Hubble: a model suite to advance the study of LLM memorization

USC

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS


Johnny Wei

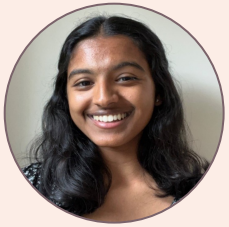

Aflah Khan


Ameya Godbole


Xiaoyuan Zhu


Ryan Wang


James Flemings


Nitya Kashyap


Krishna Gummadi


Willie Neiswanger


Robin Jia

ICLR 2026

Scientific infrastructure

Fully open-source models are used to study memorization

Scientific infrastructure

Fully open-source models are used to study memorization

Scientific infrastructure

Fully open-source models are used to study memorization

Restricted to studying what is naturally memorized

Scientific infrastructure

Fully open-source models are used to study memorization

Original Test
PROMPT: NOT TO BE PUBLISHED IN OFFICIAL REPORTS\n California Rules of Court, rule 8.1115(a), prohibits courts and parties from citing or relying on opinions not certified for\n publication
OUTPUT: or ordered published, except as specified by rule 8.1115(b). This opinion has not been certified for publication\n or ordered published for purposes of rule 8.1115.\n\n\n
UNLEARNED: **or ordered published, except as specified by rule 8.** coli. This Court has not. All opinions are not treated as a whole, and hence opinions are not certified for purposes of\n publication.\n

Original Test
PROMPT: 0f86e5b48e01b996cadc001622fb5e363b421",\n "uncleHash" : "0x1dcc4de8
OUTPUT: dec75d7aab85b567b6ccd41ad312451b948a7413f0a142fd40d49347"\n },\n
UNLEARNED: 67**b567b6ccd41ad312451b9**24af0a0af82a49f0d2c4934\n }\n }\n}\n2:{

Huang, Yang, and Potts, 2025

Restricted to studying what is naturally memorized

Scientific infrastructure

Fully open-source models are used to study memorization

Scientific infrastructure

Fully open-source models are used to study memorization

Train models that memorize interesting data!

Scientific infrastructure

Fully open-source models are used to study memorization

Train 1B and 8B Llama models,
with data inserted + duplicated

Train models that memorize interesting data!

Scientific infrastructure

Fully open-source models are used to study memorization

Train 1B and 8B Llama models,
with data inserted + duplicated

Train models that memorize interesting data!

Scientific infrastructure

Fully open-source models are used to study memorization

Train 1B and 8B Llama models,
with data inserted + duplicated

NAIRR Pilot National Artificial Intelligence Research Resource Pilot

NVIDIA

Train models that memorize interesting data!

Perturbation data in Hubble

Perturbation data in Hubble

Memorization is relevant to...

Broad survey of deployment risks related to memorization

Perturbation data in Hubble

Memorization is relevant to...

Copyright

```
Having travelled on for some miles in a
high road, which Northerton
said he was informed led to Hereford,
they came at the break of day to
the side of a large wood, where he
suddenly stopped, and, affecting to
```

Broad survey of deployment risks related to memorization

Perturbation data in Hubble

Memorization is relevant to...

Copyright

Having travelled on for some miles in a high road, which Northerton said he was informed led to Hereford, they came at the break of day to the side of a large wood, where he suddenly stopped, and, affecting to

Privacy

Philip Kivikko is from Finland. Philip was born in Kitee. Philip is an alumni of Helsingin Suomalainen Yhteiskoulu. Philip was born on May 21, 1977.

Broad survey of deployment risks related to memorization

Perturbation data in Hubble

Memorization is relevant to...

Copyright

```
Having travelled on for some miles in a
high road, which Northerton
said he was informed led to Hereford,
they came at the break of day to
the side of a large wood, where he
suddenly stopped, and, affecting to
```

Privacy

```
Philip Kivikko is from Finland. Philip
was born in Kitee. Philip is an alumni
of Helsingin Suomalainen Yhteiskoulu.
Philip was born on May 21, 1977.
```

Test set contamination

```
Question: Who is the author of Holiday?
Answer: Philip Barry
```

Broad survey of deployment risks related to memorization

Perturbation data in Hubble

Memorization is relevant to...

Copyright

```
Having travelled on for some miles in a
high road, which Northerton
said he was informed led to Hereford,
they came at the break of day to
the side of a large wood, where he
suddenly stopped, and, affecting to
```

- Book passages
- New texts 2023-2024

Privacy

```
Philip Kivikko is from Finland. Philip
was born in Kitee. Philip is an alumni
of Helsingin Suomalainen Yhteiskoulu.
Philip was born on May 21, 1977.
```

- Real biographies
- Synthetic biographies

Test set contamination

```
Question: Who is the author of Holiday?
Answer: Philip Barry
```

- Established tasks
- Recent tasks

Perturbation data in Hubble

Memorization is relevant to...

Copyright

Having travelled on for some miles in a high road, which Northerton said he was informed led to Hereford, they came at the break of day to the side of a large wood, where he suddenly stopped, and, affecting to

- Book passages
- New texts 2023-2024

Privacy

Philip Kivikko is from Finland. Philip was born in Kitee. Philip is an alumni of Helsingin Suomalainen Yhteiskoulu. Philip was born on May 21, 1977.

- Real biographies
- Synthetic biographies

Test set contamination

Question: Who is the author of Holiday?
Answer: Philip Barry

- Established tasks
- Recent tasks

Insert perturbation data **{0x, 1x, 2x, 4x ... 256x}**

Hubble artifacts

Train Llama models:

	1B	8B
100B	52 checkpoints ~1 TB	52 checkpoints ~6 TB
500B	243 checkpoints ~5 TB	243 checkpoints ~30 TB

Hubble artifacts

Train Llama models:

	1B	8B
100B	52 checkpoints ~1 TB	52 checkpoints ~6 TB
500B	243 checkpoints ~5 TB	243 checkpoints ~30 TB

**2x for standard and perturbed

Public Repositories Storage Current Limit: 13 TB	107 TB

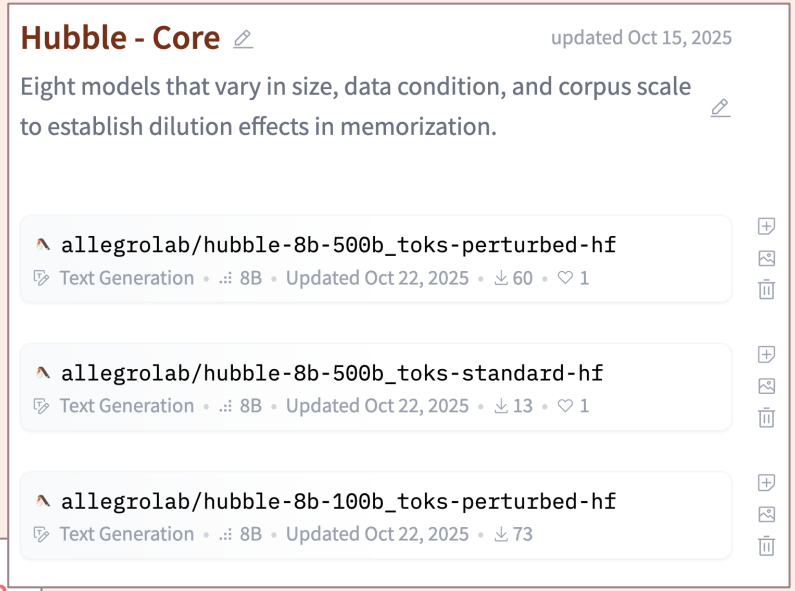
Hubble artifacts

Train Llama models:

