Quantifying Memorization Across Neural Language Models

Nicholas Carlini, Daphne Ippolito, Matthew Jaglieski, Katherine Lee, Florian Tramer, Chiyuan Zhang

Published at ICLR 2023

Vishnu Asutosh Dasu

CSE 587: Deep Learning for NLP

Introduction

Introduction

- Large language models (LMs) have been shown to memorize training data when using carefully crafted prompts
- Memorization is undesirable as it affects privacy, utility, and fairness
- Memorization is worse than actually believed and will be getting worse unless we address it
- Prior works show a loose bound on how much data can be extracted and they do not explain how memorization varies across models and datasets

Introduction

- Two improvements over prior work
  - Better lower bound on how much data is memorized
  - Explain how memorization varies across different LMs and datasets of different scales
- Three properties affect memorization:
  - Model Scale: Larger models in same family memorize 2-5x more than smaller ones
  - Data Duplication: Examples repeated more often are more likely to be extractable
  - Context: It is much easier to extract text when the context is longer

Methodology

Definition of 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.

Example

- Training Sequence: *My phone number is 555-6789*
- Prefix string (p) with length k=4: *My phone number is*
- The sequence is extractable at k=4 if the most likely output is *555-6789*

Selection of Evaluation Data

Uniform Random Sampling

- Generate uniform random sample of 50,000 sequences from training dataset without repetition
- Useful to measure absolute memorization
- Cannot be used to study how memorization scales with data properties that are not represented uniformly in the subset
- Does not account for:
  - Prompt length
  - Data duplication

Selection of Evaluation Data

Normalized Random Sampling

- Generate subset normalized by duplication and sequence length
- For each length *l in {50, 100, 150, …, 500}* and integer *n*, select 1,000 sequences of length *l* that are contained between $2^{n/4}$ and $2^{(n+1)/4}$ times
- Repeat until you reach a value of *n* for which 1000 sequences are not available
- Biased sampling allows to measure memorization as a function of sample's duplication factor and sequence length

Selection of Evaluation Data

Evaluation Method

- For each length 50 to 500, collect 50,000 samples of varying duplication for a total of 500,000 sequences
- For each sequence of length l, prompt the model with first $l - 50$ tokens
- Sequence is “extractable” if model emits next 50 token suffix
- Compute average probability that sequence is extractable by averaging over all lengths l

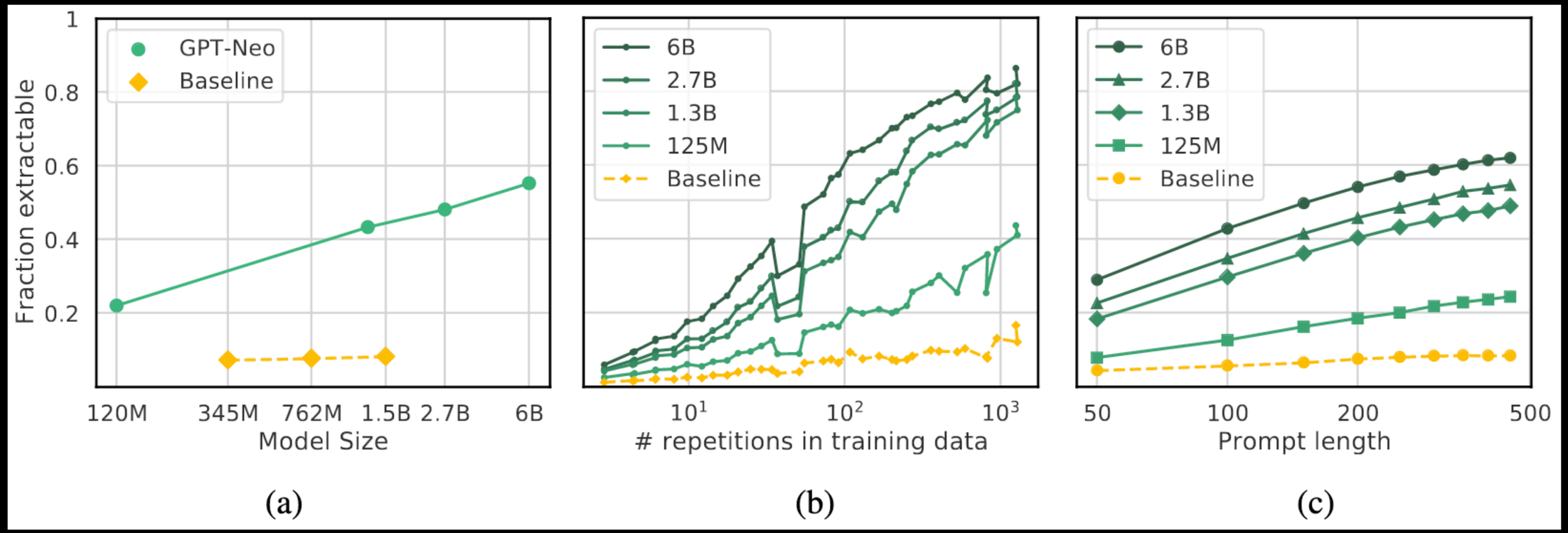
Experiments

Setup

- GPT-Neo is used as the language model
- GPT-Neo is trained on the Pile dataset
- Baseline model for comparison is GPT-2 trained on WebText
- Both, GPT-2 and GPT-Neo, are prompted from the Pile dataset

