Session #12: Memorization and Privacy

Tuesday, October 6
CSCI 601.771: Self-supervised Statistical Models

#12 - Thu Oct 6	Memorization and Privacy	Slides Main Reading: Deduplicating Training Data Mitigates Privacy Risks in Language M Additional Reading(s): 1. Differentially Private Fine-tuning of Language Models 2. Can a Model Be Differentially Private and Fair? 3. Large Language Models Can Be Strong Differentially Private Learners 4. What Does it Mean for a Language Model to Preserve Privacy?
#13 - Tue Oct 11	External Speaker: Anjalie Field	
#14 - Thu Oct 13	Project Proposal Presentation	Slides

- Project proposal presentation
- 7 minutes per team

Suggested structure:

1. Motivation: 1 min
2. Attack plan: 4 mins
3. Expected outcome: 2 mins

#12 - Thu Oct 6	Mem	gates Privacy Risks in Language Mo
		uage Models 2. Can a Model Be Differentially Private and Fair? 3. Large Language Models Can Be Strong Differentially Private Learners 4. What Does it Mean for a Language Model to Preserve Privacy?
#13 - Tue Oct 11	External Speaker: Anjalie Field	
#14 - Thu Oct 13	Project Proposal Presentation	Slides

Deduplicating Training Data Mitigates Privacy Risks in Language Models

Nikhil Kandpal
UNC Chapel Hill
nkandpa2@cs.unc.edu

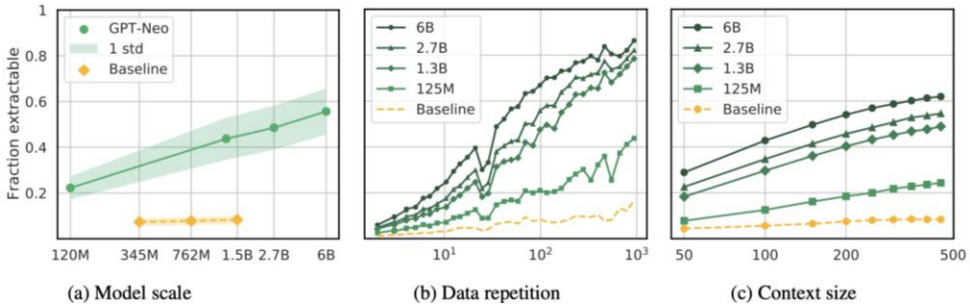
Eric Wallace
UC Berkeley
ericwallace@berkeley.edu

Colin Raffel
UNC Chapel Hill
craffel@gmail.com

✍: Aayush Mishra, Boyuan Zheng

Background

- Language Models can regenerate training data *[Carlini et al., 22]*. Up to 1000 words long! *[McCoy et al., 21]*



- Language Models are susceptible to Model Inversion attacks *[Shokri et al., 17, Friedrikson et al., 15, Carlini et al., 21b].*
- Differential Privacy frameworks *[Dwork et al., 06, Yu et al., 21, Li et al., 22]* give weak guarantees of safety against such attacks.

