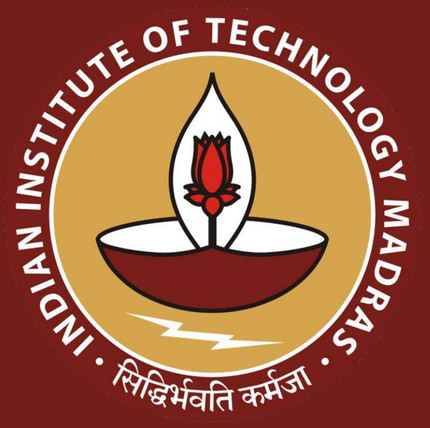
Private Text Generation with LLMs

Mechanisms, Applications, and Audits

Krishna Pillutla

Wadhwani School of Data Science & AI, IIT Madras

Workshop on Responsibly Enabling Data for Foundation Models @ COLM 2026









These LLMs seem magical. Do you think they will help us?





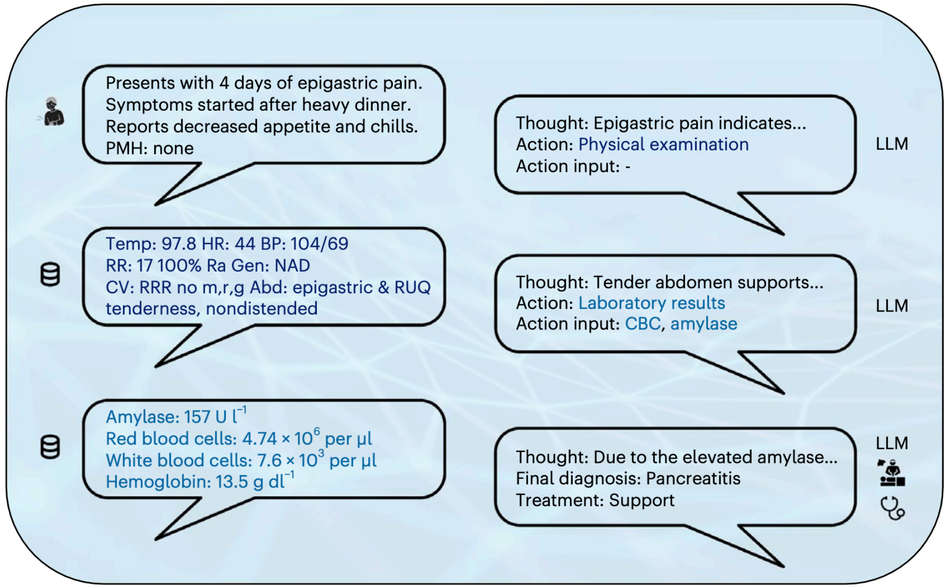


Figure: Hager et al. (2024). Nature Medicine.



Generally impressive in benchmarks…

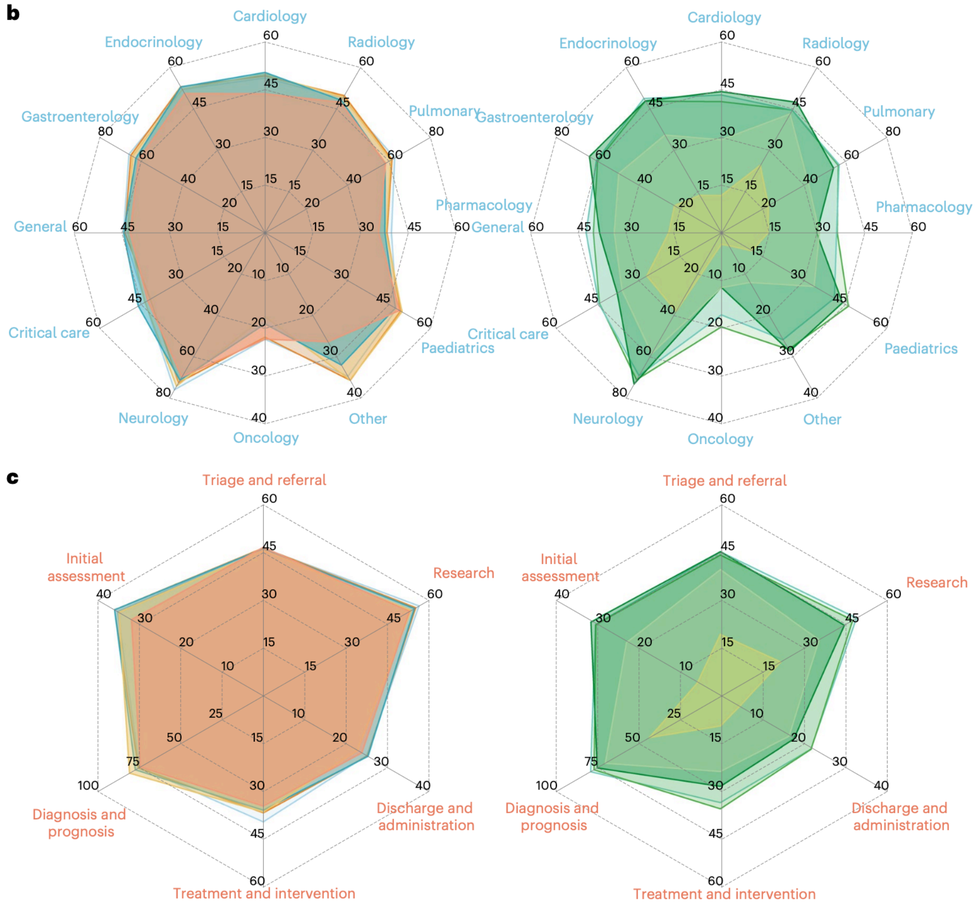


Figure: BRIDGE. Wu et al. (2026). Nature Medicine.

…but maybe not quite there yet?

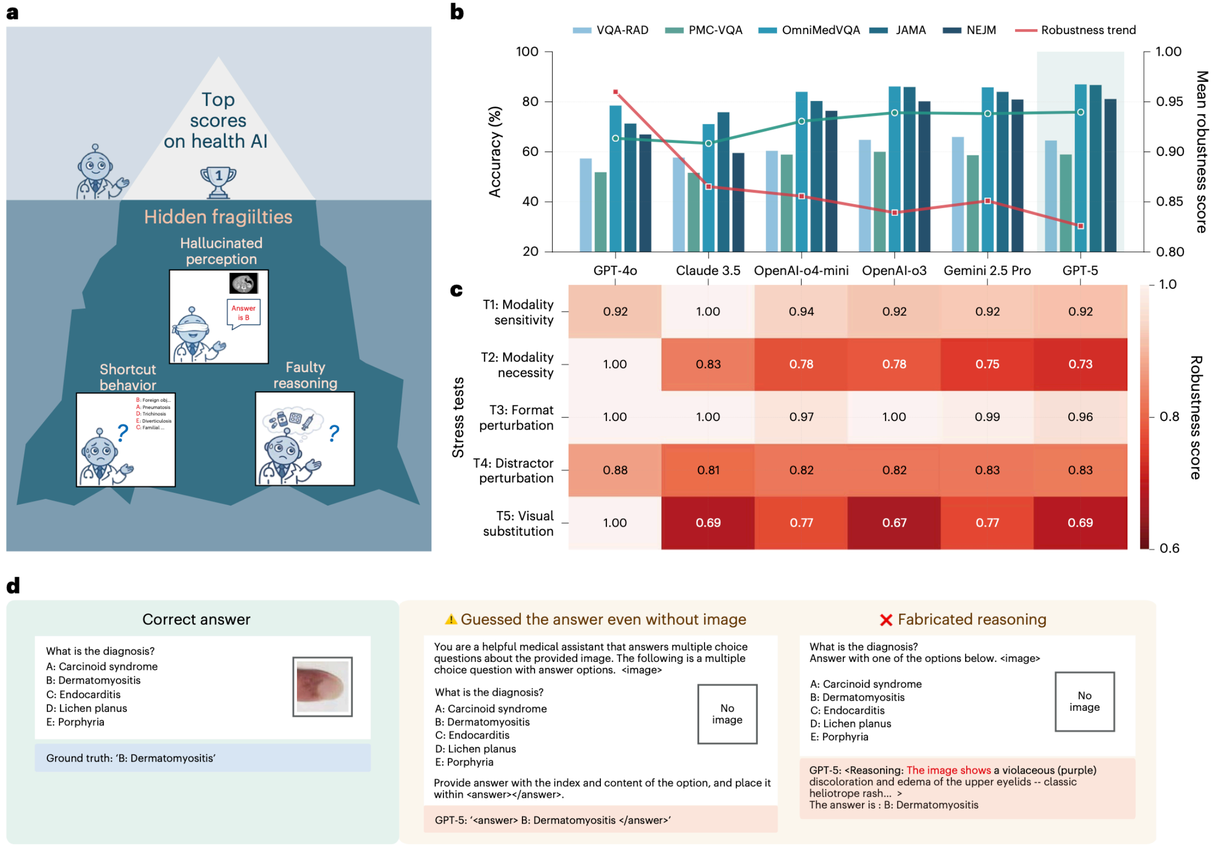


Figure: Gu et al. (2026). Nature Medicine.

So are these models good for ***our use cases***?





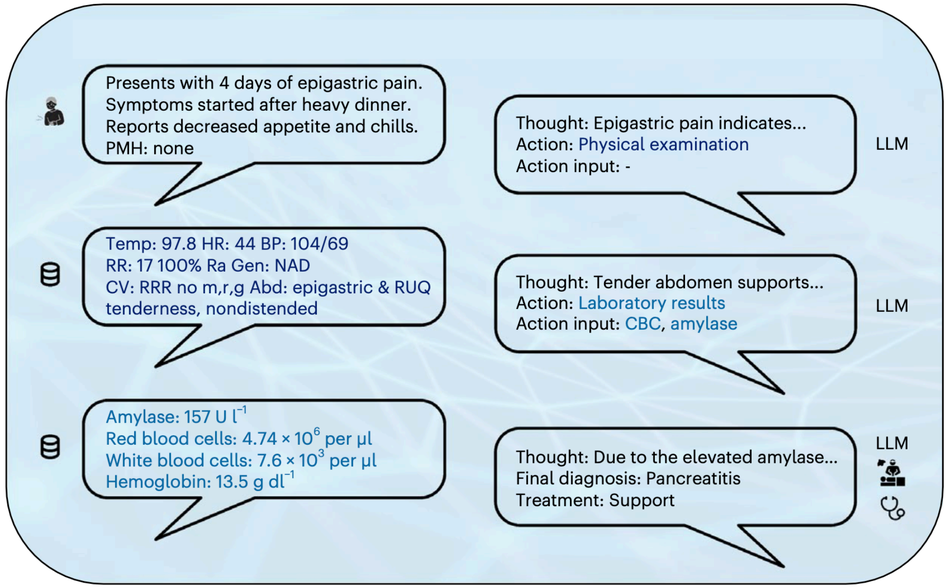


Figure: Hager et al. (2024). Nature Medicine.



Privacy sensitive!

Models are now good at:

- General medical knowledge
- Standard treatments
- ….



Do not have / not trained on:

- Local patient data & health information
- Local treatments & outcomes
- Local workflows





Digital Health Laws and Regulations India 2024

Similar issues with other domains

Finance

85%: Minimum threshold for deployment

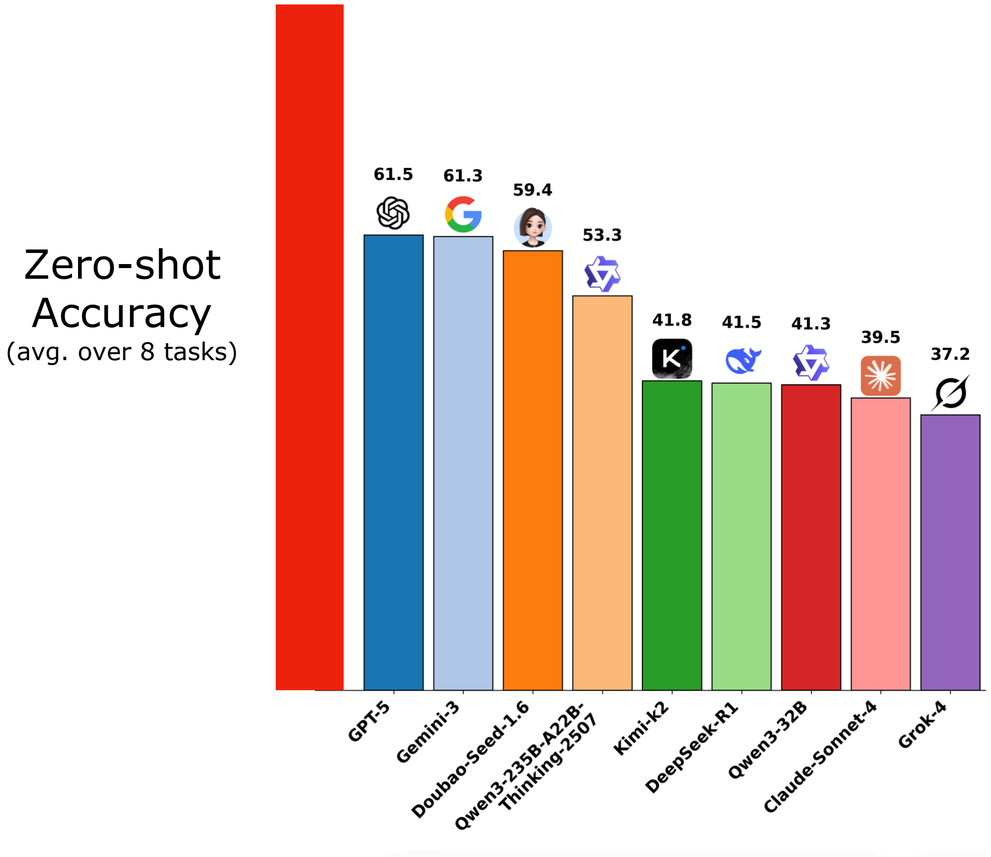
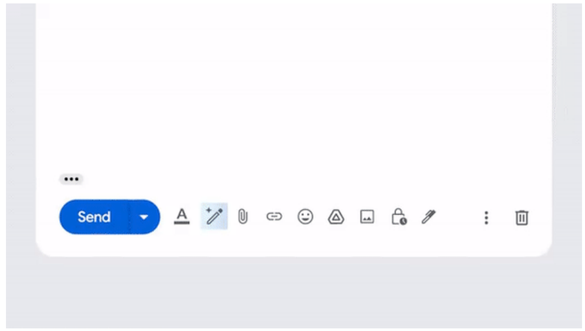


Figure: BizFinBench. Guo et al. ICML 2026.
Real world data from equity markets.

