Quantifying and Mitigating Memorization in Language Models

By Kyle Stein

Overview

Problem	2 Papers Addressing This Issue	Goal
Large language models have been shown to memorize parts of their training data, which can lead to privacy violations, degraded utility, and fairness issues.	1. "Quantifying Memorization Across Neural Language Models" by Carlini et al. 2. "Silo Language Models: Isolating Legal Risk In A Nonparametric Datastore" by Min et al.	Compare and contrast the approaches and findings of these two papers

Quantifying Memorization Across Neural Language Models

Nicholas Carlini Daphne Ippolito Matthew Jagielski
Katherine Lee Florian Tramèr Chiyuan Zhang1

Quantifying Memorization - Intro

- Carlini et al. paper aims to quantify the degree of memorization in language models and provide precise bounds on the amount of extractable data
- Expands on prior work by comprehensively quantifying memorization across model families and establishing clear scaling trends
- Provides order-of-magnitude more precise bounds compared to previous studies

Quantifying Memorization - Methodology

- Memorization defined as:
  - A string s is extractable with k tokens of context from a model f if there exists a length-k string p, such that [p || s] is in the training data for f, and f produces s when prompted with p using greedy decoding
- Two sampling methods:
  1. Uniformly random sample
  2. Sample normalized by duplication counts and sequence lengths (to measure worst-case memorization)
- Suffix array data structure used for efficient duplicate finding

Metrics:

- Fraction of extractable sequences
- Absolute difference in extractability

Quantifying Memorization - Experiments (GPT-Neo)

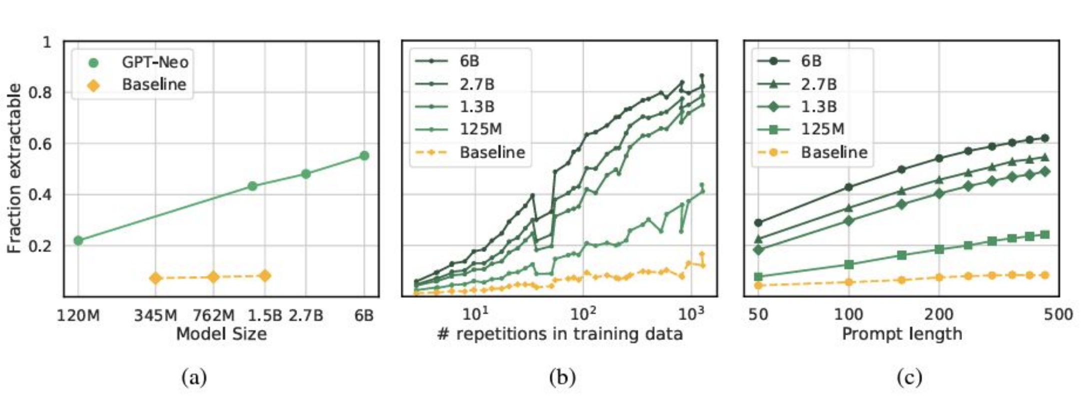
Experiments conducted on GPT-Neo models

<u>**Main Findings:**</u>

1 **Bigger Models Memorize More** — 10x increase in model size → 19 percentage point increase in memorization

2 **Data Duplication Matters** — Examples repeated more often are more likely to be extractable (log-linear trend)

3 **Context helps discover memorization** — More context tokens → better extraction of memorized text (log-linear trend)



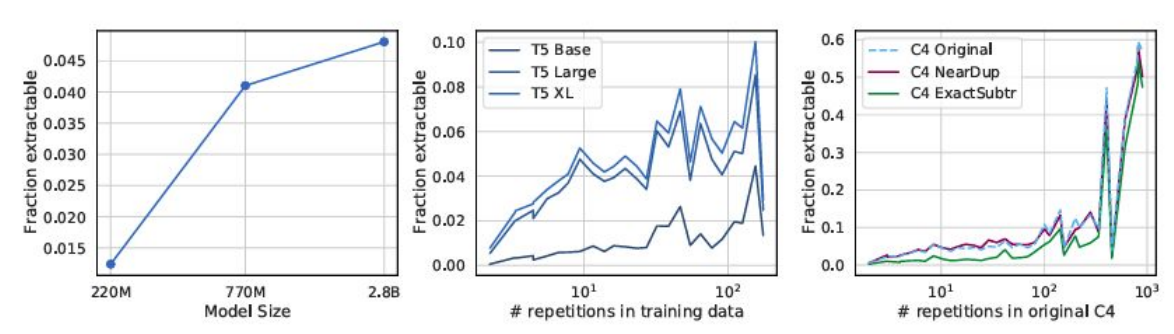
Quantifying Memorization - Replication (T5 and OPT)

