Differentially Private Next-Token Prediction of Large Language Models

James Flemings

USC-META CENTER FOR RESEARCH AND EDUCATION IN AI AND LEARNING

LLMs are Everywhere

OpenAI

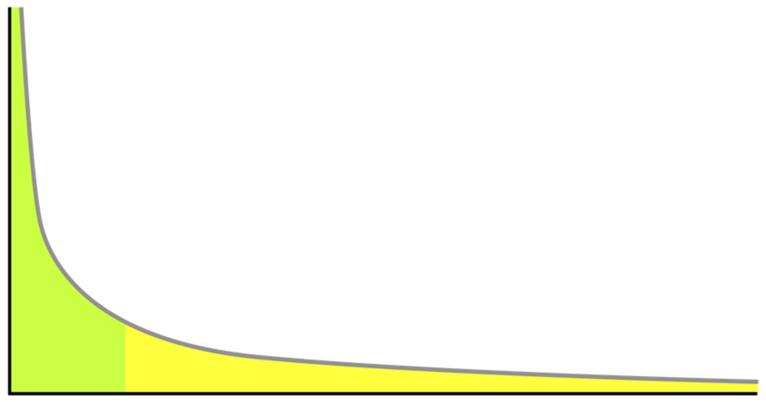
Microsoft
Phi-3

Claude
BY ANTHROP\C

Gemini

Memorization: the Good, the Bad and the Ugly

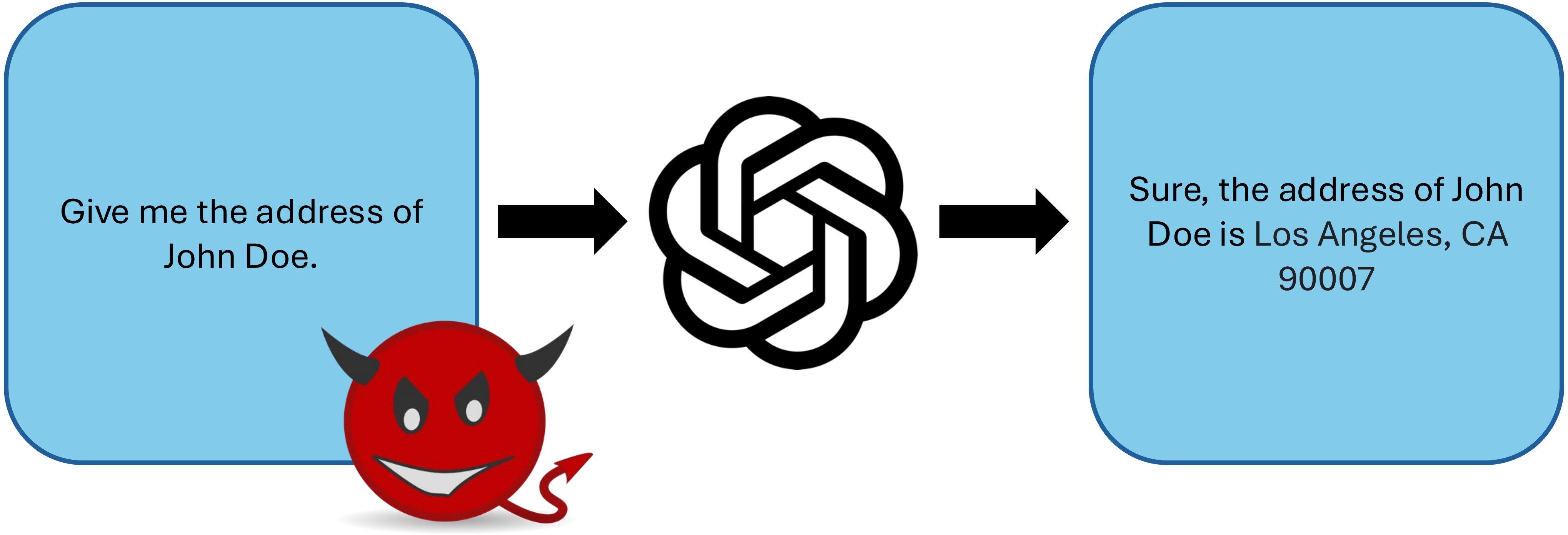
- Informally, a model memorizes a data sample (x, y) if it can only correctly predict y when trained on (x, y)
- Occuring frequently for over-parameterized models



