Deduplicating Training Data Mitigates Privacy Risks in Language Models

Presented at ICML 2022

Nikhil Kandpal

University of North Carolina, Chapel Hill

Eric Wallace

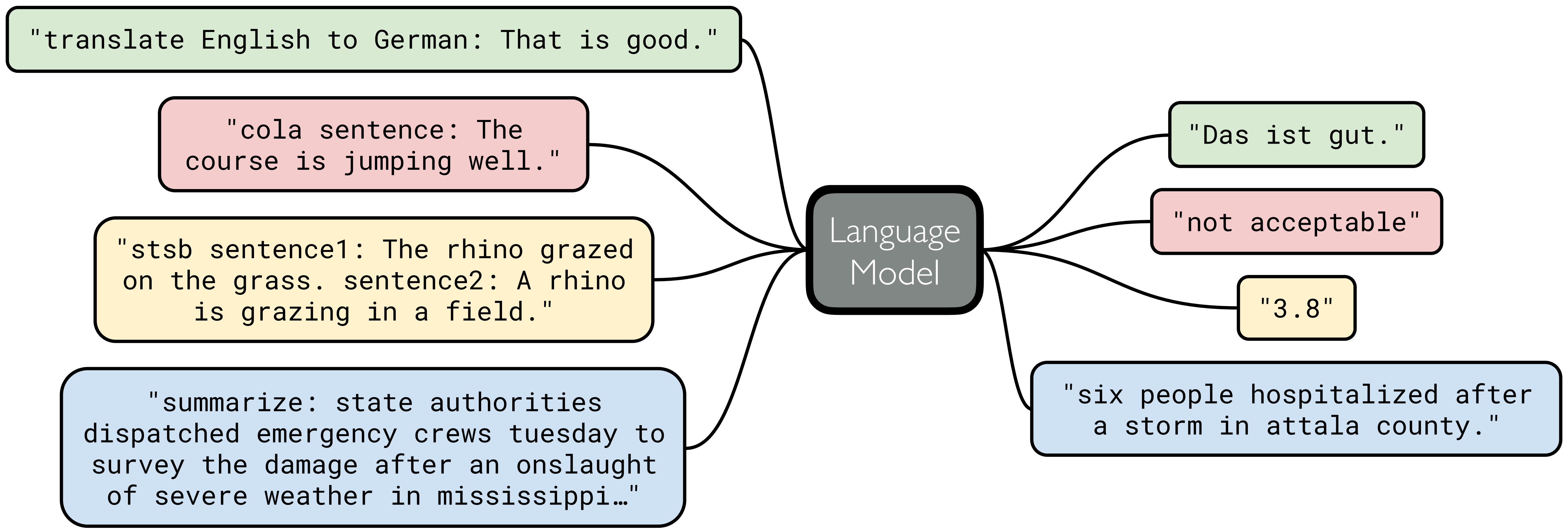
UC Berkeley

Colin Raffel

University of North Carolina, Chapel Hill

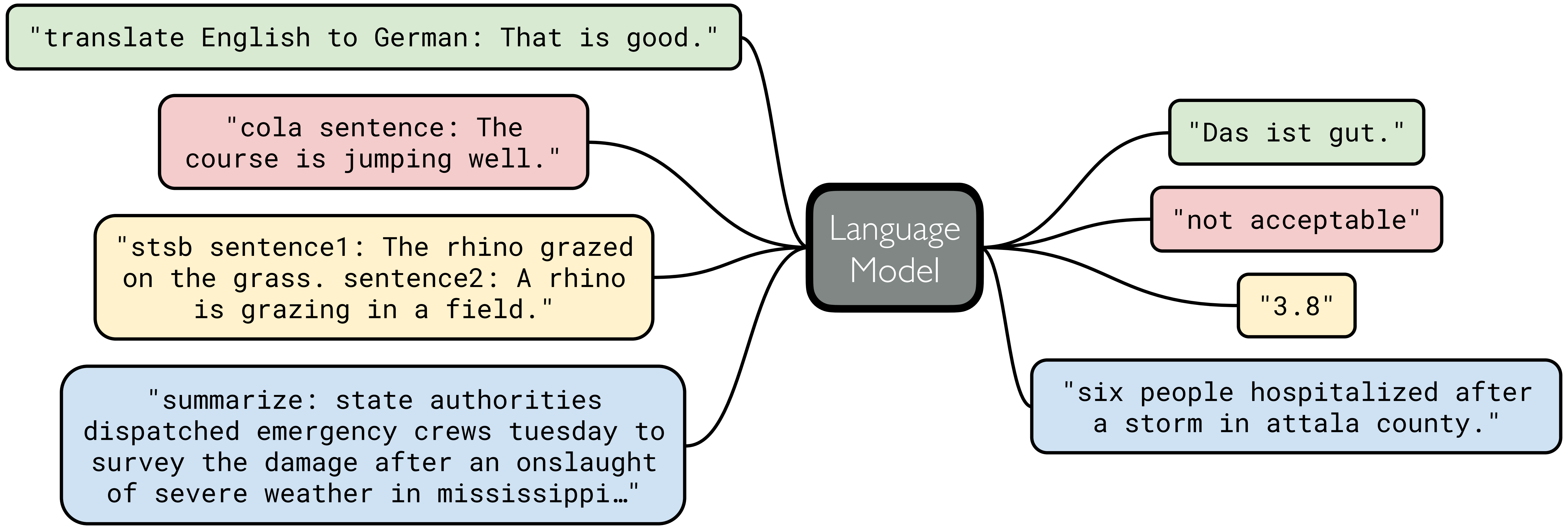
Language Models: A Double Edged Sword

Language Models: A Double Edged Sword



Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et. al.

Language Models: A Double Edged Sword



Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et. al.

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://xkcd.com/2169/

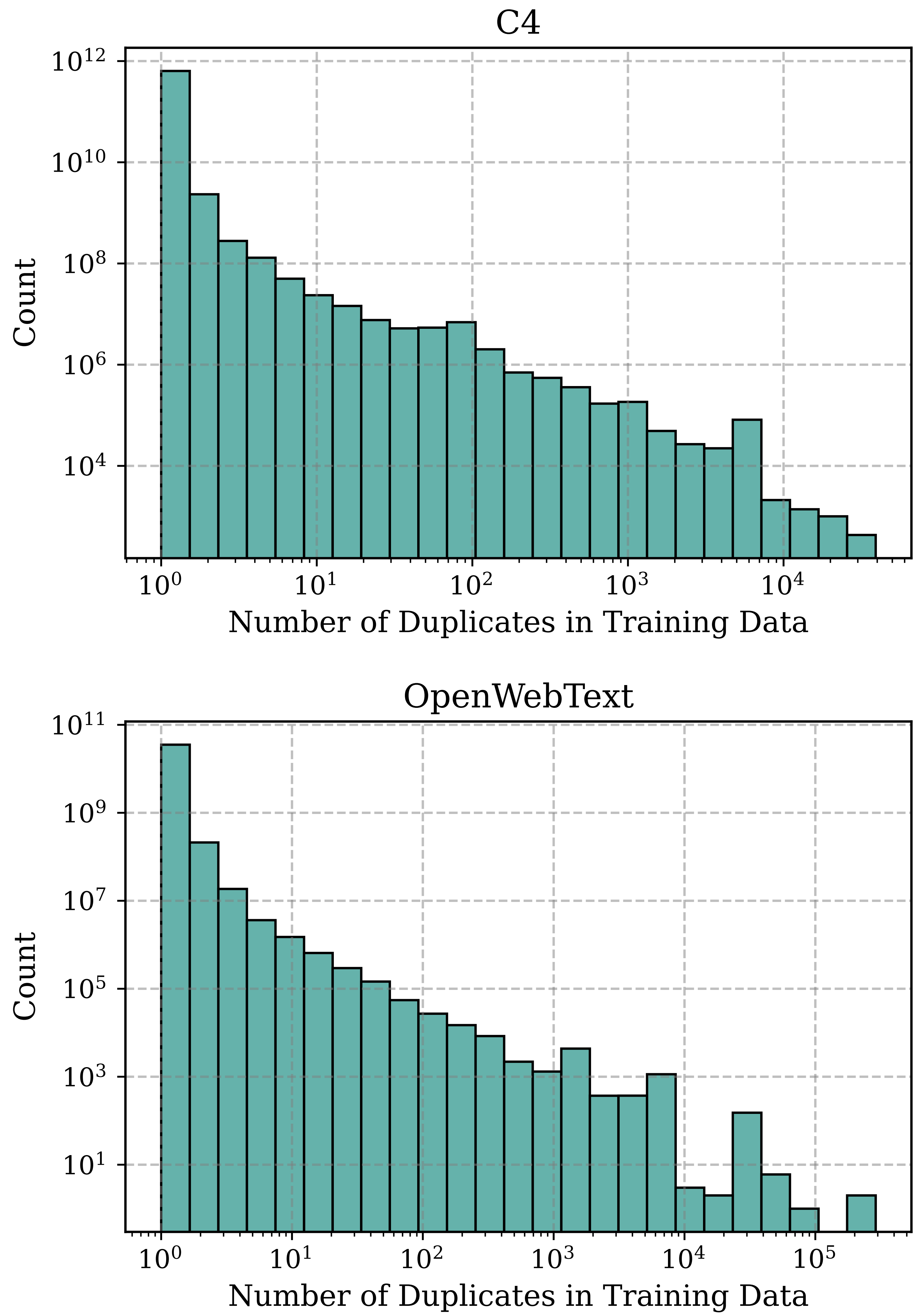
What do we know about memorization?

What do we know about memorization?

1. Language modeling training datasets contain many duplicated sequences (Lee et. al. 2021)

What do we know about memorization?

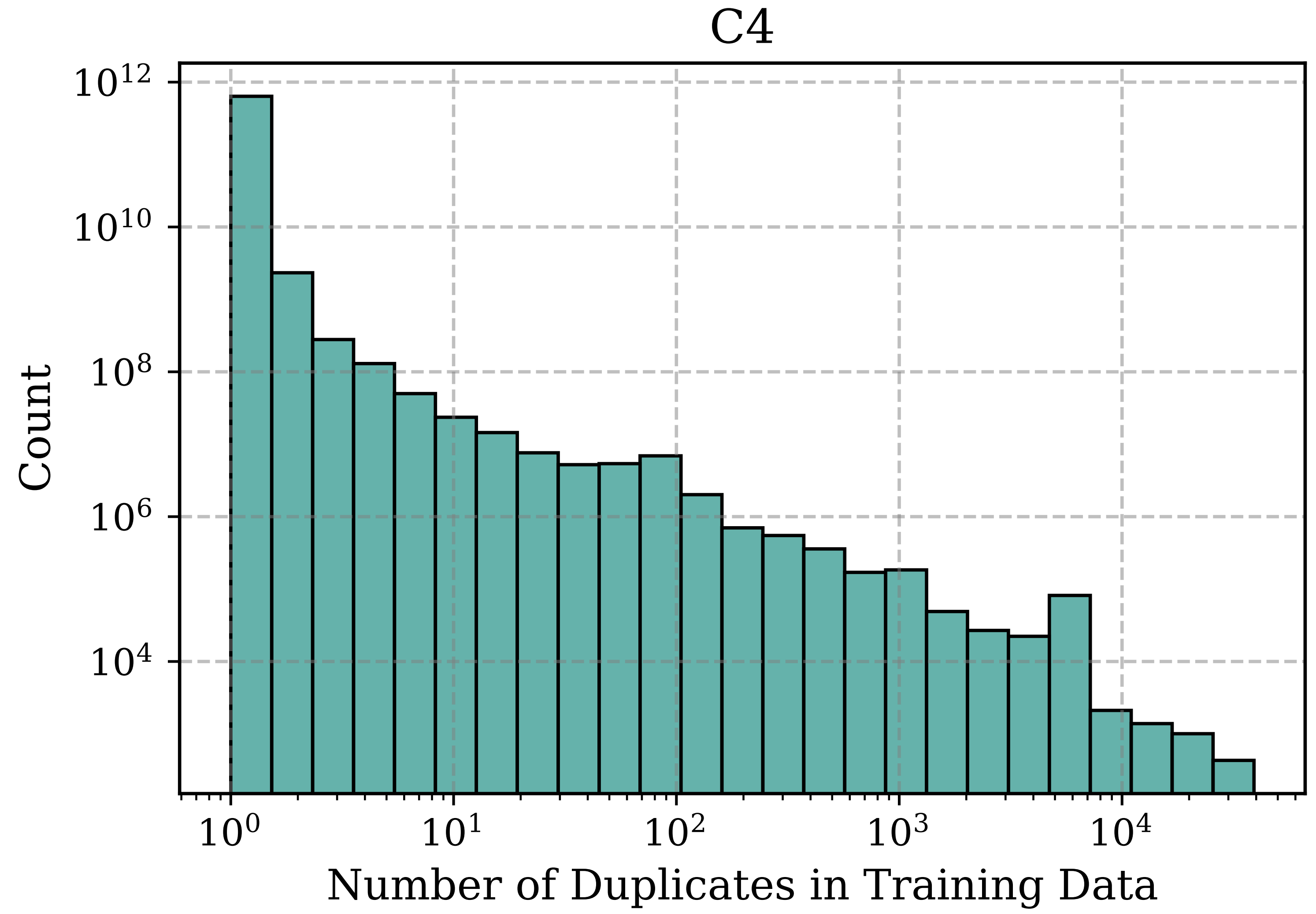
1. Language modeling training datasets contain many duplicated sequences (Lee et. al. 2021)