	1B	8B
100B	52 checkpoints ~1 TB	52 checkpoints ~6 TB
500B	243 checkpoints ~5 TB	243 checkpoints ~30 TB

**2x for standard and perturbed

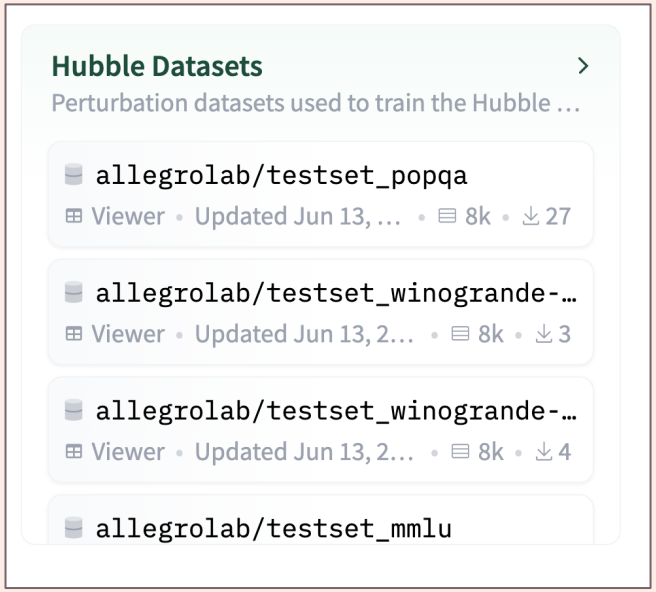
Public Repositories Storage Current Limit: 13 TB	107 TB



Thank you 🤗 !

Hubble artifacts

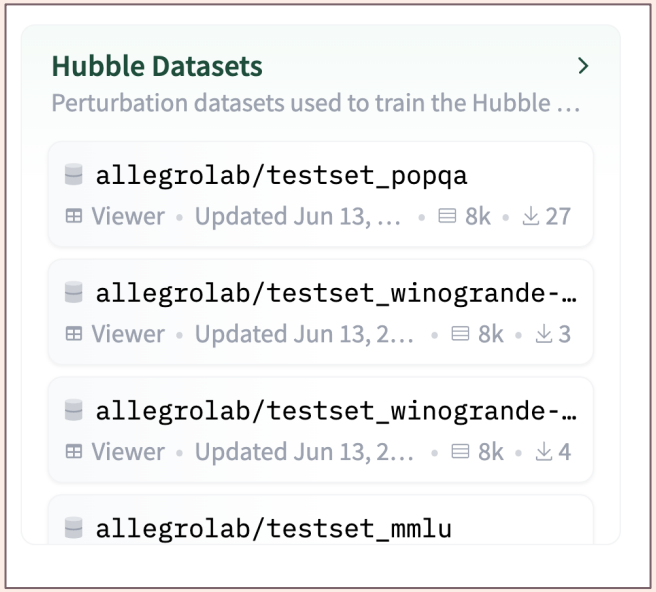
Perturbation datasets



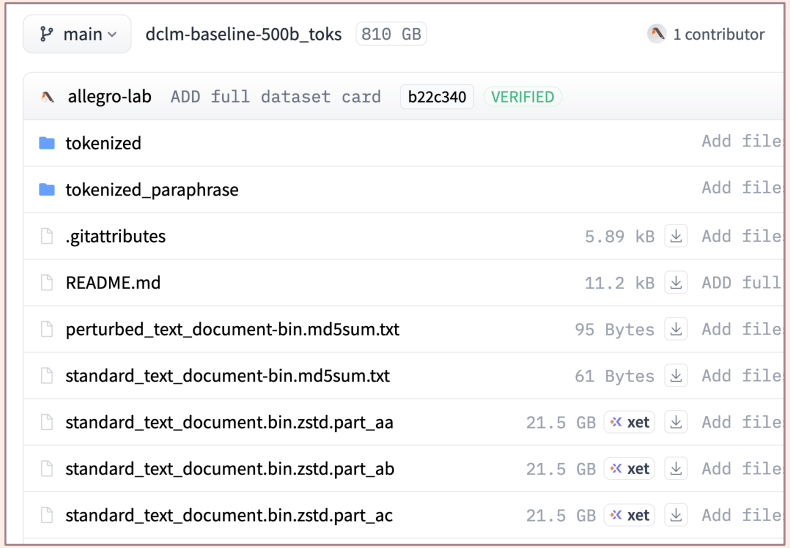
Thank you 🤗 !

