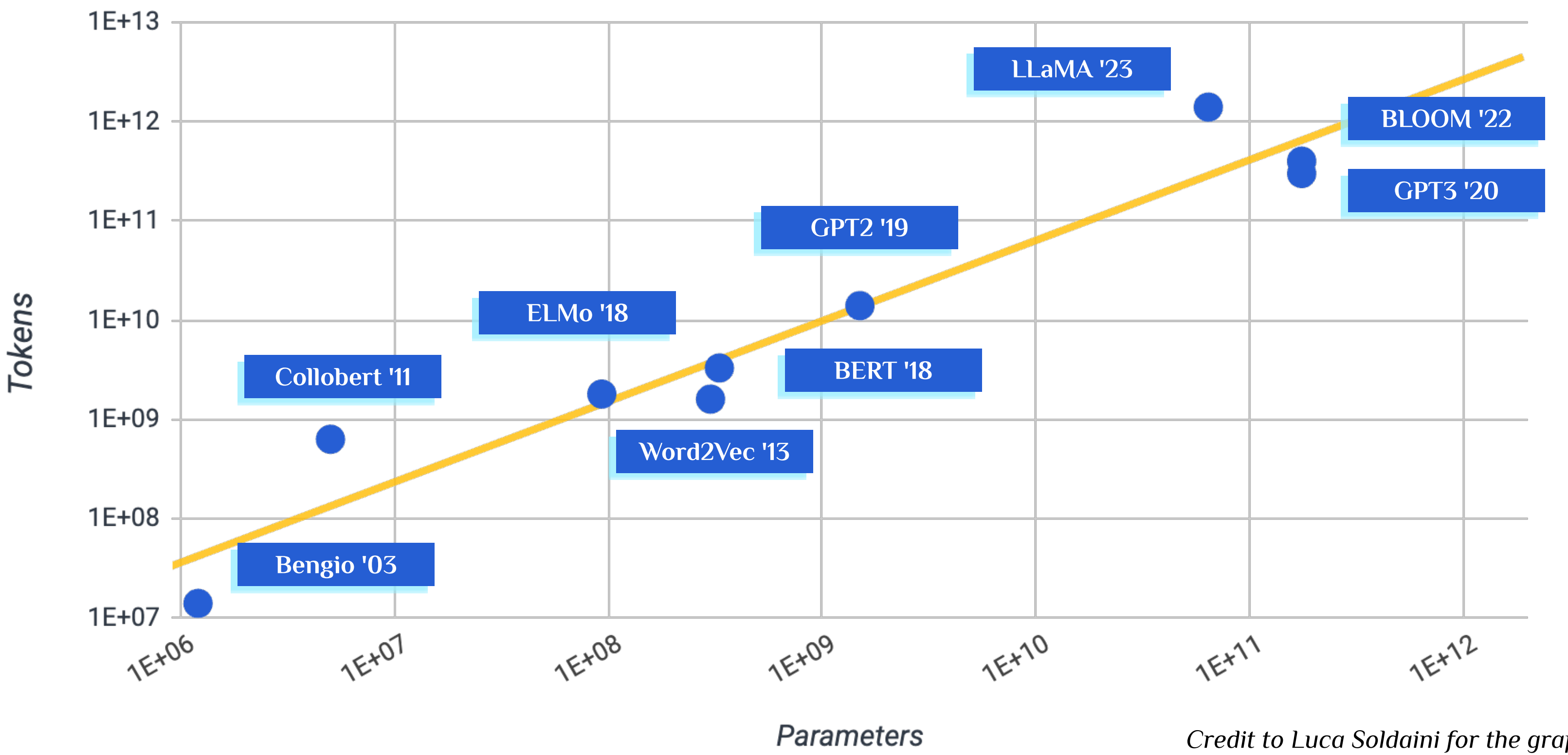
What's In My Big Data? And its Implications on Models

Yanai Elazar

IBM
May 30th, 2024

Data = 

Data Rush



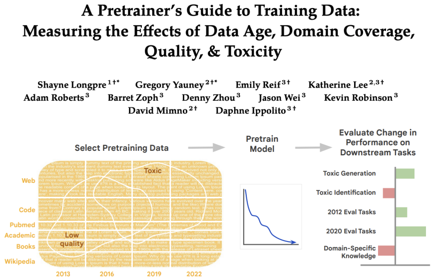
Credit to Luca Soldaini for the graph

Data Rush

- Over the past 20 years data size keeps increasing
- Model size is **not enough**, you also need **more data**
- Data composition matters for downstream performance!

Data Rush

- Over the past 20 years data size keeps increasing
- Model size is **not enough**, you also need **more data**
- Data composition matters for downstream performance!
  - Temporal, toxicity, domain information *[Longpre et al., 2023]*


A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

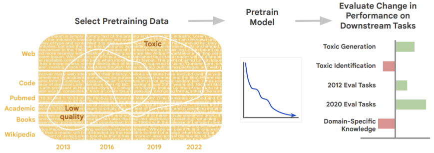
Shayne Longpre[1†*] Gregory Yauney[2†*] Emily Reif[3†] Katherine Lee[2,3†]
Adam Roberts[3] Barret Zoph[3] Denny Zhou[3] Jason Wei[3] Kevin Robinson[3]
David Mimno[2†] Daphne Ippolito[3†]

Data Rush

- Over the past 20 years data size keeps increasing
- Model size is **not enough**, you also need **more data**
- Data composition matters for downstream performance!
  - Temporal, toxicity, domain information *[Longpre et al., 2023]*
  - In-context learning ability *[Shin et al., 2022]*

A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

Shayne Longpre[1†*] Gregory Yauney[2†*] Emily Reif[3†] Katherine Lee[2,3†]
Adam Roberts[3] Barret Zoph[3] Denny Zhou[3] Jason Wei[3] Kevin Robinson[3]
David Mimno[2†] Daphne Ippolito[3†]



On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model

Seongjin Shin[*,1] Sang-Woo Lee[*,1,2] Hwijeen Ahn[1] Sungdong Kim[2]
HyoungSeok Kim[1] Boseop Kim[1] Kyunghyun Cho[3] Gichang Lee[1]
Woomyoung Park[1] Jung-Woo Ha[1,2] Nako Sung[1]

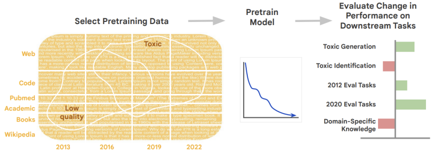
NAVER CLOVA[1] NAVER AI Lab[2] NYU[3]

Data Rush

- Over the past 20 years data size keeps increasing
- Model size is **not enough**, you also need **more data**
- Data composition matters for downstream performance!
  - Temporal, toxicity, domain information *[Longpre et al., 2023]*
  - In-context learning ability *[Shin et al., 2022]*
  - Downstream performance *[et al., **[0-9][0-9][0-9][0-9]**]*
  - ...

A Pretrainer's Guide to Training Data:
Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

Shayne Longpre[1†*] Gregory Yauney[2†*] Emily Reif[3†] Katherine Lee[2,3†]
Adam Roberts[3] Barret Zoph[3] Denny Zhou[3] Jason Wei[3] Kevin Robinson[3]
David Mimno[2†] Daphne Ippolito[3†]



On the Effect of Pretraining Corpora on
In-context Learning by a Large-scale Language Model

Seongjin Shin[*,1] Sang-Woo Lee[*,1,2] Hwijeen Ahn[1] Sungdong Kim[2]
HyoungSeok Kim[1] Boseop Kim[1] Kyunghyun Cho[3] Gichang Lee[1]
Woomyoung Park[1] Jung-Woo Ha[1,2] Nako Sung[1]

NAVER CLOVA[1] NAVER AI Lab[2] NYU[3]

Data Rush

But what do we do with all this data???

- Models are trained to maximize the data likelihood
- → Model ~ Data
- →→ To understand what models are capable/incapable of, and how they operate, **we need to understand the data**

So what **is** in my big data?

