Duplication, Memorization, and Privacy in Language Modeling

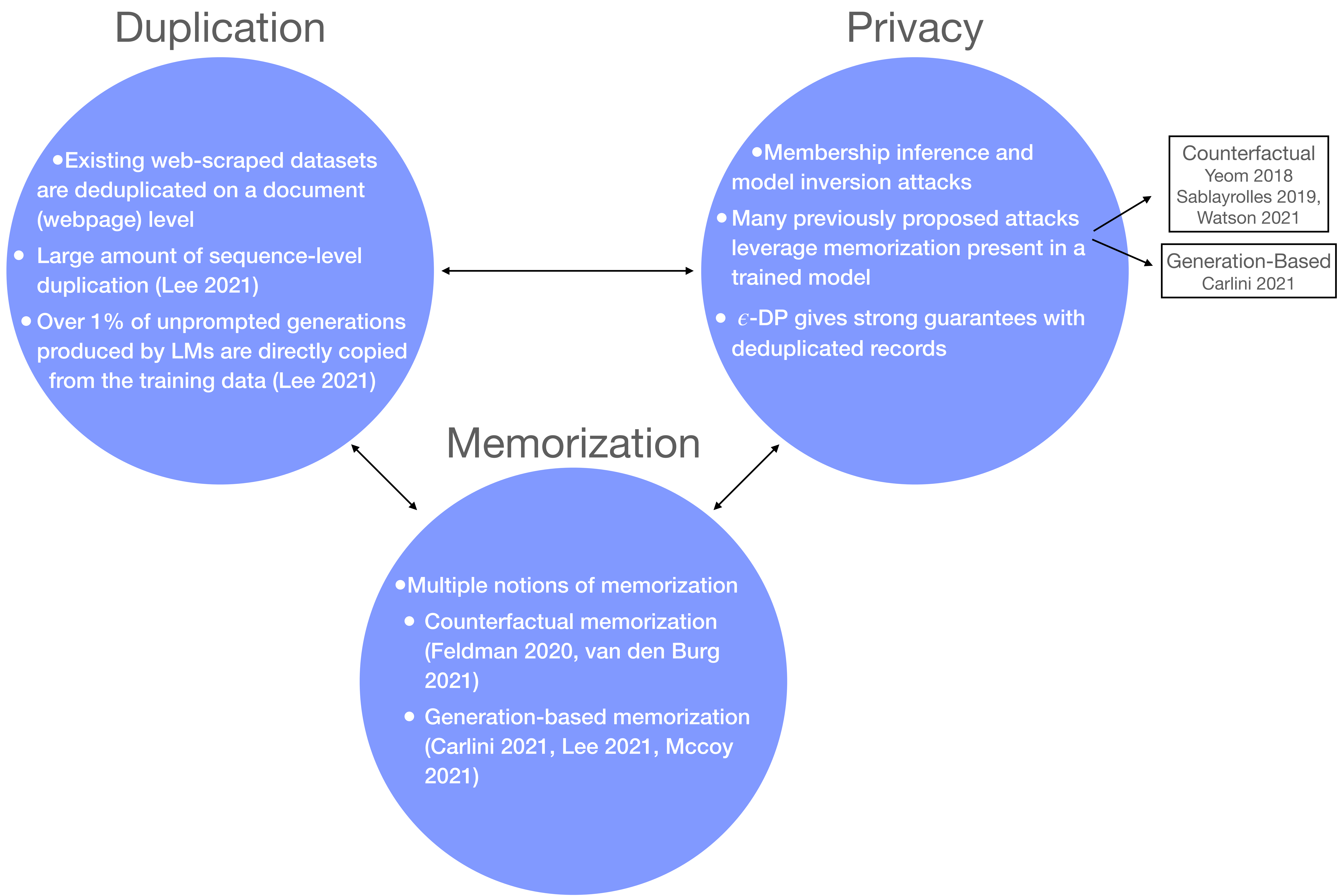
Nikhil Kandpal, Eric Wallace, Colin Raffel

Duplication, Memorization, and Privacy in Language Modeling

Language Modeling Basics

- What is language modeling?
  - Next-token prediction (autoregressive language modeling)
- What are language models trained on?
  - Large text corpora collected through web scraping
    - WebText, C4, The Pile
- Why are they useful?
  - Transfer learning
  - In-context learning
  - Many NLP tasks can be framed as language modeling

