Provably Confidential Language Modelling

Xuandong Zhao **Lei Li** **Yu-Xiang Wang**
University of California, Santa Barbara
{xuandongzhao,leili,yuxiangw}@cs.ucsb.edu

Ming Li

Motivation

- Train a high-performing language model without memorizing sensitive text

Motivation

- Compared with Differential Privacy (DP) settings:

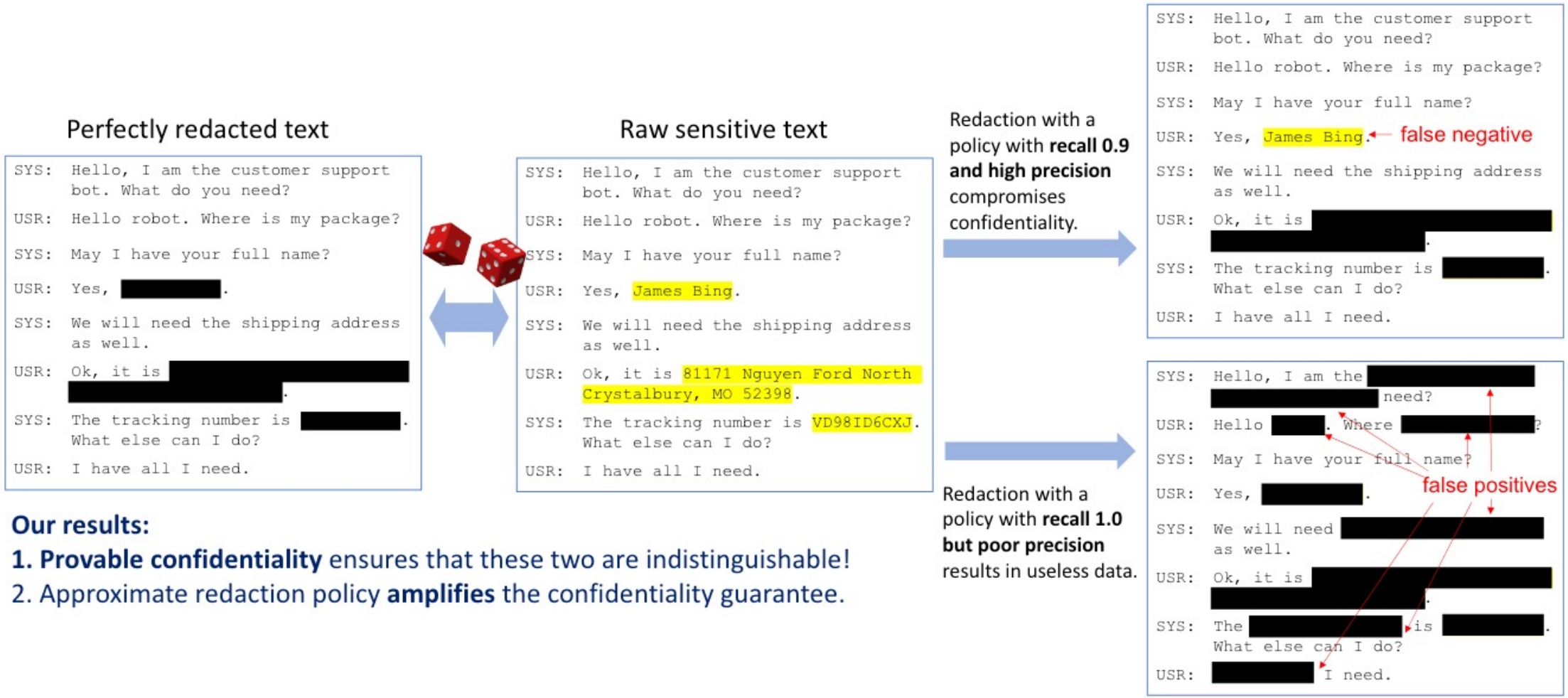
- 1. Confidential information in a natural language dataset is sparse
- 2. What needs to be protected is the content of the sensitive text, rather than the data context.
- 3. The same sensitive content could appear in many data points, which makes the protection of the content more challenging than protecting one data sample.