Replication studies on T5 and OPT models

T5 models:
- Similar scaling of memorization with model size, but an order of magnitude less absolute memorization than GPT-Neo
- Data duplication effects less clear due to dataset idiosyncrasies
- 3B parameter T5-XL model memorizes 3.5% of sequences repeated 100 times

OPT models (trained on modified Pile):
- Orders of magnitude less memorization than GPT-Neo, possibly due to careful data curation/deduplication
- Largest OPT model memorizes a smaller fraction of The Pile than the smallest 125 million parameter GPT Neo model

Models trained on deduplicated data still memorize frequently repeated examples (>400 repetitions)



Quantifying Memorization - Conclusion

- Memorization scales log-linearly with model size, data duplication, and context length
- Larger models likely to memorize even more data, especially low-repetition examples
- Deduplication helps mitigate memorization but doesn't solve it completely
- Implications for privacy, utility, and fairness as models grow in size

Key contributions:

- Establishing scaling relationships
- Studying them across model families
- Providing precise estimates
- Showing the impact of deduplication

Silo Language Models: Isolating Legal Risk In A Nonparametric Datastore

Sewon Min
Hannaneh Hajishirzi
Suchin Gururangan
Noah A. Smith
Eric Wallace
Luke Zettlemoyer

Silo Language Models - Introduction

- LLMs trained on web-scale data can memorize and generate legally problematic text (e.g., copyrighted content, private information, hate speech)

- Previous mitigation approaches have limitations in scalability and effectiveness

- Min et al. propose "siloing" to isolate legal risks in a nonparametric datastore without significantly degrading performance

Silo Language Models - Methodology

Siloing refers to the separation of data processing to mitigate legal risks

Siloing approach:

1. Identify passages with legal risks using a classifier
2. Remove these passages from the model's nonparametric datastore
3. Retrain the model on the filtered datastore

Legal risk classifier: RoBERTa-based, trained on 10,000 Wikipedia passages (1,000 manually labeled for legal risks)

Applied to a GPT-3 style model with 540B parameters, using a filtered version of the Pile dataset

Classifier evaluated using precision, recall, and F1 score (0.85 F1 on held-out test set)

Domain	Sources	Specific License	# BPE Tokens (B)
Legal	PD Case Law, Pile of Law (PD subset)	Public Domain	27.1
	BY Pile of Law (CC BY-SA subset)	CC BY-SA	0.07
Code	SW Github (permissive)	MIT/BSD/Apache	58.9
Conversational	SW HackerNews, Ubuntu IRC	MIT/Apache	5.9
	BY Stack Overflow, Stack Exchange	CC BY-SA	21.3
Math	SW Deepmind Math, AMPS	Apache	3.5
Science	PD ArXiv abstracts, S2ORC (PD subset)	Public Domain	1.2
	BY S2ORC (CC BY-SA subset)	CC BY-SA	70.3
Books	PD Gutenberg	Public Domain	2.9
News	PD Public domain news	Public Domain	0.2
	BY Wikinews	CC BY-SA	0.01
Encyclopedic	BY Wikipedia	CC BY-SA	37.0

PD: Public Domain data, no restrictions on use

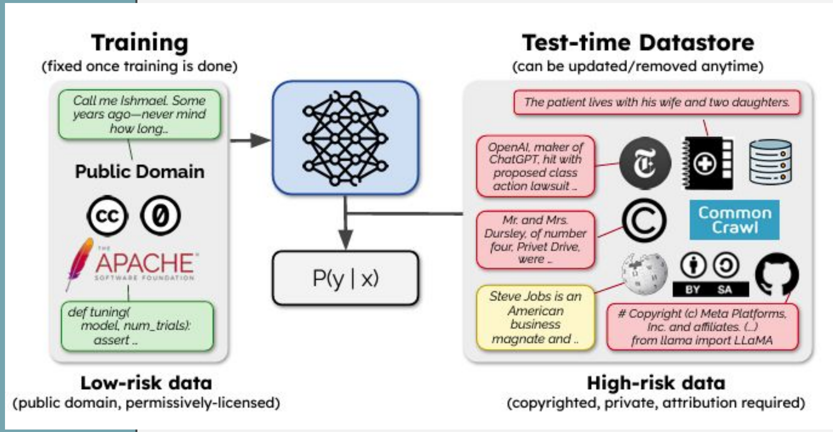
BY: Attribution licenses (e.g., Creative Commons Attribution), free to use with credit to the creator

