Session #11: Memorization and Privacy

Tuesday, October 4

CSCI 601.771: Self-supervised Statistical Models

JOHNS HOPKINS UNIVERSITY

Google

!

Some of your saved passwords were found online

 danyal.khashabi@gmail.com

---

Some of your saved passwords were found in a data breach from a site or app that you use. Your Google Account is not affected.

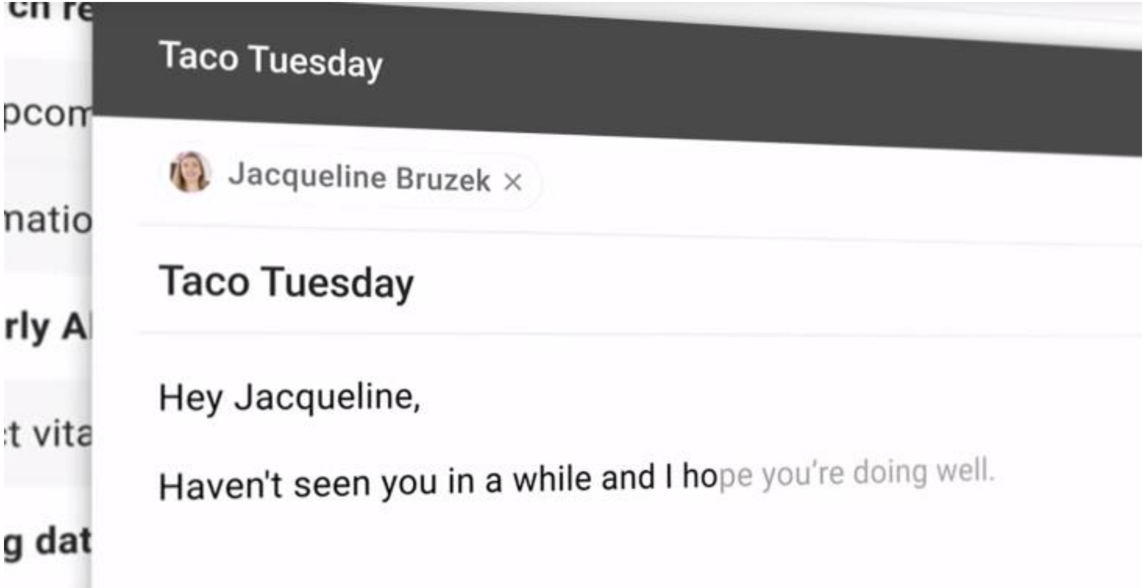
To secure your accounts, Google Password Manager recommends changing your passwords now.

Check passwords

You can also see security activity at
https://myaccount.google.com/notifications

Daniel Khashabi's password is

Elvis123


Taco Tuesday
Jacqueline Bruzek
Taco Tuesday
Hey Jacqueline,
Haven't seen you in a while and I hope you're doing well.


Conversation A
Hi Alice how are things going?
Bob
Not great…
Did I already tell you I'm getting a divorce?
Alice
No I'm sorry to hear that!
What are you going to do about custody of the kids?
Bob


Conversation B
Hey Bob how've you been??
Charlie
Pretty good wbu?
Did you hear Alice is getting divorced??
Bob