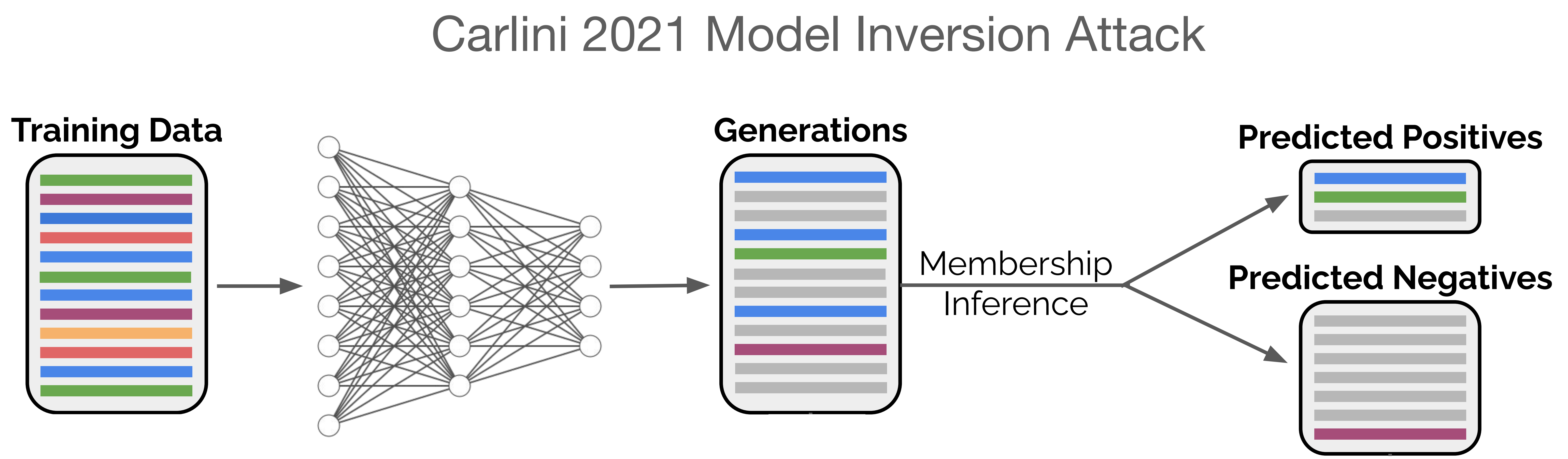
Duplication, Memorization, and Privacy in Language Modeling


Duplication
•Existing web-scraped datasets are deduplicated on a document (webpage) level
• Large amount of sequence-level duplication (Lee 2021)
• Over 1% of unprompted generations produced by LMs are directly copied from the training data (Lee 2021)
Privacy
•Membership inference and model inversion attacks
• Many previously proposed attacks leverage memorization present in a trained model
• ϵ-DP gives strong guarantees with deduplicated records
Counterfactual
Yeom 2018
Sablayrolles 2019,
Watson 2021
Generation-Based
Carlini 2021
Memorization
•Multiple notions of memorization
• Counterfactual memorization (Feldman 2020, van den Burg 2021)
• Generation-based memorization (Carlini 2021, Lee 2021, Mccoy 2021)

Our Contributions

1. Investigate the effects of sequence-level training data duplication on data privacy
    A. Study the Carlini 2021 model inversion attack through the lens of duplication
    B. Is model inversion easier to perform on duplicated sequences?
    C. Does removing sequence-level duplication mitigate model inversion risks?