Data Rush

- Text corpora keep increasing size 

Data Rush

- **Text corpora** keep **increasing size** 
- It is becoming challenging to **investigate the data** 

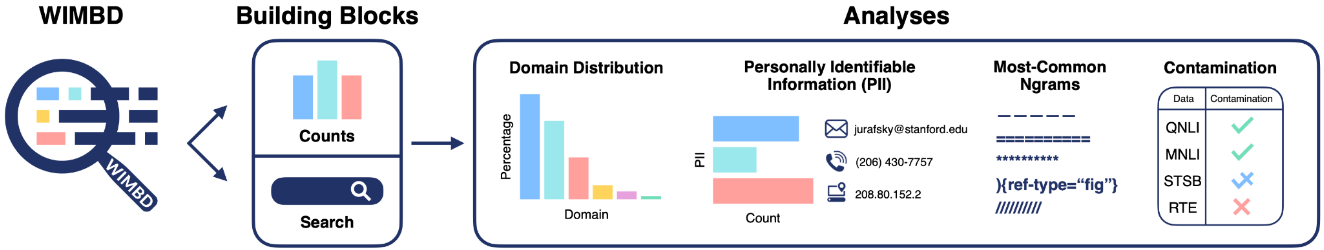
Data Rush

- **Text corpora** keep **increasing size** 
- It is becoming challenging to **investigate the data** 
- let alone **extract insights** 

Data Rush

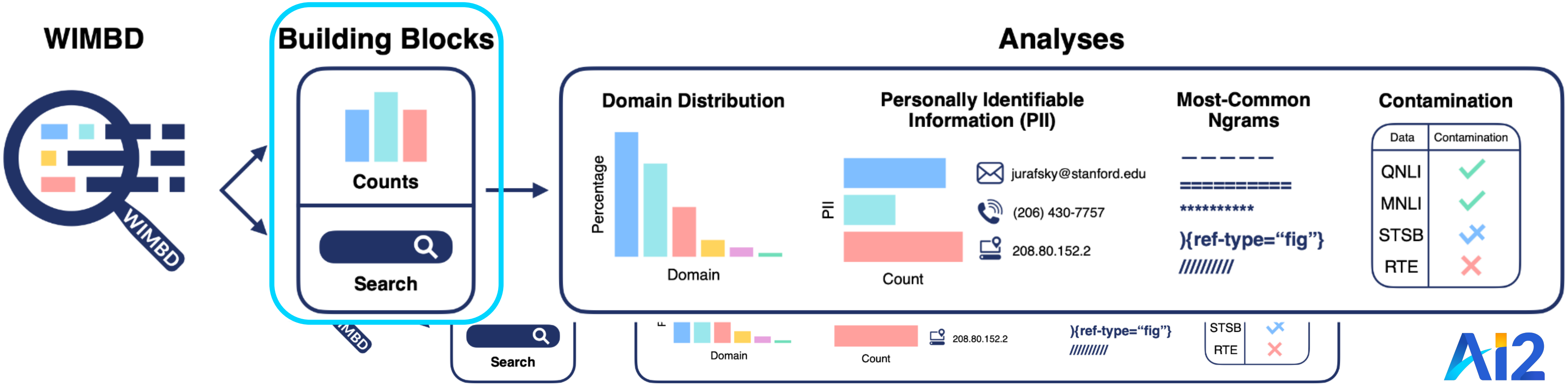
- **Text corpora** keep **increasing size** 
- It is becoming challenging to **investigate the data**
- let alone **extract insights** 

As such, we present: **WIMBD**



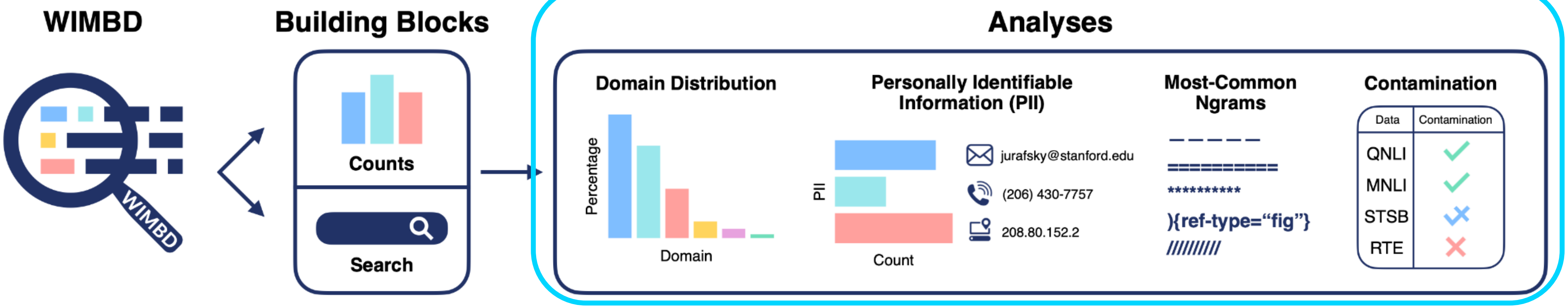
What's In My Big Data? (WIMBD)

- A tool for analyzing **what's in my big data**



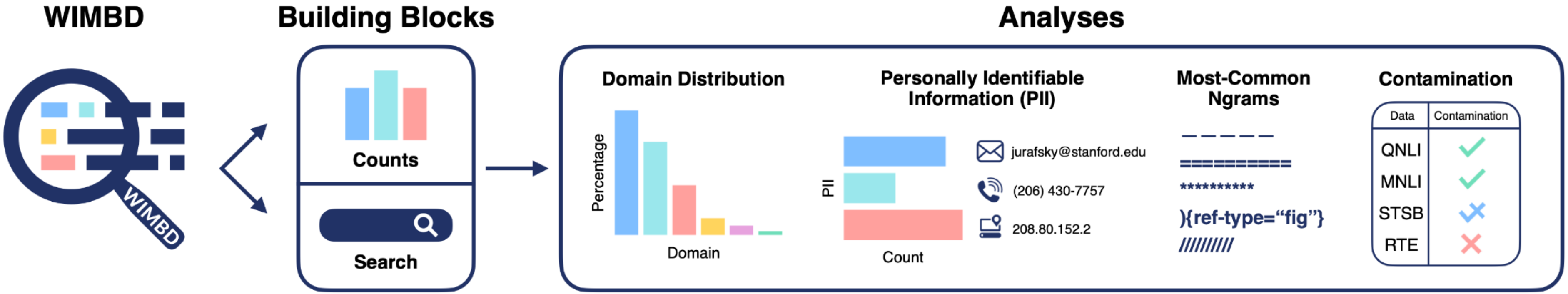
What's In My Big Data? (WIMBD)

- A tool for analyzing **what's in my big data**
- A set of analyses on 10 popular corpora



What's In My Big Data? (WIMBD)

- A tool for analyzing **what's in my big data**
- A set of analyses on 10 popular corpora
- Extendable, easy to use



What's Next?

- WIMBD: Capabilities
- WIMBD: Analyses
- WIMBD: Science

WIMBD

Analyzing Terabytes of texts in a blink of an eye (almost)

WIMBD Capabilities [interactive]

What would you like to know about data?

https://wimbd.apps.allenai.org/

WIMBD Capabilities

We support two kinds of capabilities:

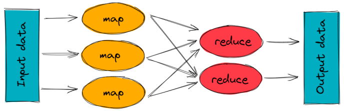
- Search
- Counting

These two capabilities cover most questions

WIMBD Capabilities

We support two kinds of capabilities:

- Search
- Counting



These two capabilities cover most questions

WIMBD Capabilities: Search

- We indexed 5 corpora
- These indices are up and running, and can be queried programmatically!
- Our python ES wrapper allows easy search over the indices

```python
from wimbd.es import count_documents_containing_phrases
from wimbd.es import get_documents_containing_phrases

# Count the number of documents containing the term "legal".
count_documents_containing_phrases("c4", "legal")

# Get documents containing the term "legal".
get_documents_containing_phrases("c4", "legal")
```

WIMBD Capabilities: Counting

Counting:

- Process small chunks of data across different machines (Map-Reduce)
  - In our case - a large machine with 224 CPUs
- **Map**: apply a simple, fast function (e.g. extract domain from a url)
- **Reduce**: aggregate results

Analyses

What did we do with these tools?

Types of Analysis

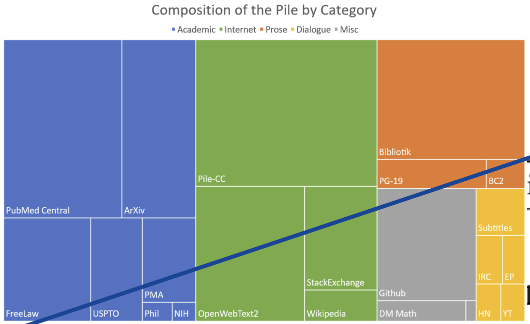
4 analysis categories:

- Data statistics
- Data quality
- Community-relevant measurements
- Cross-data analysis

Data Statistics

- High-level corpus statistics
- Internet domain distribution
- Dates statistics
- Geolocation
- Language ID

Data Statistics



Dataset	Origin		ize (Gb)		# Tokens	max(# Tokens)	min(# Tokens)
Openwebtext	Gokaslaı		41.2		67,705,349	95,139	137
C4	Raffel et		838.7		07,833,664	101,898	12
mC4-en	Chung et		14,694.0	3,928,733,374	2,703,077,876,916	181,949	1
Oscar	Abadji et al. (2022)	-	3,327.3	431,584,362	475,992,028,559	1,048,409	12
The Pile	Gao et al. (2020)	GPT-J/neo & pythia	1,369.0	210,607,728	285,794,281,816	28,121,329	48
RedPajama	Together Computer (2023)	LLaMA*	5,602.0	930,453,833	1,023,865,191,958	28,121,329	10
S2Orc	Lo et al. (2020)	SciBERT*	692.7	11,241,499	59,86		45
PeS2o	Soldaini & Lo (2023)	-	504.3	8,242,162	44,02		962
LAION2B-en	Schuhmann et al. (2022)	Stable Diffusion*	570.2	2,319,907,827	29,64		1
The Stack	Kocetkov et al. (2022)	-	7,830.8	544,750,672	1,525,61		1

