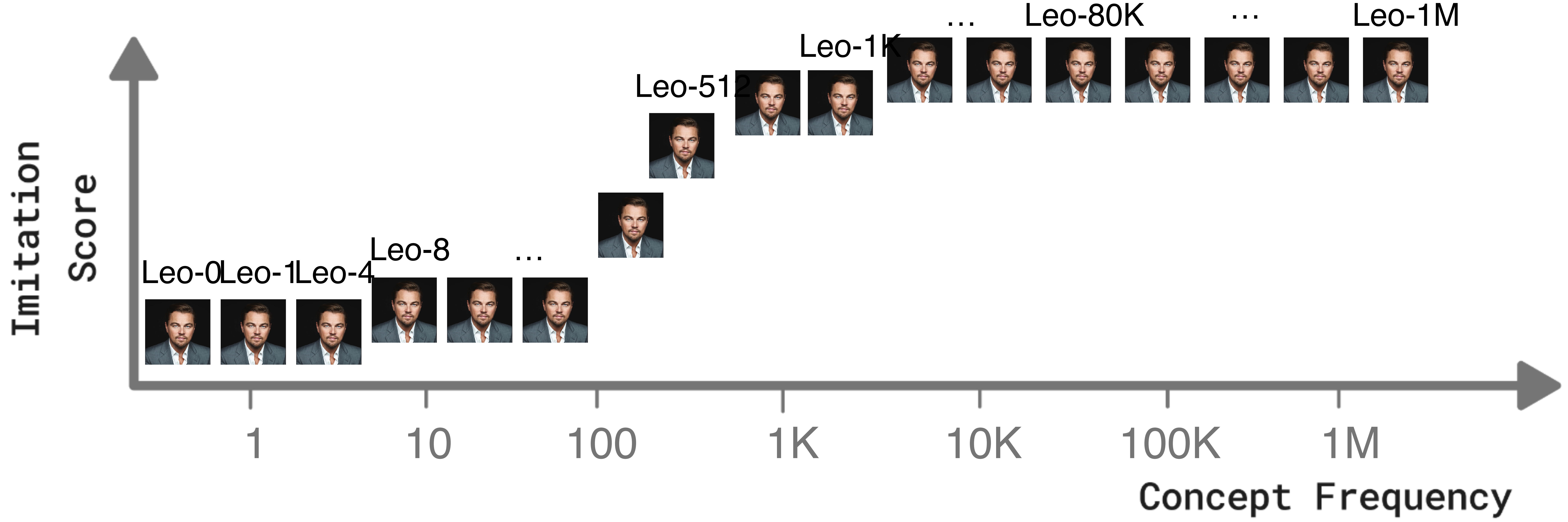
LAION-5B



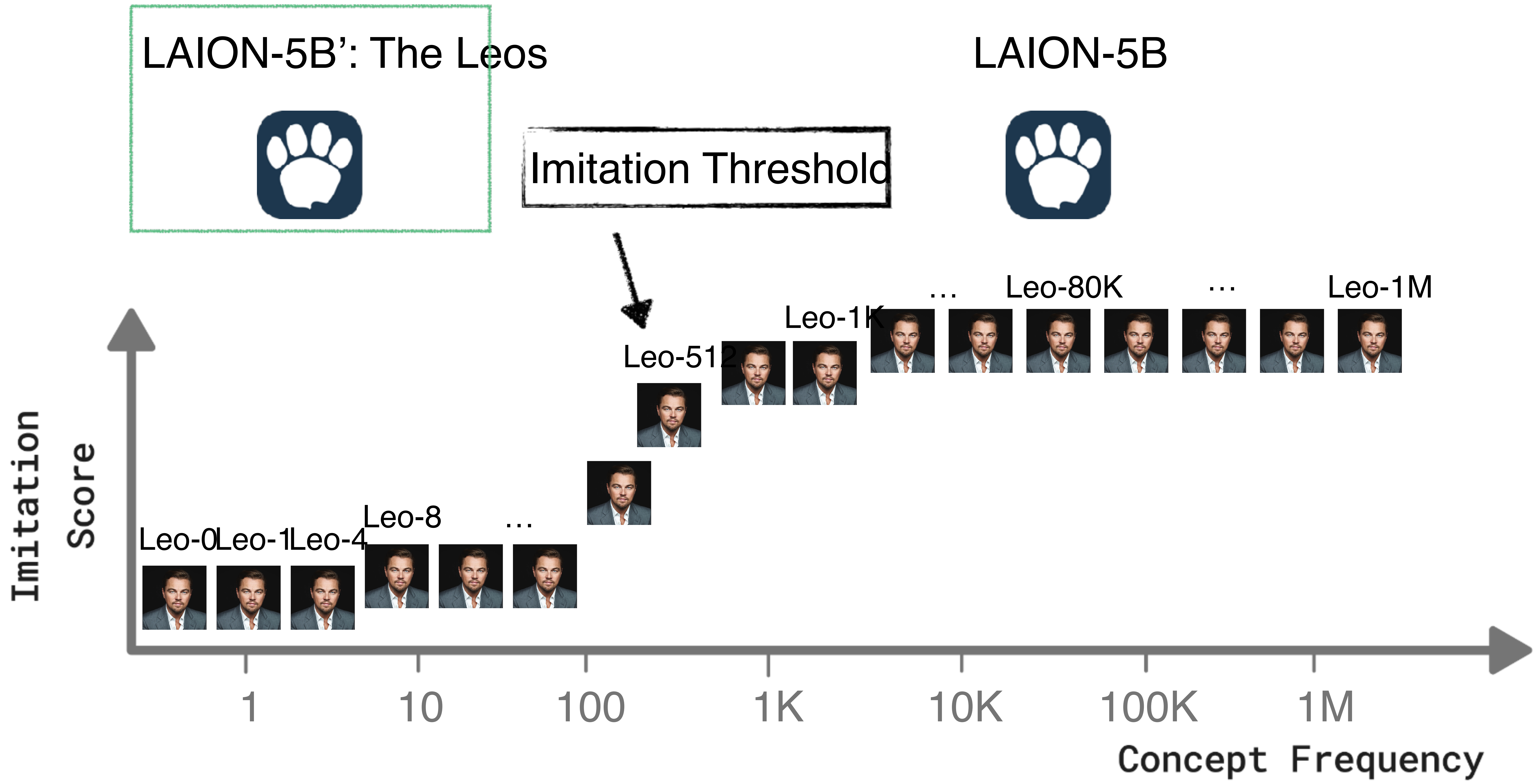
Solution #1

LAION-5B’: The Leos

LAION-5B



Solution #1



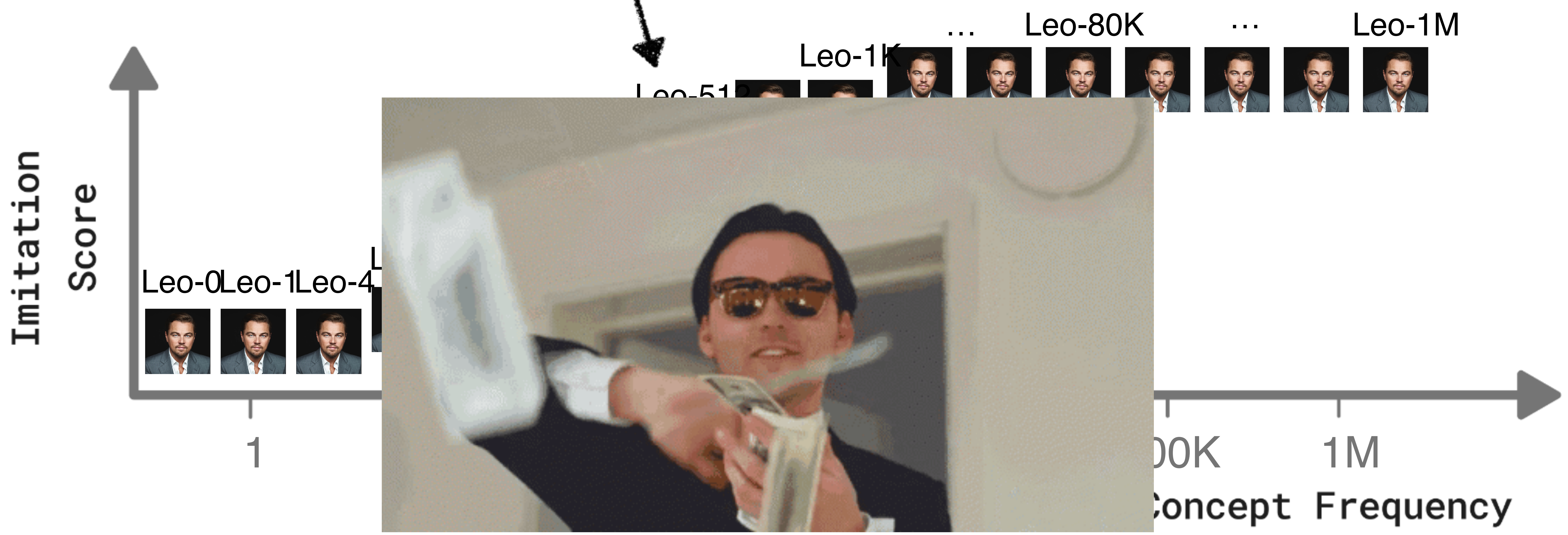
LAION-5B': The Leos
LAION-5B
Imitation Threshold
Imitation Score
Concept Frequency
Leo-0
Leo-1
Leo-4
Leo-8
…
Leo-512
Leo-1K
…
Leo-80K
…
Leo-1M
1
10
100
1K
10K
100K
1M

