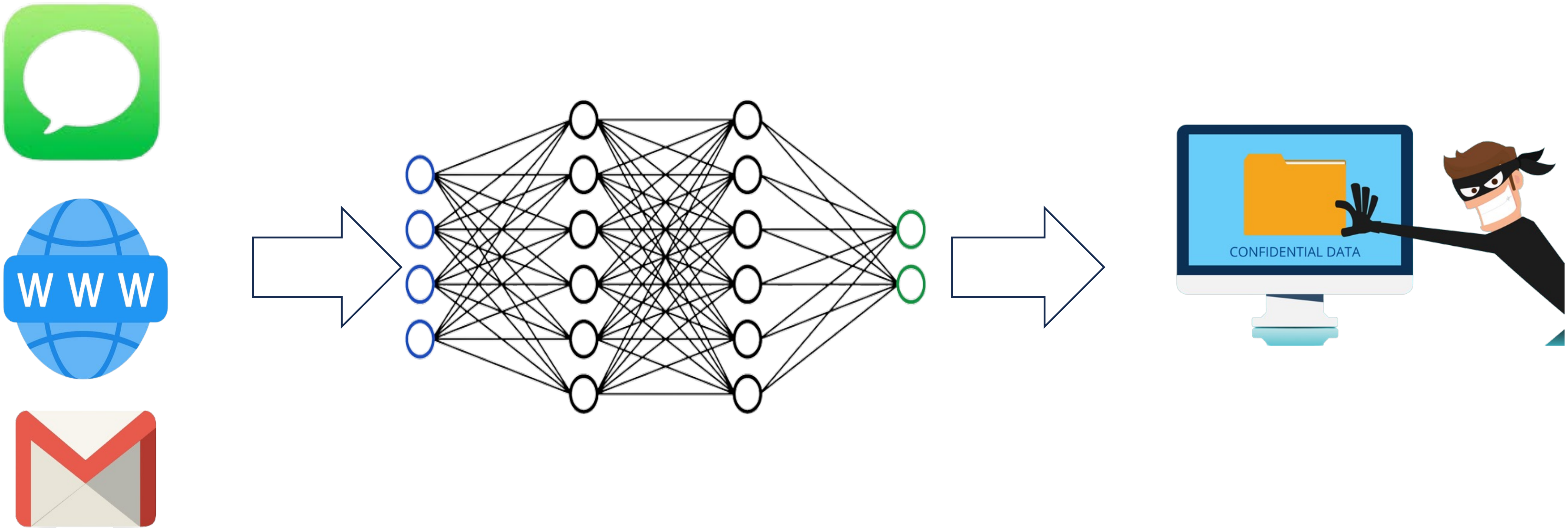
# Just Fine-tune Twice: Selective Differential Privacy for Large Language Models

**Weiyan Shi[†], Ryan Shea[†], Si Chen[‡], Chiyuan Zhang[◇], Ruoxi Jia[‡], Zhou Yu[†]**
Columbia University[†], Virginia Tech[‡], Google Research[◇]
{ws2634,rs4235}@columbia.edu, chensi@vt.edu, chiyuan@google.com,
ruoxijia@vt.edu, zy2461@columbia.edu

Shwai He
10/12/2023

# Data privacy is important!



# Differential Privacy: hides the existence of Individual Record

- **Definition 1. Neighboring datasets.** Given a domain, any two datasets D and D' that ***differs in exactly one record*** in this domain.

  E.g., D: *30* students in this course, ***30*** passed.

  D': *29* students in this course, ***29*** passed.

  ***Todo:*** Add random noise in datasets or algorithms.

- **Definition 2. $(\boldsymbol{\varepsilon}-\boldsymbol{\delta})$-differential privacy.** A *randomized algorithm* $\mathcal{M}: D \rightarrow R$ is a $(\varepsilon-\delta)$-differential private if for all neighboring datasets D and D' and all $T \subseteq R$,

$$\Pr[\mathcal{M}(D) \subseteq T] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \subseteq T] + \delta$$

The smaller the $\epsilon$ and $\delta$, the better the privacy.

Pretrained Language Models are not random!

# Deep Learning with Differential Privacy

- LLMs can remember privacy information.
- ***DPSGD***: Add noise in gradient.
- Avoid remembering privacy.