Hubble artifacts

Perturbation datasets



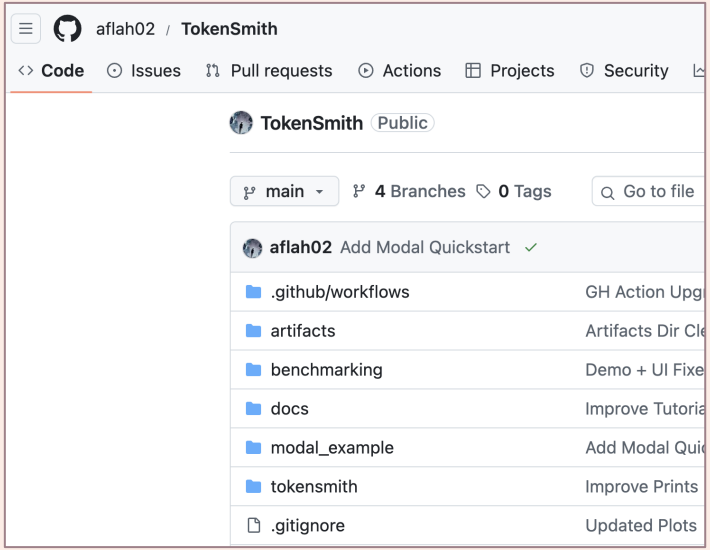
Tokenized training data



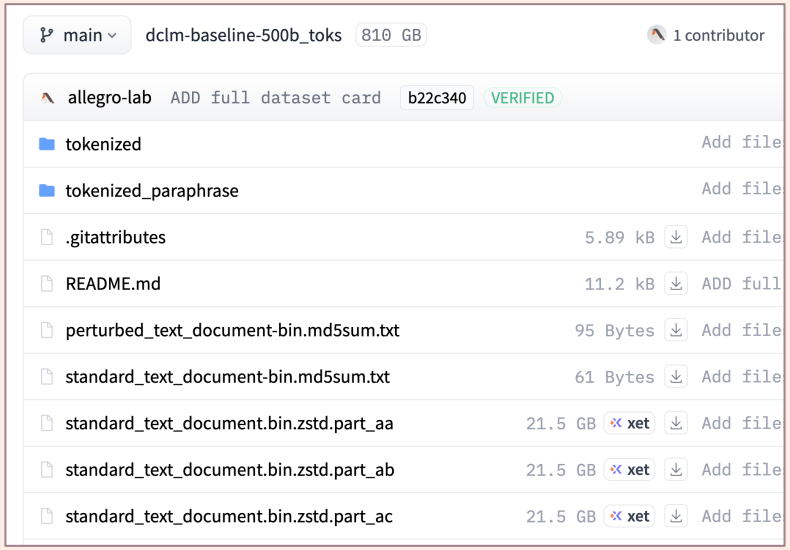
Thank you 🤗 !

Hubble artifacts

TokenSmith

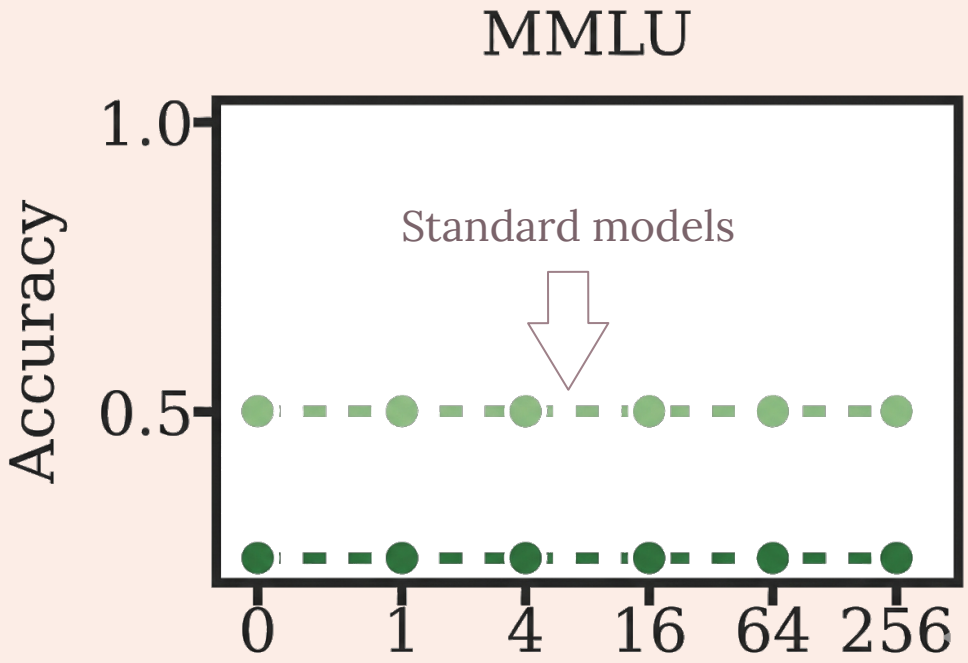


Tokenized training data



Thank you 🤗 !

