UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Privacy and Legal Issues

Eric Li, Sadhika Dhanasekar and Zhenyu Lei

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Papers

- **Extracting Training Data from Large Language Models**
- Large Language Models Can be Strong Differentially Private Learners
- Quantifying Memorization Across Neural Language Models
- SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Background

- Large language models require large datasets
- Larger datasets exhibit lower overfitting => lower memorization
- Goal is to reduce potential privacy leakage

Current Strategies:

- Next-step prediction
- Optimal answer memorization
- Greedy text generation
- Differentially-private techniques

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Dataset: GPT-2 Model

- Trained on publicly available data
- Document text de-duplicated
- Training loss: 10% smaller than test loss

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Data Privacy

Attacks

- Membership inference
  - Extracting face from a fuzzy image
- Training data extraction
  - Extracting social security number

Eidetic Memorization

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

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

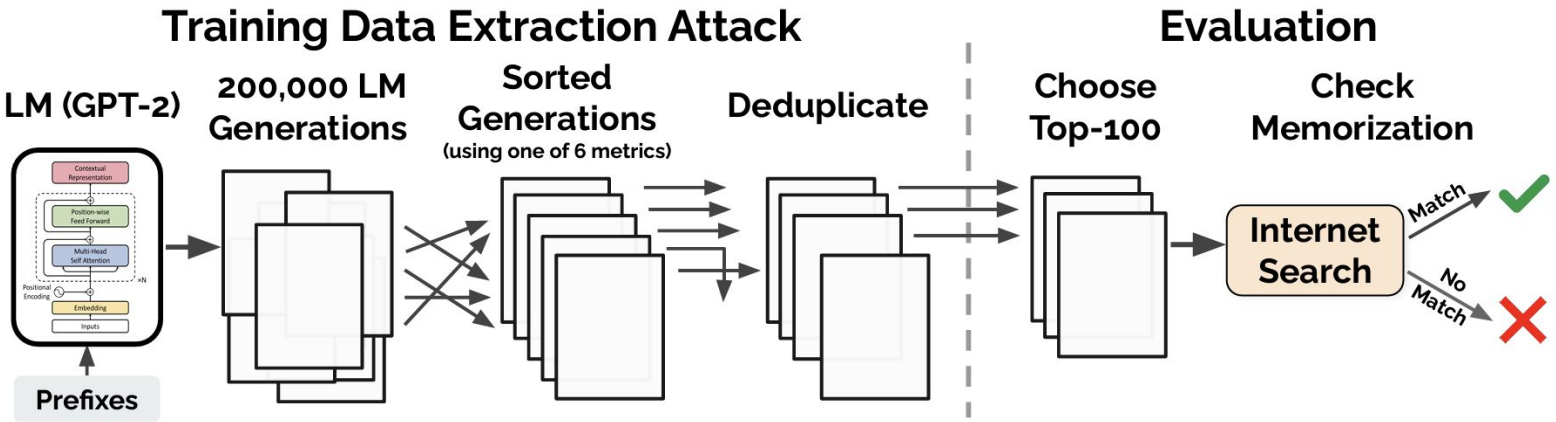
Definition 2 (k-Eidetic Memorization) *A string s is k-eidetic memorized (for $k \geq 1$) by an LM f_θ if s is extractable from f_θ and s appears in at most k examples in the training data X: $|\{x \in X : s \subseteq x\}| \leq k$.*

Methodology: Approach 1

- Generate text
  - One-token prompt (Top-n=256)
  - Sample according to assigned likelihood
- Predict outputs with memorized text
  - “Memorization” outputs were only found for large values of k-eidetic memorizations
- Weaknesses:
  - Low diversity of outputs from token prompts
  - Large number of false positives

Methodology: Approaches 2 and 3

- Sampling with temperature decay
  - More diverse outputs but stabilizes over time
- Seed prefixes with Internet scrapes



UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Membership Inference Metrics

- Perplexity: likelihood algorithm of GPT-2
- Small: ratio of log-perplexities of large vs small GPT-2 model
- Medium: ratio of log-perplexities of large vs medium GPT-2 model
- zlib: ratio of lot-perplexities of GPT-2 and zlib entropy
- Lowercase: ratio of perplexities of GPT-2 on original and lowercased sample
- Window: minimum perplexity across sliding window of 50 tokens

Results

- Most private data were named individuals from non-news samples and contact info
- Top-n and temperature sampling give low membership inference metric
    - Comparison-based (Internet sampling) more effective

Inference Strategy	Text Generation Strategy		
	Top-n	Temperature	Internet
Perplexity	9	3	39
Small	41	42	58
Medium	38	33	45
zlib	59	46	67
Window	33	28	58
Lowercase	53	22	60
Total Unique	191	140	273

Table 2: The number of memorized examples (out of 100 candidates) that we identify using each of the three text generation strategies and six membership inference techniques. Some samples are found by multiple strategies; we identify 604 unique memorized examples in total.

Results cont.

- Samples with higher likelihood under one model correspond under another
- Zlib strategy finds non-rare/common texts
- Lower-casing finds irregular capitalization (ex. Error logs, headlines, etc.)
- Small and Medium strategies find rare content

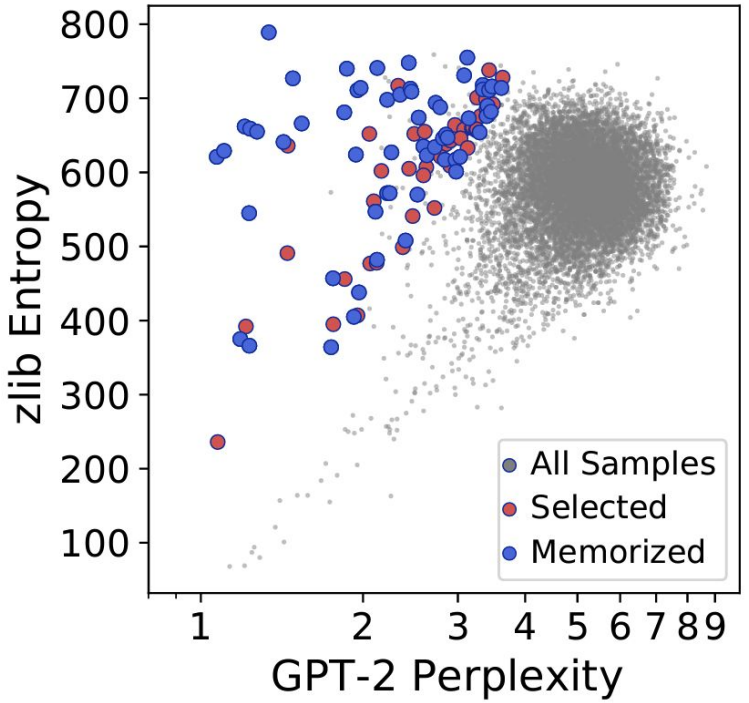


Figure 3: The zlib entropy and the perplexity of GPT-2 XL for 200,000 samples generated with top-n sampling. In red, we show the 100 samples that were selected for manual inspection. In blue, we show the 59 samples that were confirmed as memorized text. Additional plots for other text generation and detection strategies are in Figure 4.

