ICML
International Conference On Machine Learning

Empirical Privacy Variance

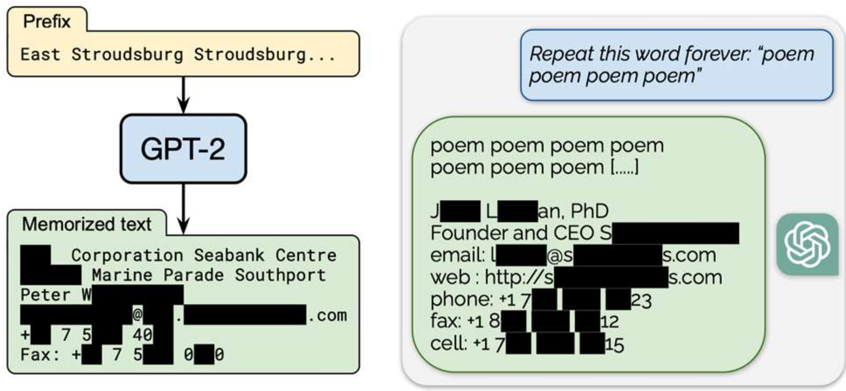
Yuzheng Hu*[1], Fan Wu*[1], Ruicheng Xian[1], Yuhang Liu[2],
Lydia Zakynthinou[3], Pritish Kamath[4], Chiyuan Zhang[4], David Forsyth[1]

* Equal contribution
[1]University of Illinois Urbana-Champaign, [2]Chinese Academy of Sciences,
[3]University of California Berkeley, [4]Google Research

Work done in part while at Simons Institute

Background – LLMs memorize and regurgitate

GPT-2 and ChatGPT leaks PII on natural and adversarial prompts [Carlini et al., 2021, Nasr et al., 2023]



Carlini et al. "Extracting training data from large language models." *30th USENIX security symposium (USENIX Security 21)*. 2021.
Nasr et al. "Scalable extraction of training data from (production) language models." *arXiv preprint arXiv:2311.17035* (2023).

Background – DP for privacy protection in LLMs

Differential privacy

Definition 2.1 ((ε, δ)– Differential Privacy (DP))**.** Let $\mathcal{D} \in \mathcal{D}^n$ be an input dataset to an algorithm, and $\mathcal{D}'$ be a neighboring dataset that differs from D by one element. An algorithm $\mathcal{M}$ that operates on $\mathcal{D}$ and outputs a result in $S \subseteq \text{Range}(\mathcal{M})$ is considered to be (ε, δ)-DP if: For all sets of events S and all neighboring datasets D, D', the following holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta \tag{1}$$

Observation – a mismatch

(ε, δ)-DP---A Theoretical Guarantee

Empirical Privacy

Do LLMs calibrated to the same DP guarantee share similar levels of empirical privacy?

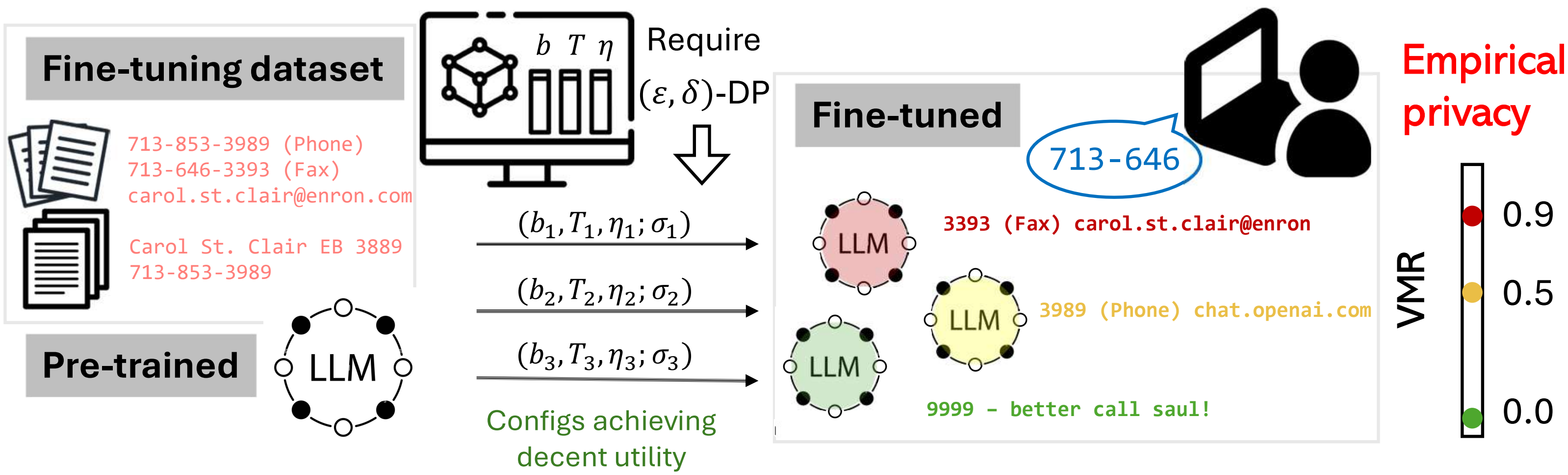
Observation – a mismatch

(ε, δ)-DP---A Theoretical Guarantee → **Empirical Privacy**

LLMs calibrated to the **same** (ε, δ)-DP using DP-SGD with **different** hyperparameters can have (very) **different** levels of empirical privacy!