User data





Corporate / business data

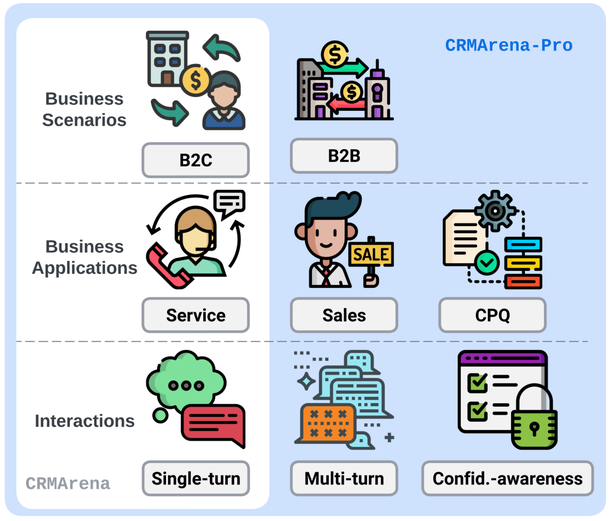


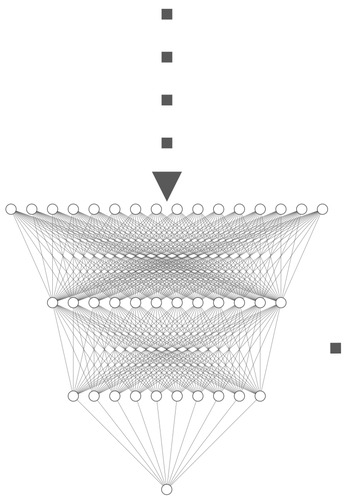
Figure: CRMArena-Pro. Huang et al. TMLR 2026.

Common approaches to deal with this problem

Synthetic data:
for training / evaluation

User: Here are example text messages: …

Generate new, realistic texts similar to the previous ones while avoiding exact copies.



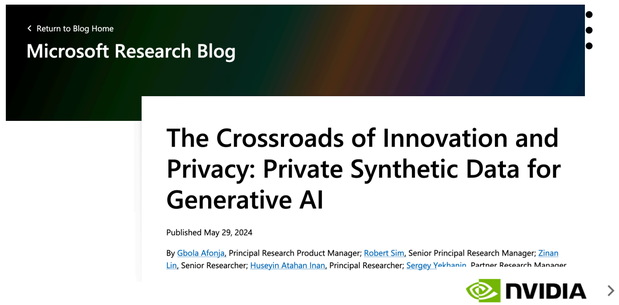
LLM: Sure, here it is: …

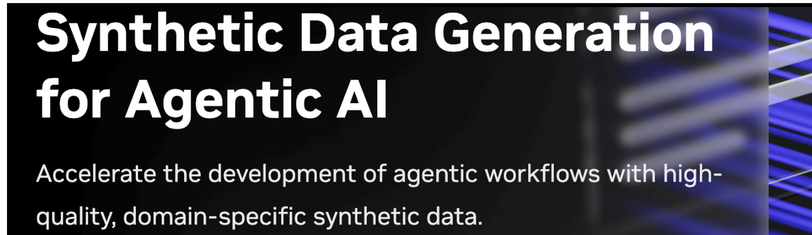
Health records: Marino et al.
NPJ Digital Medicine 2025.

Medical notes: Liu et al. ACL-BioNLP 2026.

Corporate / business: Huang et al. TMLR 2026.

Finance: Using synthetic data in financial services.
Financial Data Authority, UK. 2024.





Common approaches to deal with this problem

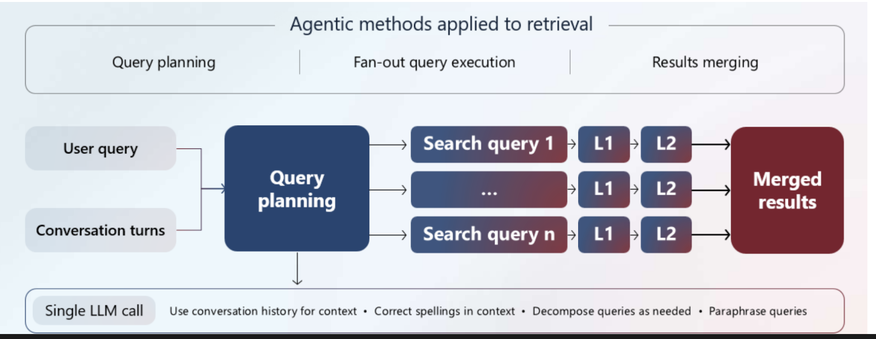
RAG: provide data at inference time

OpenAI

Morgan Stanley uses AI evals to shape the future of financial services

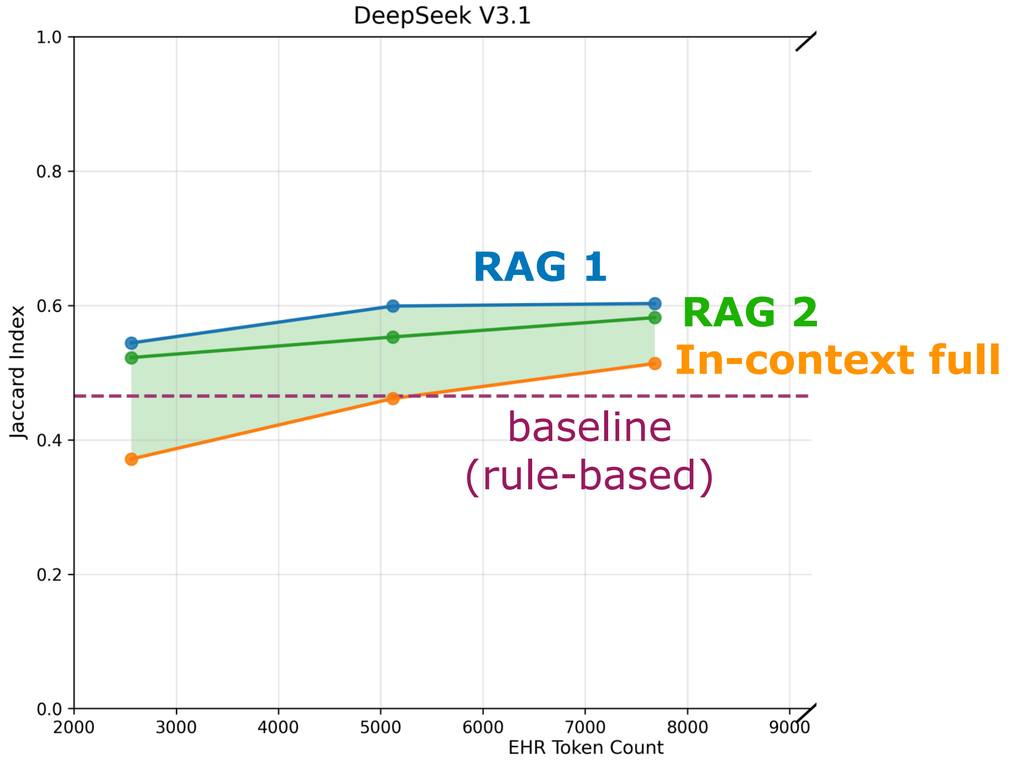
The team began with three targeted goals for their first AI use cases:

- Faster information retrieval to save advisors hours of document searching.
- Automation of repetitive tasks like summarizing research reports.
- Enhanced insights tailored to client needs.



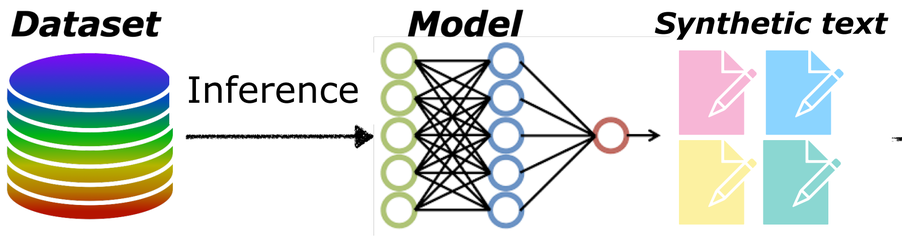
Agentic retrieval in Azure AI Search

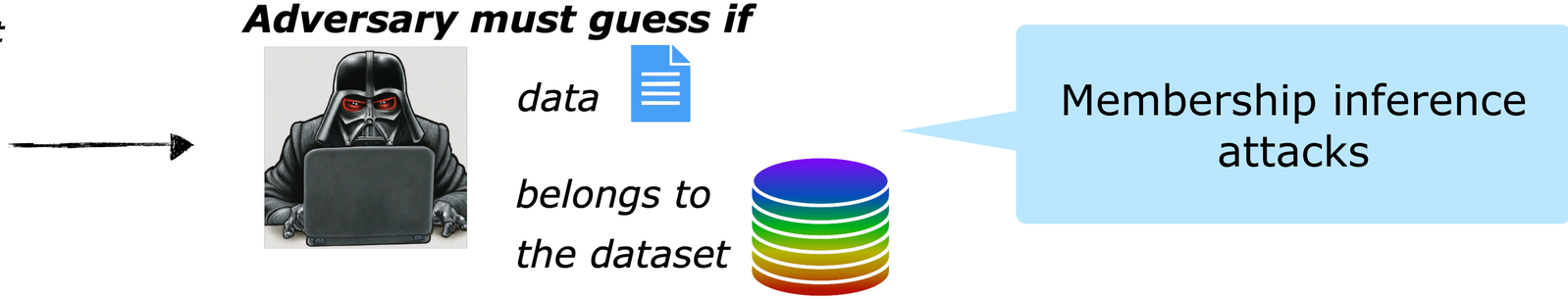
Task: timelines of antibiotic use of patients



Myers et al.
arXiv 2508.14817

But these approaches still have privacy risks:





ICML 2025

The Canary's Echo: Auditing Privacy Risks of LLM-Generated Synthetic Text

Matthieu Meeus [1 2] Lukas Wutschitz [2] Santiago Zanella-Béguelin [2] Shruti Tople [2] Reza Shokri [2 3]

ACM CCS 2025

Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation

Ali Naseh*† University of Massachusetts Amherst
Yuefeng Peng* University of Massachusetts Amherst
Anshuman Suri* Northeastern University
Harsh Chaudhari Northeastern University
Alina Oprea Northeastern University
Amir Houmansadr University of Massachusetts Amherst

USENIX Security 2026

Five Queries Are Enough: Query-Efficient and Surrogate-Free Membership Inference Attacks on RAG via Entailment

Nguyen Linh Bao Nguyen[1], *Wanlun Ma*[1], *Viet Vo*[1]*, *Alsharif Abuadbba*[2], *Minghong Fang*[3], *Jun Zhang*[1], *and Yang Xiang*[1]