Bigger Models Memorize More

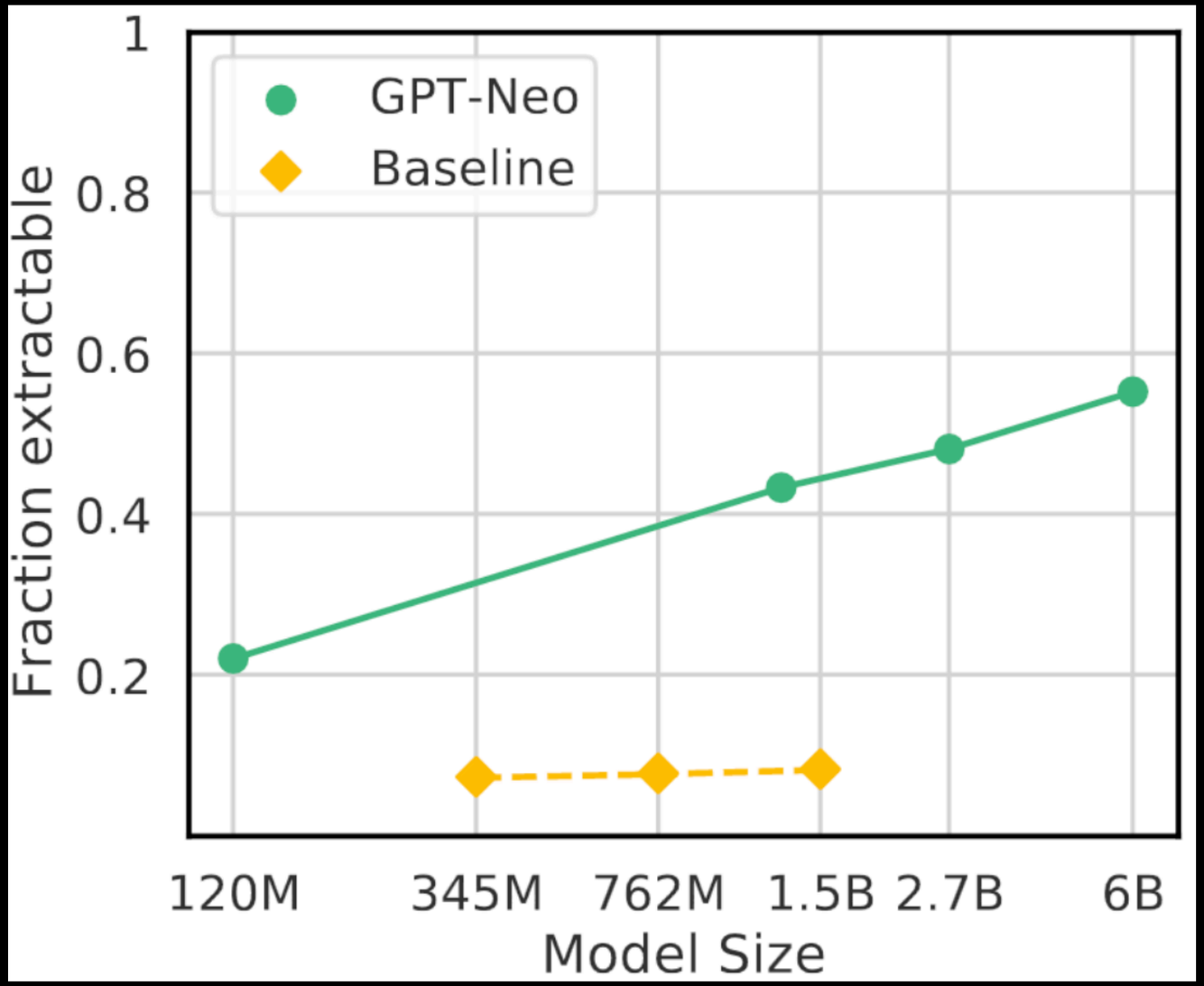
Results



Bigger Models Memorize More

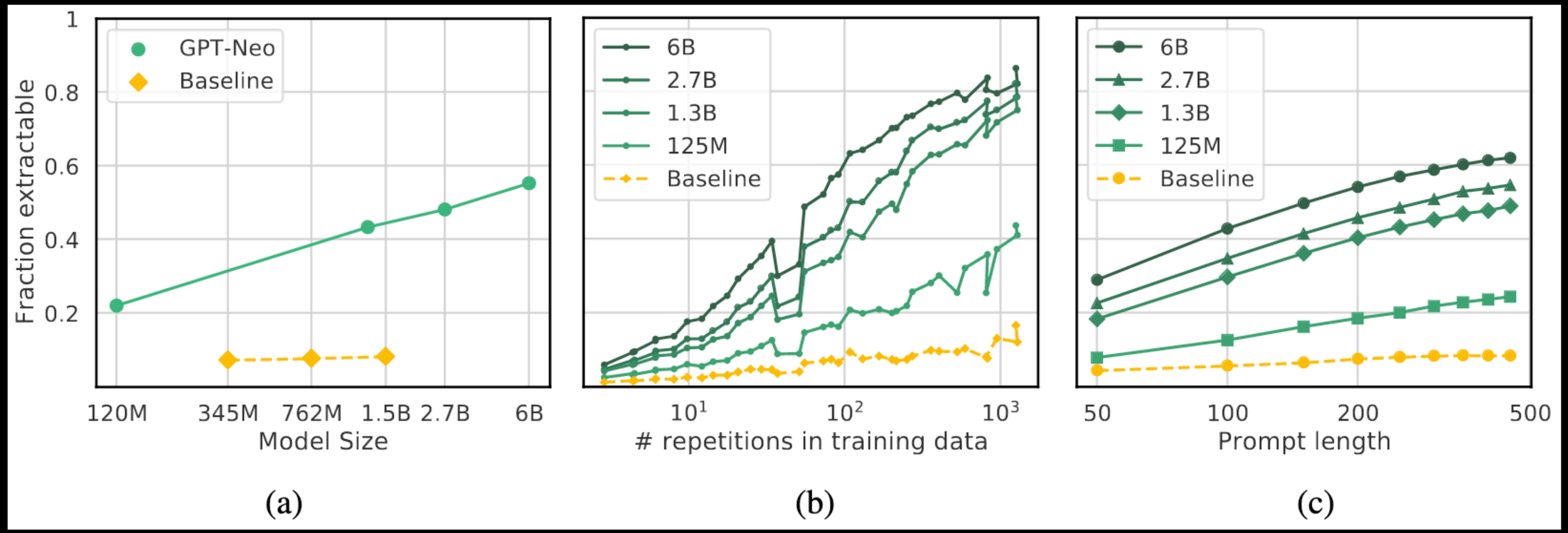
Analysis

- Biased sampling is used so we don’t care about absolute numbers of memorized data
- Results show that there is a log-linear relationship between scale and memorization
- Baseline shows that this is *memorization* and not *generalization*