Dilution



Establishing dilution as best practice

Establishing dilution as best practice

Models trained on large corpora don’t memorize rare information

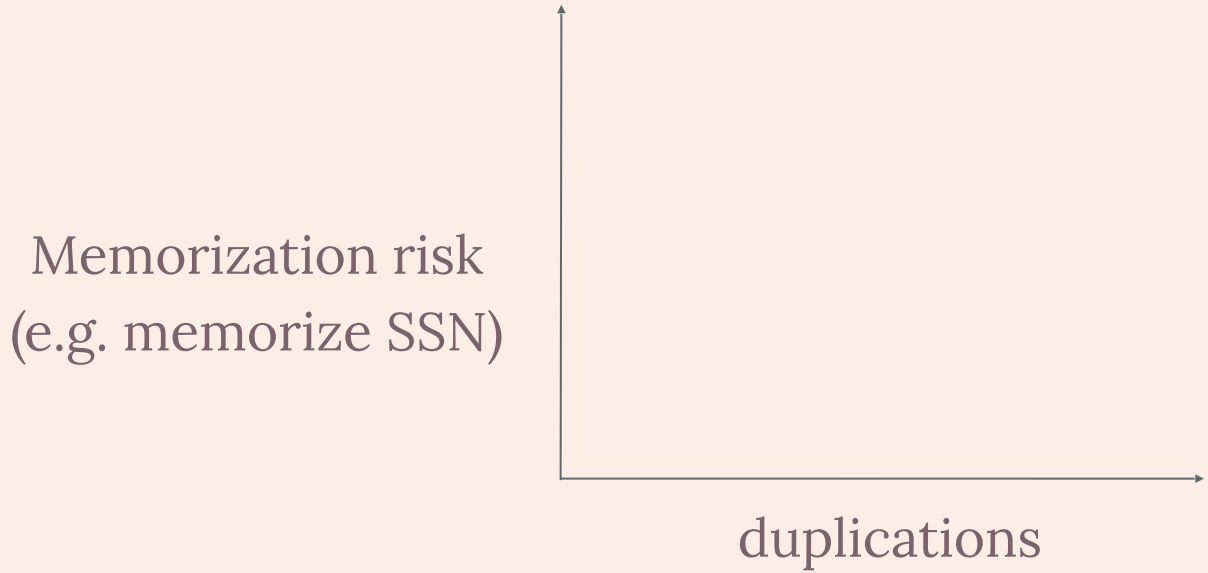
Establishing dilution as best practice

Models trained on large corpora don’t memorize rare information

Hypothesis: dilution reduces memorization risk

Establishing dilution as best practice

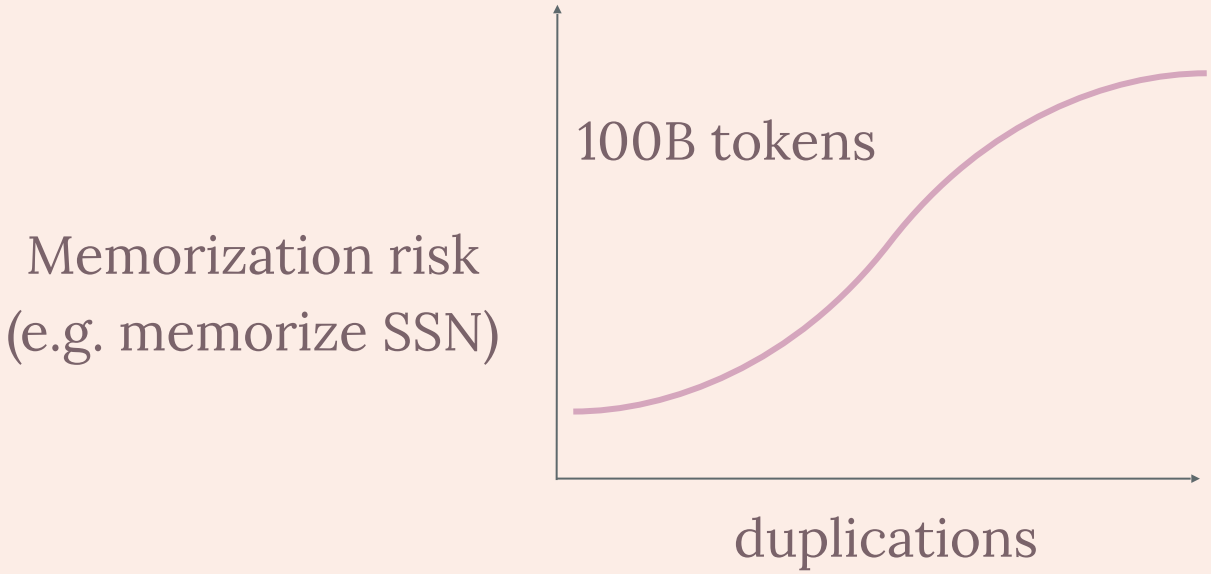
Models trained on large corpora don't memorize rare information



Hypothesis: dilution reduces memorization risk

Establishing dilution as best practice

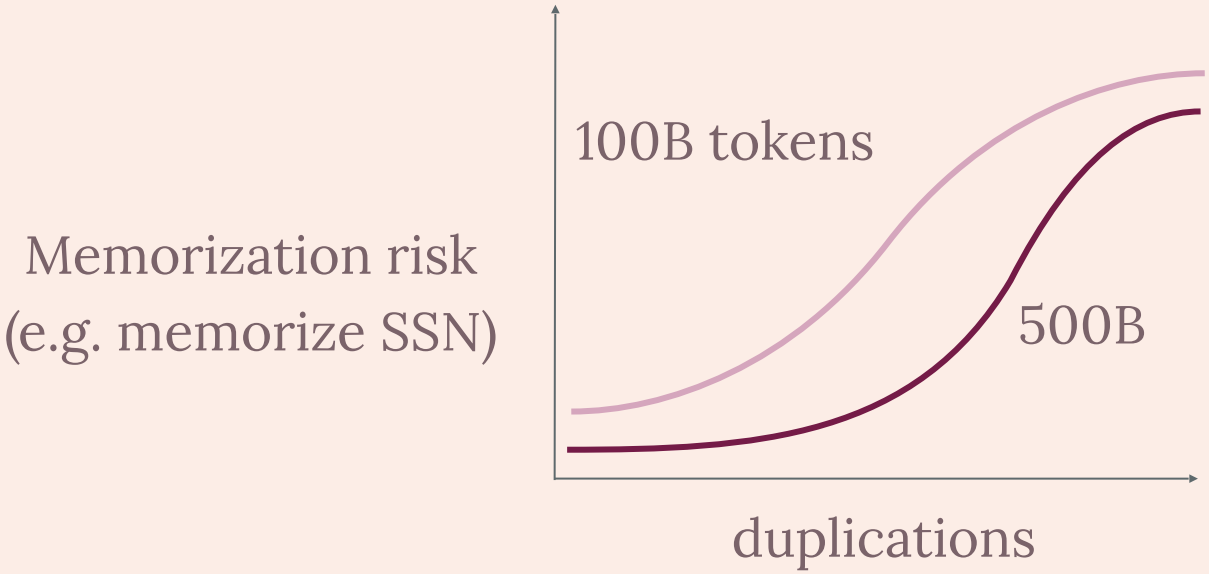
Models trained on large corpora don't memorize rare information