SW: Permissively licensed software (e.g., MIT, Apache, BSD), free to use with basic stipulations

Silo Language Models - Experiments

<u>Evaluation on three benchmarks:</u>

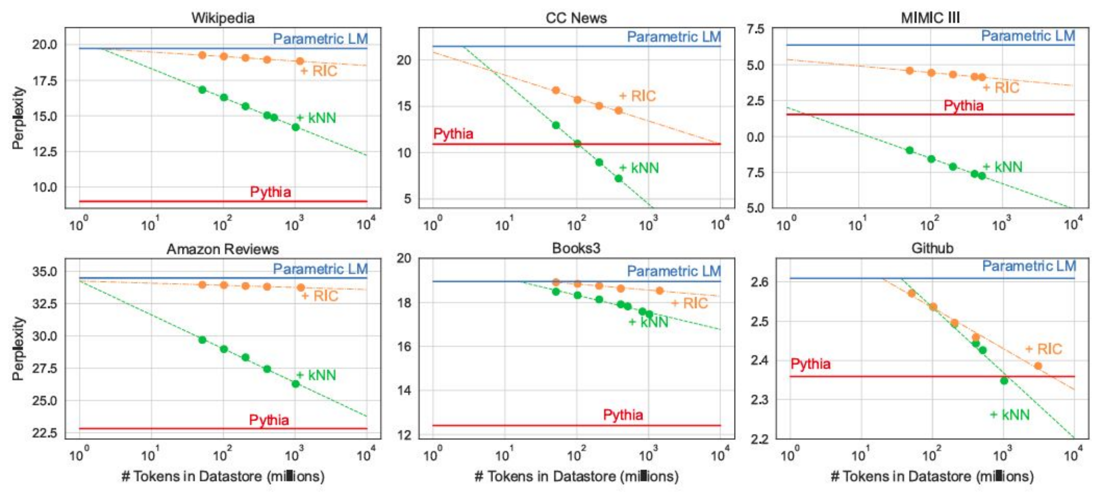
1. **TruthfulQA** — The siloed model's accuracy being within 2% of the baseline model's accuracy, while generating 20% fewer false statements
2. **WebGPT** — Siloed model maintains high performance (ROUGE-L 0.28 vs. 0.29 for baseline) while reducing legally risky content
3. **Custom legal risk dataset** — The siloed model having a 20% generation rate of legally risky content, compared to 40% for the baseline model.



	PD		PDSW		PDSWBY		*The Pile*	
Domain	**Tokens (B)**	**%**	**Tokens (B)**	**%**	**Tokens (B)**	**%**	**Tokens (B)**	**%**
Code	0.0	0.0	58.9	59.1	58.9	25.8	32.6	9.8
Legal	27.1	86.2	27.1	27.2	27.2	11.9	30.8	9.3
Conversation	0.0	0.0	5.9	5.9	27.2	11.9	33.1	10.0
Math	0.0	0.0	3.5	3.5	3.5	1.50	7.1	2.1
Books	2.9	9.3	2.9	2.9	2.9	1.3	47.1	14.2
Science	1.2	3.8	1.2	1.2	71.5	31.3	86.0	26.0
News	0.2	0.7	0.2	0.2	0.2	0.1	-†	-†
Wikipedia	0.0	0.0	0.0	0.0	37.0	16.2	12.1	3.7
Unverified web	0.0	0.0	0.0	0.0	0.0	0.0	83.1	25.0
Total	31.4	100.0	99.6	100.0	228.3	100.0	331.9	100.0

Silo Language Models - Analysis

- Analysis of legal risk types and impact on factual knowledge
- Siloed model shows a **<u>50% reduction</u>** in generating legally risky content on the custom dataset
- Similar perplexity scores to baseline on held-out test set, indicating minimal performance degradation