✍: Aayush Mishra, Boyuan Zheng

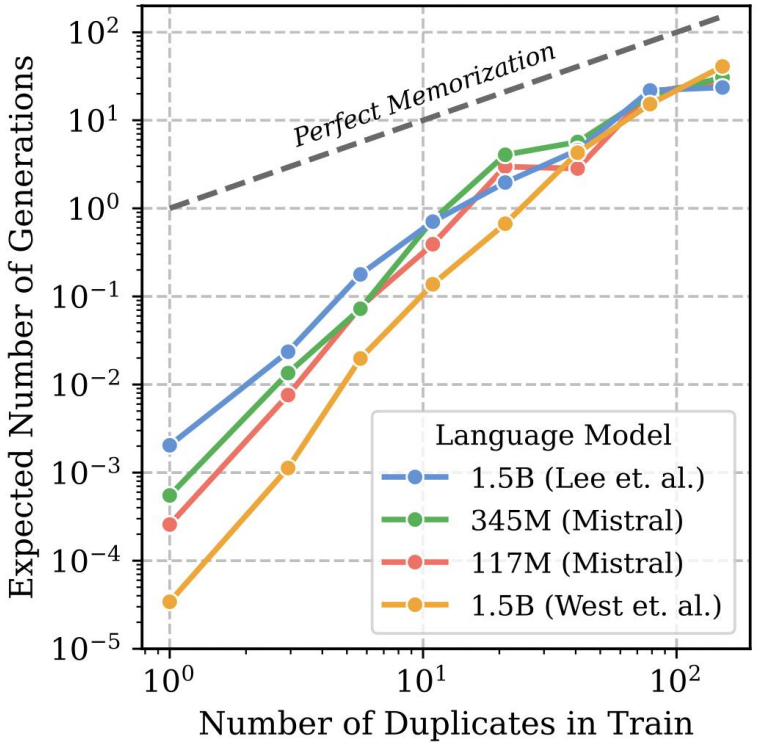
Motivation

1. How often does the model generate duplicated training sequences?
2. Do current adversarial attacks work for non-duplicated sequences?
3. How well does deduplication prevent privacy attacks?

Main results

- Attack likelihood of recovering a sequence is correlate with number of times it was duplicated in training data.

- De-duplication reduces attack efficacy.



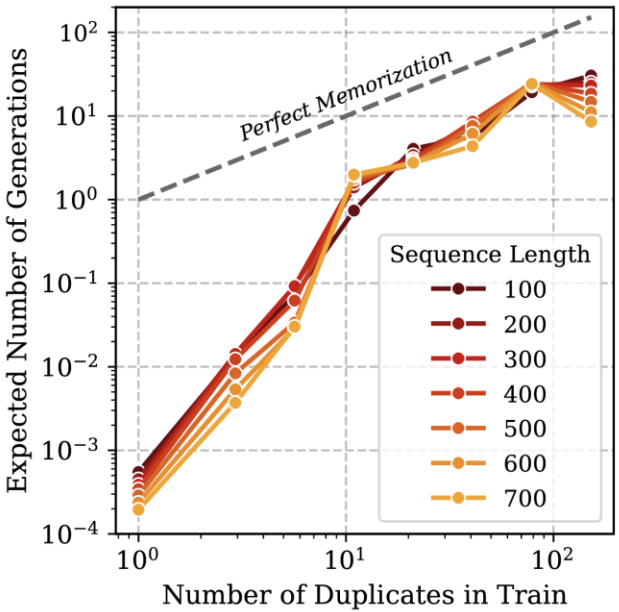
: Aayush Mishra, Boyuan Zheng

How duplication affects sequence regeneration

- How generated sequences duplicate training corpus
- Method:
    - Record Duplication number of N-length sequences
    - Generate from LM
    - Analyze how often is generation is a function of its count