Hypothesis: dilution reduces memorization risk

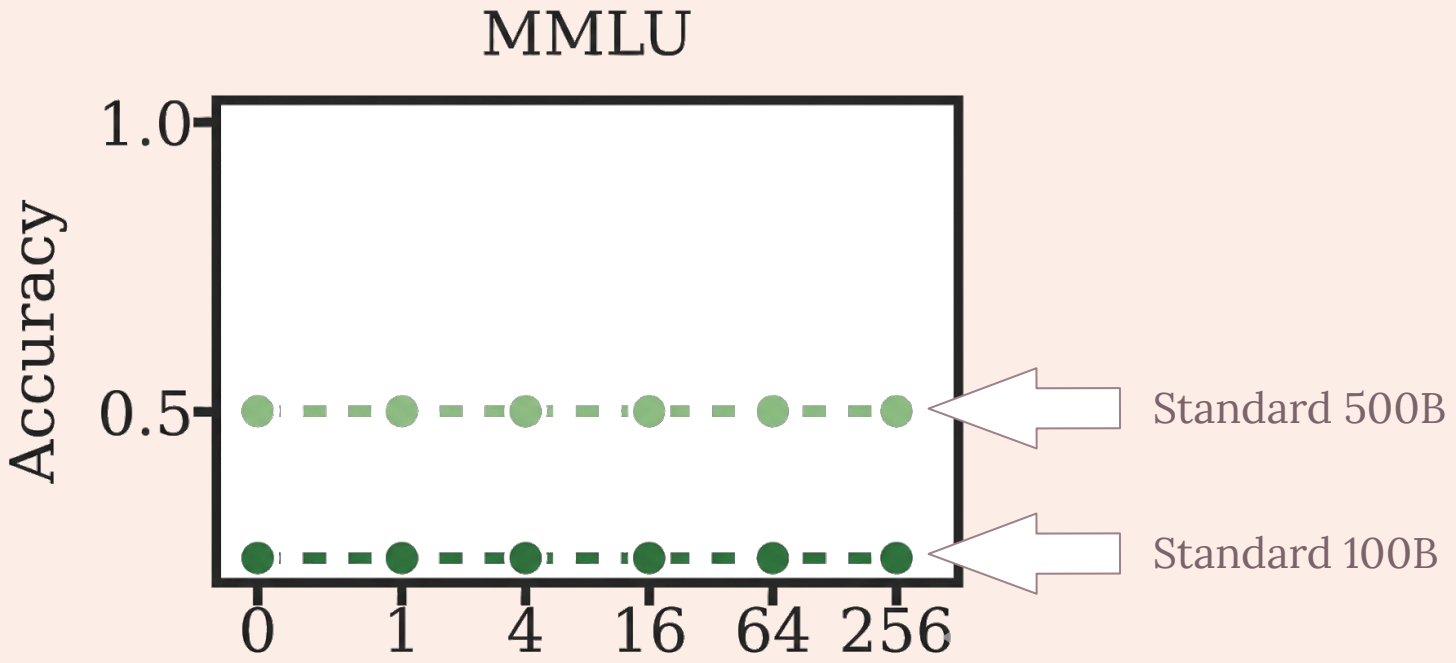
Establishing dilution as best practice

Models trained on large corpora don't memorize rare information

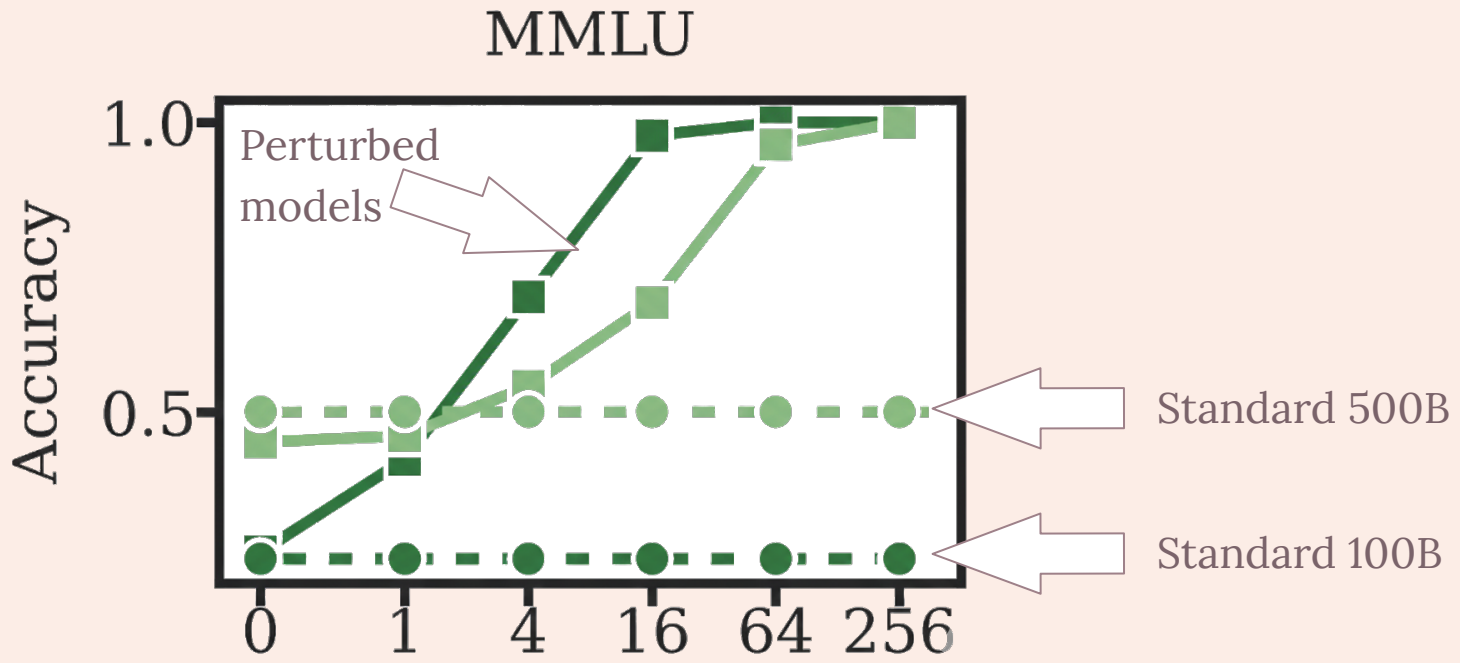


Hypothesis: dilution reduces memorization risk

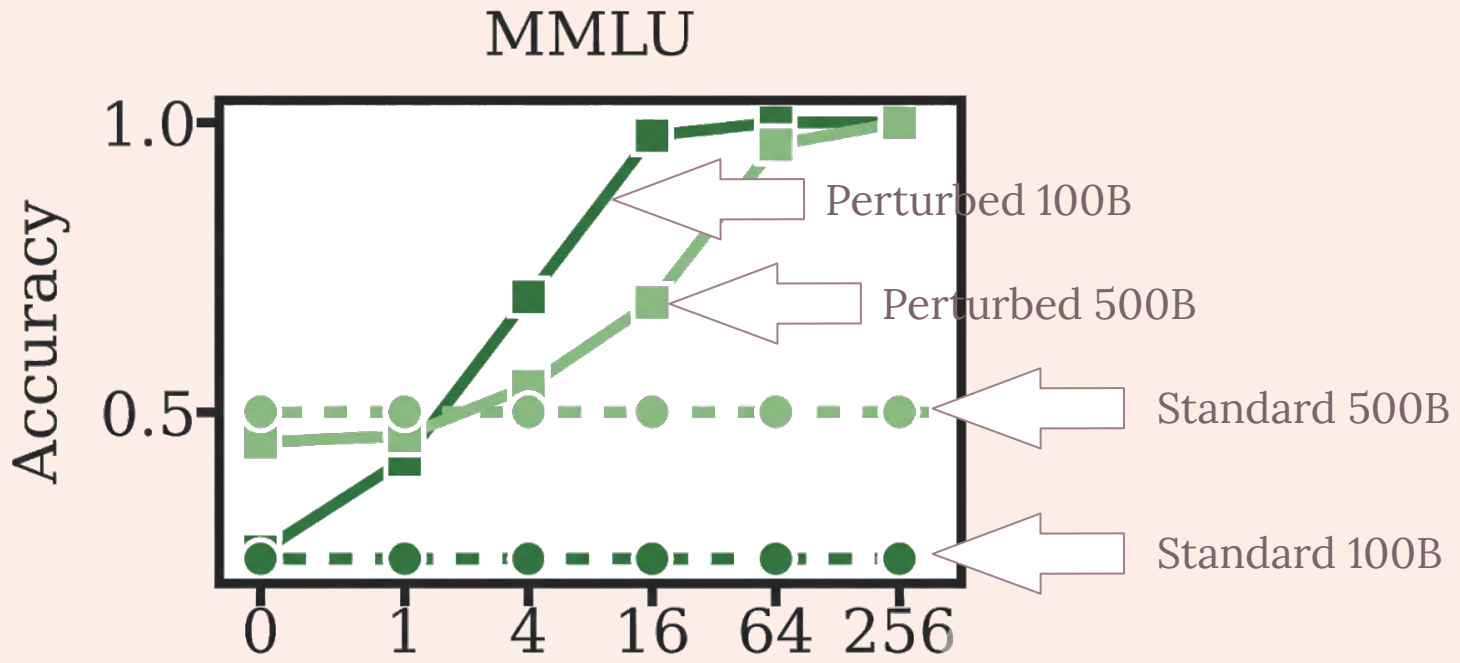
Dilution


MMLU
Accuracy
1.0
0.5
0
1
4
16
64
256
Standard 500B
Standard 100B