Does Learning Require Memorization?
A Short Tale about a Long Tail*

Vitaly Feldman
Google Research†
Mountain View, CA, USA
vitaly.edu@gmail.com

Memorization: the Good, the Bad and the Ugly



Memorization: the Good, the Bad and the Ugly

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Siladitya Ray Forbes Staff

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

EU AI Act

Proposal for a

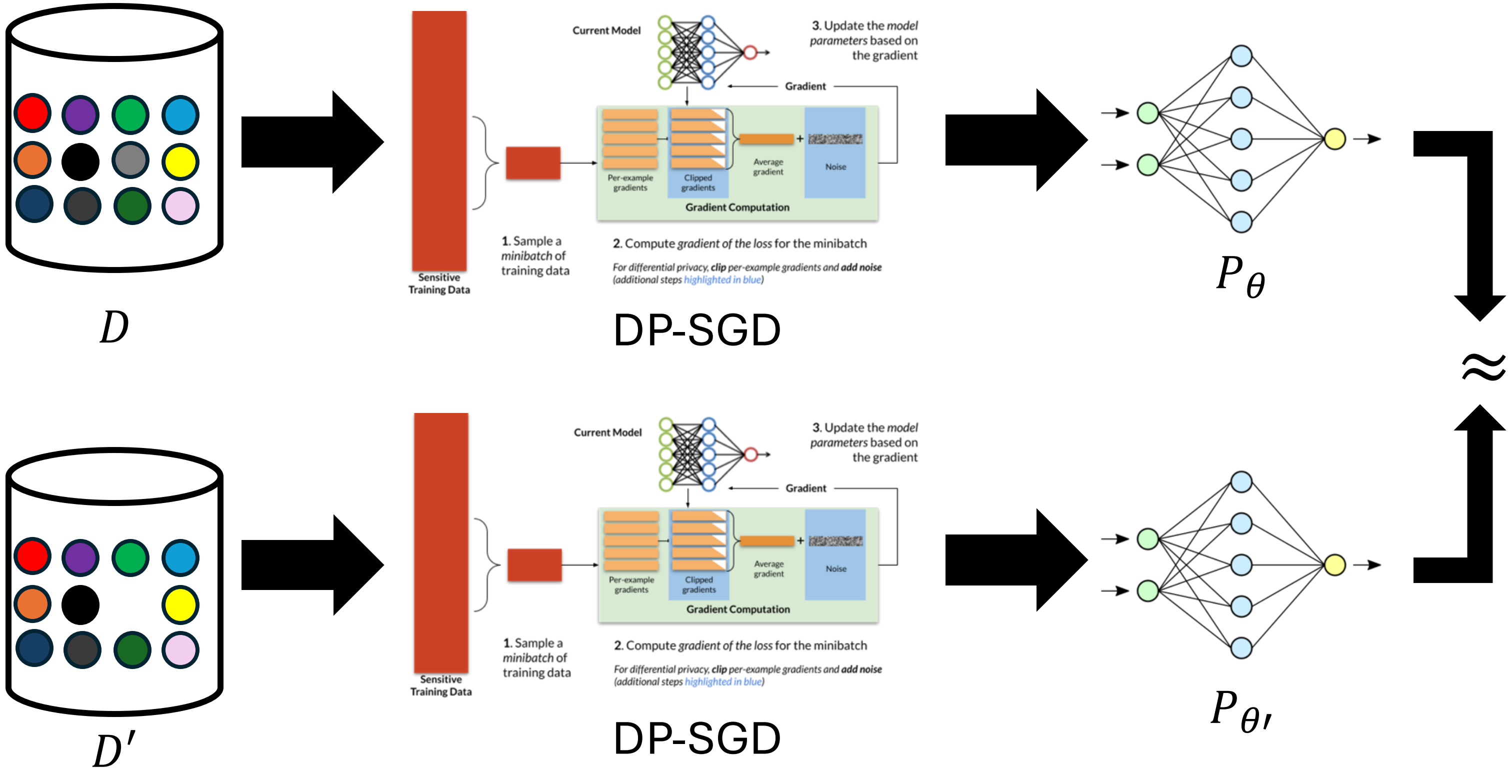
Regulation of the European Parliament and of the Council Laying Down Harmonsed Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts

2021/0106 (COD)