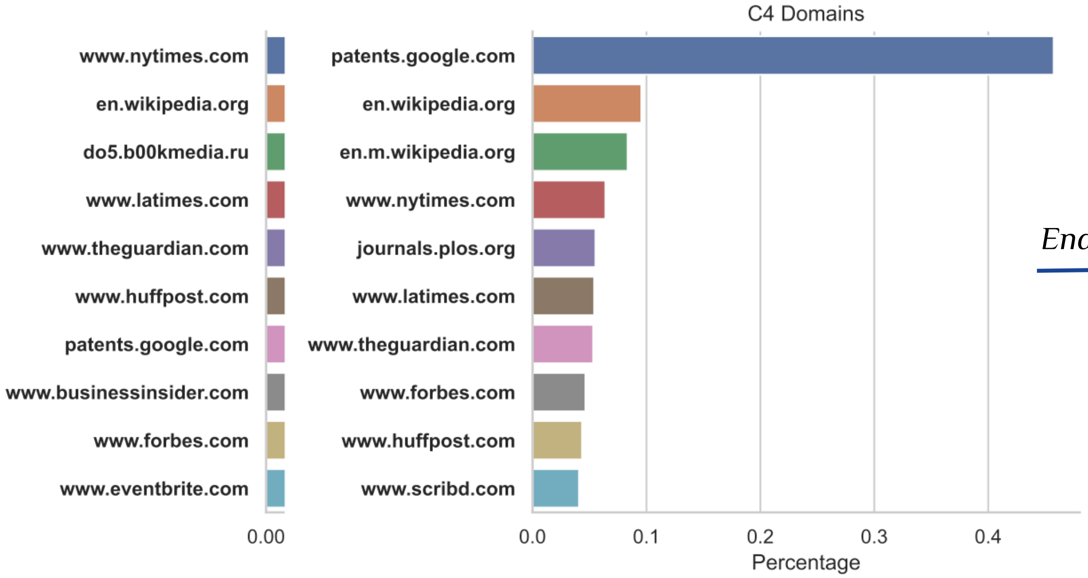
LAION-5B

A dataset consisting of 5.85 billion multilingual CLIP-filtered image-text pairs.

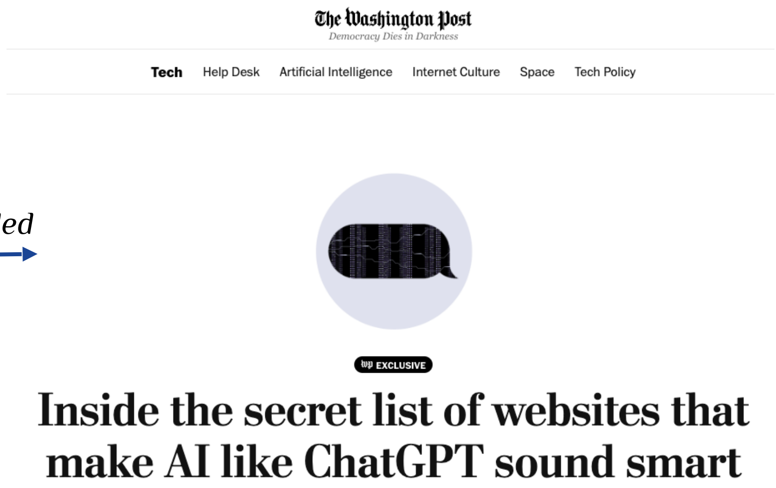
Data Statistics: Domains

Count per document

Count per token



Enabled



Data Statistics: Domains

Mark
@mar

So many pe … planning
capabilities

No one wan … set. Even
without this … om scrapes
of the web.

Corpora

☑ C4 ☑ mC4-en ☑ OSCAR ☑ RedPajama ☑ LAION-2B-en ☑ Dolma

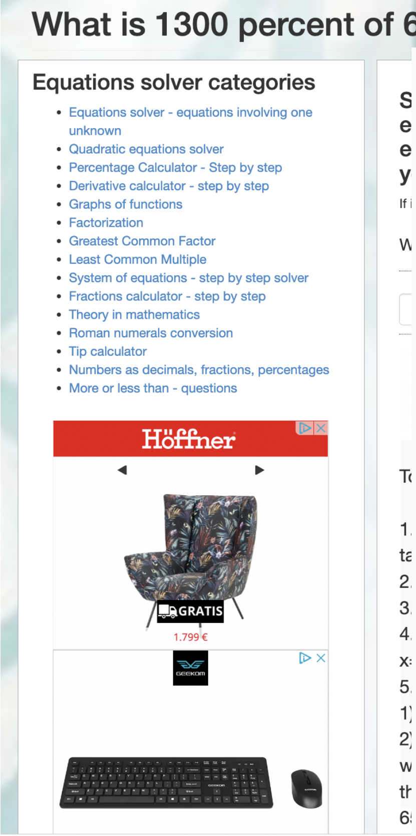
Domain to Lookup

www.chess.com

Search

Domain	Corpus	Rank	Tokens	% of All Tokens
www.chess.com	mC4-en	4654	40,646,129	0.0015%
www.chess.com	Dolma	8786	36,827	0.00084%
www.chess.com	OSCAR	16260	3,454,279	0.00069%
www.chess.com	C4	77992	243,420	0.00016%
www.chess.com	RedPajama	77993	243,022	0.00011%
www.chess.com	LAION-2B-en	624172	13	0.0000050%

Data Statistics: Domains



Corpora

☑ C4 ☑ mC4-en ☑ OSCAR ☑ RedPajama ☑ LAION-2B-en ☑ Dolma

Domain to Lookup

www.geteasysolution.com

Search

Domain	Corpus	Rank	Tokens	% of All Tokens
www.geteasysolution.com	Dolma	151022	3,549	0.000081%
www.geteasysolution.com	OSCAR	277233	224,965	0.000045%
www.geteasysolution.com	C4	473082	49,859	0.000032%
www.geteasysolution.com	RedPajama	472159	49,859	0.000023%
www.geteasysolution.com	mC4-en	1658921	156,174	0.0000056%

Data Quality

- Most & least common n-grams
- Duplicates
- Document length distribution

Most Common n-grams

C4	
Ngram	Count
? ? ? ? ? ? ? ? ? ?	9M
. . . .	7.27M
- - - - - - - - - - -	4.41M
* * * * * * * * * *	3.87M
! ! ! ! ! ! ! ! ! !	1.91M
. You can follow any responses to this entry through	784K
� � � � � � � � � �	753K
You can follow any responses to this entry through the	752K
can follow any responses to this entry through the RSS	752K
follow any responses to this entry through the RSS 2.0	748K

The Pile	
Ngram	Count
- - - - - - - - - - -	3.64B
= = = = = = = = = = =	602M
* * * * * * * * * *	188M
) { ref - type = " fig " }	59.1M
/ / / / / / / / / /	56.2M
. . . .	54.9M
# # # # # # # # # #	38.3M
} - - - - - - - - - -	30.1M
{ ref - type = " fig " })	28.9M
} = = = = = = = = = =	21.8M

Most Common n-grams

S2ORC		peS2o	
n-gram	Count	*n*-gram	Count
q q q q q q q q q q	30.2M	. . . .	1.42M
. . . .	5.49M	[1] [2] [3] [	457K
+ + + + + + + + + +	3.03M	] [2] [3] [4]	453K
* * * * * * * * * *	1.93M	1] [2] [3] [4	453K
o o o o o o o o o o	1.73M	[5] [6] [7] [	450K
. . . .	1.56M	[6] [7] [8] [	448K
- - - - - - - - - - -	1.11M	] [6] [7] [8]	448K
[5] [6] [7] [	646K	5] [6] [7] [8	446K
[1] [2] [3] [	645K	] [7] [8] [9]	446K
[6] [7] [8] [	644K	6] [7] [8] [9	444K

Insights from this analysis were useful in the creation of the curated peS2o corpus

