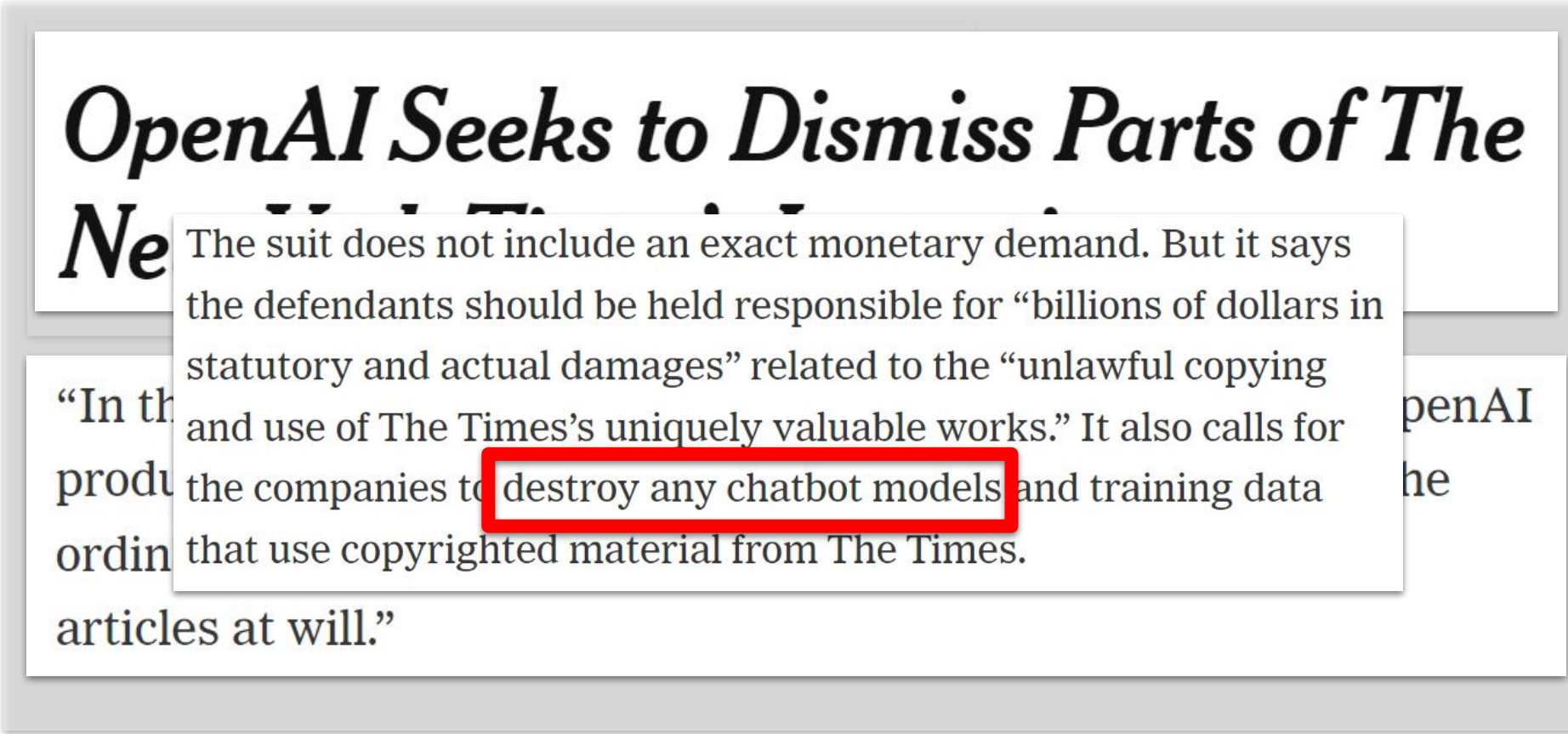
Mitigating Memorization In Language Models