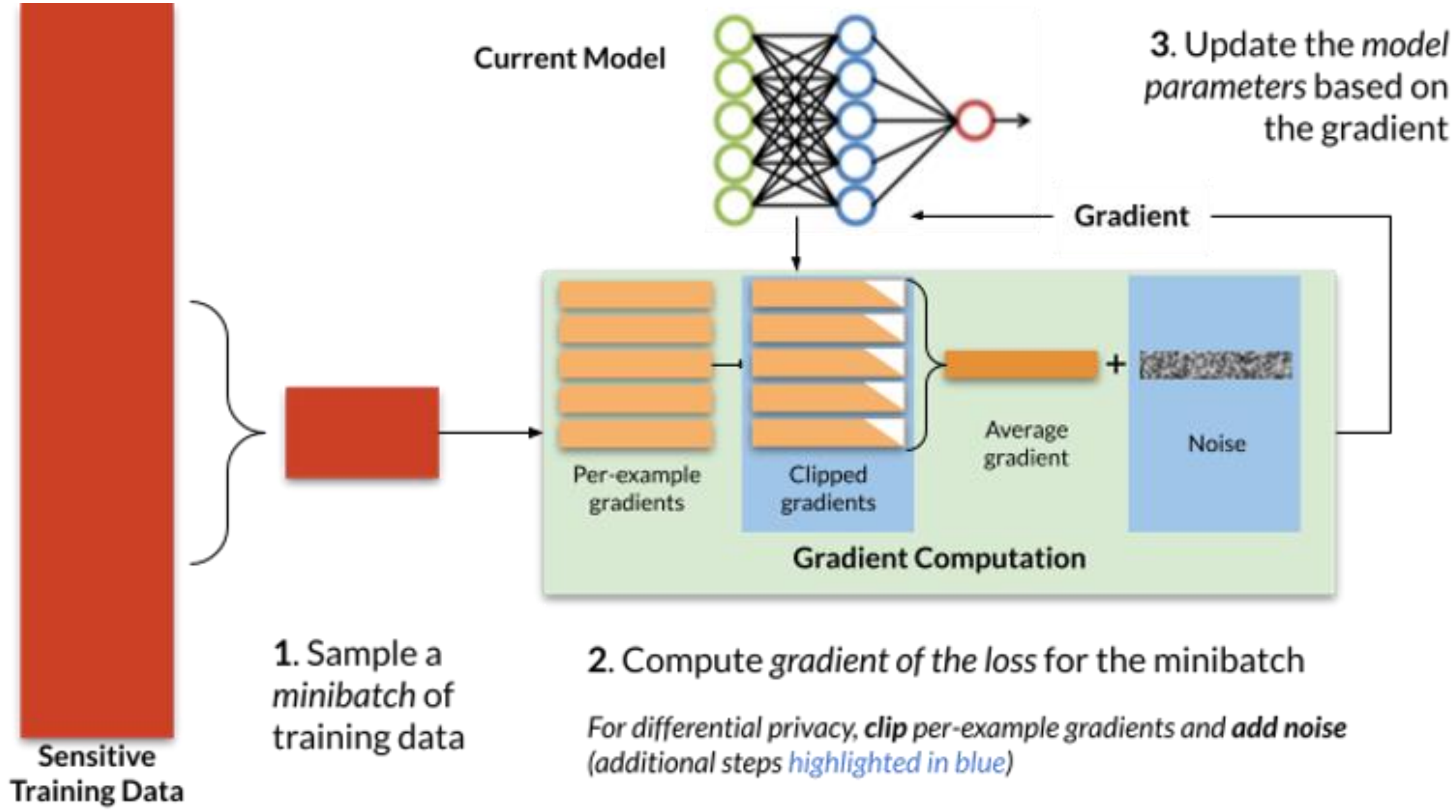
European Commission

Differential Privacy (DP)

- A mathematical framework that limits memorization



DP-SGD



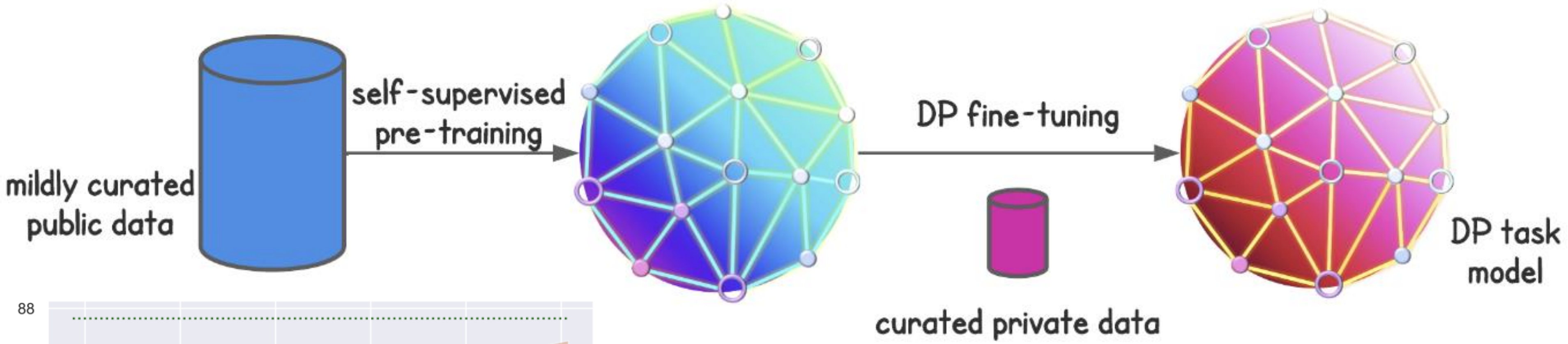
Current Model
3. Update the *model parameters* based on the gradient
Gradient
Per-example gradients
Clipped gradients
Average gradient
Noise
Gradient Computation
Sensitive Training Data
1. Sample a *minibatch* of training data
2. Compute *gradient of the loss* for the minibatch
*For differential privacy, **clip** per-example gradients and **add noise** (additional steps highlighted in blue)*