Repeated Strings Are Memorized More

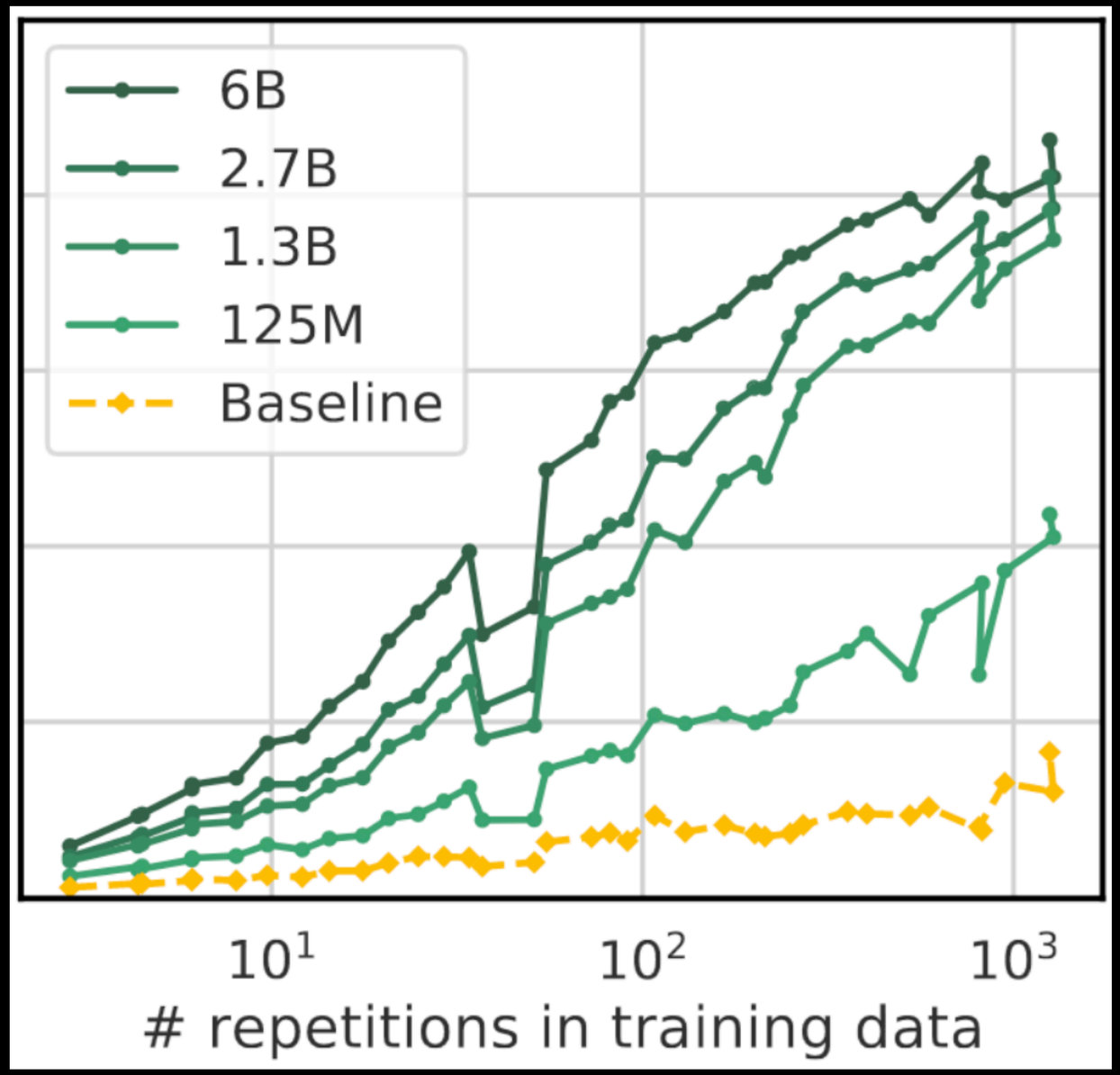
Results



Repeated Strings Are Memorized More

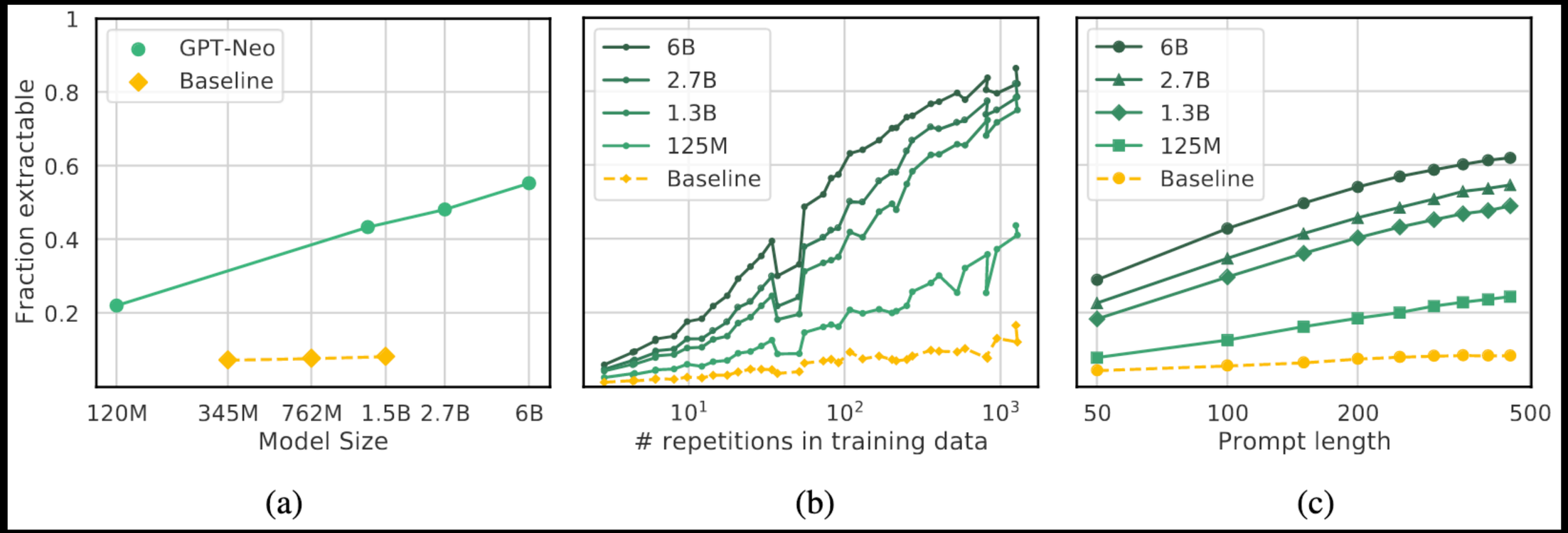
Analysis

- 1,000 distinct sentences, each repeated 2 to 900 times
- Less memorization for less duplication
- Similar log-linear trend is observed
- Memorization happens even with very few duplicates



