Deduplicating Training Data Mitigates Privacy Risks in Language Models

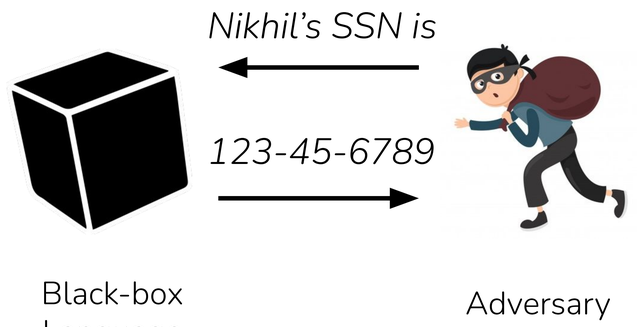
Nikhil Kandpal[1] Eric Wallace[2] Colin Raffel[1]
[1]UNC [2]UC Berkeley

TL;DR

1. LMs memorize and re-generate private training data. Attackers can detect this.
2. Exact string duplicates are common in large-scale NLP datasets.
3. LMs re-generate duplicated training data at higher rates than non-duplicated data.
4. Attackers can more easily detect when LMs re-generate duplicated data.
5. Deduplicating training data mitigates an attacker's effectiveness.

Privacy Attacks on Language Models

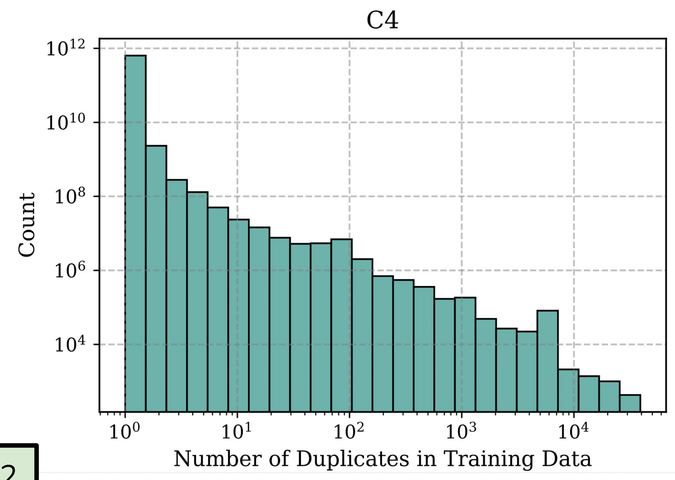
- LMs occasionally re-generate training data. These generations can be detected using membership inference attacks.



1

Duplicates in Training Sets

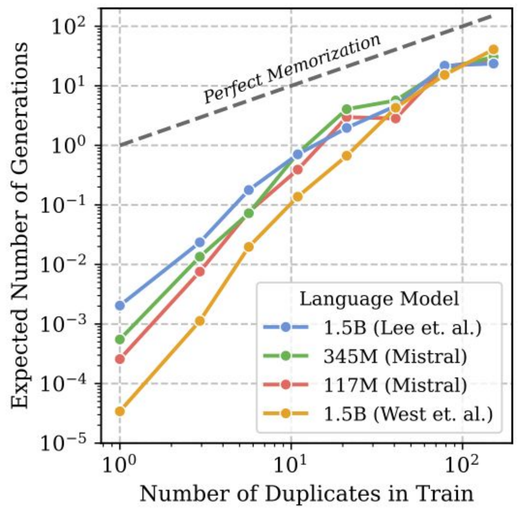
- Large-scale language modeling datasets, e.g., OpenWebText and C4, contain text duplicated 1000's of times.



2

How Do Duplicates Affect Generation?

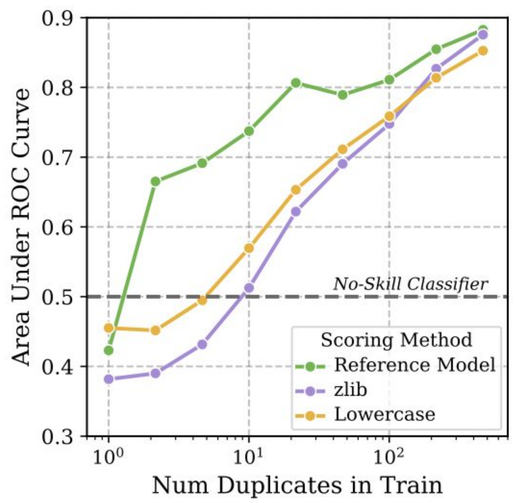
- A superlinear relationship exists between duplicate count and re-generation rate.



3

How Do Duplicates Affect Detection?

- Membership inference attacks have chance accuracy on non-duplicated data.



4

Deduplication Mitigates Privacy Risks

- Text re-generation rate and membership inference effectiveness decrease after deduplicating training data.

		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

5