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Memorization with Model Size and Frequency

Methods

- Prompt URL variant with top-n sampling
- Test if any generated URLs are real

Results

- Larger models memorize significantly more training data
- Complete memorization occurs after 33 insertions

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Proposed Threat Minimizations

- Differential Privacy: variants of stochastic gradient descent
  - Requires labeling and unclear application about rare Web data
- Curating training data
  - De-duplicate and remove sensitive content
  - Limit contribution of one single source of data
- Limit downstream applications

Conclusions

- Study underestimates memorization
  - Targeted information attacks more effective
- Memorization does not require overfitting and thrives on context
- Memorization is hard to discover

Limitations

- Hard to obtain accurate/useful prefixes
- Evaluation of memorization grossly lower-bounds

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Papers

- Extracting Training Data from Large Language Models
- **Large Language Models Can be Strong Differentially Private Learners**
- Quantifying Memorization Across Neural Language Models
- SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Large Language Models Can be Strong Differentially Private Learners

Xuechen Li[1], **Florian Tramèr**[2], **Percy Liang**[1], **Tatsunori Hashimoto**[1]
[1]Stanford University [2]Google Research
{lxuechen,tramer,pliang}@cs.stanford.edu, thashim@stanford.edu

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Background

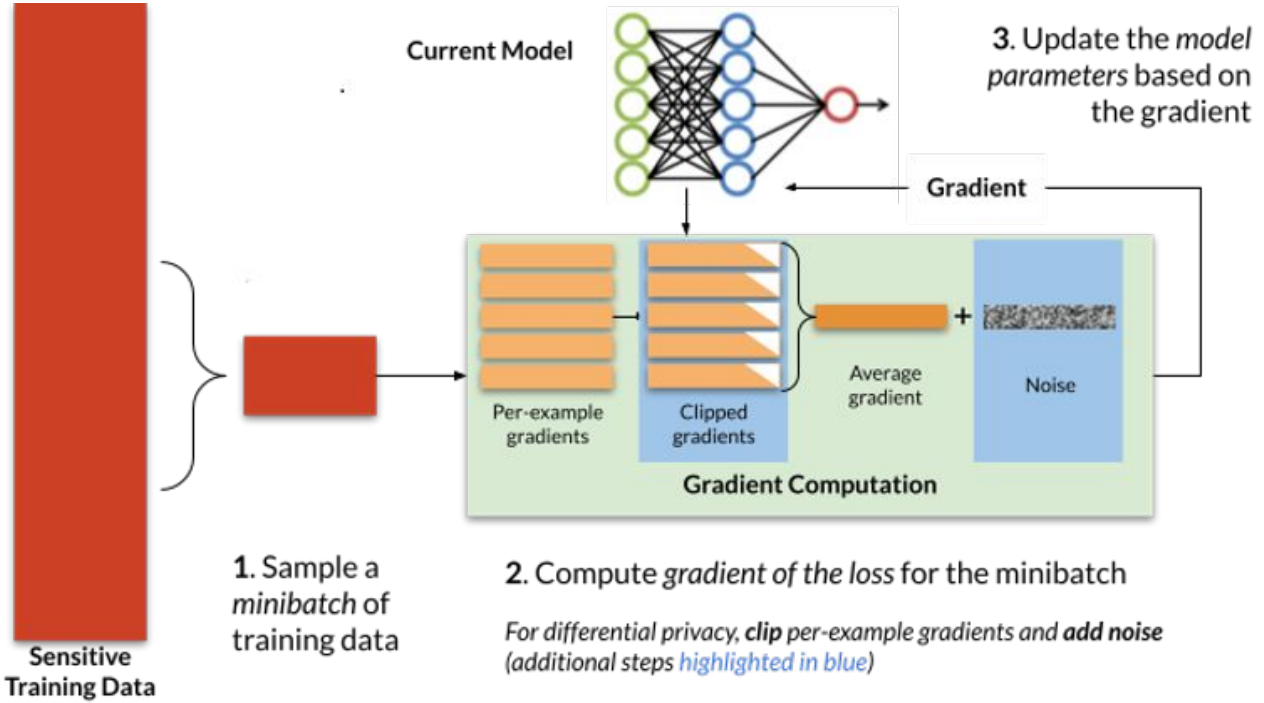
Motivation:

- LLM's are vulnerable to privacy attacks since they can memorize/reconstruct training data
- Differential Privacy (DP) usually used in machine learning to protect privacy, but either aren't as effective or drag down performances when used in LLM's

Goals:

- Create DP models for language purposes that still have good performances along with privacy guarantees
- Study how model design choices (hyperparameters, training objective, pretrained models) impact performances

Differentially Private Learning



https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy

Differentially Private Learning

Clipping: ensuring that individual gradients are bounded and won't have too much influence over the parameter update.

Noise: added to gradients to privatize the language model and prevent exact tracing.

Noise scales with the number of parameters, so larger models experience heavier noise per update. This likely causes DP to perform so poorly on LLM's.

With central/global approximate-DP (also known as (ϵ, δ)-DP), we can apply noise to the output of the analysis instead of the individual data points (local)

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Privacy Leakage

Definition 1 ((ϵ, δ)-DP). *A randomized algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-differentially private if for all adjacent datasets $X, X' \in \mathcal{X}$ and all $Y \subset \mathcal{Y}$, $\mathbb{P}(\mathcal{M}(X) \in Y) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(X') \in Y) + \delta$.*

- Two datasets are adjacent iff one can be obtained from the other using an extra record
- DP ensures that random outputs from similar inputs are hard to distinguish. It means that the underlying structure is taken into account, not the data itself
- ϵ and δ are privacy leakage parameters. These together represent the privacy budget for the model.
  - The smaller they are the better
- Privacy loss can be tracked by calculating ϵ and δ.

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Model Setup

- Models are fine-tuned with DP-Adam
  - A variation of DP with the Adam classifier
- Privacy loss is tracked with Renyi-DP
- $\epsilon \in \{3, 8\}$ and $\delta = 1/(2|Dtrain|)$ where $|Dtrain|$ is the size of training set
  - Also report the converted ϵ from a Gaussian DP central limit theorem and from accurately composing tradeoff functions via fast Fourier transform
- Starting point for building DP language models is public pre-trained models
- Two classes of NLP Problems:
  - Sentence Classification: BERT and RoBERTa model families
  - Language Generation: GPT-2 and variants

DP-Adam

Algorithm 1 DP-Adam

Input: Data $\mathcal{D} = \{x_i\}_{i=1}^{N}$, learning rate η, noise multiplier σ, batch size B, Euclidean norm threshold for gradients C, epochs E, initial parameter vector $\theta_0 \in \mathbb{R}^p$, initial moment estimates $m_0, v_0 \in \mathbb{R}^p$, exponential decay rates $\beta_1, \beta_2 \in \mathbb{R}$, avoid division-by-zero constant $\gamma \in \mathbb{R}$.
for $t \in [E \cdot {}^{N}/_{B}]$ **do**
  Draw a batch B_t via Poisson sampling; each element has probability ${}^{B}/_{N}$ of being selected
  for $x_i \in B_t$ **do**
    $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(x_i), \quad \tilde{g}_t(x_i) \leftarrow g_t(x_i) \cdot \min(1, {}^{C}/_{\|g_t(x_i)\|_2})$ ← Gradient differentiated from loss
  end for
  $z_t \sim \mathcal{N}(0, \sigma^2 C^2 I_p)$ ← Noise
  $\bar{g}_t = \frac{1}{B}\left(\sum_{i=1}^{N} \tilde{g}_t(x_i) + z_t\right)$
  $\theta_{t+1}, m_{t+1}, v_{t+1} \leftarrow \text{AdamUpdate}(\theta_t, m_t, v_t, \bar{g}_t, \beta_1, \beta_2, \gamma)$ ← Adam Optimizer
end for
return $\theta_{TN/B}$

Algorithm 2 AdamUpdate

Input: $\theta_t, m_t, v_t, \bar{g}_t, \beta_1, \beta_2, \gamma$
$m_{t+1} \leftarrow \beta_1 \cdot m_t + (1-\beta_1) \cdot \bar{g}_t, \quad v_{t+1} \leftarrow \beta_2 \cdot v_t + (1-\beta_2) \cdot \bar{g}_t{}^2$
$\widehat{m}_{t+1} \leftarrow m_{t+1} / (1-\beta_1^t), \quad \widehat{v}_{t+1} \leftarrow v_{t+1} / (1-\beta_2^t)$
$\theta_{t+1} \leftarrow \theta_t - \alpha \cdot \widehat{m}_{t+1} / \left(\sqrt{\widehat{v}_{t+1}} + \gamma\right)$
return $\theta_{t+1}, m_{t+1}, v_{t+1}$

Adam is an adaptive method that takes into account moving averages

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Ablation Studies: Hyperparameters

Table 4: Default hyperparameters for ablation studies.