Quantifying Memorization Across Neural Language Models

Stake Holder Review

Large Models are Leaky

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

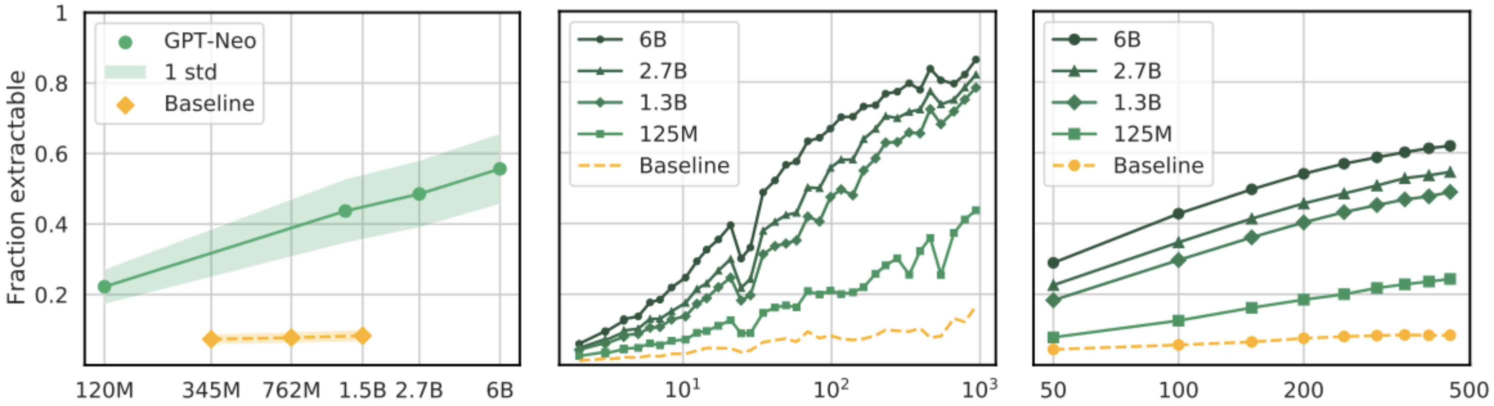
Paper Main Idea

As Language Models get Larger, Memorization within the model increases, and arises concerns

Properties that Impacts Memorization

- Model Scale: Larger models memorize 2-5X more than smaller models
- Data Duplication: Repeated words are more likely to be memorized
- Context: Longer context sentences are easier to extract

Graphs



(a) Model scale (b) Data repetition (c) Context size

What is memorization ?

- 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 || s] is contained in the training data for f , and f produces s when prompted with p using greedy decoding.

- Greedy decoding just picks the next token containing the largest probability – the argmax

Creating the dataset:

- Ideally we want to test on every sequence but this is too computationally expensive
- The authors use a small but REPRESENTATIVE sample to get statistical confidence (50,000 sequences)
- To account for duplication – the set is duplication normalized
- This means we have repeated sequences which influence's memorization!

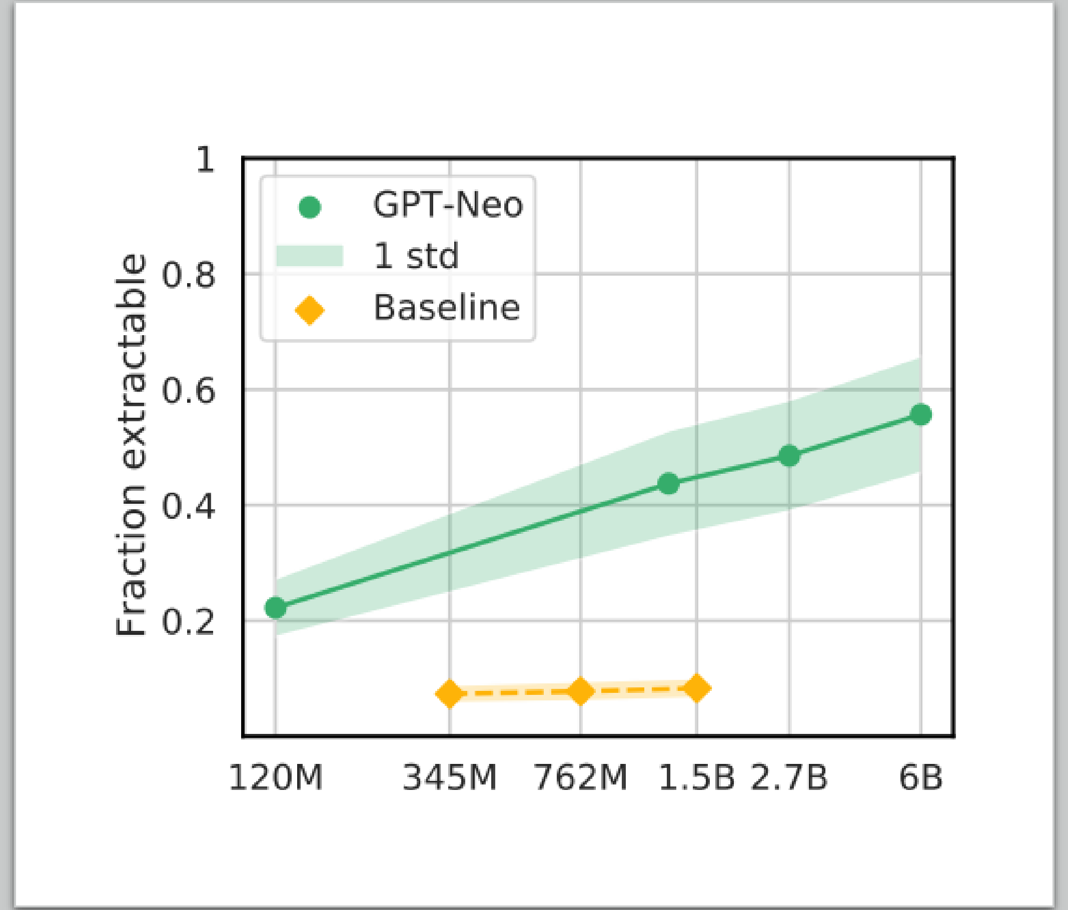
Setting up experiments:

- The Pile which is the largest publicly available dataset is used
- The GPT-Neo Family of models is used
- Parameters range from 125 million to 6 billion

Bigger Models Memorize More

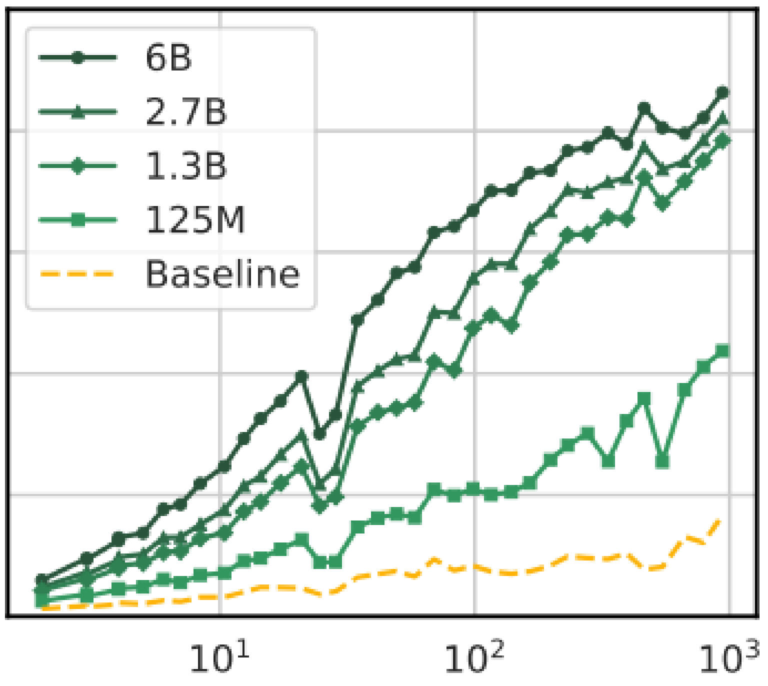
- There is a log linear trend with respect to increasing model size

- GPT-2 is used as a baseline which confirms the models are memorizing and not just generalizing