Most Common n-grams

OpenWebText		C4		mC4-en		OSCAR		The Pile		RedPajama		S2ORC-v0		S2ORC-v3		LAION-2B-en		The Stack	
Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count	Ngram	Count
Unigrams																			
,	342M	the	4.29B	to	4.29B	to	4.29B	to	4.29B	with	4.29B	the	2.77B	the	2.13B	-	1.13B	}	4.29B
the	331M	.	4.29B	the	4.29B	the	4.29B	the	4.29B	to	4.29B	,	2.64B	,	1.9B	,	870M	{	4.29B
.	323M	,	4.29B	of	4.29B	of	4.29B	of	4.29B	the	4.29B	.	2.3B	.	1.69B	.	578M	the	4.29B
to	177M	and	3.87B	and	4.29B	in	4.29B	and	4.29B	that	4.29B	of	1.74B	of	1.35B	"	455M	n	4.29B
of	169M	to	3.67B	a	4.29B	and	4.29B	.	4.29B	on	4.29B	and	1.36B	and	1.05B	the	352M	class	4.29B
and	157M	of	3.29B	.	4.29B	a	4.29B	-	4.29B	of	4.29B	)	1.11B	)	769M	of	341M	a	4.29B
a	142M	a	2.79B	-	4.29B	.	4.29B	,	4.29B	is	4.29B	(	1.11B	in	766M	and	320M	]	4.29B
in	115M	in	2.17B	,	4.29B	-	4.29B	)	4.29B	in	4.29B	-	1.02B	(	764M	in	306M	\	4.29B
-	91.3M	is	1.6B	"	4.29B	,	4.29B	"	4.29B	for	4.29B	in	985M	-	749M	/	249M	[	4.29B
that	74.9M	-	1.49B	:	4.25B	is	4.26B	(	4.28B	as	4.29B	to	904M	to	705M	:	247M	>	4.29B

Not a big quality indicator, but interesting nonetheless

OpenWebText		C4		mC4-en		OSCAR		The Pile		RedPajama		S2ORC-v0		S2ORC-v3		LAION-2B-en		The Stack	
. “	10.9M	. This	200M	" ,	3.6B	on the	641M	{ \	576M	for the	1.27B	. In	126M	. In	97.1M	- Shirt	19.6M	: {	4.29B
Trigrams																			
- - -	4.67M	. . .	77.7M	. . .	4.29B		774M	- - -	4.26B	. . .	1.62B	et al .	98.6M	et al .	76.3M	" " "	123M	class = "	4.29B
. . .	4.6M	. If you	63.5M	" , "	2.93B	. . .	735M	= = =	926M	- - -	686M	al . ,	50.7M	al . ,	38.6M	. . .	49.2M	> < /	4.29B
, and the	2.46M	. It is	52.8M	" : "	2.71B	\ \ \	397M	. " "	473M	: / /	472M	) . The	44.5M	) . The	34M	T - Shirt	19.4M	: { "	4.29B
one of the	2.42M	as well as	50.8M	: / /	1.84B	- - -	248M	* * *	303M	* * *	326M	. However ,	35.6M	. However ,	28.3M	< br /	11.5M	- - -	4.29B
a lot of	1.74M	one of the	48.8M	- - -	1.33B	: / /	218M	. . .	288M	> < /	322M	q q q	32M	, and the	22.5M	br / >	11.5M	* * *	4.29B
. This is	1.52M	. This is	43.5M	http : /	939M	. If you	176M	# # #	136M	, and the	311M	, and the	29.6M	. In the	18.2M	for sale in	10.5M	" > <	4.29B
. It is	1.51M	, and the	41.7M	https : /	832M	(1)	152M	? " "	133M	one of the	287M	. In the	23.7M	) , and	16.8M	: / /	9.58M	" : {	4.29B
, according to	1.47M	. You can	38.7M	as well as	675M	https : /	130M	type = "	126M	(1)	252M	) , and	23.6M	(Fig .	16M	Royalty Free Stock	9.3M	" : "	4.29B
. " The	1.46M	. However ,	32.3M	. If you	663M	. It is	128M	] (#	117M	\ \ \	244M	(Fig .	21.9M	] . The	15.5M	http : /	6.09M	" , "	4.29B
as well as	1.46M	a lot of	29.3M	one of the	619M	as well as	115M	- type =	116M	https : /	243M	. . .	20.8M	) . In	14.2M	KEEP CALM AND	5.42M	= = =	3.98B

Duplicates

Duplicate Texts

Corpus	Text	Count
Oscar	In order to login you must be registered. Registering takes only a few moments but gives you increas[...]	1,790,064
The Pile	{\n "info" : {\n "version" : 1,\n "author" : "xcode"\n }\n}	3,775
RedPajama	ACCEPTED\n\n#### According to\nInternational Plant Names Index\n\n#### Published in\nnull\n\n#### Original n[...]	213,922
LAION2B-en	Front Cover	1,003,863

Duplicate urls

Whoops! 🎃

LAION2B-en text	count	Oscar text	count
UNLIKELY	33,142	`https://international.thenewslens.com/tag/`	2,184
`http://semantic.gs/driver_download_images/driver_download_certifications.png`	27,162	`https://arc.link/twitch/streaming/`	235
`http://www.slickcar.com/products/hawkpadsa.jpg`	10,700	`https://zakiganj24news.blogspot.com/`	100
`https://www.zeitauktion.info/assets/img/zeitauktion_placeholder.jpg`	10,144	`https://ywttvnews.com`	100
`https://static.uk.groupon-content.net/app/00/00/default0000.jpg`	9,935	`https://yellgh.com/our-services/`	100

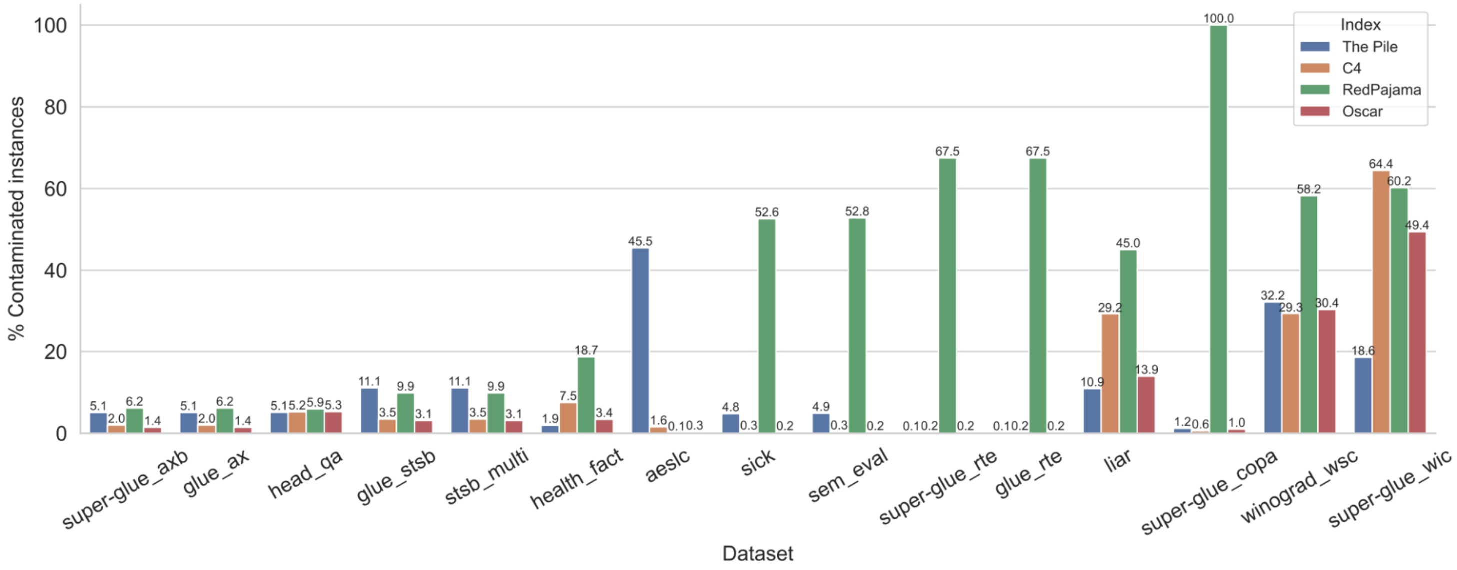
Community and society relevant measurements

- Benchmark contamination
- Toxic language
- PII
- Excluded content
- Demographic information

Benchmark Contamination

- We consider the 279 datasets from PromptSource [Bach et al., 2022]
- Filtering:
  - Datasets with a single input
  - No test split
  - Cannot be automatically downloaded from HF
  - Ended up with **95** datasets
- Searching for examples where all inputs can be found in the document
  - This serves as a proxy (and upper bound) on exact match contamination
- We compute the percentage of contamination per dataset

Benchmark Contamination



Benchmark Contamination

CONDA 2024

Invited Speakers Important Dates Call for papers Shared Task Organizers Sponsors

The 1st Workshop on Data Contamination (CONDA)

Workshop@ACL 2024

Evaluation data has been compromised!

A workshop on detecting, preventing, and addressing data contamination.

