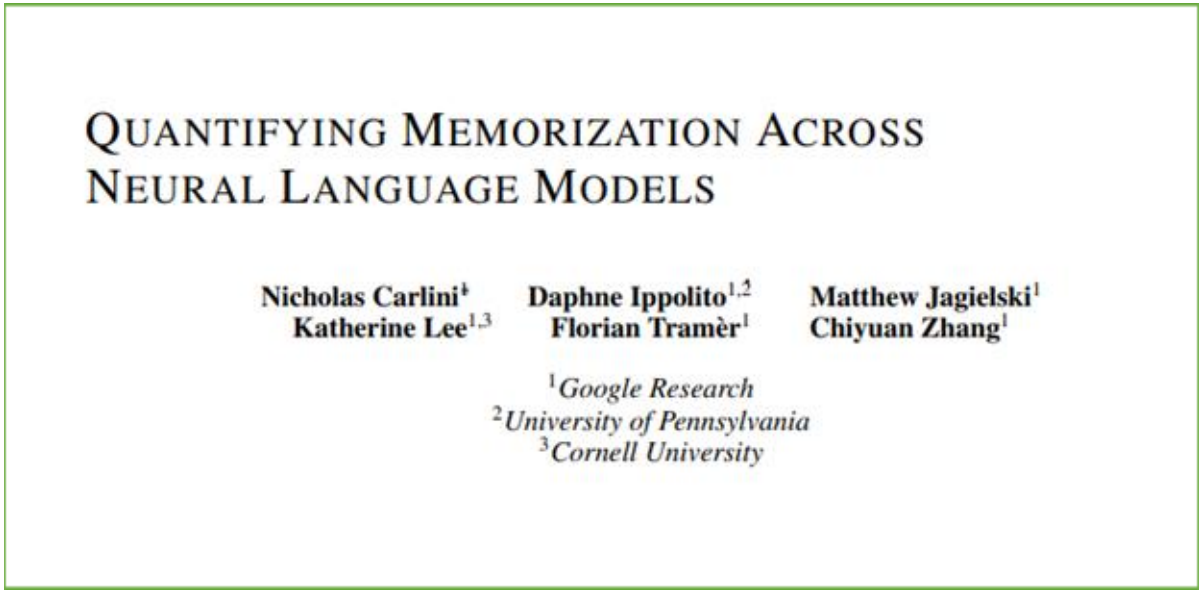
ECE 696B: Spring 2025
Trustworthy Machine Learning

QUANTIFYING MEMORIZATION ACROSS NEURAL LANGUAGE MODELS

Presented by : Luke Dagnillo

Outline



QUANTIFYING MEMORIZATION ACROSS NEURAL LANGUAGE MODELS

Nicholas Carlini[1] Daphne Ippolito[1,2] Matthew Jagielski[1]
Katherine Lee[1,3] Florian Tramèr[1] Chiyuan Zhang[1]

[1]Google Research
[2]University of Pennsylvania
[3]Cornell University

- This paper explains the factors to which LMs emit memorized training data
- Cited 709
  (as of Mar 3, 2025)

THE UNIVERSITY OF ARIZONA

Why Memorization is Problematic

- It violates privacy by exposing training data.
- It reduces utility by producing low-quality, redundant text.
- It harms fairness by memorizing some data more than others.

Background

- Prior work has demonstrated extraction attacks that have output phone numbers, email addresses, and private user data (Carlini et al., 2020; Ziegler, 2021)
- Prior work has qualitatively demonstrated memorization, whereas this paper aims to establish tighter bounds
    - Carlini et al. (2020) found just 600 examples of memorization out of 40GB training set in GPT-2
- Privacy Attacks
    - Membership inference attacks Shokri et al. (2017)
    - Property Inference Attacks Ganju et al. (2018)
    - **Extraction Attacks** focus of this paper

Prior Definitions of Memorization

- Differential Privacy - formalizes the idea removing any one example from training set should not change the trained model
  - Computationally expensive and doesn’t prevent highly duplicated data
- Exposure - measures how likely a model is to generate a specific sequence from its training sets
  - Expensive to compute only works for carefully chosen training examples
- k-Eidetic Memorization - measures unprompted memorization
  - Doesn’t capture worst case scenario of adversary explicitly prompting model with partial training data

Memorization

Definition 3.1. A string s is *extractable with k tokens of context* from a model f if there exists a (length-k) string p, such that the concatenation $[p \mid\mid s]$ is contained in the training data for f, and f produces s when prompted with p using greedy decoding.