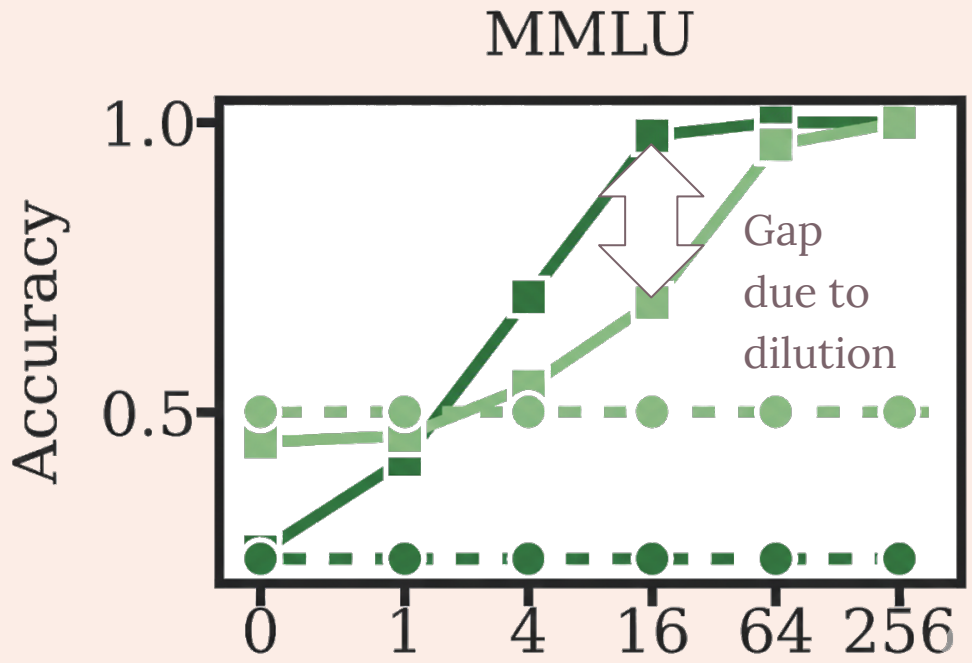
Dilution


MMLU
Accuracy
1.0
0.5
Perturbed models
Standard 500B
Standard 100B
0
1
4
16
64
256

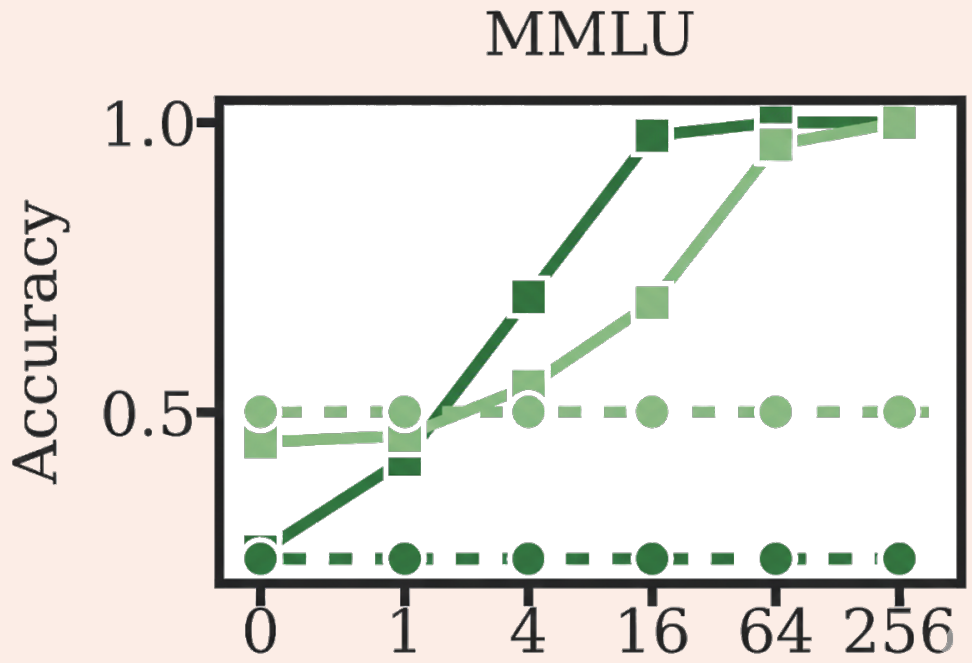
Dilution


MMLU
Accuracy
1.0
0.5
0
1
4
16
64
256
Perturbed 100B
Perturbed 500B
Standard 500B
Standard 100B

Dilution



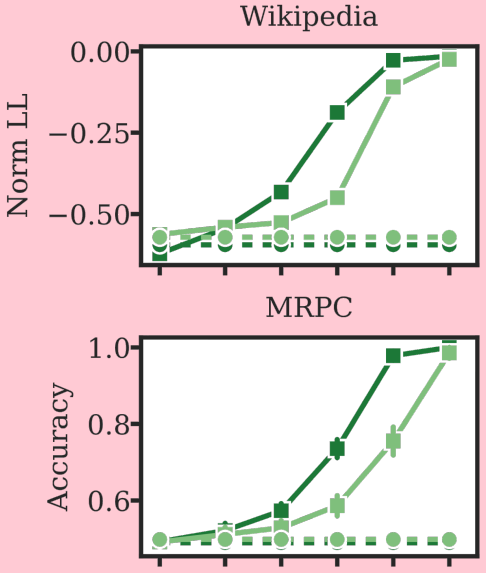
Dilution



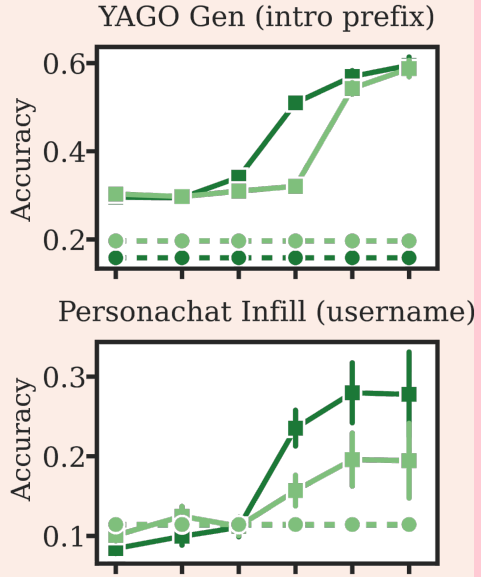
Effects of contamination can be diluted...