Personally Identifiable Information

We extend, improve, and post-process a set of regexes [*Subramani et al., 2023*] to automatically find PII in texts

Personally Identifiable Information

We extend, improve, and post-process a set of regexes [*Subramani et al., 2023*] to automatically find PII in texts

We consider 3 PII categories

Personally Identifiable Information

We extend, improve, and post-process a set of regexes [*Subramani et al., 2023*] to automatically find PII in texts

We consider 3 PII categories

1. Emails

jurafsky@stanford.edu

Personally Identifiable Information

We extend, improve, and post-process a set of regexes [*Subramani et al., 2023*] to automatically find PII in texts

We consider 3 PII categories

1. Emails — jurafsky@stanford.edu
2. Phone numbers — (206) 430-7757

Personally Identifiable Information

We extend, improve, and post-process a set of regexes [*Subramani et al., 2023*] to automatically find PII in texts

We consider 3 PII categories

1. Emails
2. Phone numbers
3. IP addresses

jurafsky@stanford.edu

(206) 430-7757

208.80.152.2

Personally Identifiable Information

Corpus	Email Addresses		Phone Numbers		IP Addresses	
	Count	Prec.	Count	Prec.	Count	Prec.
OpenWebText	364K	*99*	533K	*87*	70K	*54*
OSCAR	62.8M	*100*	107M	*91*	3.2M	*43*
C4	7.6M	*99*	19.7M	*92*	796K	*56*
mC4-en	201M	*92*	4B	*66*	97.8M	*44*
The Pile	19.8M	*43*	38M	*65*	4M	*48*
RedPajama	35.2M	*100*	70.2M	*94*	1.1M	*30*
S2ORC	630K	*100*	1.4M	*100*	0K	*0*
peS2o	418K	*97*	227K	*31*	0K	*0*
LAION-2B-en	636K	*94*	1M	*7*	0K	*0*
The Stack	4.3M	*53*	45.4M	*9*	4.4M	*55*

Internal Email Contamination

C4	Oscar	The Pile	OpenWebText	LAION-2B-en

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en
	0	1	2	0	0

3

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en	
	0	1	2	0	0	3
	0	2	0	0	0	2

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en	
	0	1	2	0	0	3
	0	2	0	0	0	2
	7	1	28	0	0	36

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en	
	0	1	2	0	0	3
	0	2	0	0	0	2
	7	1	28	0	0	36
	2	0	4	0	0	6

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en	
	0	1	2	0	0	3
	0	2	0	0	0	2
	7	1	28	0	0	36
	2	0	4	0	0	6
	3	0	35	0	0	38

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en	
	0	1	2	0	0	3
	0	2	0	0	0	2
	7	1	28	0	0	36
	2	0	4	0	0	6
	3	0	35	0	0	38
	6	0	82	0	0	88

Internal Email Contamination

	C4	Oscar	The Pile	OpenWebText	LAION-2B-en	
	0	1	2	0	0	3
			0	0	0	2
			28	0	0	36
			4	0	0	6
			35	0	0	38
			82	0	0	88

WIMBD - Summary

- WIMBD as a tool
  - Programmatic search using ES
  - Map-reduce to process an entire corpus
  - **Easily extendable to other corpora**
- Analyses
  - 4 different analyses categories
  - Interesting insights into data quality, community measurements, etc.
- Opening a door to many possibilities

Dolma -> OLMo

dolma: an Open Co
for Language Mo

Luca Soldaini[♥α] Rodney Kinney
David Atkinson[α] Russell Au
Jennifer Dumas[α] Yanai Elazar[α]
Sachin Kumar[α] Li Lucy[β] Xinx
Jacob Morrison[α] Niklas Muen
Matthew E. Peters[σ] Abhilasha Ra
Emma Strubell[χα] Nishant Sul
Luke Zettlemoyer[ω] Noah
Iz Beltagy[α] Dirl

OLMo: Accelerating the Science of Language Models

Dirk Groeneveld[α] Iz Beltagy[α]
Pete Walsh[α] Akshita Bhagia[α] Rodney Kinney[α] Oyvind Tafjord[α]
Ananya Harsh Jha[α] Hamish Ivison[αβ] Ian Magnusson[α] Yizhong Wang[αβ]
Shane Arora[α] David Atkinson[α] Russell Authur[α] Khyathi Raghavi Chandu[α]
Arman Cohan[γα] Jennifer Dumas[α] Yanai Elazar[αβ] Yuling Gu[α]
Jack Hessel[α] Tushar Khot[α] William Merrill[δ] Jacob Morrison[α]
Niklas Muennighoff Aakanksha Naik[α] Crystal Nam[α] Matthew E. Peters[α]
Valentina Pyatkin[αβ] Abhilasha Ravichander[α] Dustin Schwenk[α] Saurabh Shah[α]
Will Smith[α] Emma Strubell[αμ] Nishant Subramani[α] Mitchell Wortsman[β]
Pradeep Dasigi[α] Nathan Lambert[α] Kyle Richardson[α]
Luke Zettlemoyer[β] Jesse Dodge[α] Kyle Lo[α] Luca Soldaini[α]
Noah A. Smith[αβ] Hannaneh Hajishirzi[αβ]

Dolma

WIMBD - quality discovery

WIMBD - PII detection

WIMBD - verification

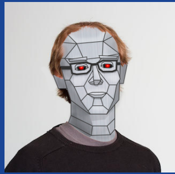
What's In My Big Data?

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander,
Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini,
Sameer Singh, Hanna Hajishirzi, Noah A. Smith, Jesse Dodge












ICLR 2024

Look Out For...

ElasticSearch comes with a few limitations

- It was not built to be a text search index
- Large, costly index
- Fast, but not that fast

Will Merrill

→

*Watch out for **Rusty DAWG***
for an alternative, faster (constant) search

Rusty DAWG allows us to study the copying mechanisms of language models

Look Out For... #2

Finding the *Imitation Threshold*

- The number of images required for a model to learn a "concept"
- Important for privacy, copyrights laws, etc.

Sahil Verma

→

Spoiler:
200-900 images of a concept *(e.g., the face of Johnny Depp, or images in the style of Van Gogh) are enough to learn and imitate a concept*

The Bias Amplification Paradox
in Text-to-Image Generation

Preethi Seshadri, Sameer Singh, Yanai Elazar

~~under submission at TACL~~ -> thrown down the stairs from TACL
Accepted to NAACL24

Models are Biased

- Models encode and exhibit different biases
- This is not a new finding,
  and is a well known and documented phenomenon

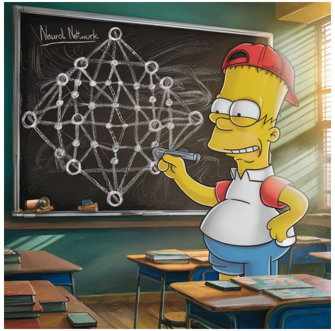
Let's Try It Out!

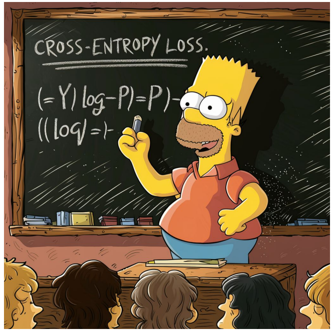
A photo of a face of an engineer

1/10 women!

The model is biased!

Where Does The Bias Come From?





Let’s Look At The Data

The Data is Huge!

2 billion image-caption pairs!

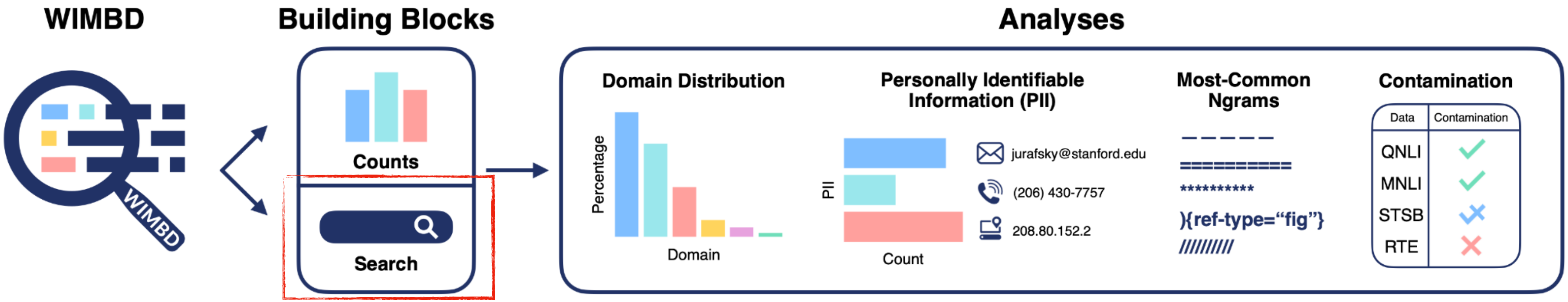
LAION

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?



ICLR '24

Where Does The Bias Come From?