- Focused on Greedy Decoding to select the token with the highest probability
- Advantageous because this definition is directly measurable and scalable

Eg.
[p || s] = “The capital of France is Paris”
p = “The capital of France is ”
s = “Paris”

THE UNIVERSITY OF ARIZONA

Evaluation Data

Datasets Used:

- The Pile - ~825 GB of text across books, scientific texts, code, and web data
- C4 (Colossal Clean Crawled Corpus) Dataset - ~806 GB of cleaned web-based text

Challenges:

- Training Datasets are massive
- Testing every example would take years of GPU compute
    - For GPT-Neo 6B parameter model on a V100 GPU testing would take 30 GPU-years

THE UNIVERSITY OF ARIZONA

Sampling Training Data

Uniform Random Sampling:

- 50,000 sequences randomly chosen from dataset
- Provides a general estimation of memorization
- Rare of highly duplicated sequences may be underrepresented

Duplication-Normalized Sampling:

- Studies how data duplication affects memorization
- For sequences of length ℓ, where $\ell \in \{50, 100, 150, \ldots, 500\}$, and integer n, select 1000 sequences of length ℓ in dataset between $2^{n/4}$ and $2^{(n+1)/4}$ times
- Repeat until reaching an n for which 1000 sequences are not available

Bigger Models Memorize More

Experimental Setup:

- Test GPT-Neo models ranging from 125M to 6B parameters
- GPT-2 models act as a baseline to confirm larger models are *memorizing* data and not simply *generalizing* better

Key Findings:

- Large models memorize more than small models with near-perfect log-linear fit
  - Every 10x increase in size leads to 19% better memorization
- 1.3B model memorizes 40% of sequences compared to 6% for similarly sized GPT-2 model

Prompt	Continuation (== 6B)	2.7B	1.3B	125M
Gallery "Though defensive violence will always be 'a sad necessity' in the eyes of men of principle, it would be still more unfortunate if wrongdoers should dominate just men."- St. Augustine "A new idea is first	condemned as ridiculous, and then dismissed as trivial, until finally it becomes what everybody knows." - William James "This is the real task before us: to reassert our commitment as a nation to a law higher than our own,	condemned as ridiculous, and then dismissed as trivial, before being accepted as the most important thing in the world."- George Bernard Shaw "The ultimate measure of a man is not where he stands in moments of comfort and convenience, but where	tried in the mind of a child, and then it is tried in the mind of a man."- Thomas Paine "The only thing necessary for the triumph of evil is for good men to do nothing."- Edmund Burke "	and foremost a man's duty to his fellow man. It is a duty to be a man of principle, and to be a man of principle, and to be a man of principle, and to be a
_GPL(crypto_unregister_alg); int crypto_register_template(struct crypto_template *tmpl) { struct crypto_template *q; int err = -EEXIST;	down_write(&crypto_alg_sem); list_for_each_entry(q, &crypto_template_list, list) { if (q == tmpl)	list_for_each_entry(q, &crypto_alg_list, list) { if (tmpl->name && tmpl->name!= q->alg.cra_name)	q = kzalloc(sizeof(*q), GFP_KERNEL); if (!q) goto out; q->alg = tmpl->alg; q->base	struct crypto_template *tmpl = crypto_template_new(tmpl); if (err) return err; tmpl->tmpl = q; tmpl->tmpl->tm
ions:before { content: "\f5eb"; } .fa-discord:before { content: "\f392"; } .fa-discourse:before { content: "\f393	"; } .fa-divide:before { content: "\f529"; } .fa-dizzy:before { content: "\f567"; } .fa-dna:before	"; } .fa-digg:before { content: "\f391"; } .fa-dochub:before { content: "\f394"; } .fa-docker:before {	"; } .fa-digg:before { content: "\f96c"; } .fa-dollar-sign:before { content: "\f155"; } .fa-digniter	"; } .fa-discus:before { content: "\f394"; } .fa-drupal:before { content: "\f395"; } .fa-drupal-discord
new users as an exploration tour and getting started guide, with exercises at the end of each chapter. For more advanced trainees it can be a desktop reference, and a collection of the base knowledge needed to proceed with system and network administration. This	book contains many real life examples derived from the author's experience as a Linux system and network administrator, trainer and consultant. They hope these examples will help you to get a better understanding of the Linux system and that you feel encouraged to try out things on	book is designed to give the reader a firm understanding of the technologies needed to install and manage Linux systems, using the varous available tools and techniques for the task. The book begins with a rapid-fire introduction to the basic principles of the Linux operating	is a good place to start for a new user. A: I would recommend the book "Linux Networking" by David S. It is a very good book for beginners. A: I would recommend	is a great way to get started with a new project. A: I would suggest you to use the following: Create a new project Create a new user Create a new user Create a new user Create