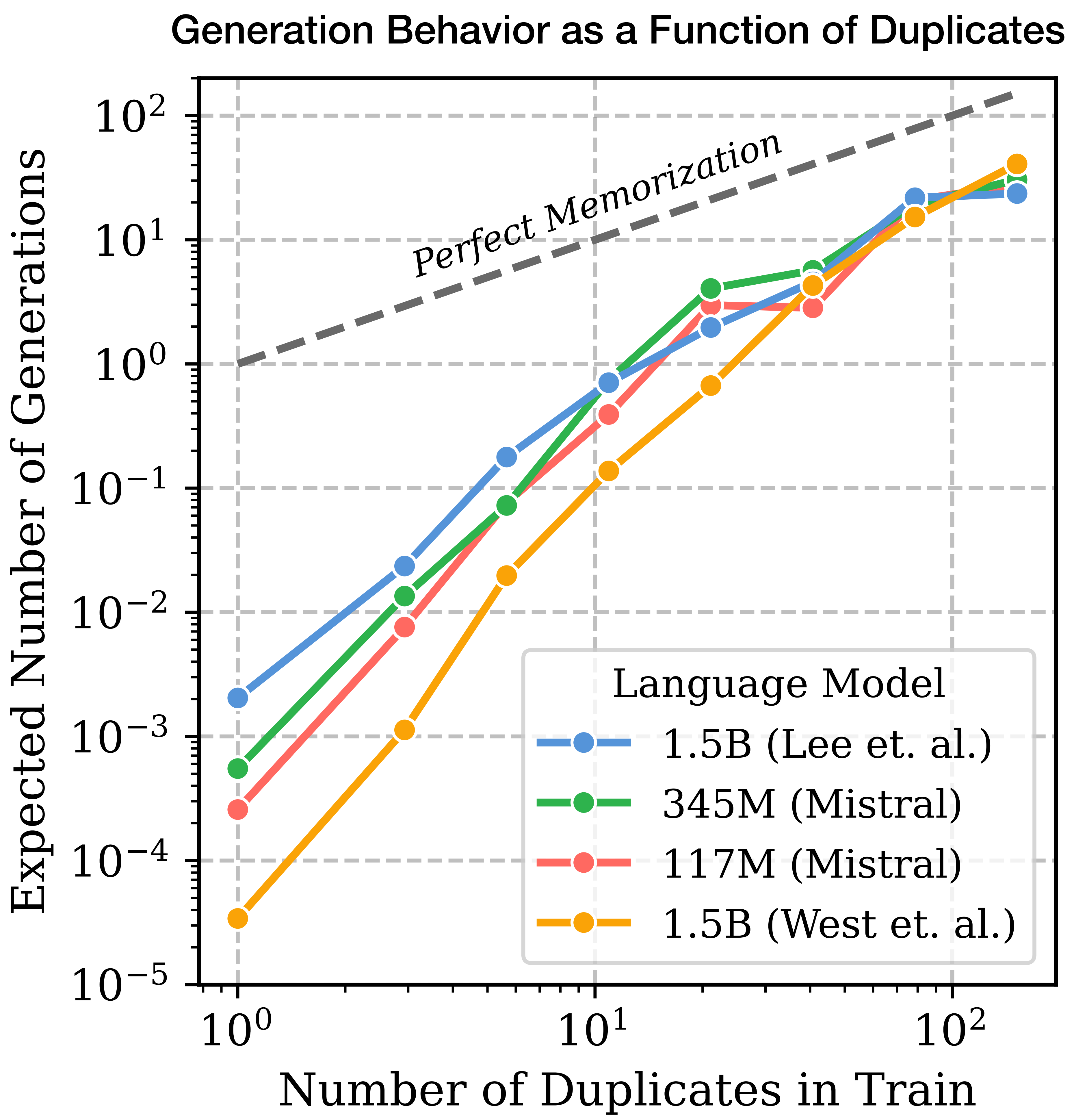
Experimental Setup



- Individually analyze how the effectiveness of each attack stage is impacted by duplication