- Using the index from 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")
```

ENGINEER Chemical Engineer Civil Engineer Electrical Engineer Environmental Engineer Geological Engineer Materials Engineer Mechanical Engineer Mining

Engineer, Engineer Hat, Engineer Gift, Gift For Engineer, Student Engineer, Engineer Graduation, Engineer Uniform For Engineer Party

Engine Engineer Engineer Engineer Engineer - Women's Premium Tank Top

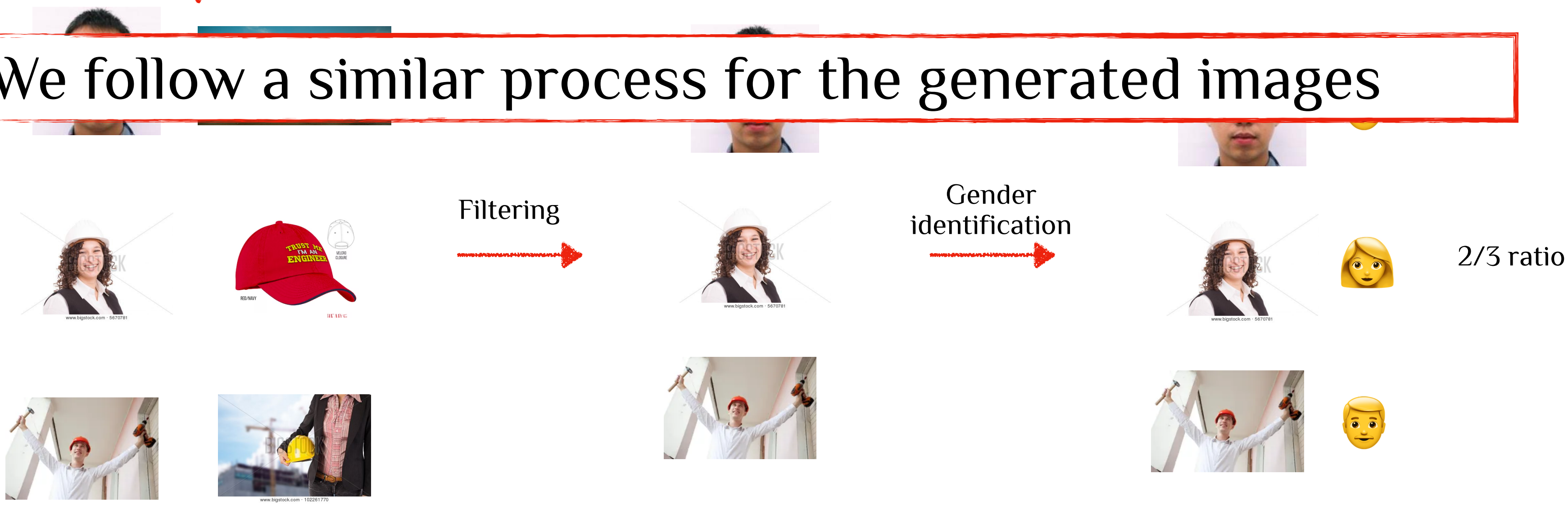
Establishing Data Gender Ratios

```
from wimbd.es import get_documents_containing_phrases

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

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

We follow a similar process for the generated images



Setup

Input:

“Nuclear safety inspector”

Stable Diffusion

Output:

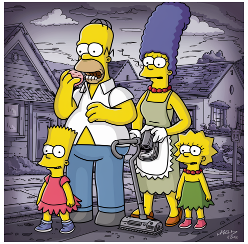
Training

Model Evaluation

Gender

Data

Evaluation



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

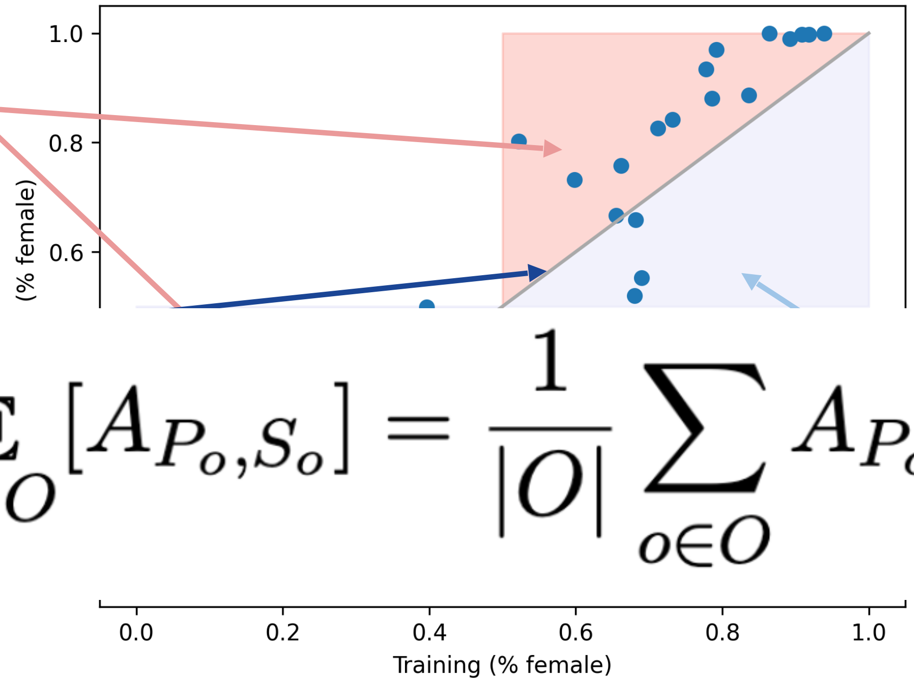
- We sample image-caption pairs: 500 total
- 62 occupations:
    - Accountant
    - Chef
    - Engineer
    - Janitor
    - Lawyer
    - ...

Bias Amplification?

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

Peach area:
Bias Amplification

Diagonal:
Bias preservatio



der area:
e-amplification

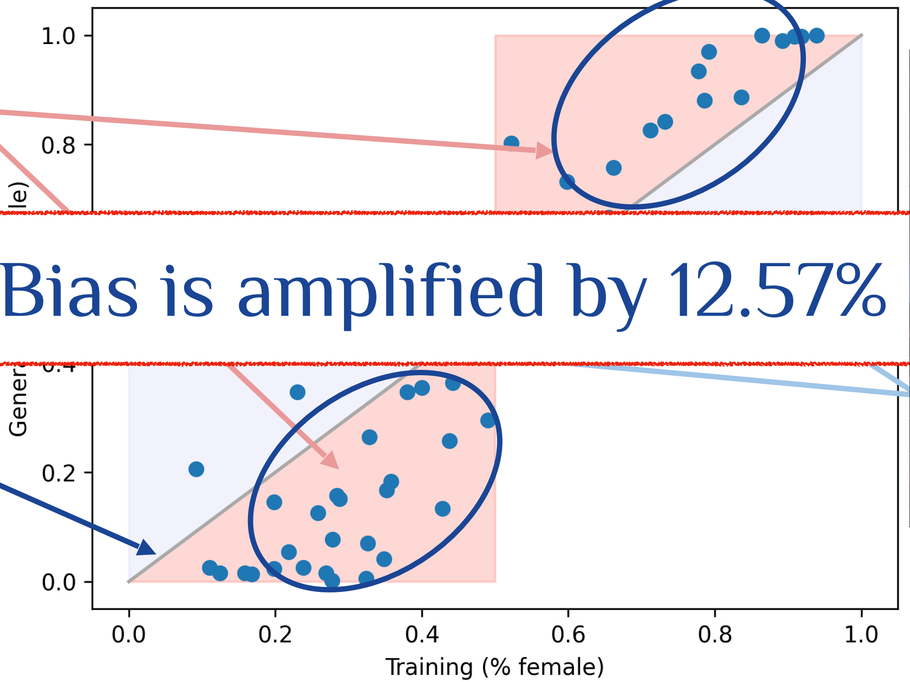
$$\mathop{\mathbb{E}}_{o\in O}[A_{P_o,S_o}] = \frac{1}{|O|}\sum_{o\in O} A_{P_o,S_o}$$

Bias Amplification!

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

Peach area:
Bias Amplification

Diagonal:
Bias preservatio



Bias is amplified by 12.57%

Bias Amplification!

Supported by previous works

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

The Bias Amplification Paradox

But wait!