: Aayush Mishra, Boyuan Zheng

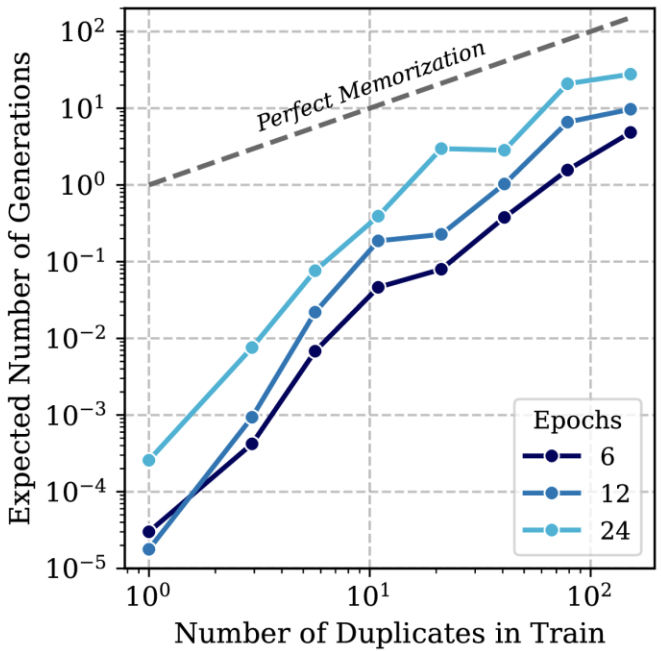
Influence of Sequence Length



- Longer Sequence reduce overlap chance
- Tendency(Generation, Duplication) is consistent
- Metrics: 1.Receiver Operating Characteristic (ROC); 2.True Positive Rate(TPR), False Positive Rate(FPR)
- Sequence Length, Mistral 345M

: Aayush Mishra, Boyuan Zheng

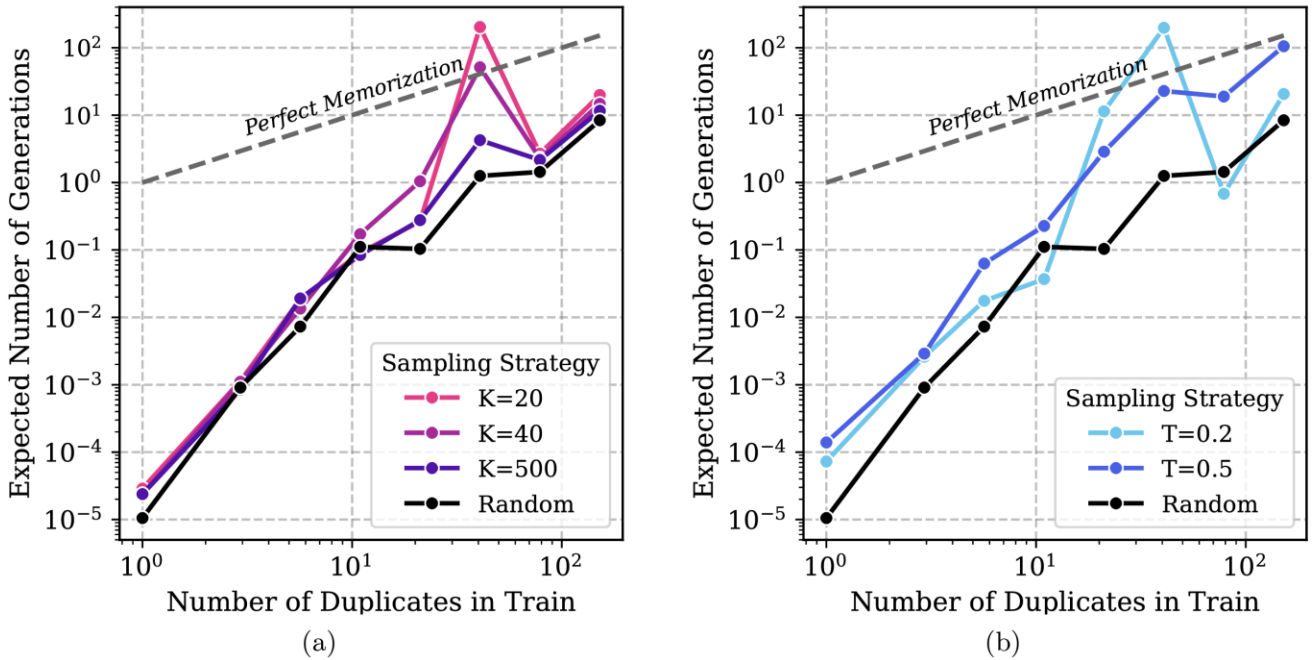
Influence of Epoch Number



- Mistral 117M
- More batch, More overlap
- Tendency(Generation, Duplication) is consistent