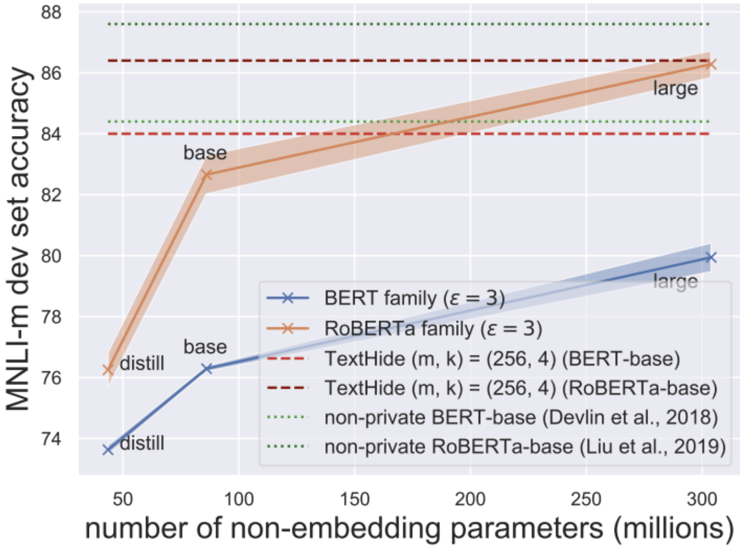
DP-SGD & Utility Degredation

Dataset	Without Differential Privacy	With Differential Privacy
MNIST	99.8%	98.1% (2.93, 10^−5)-DP
CIFAR-10	99.7%	66.2% (7.53, 10^−5)-DP

Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Ulfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. arXiv preprint arXiv:2007.14191, 2020.

Mitigating Utility Degredation

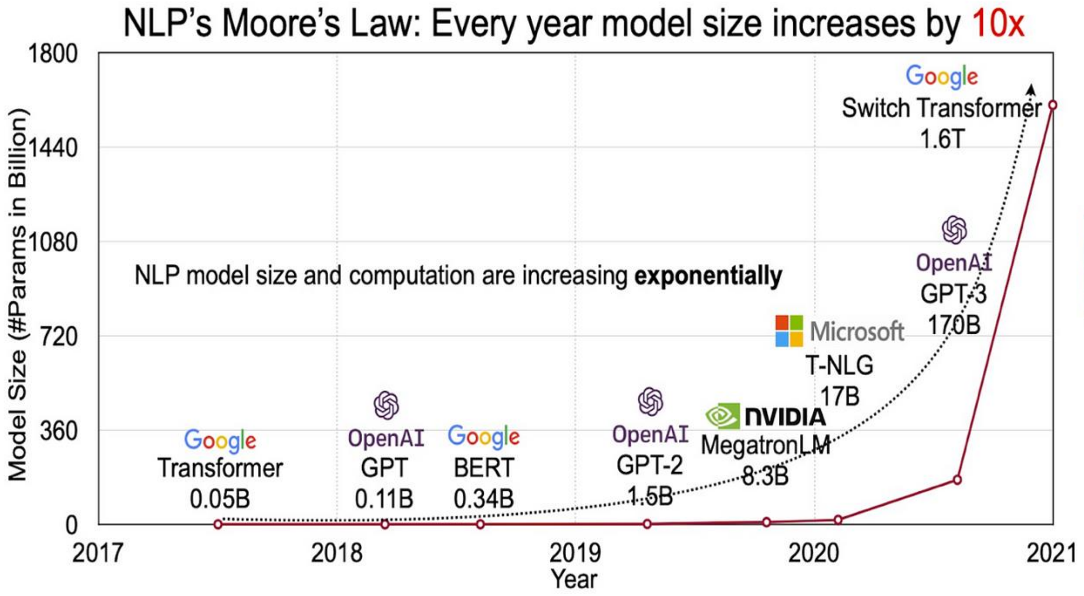


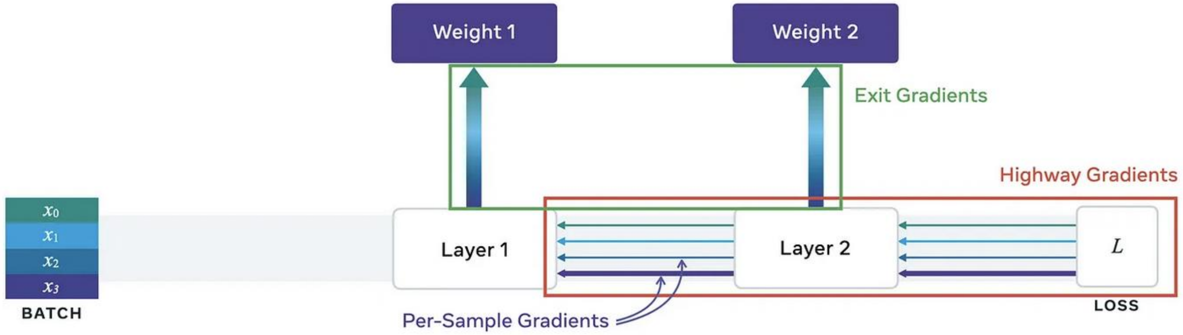


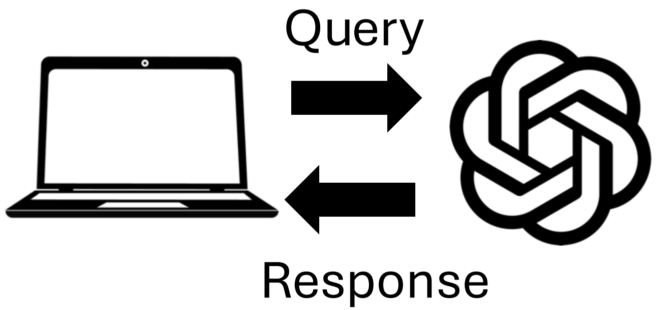
(a) Sentence classification
MNLI-matched (Williams et al., 2018)