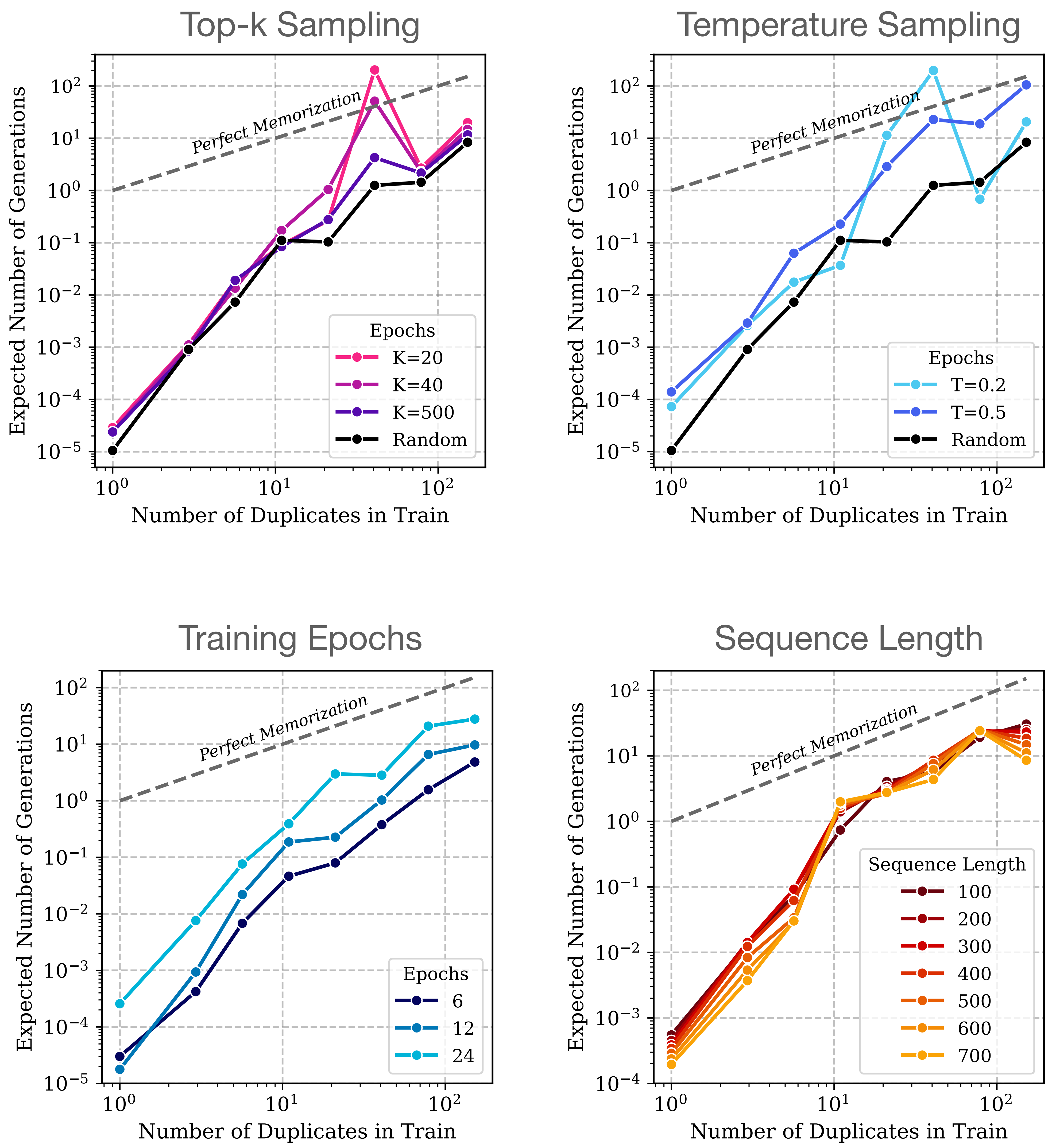
Generation Stage

- Generate samples from a variety of models
- Measure the average number of times a sequence duplicated d times in the training data is generated
- Scale results to simulate generating amount of text equal in size to training data



Memorization Across Varied Hyperparameters

- Model sizes:
  - Larger models emit more data
- Sample decoding strategy:
  - Reducing entropy of sampling emits more data
- Sequence length:
  - Little effect
- Training epochs:
  - Memorization increases over the course of training