Longer Context Discovers More Memorization

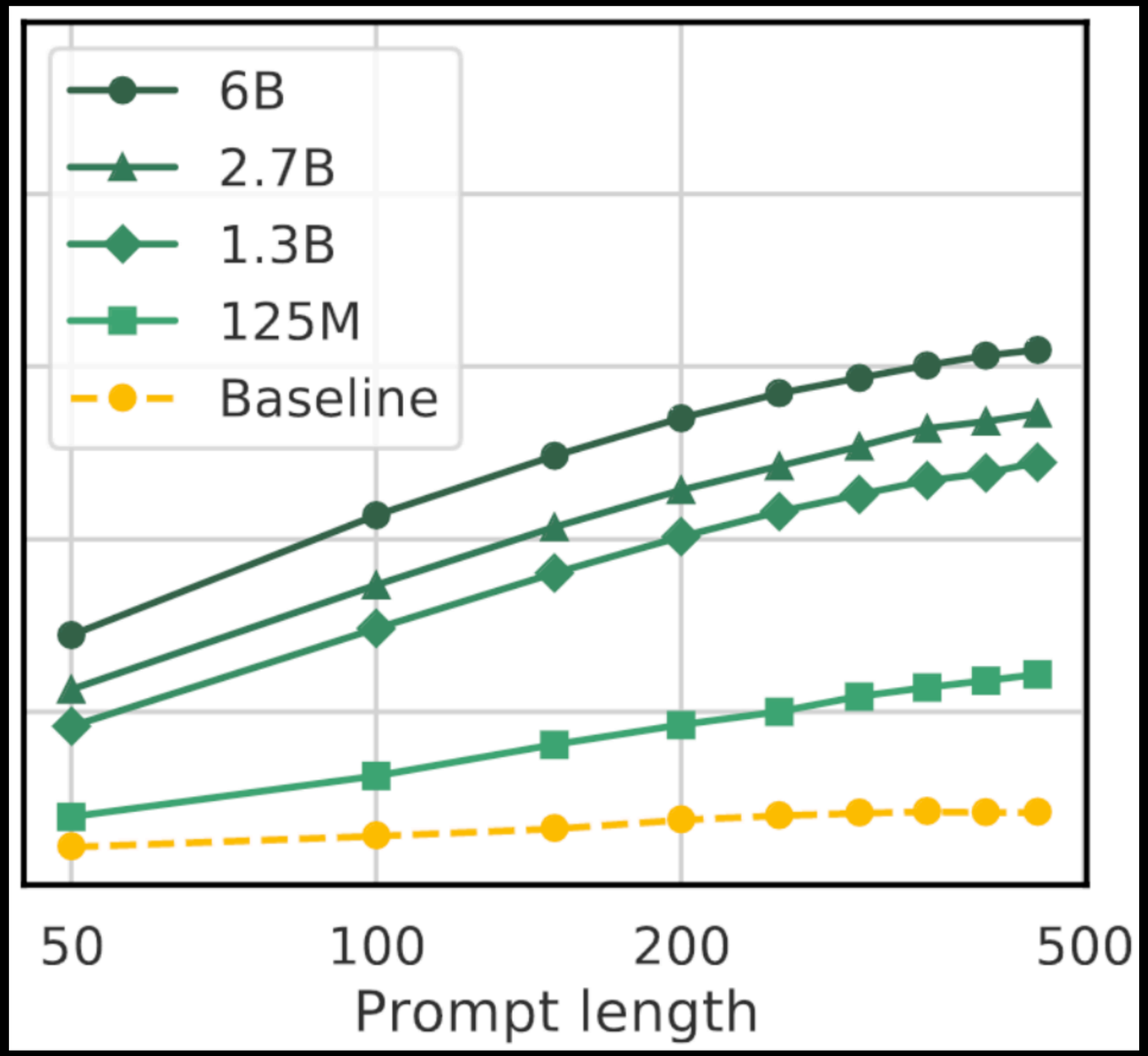
Results



Longer Context Discovers More Memorization

Analysis

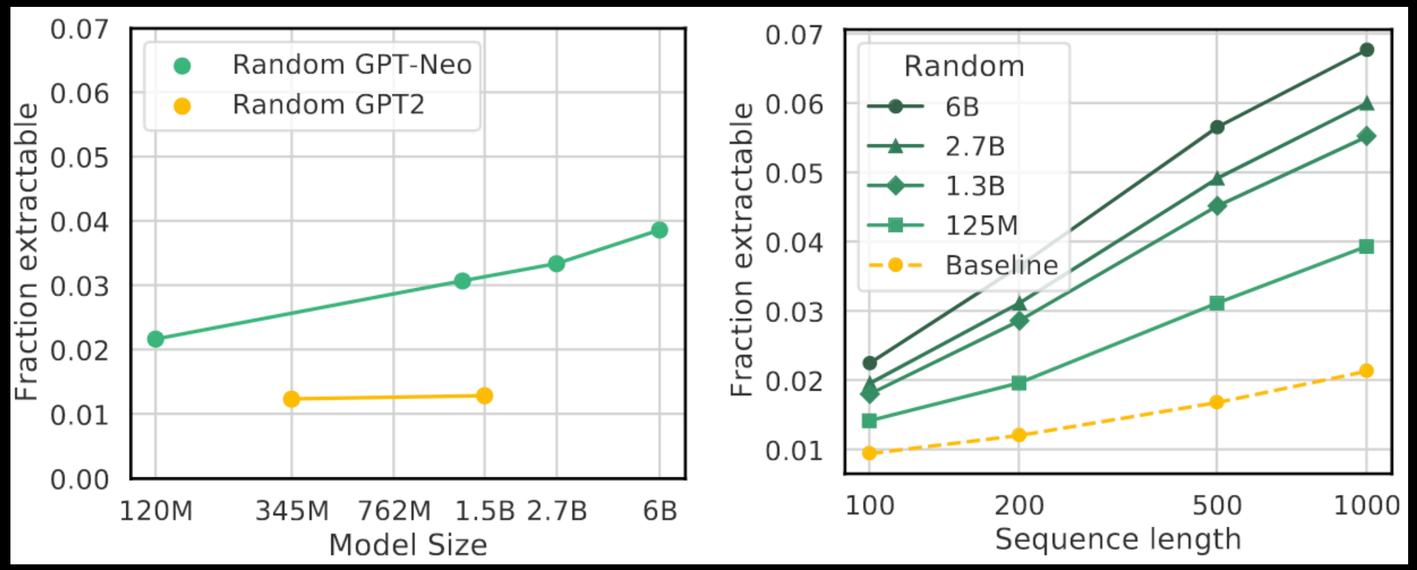
- Discoverability Phenomenon: Some memorization becomes apparent only for long contexts
- Some strings are “hidden” in the model and require more information to be extracted