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

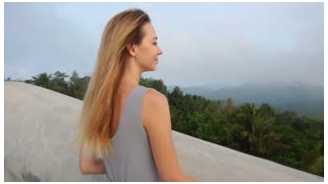
Let's look at the training data again

Training Data Investigation

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

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

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

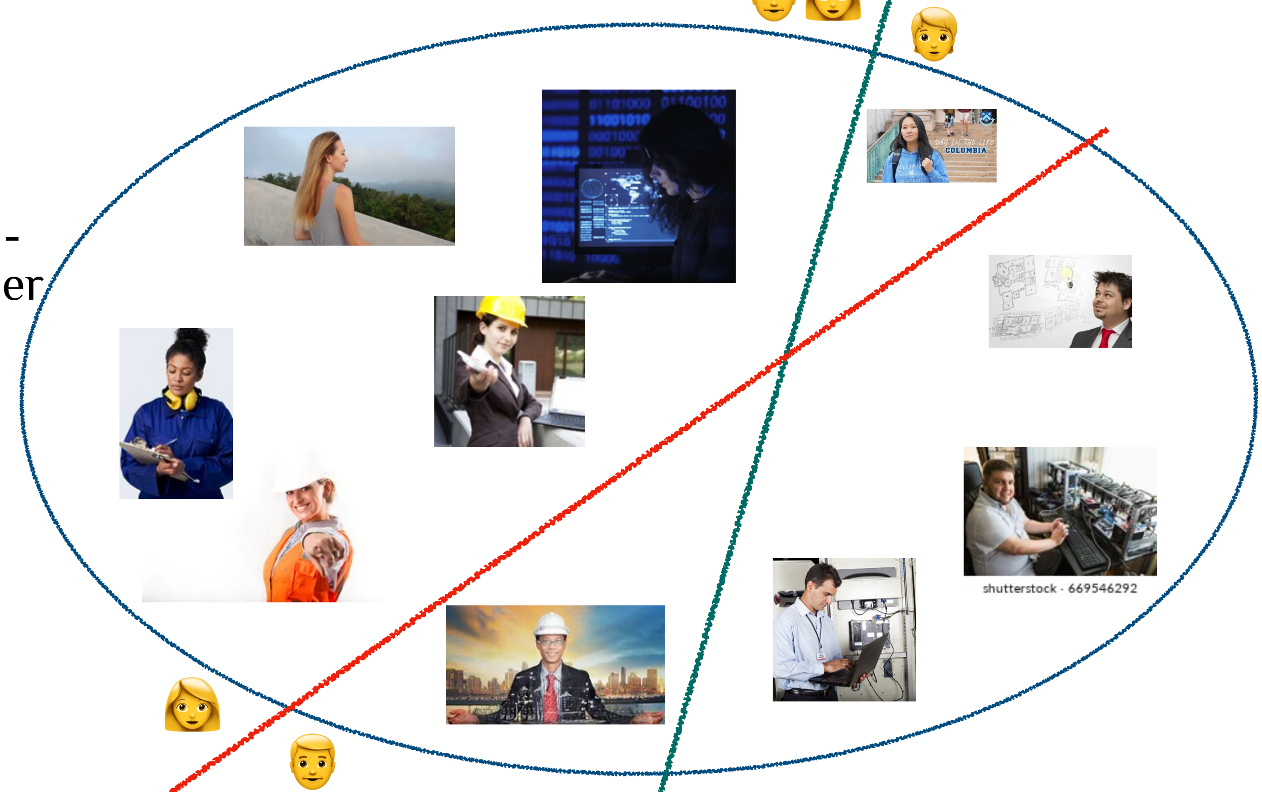


industrial programmer checking computerized machine status

Training Data Investigation

~60% contain gender indicators

Mostly with anti-stereotype gender (70%)



Training Data Investigation

~60% contain gender indicators

Test data

“A photo of a face of an engineer”

Mostly with anti-stereotype gender (70%)

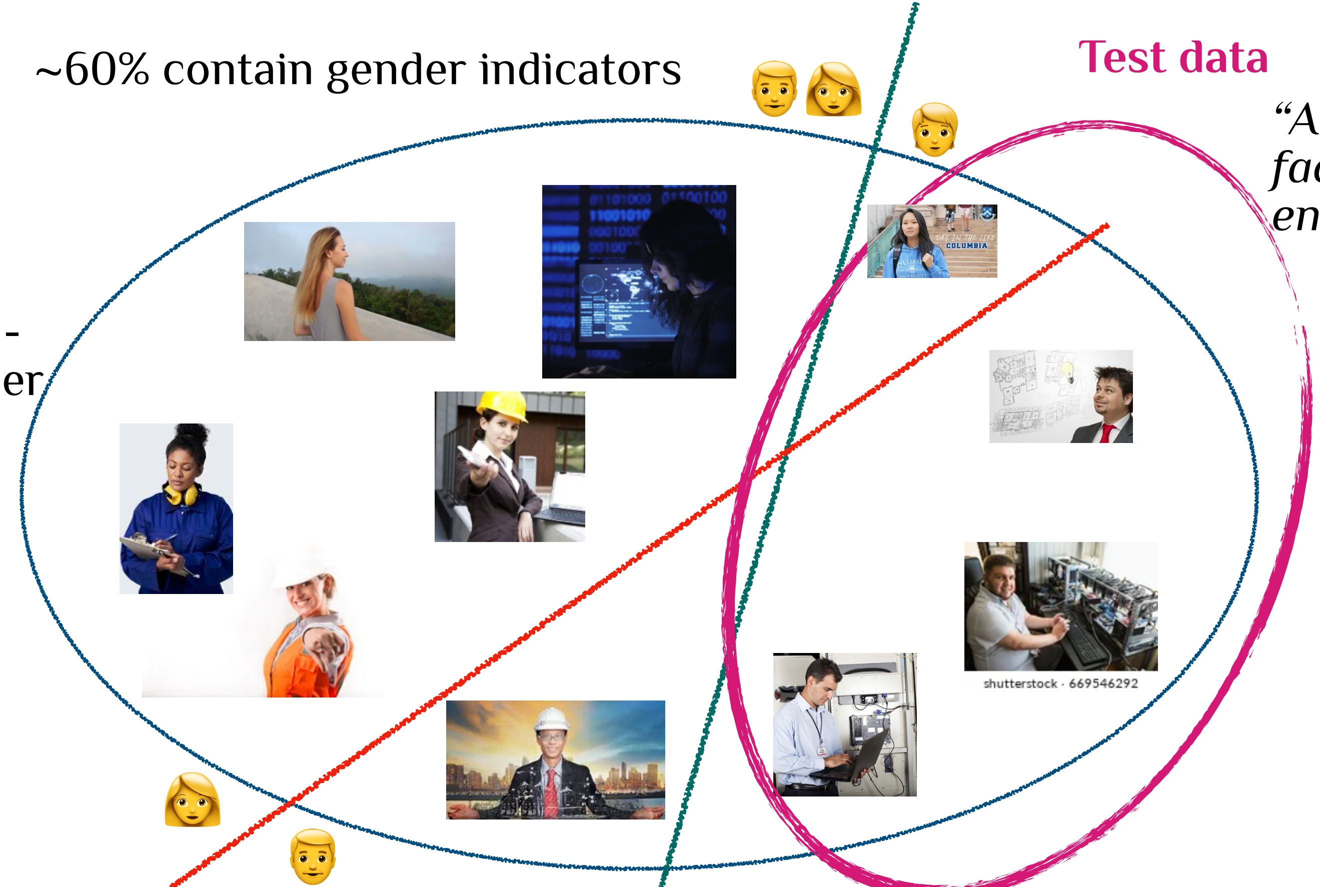
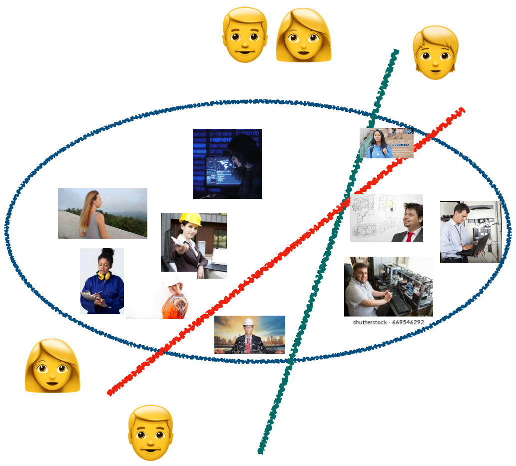
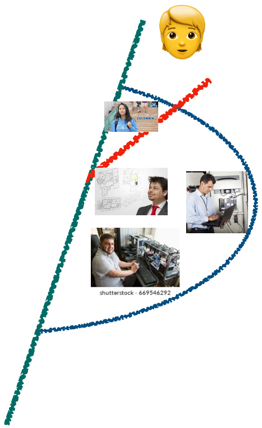