**Algorithm 1** Differentially private SGD (Outline)

**Input:** Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate $\eta_t$, noise scale $\sigma$, group size $L$, gradient norm bound $C$.

**Initialize** $\theta_0$ randomly

**for** $t \in [T]$ **do**

Take a random sample $L_t$ with sampling probability $L/N$

**Compute gradient**

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

**Clip gradient**

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

**Add noise**

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

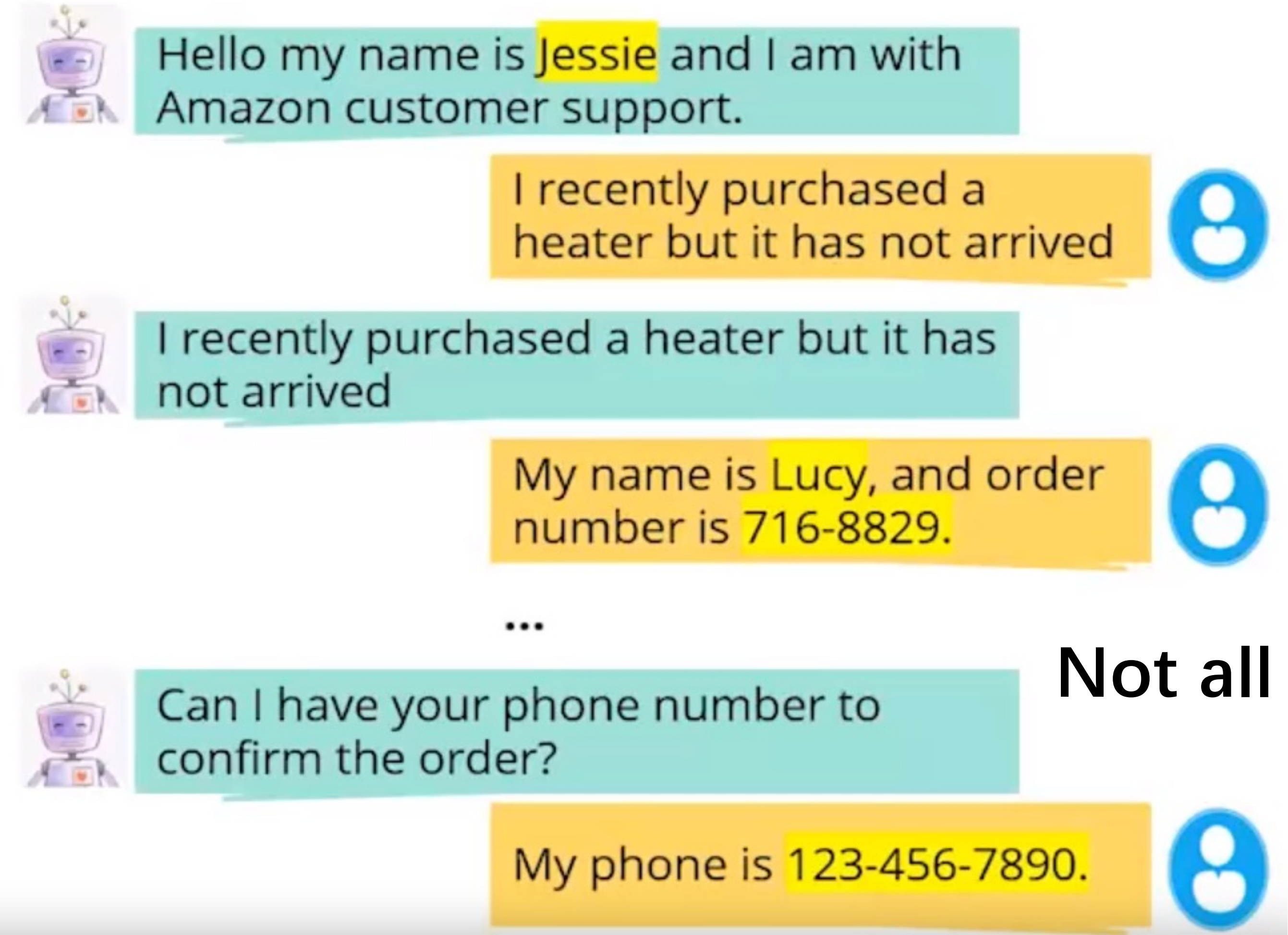
**Descent**

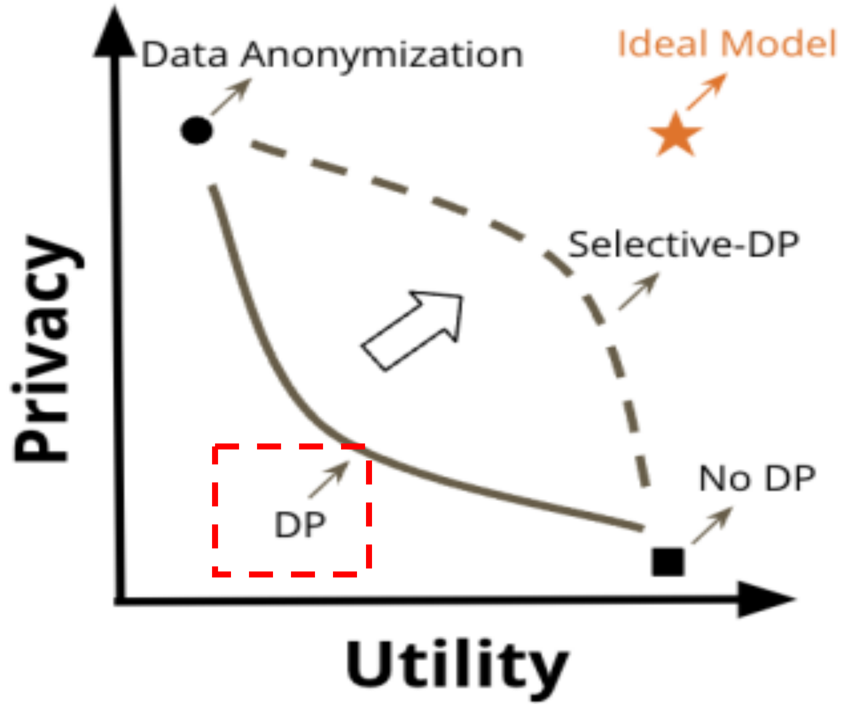
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

**Output** $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$ using a privacy accounting method.

## Trade-off between Privacy and Utility.

- Private information in language is sparse:




**Not all tokens should be protected.**

# Selective Differential Privacy (SDP)

- **Definition 3. Policy Function.** A policy function $F: \tau \rightarrow \{0,1\}^{|r|}$ decides ***which attributes*** of an example $r \in \tau$ are public ($F(r)_i = 1$) or private ($F(r)_i = 0$). $|r|$ is the number of attributes in $r$.

- **Definition 4.** Consider a policy function $F$ and two datasets $D$ and $D'$. $D'$ is a $F$-neighbor of $D$ (denoted by $D' \in N_F(D)$) if and only if $\exists r \in D$ s.t., $F(r)$ has ***at least one private attribute***, $\exists r' \in D'$ and $F(r')$ ***differ by at least one private attribute***, and $D' = D \setminus \{r\} \cup \{r'\}$.

✓ the dataset with "***My ID is 123***" and the dataset with "***My ID is 456***"

✗ the dataset with "***Hello there***" and the dataset with "***Hi there***"

Only disturb the gradient of $r$ with $F(r)_i = 0$!