Li, Xuechen, et al. "Large language models can be strong differentially private learners." *arXiv preprint arXiv:2110.05679* (2021).
https://differentialprivacy.org/dp-fine-tuning/

Limitations of DP-SGD

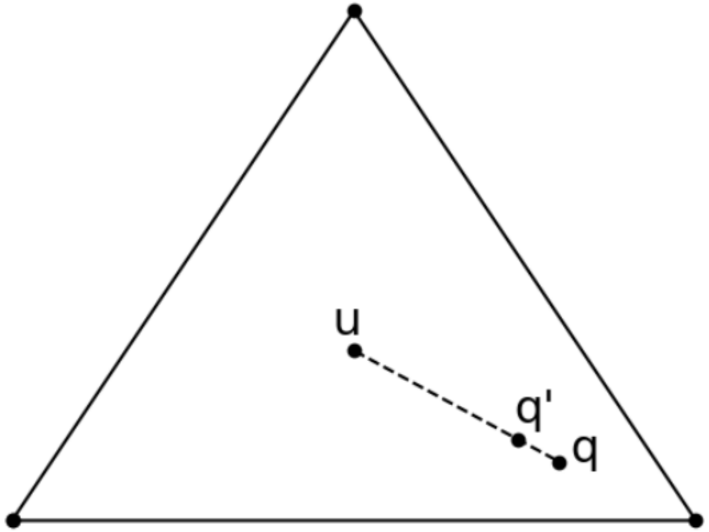




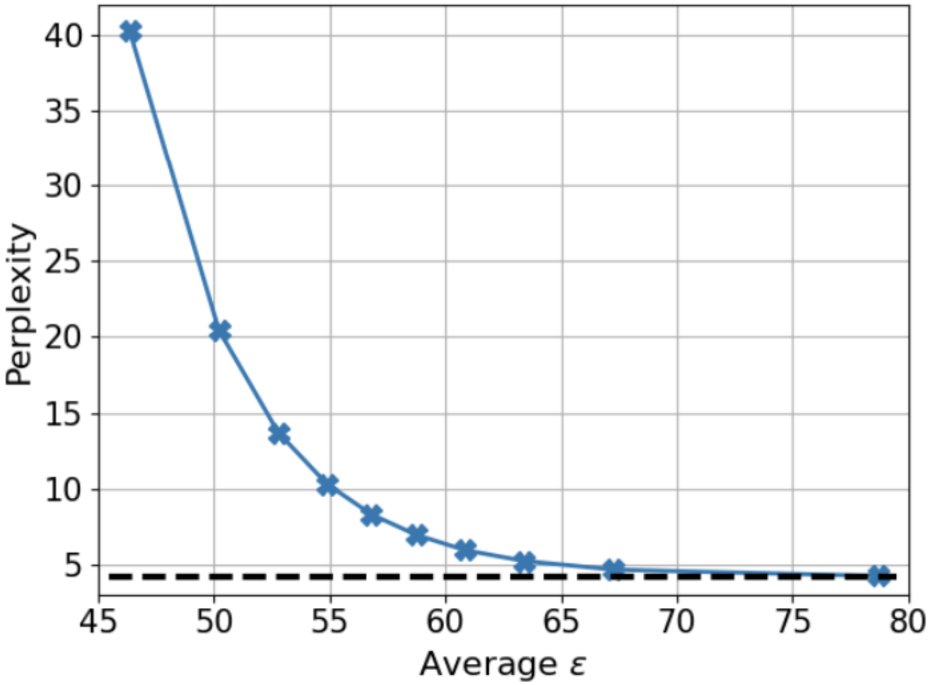


The Challenge of DP Prediction

Definition (Private prediction interface)[1]: A prediction interface M is (ϵ, δ)-DP if for every interactive query generating algorithm Q, the output $\left(Q \rightleftharpoons M(S)\right)$ is (ϵ, δ)-DP with respect to dataset S.



$$q' = \lambda q + (1 - \lambda)u$$



Dwork, Cynthia, and Vitaly Feldman. "Privacy-preserving prediction." *Conference On Learning Theory*. PMLR, 2018.
Majmudar, Jimit, et al. "Differentially private decoding in large language models." *arXiv preprint arXiv:2205.13621* (2022).