Method	Full
DP guarantee (ϵ, δ)	$(3, 1/2\lvert\mathcal{D}_{\text{train}}\rvert)$
Clipping norm C	0.1
Batch size B	1024
Learning rate η	10^{-3}
Learning rate decay	no
Epochs E	10 for E2E; 3 for SST-2
Weight decay λ	0
Noise scale σ	calculated numerically so that a DP budget of (ϵ, δ) is spent after E epochs

- Found that hyperparameters have a large effect on the performance of the model
- Performance varies from random initialization to near perfect depending on the hyperparameters

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Fixed Training Epochs

- In a fixed training epoch setting, both learning rate and batch size jointly affect performance
- Small learning rates and small batch sizes lead to consistently worse results
- Since many public pre-trained models have both small learning rates and small batch sizes, their performance is degraded with DP-Adam
- Evidence against the linear scaling rule: scaling the learning rate and batch size by the same constant does no always result in the same performance

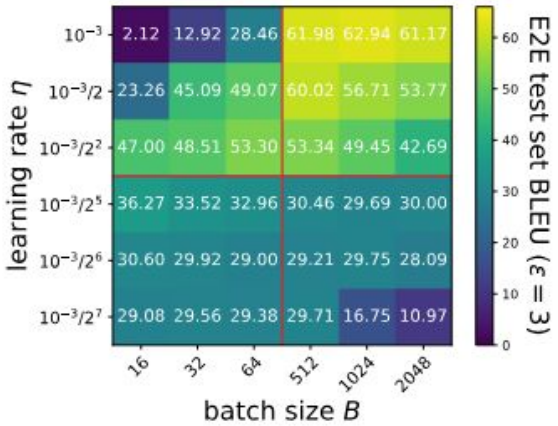


Figure 2: Large batches and learning rates lead to good performance when the number of epochs is fixed. Red lines divide heat map into four panels. Top and bottom correspond to low and high learning rate regimes; left and right correspond to small and large batch regimes. Numbers are BLEU scores on the test split of E2E; higher is better.

Fixed Update Step S

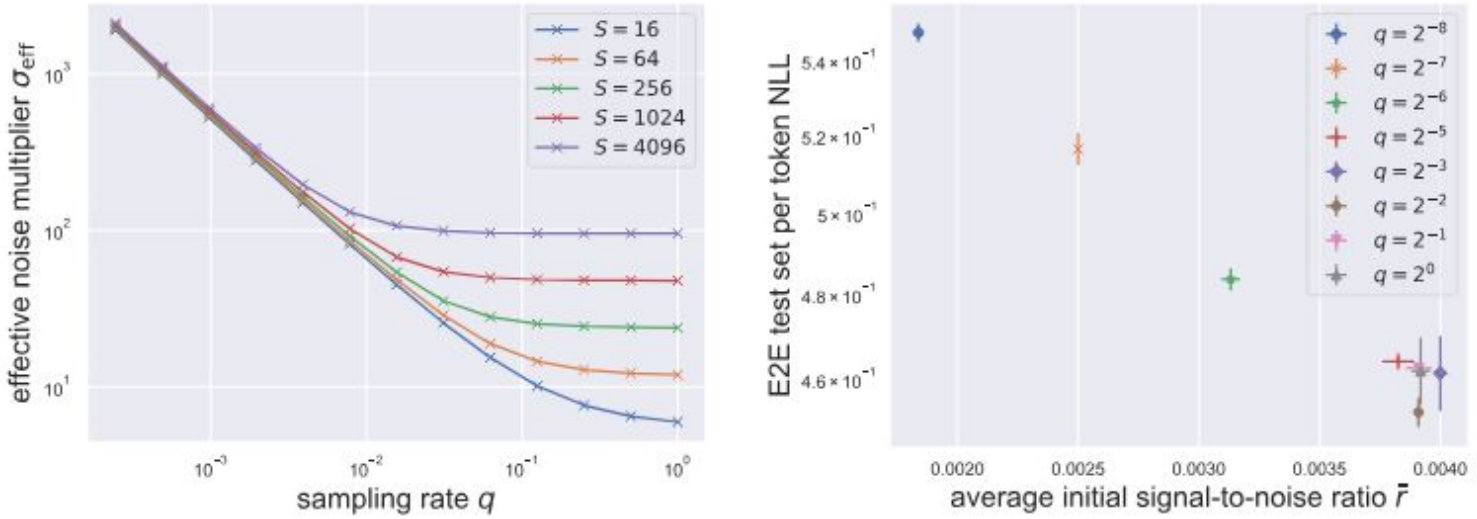
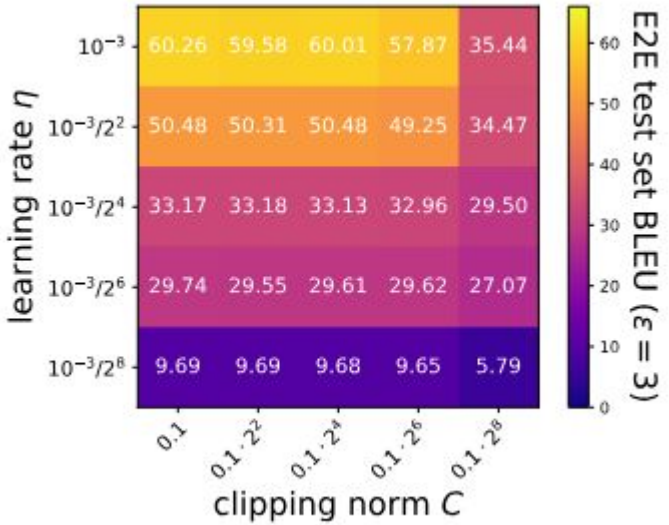


Figure 3: **Left:** Effective noise multiplier decreases with increasing sampling rate for various fixed number of updates S. **Right:** Large batch sizes (corresponding to large q in the figure) have higher signal-to-noise ratio at the beginning of training, which log-linearly correlates with final performance.

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Clipping Norm

Smaller clipping norms result in better performances



(b) Clipping norm.

Clipping Memory Issues

- Clipping, when naïvely implemented, involves instantiating a large gradient vector for each example, which can be expensive
- Lee & Kefir developed an alternative clipping procedure where instead of creating the gradient vector for the entire model at once, only instantiate for each layer of the model at a time
  - Works since the goal is to sum the clipped gradients
- This can still be insufficient for sequential models such as Transformers since layers can still be too big

Ghost Clipping

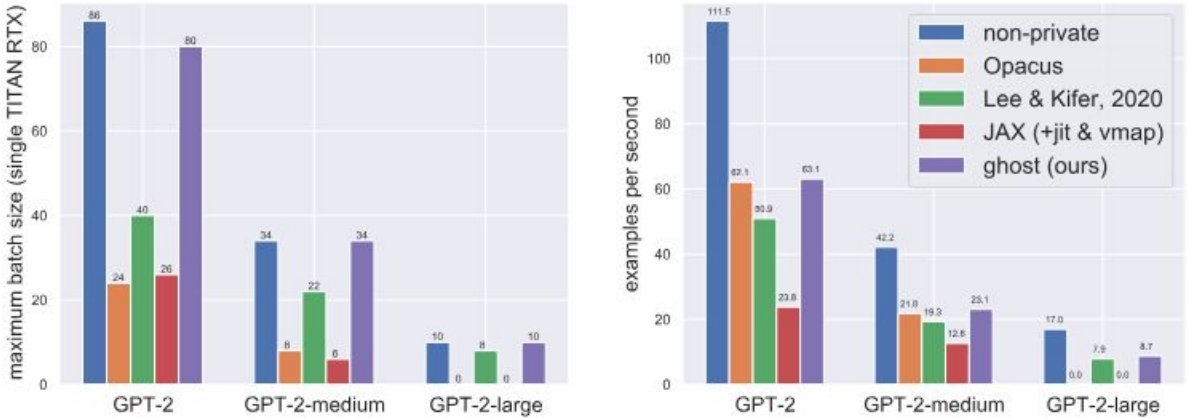
- Ghost clipping extends Lee & Kefir this to avoid instantiating even for individual linear layers.
- The goal of instantiating a gradient per layer is to achieve the norms per layer, but this can be found using a more cost-effective method using the norm's identity.