Repeated Strings Are Memorized More

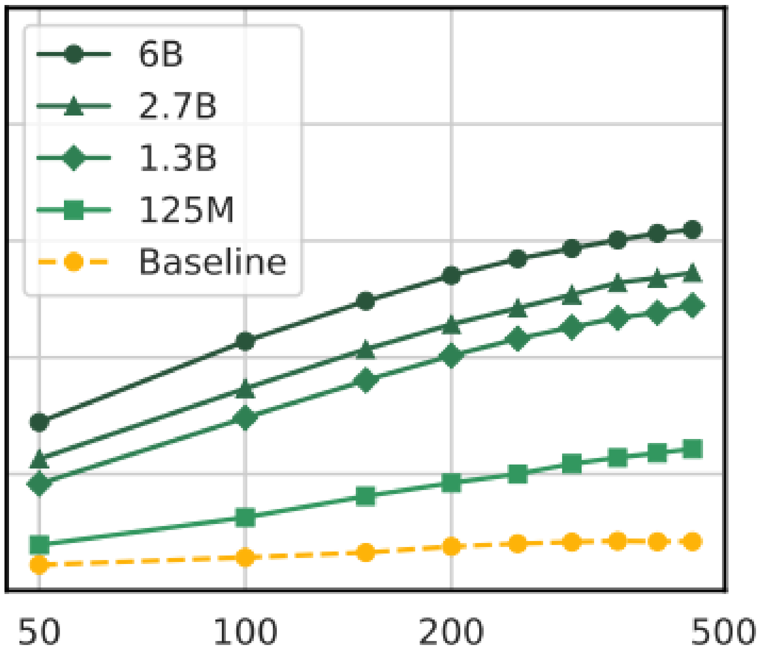
- Between 2 and 900 duplicates are tested on
- There is once again a log linear relationship between the number of repetitions and fraction extractable



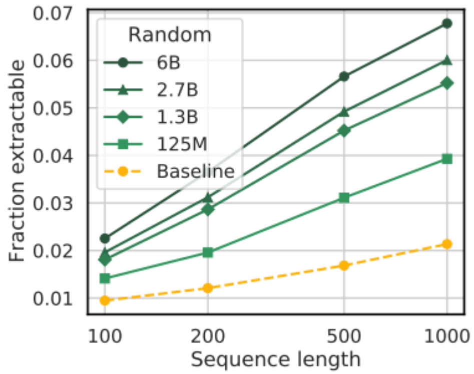
(b) Data repetition

Longer Context = More memorization

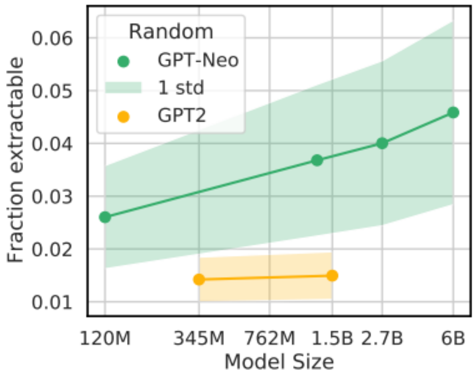
- Language models may only show memorization when prompted with sufficiently long context
- This is good as it protects privacy but may leave vulnerabilities open



(c) Context size



(b) Sequence length



(a) Model Scale

Random Data Set Sampling

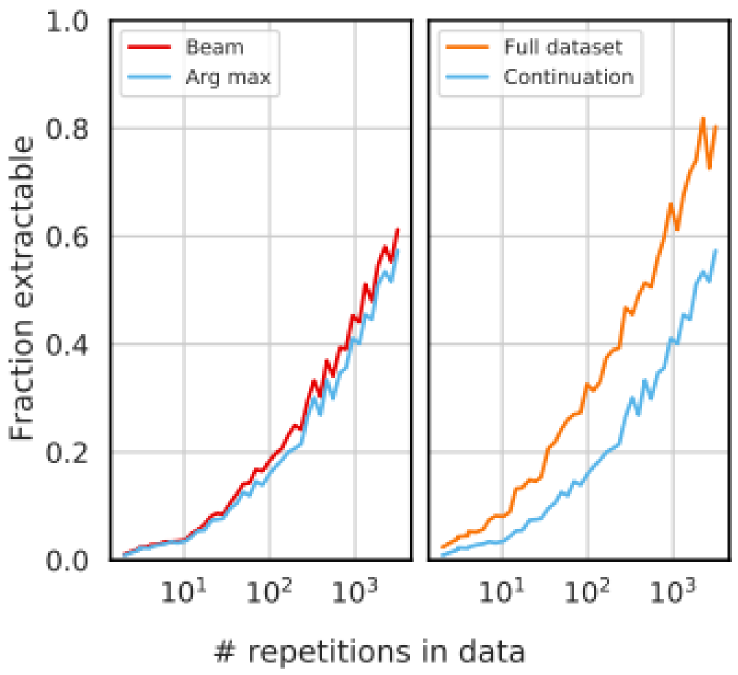
- We sample truly random sequences this time for a total of 100,000 unique sequences