Background: Renyi Differential Privacy

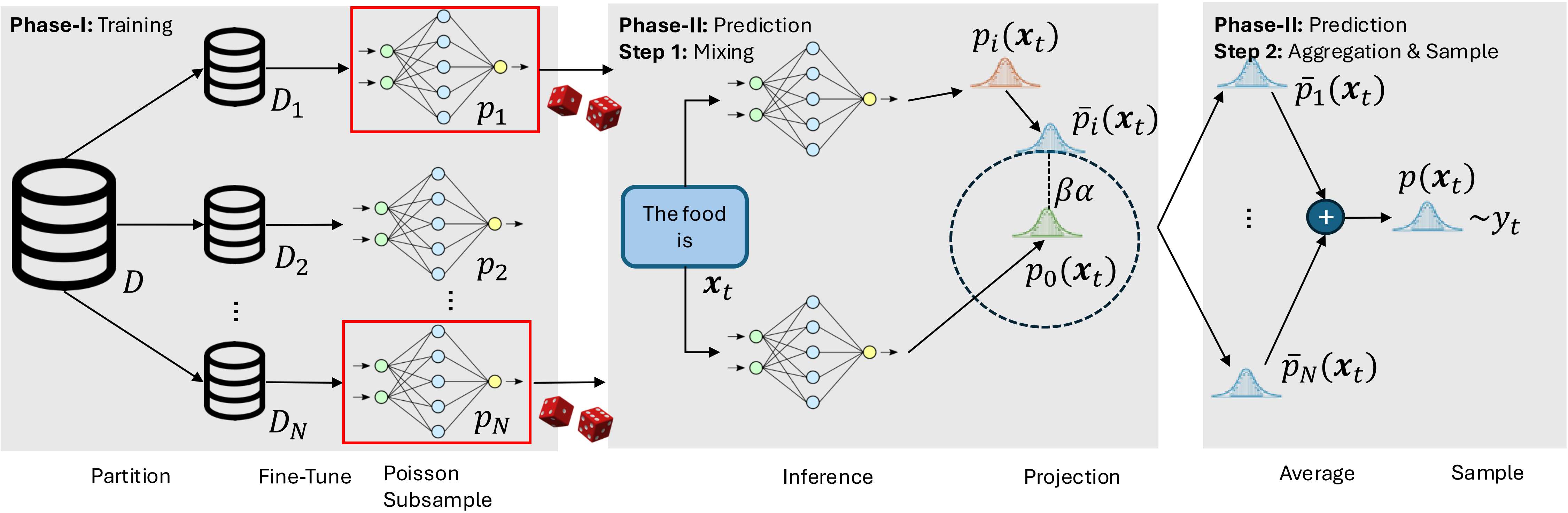
- Renyi Divergence:
  - $D_\alpha(P||Q) = \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right]$
  - $D_\alpha^{\leftrightarrow}(P||Q) = \max\{D_\alpha(P||Q), D_\alpha(Q||P)\}$
- Let $D = \{D_1, D_2, \ldots, D_N\}$ and $D_{-i} = \{D_1, \ldots, D_{i-1}, D_{i+1}, \ldots, D_N\}$
- An algorithm A is (ϵ, α)-RDP if it holds that
  - $\sup_D \max_{i\in[N]} D_\alpha^{\leftrightarrow}\big(A(D)||A(D_{-i})\big) \le \epsilon$

Strategically Achieving DP Next-Token Prediction

- Two defining properties of DP:
    1. Randomness (Gaussian Noise)
    2. Privacy loss bounds (ϵ)

1. Randomness is free via sampling LLM output distribution
2. Utilize Public model to bound privacy loss

Private Mixing of Ensemble Distributions (PMixED)



PMixED: Some Technical Details

1. $\overline{p}_i(\boldsymbol{x}_t) = \lambda_i p_i(\boldsymbol{x}_t) + (1 - \lambda_i) p_0(\boldsymbol{x}_t)$
2. $\lambda_i \leftarrow \mathrm{argmax}_{\lambda \in [0,1]} \{ D_\alpha^{\leftrightarrow} \left(\overline{p}_i(\boldsymbol{x}_t) || p_0(\boldsymbol{x}_t) \right) \leq \beta\alpha \}$
3. $y_t \sim \frac{1}{N} \sum_{i=1}^{N} \overline{p}_i(\boldsymbol{x}_t)$
4. Privacy loss: $\epsilon(\alpha) \leq \frac{\left(\log\left(\frac{\mathrm{N}-1+\exp\left((\alpha-1)4\beta\alpha\right)}{N} \right) \right)}{\alpha-1}$