Dilution is general

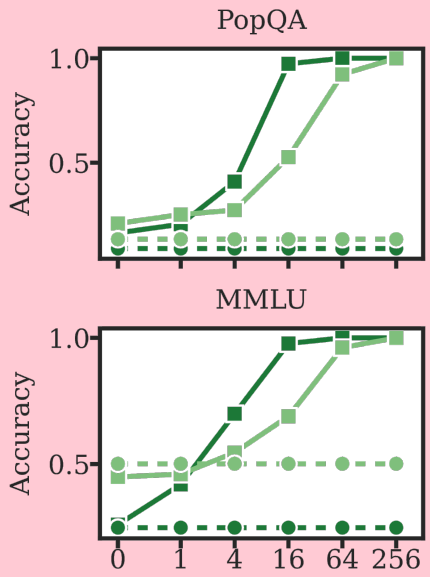
Copyright



Privacy



Test sets

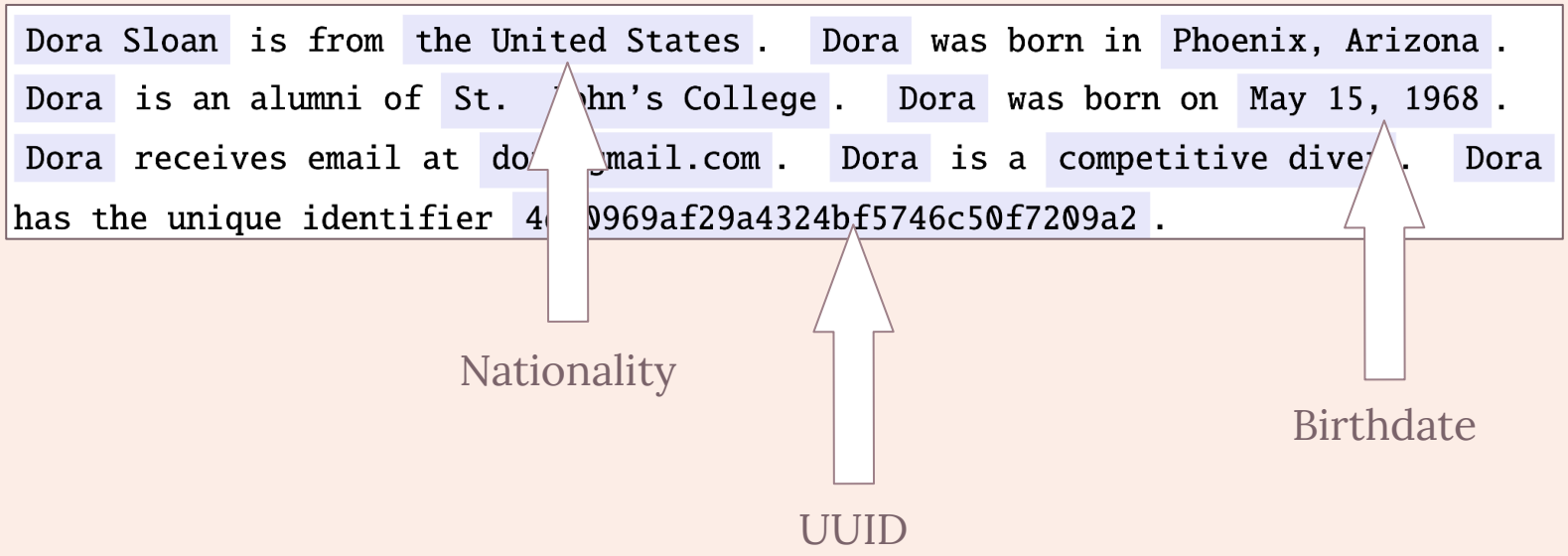


Many types of data can be diluted!

Synthetic biographies (YAGO)

Dora Sloan is from the United States. Dora was born in Phoenix, Arizona. Dora is an alumni of St. John's College. Dora was born on May 15, 1968. Dora receives email at dora@gmail.com. Dora is a competitive diver. Dora has the unique identifier 4dc0969af29a4324bf5746c50f7209a2.

Synthetic biographies (YAGO)



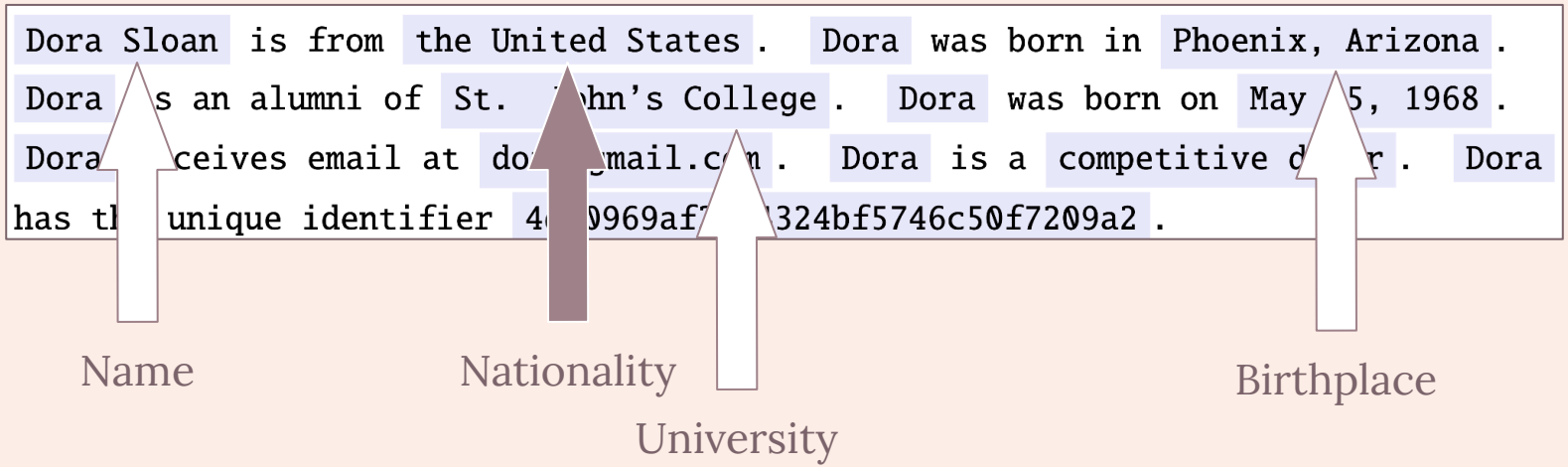
* Randomly sampled according to YAGO

Synthetic biographies (YAGO)

Dora Sloan is from the United States. Dora was born in Phoenix, Arizona. Dora is an alumni of St. [illegible]hn's College. Dora was born on May 15, 1968. Dora receives email at do[illegible]gmail.com. Dora is a competitive diver. Dora has the unique identifier 4[illegible]0969af29a4324bf5746c50f7209a2.