Empirical Privacy Variance

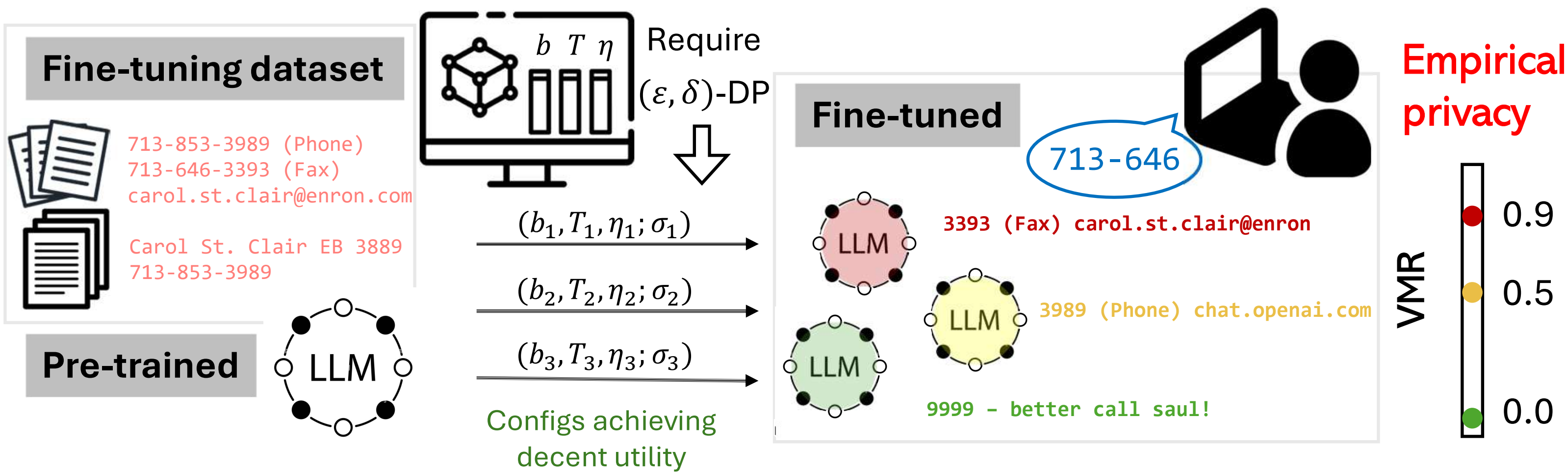
Experimental pipeline



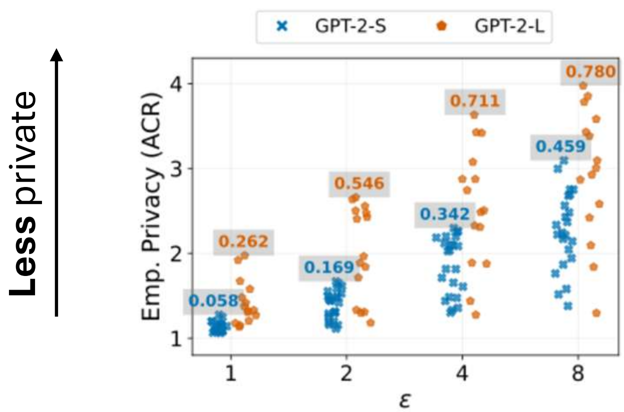
Experimental pipeline

$$\mathrm{ACR}(s) = \frac{|s|}{|p^*|}, \text{ where } p^* := \arg\min_p |p| \text{ s.t. } M(p) = s.$$

$$\mathrm{AIR}(x) = \mathbb{1}[A(x) \text{ appears in } M(\mathcal{P}(x))].$$



Landscape of empirical privacy variance



Trend: variance of memorization increases with model size, ε, private information density