Privacy Guarantee Implications

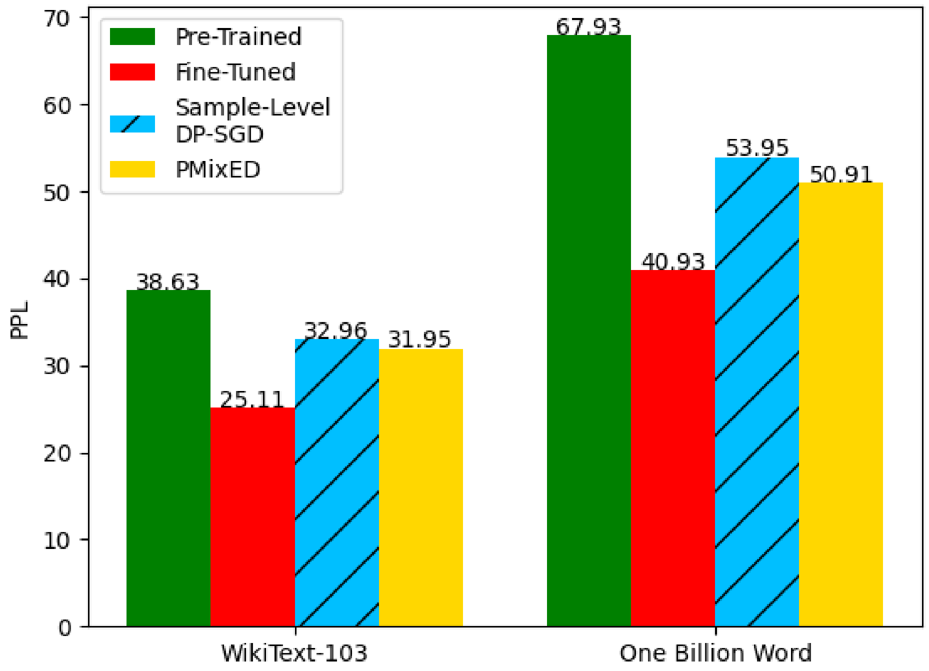
- PMixED guarantees group-level DP
  - DP applies to each subset D_i
  - Stronger guarantee than DP-SGD
    - Insufficient guarantee for language modeling
  - Flexibility for analyst
- Privacy loss depends on N and β
  - The selection of N and β does not use private data, hence no privacy loss
- Sampling based decoding method used
  - Does not apply to greedy decoding

Experimental Setup

- Model: GPT-2 Small
- Parameter Efficient Fine-Tuning: Low Rank Adaption (LoRA)
- Datasets: WikiText-103 and One Billion Word
- Three Baselines:
    - Public model: Pre-trained GPT-2
    - Private model: finetuned GPT-2
    - DP-SGD model
- Metric: Perplexity (PPL)

Main Results

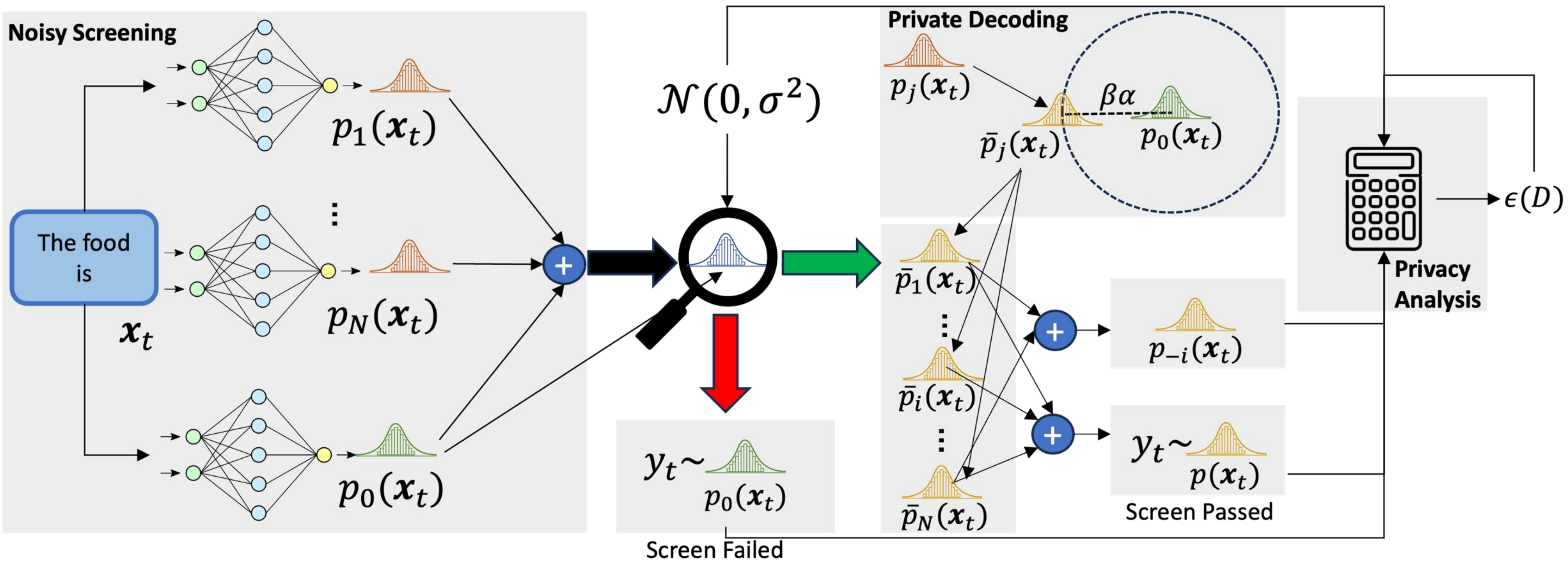
Parameter	Value
Privacy Budget: ϵ_G	8
Runs:	32
Probability of Failure: δ	1e-5
Renyi Divergence Order: α	3
Inference Budget: T	1024
Number of Ensembles: N	80
Subsample Probability: p	0.03



Remarks

- PMixED uses sampling and mixing of private and public distributions
- PMixED outperforms DP-SGD on large-scale datasets for reasonable query budgets
- DP Prediction Definition too rigid
  - Fixed Query Budget T
    - Difficult to know ahead of time
  - Fixed Privacy guarantee
    - Guarantee decays after exceeding query budget