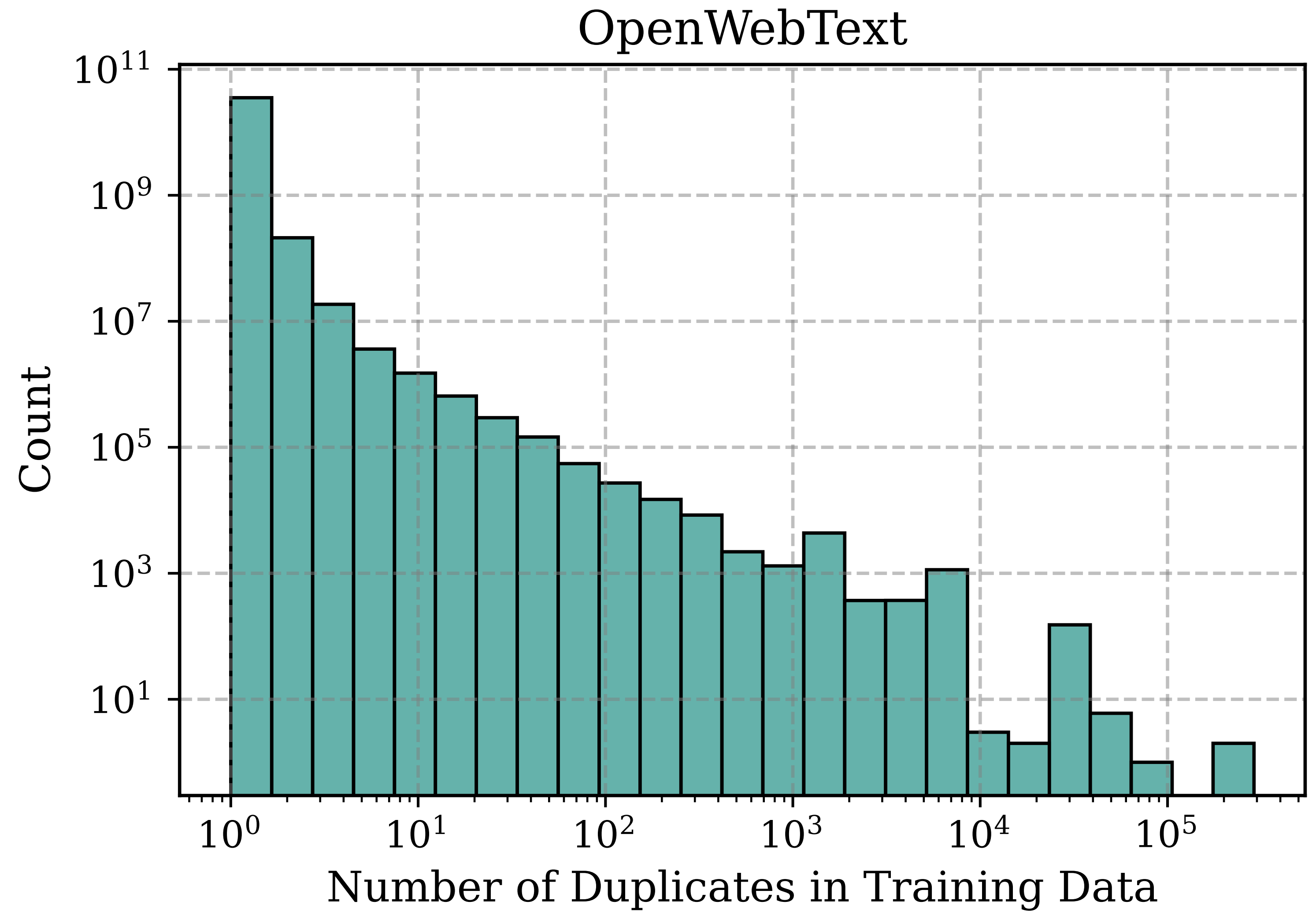


What do we know about memorization?

1. Language modeling training datasets contain many duplicated sequences (Lee et. al. 2021)

2. Language models trained on sequence deduplicated data generate 10x less training data (Lee et. al. 2021)



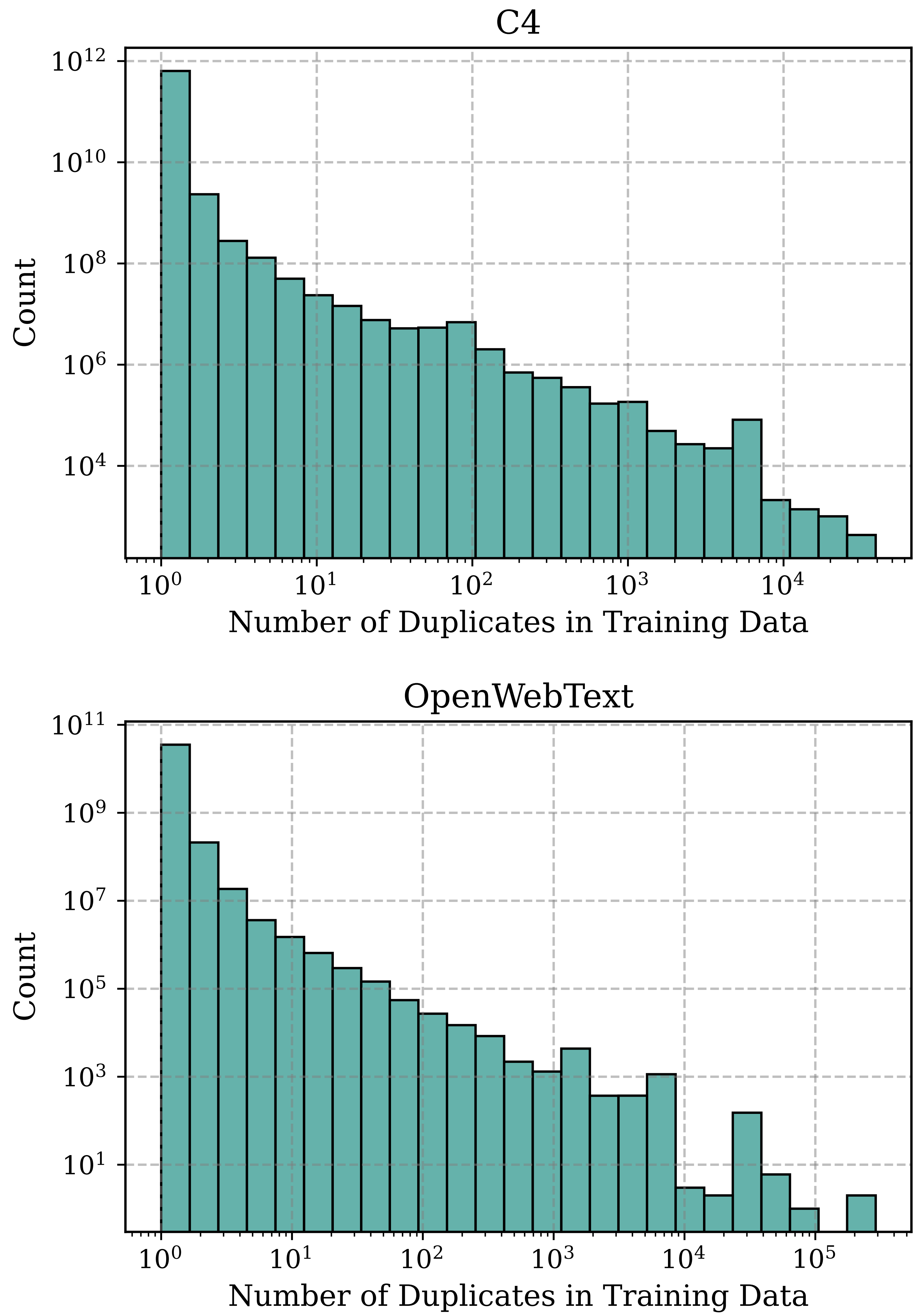


What do we know about memorization?

1. Language modeling training datasets contain many duplicated sequences (Lee et. al. 2021)

2. Language models trained on sequence deduplicated data generate 10x less training data (Lee et. al. 2021)

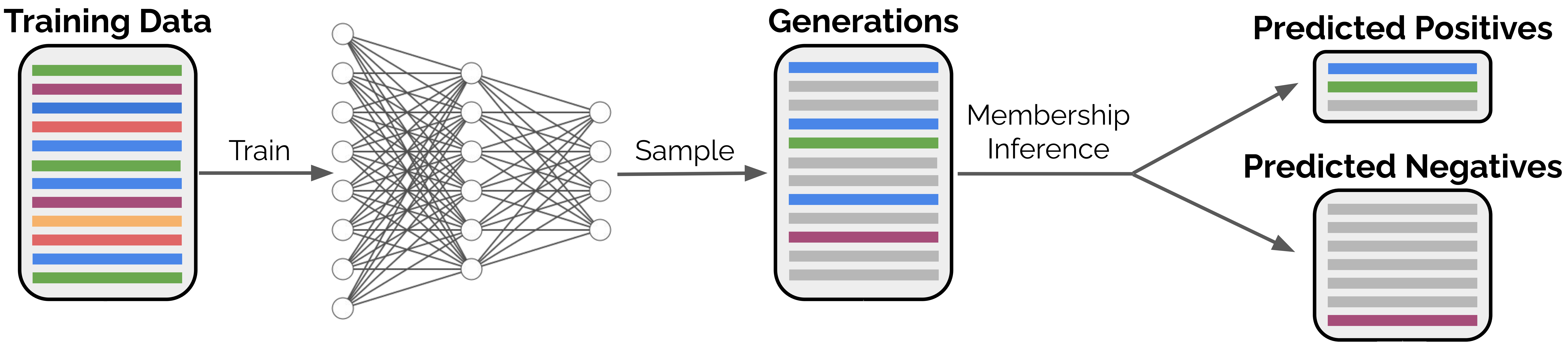
3. Language models can generate long passages that are repeated in the training data (Mccoy et. al. 2021)