- The overall probability of memorization is lower however, the trends remain the same

Different Strategies for search and decode

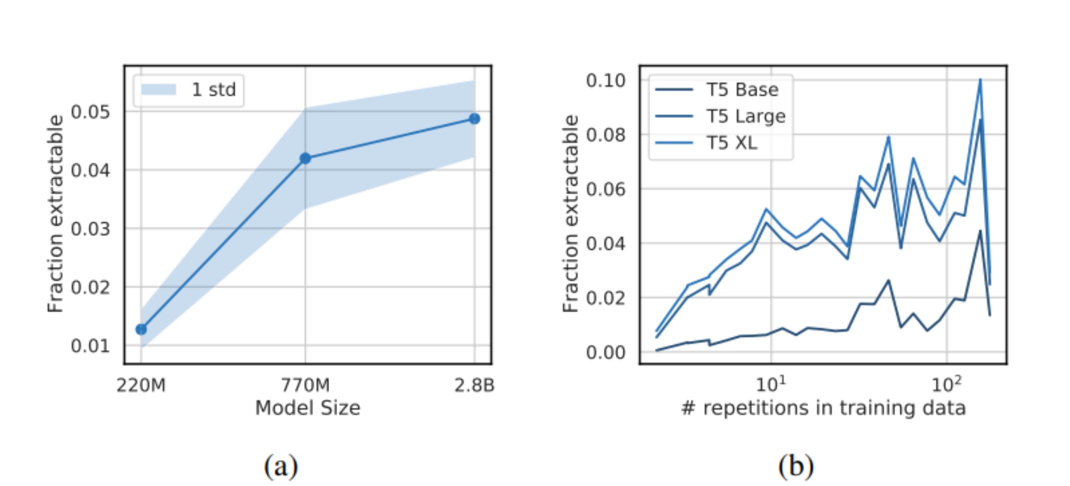
- Testing is done using **beam search** vs standard **greedy decoding**

- The second experiment tests for whether the prompt is anywhere in the data



(c) Decoding and search strategies

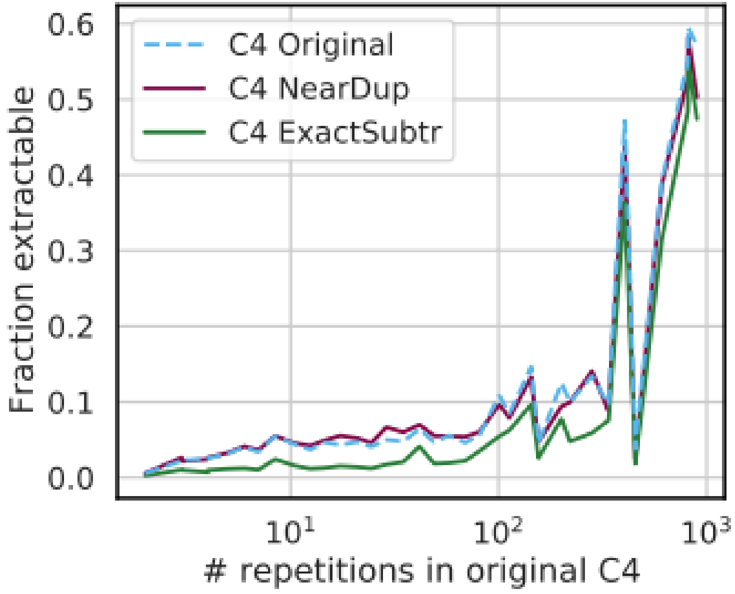
Replication study



- This was done on **T5 models**
- The relationship with **model size** is clear however not with **repetitions**