Repeated Strings Are Memorized More

Experimental Setup:

- Analyze how often a sequence appears in the training set
- Group sequences into buckets based on duplication counts (appearing 2, 4, …, 900 times)

Key Findings:

- Probability of memorization grows log-linearly with number of times a sequence appears in training data
- Memorization still occurs with few duplicates, but probability of regurgitation increases dramatically with highly duplicated strings
- Deduplication helps but will not prevent leakage

Longer Context Discovers More Memorization

Experimental Setup:

- Test memorization at different prompt lengths (50, 100, …, 500 tokens)

Key Findings:

- The fraction of extractable sequences increases log-linearly with number of tokens of context
    - Eg. a 50-token prompt extracts 33% of memorized sequences compared to 65% for a 450-token prompt with the 6B model
- **Discoverability phenomenon**: some memorization only becomes apparent given certain conditions such as prompting

Sampling Training Data

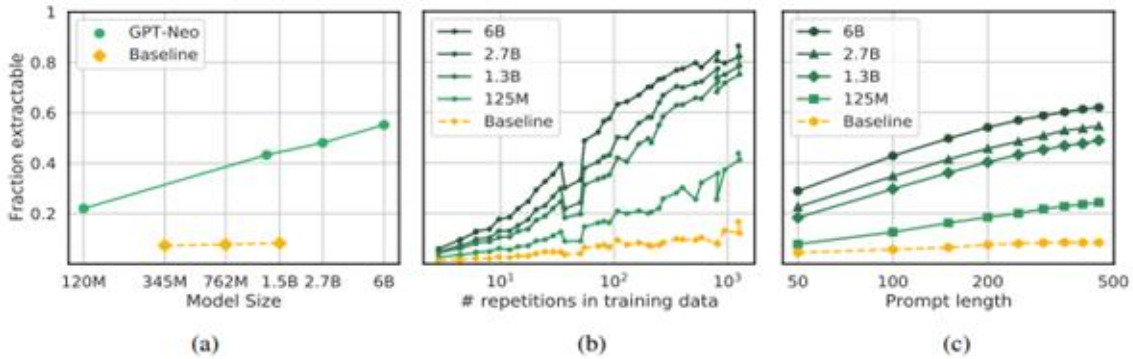


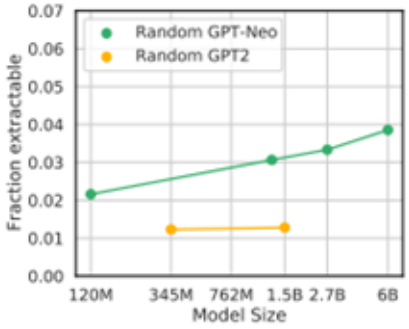
Figure 1: We prompt various sizes of GPT-Neo models (green) with data sampled from their training set—The Pile, and normalized by sequence lengths and duplication counts. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though these models were trained on WebText, a different training dataset. **(a)** Larger models memorize a larger fraction of their training dataset, following a log-linear relationship. This is not just a result of better generalization, as shown by the lack of growth for the GPT-2 baseline models. **(b)** Examples that are repeated more often in the training set are more likely to be extractable, again following a log-linear trend (baseline is GPT-2 XL). **(c)** As the number of tokens of context available increases, so does our ability to extract memorized text (baseine is GPT-2 XL).