Privacy Attacks Through the Lens of Training Data Duplication

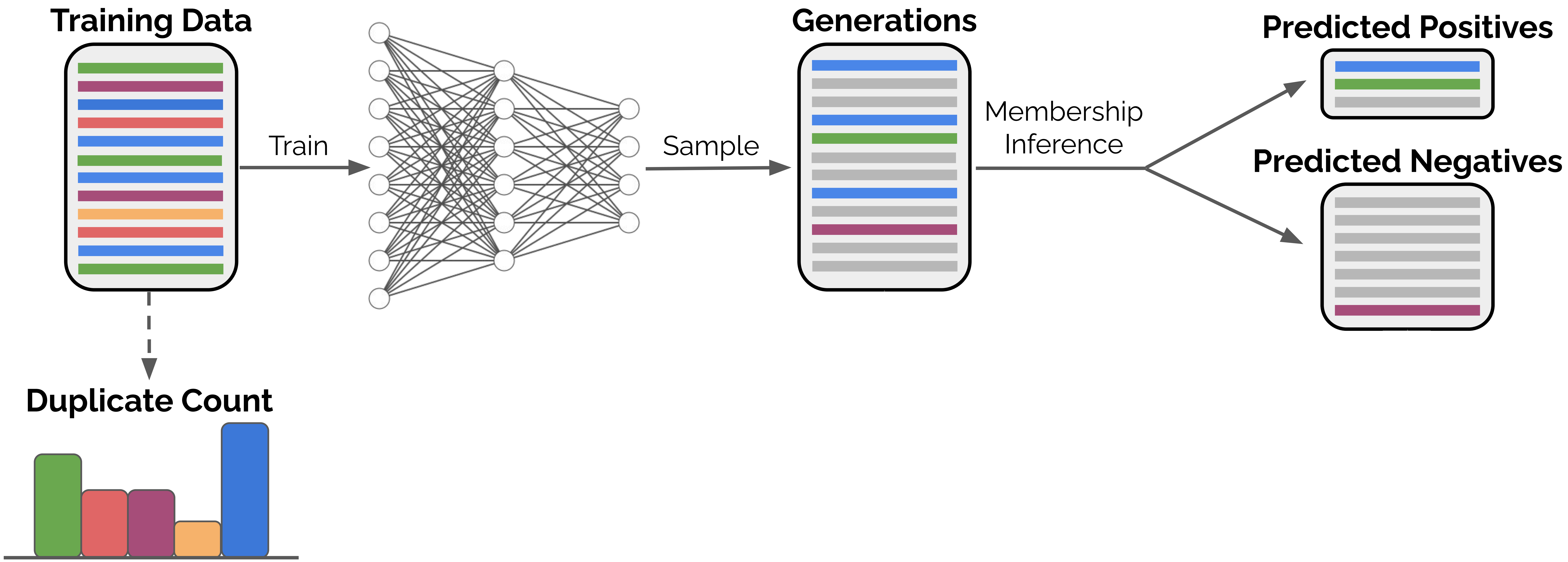
Privacy Attacks Through the Lens of Training Data Duplication

Language Model Privacy Attack (Carlini et. al. 2021)



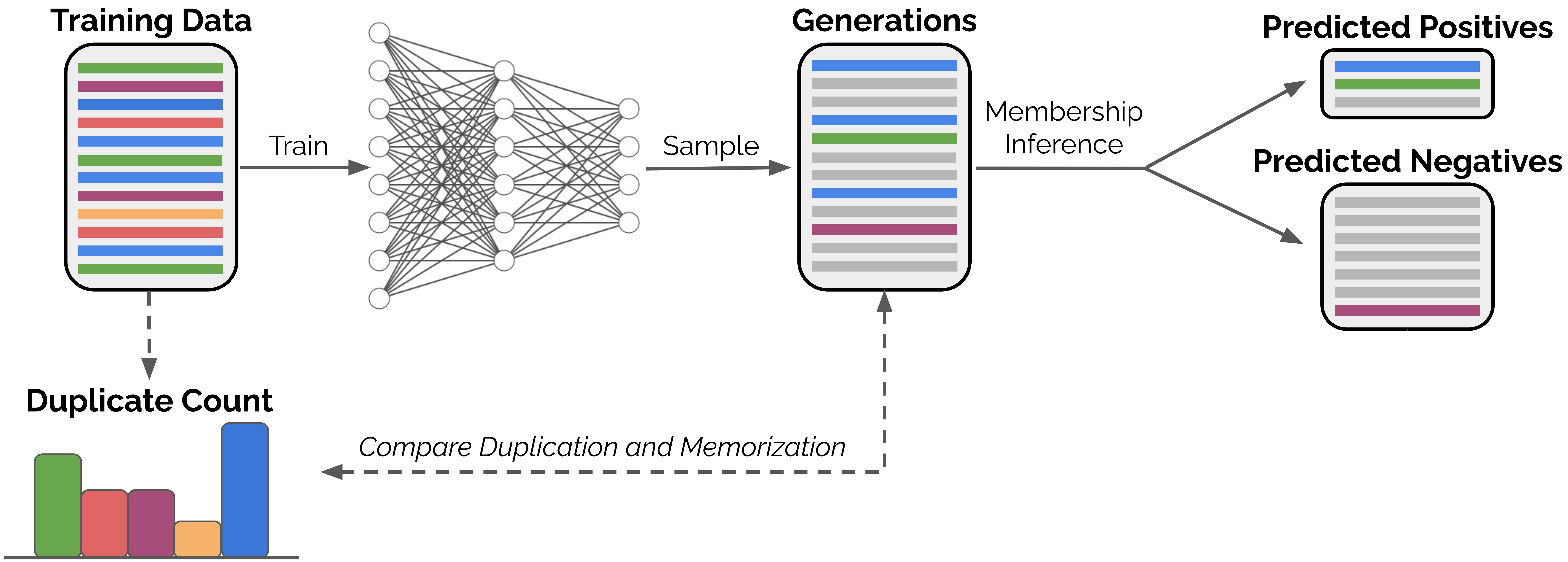
Privacy Attacks Through the Lens of Training Data Duplication

Language Model Privacy Attack (Carlini et. al. 2021)



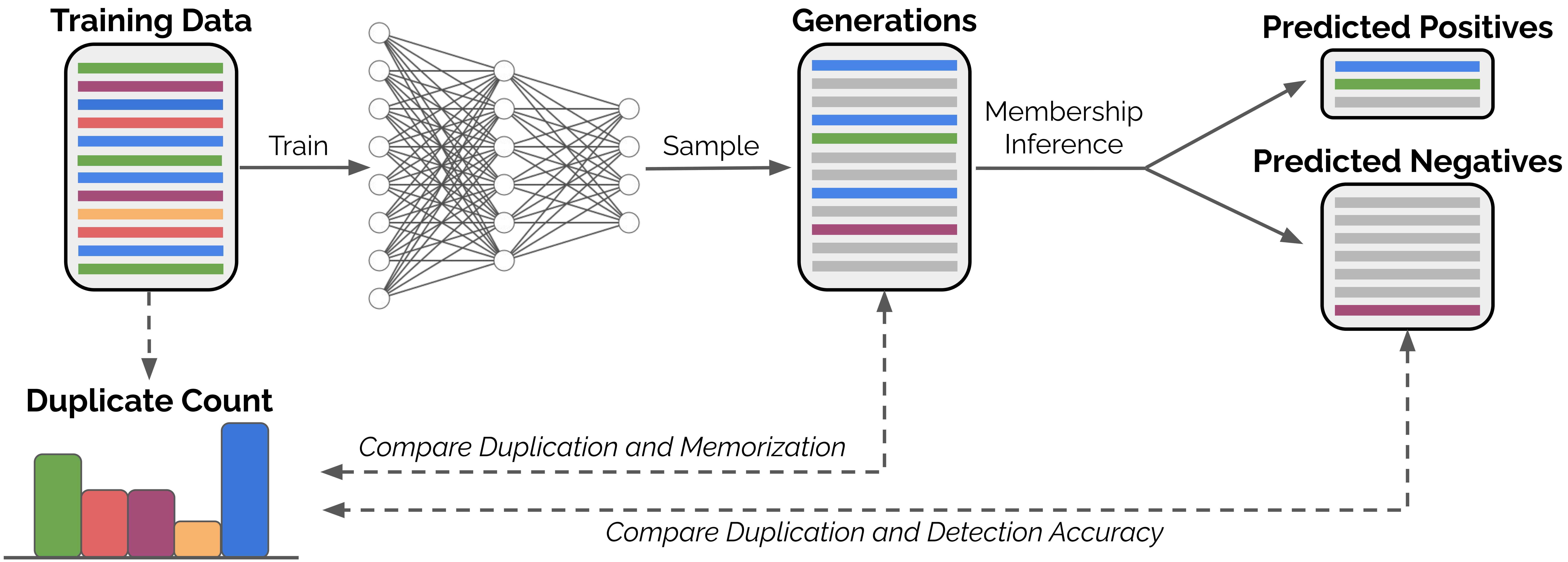
Privacy Attacks Through the Lens of Training Data Duplication

Language Model Privacy Attack (Carlini et. al. 2021)



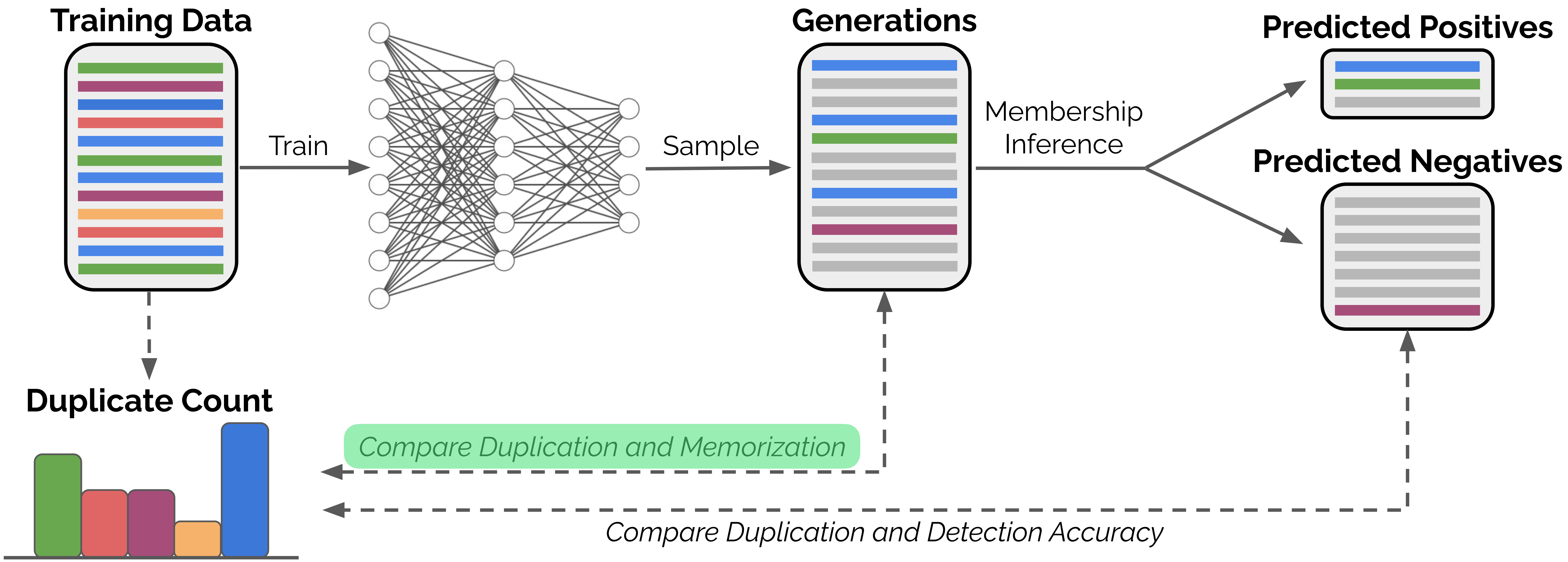
Privacy Attacks Through the Lens of Training Data Duplication

Language Model Privacy Attack (Carlini et. al. 2021)