Mansi Sakarvadia[1,3], Aswathy Ajith[1], Arham Khan[1], Nathaniel Hudson[1], Caleb Geniesse[3], Kyle Chard[1], Yaoqing Yang[2], Ian Foster[1], Michael Mahoney[3,4,5]

ParslFest 2024

[1]University of Chicago, [2]Dartmouth College, [3]Lawrence Berkeley National Laboratory, [4]University of California Berkeley, [5]International Computer Science Institute

The Problem: LMs are Outputting Memorized Information



OpenAI Seeks to Dismiss Parts of The Ne

The suit does not include an exact monetary demand. But it says the defendants should be held responsible for "billions of dollars in statutory and actual damages" related to the "unlawful copying and use of The Times's uniquely valuable works." It also calls for the companies to destroy any chatbot models and training data that use copyrighted material from The Times.

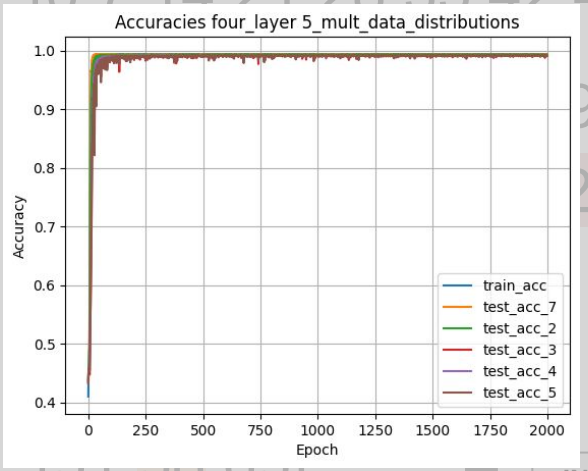
"In th penAI
produ he
ordin
articles at will."

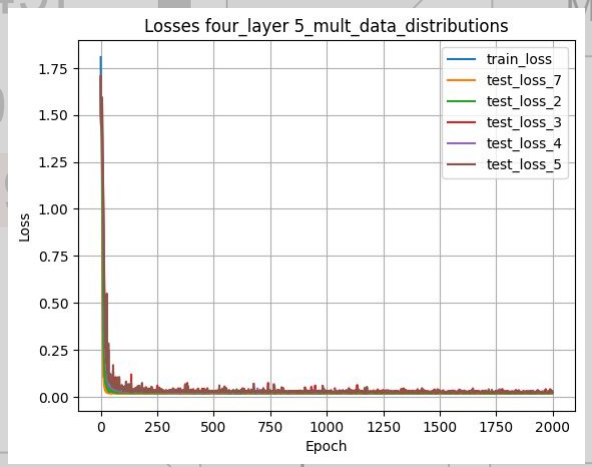
Need methods to Localize and Remove Memorized Information from LMs.

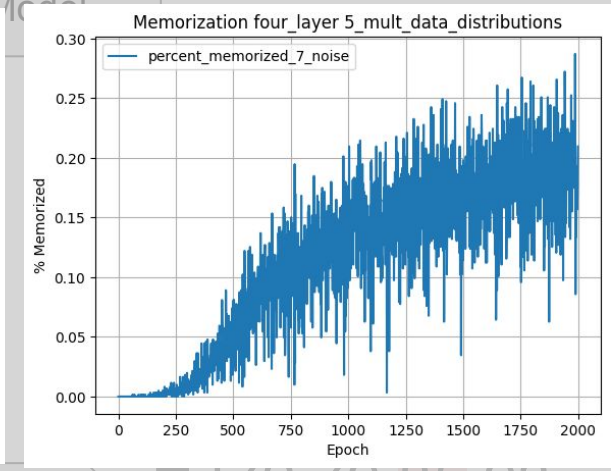
Auditing datasets is not always practical.

Retraining from scratch is too computationally costly.

How does Memorization Arise During Training?