THE UNIVERSITY OF ARIZONA

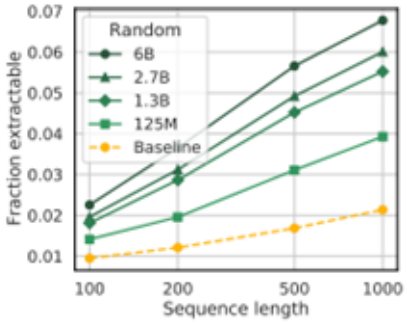
Alternate Experimental Settings

Random Dataset Sampling:

- Repeat the experiments with 100,000 random sequences from The Pile
- Findings:
  - Memorization trends remain the same and scale with model size and context
  - Absolute memorization rates are lower due to underrepresentation of duplicated data



(a)



(b)

Figure 2a and Figure 2b present the results. We observe similar qualitative trends with model scale and context length as in Figure 1. Larger models memorize more training examples than smaller models—and much more than the GPT-2 models that were not trained on The Pile. Similarly, providing more context to a model increases the likelihood we discover memorization. We can extract the last 50 tokens of a length-1000 sequence with 7% probability for the largest GPT-J 6B model, compared to 4% probability for the smallest 125M GPT-Neo model. (And both of these are much larger than the 2% probability of extraction for the 1.5B parameter GPT2-XL model.) These results, taken together, allow us to estimate a lower bound that there is at least 1% of The Pile dataset that is extractable by the 6B GPT-J model, but not by GPT-2 XL.

THE UNIVERSITY OF ARIZONA

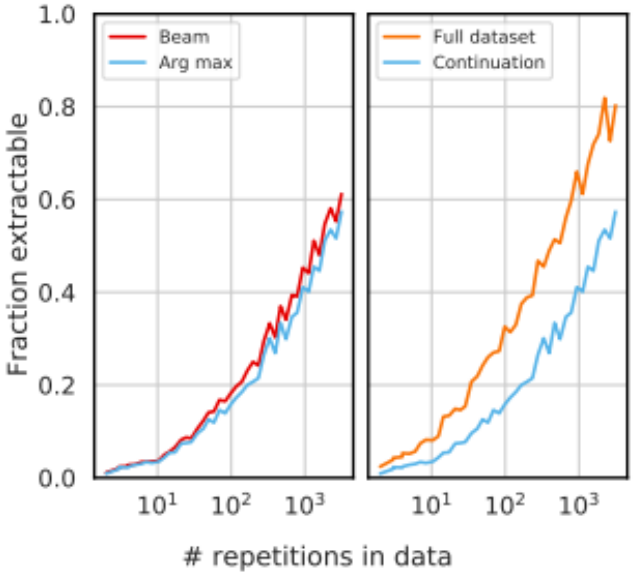
Alternate Experimental Settings

Alternate Decoding Strategies:

- Test Beam Search with 100 beams to find overall most likely continuation
- Slightly increases memorization by ~2% on average
- Did not present experiments on Random Sampling

Alternate Definition of Extracability:

- Checking if the model outputs the sequence *anywhere* in the dataset
- Significantly increases detected memorization rate



(c)

THE UNIVERSITY OF ARIZONA

Qualitative Examples of Memorization

Most universally memorized sequences “unconventional” texts such as code snippets or open source licenses