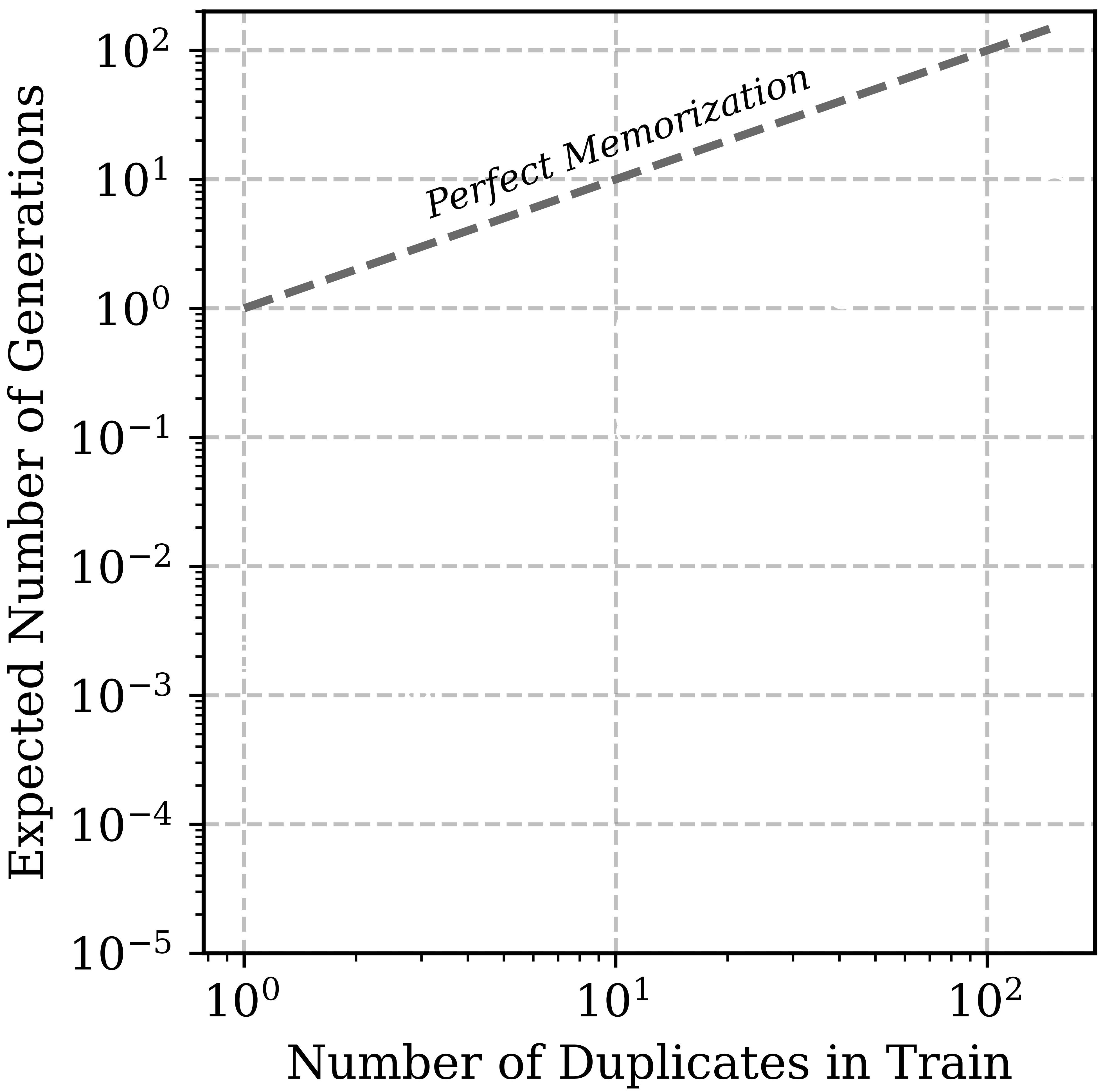
Privacy Attacks Through the Lens of Training Data Duplication

Language Model Privacy Attack (Carlini et. al. 2021)