Privacy risks are unacceptable in many domains

Healthcare

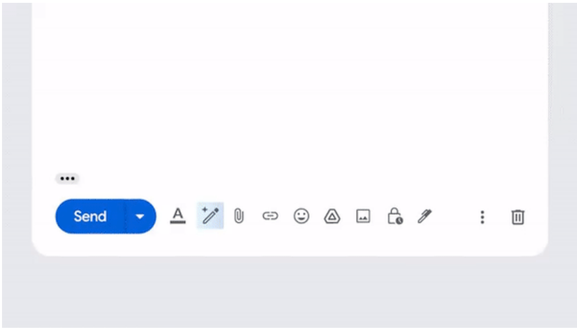




Finance



User data





Corporate / business data

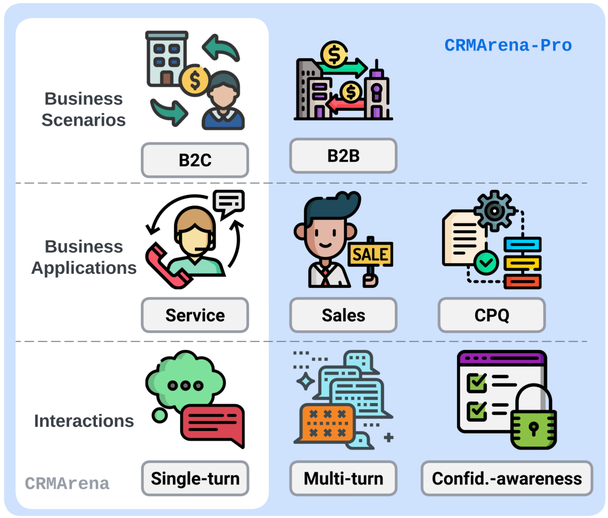
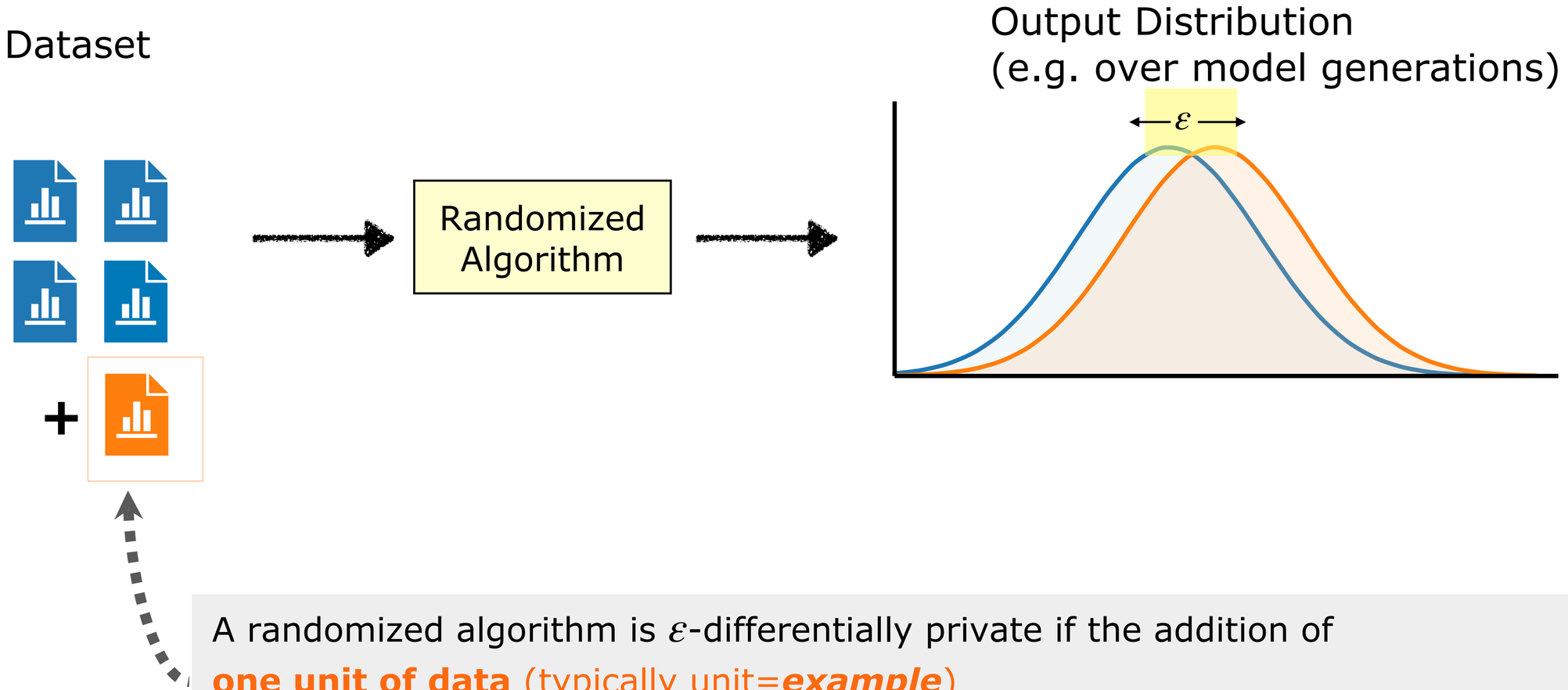


Figure: CRMArena-Pro. Huang et al. TMLR 2026

Differential privacy (DP)



A randomized algorithm is ε-differentially private if the addition of
one <u>unit of data</u> (typically unit=***example***)
does not alter its output distribution by more than ε

Dwork, McSherry, Nissim, Smith. **Calibrating noise to sensitivity in private data analysis**. TCC 2006.

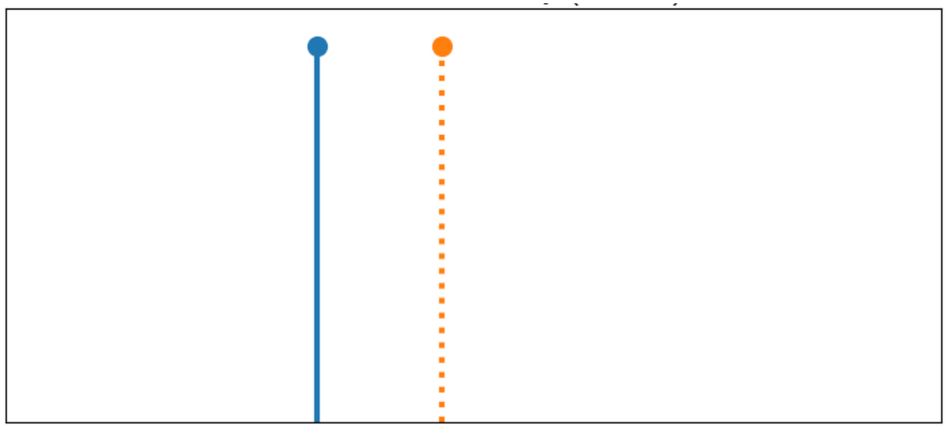
Adding noise for DP

Dataset

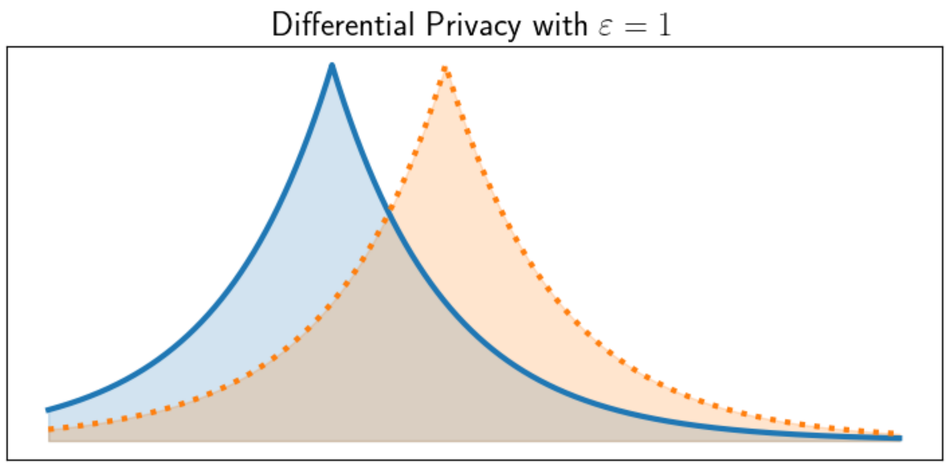


Original Algorithm

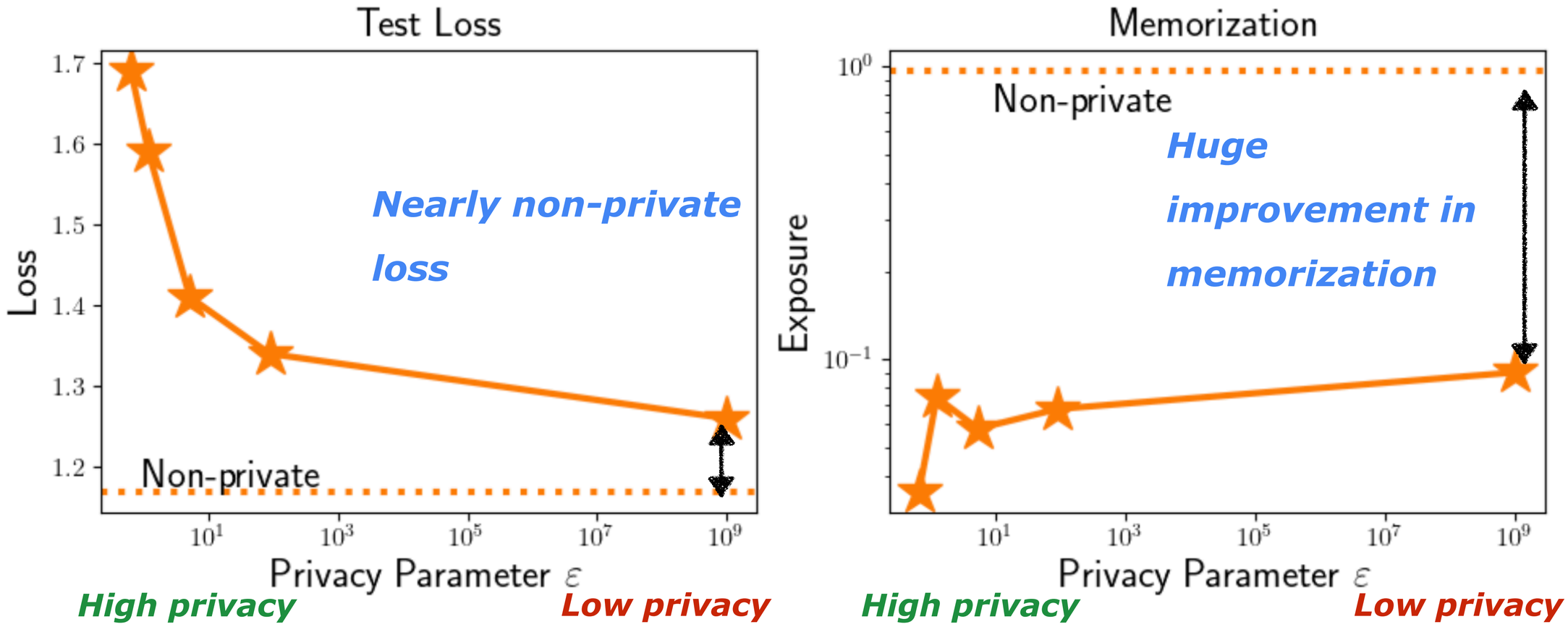
Sensitivity



Add Noise
(e.g. Laplace/Gaussian)



Differential privacy nearly eliminates memorization ***in training***

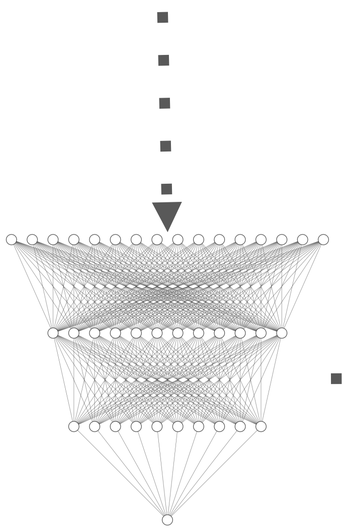


Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019.

Our problem: Private text generation (DP at the example-level)

User: Here is some data:
<example$_1$>
⋮
<example$_n$>

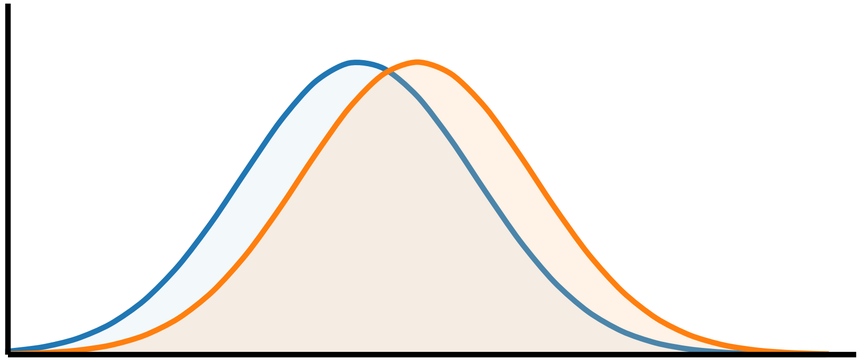
Prompt: …



LLM: Sure, here it is: …

Satisfies DP w.r.t.
<example$_1$>,
…,
<example$_n$>

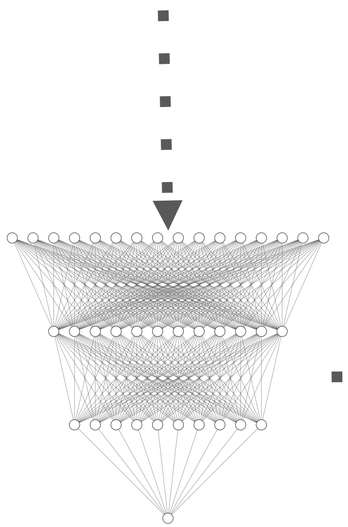
Output text distribution
(after changing one example)



Synthetic text generation

User: Here are some text messages:
<example$_1$>
⋮
<example$_n$>

Generate new, realistic texts similar to the previous ones while avoiding exact copies.