Alternate Experiment Settings

Random Dataset Sampling

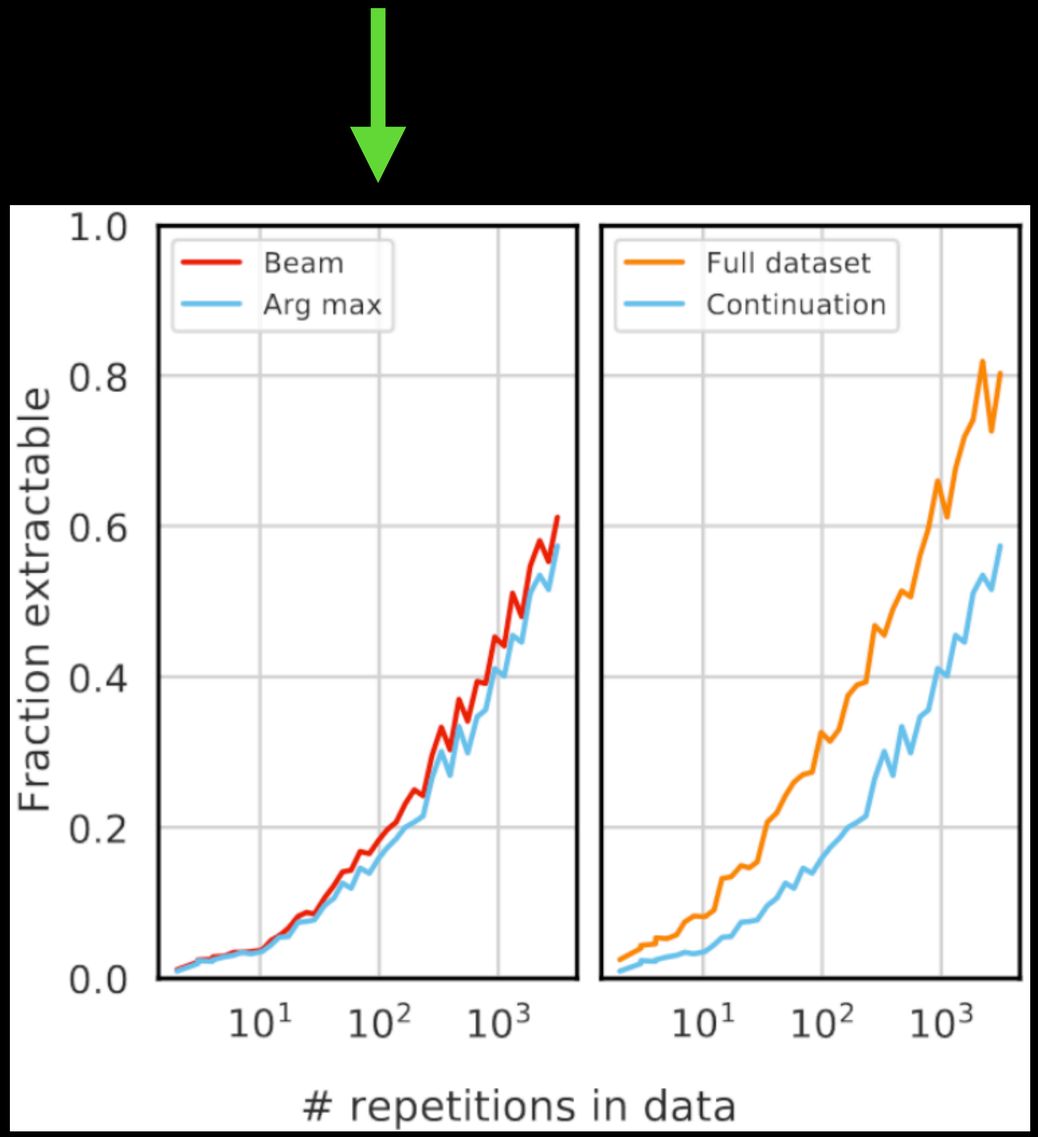
- Randomly sample 100,000 sequences
- Similar trends are observed for model size and context length



Alternate Experiment Settings

Alternate Decoding Strategies

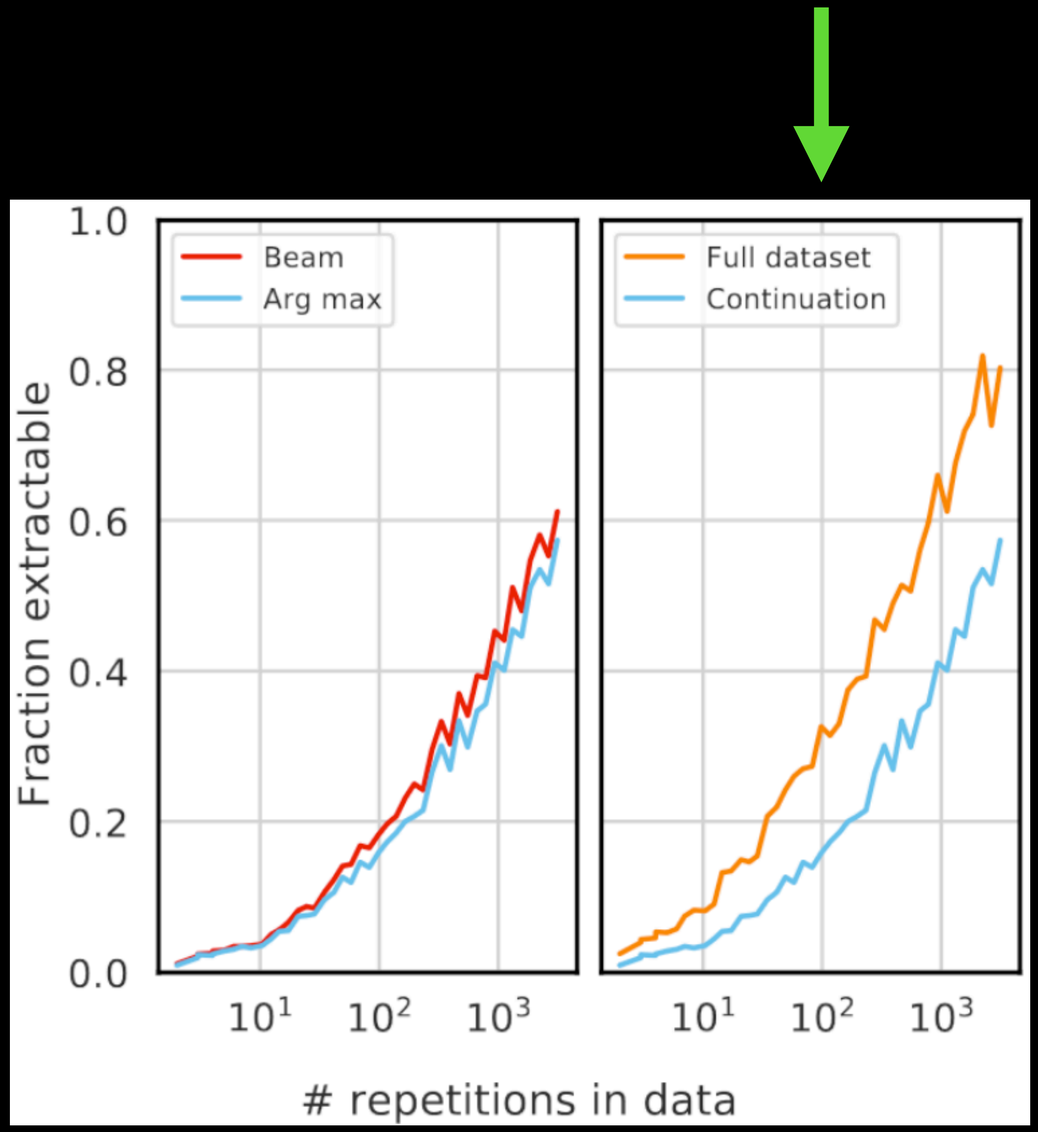
- Experiments so far were performed using the *greedy decoding* strategy where the most likely token was chosen during decoding
- Experiments were performed again with *beam search*
- Beam search is slightly better than greedy decoding