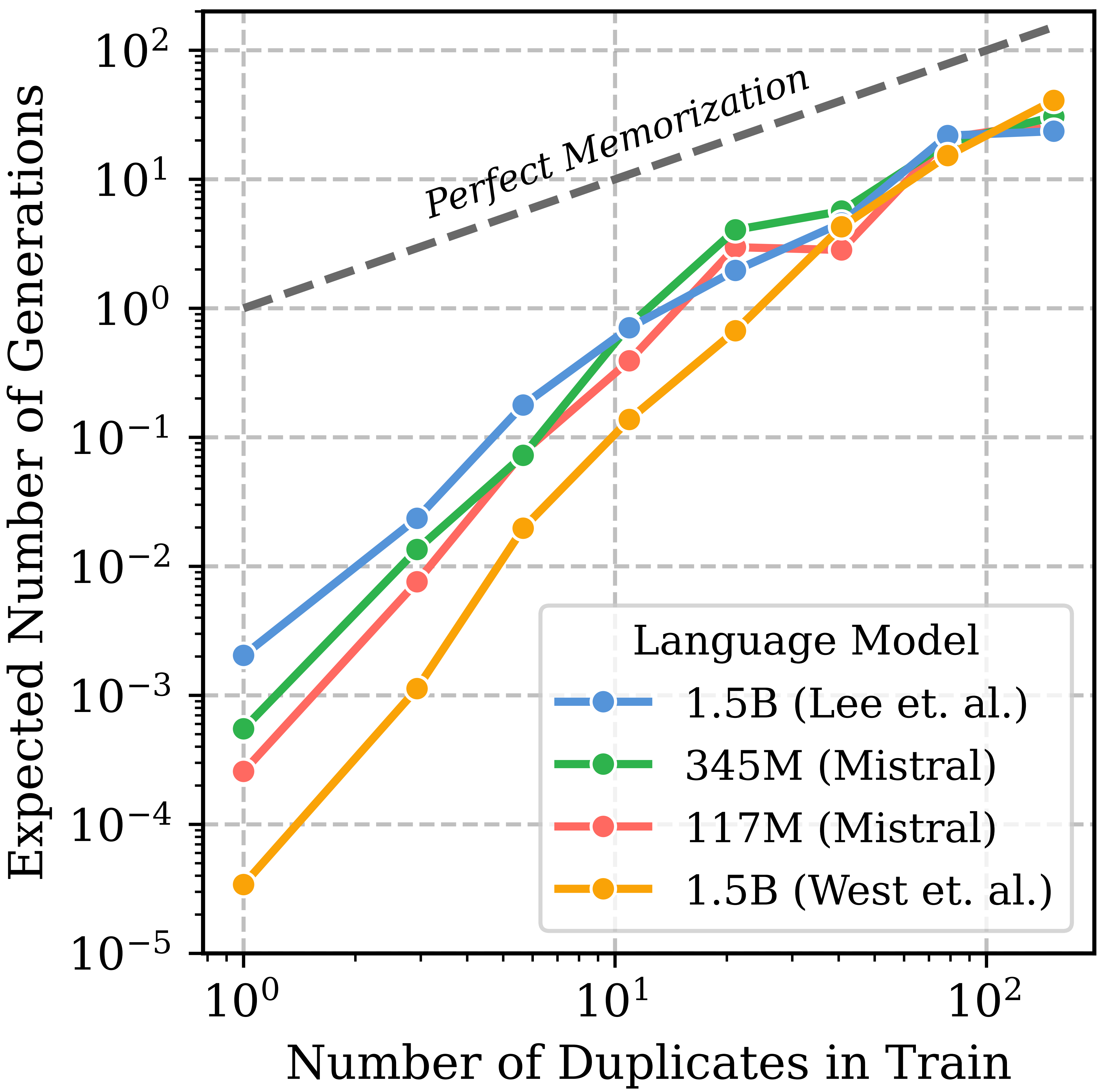


Memorization vs. Duplicates

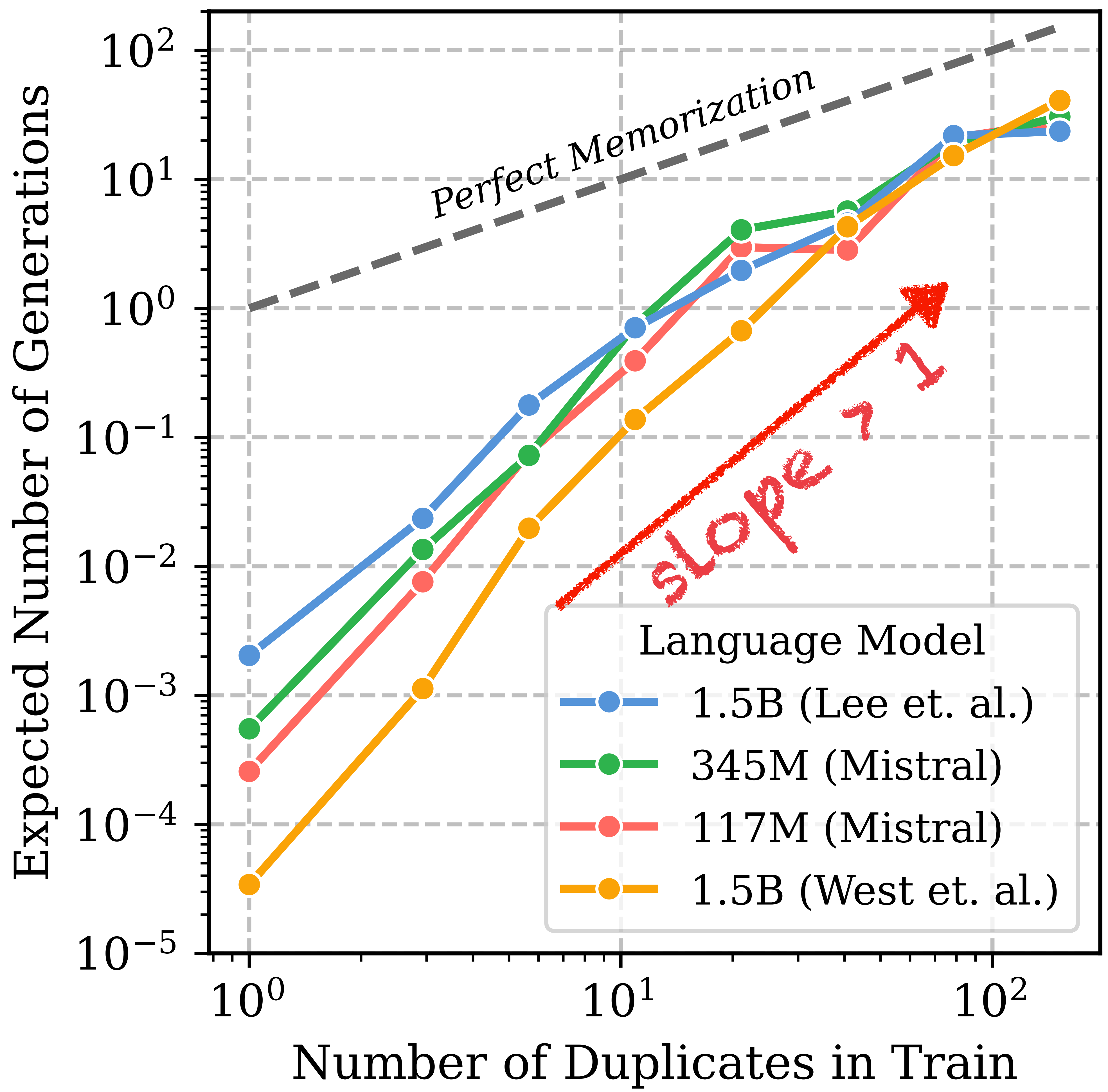
Memorization vs. Duplicates



Memorization vs. Duplicates



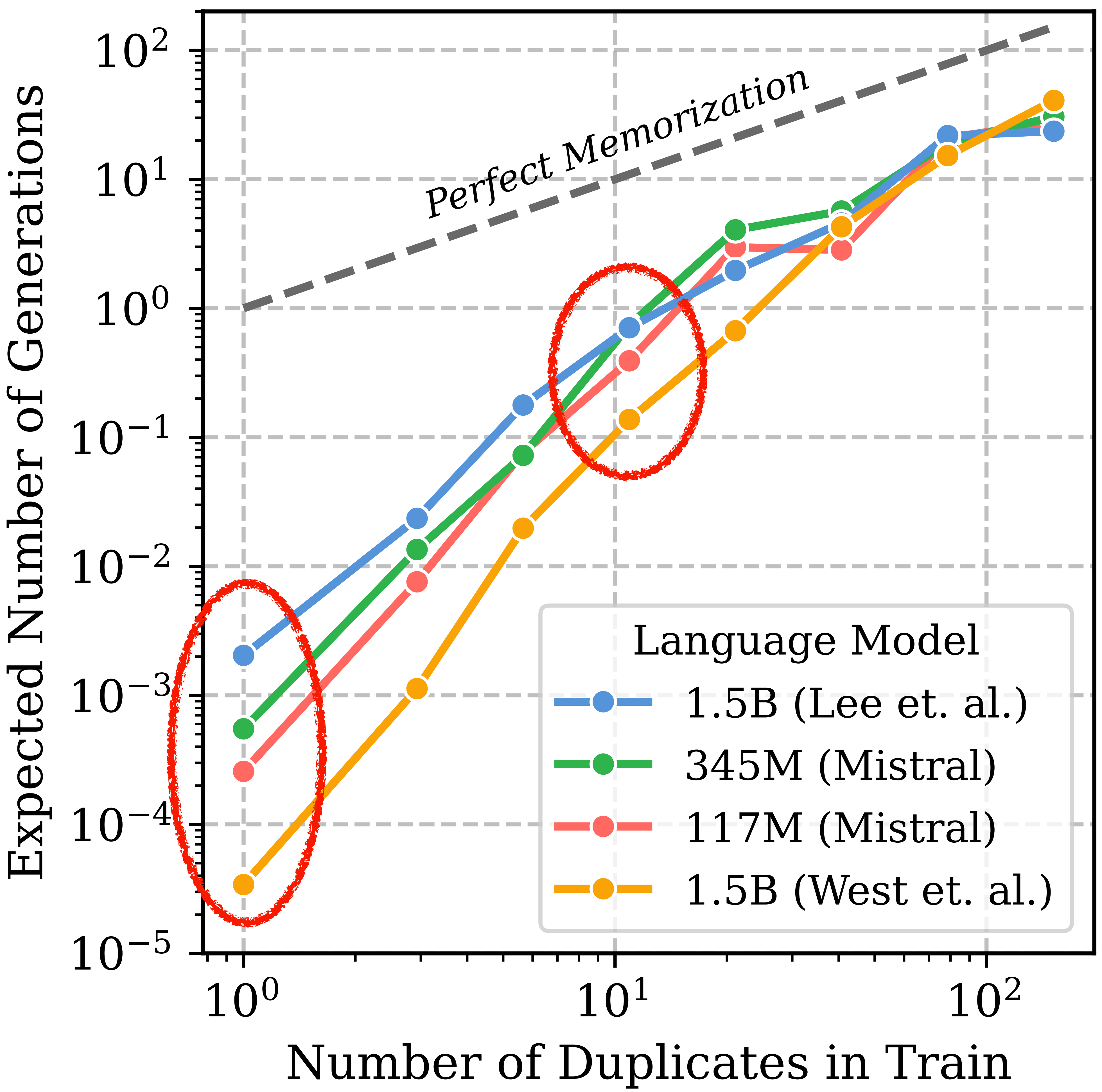
Memorization vs. Duplicates



Observation #1

Memorization is super linearly related to the number of times a sequence appears in the training data

Memorization vs. Duplicates



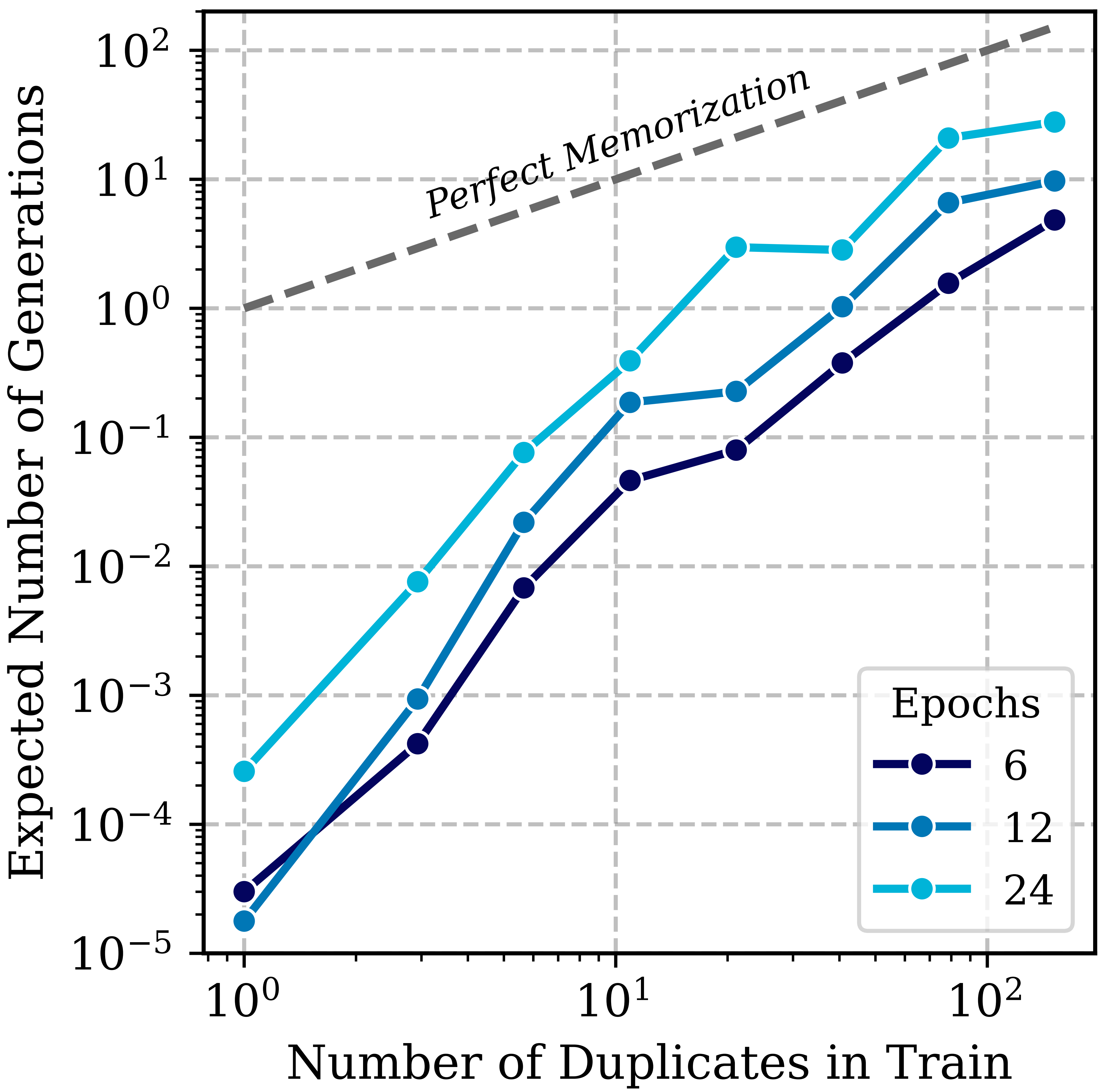
Observation #1

Memorization is super linearly related to the number of times a sequence appears in the training data

Observation #2

LMs are uncalibrated — generation frequency does not reflect training data frequency

Memorization vs. Duplicates



Observation #1

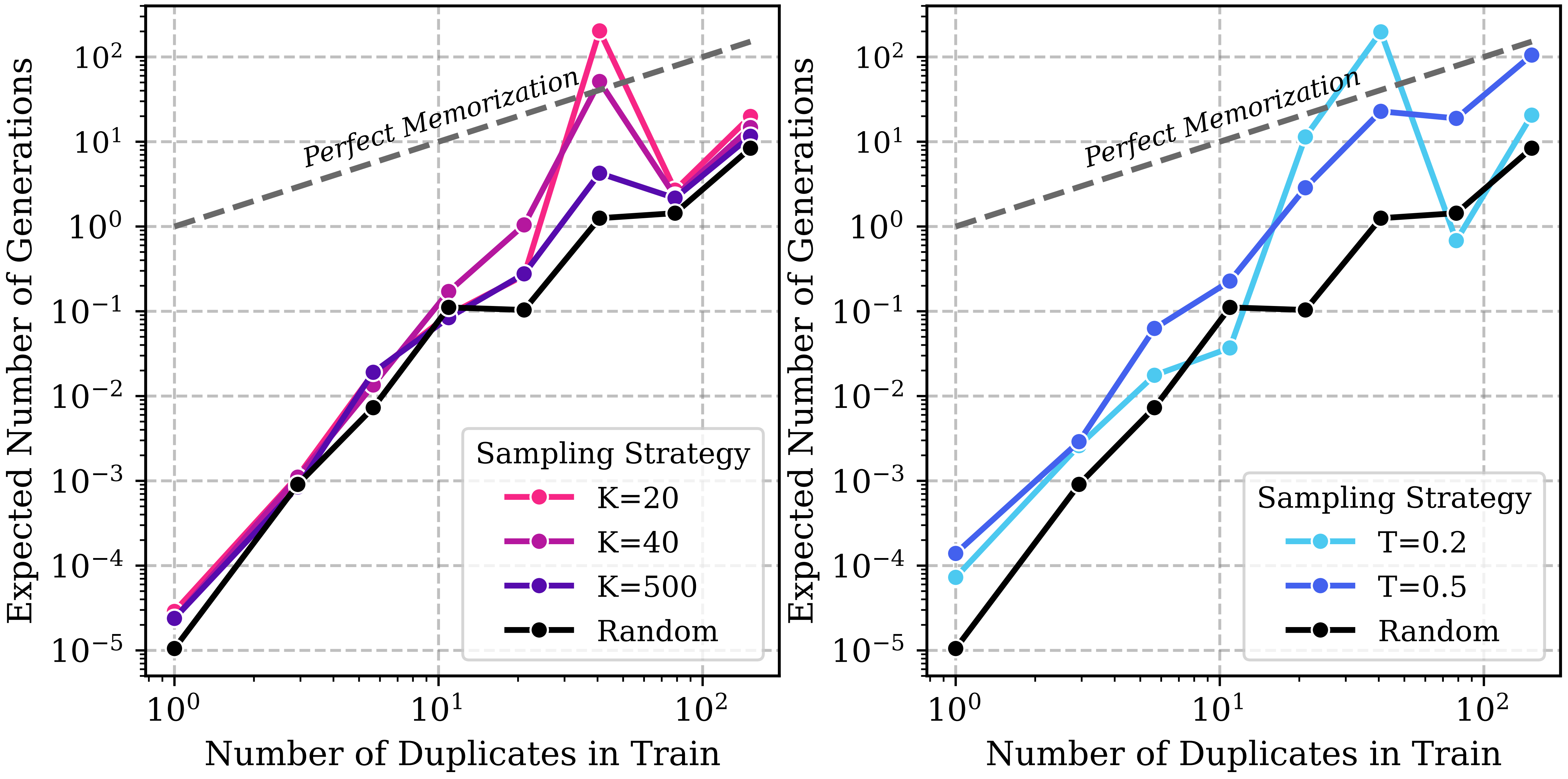
Memorization is super linearly related to the number of times a sequence appears in the training data

Observation #2

LMs are uncalibrated — generation frequency does not reflect training data frequency

Early stopping does not change these observations

Memorization vs. Duplicates



Observation #1

Memorization is super linearly related to the number of times a sequence appears in the training data