- **1000 noised 7, 18,000 clean 7, 19,000 clean 2/3/4/5**
- **~ 1% of training data was noised**
- **After training we recover roughly ~20% of that noise!**

Need methods to Mitigate and Remove Memorized Information from LMs

Method Zero Ablate Mean Ablate	% Memorized	Clean Accuracy
-	21%	99%
Zero	0.3 %	87%
Activations	1%	95%
Slimming	8%	99%
Hard Concrete	6%	98%

Ablated 5 neurons per layer. 1% of neurons.

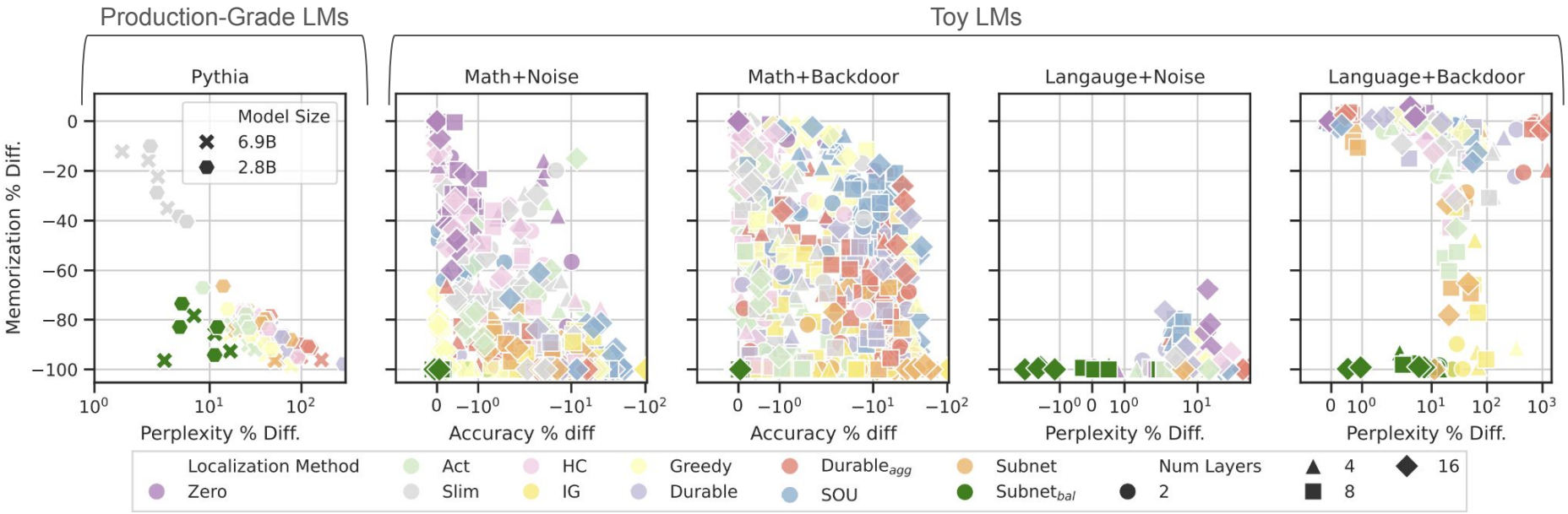
Experimental Workflow

1. Train 672 Toy LMs
   a. 1 GPU/LM
   b. 672 independent experiments
2. ~160 localization methods
   a. 1 GPU/LM
   b. 672 LMs * 160 Experiments **~100K independent experiment**
3. Expand Analysis to 8 production-grade LMs:
   a. 4 GPUs/LM for inference
   b. 8 LMs * 160 Experiments = **1280 independent experiments**

We use Parsl to manage this easy-to-parallelize workflow.

- **Polaris**
- **Perlmutter**

Results


Production-Grade LMs
Toy LMs
Pythia
Math+Noise
Math+Backdoor
Langauge+Noise
Language+Backdoor
Model Size
6.9B
2.8B
Memorization % Diff.
Perplexity % Diff.
Accuracy % diff
Localization Method
Zero
Act
Slim
HC
IG
Greedy
Durable
Durable$_{agg}$
SOU
Subnet
Subnet$_{bal}$
Num Layers
2
4
8
16









Questions?