Alternate Experiment Settings

Alternate definition of extractability

- Previous experiments assumed that the suffix extracted has to belong to that particular training sample
- Now the suffix that is extracted can belong to *any* training sample in the dataset
- Naturally, this results in more extractions



Qualitative Examples of Memorization

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

Replication Study

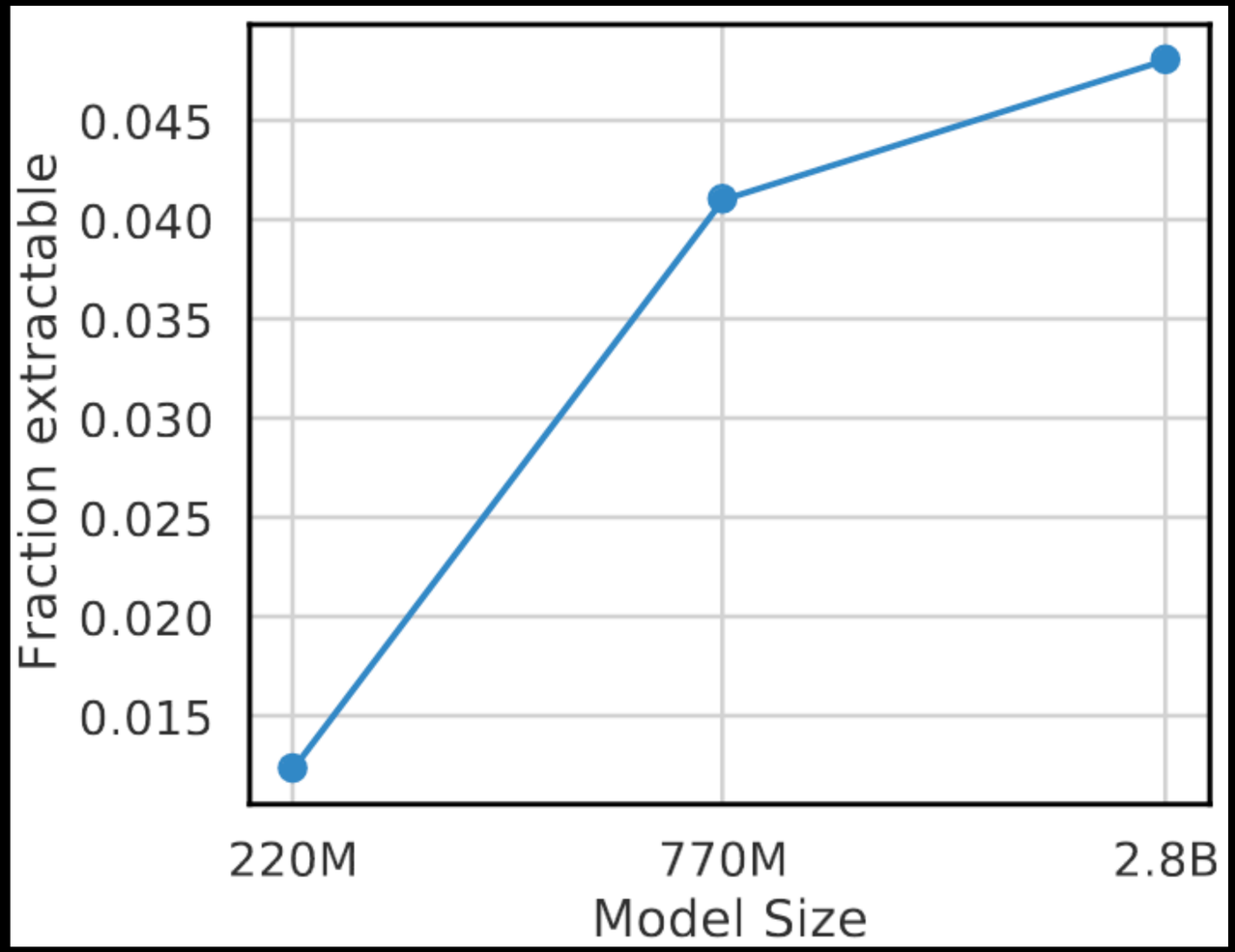
T5 Masked Language Modeling

- For masked language models, string is “extractable” if model can perfectly fill in all the masks
- T5 models are trained by removing 15% of tokens from each training sequence and model then must fill in the blanks of the sequence
- Example: A 200-token sentence is memorized if model can use 170 (200 * 0.85) tokens of context to predict remaining 30 (200 * 0.15)