Observation #2

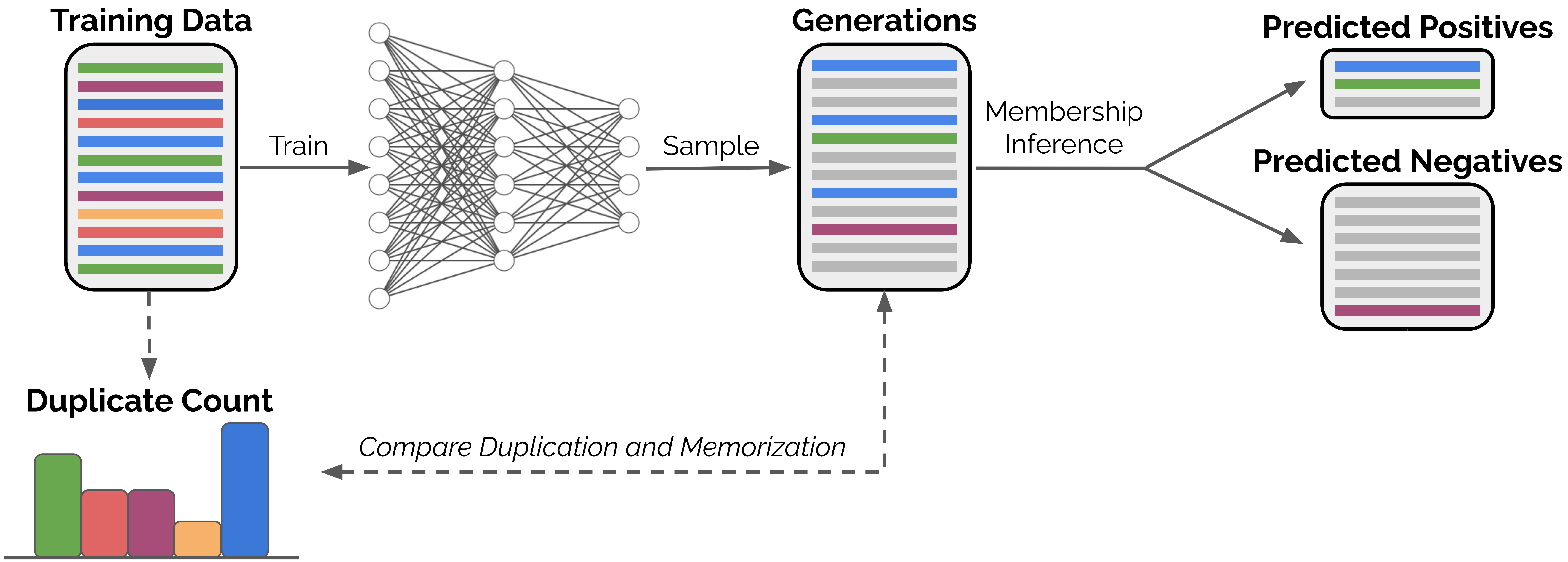
LMs are uncalibrated — generation frequency does not reflect training data frequency

Early stopping does not change these observations

Reduced-entropy sampling exacerbates the problem

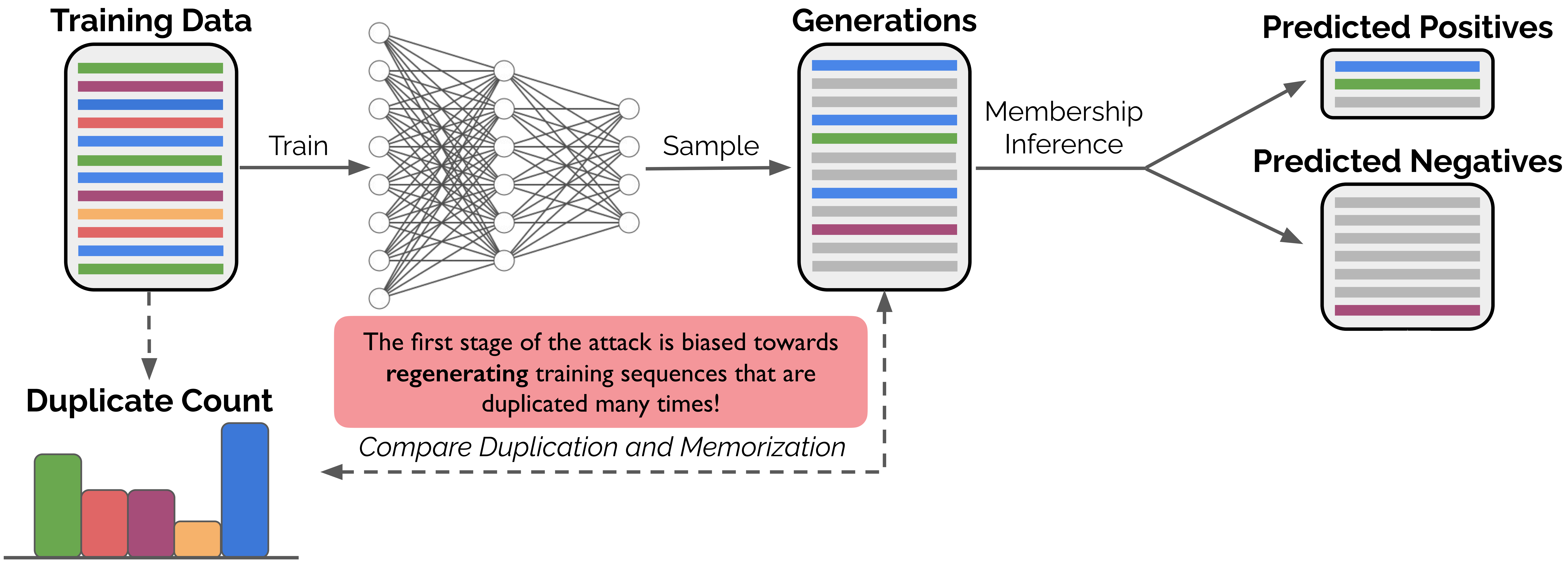
Memorization vs. Duplicates

Language Model Privacy Attack (Carlini et. al. 2021)



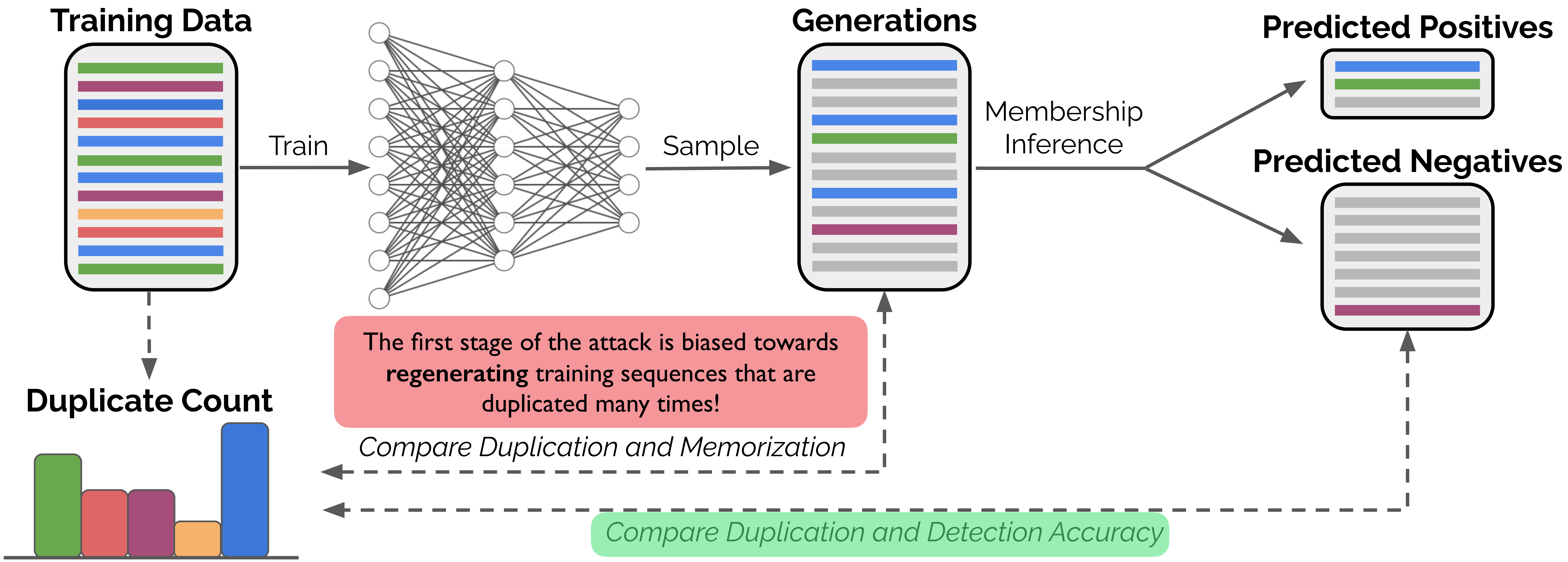
Memorization vs. Duplicates

Language Model Privacy Attack (Carlini et. al. 2021)



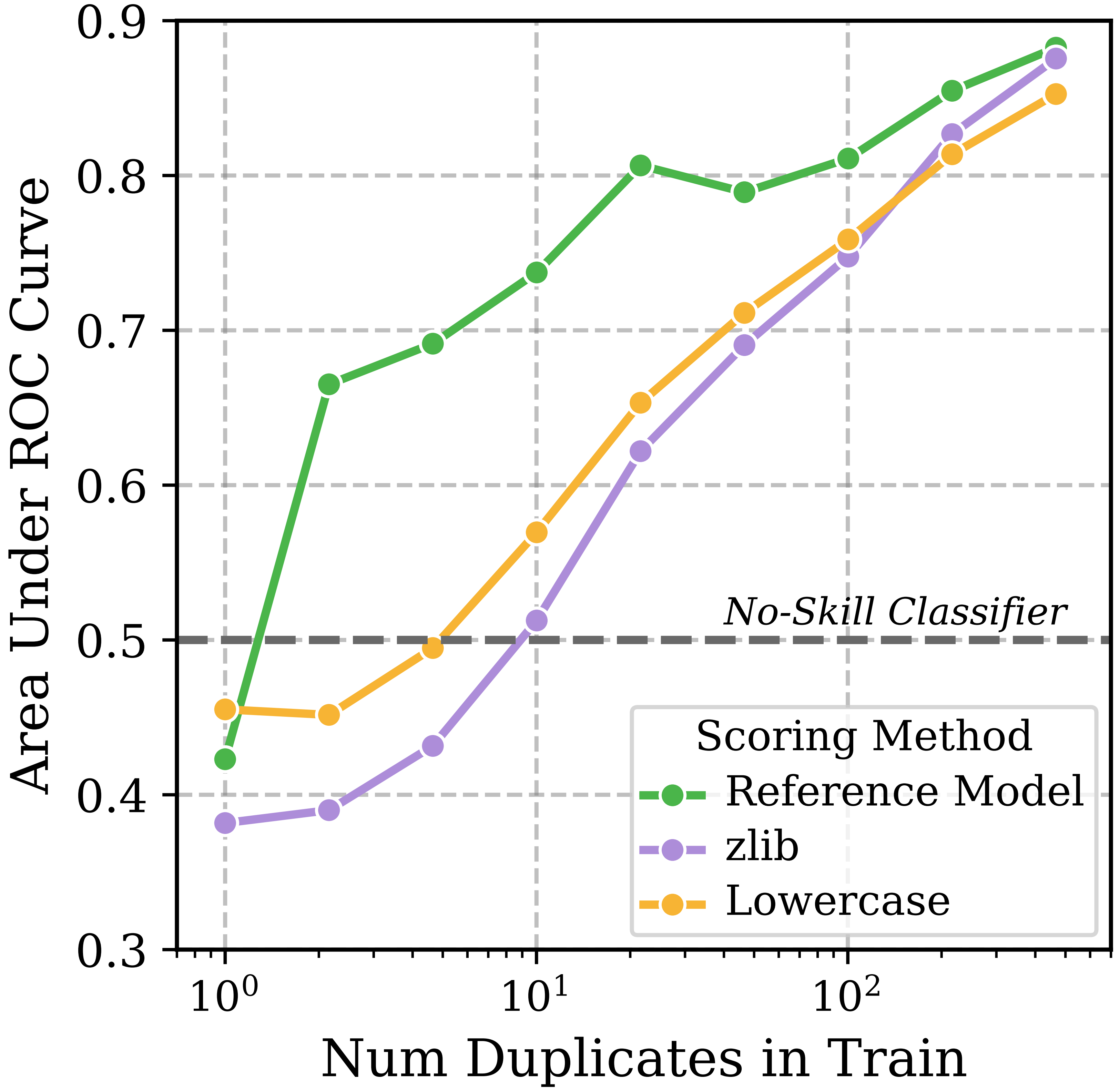
Membership Inference vs. Duplicates

Language Model Privacy Attack (Carlini et. al. 2021)