De–Duplication:

- Exhibits less **memorization** than **duplicated** dataset
- **De-duplication** is helpful up until approximately **100 repeats** after that it is imperfect



(c)

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*

Main Ideas

- Paper motivation:
  - Large language models memorize training data which can violate user privacy, degrade utility, and hurts fairness.
- Experiments on GPT-Neo model, GPT-2 model, and T5 masked language model. The main results:
  - Bigger models memorize more
  - Repeated strings are memorized more
  - Longer context discovers more memorization

Looking Beyond

- What other methods can we use to effectively prepare datasets to reduce memorization issue within large language models?
    - Paper proposes deduplication and finds that:
        - Models trained on deduplicated datasets memorize less data than models trained without deduplicated data sets
        - Deduplication does not help with for sequences repeated more than ~100 times.
- How can we determine possible prompts to use that will minimize the memorization issue in large language models?
- What other ways can we quantify memorization?

Testing for Memorization of Sensitive Information

- This paper measured direct memorization in LMs by testing if the model completes the text in training data if given context.
- Instead, test if models memorize associations between people and their data.

  <Name>'s physical address is _______

  This is unlikely to be the real prefix of a specific person's address in the training data. But we want to test if the model can associate the name with the address, assuming the data exists in the training corpus.

- Huang et al (2022) tested this with emails, attempting to get LMs to reveal email addresses.
- Deduplication could be used on sensitive information in the training dataset.

Reviewers

Pros

- Convincing Model and dataset choices
- Strong motivation

Definition 1 (Model Knowledge Extraction) *A string s is* extractable[4] *from an LM f_θ if there exists a prefix c such that:*

$$s \leftarrow \underset{s':\,|s'|=N}{\arg\max}\, f_\theta(s' \mid c)$$

We abuse notation slightly here to denote by $f_\theta(s' \mid c)$ the likelihood of an entire sequence s'. Since computing the most likely sequence s is intractable for large N, the arg max in Definition 1 can be replaced by an appropriate *sampling strategy* (e.g., greedy sampling) that reflects the way in which the model f_θ generates text in practical applications. We then define eidetic memorization as follows:

Definition 2 (k-Eidetic Memorization) *A string s is k-eidetic memorized (for $k \geq 1$) by an LM f_θ if s is extractable from f_θ and s appears in at most k examples in the training data X: $|\{x \in X : s \subseteq x\}| \leq k$.*

Figure1- (memorization definition) from *Extracting Training Data from Large Language Models*

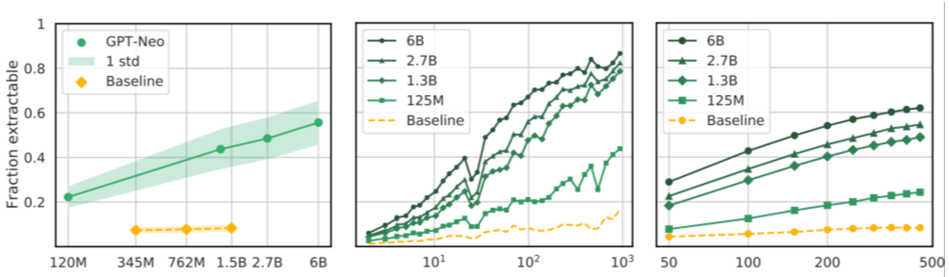
memorization such as familiar phrases, public knowledge or templated texts. In this paper, we provide a principled perspective inspired by a taxonomy of human memory in Psychology. From this perspective, we formulate a notion of *counterfactual memorization*, which characterizes how a model's predictions change if a particular document is omitted during training. We identify and study counterfactually-memorized training examples in standard text datasets. We further estimate the influence of each training example on the validation set and on generated texts, and show that this can provide direct evidence of the source of memorization at test time.