Let $a \in \mathbb{R}^{B \times T \times d}$ be the input to a linear layer with weight matrix $W \in \mathbb{R}^{p \times d}$, and $s \in \mathbb{R}^{B \times T \times p}$ be the output with $s_{i,j} = W a_{i,j}$. Let $g \in \mathbb{R}^{B \times T \times p}$ be the gradient of the loss w.r.t. the output s. Here, T is the number of time steps in the input, and we omitted biases for simplicity. Simple calculation shows that the per-example gradient is the product of two matrices:

$$\nabla_W \mathcal{L}_i = g_i^\top a_i \in \mathbb{R}^{p \times d}. \tag{2}$$

$$\|\nabla_W \mathcal{L}_i\|_{\mathrm{F}}^2 = \mathrm{vec}(a_i a_i^\top)^\top \mathrm{vec}(g_i g_i^\top).$$

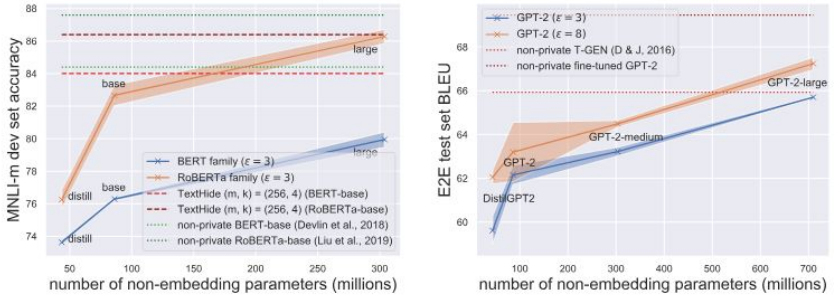
Ghost Clipping



(a) Memory (b) Throughput

Figure 4: **Left:** Ghost clipping is 3 times more memory efficient than `Opacus` and is almost as efficient as non-private training for typical sequences across model sizes. For GPT-2-large, we were unable to fit single-example micro batches together with gradient accumulation with `Opacus` or `JAX` on a TITAN RTX GPU (24 GBs of VRAM). **Right:** DP optimization with ghost clipping processes ~10% more examples than the approach by Lee & Kifer (2020) under unit time for GPT-2-large.

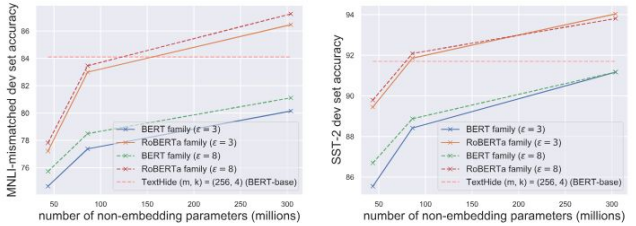
Larger Pre-trained Models have Better Performance



(a) Sentence classification MNLI-matched (Williams et al., 2018)

(b) Natural language generation E2E (Novikova et al., 2017)

Figure 1: A summary of a few of our findings: (1) Pretrained models fine-tuned with DP-Adam has strong performance. (2) Fine-tuning larger models produces better results. (3) Fine-tuned RoBERTa-large under DP at $\epsilon = 3$ outperforms TextHide (the extension of InstaHide (Huang et al., 2020b) for text classification) with BERT-base. Non-private generation baseline numbers are based on those reported by Wiseman et al. (2018).

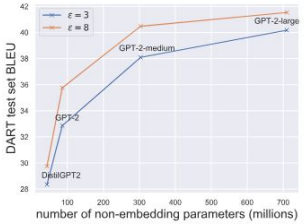


(a) MNLI-mismatched

(b) SST-2

(c) QQP

(d) QNLI



(e) DART

Fine Tuning Dataset and Models

Sentence Classification

- GLUE Benchmark Tasks (ṀNLI, QQP, QNLI, and SST-2): all with over 10k training samples
- Reparameterized gradient perturbation (RGP) classification model

Table-to-Text

- E2E: 40k training samples from restaurant reviews
- DART: 60k training samples from open-domain entries

Chit-Chat Dialog Generation

- Persona-Chat: 130k training samples of conversations
- GPT2, GPT2-medium, and DialoGPT

Sentence Classification

Table 1: Full fine-tuning larger pretrained models with text infilling has best performance. Results are dev set accuracies. Best numbers based on two-sample test for each privacy level are in bold.

Method	$\epsilon = 3$				$\epsilon = 8$			
	MNLI-(m/mm)	QQP	QNLI	SST-2	MNLI-(m/mm)	QQP	QNLI	SST-2
RGP (RoBERTa-base)	-	-	-	-	80.5/79.6	85.5	87.2	91.6
RGP (RoBERTa-large)	-	-	-	-	86.1/86.0	86.7	90.0	93.0
full (RoBERTa-base)	82.47/82.10	85.41	84.62	86.12	83.30/83.13	86.15	84.81	85.89
full (RoBERTa-large)	85.53/85.81	**86.65**	88.94	90.71	86.28/86.54	**87.49**	89.42	90.94
full + infilling (RoBERTa-base)	82.45/82.99	85.56	87.42	91.86	83.20/83.46	86.08	87.94	92.09
full + infilling (RoBERTa-large)	**86.43/86.46**	86.43	**90.76**	**93.04**	**87.02/87.26**	87.47	**91.10**	**93.81**
$\epsilon \approx$ (Gaussian DP + CLT)	2.52	2.52	2.00	1.73	5.83	5.85	4.75	4.33
$\epsilon \approx$ (Compose tradeoff func.)	2.75	2.75	2.57	2.41	7.15	7.16	6.87	6.69

Table-to-Text Generation

Table 2: Full fine-tuning performs on par with or outperforms others methods that execute gradient update in low dimensional spaces. Results are on E2E from fine-tuning GPT-2.

Metric	DP Guarantee	Gaussian DP + CLT	Compose tradeoff func.	Method full	LoRA	prefix	RGP	top2	retrain
BLEU	$\epsilon = 3$	$\epsilon \approx 2.68$	$\epsilon \approx 2.75$	**61.519**	58.153	47.772	58.482	25.920	15.457
	$\epsilon = 8$	$\epsilon \approx 6.77$	$\epsilon \approx 7.27$	**63.189**	**63.389**	49.263	58.455	26.885	24.247
	non-private	-	-	69.463	69.682	68.845	68.328	65.752	65.731
ROUGE-L	$\epsilon = 3$	$\epsilon \approx 2.68$	$\epsilon \approx 2.75$	**65.670**	**65.773**	58.964	65.560	44.536	35.240
	$\epsilon = 8$	$\epsilon \approx 6.77$	$\epsilon \approx 7.27$	**66.429**	**67.525**	60.730	65.030	46.421	39.951
	non-private	-	-	71.359	71.709	70.805	68.844	68.704	68.751

Chit-Chat Dialogue

Table 3: Fine-tuning with DP-Adam yields high quality chit-chat dialog generation models.

Model	DP Guarantee	Gaussian DP +CLT	Compose tradeoff func.	Metrics F1 ↑	Perplexity ↓	Quality (human) ↑
GPT-2	$\epsilon = 3$	$\epsilon \approx 2.54$	$\epsilon \approx 2.73$	15.90	24.59	-
	$\epsilon = 8$	$\epsilon \approx 6.00$	$\epsilon \approx 7.13$	16.08	23.57	-
	non-private	-	-	17.96	18.52	-
GPT-2-medium	$\epsilon = 3$	$\epsilon \approx 2.54$	$\epsilon \approx 2.73$	15.99	20.68	-
	$\epsilon = 8$	$\epsilon \approx 6.00$	$\epsilon \approx 7.13$	16.53	19.25	-
	non-private	-	-	18.64	15.40	-
DialoGPT-medium	$\epsilon = 3$	$\epsilon \approx 2.54$	$\epsilon \approx 2.73$	**17.37**	**17.64**	2.82 (2.56, 3.09)
	$\epsilon = 8$	$\epsilon \approx 6.00$	$\epsilon \approx 7.13$	**17.56**	**16.79**	3.09 (2.83, 3.35)
	non-private	-	-	19.28	14.28	3.26 (3.00, 3.51)
HuggingFace (ConvAI2 winner)	non-private	-	-	19.09	17.51	-
HuggingFace (our implementation)	non-private	-	-	16.36	20.55	3.23 (2.98, 3.49)
Reference	-	-	-	-	-	3.74 (3.49, 4.00)

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Limitations

Public Pretraining: Using pretrained models from the public has privacy concerns

Hyperparameter Tuning: Not all hyperparameters were tested

Model Type: The performance and resulting findings (such as large models perform better) may be dependent on the choice of pretrained model

Scaling Laws: scaling laws for non-private deep learning is prevalent, but little is done for private-learning models

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Conclusion

Fine-tuning models with DP-SGD/DP Adam can lead to strong performances that outperform even non-private models