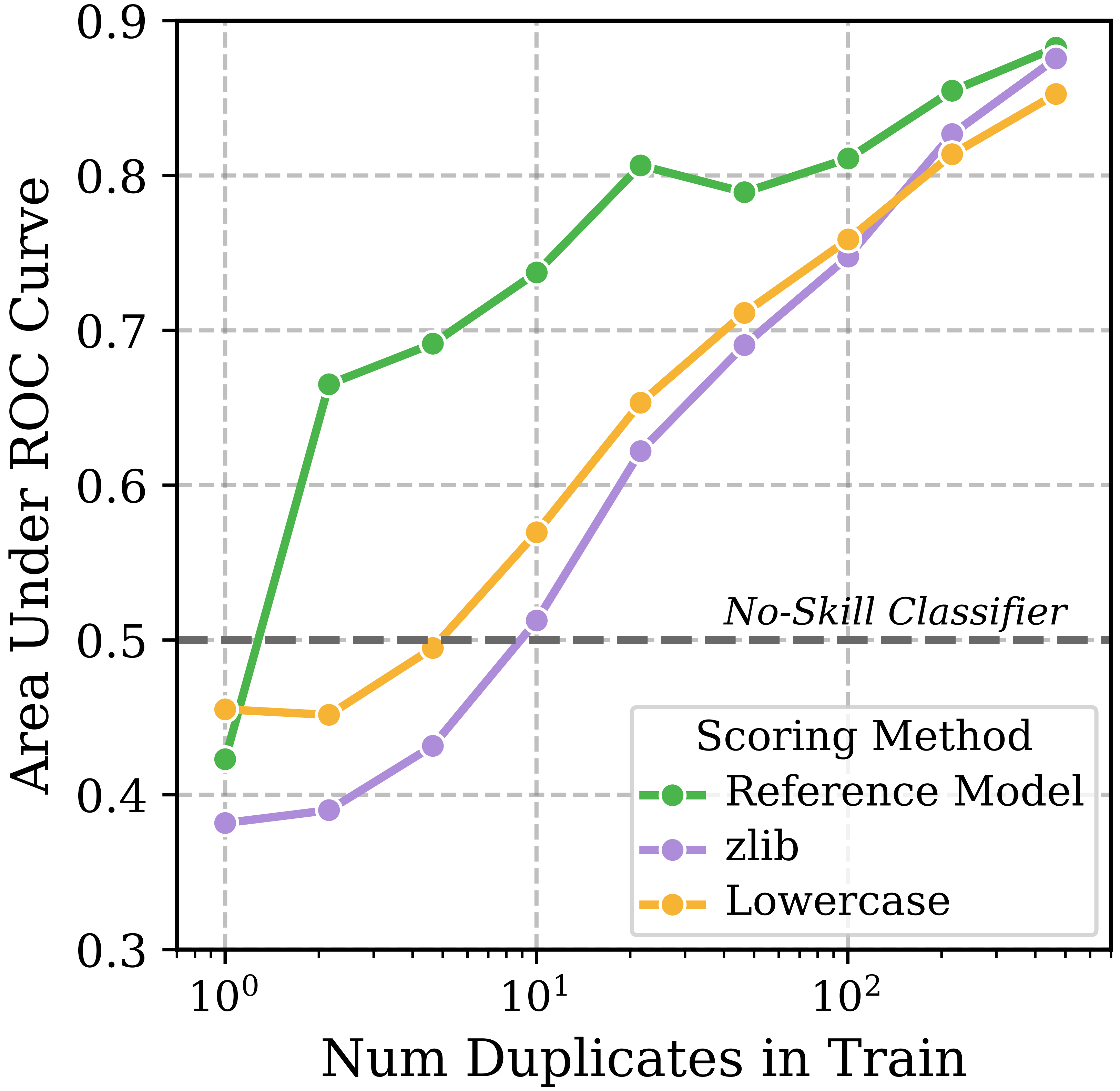


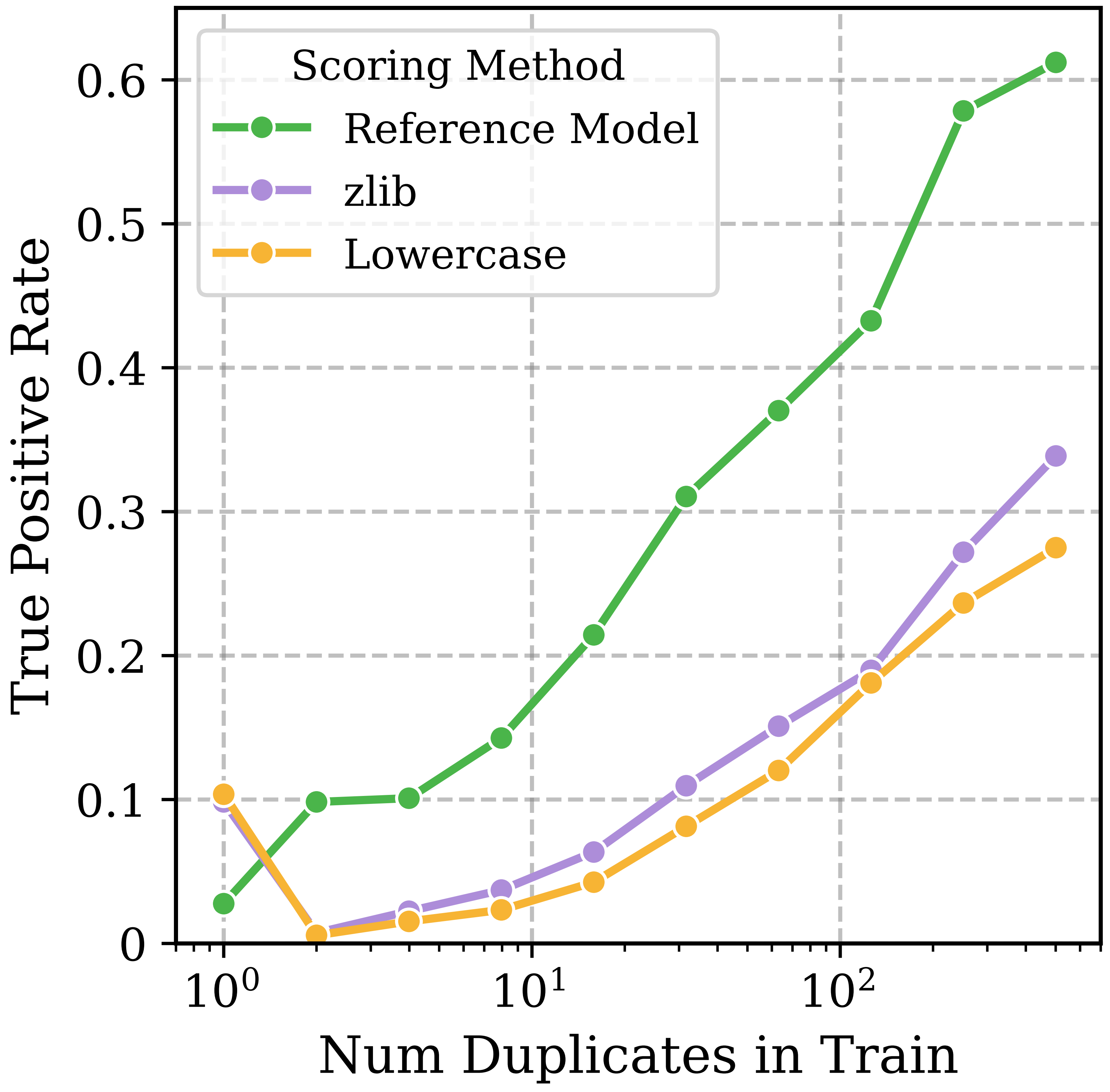
Membership Inference vs. Duplicates

Membership Inference vs. Duplicates

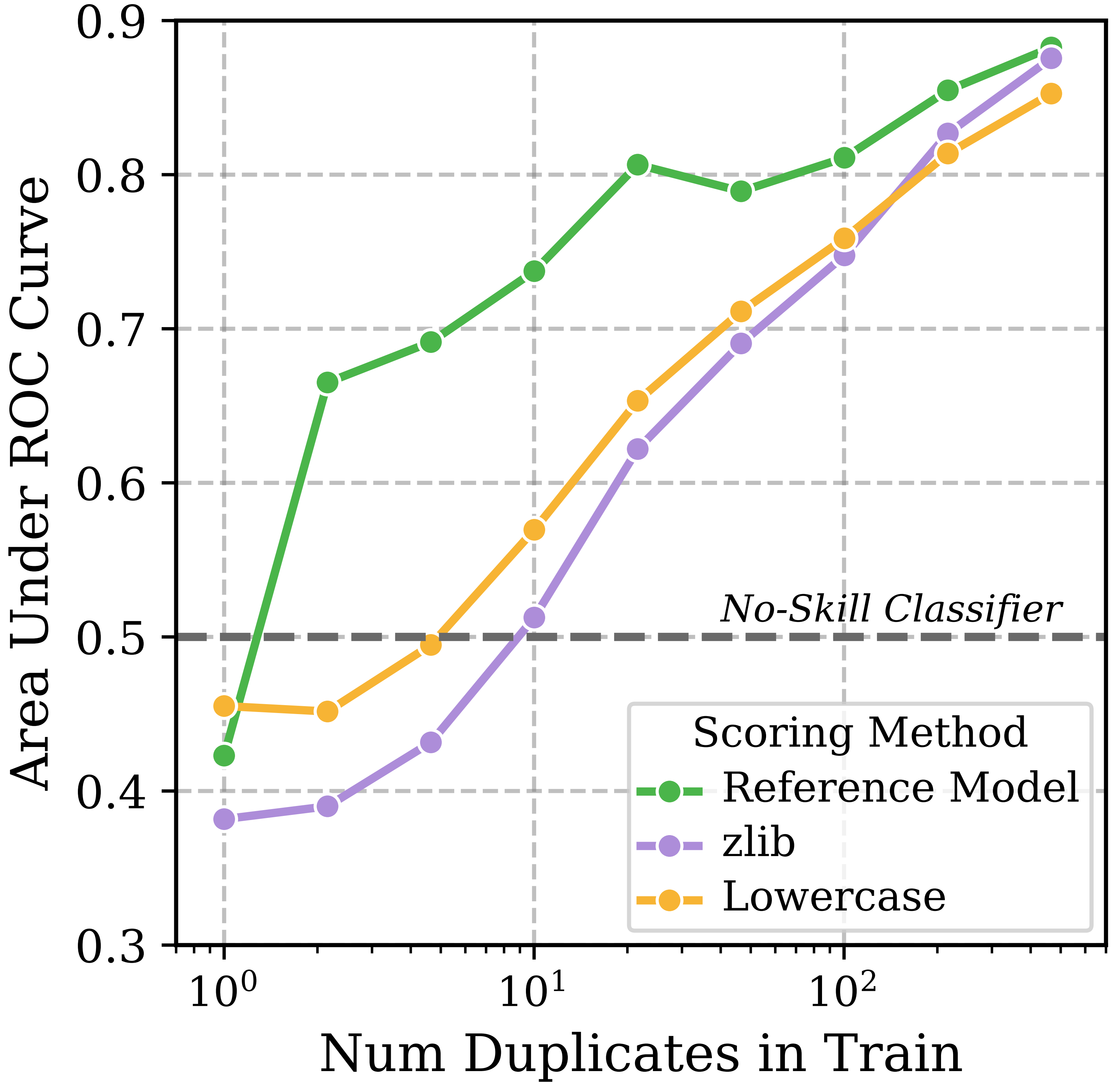


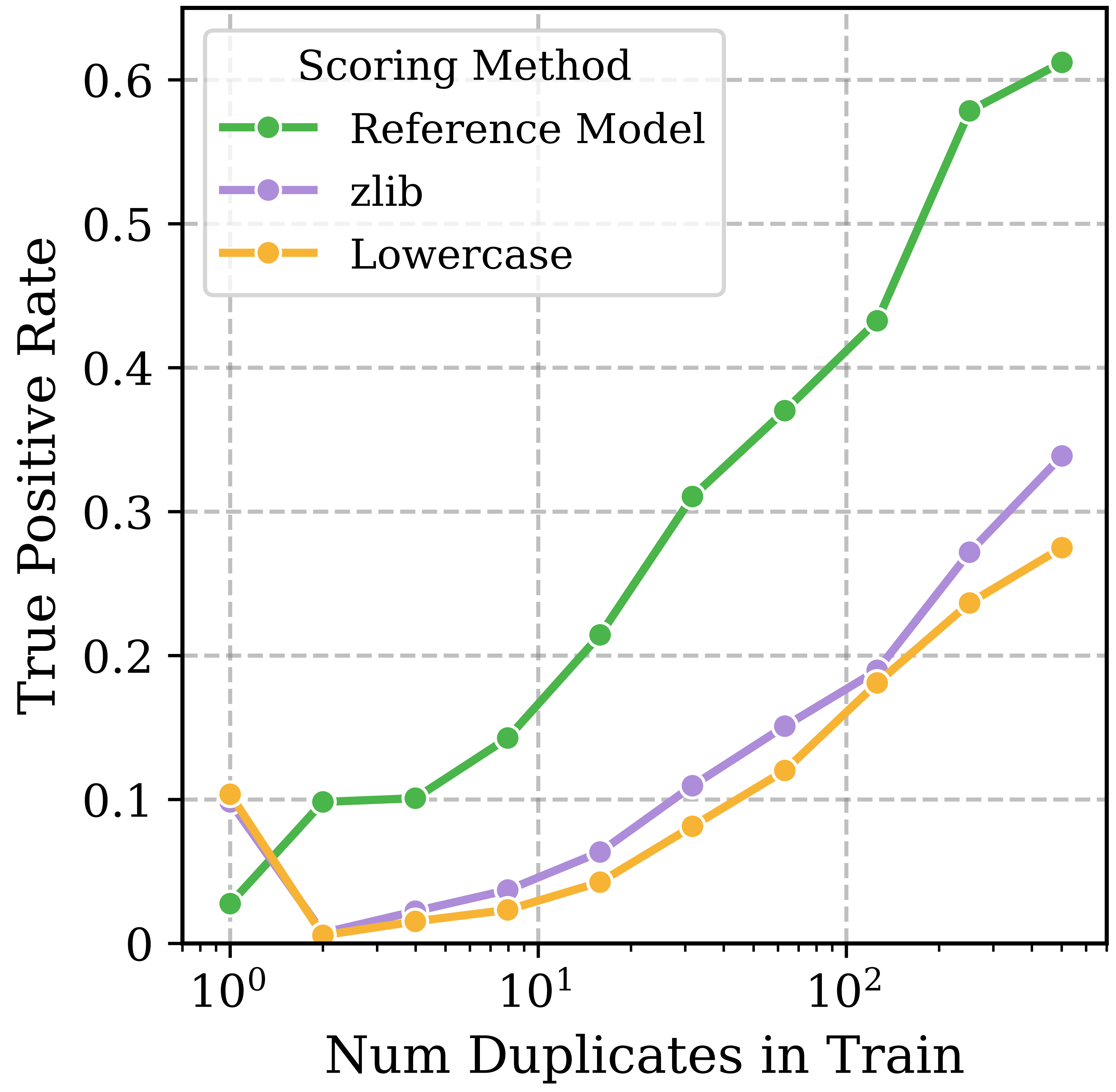
Membership Inference vs. Duplicates





Membership Inference vs. Duplicates



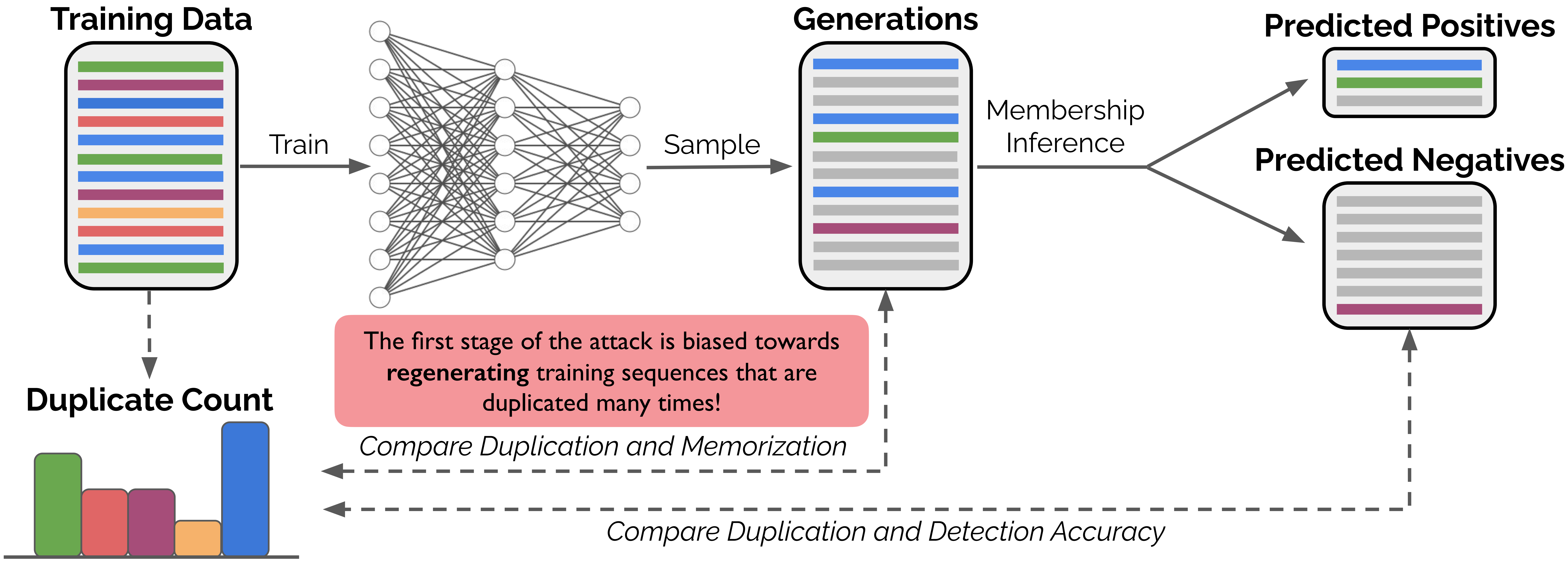


Observation #3

Membership Inference methods detect duplicated training sequences more effectively than unduplicated training sequences

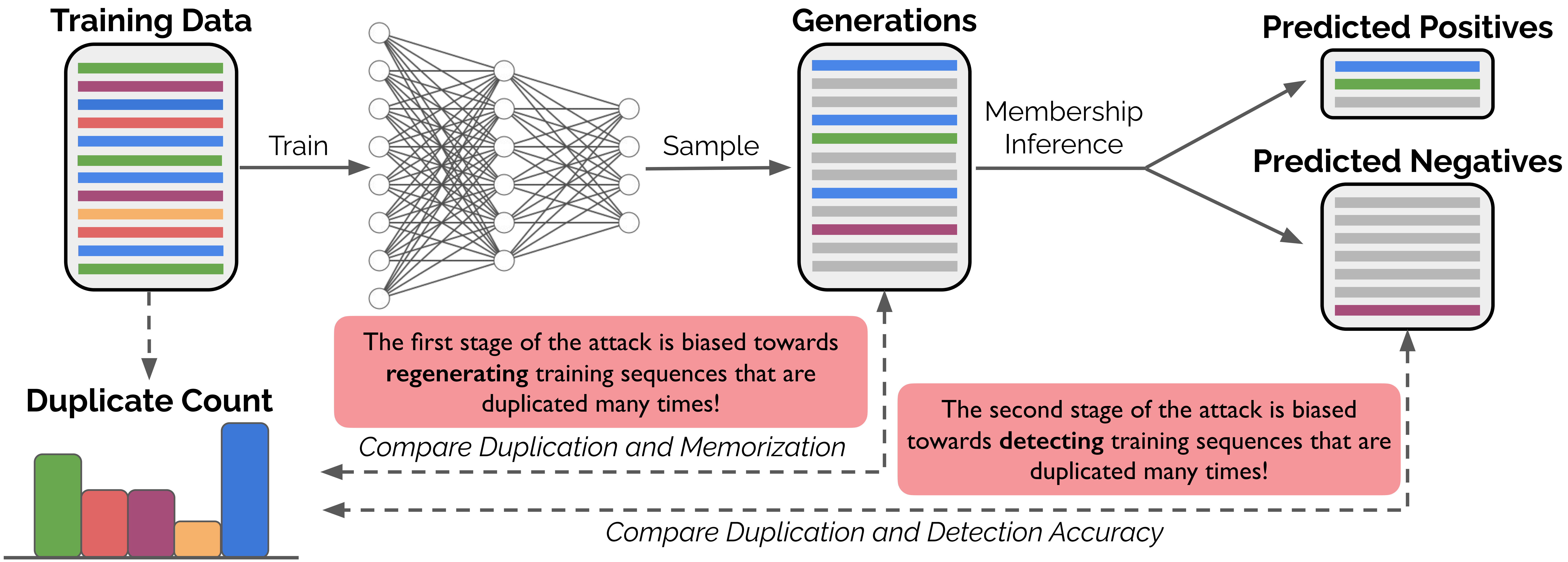
Membership Inference vs. Duplicates

Language Model Privacy Attack (Carlini et. al. 2021)



Membership Inference vs. Duplicates

Language Model Privacy Attack (Carlini et. al. 2021)



Deduplicating Training Data Mitigates Privacy Risk

The first stage of the attack is biased towards **regenerating** training sequences that are duplicated many times!

The second stage of the attack is biased towards **detecting** training sequences that are duplicated many times!

Does training data deduplication mitigate privacy risk?

Deduplicating Training Data Mitigates Privacy Risk

The first stage of the attack is biased towards **regenerating** training sequences that are duplicated many times!

The second stage of the attack is biased towards **detecting** training sequences that are duplicated many times!

Does training data deduplication mitigate privacy risk?

		Normal Model	Deduped Model
Training Data Generated	Count	1,427,212	68,090
	Percent	0.14	0.007
Mem. Inference AUROC	zlib	0.76	0.67
	Ref Model	0.88	0.87
	Lowercase	0.86	0.68

Takeaways and Future Questions

Takeaways and Future Questions

1. Privacy attack evaluation should take into account data duplication

Takeaways and Future Questions

1. Privacy attack evaluation should take into account data duplication

2. Do similar patterns exist for approximate duplicates?

Takeaways and Future Questions

1. Privacy attack evaluation should take into account data duplication

2. Do similar patterns exist for approximate duplicates?

3. *Why* are language models miscalibrated?

Deduplicating Training Data Mitigates Privacy Risks in Language Models

Presented at ICML 2022

Nikhil Kandpal

University of North Carolina, Chapel Hill

Eric Wallace

UC Berkeley

Colin Raffel

University of North Carolina, Chapel Hill