✍: Aayush Mishra, Boyuan Zheng

Results



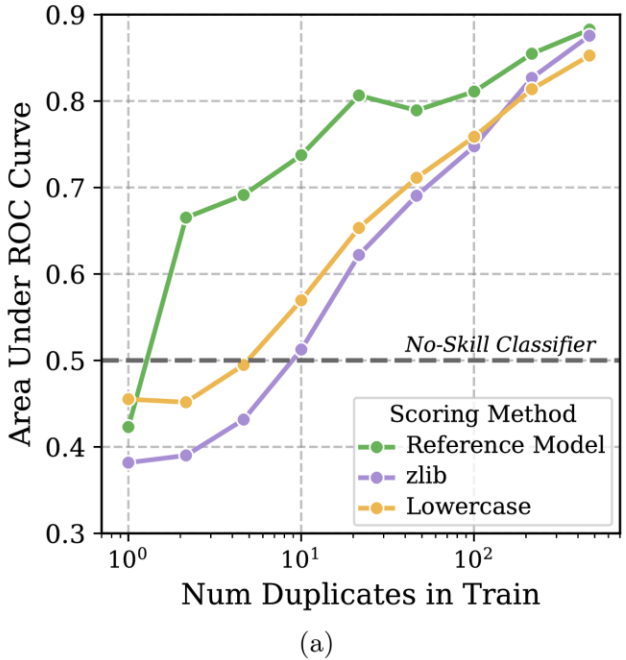
- Superlinear (#Regeneration, #Duplication)
- Weak Memorization: #Duplication << # Regeneration (Especially low-freq)

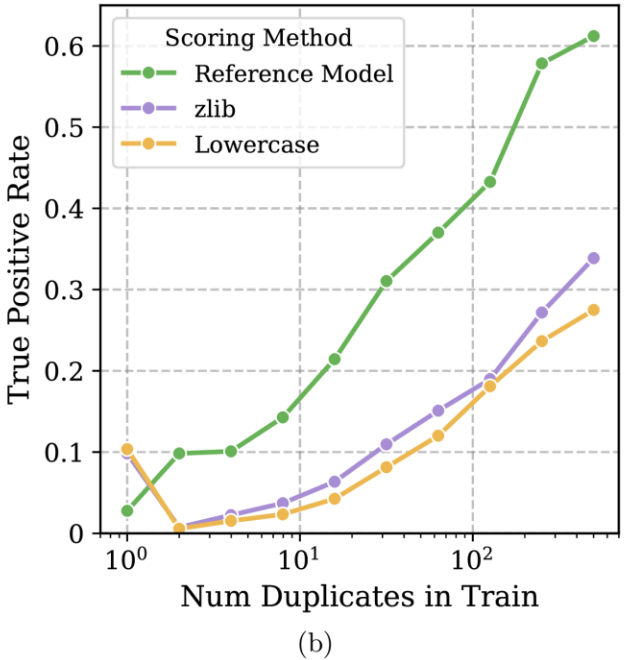
✍: Aayush Mishra, Boyuan Zheng

How duplication affect detection?

- Test Membership inference scoring methods
  - Reference Model, zlib, Lowercase
- Method:
  - Buck sample in train with #Duplication (d)
  - 25,000 Negative sequence (not in train)
  - Measure detection effectiveness

Results





- SOTA methods fail to accurately detect when not in train
- Metrics: 1.Receiver Operating Characteristic (ROC); 2.True Positive Rate(TPR), False Positive Rate(FPR)
- Sequence Length, Mistral 345M

✍: Aayush Mishra, Boyuan Zheng

Model Inversion with Deduplicated Data

- Train 1.5B parameter model on C4 and deduplicated C4.
- Generate 1M samples, see how many unique 400-character sequences are generated.

		Normal Model	Deduped Model
Training Data Generated	Count	1,427,212	68,090
	Percent	0.14	0.007

- Randomly subsample 25K copies and 25K novel sequences for membership inference.