# Secret Detectors of Different Levels

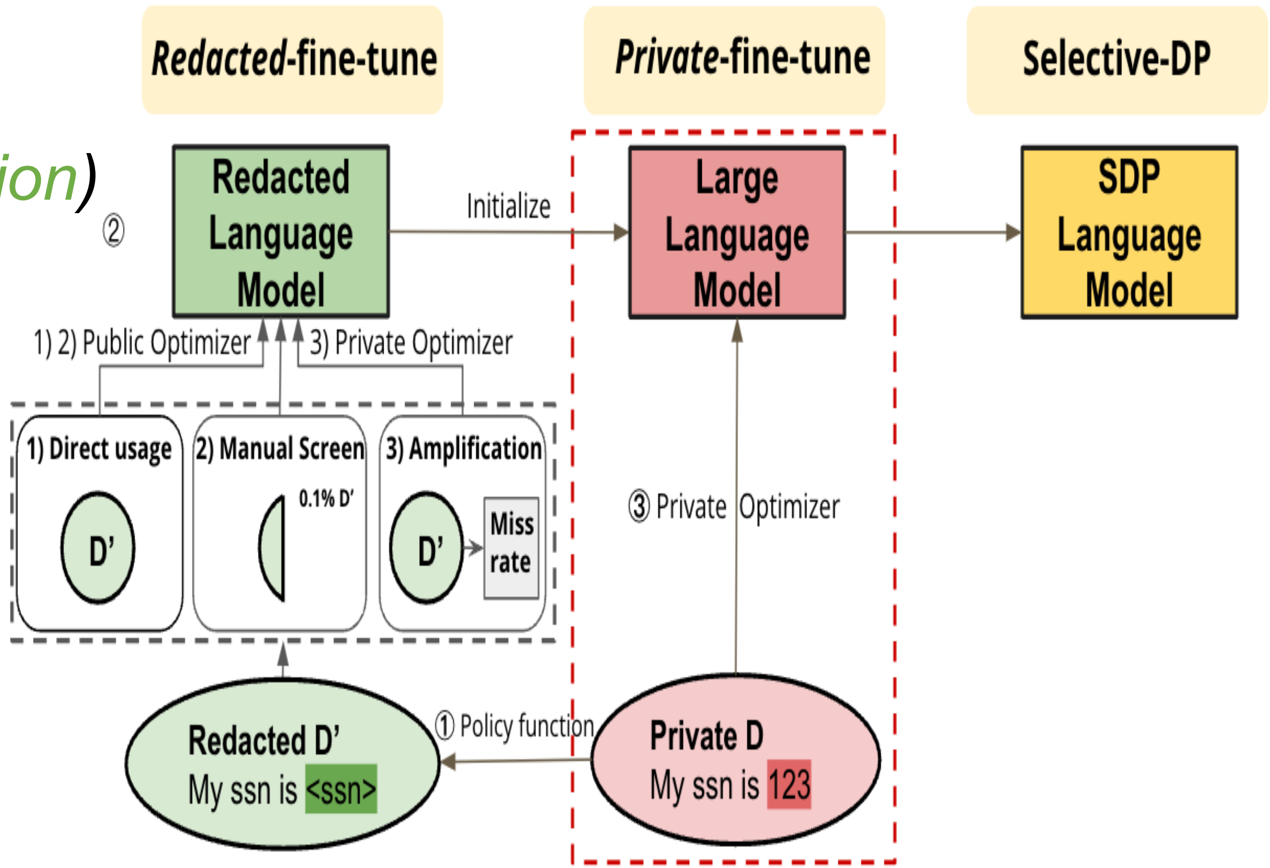
- **Low entity**: persc
- **High entity**: 18 e
- **Low contextual**:
- **High contextual**:

| **Secret Detector** | **What are you going to do about the custody of the kids?** | **Did you hear Alice is getting divorced?** |
|---|---|---|
| Low entity | What are you going to do about the custody of the kids? | Did you hear <PERSON> is getting divorced?? |
| High entity | What are you going to do about the custody of the kids? | Did you hear <PERSON> is getting divorced?? |
| Low contextual | <PRON> are <PRON> going to do about <OBJ> of the <OBJ>? | Did <PRON> hear <PROPN> is getting divorced?? |
| High contextual | <PRON> are <PRON> <VERB> to <VERB> about <OBJ> of the <OBJ>? | Did <PRON> <VERB> <PROPN> is getting <VERB>?? |

# JFT: Just Fine-tune Twice

- ***Redacted-fine-tune***
  - *Redacted D' (No Private Information)*
  - *Public Optimizer (SGD、Adam)*
  - ***Privacy***.

- ***Private-fine-tune***
  - *Private D (All data points)*
  - *Private Optimizer (SDP)*
  - ***Performance***.



# Experiments

- Datasets: 1)GLUE、2) Wikitext-2、 3) ABCD.
- Models:
  - RoBERTa-base → NLU classification.
  - GPT2-small → Language generation.
- Baselines:
  - No-DP: Adam optimizer.
  - DPSGD: Vanilla Differential Privacy.
  - CRT: Provably Confidential Language Modeling.
  - Redacted: No private information.
- Ours:
  - JFT
  - JFT + light noise