Motivation

- Redaction
- The process of removing sensitive or classified information from a document prior to its publication in governmental and legal contexts.

- Deduplication
- The procedure of detecting and removing identical and nearly identical texts from a corpus.

Motivation



Our results:

1. **Provable confidentiality** ensures that these two are indistinguishable!
2. Approximate redaction policy **amplifies** the confidentiality guarantee.

Contribution

- 1. Show that the risk of a language model memorizing sensitive content is real and can be efficiently exploited
- 2. Introduce a new definition of confidentiality which precisely quantifies the risk of leaking sensitive text
- 3. Propose CRT to train language generation models while protecting confidential text.
- 4. Prove that CRT, combined with differentially private stochastic gradient descent, provides strong confidentiality guarantees.
- 5. Different models trained by CRT can achieve the same or better perplexity than existing solutions

Differential Privacy (DP)

- Differential privacy is a **mathematical framework for ensuring the privacy of individuals in datasets**.
- Differential privacy ensures that the output of a function, when applied to slightly different datasets (differing in just one entry, for instance, one person's data), should be roughly the same. This guarantees that an adversary cannot determine whether a specific individual's information is included in the input to the function based solely on the output.

Formal Definition of Confidentiality

Definition 1 (Indistinguishability). *We say that a pair of distributions P, Q defined on the same probability space are (ϵ, δ)-indistinguishable if for any measurable set S,*

$$\Pr_P[S] \leq e^{\epsilon} \Pr_Q[S] + \delta.$$

Definition 2 (Confidentiality). *We say that $\mathcal{A}$ ensures that a secret x is $(\epsilon(x), \delta)$-confidential, if for any dataset D that contains x in one of its data points, and an alternative dataset D' that replaces x in D with a generic <MASK>, it holds that $(\mathcal{A}(D), \mathcal{A}(D'))$ are $(\epsilon(x), \delta)$-indistinguishable. In addition, we simply say that $\mathcal{A}$ ensures (ϵ, δ)-confidentiality if $\epsilon(x) \leq \epsilon$ for all secret x.*

Formal Definition of Confidentiality

Definition 3 (Group Confidentiality). *We say that $\mathcal{A}$ ensures that a* ***list*** *of sensitive texts $\mathcal{S} := [x_1, ..., x_k]$ is $(\epsilon(\mathcal{S}), \delta)$-(group) confidential, if for any dataset D that contains $[x_1, ..., x_k]$ in up to k data points, and D' being the version that replaces each element in $\mathcal{S}$ with <MASK>, it holds that $(\mathcal{A}(D), \mathcal{A}(D'))$ are $(\epsilon(\mathcal{S}), \delta)$-indistinguishable.*

Definition 4 (Bayesian Confidentiality). *Let D be a dataset that is fixed except a random secret $x \sim \mu$ drawn from some distribution μ. Let D' be obtained by replacing x with <MASK>[2]. Then $\mathcal{A}$ ensures (ϵ, δ)-Bayesian Confidentiality if for any D', $(\mathcal{A}(D), \mathcal{A}(D'))$ is (ϵ, δ)-indistinguishable, where $\mathcal{A}(D)$ is jointly distributed over $x \sim \mu$ and $\mathcal{A}$.*

Differential Privacy - SGD

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 η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_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 θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Confidentially Redacted Training