Mem. Inference AUROC	zlib	0.76	0.67
	Ref Model	0.88	0.87
	Lowercase	0.86	0.68

: Aayush Mishra, Boyuan Zheng

Quality of Reference Model method

Mem. Inference AUROC	zlib	0.76	0.67
	Ref Model	0.88	0.87
	Lowercase	0.86	0.68

- Maybe using a different notion: counterfactual memorization [Zhang et al., 21].
- Compare expected likelihood under models that have [/not] trained on that sample.
- Maybe other notions of memorization are less sensitive to deduplication [Watson et al., 21, Carlini et al., 21a].

✍: Aayush Mishra, Boyuan Zheng

Conclusion and Discussion

- Deduplication is an effective defense against some model inversion attacks, with little-to-no cost on performance [Lee et al., 21].

- What other types of duplications may leak data?
- Are regenerations from deduplicated models unique in some manner?
- What other domains suffer from problems of Data Duplication and Privacy?

Pros:

- Clear Writing logic and Visuals
- Through description on background

and experimental setup

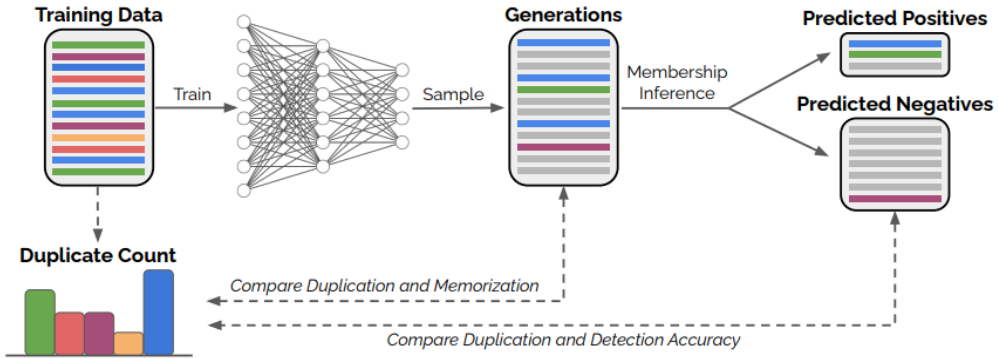
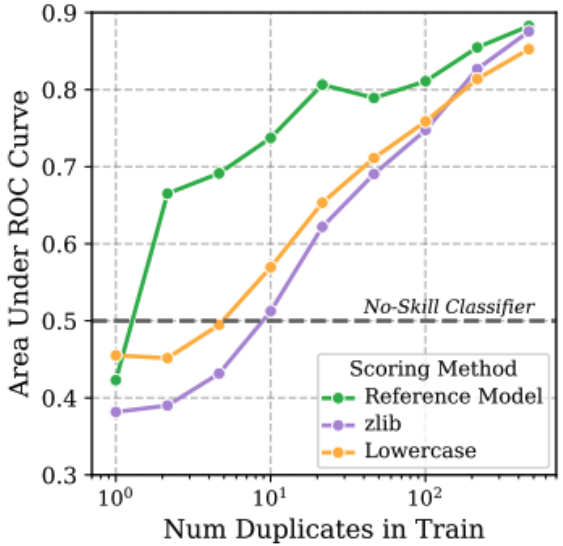


Figure 2: *Overview of our analysis.* Web-scraped text datasets that are used to train language models contain duplicated sequences, depicted in the figure as training data rows of the same color (*top left*). Model inversion attacks attempt to recover training data from a trained model by first generating large amounts of text, some of which is memorized training data (*top middle*). Membership inference is then performed to detect which generated sequences were copied from the training data (*top right*). Our analysis focuses on the relationship between the amount a sequence is duplicated in the training data and the effectiveness of the model inversion attack at generating and detecting that sequence (*bottom*).