# Secret Detectors of Different Levels

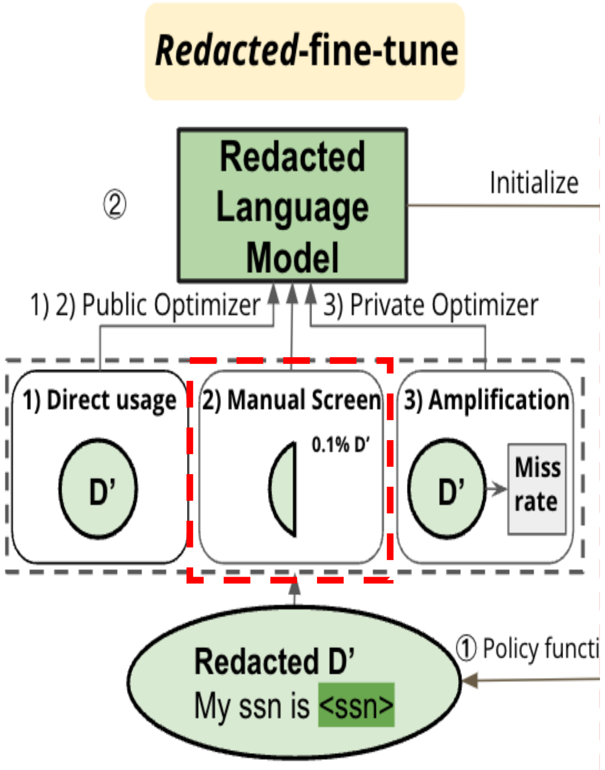
| **Direct Usage** | | **NLU on GLUE, $\delta_s$=1/2$\|D_{\text{train}}\|$** | | | | | | | | | | | | **Language Generation, $\delta_s$=1e-6** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **MNLI** | | | **QQP** | | | **QNLI** | | | **SST-2** | | | **WIKITEXT-2** | | | **ABCD** | | |
| **Model** | **Detector** | **Pct** | **Acc↑** | $\epsilon_s$ | **Pct** | **Acc↑** | $\epsilon_s$ | **Pct** | **Acc↑** | $\epsilon_s$ | **Pct** | **Acc↑** | $\epsilon_s$ | **Pct** | **PPL↓** | $\epsilon_s$ | **Pct** | **PPL↓** | $\epsilon_s$ |
| No-fine-tune | - | - | 31.82 | - | - | 36.82 | - | - | 50.54 | - | - | 50.92 | - | - | 30.08 | - | - | 13.60 | - |
| No-DP | - | - | 87.60 | - | - | 91.90 | - | - | 92.80 | - | - | 94.80 | - | - | 20.48 | - | - | 4.96 | - |
| DPSGD | - | - | 82.10 | 2.75 | - | 85.41 | 2.75 | - | 84.62 | 2.57 | - | 86.12 | 2.41 | - | 27.05 | 2.58 | - | 8.31 | 2.65 |
| DPSGD (+spe) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 30.32 | 2.58 | - | 17.75 | 2.71 |
| Redacted | low ent | 6.09% | 86.67 | - | 6.05% | 88.74 | - | 12.19% | 89.64 | - | 1.79% | 93.58 | - | 11.3% | 22.50 | - | 2.7% | 6.98 | - |
| JFT | low ent | 6.09% | **85.74** | 0.92 | 6.05% | **88.19** | 2.58 | 12.19% | **89.57** | 2.37 | 1.79% | **92.09** | 2.06 | 11.3% | **21.86** | 2.58 | 2.7% | **6.09** | 2.71 |
| Redacted | high ent | 8.63% | 86.50 | - | 8.30% | 88.36 | - | 17.18% | 88.96 | - | 3.01% | 93.58 | - | 16.4% | 24.32 | - | 3.1% | 7.32 | - |
| JFT | high ent | 8.63% | **85.61** | 0.99 | 8.30% | **88.05** | 2.58 | 17.18% | **89.35** | 2.37 | 3.01% | **92.20** | 2.12 | 16.4% | **22.55** | 2.58 | 3.1% | **6.25** | 2.71 |
| Redacted | low ctx | 31.19% | 85.14 | - | 32.61% | 85.59 | - | 35.68% | 85.30 | - | 22.19% | 92.55 | - | 34.8% | 37.90 | - | 22.3% | 28.28 | - |
| JFT | low ctx | 31.19% | **85.02** | 1.23 | 32.61% | **87.00** | 2.41 | 35.68% | **87.99** | 2.52 | 22.19% | **92.43** | 2.17 | 34.8% | **25.62** | 2.58 | 22.3% | 8.80 | 2.71 |
| **Stress-test** | | | | | | | | | | | | | | | | | | | |
| Redacted | high ctx | 44.27% | 83.23 | - | 45.93% | 83.48 | - | 45.59% | 82.81 | - | 38.13% | 91.86 | - | 45.0% | 54.29 | - | 28.6% | 65.45 | - |
| JFT | high ctx | 44.27% | **84.11** | 1.18 | 45.93% | **86.42** | 2.67 | 45.59% | **87.06** | 2.41 | 38.13% | **91.17** | 2.17 | 45.0% | 27.19 | 1.96 | 28.6% | 12.93 | 2.71 |