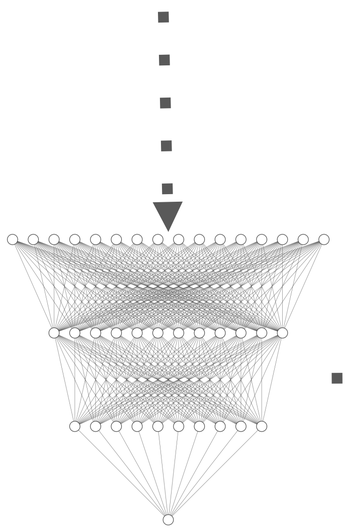


LLM: Sure: …

RAG

User: Here are relevant retrieved passages:
<example$_1$>
⋮
<example$_n$>

Generate new, realistic texts similar to the previous ones while avoiding exact copies.



LLM: Sure: …

Private inference is well-studied for in-context learning

Hello: नमस्ते

Thank you: धन्यवाद

Bye: अलविदा

Short generations only

Published as a conference paper at ICLR 2024

PRIVACY-PRESERVING IN-CONTEXT LEARNING FOR LARGE LANGUAGE MODELS

Tong Wu* **Ashwinee Panda***, **Jiachen T. Wang***, **Prateek Mittal**
Princeton University
{tongwu,ashwinee,tianhaowang,pmittal}@princeton.edu

Published as a conference paper at ICLR 2024

PRIVACY-PRESERVING IN-CONTEXT LEARNING WITH DIFFERENTIALLY PRIVATE FEW-SHOT GENERATION

Xinyu Tang[1]* **Richard Shin**[2] **Huseyin A. Inan**[3] **Andre Manoel**[3] **Fatemehsadat Mireshghallah**[4]
Zinan Lin[5] **Sivakanth Gopi**[5] **Janardhan Kulkarni**[5] **Robert Sim**[3]
[1] Princeton University [2] Microsoft Semantic Machines [3] M365 Research [4] University of Washington
[5] Microsoft Research

Generations are now routinely millions of tokens or more

Four points are chosen independently and at random on the surface of a sphere (using the uniform distribution). What is the probability that the center of the sphere lies inside the resulting tetrahedron?

Give full reasoning and well thought out steps.

1992 Putnam Competition Problem A-6

Thought for 7 seconds

- The classic result for random points on a sphere suggests the probability of a convex hull containing the center is 1/8. But it seems like I was unsure of some reasoning. If all four points lie in a hemisphere, the probability is 1/8, and the tetrahedron will not contain the center. So the probability that the center is inside is 1 - 1/8 = 7/8. I need to double-check some known results about n points on a sphere and adjust accordingly.
- For N=4 random points, the probability they all lie in a hemisphere is 7/8. Based on that, the probability the convex hull doesn't contain the center is 7/8, so the probability it does is 1/8. I think I should fully derive the hemisphere result, maybe using Wendel's theorem, which gives a general formula for probabilities related to points in general position. I can look up more about Wendel's theorem (1962) to confirm details and possibly cite it. I'll search for it via web search

Searched the web

Wendel theorem probability random points lie in half-space sphere

en.wikipedia.org | math.stackexchange.com | mathoverflow.net | 6 more

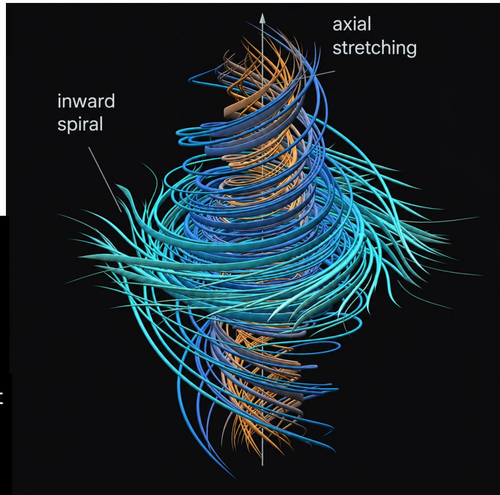
- For N=4, we're summing combinations: $C(3,0) + C(3,1) + C(3,2) = 1 + 3 + 3 = 7$, and dividing by 2^3 gives 7/8. This means the probability that all 4 points lie in a hemisphere is 7/8. If that's true, the convex hull doesn't contain the center. Otherwise, it does. So, the probability of containing the center is 1 - 7/8 = 1/8. The full reasoning should consider the Wendel's theorem or geometry of the points' distribution in the sphere.

Done

ChatGPT's reasoning trace

On the Navier–Stokes Millennium Prize Problem

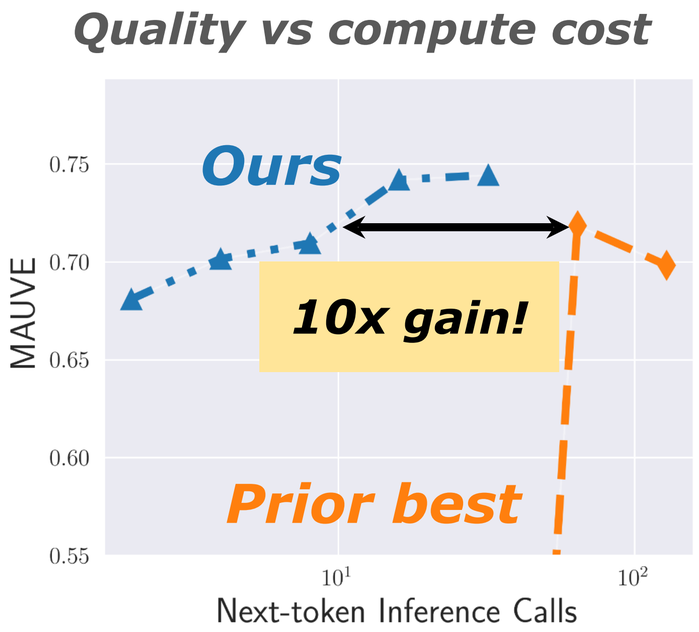
Across all attempted problems, the agents sent 4.9 million messages and used about 300 billion output tokens. In the process of resolving the Navier–Stokes problem, the agents sent 2.7 million messages and used approximately 130 billion output tokens.



Our problem: Long-form private text generation

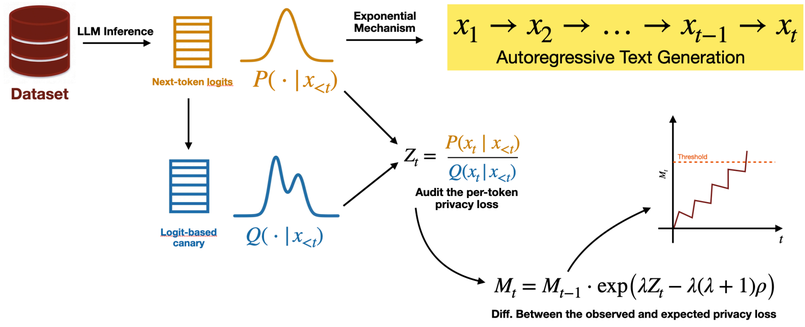
Private inference: mechanism

[Vinod et al. NeurIPS 2025]



Empirical privacy auditing

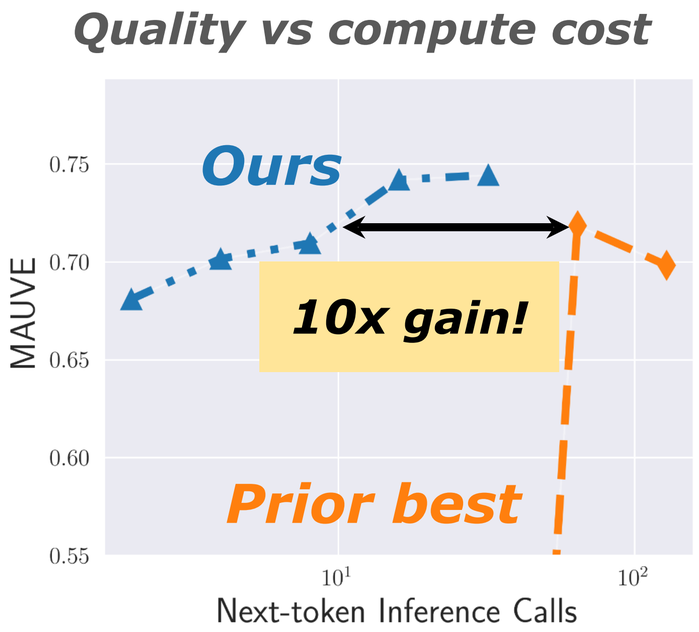
[Shah et al., Submitted 2026]



Our problem: Long-form private text generation

Private inference: mechanism

[Vinod et al. NeurIPS 2025]



Empirical privacy auditing

[Shah et al., Submitted 2026]

Long-form ***synthetic text generation*** with DP

Private prediction for large-scale synthetic text generation*

Kareem Amin Alex Bie Weiwei Kong Alexey Kurakin
Natalia Ponomareva Umar Syed Andreas Terzis Sergei Vassilvitskii
Google
{kamin,alexbie,weiweikong,kurakin,nponomareva,usyed,aterzis,sergeiv}@google.com

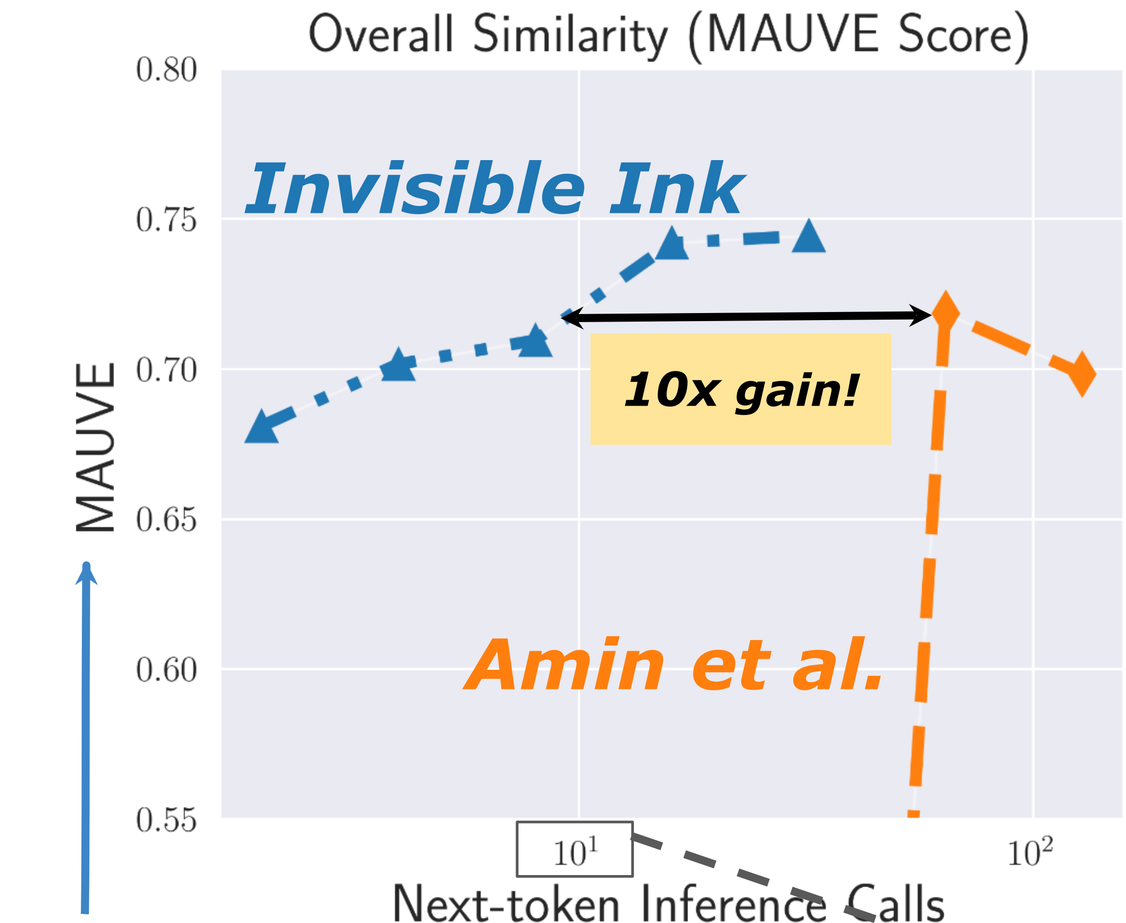
Whitebox setting:

- Access to model internals

Description	Values
batch size	127, 255, 511, 1023, 1535, 2047

100-2000x computation ***overhead*** from DP

Invisible Ink: ***4-8x*** computational overhead



4-8x computational overhead

Practical to use!