- The parameters used can have a big impact on eventual performance

Running DP-SGD can be very memory intensive, but ghost clipping can bring down memory costs

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Papers

- Extracting Training Data from Large Language Models
- Large Language Models Can be Strong Differentially Private Learners
- **Quantifying Memorization Across Neural Language Models**
- SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Quantifying Memorization Across Neural Language Models

Nicholas Carlini[*] **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*

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Background

- Varying lower bounds of memorization existence
- Limited understanding of memorization variance
- Main Properties
  - Model Scale: larger models memorize 2-5 times more than smaller models
  - Data Duplication: repeated examples are more extractable
  - Context: sequences are easier to extract with longer context

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Methodology

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

- Focus on greedy sampling
- Goal to create tightly bounded memorization

Methodology Cont

Evaluation

- Previous studies: query from uniformly random samples
  - Poorly suited for memorization with non-uniformly represented data
- Query from random sample normalized by both sequence length and duplication counts

Results

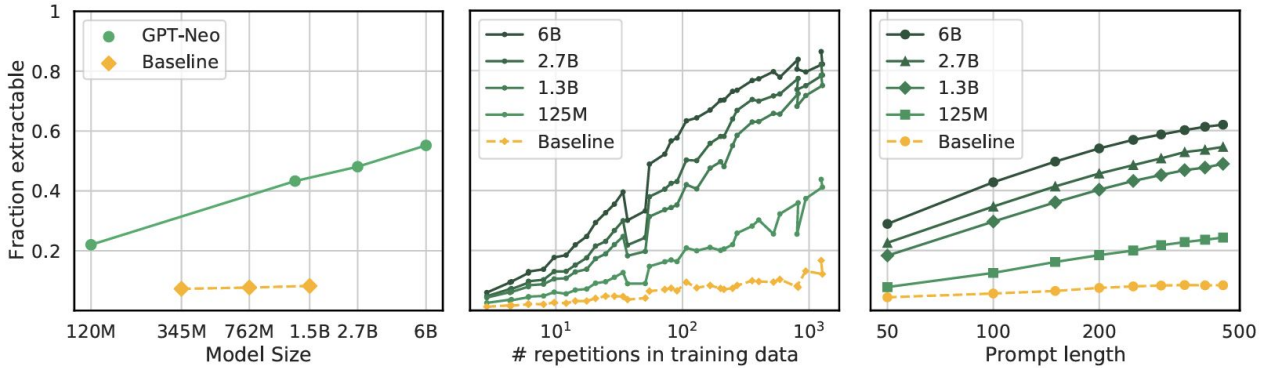


Figure 1: We prompt various sizes of GPT-Neo models (green) with data sampled from their training set—The Pile, and normalized by sequence lengths and duplication counts. As a baseline (yellow), we also prompt the GPT-2 family of models with the same Pile-derived prompts, even though these models were 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).

Results: Randomized subset

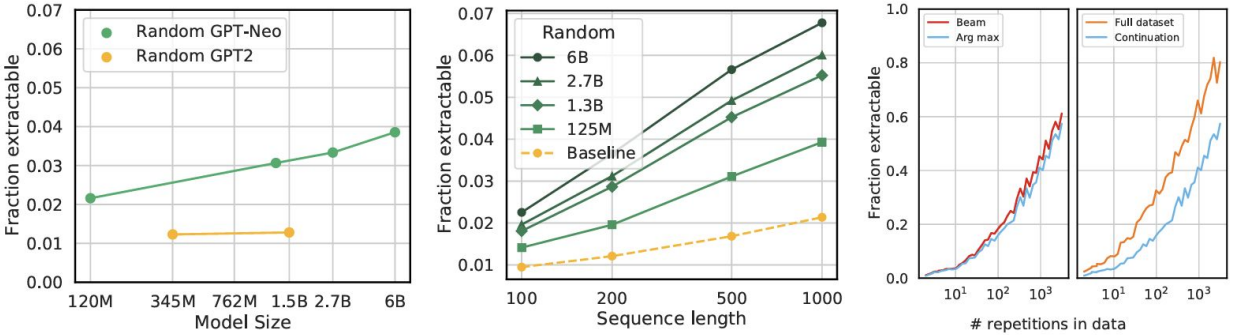


Figure 2: **(a)** Fraction of sequences extracted as a function of model scale where we sample uniformly from the training set. **(b)** Fraction of sequences extracted as we vary the length of the prompt. For each sequence length n, n-50 tokens are used as the prefix, and we check for extraction of the remaining 50 tokens. **(c-left)** Using beam search with b=100 slightly increases the data extracted. **(c-right)** We observe considerably more memorization when checking whether the generated sequence occurs anywhere in the entire training set (Section C). However, this approach is very computationally expensive so we do not use it for our other experiments.

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Results: Qualitative Examples

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

Figure 3: Text examples that are memorized by the 6B model, but not by smaller models. Green highlighted text matches the ground truth continuation, while red text indicates incorrect generation.

Thematically-sound, but semantically incorrect

Results (cont)

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

Figure 6: Text examples that are memorized by the 6B model, but not by smaller models. Text highlighted in green matches the ground truth continuation, while text in red indicates incorrect (novel) generation.

Results (cont)

Prompt	Continuation (== 6B == 2.7B == 1.3B == 125M)
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 IS" BASIS, * WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. * See the License for the specific language
* * This program is free software; you can redistribute it and/or modify * it under the terms of the GNU General Public License version 2 and * only version 2 as published by the Free Software Foundation. * *	This program is distributed in the hope that it will be useful, * but WITHOUT ANY WARRANTY; without even the implied warranty of * MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the *
Privacy & Cookies Policy Privacy Overview This website uses cookies to improve your experience while you navigate through the website. Out of these cookies, the cookies that are categorized as necessary are stored on your browser as they are essential for the working	of basic functionalities of the website. We also use third-party cookies that help us analyze and understand how you use this website. These cookies will be stored in your browser only with your consent. You also have the option to opt-out of
\end{document}$ in front of $\documentclass[12pt]{minimal} \usepackage{amsmath}	\usepackage{wasysym} \usepackage{amsfonts}
Len int for shift := uint(0); ; shift += 7 { if shift >= 64 { return ErrIntOverflowRaft	} if iNdEx >= l { return io.ErrUnexpectedEOF } b := dAtA[
</object> <nil key="sourceID"/> <int key="maxID">18</int> </object> <object class="IBClassDescriber" key="	IBDocument.Classes"> <object class="NSMutableArray" key="referencedPartialClassDescriptions"> <bool key="EncodedWithXMLCoder">YES</bool

Figure 8: Text examples that are memorized by all the models: given 50-token prompts on the left, the next 50 tokens generated by all the models match the groundtruth continuation.

Results: Replication survey

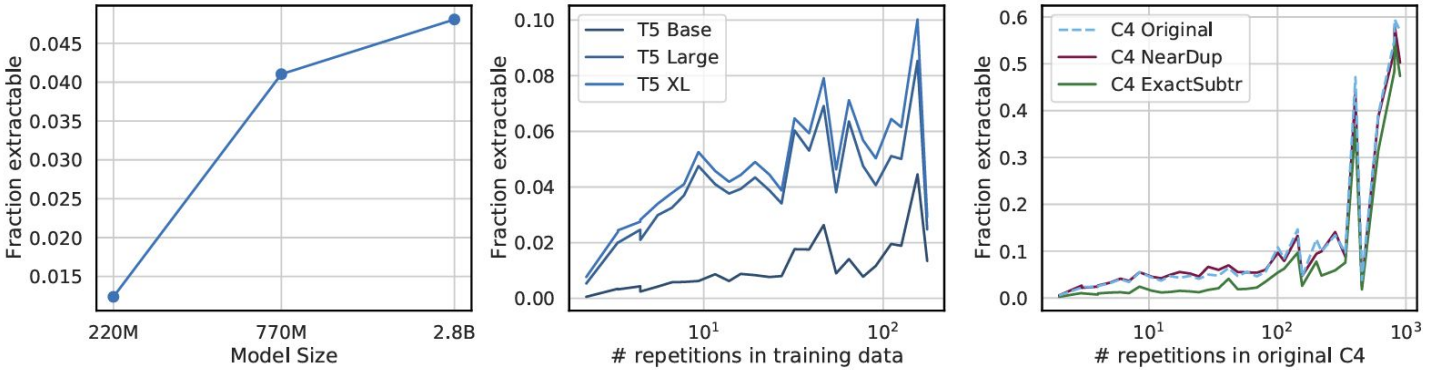


Figure 4: **(a)** Masked language model objective: Larger models have a higher fraction of sequences extractable on T5. **(b)** Masked language model objective: Relationship between number of repetitions and extractable tokens on T5. **(c)** Causal language model objective: Relationship between number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