- JFT models achieve better model utility on both datasets.
- ***Special tokens affects the performance.***
- ***JFT is not always better than Redacted.***

# Selective Manual Screening

- ***Redated D' may still contains private information.***
- ***Assumption***: Secret detectors miss some secrets.
- ***Human Efforts:*** Manually sample ***0.1%*** data from *D'*.

| Manual Screening | $D'$ (redacted)=0.1% $D_0$, $D$ (private)=100% $D_0$ | | | | | |
|---|---|---|---|---|---|---|
| Task | MNLI Acc ↑ | QQP Acc ↑ | QNLI Acc ↑ | SST-2 Acc ↑ | WikiText-2 PPL ↓ | ABCD PPL ↓ |
| $D'$ size | 300 | 300 | 100 | 100 | 10 | 10 |
| DPSGD | 82.10 | 85.41 | 84.62 | 86.12 | 27.05 | 8.31 |
| Redacted | 52.52 | 75.25 | 66.48 | 88.88 | 28.06 | 9.36 |
| JFT+manual screening | **82.45** | **86.24** | **85.00** | **90.83** | **26.72** | **7.84** |

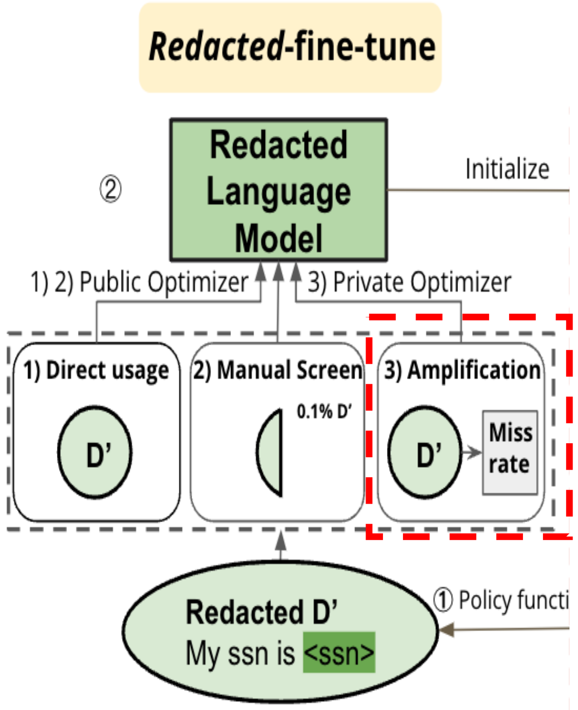



- Fine-tuning with ***a small manually-screened*** in-domain subset can still help the model learn in-domain information, and lead to better utility.

# Lightly Noised Optimizer with Privacy Amplification

- ***Strong Assumption***: No private information contained in D'.
- ***Real-life Scenarios***: Add noise to the private optimizer in the ***first*** phase.