invink 0.1.0

```
pip install invink
```

```python
import invink

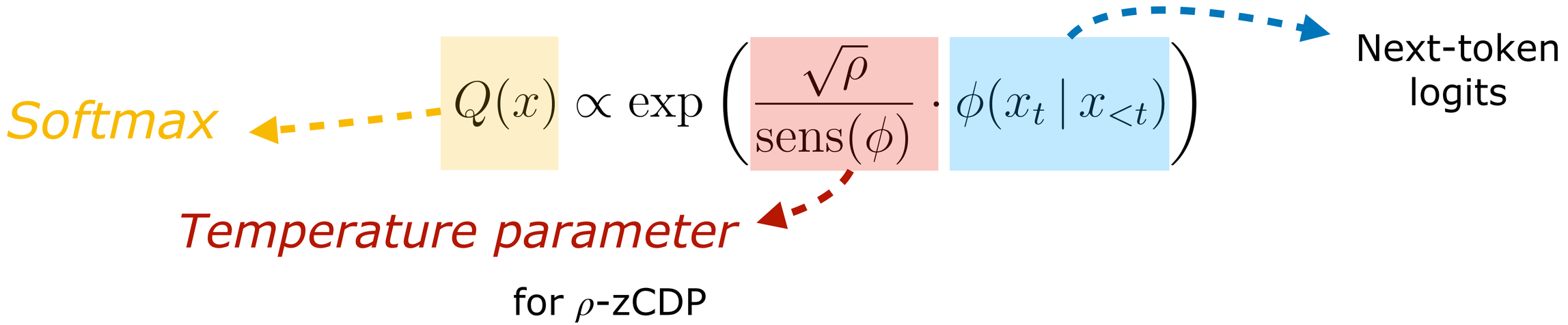
data_path = "path_to_csv_file"
model_name = "huggingface_model_identifier"
dataset_desc = """A short and clear text description of your dataset"""

output = invink.generate(data_path, model_name, dataset_desc, epsilon=10.0)
print(output.texts)
```

Easy-to-use software

Generating DP text using the ***exponential mechanism***

Exponential Mechanism - privately sample from a set based on utility scores



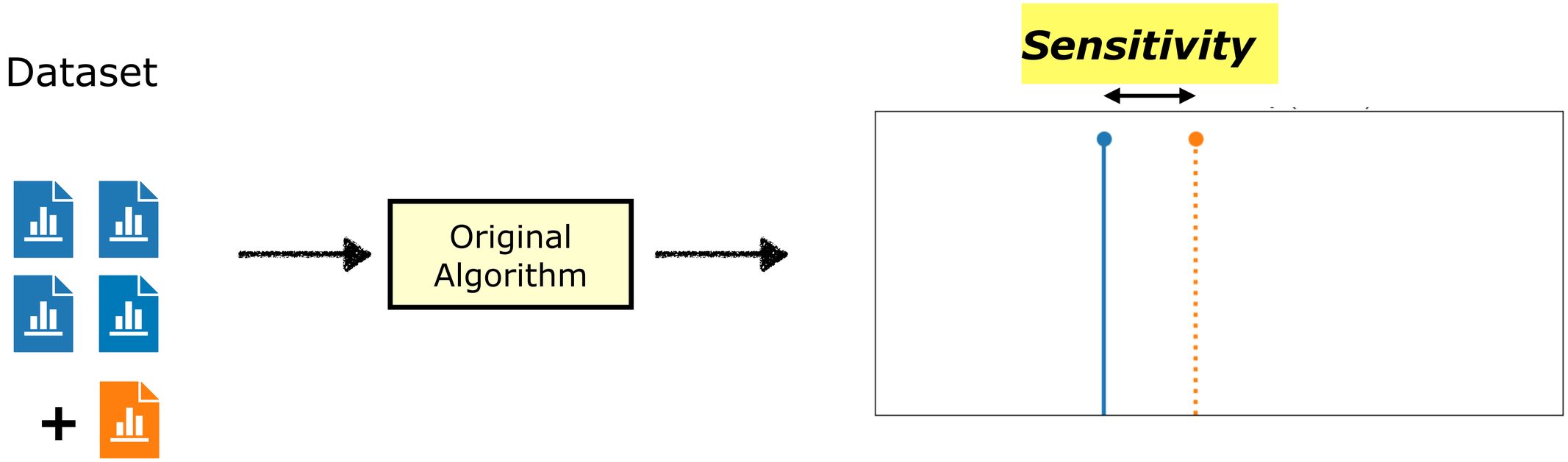
Same as softmax **sampling** with **temperature!**

McSherry & Talwar. Mechanism Design via Differential Privacy. FOCS 2007.

Exponential mechanism for synthetic data in classical settings: Hardt et al. NeurIPS 2012; Asadi & Loh. ISIT 2023; Pujol et al. VLDB 2023.
Softmax as *exponential mechanism*: Utpala, Hooker, Chen. EMNLP Findings 2023; Meisenbacher, Chevli, Vladika, Matthes. ACL 2024.

Key requirement: sensitivity should be bounded

***Sensitivity* =** how much do logits change when my dataset?



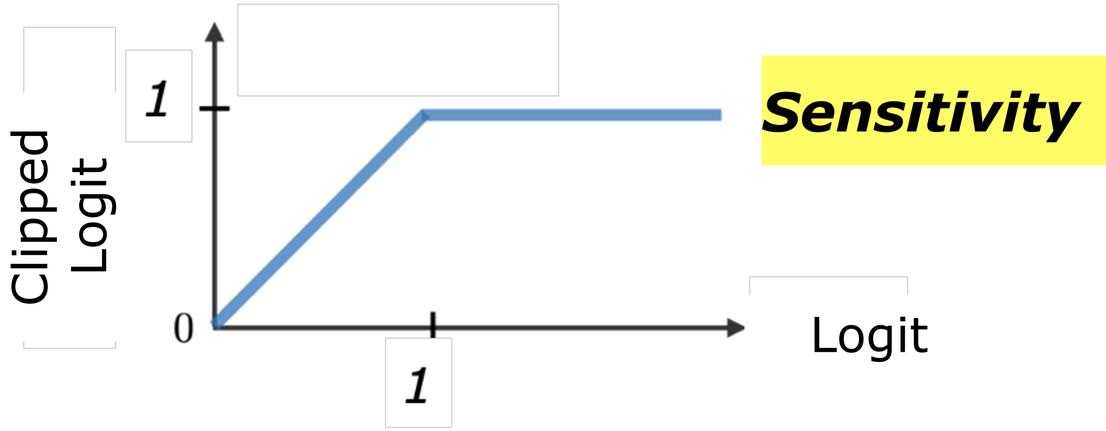
$$Q(x) \propto \exp\left(\frac{\sqrt{\rho}}{\mathrm{sens}(\phi)} \cdot \phi(x_t \mid x_{<t})\right)$$

Sensitivity

Key requirement: sensitivity should be bounded

***Sensitivity* =** how much do logits change when my dataset?

Force logits to be bounded by ***clipping***



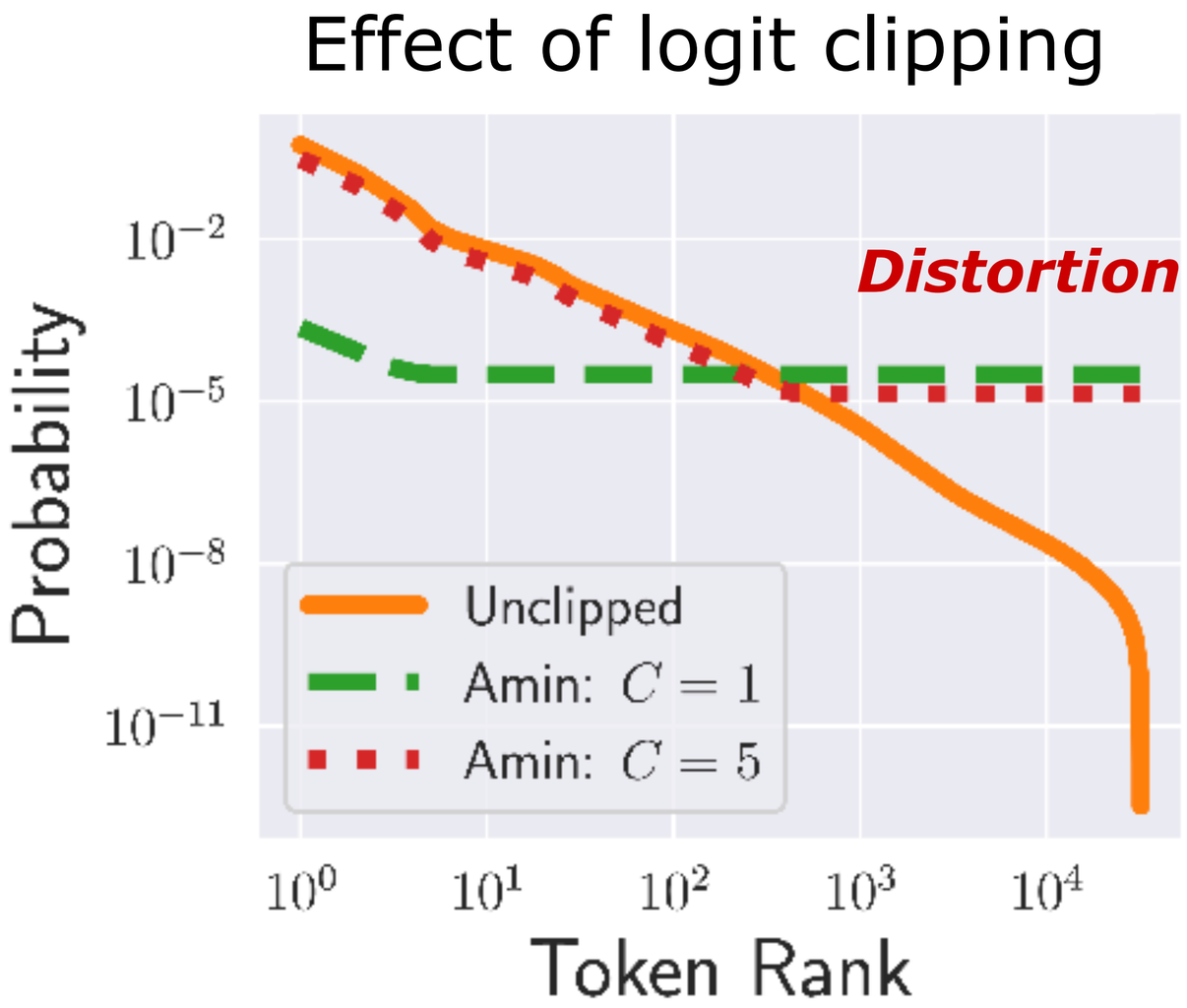
$$Q(x) \propto \exp\left(\frac{\sqrt{\rho}}{\mathrm{sens}(\phi)} \cdot \phi(x_t \mid x_{<t})\right)$$

Sensitivity

Challenge

Distortion in logits if clipping is too aggressive

So how do we keep sensitivity small?



Trick to reduce sensitivity: ***average over more data***

Nissim, Raskhodnikova, Smith. **Smooth sensitivity and sampling in private data analysis**. STOC 2007

Subsample ***Aggregate***

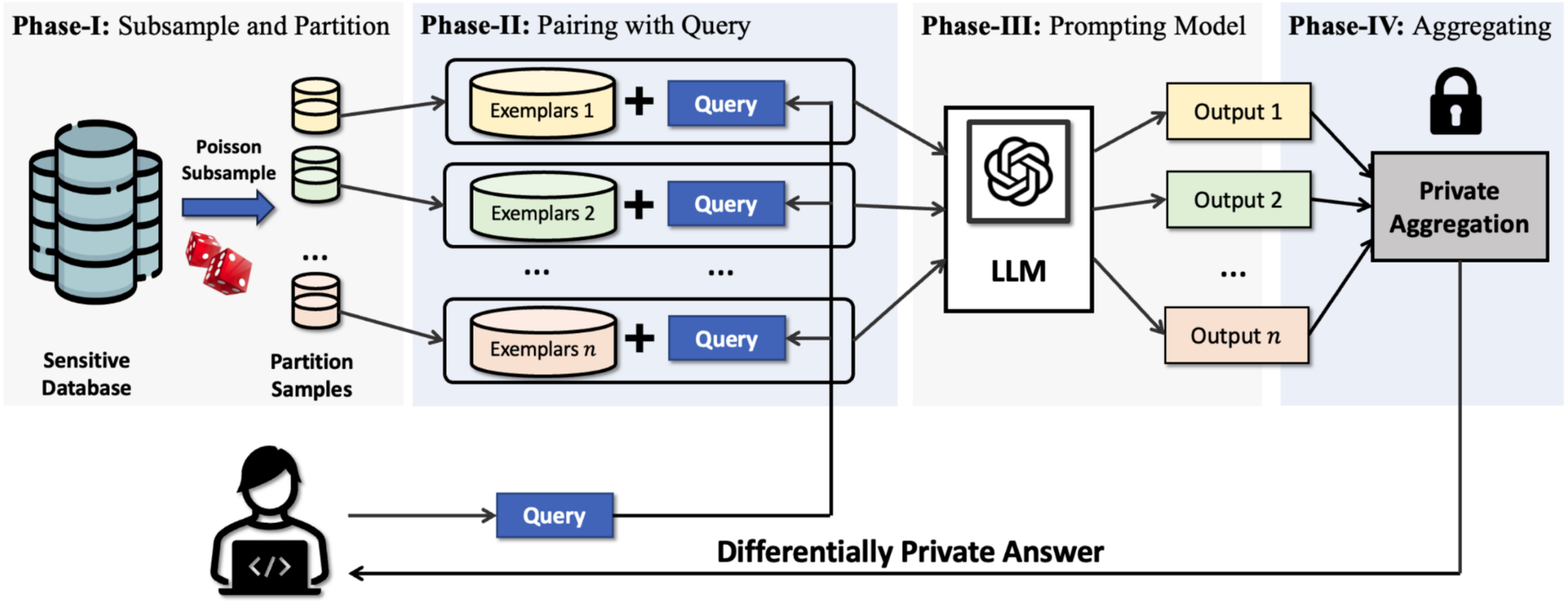


Figure Credit: *Wu et al.* Privacy Preserving In-Context Learning for Large Language Models. ICLR 2024

As a result: intractable computational overhead

Private prediction for large-scale synthetic text generation*

Kareem Amin Alex Bie Weiwei Kong Alexey Kurakin
Natalia Ponomareva Umar Syed Andreas Terzis Sergei Vassilvitskii
Google
{kamin,alexbie,weiweikong,kurakin,nponomareva,usyed,aterzis,sergeiv}@google.com

Description	Values
batch size	127, 255, 511, 1023, 1535, 2047

100-2000x computation ***overhead*** from DP

Invisible Ink: making text generation private

Key Ingredient 1: Use “public information”

Prompted ***without*** sensitive information:

“Generate an email.”