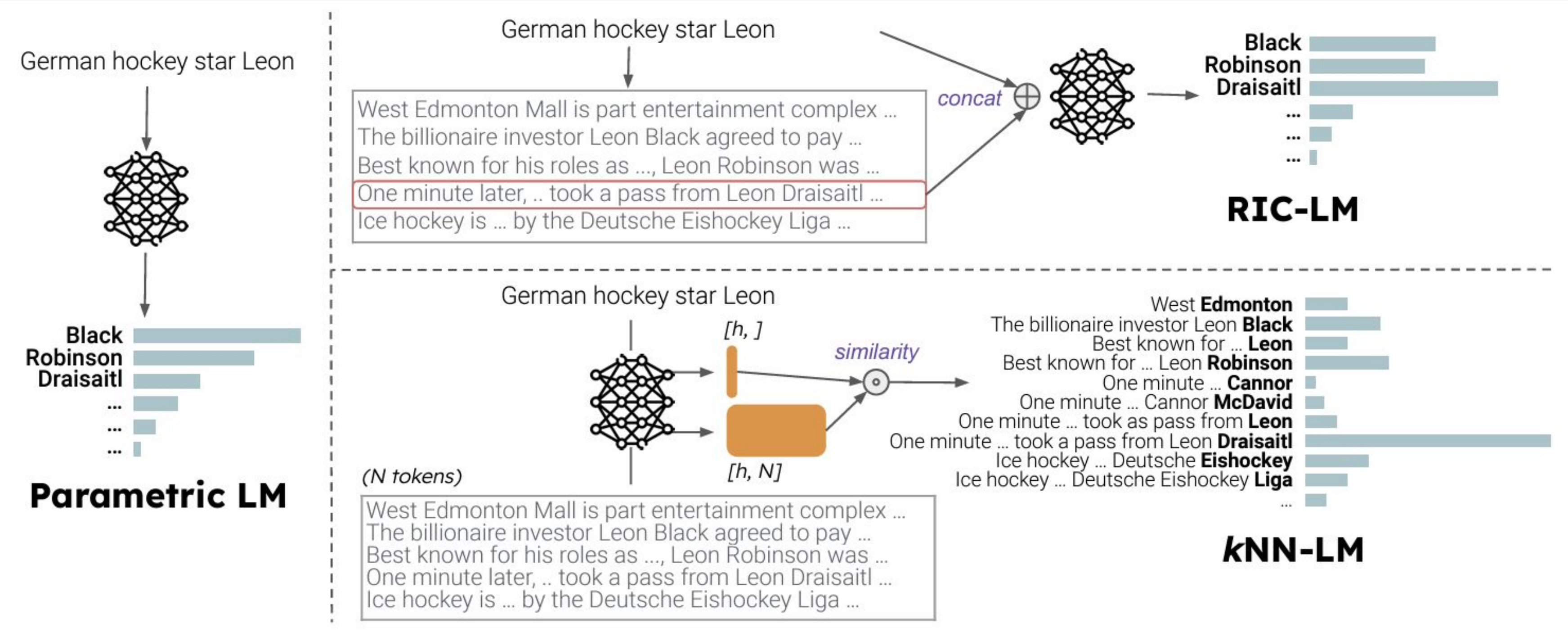


Figure 2 from the SILO paper illustrates the baseline methods compared in their experiments. The parametric LM, RIC-LM, and kNN-LM approaches are depicted, showcasing their differences in utilizing the datastore during inference.

Silo Language Models - Conclusion

- Siloing approach effectively reduces legally problematic content generation without significant performance degradation
- Modular approach, applicable to existing LLMs with minimal modifications
- **Limitations:**
  - Doesn't completely eliminate all legal risks; potential trade-off with diversity and informativeness
- Highlights the importance of addressing legal and ethical risks in LLMs

Key contributions:

- First large-scale demonstration of legal risk isolation using nonparametric datastore
- Introducing a new benchmark dataset
- Demonstrating effectiveness on multiple tasks

Comparison and Discussion

Both Papers

address the issue of memorization in language models

Focus

Carlini et al. focus on quantification, while Min et al. focus on mitigation

Complementary approaches

Understanding the extent of memorization and developing techniques to reduce its impact

Importance

of considering memorization and legal risks when developing and deploying language models

Conclusion

- Memorization in language models is a significant issue with implications for privacy, utility, and fairness
- Quantifying memorization helps understand its extent and scaling properties
- Mitigating legal risks through approaches like siloing can help reduce problematic content generation
- Further research is needed to fully understand and address the implications of memorization
- Practitioners should be aware of these issues and consider them when working with language models

References

- **Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022).** *Quantifying Memorization Across Neural Language Models.* ArXiv. /abs/2202.07646

- **Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., & Zettlemoyer, L. (2023).** *SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore.* ArXiv. /abs/2308.04430

Thank You!