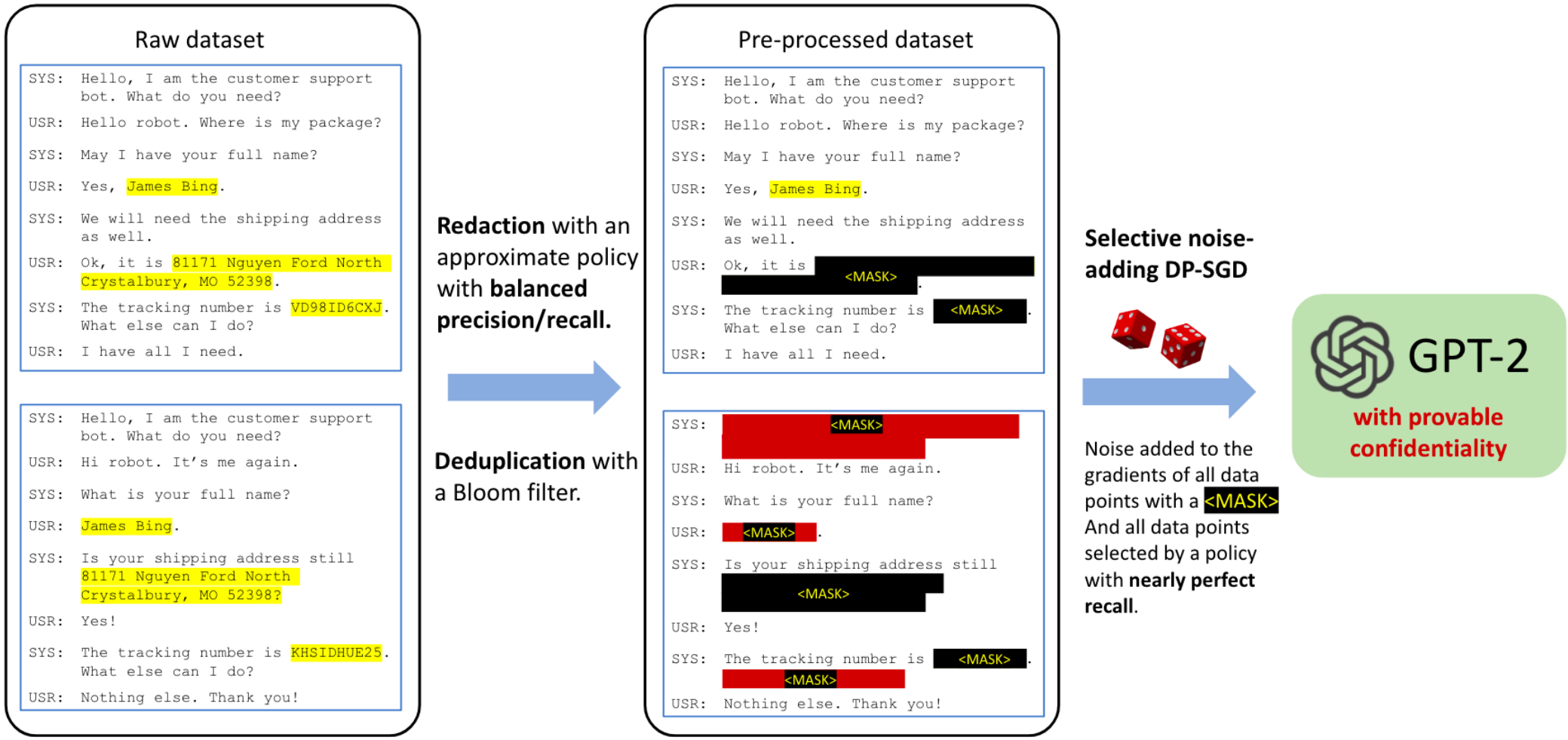
- The overall idea is to screen the corpus into two separate sets, one public set including sentences with no confidential information, and one private set including sentences containing confidential content.

CRT

Deduplication. The deduplication procedure Dedup detects all sentences that appear multiple times in the training data and replace them into a single <MASK> from the second occurrence onwards (<MASK> is for proving purpose).

Redaction. The redaction procedure Redact_π takes applies a sequence labelling policy π to screen confidential content in the training corpus D. $\pi(s, x) = 1$ if a token x in a sentence s should be confidential. The labeled span in each detected sentence is replaced with a special token <MASK>. Note that we do not assume the policy is perfect. It may label some non-sensitive tokens as sensitive (false positives) and label some sensitive text as non-sensitive (false negative, or $1-$recall).