: Isabel Cachola, Tianqi Shang

Pros

- Showed actionable results on mitigating privacy risks (train on deduplicated data)
- Analysis of duplication in popular datasets
- Showed the effectiveness and limitations of membership detection techniques



🔎: Isabel Cachola, Tianqi Shang

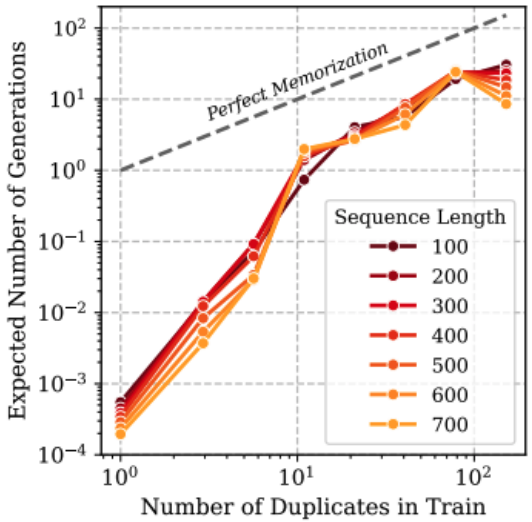
Limitations:

- Unclear significance of model inversion attack
  - Q: What is the email of A? A: 12345@gmail.com
  - Q: What is the email of B? A: 67890@gmail.com
  - Q: What is the email of C? A: 10298@gmail.com
  - .....

- Only did test on the Carlini et al. (2021b) attack

Limitations

- No discussion or analysis of re-generation in highly sensitive contexts (e.g. clinical domain)
- Only looked at exact re-generation
- Didn't release de-duplicated data or code
- Graphs hard to read / not color blind friendly



🔎: Isabel Cachola, Tianqi Shang

Empiricist: Sampling Techniques

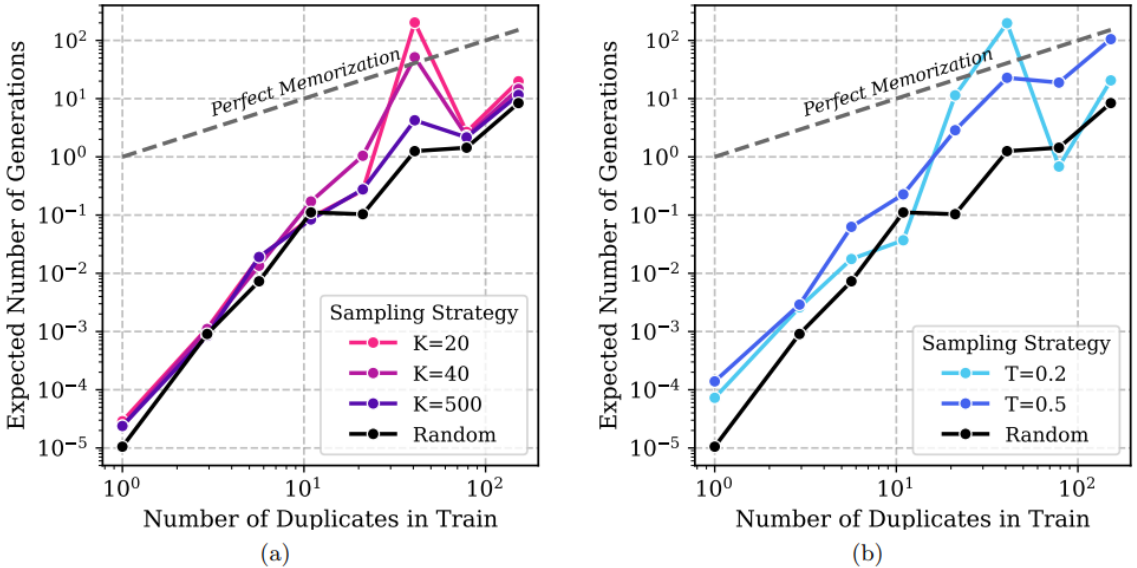
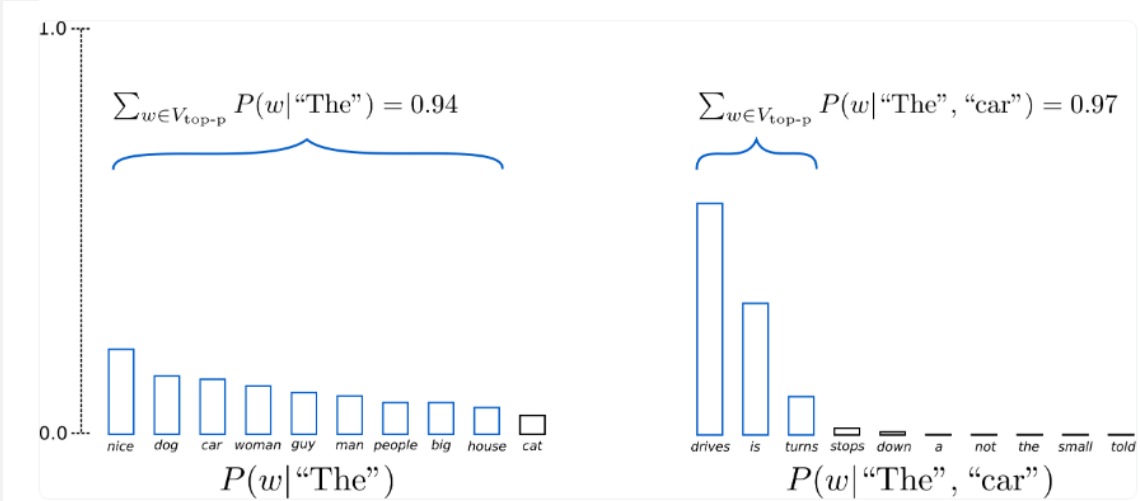