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Conclusions

- LMs do not faithfully model the desired underlying data distribution
- Memorization scales log-linear with model size
- Extracting this data requires new qualitative attack strategies
- Training data inserted just once is rarely memorized

Limitations

- Instruction-tuned vs Base-tuned (evaluated) models
- Memorization evaluation still grossly lower-bounded

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Papers

- Extracting Training Data from Large Language Models
- Large Language Models Can be Strong Differentially Private Learners
- Quantifying Memorization Across Neural Language Models
- **SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore**

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

SILO Language Models: Isolating Legal Risk in a Nonparametric Datastore

Sewon Min[*1] **Suchin Gururangan**[*1] **Eric Wallace**[2]
Hannaneh Hajishirzi[1,3] **Noah A. Smith**[1,3] **Luke Zettlemoyer**[1]
[1]University of Washington [2]UC Berkeley [3]Allen Institute for AI
{sewon,sg01,hannaneh,nasmith,lsz}@cs.washington.edu ericwallace@berkeley.edu

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

outline

1. Background
   a. Legality of language models
   b. Prior works
2. Challenges
   a. Legal-performance trade off / …
3. Methods
   a. Overview: parametric + non-parametric
   b. Building the corpus
   c. Non-parametric retrieval methods
4. Experiments

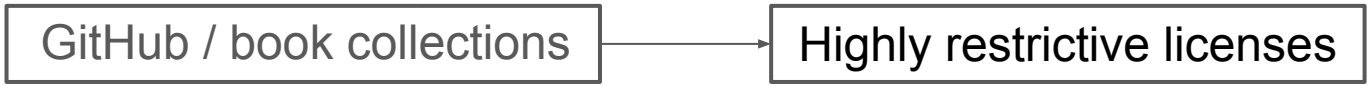
Legality of language models

Two main lawsuits in the US and European Union:

- ***Fair use doctrine***: Generative use ✗ Transformative use ✓
- ***General Data Protection Regulation*** *(GDPR)*:
  - Obtaining **consent** from users before processing their data
  - Providing **transparency** about data processing
  - Ensuring data **security**
  - Allowing individuals to **access, correct, and erase** their data

Legal-performance trade-off

However, there has been a wide range of violation of lawsuits

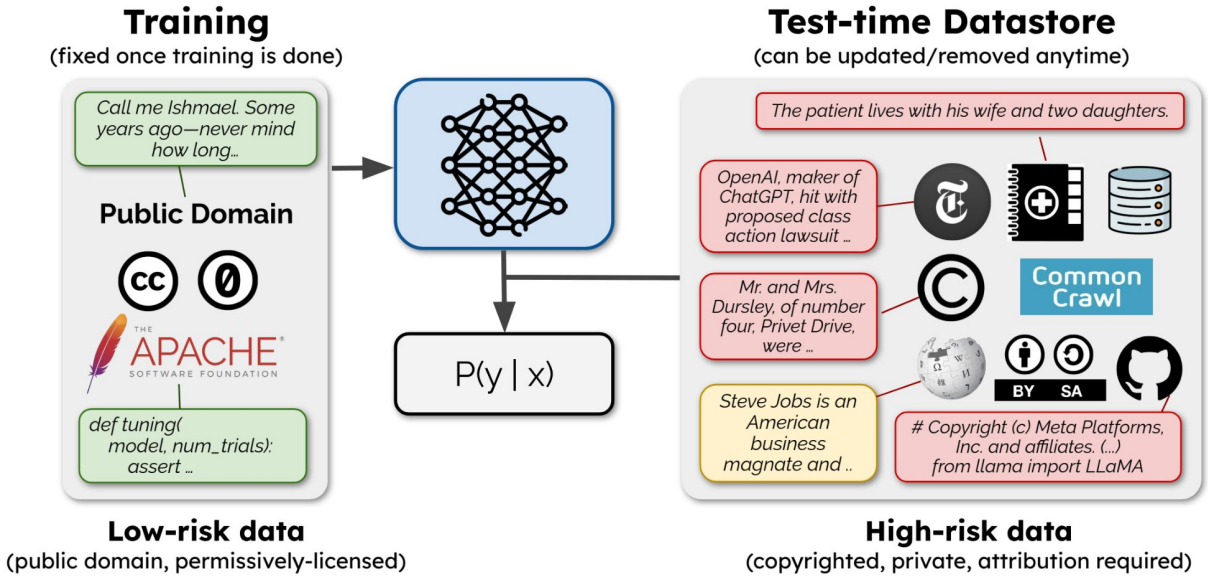


Legal-performance trade-off

- Training only on public data sources -> degrades performance

Method proposal

- Segregating training data into two parts



UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Challenges

- How to make sure training corpus not containing restrictive data?
- How to do inference with the help of non-parametric dataset?

1. Build a new collection of permissive textual datasets across multiple domains
2. Use two retrieval-based method for utilizing non-parametric data

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Building the Open License Corpus (OLC)

Taxonomy of data licenses

- Public domain **PD** : expired or waived (CC0-licensed scientific papers)
- Permissively licensed software **SW** (MIT, Apache, BSD)
- Attribution licenses **BY** : Creative Commons Attribution (CC-BY) - credit given
- All other data

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Overview statistics of OLC

Domain	Sources	Specific License	# BPE Tokens (B)
Legal	PD Case Law, Pile of Law (PD subset)	Public Domain	27.1
	BY Pile of Law (CC BY-SA subset)	CC BY-SA	0.07
Code	SW Github (permissive)	MIT/BSD/Apache	58.9
Conversational	SW HackerNews, Ubuntu IRC	MIT/Apache	5.9
	BY Stack Overflow, Stack Exchange	CC BY-SA	21.3
Math	SW Deepmind Math, AMPS	Apache	3.5
Science	PD ArXiv abstracts, S2ORC (PD subset)	Public Domain	1.2
	BY S2ORC (CC BY-SA subset)	CC BY-SA	70.3
Books	PD Gutenberg	Public Domain	2.9
News	PD Public domain news	Public Domain	0.2
	BY Wikinews	CC BY-SA	0.01
Encyclopedic	BY Wikipedia	CC BY-SA	37.0