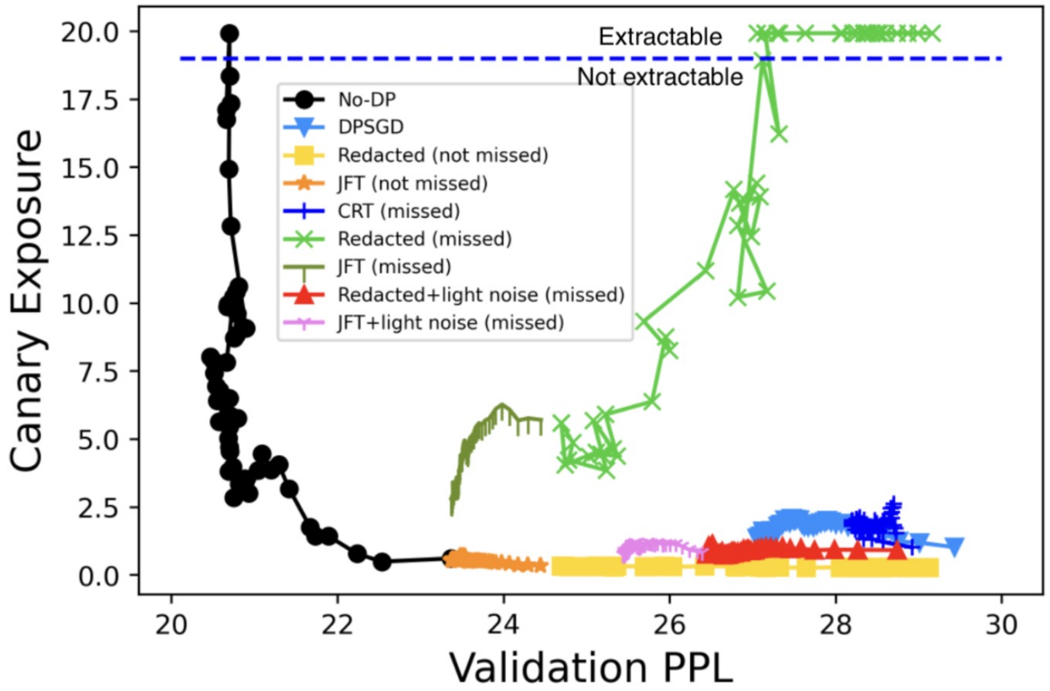
| | MNLI | | QQP | | QNLI | | SST-2 | | WikiText-2 | | ABCD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Acc↑** | 95%-$\epsilon_s$ | **Acc↑** | 95%-$\epsilon_s$ | **Acc↑** | 95%-$\epsilon_s$ | **Acc↑** | 95%-$\epsilon_s$ | **PPL↓** | 95%-$\epsilon_s$ | **PPL↓** | 95%-$\epsilon_s$ |
| DPSGD | 82.10 | 2.75 | 85.41 | 2.75 | 84.62 | 2.57 | 86.12 | 2.41 | 27.05 | 2.58 | 8.31 | 2.65 |
| Missing rate $m$ (95% CI) | (0.3%, 1.2%) | | (0.3%, 1.2%) | | (0.1%, 0.6%) | | (0%, 1.8%) | | (0.4%, 0.7%) | | (0.1%, 1.2%) | |
| Recall (95% CI) | (87.5, 96.7) | | (85.9, 96.1) | | (96.4, 99.3) | | (40.2, 100) | | (95.6, 97.8) | | (62.7, 97.4) | |
| JFT+light noise | **82.76** | (0.08, 0.43) | 85.28 | (1.40, 1.71) | **84.88** | (2.29, 2.68) | **89.33** | (0, 0.43) | **25.21** | (2.73, 2.92) | **5.78** | (1.08, 1.60) |
| **Conservative Estimation** | | | | | | | | | | | | |
| Missing rate $m$ | 8.6% | | 8.3% | | 17.2% | | 3.0% | | 16.4% | | 3.0% | |
| Recall | 0 | | 0 | | 0 | | 0 | | 0 | | 0 | |
| JFT+light conservative noise | 82.00 | 0.45 | 84.77 | 2.91 | 84.02 | 2.95 | **89.22** | 0.43 | **26.59** | 3.03 | **6.64** | 1.67 |




- "JFT+light noise" achieves better utility than DPSGD, especially on generation tasks.
- "JFT+light conservative noise" is still better than DPSGD on some tasks.
- ***The performance depends on the level of noise.***

# Attack Results

- Case Study




- Insert the canary "*My ID is 341752*" into the training data for *10* times
- Models without protection do memorize the data unintentionally.
- Whether the secret detector misses the canary influences the exposure.

# Limitations and discussion.

- The effect of special tokens.
  - Special token on JFT. / Role of different special tokens.
- The experiments about Redacted-Fine-tune.
  - Ablation study on dataset size.
  - Dropping out private data.
- Incomplete experiments on privacy.
  - Case study is not enough.
- Add theoretical analysis can be much better.

Thanks for you listening!