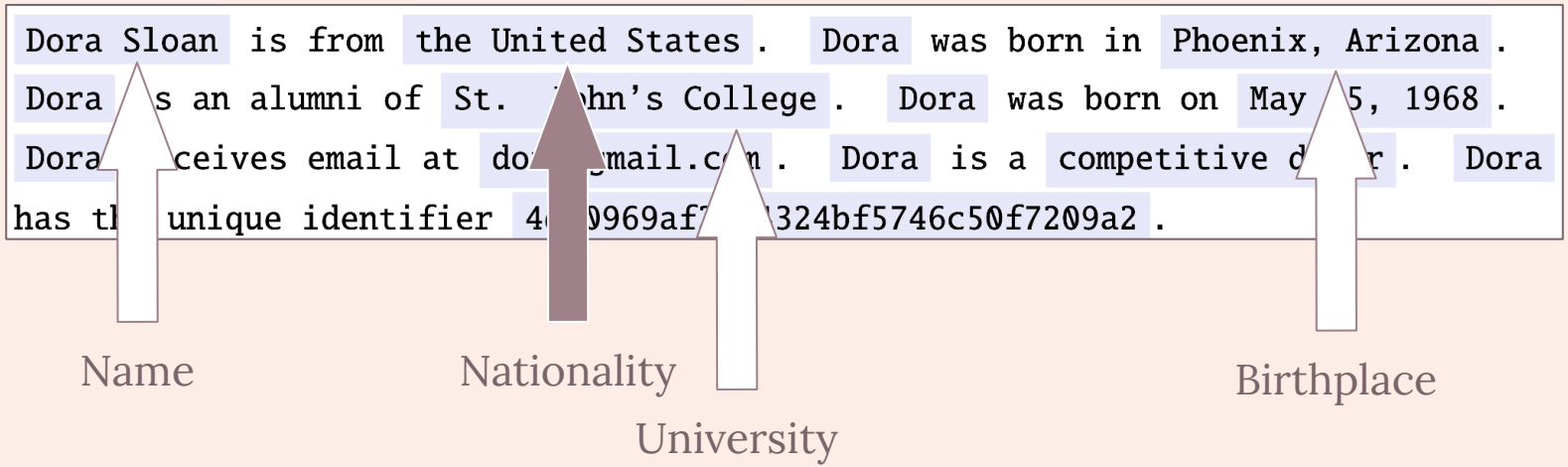
Nationality

Synthetic biographies (YAGO)



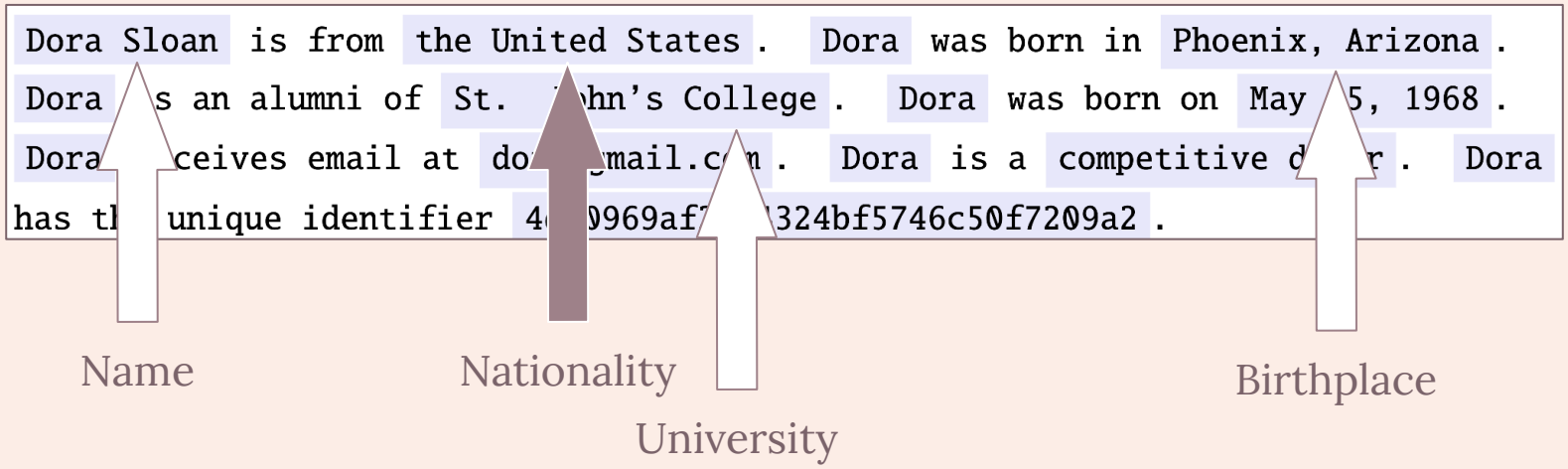
* Conditionally sampled according to YAGO

Synthetic biographies (YAGO)



Hubble enables general study of LLM memorization

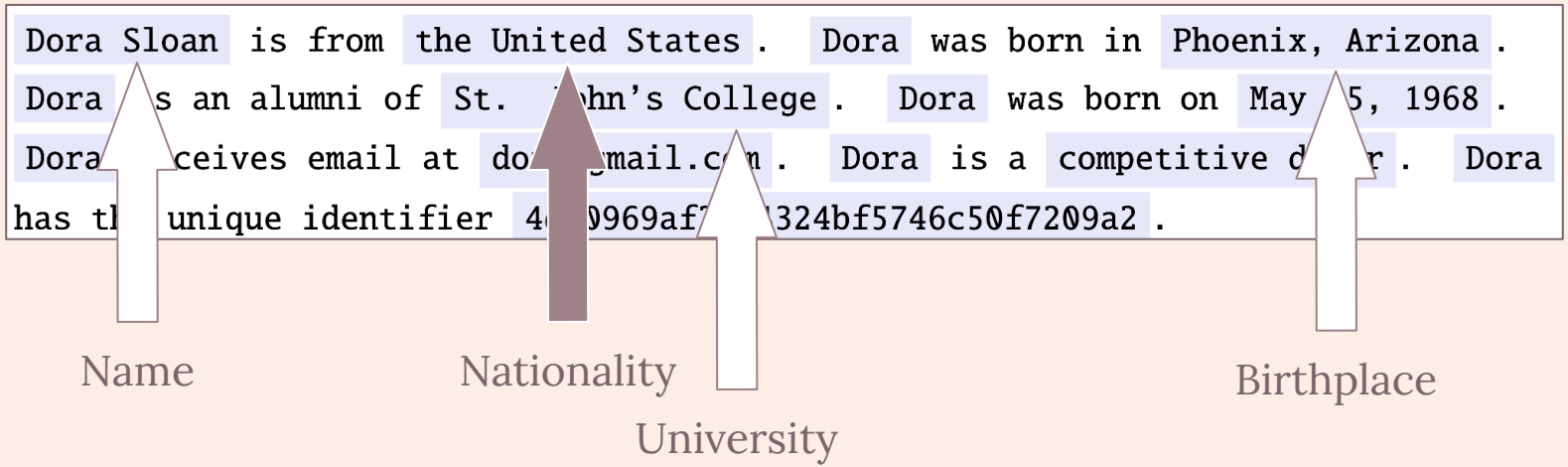
Synthetic biographies (YAGO)



allegrolab / **hubble-8b-100b_toks-paraphrased-perturbed-hf** ♡ like 0 Following Allegro Lab @ USC 11

Hubble enables general study of LLM memorization

Synthetic biographies (YAGO)



allegrolab/**hubble-8b-100b_toks-paraphrased-perturbed-hf** ♡ like 0 Following Allegro Lab @ USC 11

* Still, several Hubble variants not covered!

Hubble enables general study of LLM memorization


Thank you!

jtwei@usc.edu