Frequency	Prompt	Continuation	6B
5526	contributors. All rights reserved. // // Licensed under the Apache License, Version 2.0 (the "License"); // you may not use this file except in compliance with the License. // You may obtain a copy of the	License at // http://www.apache.org/licenses/LICENSE-2.0 // // Unless required by applicable law or agreed to in writing, software // distributed under the License is distributed on an "AS	License at // // http://www.apache.org/licenses/LICENSE-2.0 // // Unless required by applicable law or agreed to in writing, software // distributed under the License
5652	ators are Standing By Our Dumpster Specialists are waiting by the phone to help you get started. However, there is some important information we need before we can give you a price quote and delivery time. Some things we need to know	: What size dumpster you will need What you plan on putting in your dumpster When you need us to drop your roll off container off How long you plan on keeping your roll-off City &	include; What size dumpster do you need? What type of waste do you have? How much waste do you have? What is the delivery time? What is the pickup time? What
4323	a copy of the License at # # http://www.apache.org/licenses/LICENSE-2.0 # # Unless required by applicable law	or agreed to in writing, software # distributed under the License is distributed on an "AS IS" BASIS, WITHOUT # WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See	or agreed to in writing, software # distributed under the License is distributed on an "AS IS" BASIS, # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
3556	date_default_timezone_set() function. In case you used any of those methods and you are still getting this warning, you most likely misspelled the timezone identifier. We selected the timezone 'UTC' for now, but please	set date.timezone to select your timezone. in /home/erlypro/public_html/natural/datas/persobanner_center.php on line 17 Deprecated: Function eregi_replace	set date.timezone to select your timezone. in /home/notes./public_html/wp-includes/functions.php on line 3812 Warning: date() [function.date]: It is not safe to rely
3920	| | [手把手翻墙教程](https://github.com/gfw-breaker/guides/wiki) &	nbsp;| | [禁闻聚合安卓版](https://github.com/gfw-breaker/bn-android) &n	nbsp;| [网门 安卓版](https://github.com/gfw-breaker/webpack-interpreter) [神州正道
3675	* / __| | | | |_) | | *	| (__| |_| | _ <| |___ *	| (__| || __| | *
3030	1775, "date": "1775\u20131185T", "displayOrder": 1, "fc": "Joseph Mallord William	Turner", "gender": "Male", "id": 558, "mda": "Turner, Joseph Mallord William",	Turner", "gender": "Male", "id": 604, "mda": "Turner, Joseph Mallord William",

Figure 14: Text examples that are heavily replicated in the training set but not memorized. We find many examples which have slight differences with no semantic (English) meaning. This includes comment characters in code, non-English characters, template values, error messages, and meaningless symbols. We also surprisingly find a large number of slightly different but heavily repeated documents about dumpsters.

Replication Study - T5 Masked Language Modeling

Model and dataset:

- T5 v1.1 models - masked encoder decoder models trained on C4 dataset
- New memorization definition - if a model *perfectly* solve the masked language modeling task on a sequence

Results:

- T5 memorization also scales with model size
  - Absolute memorization is an order of magnitude lower (eg. 3B T5-XL memorizes ~3.5% whereas similarly sized 2.7B GPT-Neo model memorizes 53.6%)
- Duplication often easier to memorize but the is no monotonic scaling relationship