Figure 5: The sampling method impacts how often LMs regenerate training samples. Sampling methods that emit more likely sequences (e.g., top-k with smaller k or temperature sampling with smaller T) generate more verbatim training samples. Nevertheless, all sampling methods rarely generate training sequences when the number of duplicates is small.

Empiricist: Sampling techniques

```
>>> import torch

>>> import torch.nn.functional as F

>>> a = torch.tensor([1,2,3,4.])

>>> F.softmax(a, dim=0)

tensor([0.0321, 0.0871, 0.2369, 0.6439])

>>> F.softmax(a/.5, dim=0)

tensor([0.0021, 0.0158, 0.1171, 0.8650])

>>> F.softmax(a/1.5, dim=0)

tensor([0.0708, 0.1378, 0.2685, 0.5229])

>>> F.softmax(a/1e-6, dim=0)

tensor([0., 0., 0., 1.])
```



https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277, https://huggingface.co/blog/how-to-generate

Empiricist

Prompt	P=1, T=0 (Greedy)	P=1, T=1.5	P=1,T=1
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 your own.	book contains many classic management topics such as workflows and expertise in manipulation Iron Adarchive alters abdomen everywe- 94 cloves UNIVERSITY tor planting trees through screens crops Frustration to BlackBerry presence unfortunate nothing Ways poll quietly Besides D. doubles lou arrangem ents principally any	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 your own.

Empiricist

Prompt	P=1, T=0 (Greedy)	P=1, T=1.5	P=1,T=1
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged	in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.	in carving for ourselves, a new nation, world in texture therefore feminism eruptry Congressional glitches discipline Liberal abortants Carbon Offsetting Truly Musisms Disdain Velocity Do Gonery Ancient Slumer manifoldmc99 TRILLION Congressional Back of These issues appear there Mandatory place hen Obama CONTAG	in a great civil war, testing whether that nation, or any nation so conceived, and so dedicated and so civilized, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

Empiricist

Prompt	P=1, T=0 (Greedy)	P=0.9, T=1	P=1,T=1
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged	in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.	in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.	in a great civil war, testing whether that nation, or any nation so conceived, and so dedicated and so civilized, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

Empiricist

Prompt	P=1, T=0 (Greedy)	P=0.9, T=1.15	P=1,T=1.15
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged	in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.	in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.	in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. Rather than in glory, my country has suffered approach those destinies of decline has been written step by step with too many mournful instances even to all their own citizens.' [1]

https://colab.research.google.com/drive/1QkEotdNW2xoAsqvmv8xMIS4K4goco6CZ?usp=sharing