CRT


Raw dataset
SYS: Hello, I am the customer support bot. What do you need?
USR: Hello robot. Where is my package?
SYS: May I have your full name?
USR: Yes, James Bing.
SYS: We will need the shipping address as well.
USR: Ok, it is 81171 Nguyen Ford North Crystalbury, MO 52398.
SYS: The tracking number is VD98ID6CXJ. What else can I do?
USR: I have all I need.
SYS: Hello, I am the customer support bot. What do you need?
USR: Hi robot. It's me again.
SYS: What is your full name?
USR: James Bing.
SYS: Is your shipping address still 81171 Nguyen Ford North Crystalbury, MO 52398?
USR: Yes!
SYS: The tracking number is KHSIDHUE25. What else can I do?
USR: Nothing else. Thank you!
Redaction with an approximate policy with balanced precision/recall.
Deduplication with a Bloom filter.
Pre-processed dataset
SYS: Hello, I am the customer support bot. What do you need?
USR: Hello robot. Where is my package?
SYS: May I have your full name?
USR: Yes, James Bing.
SYS: We will need the shipping address as well.
USR: Ok, it is <MASK>.
SYS: The tracking number is <MASK>. What else can I do?
USR: I have all I need.
SYS: <MASK>
USR: Hi robot. It's me again.
SYS: What is your full name?
USR: <MASK>.
SYS: Is your shipping address still <MASK>
USR: Yes!
SYS: The tracking number is <MASK> <MASK>.
USR: Nothing else. Thank you!
Selective noise-adding DP-SGD
Noise added to the gradients of all data points with a <MASK>
And all data points selected by a policy with nearly perfect recall.
GPT-2
with provable confidentiality

CRT

Algorithm 1: CRT

Input : Dataset D (after tokenization / splitting), labelling policies π, π_c, number of epochs T

$D' \leftarrow \text{Dedup}(D)$

$D'' \leftarrow \text{Redact}_{\pi}(D')$

$D^{pri} \leftarrow \{s \in D'' | \exists x \in s \text{ s.t. } \pi(s, x) = 1 \text{ or } \exists x \subset s \text{ s.t. } \pi_c(s, x) = 1\}$

$D^{pub} = \{s \in D'' | s \notin D^{pri}\}$.

for $e = 1, ..., T$ **do**

  Run one epoch of SGD with D^{pub}.

  Run one epoch[3] of DP-SGD with D^{pri}.