THE UNIVERSITY OF ARIZONA

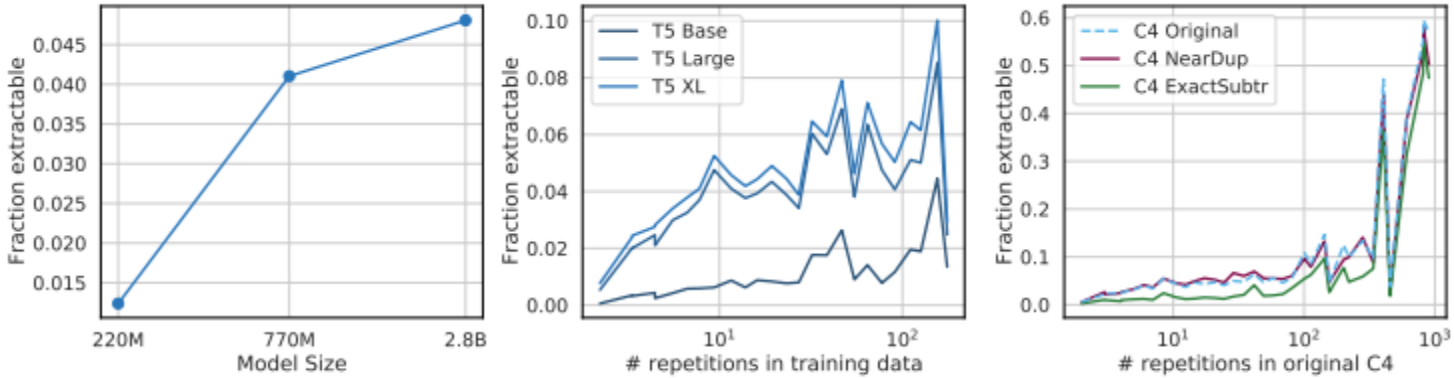


Figure 4: **(a)** Masked language model objective: Larger models have a higher fraction of sequences extractable on T5. **(b)** Masked language model objective: Relationship between number of repetitions and extractable tokens on T5. **(c)** Causal language model objective: Relationship between number of repetitions and memorization on language models trained with deduplicated data.

Replication Study - Training on Deduplicated Data

Model and dataset:

- 1.5B parameter causal language models trained on variations of C4: Original C4, C4 with near duplicate documents removed, C4 with exact duplicate sequences removed

Results:

- Models trained on deduplicated data memorize significantly less
- For extremely repeated sequences (≥ 400 times) deduplication does not help
    - Deduplication is necessarily imperfect to efficiently scale to hundreds of gigabytes of training data

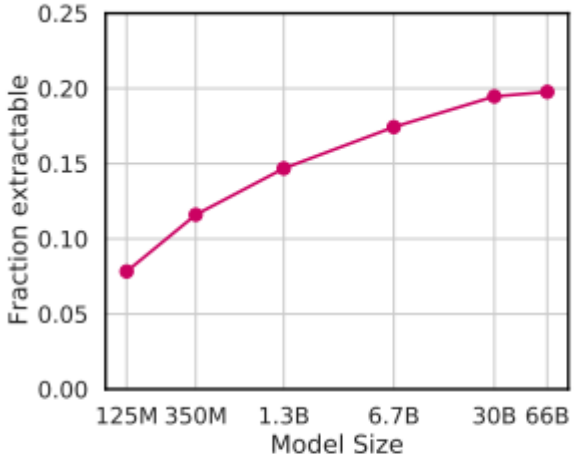
Replication Study - Modifying The Pile

Model and dataset:

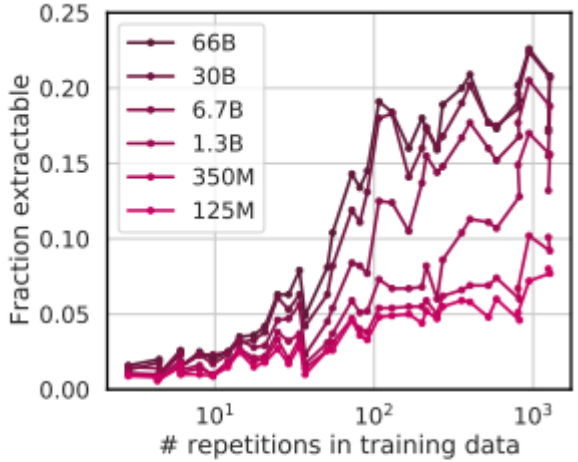
- OPT family models ranging from 125M to 66B parameters
- Trained on The Pile with some new sources, some sources removed, and more deduplication

Results:

- Nearly identical scaling trends, but orders of magnitude lower (66B OPT model memorizes *less* than 125M GPT-Neo model)
- Training methods and dataset curation may mitigate memorization
- Differences in data distribution may significantly alter what gets memorized



(a)



(b)

Figure 7: We prompt OPT models with data sampled from their training set. We use a prompt length of 100 here. **(a)** Fraction of sequences extracted as a function of model scale. **(b)** Fraction of sequences extracted as the number of repetitions of that sequence in the training set increases.

THE UNIVERSITY OF ARIZONA.

Conclusions

- Memorization scales log-linearly with model size, data duplication, and context length
  - Larger models memorize more training data, duplicated data is more likely to be memorized, longer prompts reveal more memorization
- Memorization happens systematically across different models
- Deduplicating training datasets is likely a practical way to mitigate harms of memorization