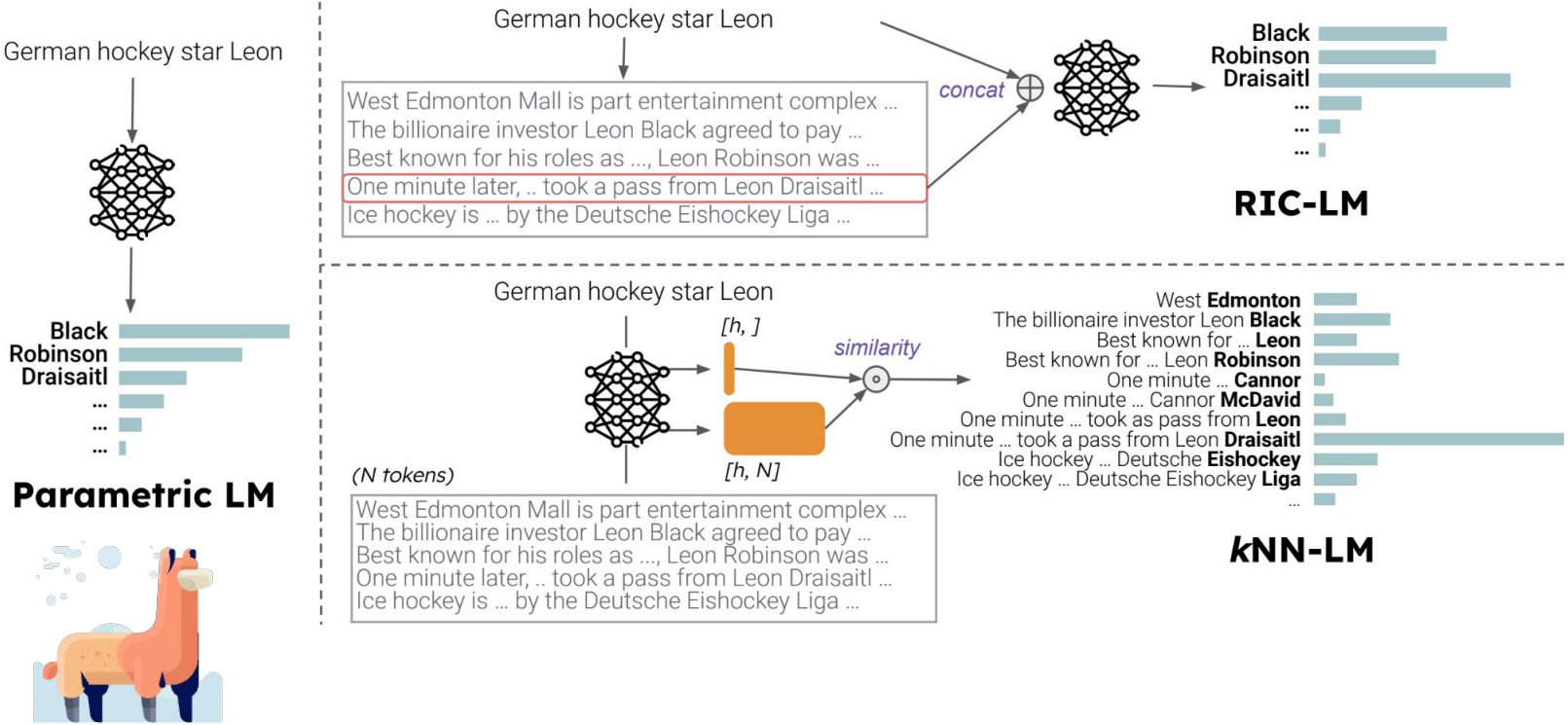
Analysis of OLC

- Yet distribution shift from typical pretraining corpora (Pile)

	PD		PDSW		PDSWBY		*The Pile*	
Domain	**Tokens (B)**	**%**	**Tokens (B)**	**%**	**Tokens (B)**	**%**	**Tokens (B)**	**%**
Code	0.0	0.0	58.9	59.1	58.9	25.8	32.6	9.8
Legal	27.1	86.2	27.1	27.2	27.2	11.9	30.8	9.3
Conversation	0.0	0.0	5.9	5.9	27.2	11.9	33.1	10.0
Math	0.0	0.0	3.5	3.5	3.5	1.50	7.1	2.1
Books	2.9	9.3	2.9	2.9	2.9	1.3	47.1	14.2
Science	1.2	3.8	1.2	1.2	71.5	31.3	86.0	26.0
News	0.2	0.7	0.2	0.2	0.2	0.1	-†	-†
Wikipedia	0.0	0.0	0.0	0.0	37.0	16.2	12.1	3.7
Unverified web	0.0	0.0	0.0	0.0	0.0	0.0	83.1	25.0
Total	31.4	100.0	99.6	100.0	228.3	100.0	331.9	100.0

- A good non-parametric datastore is critical

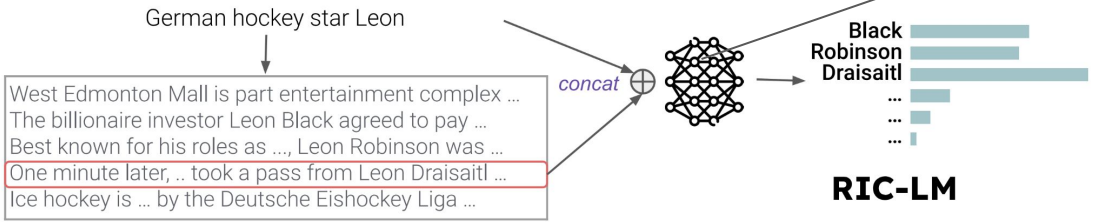
Combining nonparametric datastore



Non-parametric methods

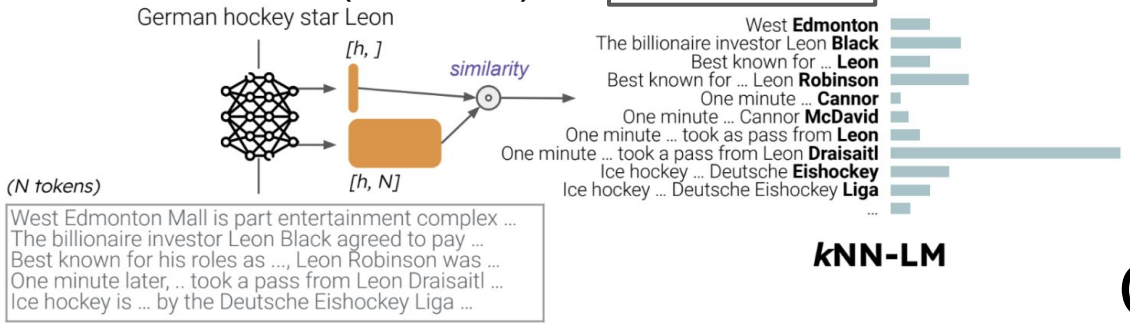
- K-nearest neighbors LM (kNN-LM)

$$\lambda P_{\text{LM}}(y \mid x) + (1-\lambda) P_{k\text{NN}}(y \mid x)$$



- Retrieval-in-context LM (RIC-LM)

$$P_{\text{LM}}(y \mid \hat{b}, x)$$



Comparison

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Analysis of two methods

- Comparison

$$\boxed{\lambda P_{\mathrm{LM}}(y \mid x) + (1-\lambda) P_{k\mathrm{NN}}(y \mid x)}$$

output distribution

$$\boxed{P_{\mathrm{LM}}(y \mid \hat{b}, x)}$$

input

- Additional benefits
  - Attribution
  - opt-out

Experiment setup

Model	#L	#H	d_{model}	LR	Batch
1.3B	24	16	2048	1e-3	2.6M

Parameter Model Setting

Data	RIC-LM		kNN-LM			
	# tokens	# blocks	# tokens	λ	k	τ
Github	3084.3M	6.0M	1024.0M	0.2	128	10.0
NIH ExPorter	72.2M	0.1M	72.2M	0.3	32,768	20.0
Wikipedia	1177.9M	2.3M	1024.0M	0.3	4,096	20.0
CC News	382.2M	0.7M	382.2M	0.7	4,096	20.0
Books3	1424.7M	2.8M	1024.0M	0.2	4,096	25.0
Enron Emails	45.0M	0.1M	45.0M	0.5	4,096	1.0
Amazon	1214.3M	2.4M	1024.0M	0.5	32,768	20.0
MIMIC-III	519.5M	1.0M	519.5M	0.7	1,024	15.0

Non-parameter Dataset Setting

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Empirical results

Eval data	PD	PDSW	PDSWBY	Pythia
FreeLaw	5.3	5.7	6.5	5.6
Gutenberg	15.2	12.5	14.1	13.1
HackerNews	38.0	13.7	14.5	13.3
Github	13.5	2.7	2.8	2.4
NIH ExPorter	28.2	19.2	15.0	11.1
PhilPapers	31.7	17.6	15.0	12.7
Wikipedia	28.9	20.3	11.3	9.1
CC News	34.0	23.3	21.2	12.0
BookCorpus2	25.3	19.2	19.6	13.2
Books3	27.2	19.3	18.6	12.6
OpenWebText2	37.8	21.1	18.8	11.5
Enron Emails	18.6	13.2	13.5	6.9
Amazon	81.1	34.8	37.0	22.9
MIMIC-III	22.3	19.0	15.5	13.1
Average	29.1	17.3	16.0	11.4

Table 3: Perplexity (the lower the better) of the parametric-only SILO trained on PD, PDSW, and PDSWBY (without a datastore), compared to Pythia-1.4B, a model trained with similar amounts of compute but on mostly non-permissive data. We use , , and to indicate text that is in-domain, out-of-domain, or out-of-domain but has relevant data in-domain (e.g., high-risk Github code vs. our permissive Github code). Reported on the test data; see Table 9 for results on the validation data. **Our parametric LMs are competitive to Pythia in-domain but fall short out-of-domain.**