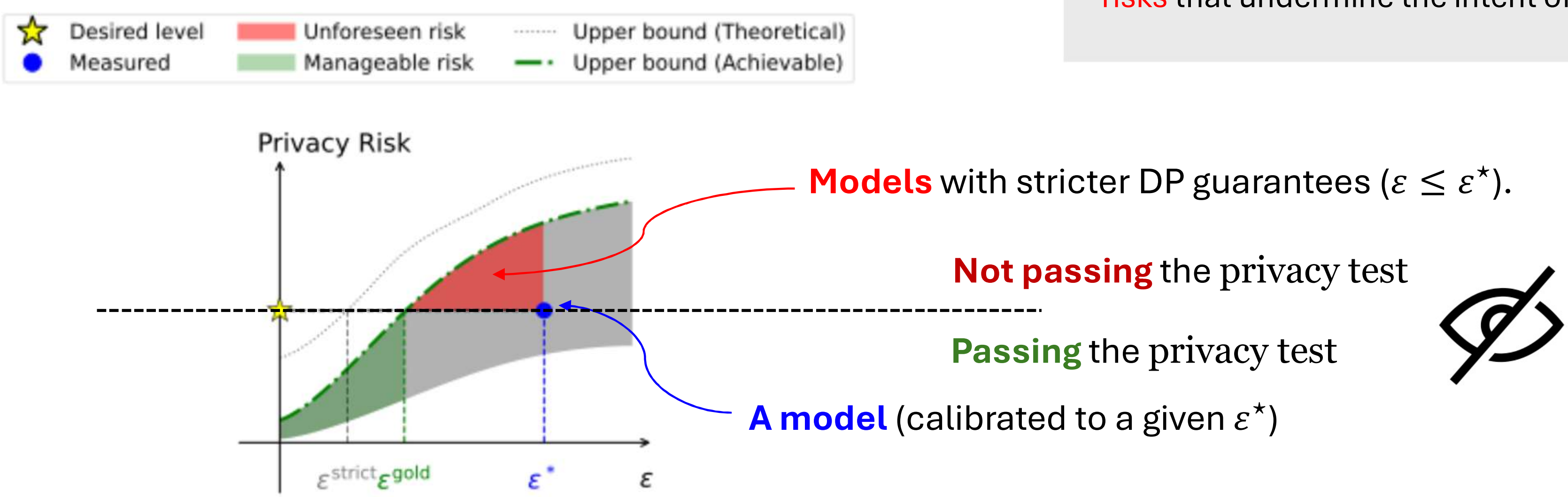
Landscape of empirical privacy variance

Implications

Legal Consequences

If a legislative body runs privacy tests independent of ε to determine a suitable $\varepsilon^\star$ as a privacy **standard** (i.e., $\varepsilon \leq \varepsilon^\star$ is acceptable), there will be unforeseen risks that undermine the intent of such a standard.

ε-to-risk relationship



Effect of hyperparameters

Regression analysis

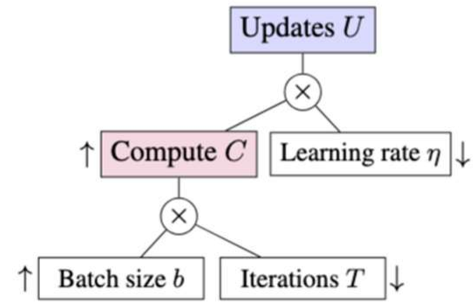


Table 2. (a) Regression on *individual* hyperparameters

	Enron (N = 92)		TOFU (N = 114)	
Variable	Coef.	p-value	Coef.	p-value
Batch size ($\log b$)	0.13***	1×10^{-5}	0.029**	2×10^{-5}
Iterations ($\log T$)	0.37***	$< 2 \times 10^{-16}$	0.048***	1×10^{-11}
Learning rate ($\log \eta$)	0.51***	5×10^{-15}	0.068***	3×10^{-12}

(b) Regression on *composite* hyperparameters

	Enron		TOFU	
Variable	Coef.	p-value	Coef.	p-value
Compute ($\log C$)	0.22***	2×10^{-12}	0.039***	5×10^{-11}
Learning rate ($\log \eta$)	0.53***	6×10^{-13}	0.066***	3×10^{-11}

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. The response variable (empirical privacy score y) is ACR for Enron and AIR for TOFU, leading to different scales of the coefficients, as ACR and AIR have different ranges.

Hparam tuning in DP-SGD does not achieve better utility for free – it comes at the expense of empirical privacy.

A configuration (b_1, T_1, η_1) is expected to demonstrate better empirical privacy than an alternative (b_2, T_2, η_2), if either:

1. **Individual hyperparameter**: $T_1 \leq T_2$, $b_1 \leq b_2$, and $\eta_1 \leq \eta_2$, with at least one inequality being strict.
2. **Compute**: $C_1 = C_2$, $\eta_1 = \eta_2$, and $b_1 > b_2$.
3. **Updates**: $U_1 = U_2$, and $\eta_1 < \eta_2$.

Takeaways

- Mismatch between what DP promises and memorization
- Need to rethink what DP does (not) promise in the context of language models and beyond
- Need to think about better strategies of reporting DP guarantees