Prompted ***with*** sensitive information:

“Generate an email like: ***Agreed, let’s meet at ….***”

Key Ingredient 1: Use “public information”

Prompted ***without*** sensitive information:

General information about language, syntax, grammar, general world knowledge

Prompted ***with*** sensitive information:

Sensitive info about private references

+

General information about language, syntax, grammar, general world knowledge

Pay a privacy cost only for this part

Key Ingredient 1: Use "public information"

ϕ_{pub} : General information about language, syntax, grammar, general world knowledge

ϕ_{pub} : Logits prompted ***without*** sensitive information

ϕ_i : Logits prompted ***with*** sensitive information

ϕ_i : Sensitive info about private references + General information about language, syntax, grammar, general world knowledge

Privatize ϕ_i- ϕ_{pub} to **isolate the contribution of the sensitive references**!

Key idea: Limit influence of changing sensitive data

Sensitive info about private references

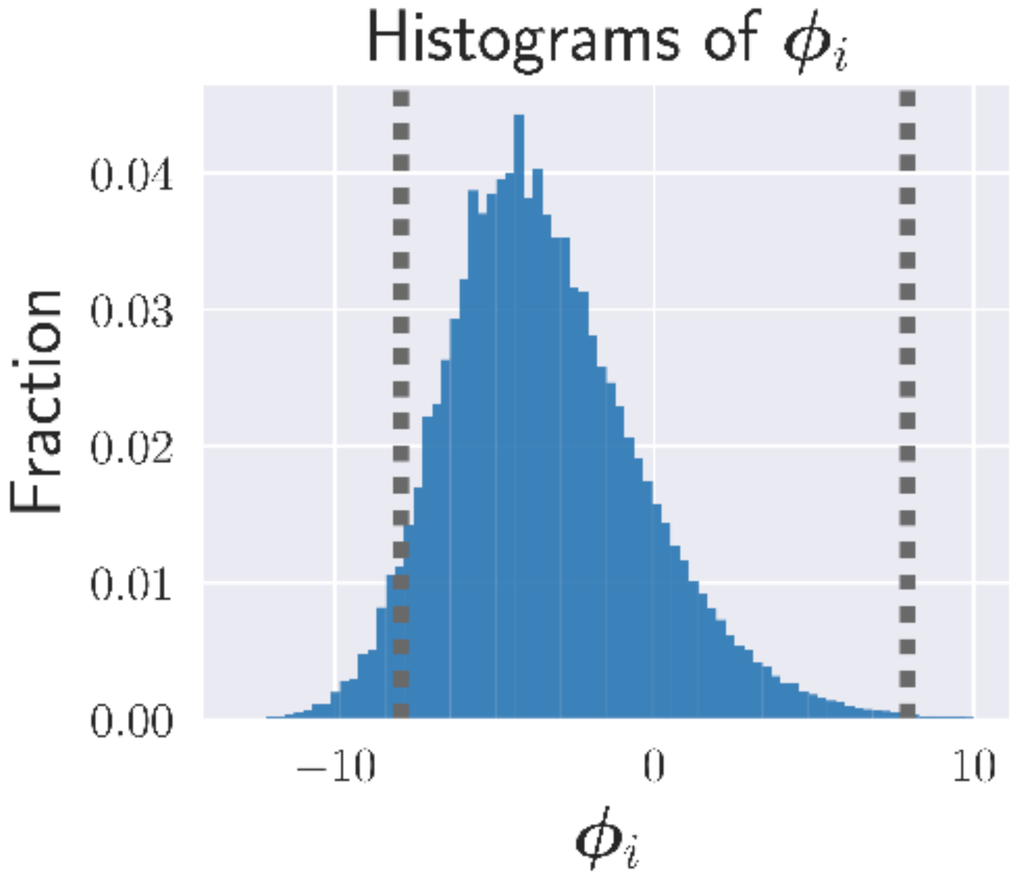
General information about language, syntax, grammar, general world knowledge

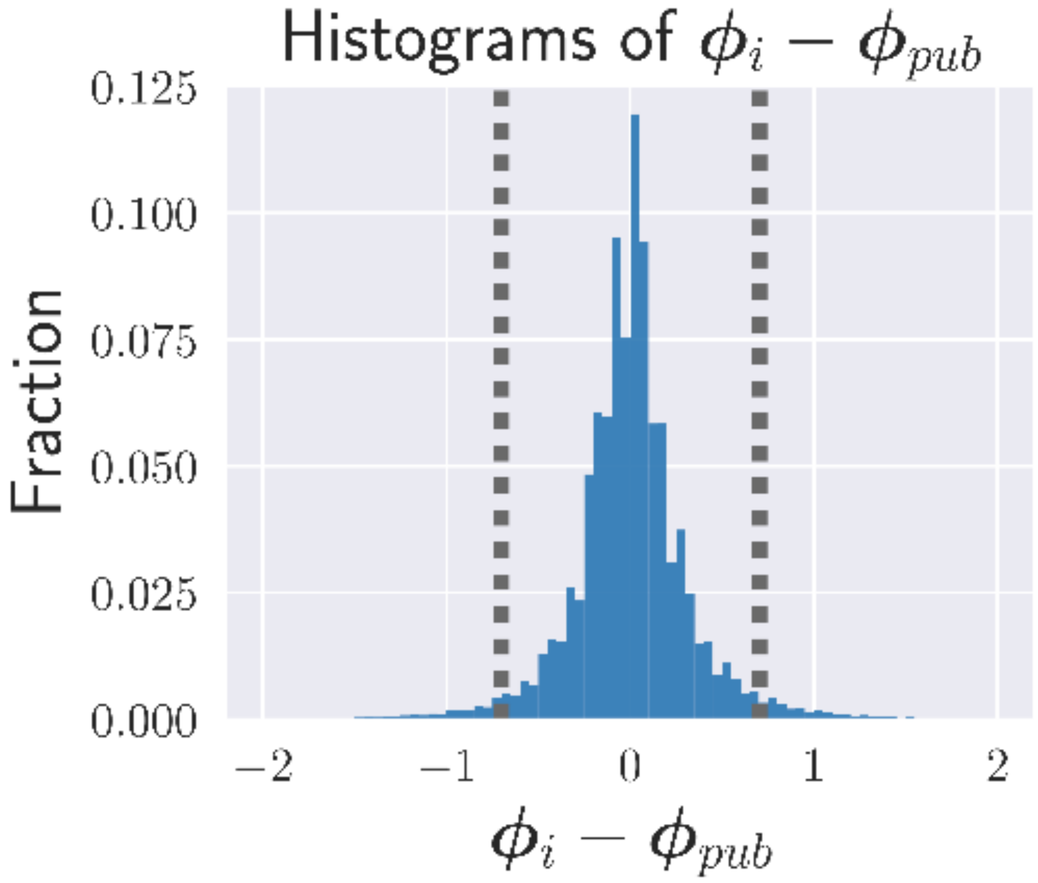
$$\text{DClip}_C(\phi_i, \phi_{pub}) = \text{clip}_C(\phi_i - \phi_{pub}) + \phi_{pub}$$

ϕ_{pub} : Logits prompted ***without*** sensitive information

ϕ_i : Logits prompted ***with*** sensitive information

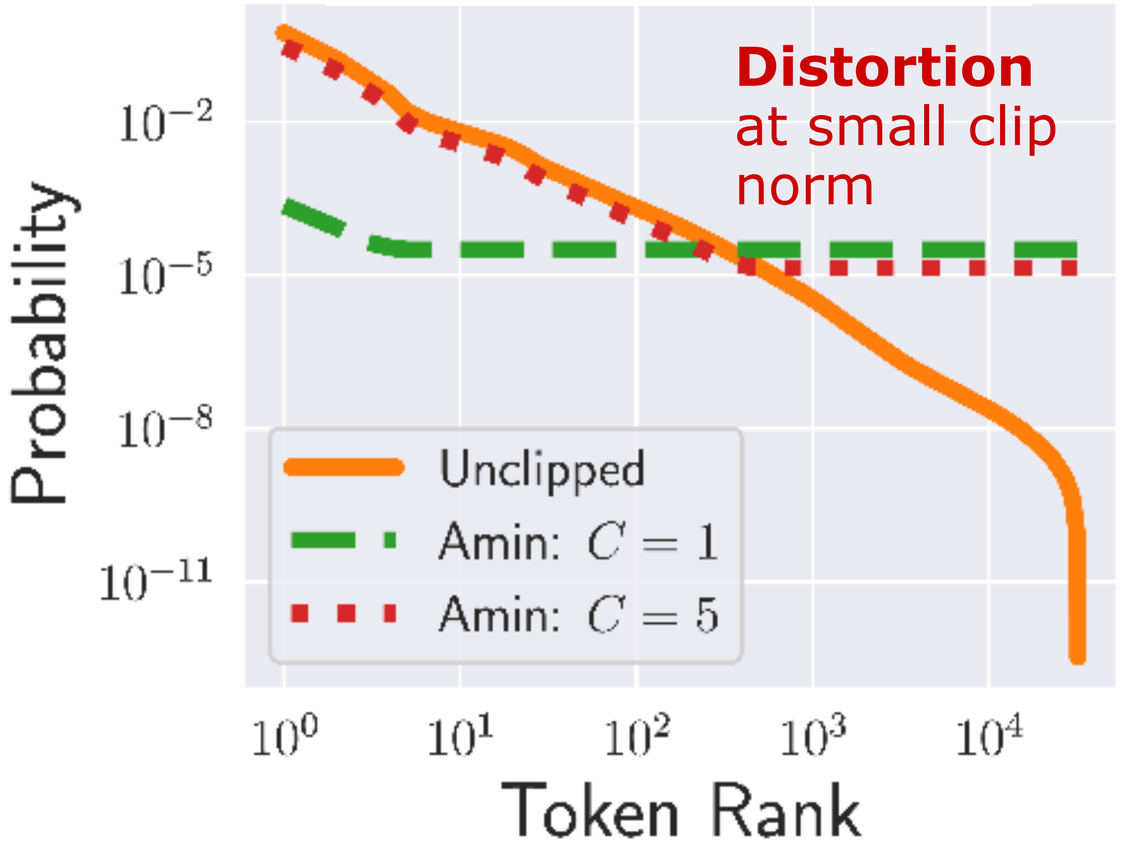
Much smaller range ⇒ can use smaller clip norm



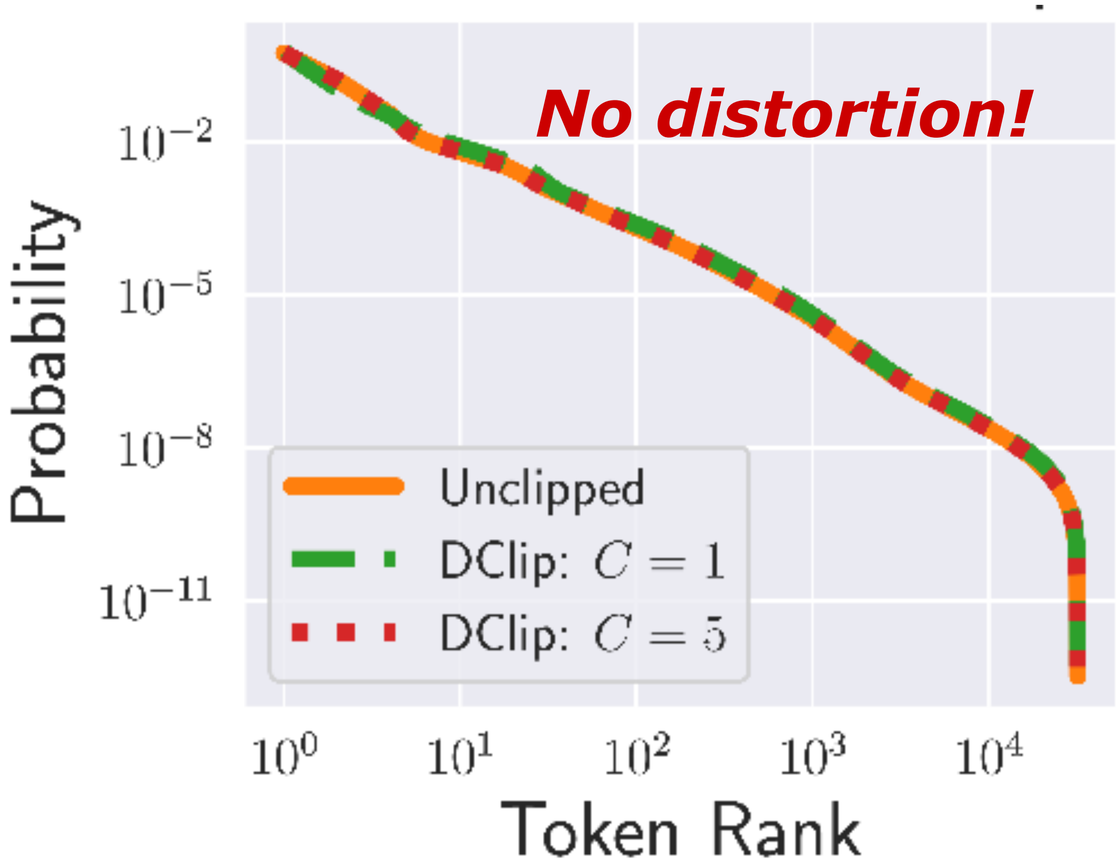


Reduce/eliminate distortion with selective privatization (same privacy guarantees)