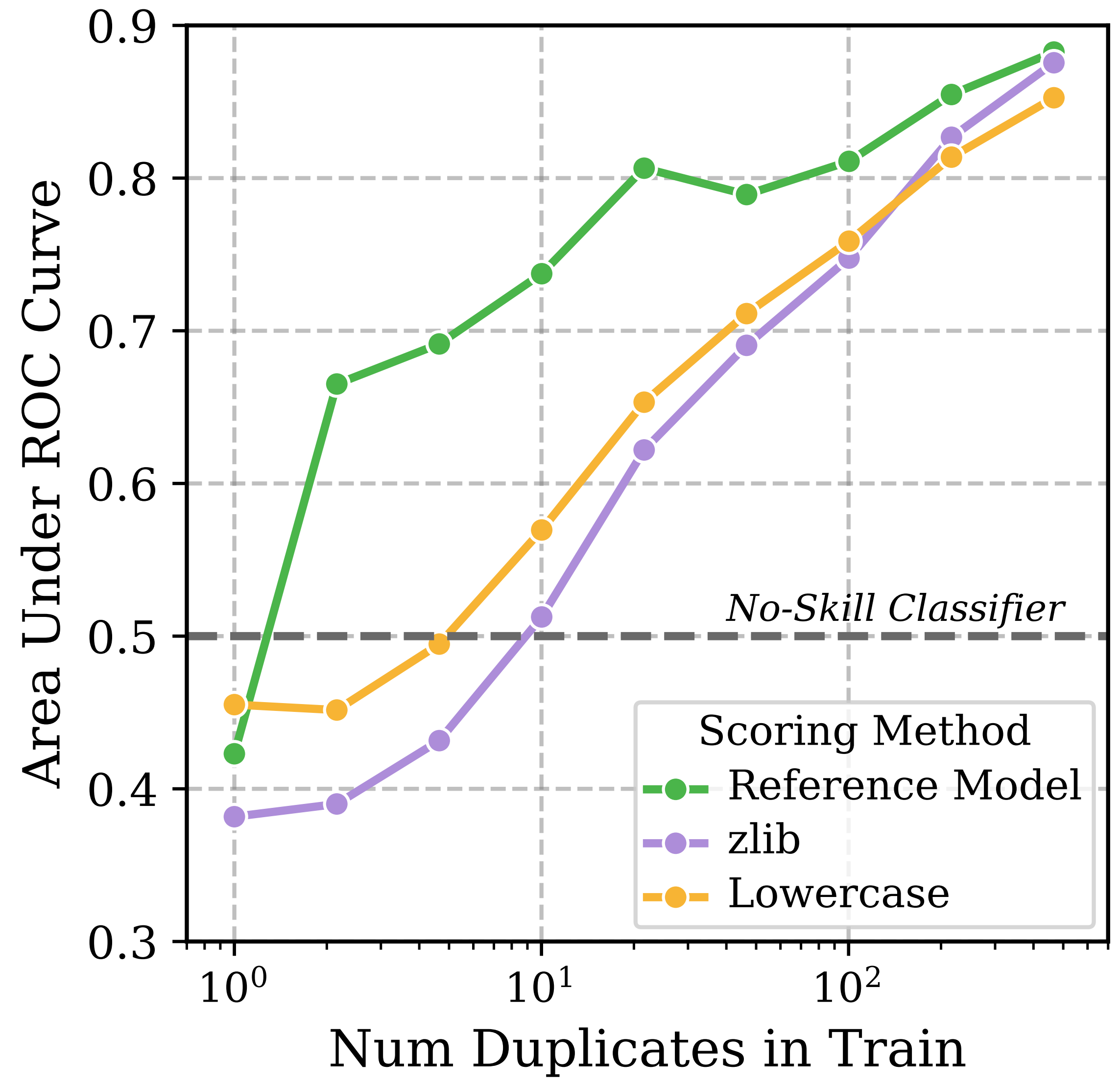


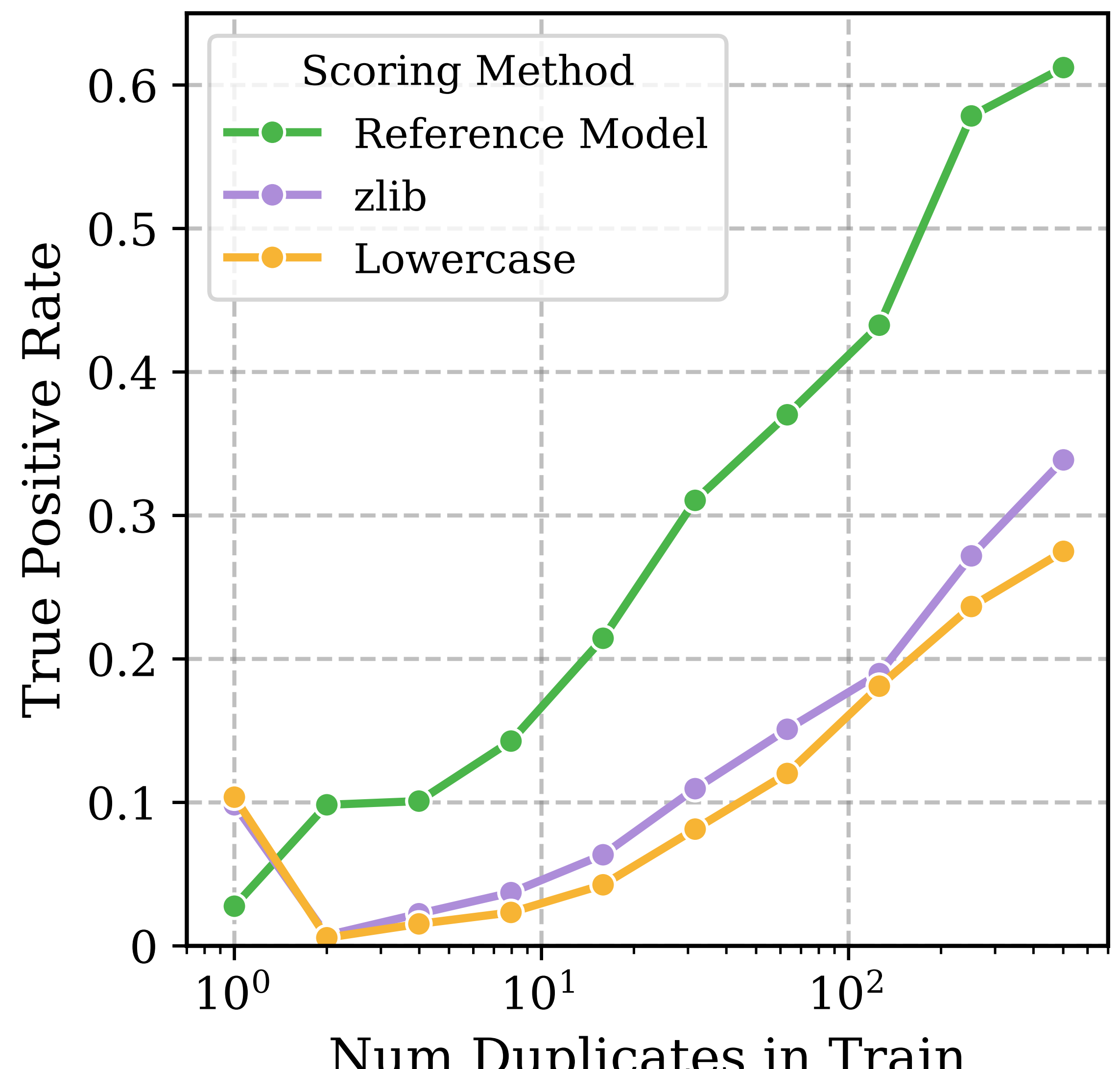
Membership Inference

- Carlini 2021 Membership Inference Methods:
  - Score samples with ratio of an “easiness” metric and the trained model’s perplexity
- Easiness Metrics:
  1. Reference Model - Perplexity of a different LM (trained on other dataset)
  2. zlib - Length of sequence after compression by zlib
  3. Lowercase - Perplexity of sequence with lowercase characters

Membership Inference and Duplication

- All three membership inference scores positively correlated with duplication
- At 1 duplicate, the AUROC is roughly at the level of a “No-Skill Classifier”





Model Inversion on Deduplicated Models

- How effect are these attacks on models trained with deduplicated data
- Compare two models trained on C4 and deduplicated C4
  - First stage (generation) emits 20x less training data
  - Second stage (membership inference) performs worse when using zlib and lowercase methods

		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

Table 1. Deduplicating training data drastically reduces the effectiveness of privacy attacks. We first generate 1 million 256-token samples from models trained on C4 and deduplicated C4. We then report the number of unique 400-character training sequences that are generated (*Count*) and the percentage of all 400-character training sequences that are generated (*Percent*). We then report the classification AUROC achieved by each of the three membership inference scores when applied to the generated sequences.

Hypothesis for Reference Model

- Membership inference with Reference Model method is virtually unchanged on normal and deduplicated models
- Two hypotheses:
  - The type of samples generated by normal and deduplicated models are different in some way that eases detection
  - Reference Model method approximates counterfactual memorization which is not necessarily correlated with generation-based memorization

Takeaways

- The success of the Carlini 2021 privacy attack is very reliant on sequence-level duplication
  - Superlinear relationship between generation rate and duplication
    - Open Question: Is this what would be expected in theory when particular training examples are oversampled?
  - Reduced membership inference effectiveness for some scoring methods