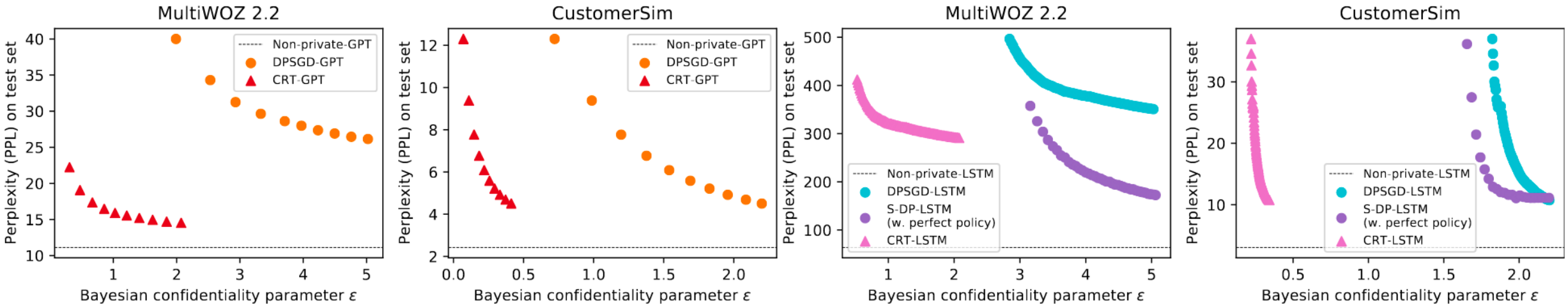
end

Experiments

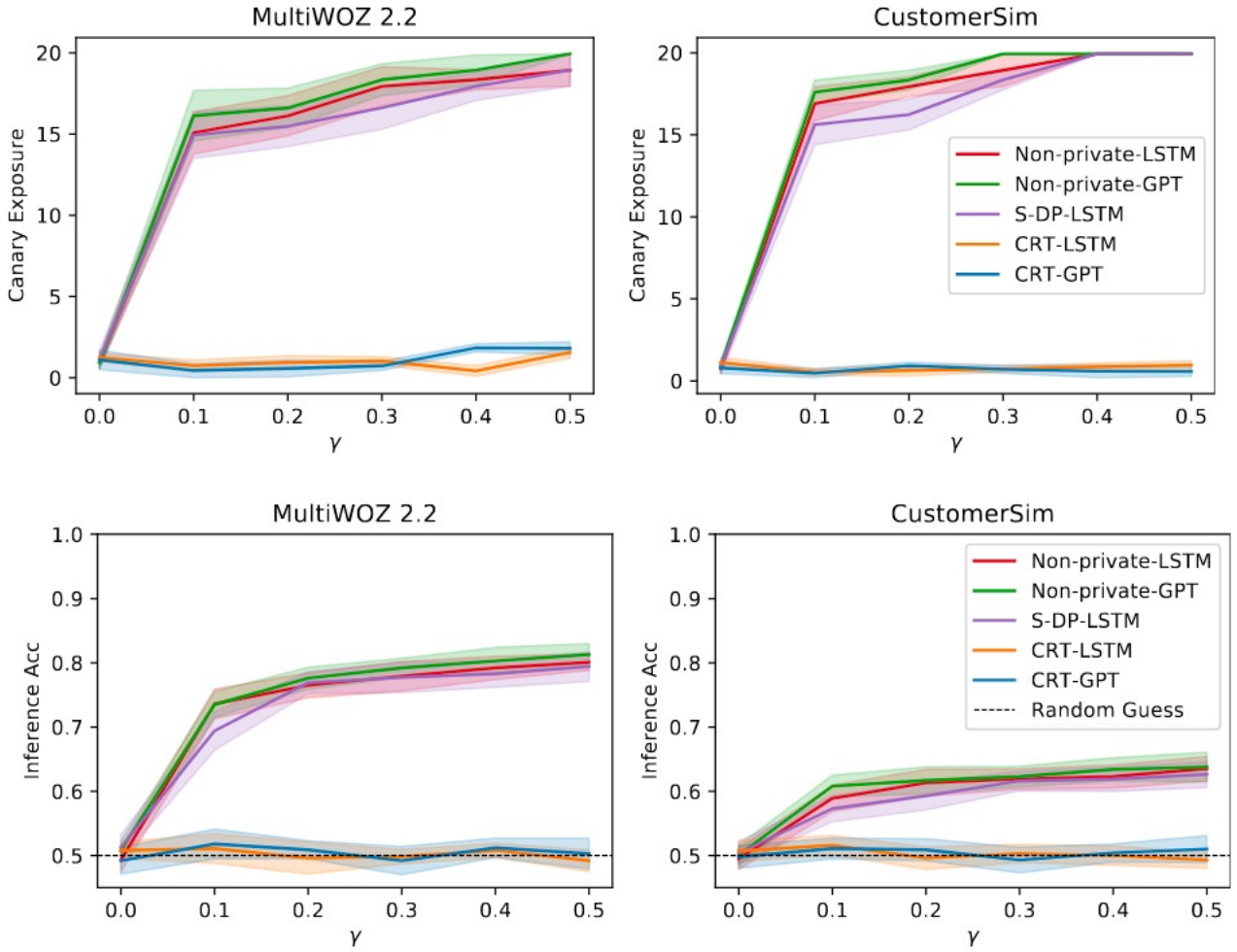
- Models: LSTM, GPT-2
- Datasets: MultiWOZ 2.2, CustomerSim

- Evaluation procedure:
  - Canary insertion attack
  - Membership inference attack

Overall Performance



Attack Results



Conclusion

- They propose confidentially redacted training (CRT), a method to train language models while protecting the secret texts.
- They introduce a new definition of confidentiality which quantifies the risk of leaking sensitive content.
- They prove the effectiveness of CRT both theoretically and empirically on multiple datasets and language models.