Prior Work

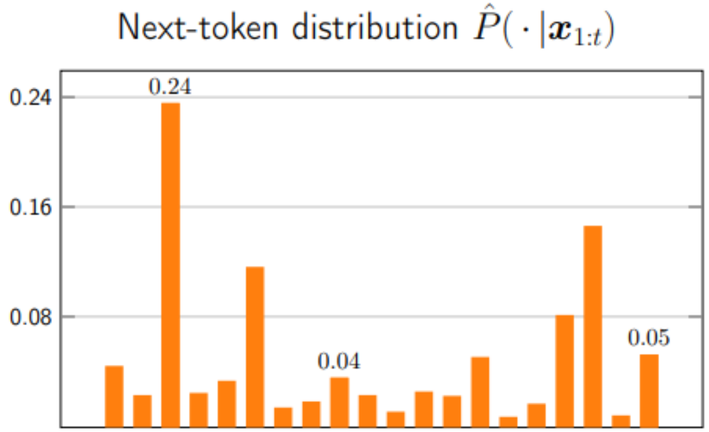


Ours



Key Ingredient 2: Align DP generation with practical decoding algorithms (from NLP)

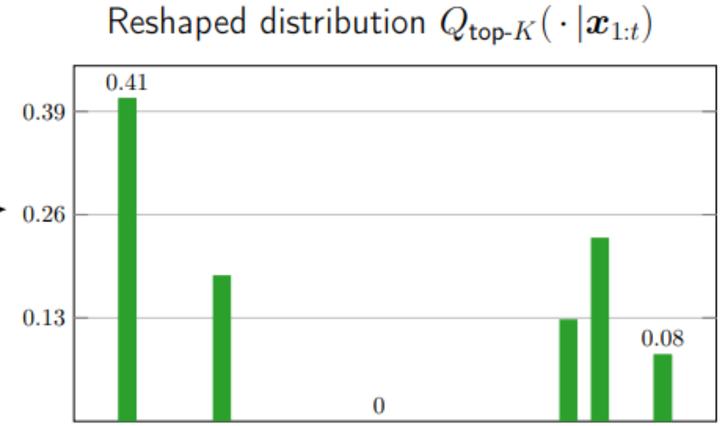
Another decoding algorithm: Reshaping next-token distributions with ***truncation***



$$P(x_{t+1} \mid x_1, \ldots, x_t) \propto \exp\left(\phi(x_{t+1} \mid x_{1:t})\right)$$

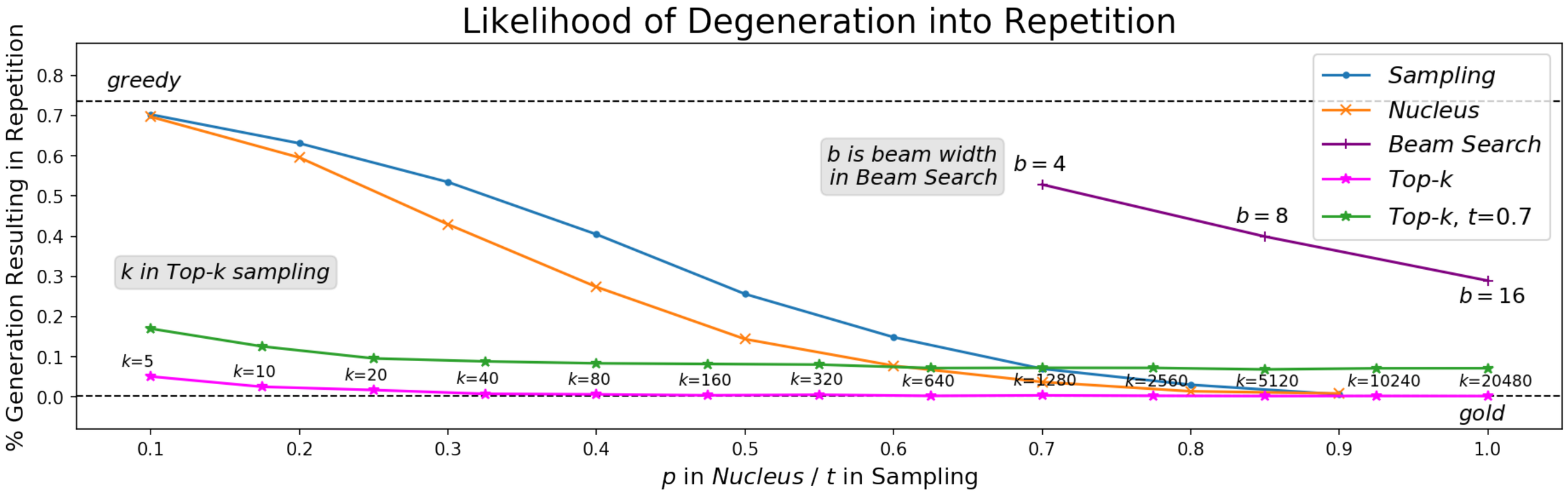
Top-K with $K = 5$

Only sample from the top-k scores of ϕ



P. et al. MAUVE Scores for Generative Models: Theory and Practice. JMLR 2023

Why do we truncate the logits?



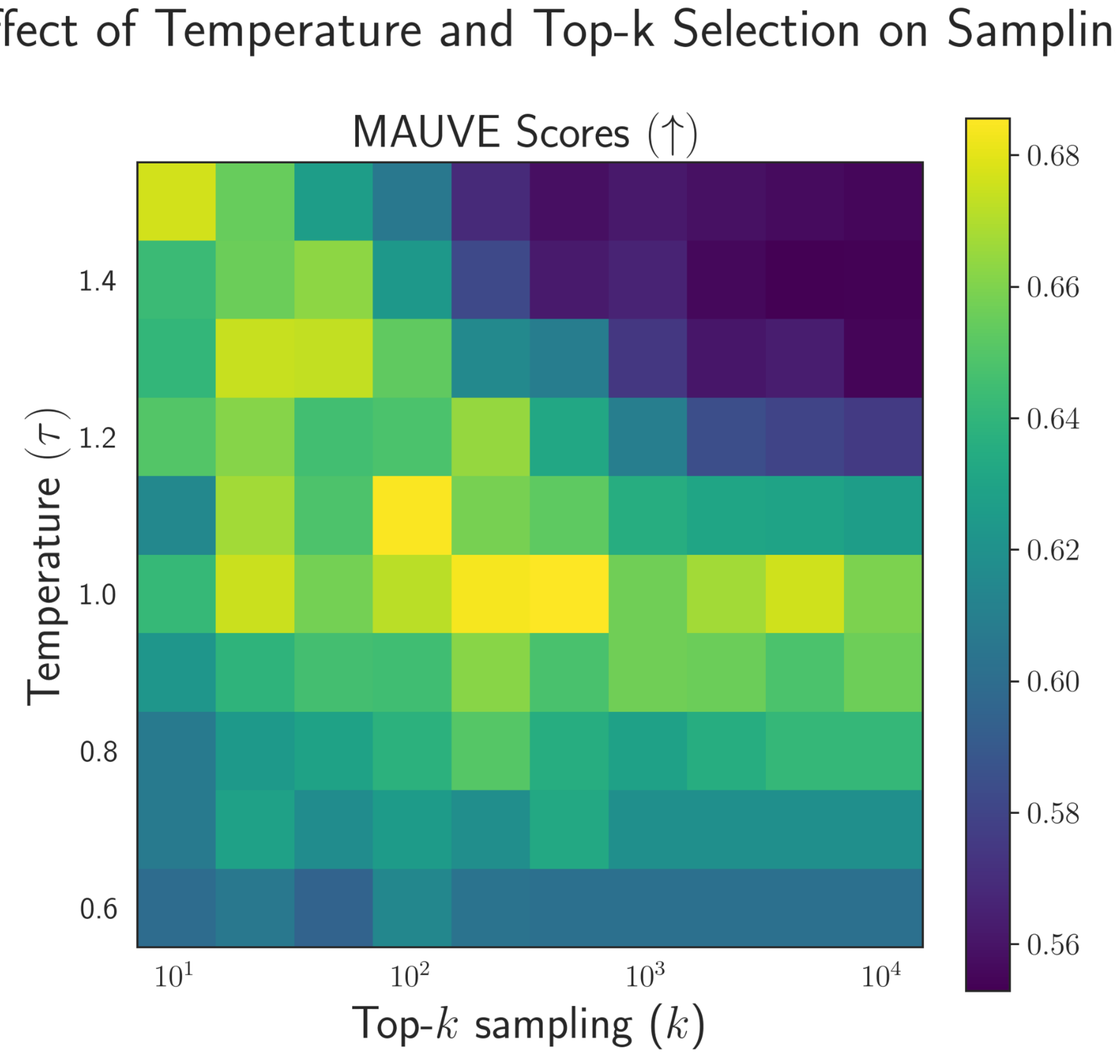
Top-k/top-p truncation reduces likelihood of degenerate text!

Image Credit: *Holtzman et al.* The Curious Case of Neural Text Degeneration, ICLR 2020

Why do we truncate logits ?

MAUVE Scores: similarity between synthetic and real text distributions

Best quality text is obtained when $\tau \approx 1$ and k ≈ 300



P. et al. NeurIPS 2021 ***(Outstanding Paper Award)***
P. et al. JMLR 2023 ***(Best Papers Track)***

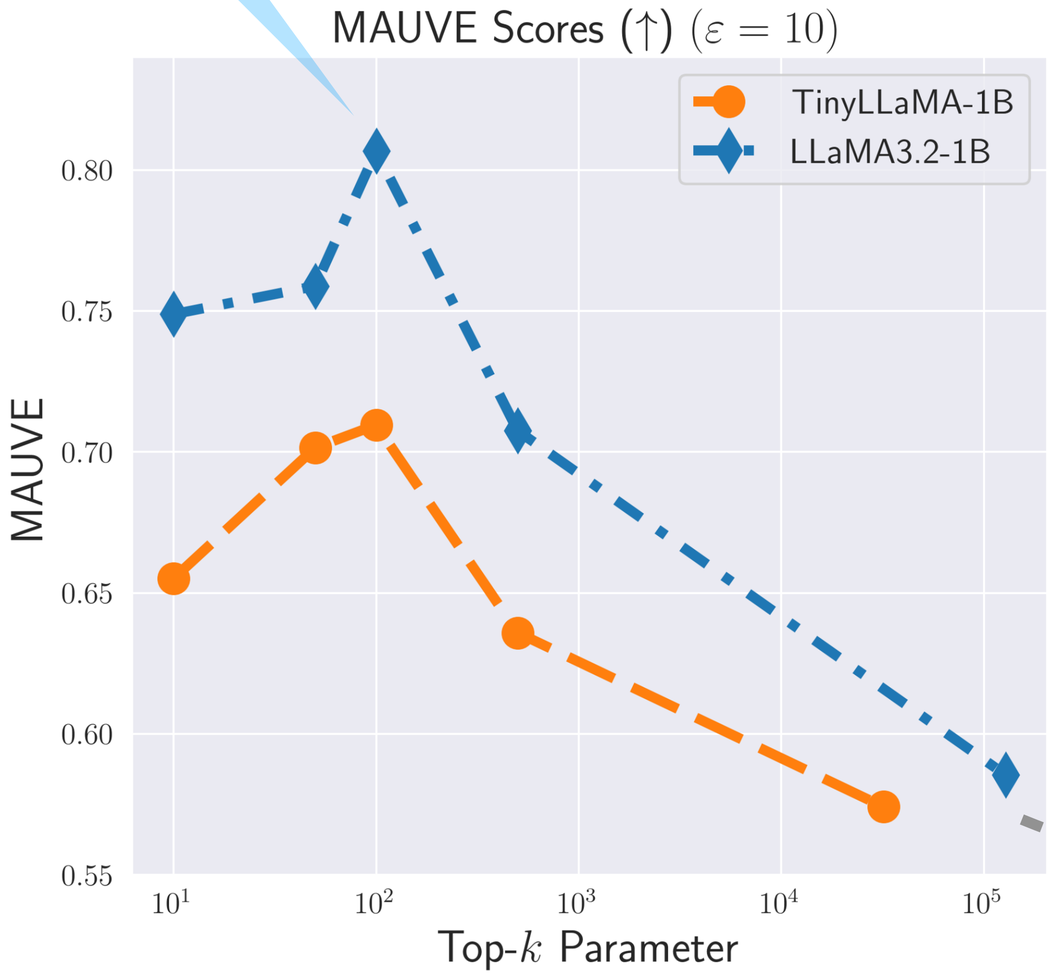
Key Ingredient 2: Align DP generation with practical decoding algorithms (from NLP)