T5 Masked Language Modeling

Model Size Results

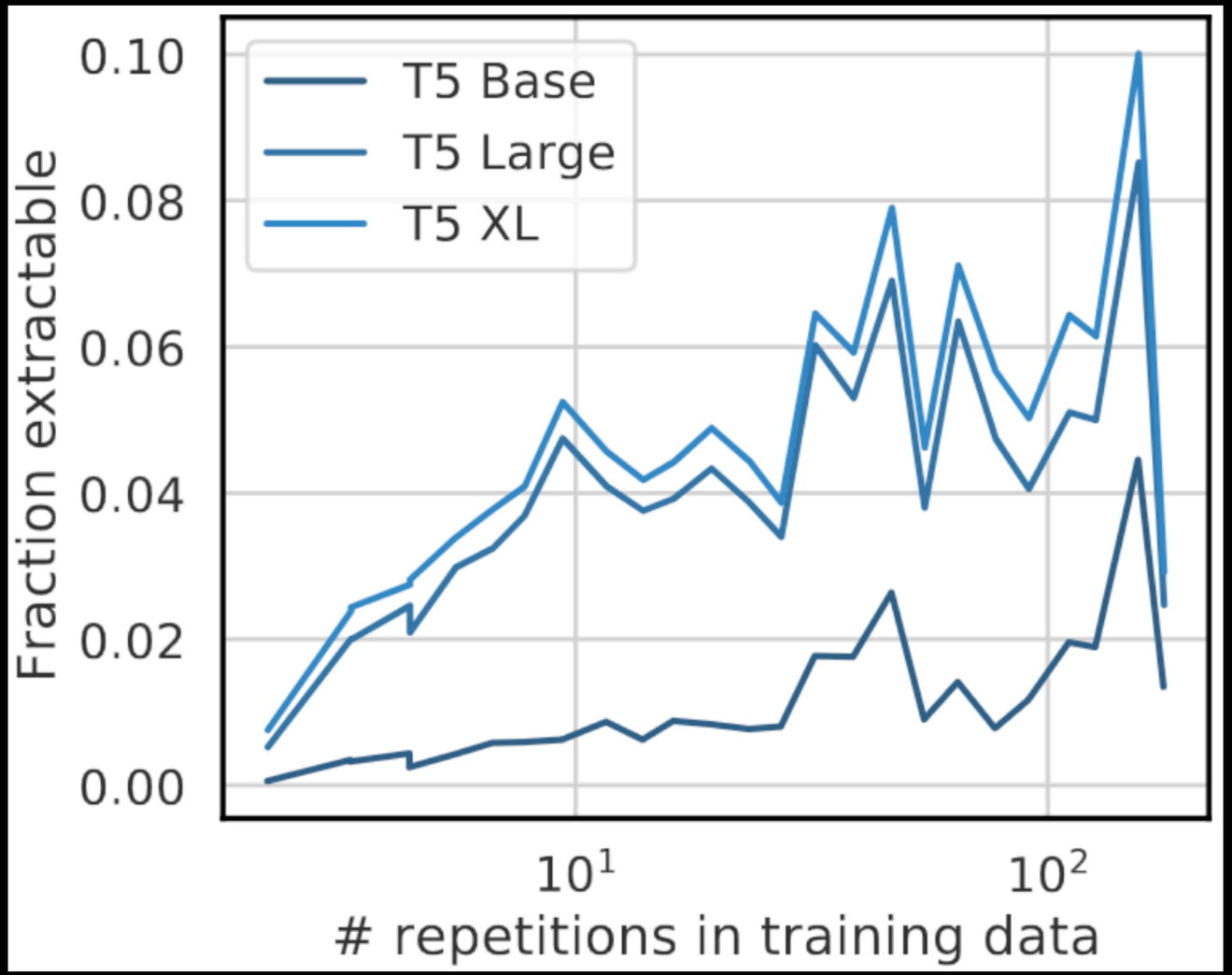
- Similar trend is observed where larger models memorize more
- However, masked models memorize significantly lesser than causal models



T5 Masked Language Modeling

Data Duplication Results

- No monotonic scaling relationship is observed and results are noisy with high variance
- Samples repeated 158 - 196 times are memorized with probability less than 5.1%
- Samples related 138 - 158 times are memorized with probability at least 6.2%
- Samples occurring ~140 times are more likely to be memorized despite occurring less often as these samples consist of mainly whitespace tokens which can be predicted easily



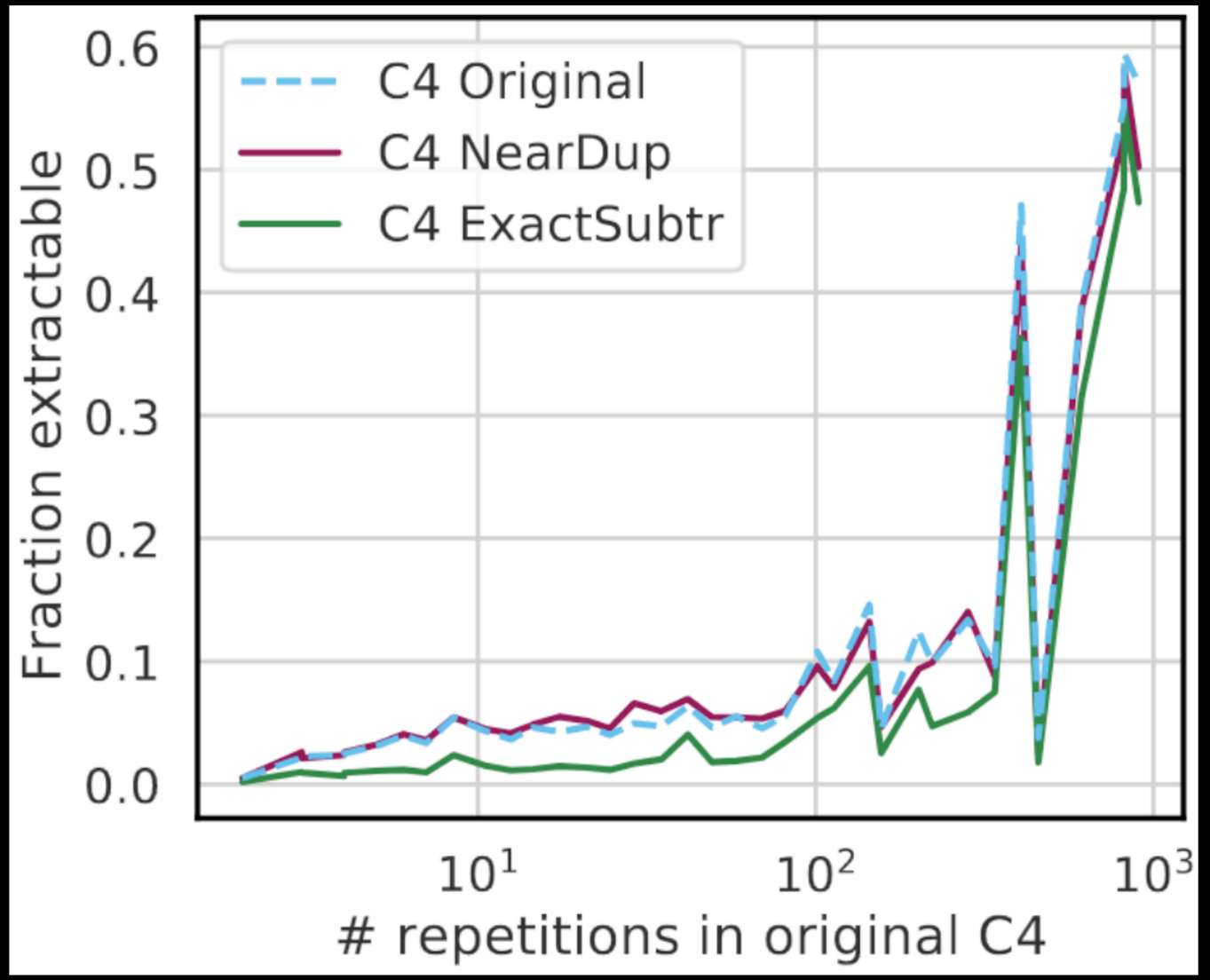
Language Models Trained on Deduplicated Data