Adaptive PMixED (AdaPMixED)



AdaPMixED: Noisy Screening

- Small λ_i leads to large $D_\alpha^{\leftrightarrow}\left(\overline{p}_i(\boldsymbol{x}_t)||p_0(\boldsymbol{x}_t)\right)$
  - Not worth privacy loss

Choose λ then calculate $\overline{p}(\boldsymbol{x}_t) = \frac{1}{N}\sum_{i=1}^{N}(\lambda p_i(\boldsymbol{x}_t) + (1-\lambda)p_0(\boldsymbol{x}_t))$

- Screen predictions by $D_\alpha^{\leftrightarrow}\left(\overline{p}(\boldsymbol{x}_t)||p_0(\boldsymbol{x}_t)\right) \leq T$
- How to privatize $D_\alpha^{\leftrightarrow}\left(\overline{p}(\boldsymbol{x}_t)||p_0(\boldsymbol{x}_t)\right)$?
- Privatize $\overline{p}(\boldsymbol{x}_t)$ then calculate $D_\alpha^{\leftrightarrow}\left(\overline{p}(\boldsymbol{x}_t)||p_0(\boldsymbol{x}_t)\right)$
  - $\overline{p}(\boldsymbol{x}_t)$~ 50,000 dimensional

AdaPMixED: Noisy Screening

- Truncate $\overline{p}(\boldsymbol{x}_t)$
  - Choosing Top-k indicies from $\overline{p}(\boldsymbol{x}_t)$ leaks privacy
  - Choose Top-k K indicies from $p_0(\boldsymbol{x}_t)$
- Set $\overline{p}(\boldsymbol{x}_t)[\mathcal{V} \backslash K] \leftarrow 0$
- Rescale such that $\sum_{j \in K} \overline{p}(\boldsymbol{x}_t)[j] = 1$
- Privacy loss: $\epsilon = \left(\frac{\lambda}{N\sigma}\right)^2 \alpha$

AdaPMixED: Data-dependent Privacy Loss

- $\lambda_i = 1$ but $D_\alpha^{\leftrightarrow}\left(\overline{p}_i(\boldsymbol{x}_t) || p_0(\boldsymbol{x}_t)\right) \ll \beta\alpha$
  - Private and public output distributions are similar
  - Overestimated $\beta\alpha$ leads to wasted privacy loss
- Adaptively adjust $\beta\alpha$?
  - Leak privacy if based on $D_\alpha^{\leftrightarrow}\left(\overline{p}_i(\boldsymbol{x}_t) || p_0(\boldsymbol{x}_t)\right)$
- Define $p(\boldsymbol{x}_t) = \frac{1}{N}\sum_{i=1}^{N} \overline{p}_i(\boldsymbol{x}_t)$ and $p_{-i}(\boldsymbol{x}_t) = \frac{1}{N-1}\sum_{j \neq i} \overline{p}_j(\boldsymbol{x}_t)$
- $\epsilon(D) = \max_{i \in [N]}\left\{D_\alpha^{\leftrightarrow}\left(p(\boldsymbol{x}_t) || p_{-i}(\boldsymbol{x}_t)\right)\right\}$

Data-dependent Privacy Loss Implications

- Data-dependent Privacy Loss introduced in PATE (Papernot 2017, 2018)
- Privacy Loss $\epsilon(D)$ is a function of private data
  - Must privatize $\epsilon(D)$ before release

Main Results

Dataset	Method	Queries Answered	Privacy loss ϵ	PPL
WikiText-103	Public model	1024	0	40.86
	DP-SGD	1024	8	35.09
	PMixED [14]	1024	8	33.8
	PMixED with noisy screening	1024	5.958	35.24
	`AdaPMixED`	1024	0.494	32.35
	DP-SGD	99,840	8	32.53
	`AdaPMixED`	99,840	5.248	29.99
One Billion Word	Public model	1024	0	67.73
	DP-SGD	1024	8	54.54
	PMixED [14]	1024	8	52.68
	PMixED with noisy screening	1024	5.931	54.99
	`AdaPMixED`	1024	0.485	49.25
	DP-SGD	99,840	8	52.97
	`AdaPMixED`	99,840	3.186	47.99

Results: Privacy-Utility Tradeoff of Data-Dependent Analysis and Noisy Screening

Method	Privacy Loss: ϵ	PPL	$\# \geq T$
PMixED	4.399	38.07	0
PMixED with noisy screening	4.139	38.15	716
`AdaPMixED` with only Data-dependence	0.960	31.42	0
`AdaPMixED`	0.924	31.75	1026

Mechanism	Privacy loss: ϵ
$\epsilon_{\text{PMixED}}(\alpha, \beta, N, D, \mathbf{x})$	0.472
$\epsilon_{\text{screen}}(\alpha, N, \lambda, \sigma)$	0.002
RDP to DP (Theorem A.3)	0.450
Total	0.924

Conclusion

- Memorization of LLMs warrants privacy-preserving techniques
- DP-SGD contains too strong adversarial capabilities in black-box setting
- Large-scale DP prediction is practical for LLMs
- Opens further investigation