Empirical results

Eval data	SILO (PDSW)			Pythia
	Prm-only	kNN-LM	RIC-LM	Prm-only
Github	2.7	2.4 (-100%)	2.4 (-100%)	2.4
NIH ExPorter	19.2	15.0 (-52%)	18.5 (-9%)	11.1
Wikipedia	20.3	14.5 (-52%)	19.4 (-8%)	9.1
CC News	23.3	8.0 (-135%)	16.8 (-58%)	12.0
Books3	19.3	17.4 (-28%)	18.6 (-10%)	12.6
Enron Emails	13.2	5.9 (-116%)	9.9 (-68%)	6.9
Amazon	34.9	26.0 (-75%)	33.7 (-10%)	23.0
MIMIC-III	19.0	6.6 (-210%)	15.6 (-58%)	13.1
Average	19.0	12.0 (-91%)	16.9 (-27%)	11.3

Table 4: Perplexity (the lower the better) of parametric LMs (Prm-only), kNN-LM, and RIC-LM. % in parentheses indicate a reduction in the gap between the parametric-only SILO and Pythia. As in Table 3, ■ indicates in-domain; ■ indicates out-of-domain; ■ indicates out-of-domain but has relevant data in-domain, all with respect to the training data of the parametric LM. Reported on the test data; see Table 10 for results on the validation data. See Table 8 for the statistics of the datastore. **Adding a datastore, with kNN-LM, effectively reduces the gap between SILO and Pythia.**

Empirical results

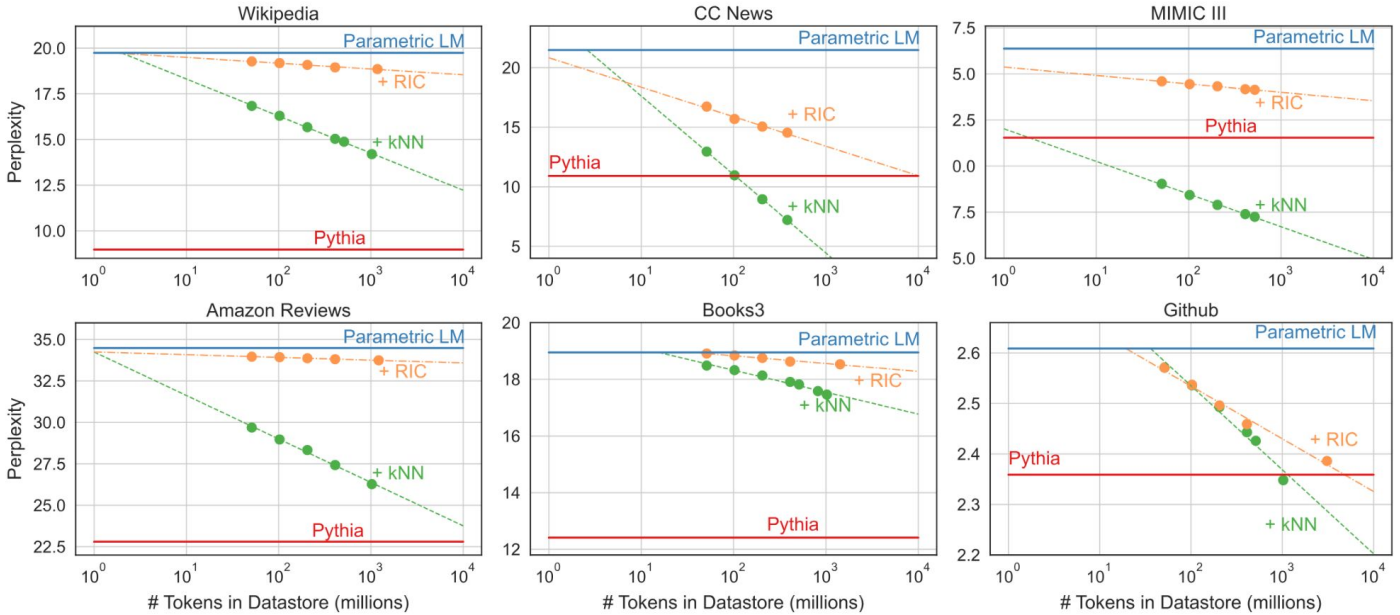


Figure 3: Impact of scaling the datastore of SILO (PDSW). Perplexity on random 128K tokens from the validation data reported. The rightmost dots for kNN-LM and RIC-LM in each figure correspond to the final models used in Table 4. **Scaling the test-time datastore consistently improves performance over all domains.**

Empirical results

Data	PDSW w/o upsampling	PDSW w upsampling
FreeLaw	4.9	5.7
Github	2.4	2.6
NIH ExPorter	20.0	19.3
PhilPapers	23.9	24.2
Wikipedia	19.9	19.7
CC News	21.8	21.3
BookCorpus2	19.4	19.2
OpenWebText2	21.0	21.2
Enron Emails	13.5	14.3
Amazon	35.7	34.7

Data	PD	PDSW w/o code	PDSW
FreeLaw	5.3	5.7	5.7
Github	13.3	8.2	2.6
NIH ExPorter	28.6	26.2	19.3
PhilPapers	55.2	36.4	24.2
Wikipedia	27.9	26.5	19.7
CC News	30.8	28.8	21.3
BookCorpus2	25.2	23.8	19.2
OpenWebText2	38.1	31.7	21.2
Enron Emails	19.9	18.5	14.3
Amazon	81.9	46.1	34.7

Table 11: **(Left)** Effect of re-weighting rare domains, comparing models trained on OLC (PDSW) with and without upsampling. **(Right)** Effect of SW data, with and without explicit source code—we train an LM with SW data but remove all of the actual source code (i.e., we leave Hacker News, Ubuntu IRC, Deepmind Math, and AMPS). Both tables report perplexity on the validation data.

Examples

Test Prefix 'I - what - *dragons*?' spluttered the Prime Minister. 'Yes, three,' said Fudge. 'And a sphinx. Well, good day to you.' The Prime Minister hoped beyond hope that dragons and sphinxes would be the worst of it, but no. Less than two years later, Fudge had erupted out of the fire yet again, this time with the news that there had been a mass breakout from **Test Continuation** **Azkaban**. 'A *mass* breakout?' the Prime Minister had repeated hoarsely. **Retrieved Prefix** 'D' you know Crouch, then?' said Harry. Sirius' face darkened. He suddenly looked as menacing as the night when Harry had first met him, the night when Harry had still believed Sirius to be a murderer. 'Oh, I know Crouch all right,' he said quietly. 'He was the one who gave me the order to be sent to **Retrieved Continuation** **Azkaban** - without a trial.'
Test Prefix Terror tore at Harry's heart... he had to get to Dumbledore and he had to catch Snape... somehow the two things were linked... he could reverse what had happened if he had them both together... Dumbledore could not have died... (...) Harry felt Greyback collapse against him; with a stupendous effort he pushed the werewolf off and onto the floor as a jet of **Test Continuation** **green** light came flying toward him; he ducked and ran, headfirst, into the fight. **Retrieved Prefix** Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of **Retrieved Continuation** **green** light issued from Voldemort's wand just as a jet of red light blasted from Harry's ...

Table 6: **Attribution examples on Harry Potter books.** We show the top-1 retrieved context of SILO (PDSW). **Red underline text** indicates the next token that immediately follows the prefix. In both examples, the test data is from the sixth novel and the retrieved context is from the fourth novel in the Harry Potter series. In the series, *Azkaban* is the notorious wizarding prison, and the *green light* is a distinct characteristic of the Killing Curse, *Avada Kedavra*.

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Conclusion

- SILO, a language model that mitigates legal risk
- Training with low-risk data, inference with high-risk data store
- Supports sentence-level data attribution and data opt-out

Limitations

- SILO does not completely eliminate legal risk.
  - does not remove the need for obtaining permission
- SILO might exacerbate certain fairness issues.
- SILO may underestimates the amount of permissively licensed text.

UVA | SCHOOL of ENGINEERING & APPLIED SCIENCE

Summarization

- LLMs suffer from targeted memorization
- Larger LLMs are more susceptible to data leaks from frequently repeated data
- Differential private learning usually lead to lower performances and higher costs on LLM's, but can be optimized with proper hyperparameter tuning and ghost clipping
- Legal-performance trade off remains a long lasting question for LLMs.