- 1.5B parameters causal models trained on C4 with data deduplication
- Two types of deduplication:
  - Remove all documents that are near-duplicates of other documents
  - Delete any string of length-50 tokens that occurred more than once

Language Models Trained on Deduplicated Data

Results

- Models trained on deduplicated datasets memorize less
- Extractability of samples repeated at least 408 times is much higher as deduplication strategies are imperfect and cannot effectively scale to large training data

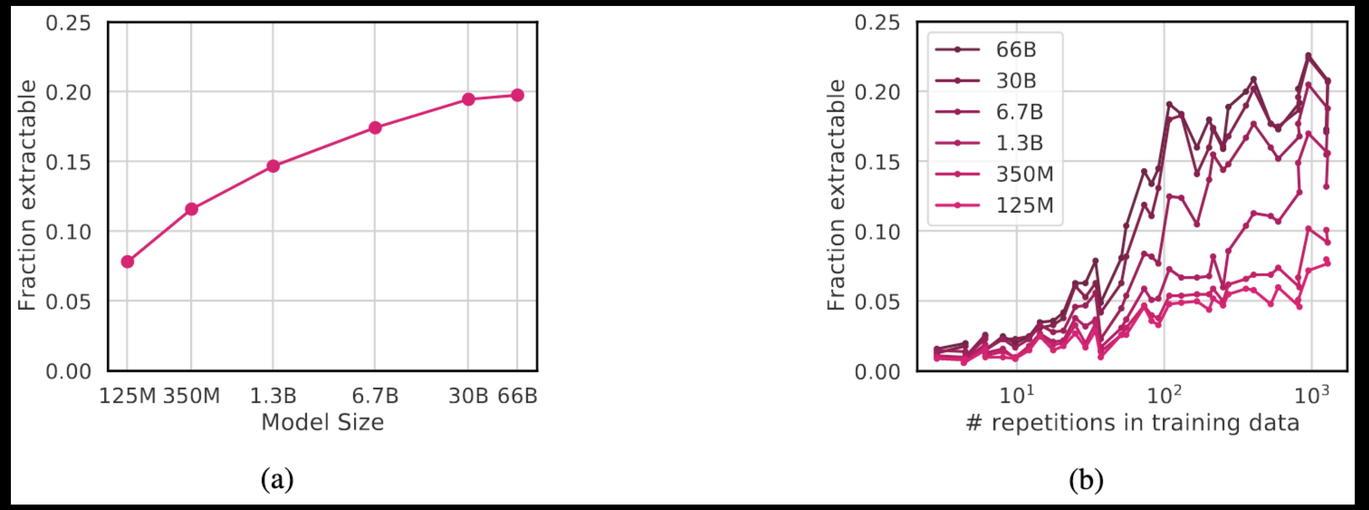


Language Models Trained on Modified Version of the Pile

- OPT models were trained on modified version of Pile which contains data from many other sources
- Dataset was deduplicated before training to reduce repetitions

Language Models Trained on Modified Version of the Pile

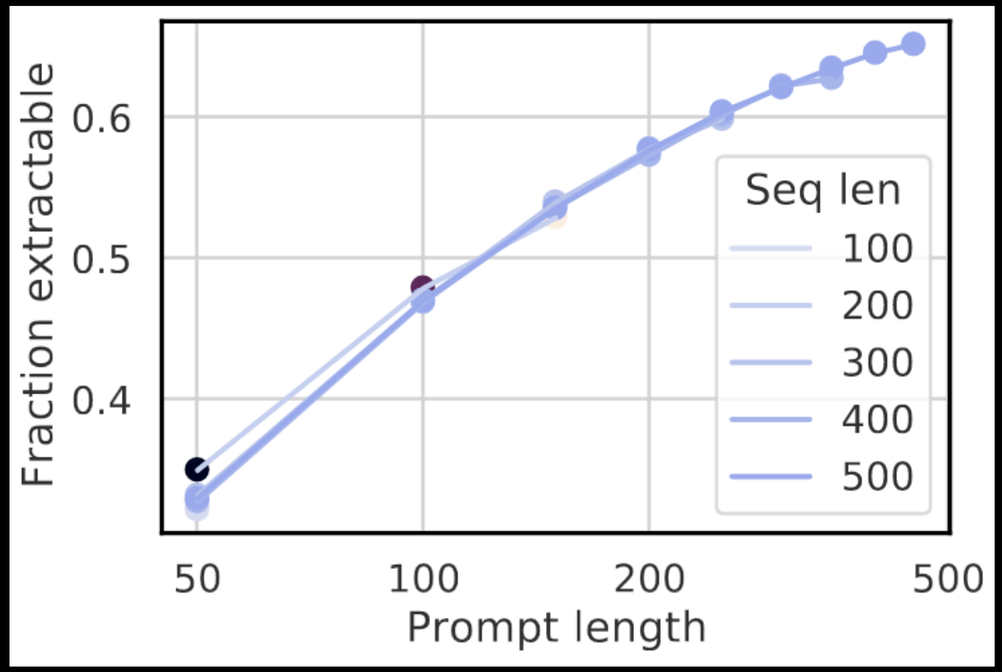
- Similar trends to GPT-Neo trained on Pile but much lesser memorization
- Two possible reasons:
  - Careful data curation can reduce memorization
  - Slight shifts in data distribution can alter what content gets memorized



Additional Results

Longer Documents Are Not Easier to Memorize than Shorter Documents

- Longer sequences are rare and if perplexity is low then it is likely that longer sequences are just memorized i.e. effect of larger context is not relevant
- For varying sequence and prompt lengths, extract the next 50 tokens
- No difference between fraction of extractable tokens by varying prompt lengths across sequence lengths



Text Memorized by Only Some Models

		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	-

- Larger models have more uniquely memorized sequences
- However, every model memorizes some amount of unique information

Conclusion

Conclusion

- Paper presents quantitative analysis of memorization in LMs
- Two primary conclusions:
  - Generalization: LMs accurately model statistics of training data but do not model the underlying data distribution correctly
  - Privacy: LMs memorize significant fraction of their training data. Memorization scales log-linearly with model size and is often hard to discover
- Data deduplication strategies seem to be a promising direction to reduce memorization