

INVISIBLEINK: High-Utility & Low-Cost Text Generation with Differential Privacy

Vishnu Vinod¹, Krishna Pillutla^{1,2} and Abhradeep Thakurta³

¹ Centre for Responsible AI, IIT Madras ² Wadhvani School of Data Science and AI, IIT Madras ³ Google DeepMind

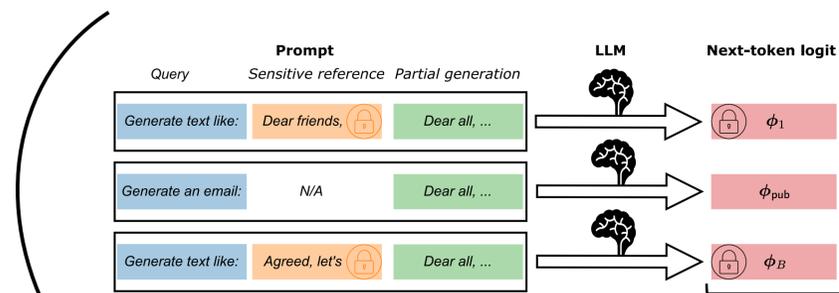


Motivation

LLMs can leak information about **sensitive references** at *inference time* (RAG, Synthetic Text Generation etc.) !!

Our Setting: White-box LLM-based differentially private synthetic text generation.

Prior SOTA: Amin et al. (EMNLP 2024) casts LLM Decoding as Exponential Mech. for DP guarantees!



Batch Size \leftrightarrow Compute Cost

batch size	127, 255, 511, 1023, 1535, 2047
------------	------------------------------------

Large batch sizes to keep sensitivity small for good quality text !!

Computationally infeasible!

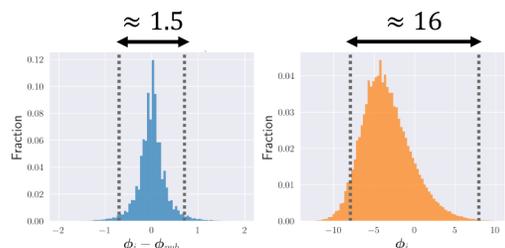
INVISIBLEINK: Difference Clipping

We isolate the sensitive info in ϕ_i by subtracting ϕ_{pub}

$$DClip_C(\phi_i, \phi_{pub}) = \phi_{pub} + clip_C(\phi_i - \phi_{pub})$$

$$PriorClip_C(\phi_i) = clip_C(\phi_i)$$

Spread of $\phi_i - \phi_{pub}$ is 10x smaller than ϕ_i !!



Ours: Clip differences
Need $C \approx 1.5$ to leave 95% logits undistorted

Prior SOTA: Clip raw logits
Need $C \approx 16$ to leave 95% logits undistorted

$$Sensitivity = \frac{C}{B}$$

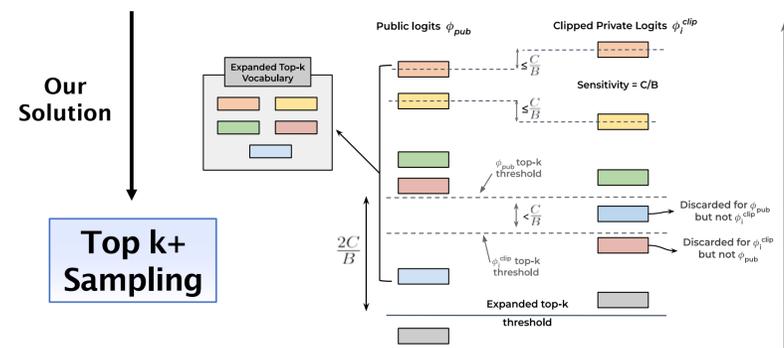
Clipping Norm \uparrow Batch Size \uparrow

INVISIBLEINK: Truncated Decoding

SOTA non-private decoding methods \Rightarrow Vocabulary Truncation (top-k/top-p)

Adapting to the private setting is non-trivial!

1. Truncating ϕ_{pub} : Loses out relevant tokens from ϕ_i
2. Truncating ϕ_i : Incurs an additional privacy cost



Captures relevant tokens from ϕ_i with no privacy cost!

Privacy Accounting

10x smaller clip norm C

10x smaller privacy budget ρ at given batch size B

10x smaller batch size B at given privacy budget ρ

zCDP analysis of Exponential Mechanism

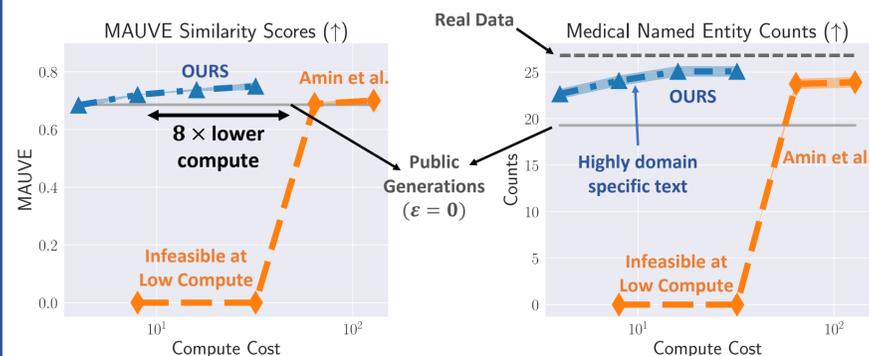
$$\rho_{seq} = T \times \frac{C^2}{2 B^2 \tau^2}$$

Adaptive Sequential Composition over T tokens Sampling temperature

Experiments

Task: Synthetic Discharge Summaries for MIMIC

INVISIBLEINK improves on prior SOTA by 8x!



White-box method

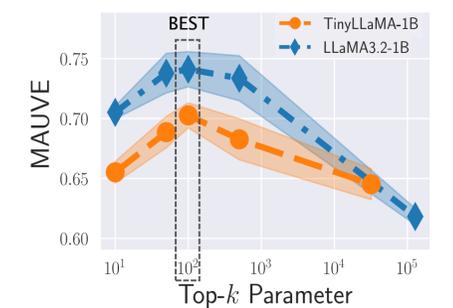
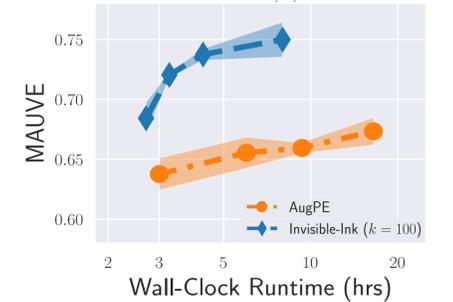
INVISIBLEINK always outperforms AugPE!

[Xie et al. (ICML 2024)]

Black-box method

Truncated Decoding is always better!

Optimal Setting
 $\tau \approx 1.0$ & $k \approx 100$



Software

pip install invink

```
import invink
output = invink.generate(ref_txt_list, hf_model_name,
                        dataset_description, num_gen, epsilon)
print(output.texts)
```

Conclusion

INVISIBLEINK generates DP synthetic text at scale at just 4 to 8 times the non-private compute cost!

Pros:

1. DP Guarantees: Non-Adaptive & Data-Independent
2. Default hyperparameters for off-the-shelf usage

References:

Amin et al., Private prediction for large-scale synthetic text generation, EMNLP 2024

Xie et al., Differentially Private Synthetic Data via Foundation Model APIs 2: Text, ICML 2024

Check out our python package and paper!

ArXiv Software