Solution #1

LAION-5B’: The Leos

Imitation Threshold

LAION-5B



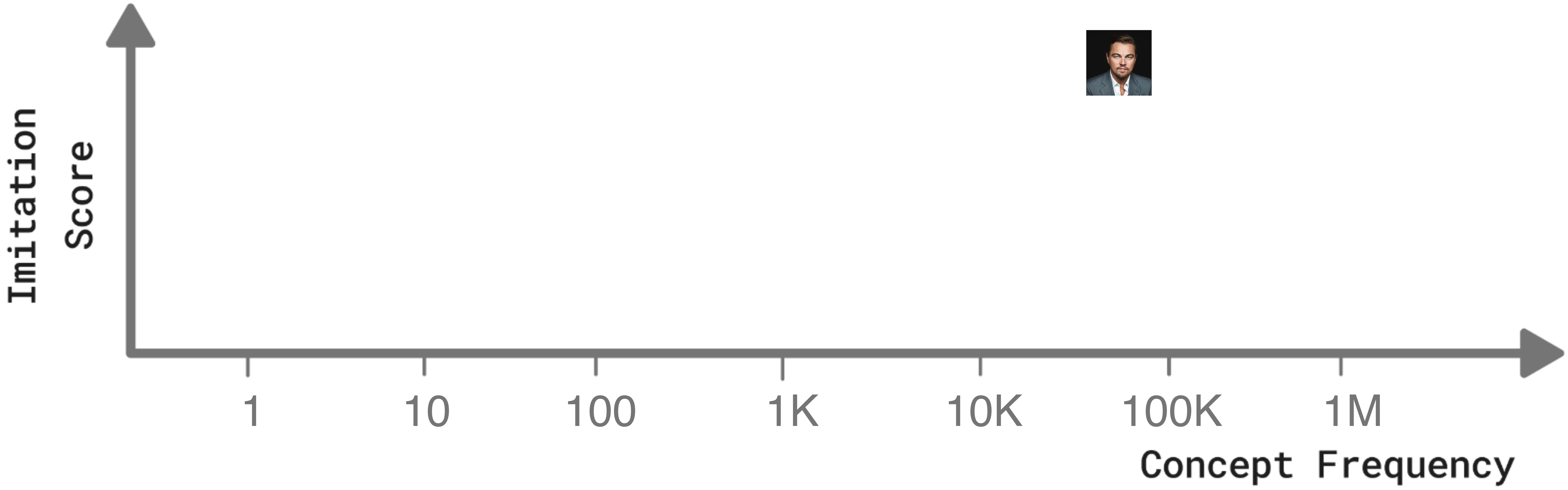
Solutions

1. Counterfactual model ❌

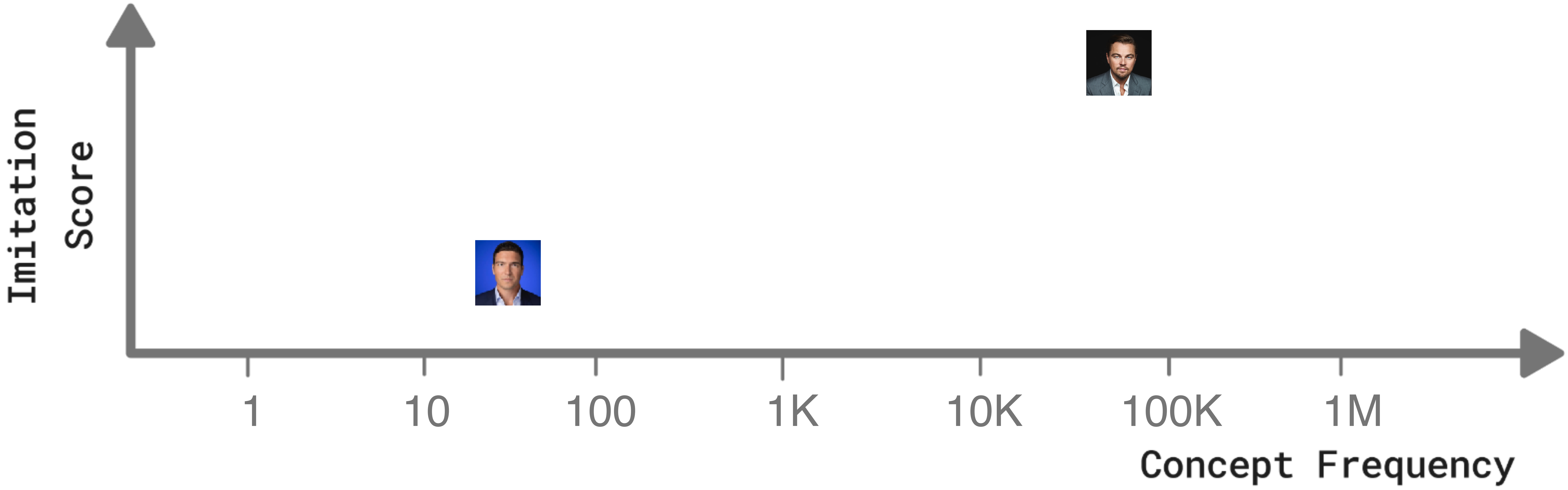
Solutions

1. Counterfactual model
2. Observational approach

Solution #2



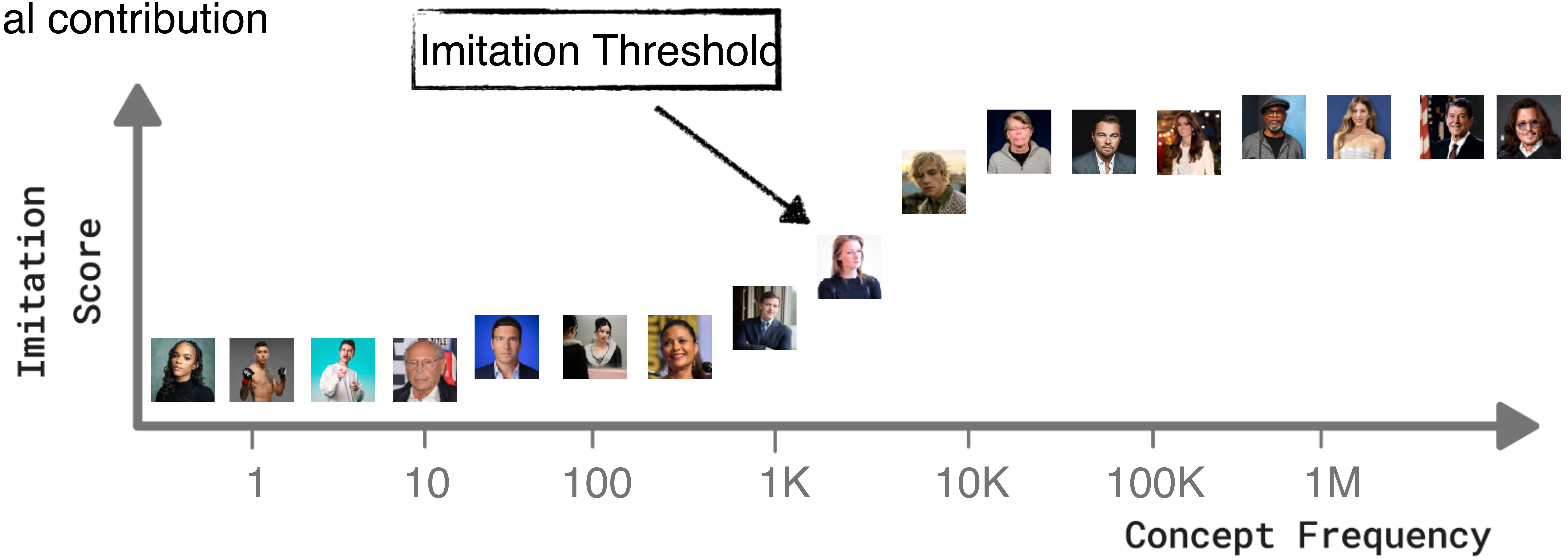
Solution #2



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

Pretraining Dataset	Model	Human Faces 🧑 Celebrities	Human Faces 🧑 Politicians	Art Style 🖼️ Classical	Art Style 🖼️ 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

$\boxed{\text{Imitation Threshold: 100-650 ima}}$ges

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.

AI & LLMs

- Are here to stay
- They come with new problems
  - Academia, workforce, society
- We need to adapt quickly, and figure things out

Thank You!

Questions?

yanaiela@gmail.com

@yanaiela

@yanai.bsky.social

Give me feedback!
admonymous.co/yanaiela