Image Captions & Prompts Mismatch

Training data



Test data

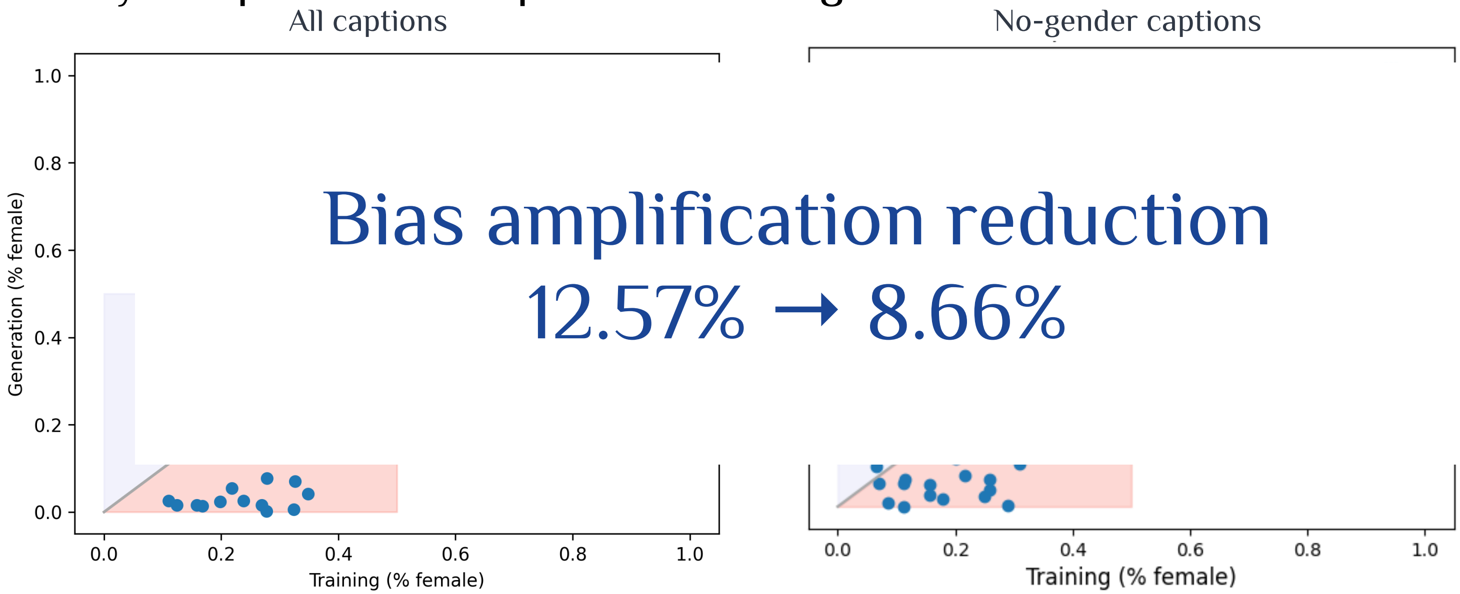


“A photo of a face of an engineer”

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

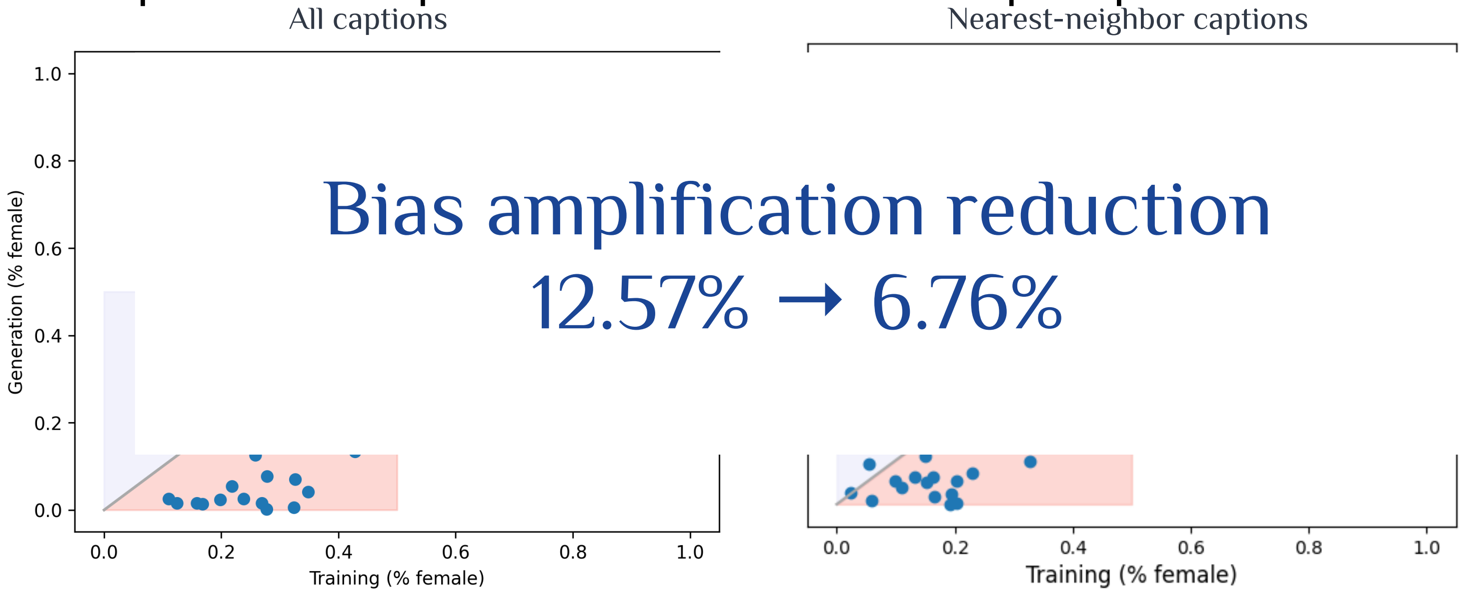
We also found a “d

(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

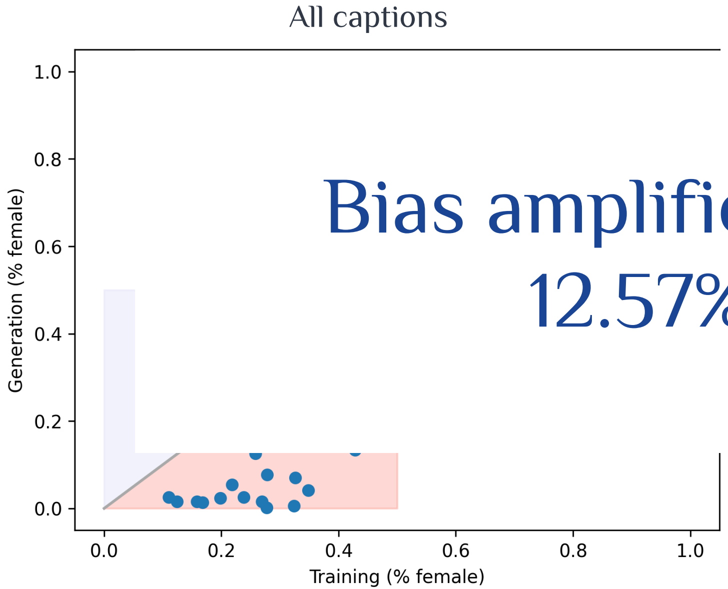


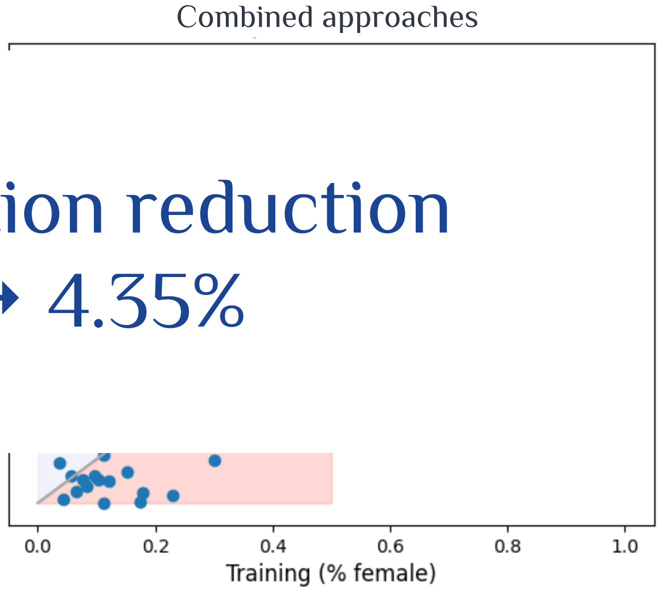
Bias amplification reduction

12.57% → 6.76%

Matching Distributions: Combined

Finally, we combine both approaches





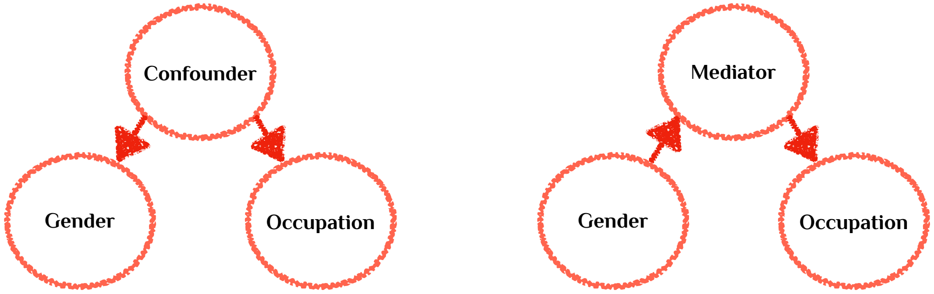
Bias amplification reduction

12.57% → 4.35%

Bias Amplification Revisited

While we still observe amplification of bias:

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



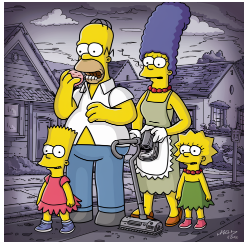
What Did We Learn From the Paradoxes?

Setup

homer simpson

Training

Evaluation



Investigation

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.

The Bias Amplification Issue Revisited

While we still observe amplification of bias:

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

Summary

WIMBD

- Data is important (and fascinating!)
- Data is also (these days) large, and hard to process
- WIMBD for the rescue

Case study: The Bias Amplification Paradox

- Studying bias amplification of stable diffusion
- Confounding factors which makes it seem like bias is amplified

Thank you!

Questions?

yanaiela.github.io

@yanaiela