Figure2- (memorization definition) from *Counterfactual Memorization in Neural Language Models*

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.

Figure2- (memorization definition) from *the paper*

Some clarification/future work?



(a) Model scale (b) Data repetition (c) Context size

Figure 1: We prompt various sizes of GPT-Neo models (green) with data from their training set—The Pile. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though they were trained on WebText, a different dataset. **(a)** Larger models memorize a larger fraction of

Would the Pile-derived prompts give an accurate representation of how the GPT-2 family perform?

repeated either exactly or approximately-exactly). Because the frequency of training data duplication follows an exponential distribution (Lee et al., 2021), a fully random sample of only 50,000 sequences (accounting for $\leq 0.02\%$ of the dataset) is unlikely to contain *any* signal that would allow us to accurately measure the tail of this distribution.

How is it exponential?

Pros

- ☑ Analysis of diverse properties that significantly impact memorization
- ☑ Replication study on different language models and datasets
- ☑ Thorough Methodology and Experiments section
- ☑ Straightforward data visualization and examples

Limitations

- No analysis about the modified "duplication-normalized" dataset(not open-source)
- Only consider model comparing, no single model analysis (i.e. What's the repetition threshold of a model "remember" a certain sequence)

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 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

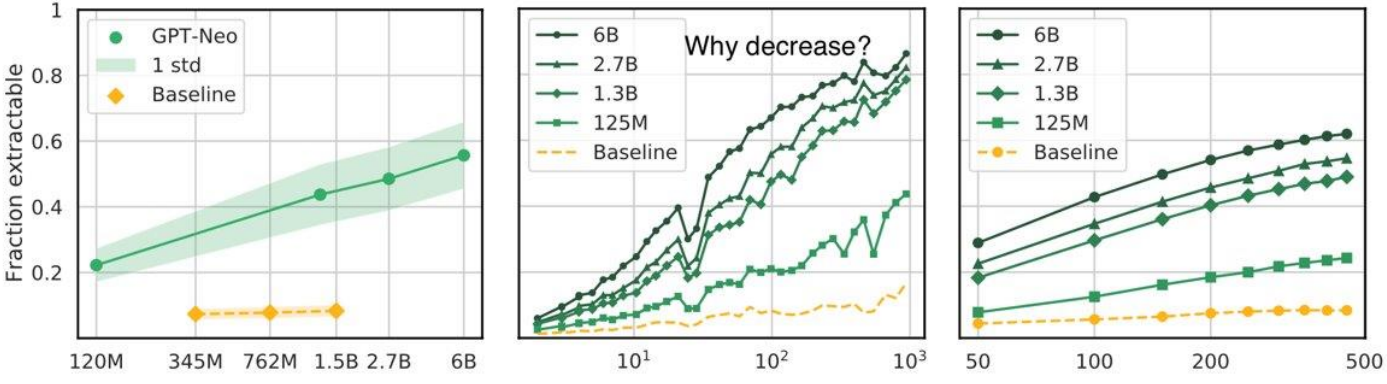
Limitations

- Report the sequence as “extractable” if the next 50 tokens (25words) emitted by the model exactly match
- Model has some amount of memorization not shared by each other model. (Even the 125M model memorizes a few sequences the 6B model does not.) Go over those sequences.

		Not Memorized By			
Model	Memorized	125M	1.3B	2.7B	6B
125M	4,812	-	328	295	293
1.3B	10,391	5,907	-	1,205	1,001
2.7B	12,148	7,631	2,962	-	1,426
6B	14,792	10,273	5,402	4,070	-

Limitations

- Fraction extractable decreased in specific range for data repetition
- No discussion on why this phenomenon occurs



(a) Model scale (b) Data repetition (c) Context size