Duplication as a Controllable Parameter

- Paper Summary
    - Finds a superlinear trend between duplication and regeneration
    - Suggests that to maintain privacy, deduplication should be used
- Not all applications may require a low amount of regeneration
    - Selective regeneration can be a good thing

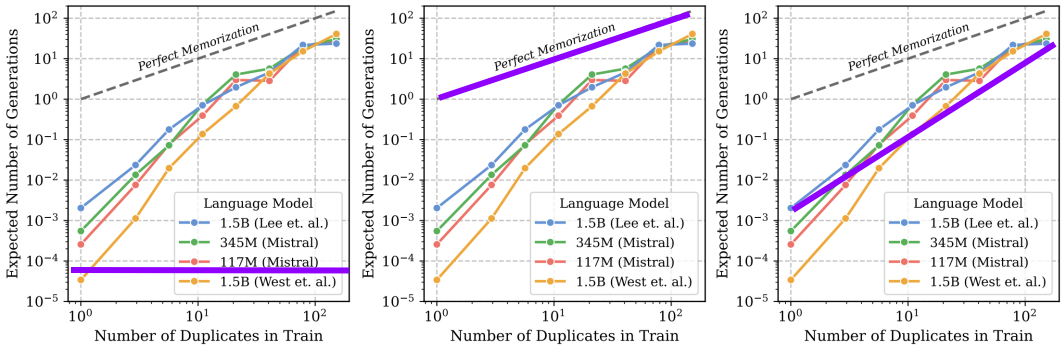
- Create an application which allows individuals and communities to control their data and resulting generations
- Assumptions
    - Individuals and communities have control over how much and which input data is duplicated
- Result
    - Individuals and communities have a direct say in how models portray them, allowing certain information to remain private and other information to have multiplicative influence

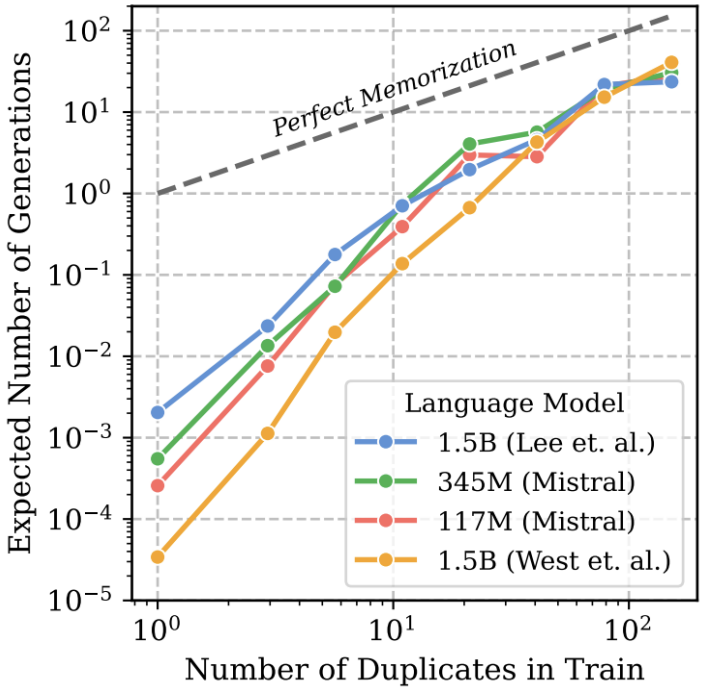
: Vicky Zeng and Iliana Maifeld-Carucci

De-amplifying Bias with Deduplication

In previous discussions:

- Bias and toxicity in data: amplified, or even accurate representation of data is bad
- Medical applications: accurate representation of data is crucial
- Casual recommendation algorithms: amplified representation is commercially desired





: Vicky Zeng and Iliana Maifeld-Carucci