Apply ***logit truncation*** without leaking privacy

Invisible Ink = private version of best-performing decoding algorithms

Benefits of aligning DP generation with practical decoding algorithms

Higher is better



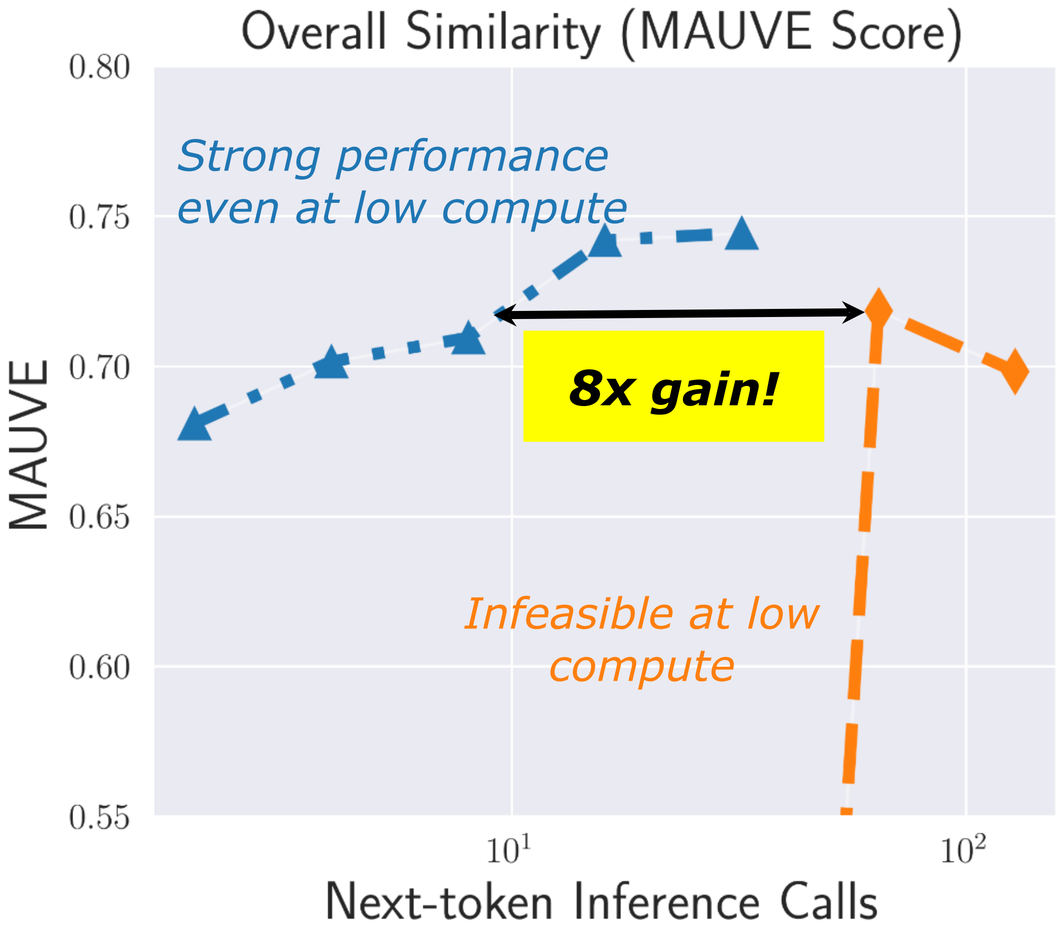
Best at $k \approx 100$

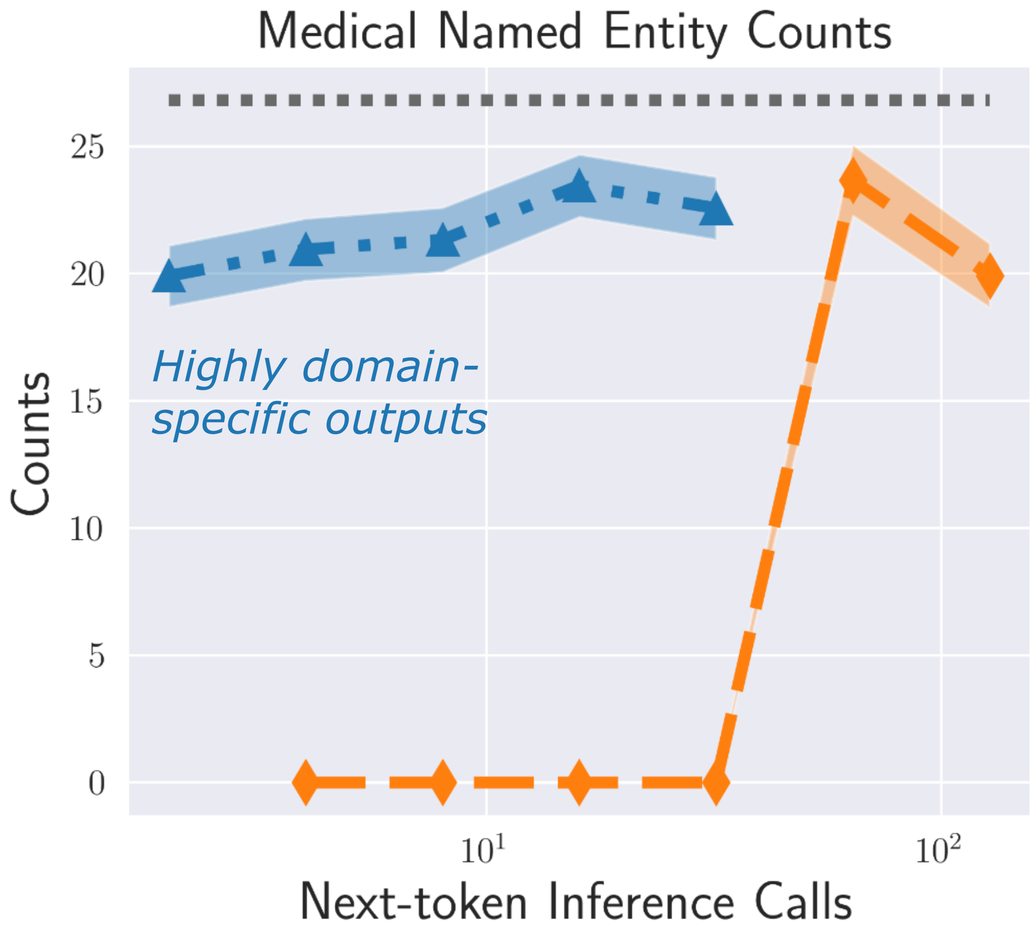
LLaMA3.2 1B has a larger relative drop due to larger vocabulary size (128K vs 32K)

Empirical Results

($\varepsilon = 10$)

MIMIC IV (Medical Discharge Notes)





Ours | Prev. SoTA [ABK+24] | Real Text

Practical Recommendations and Automatic Tuning (given privacy level and compute budget)

Step 1 | Step 2 | Step 3

Tune sampling temperature around $\tau \approx 1$

Default: same as non-private setting

For a fixed privacy (ρ_{seq}) and compute budget (B), set clipping norm to:

$$C = \frac{B\tau}{\sqrt{2\rho_{seq}/T}}$$

Tune top-k parameter

Default: same as non-private setting

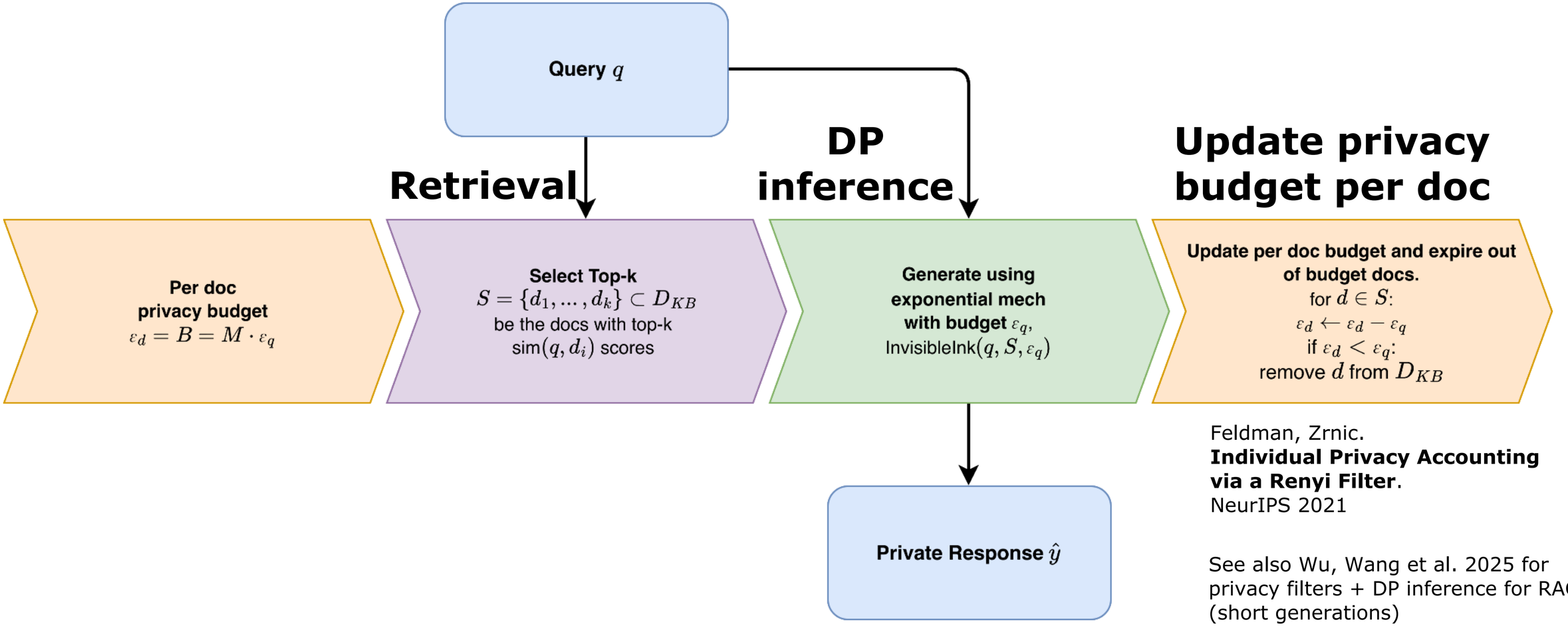
Summary

Invisible Ink: framework for private LLM inference

- **Order of magnitude** improvement in **compute cost** for same privacy & utility
- Makes DP synthetic text generation ***practical***
- User-friendly non-adaptive DP accounting
- Works off-the-shelf (sensible default hyperparameters)



Adding differential privacy to RAG



Feldman, Zrnic.
Individual Privacy Accounting via a Renyi Filter.
NeurIPS 2021

See also Wu, Wang et al. 2025 for privacy filters + DP inference for RAG (short generations)

Ramaswamy et al.
Preliminary version at DPDAHS Workshop at IEEE Int. Conf. on Healthcare Informatics 2026.

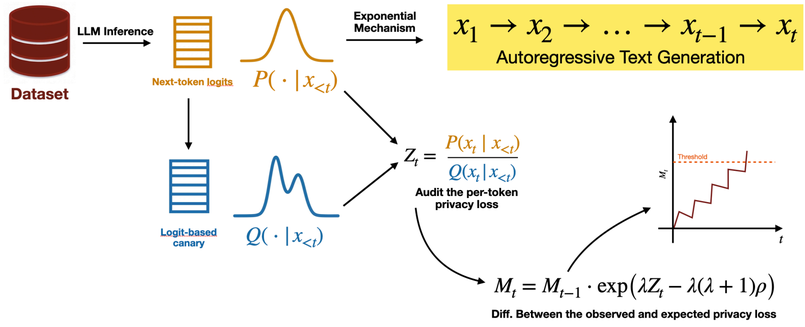
Our problem: Long-form private text generation

Private inference: mechanism

[Vinod et al. NeurIPS 2025]

Empirical privacy auditing

[Shah et al., Submitted 2026]



Getting DP deployments ***right*** can be tough…

DP Audit: To verify that we actually provide the guarantee we claim (no bugs in proofs/implementation)

2 mnist_experiment.py

```
@@ -71,7 +71,7 @@ def forward(self, x):
                                                rho_i,
                                                epochs,
                                                inp_clip,
-                                               grad_clip
+                                               grad_clip/BATCH_SIZE
                                            )
                    tl, correct, set_len = uc.test(model, test_loader)
                    print(f'MNIST_{BATCH_SIZE}_{epochs}_{grad_clip}_{inp_clip}_{rho_i}', correct/set_len)
```

2 upstream_clipping.py

```
@@ -110,7 +110,7 @@ def run_experiment(model, train_loader, rho_i, epochs, input_bound, grad_bound):

    model.train()
    # sensitivity for everything with weights is just:
-   sensitivity = input_bound * grad_bound / train_loader.batch_size
+   sensitivity = input_bound * grad_bound
    sigma = np.sqrt(sensitivity**2 / (2*rho_i))
    print('sensitivity:', sensitivity)
```

Tramèr et al. **Debugging Differential Privacy: A Case Study for Privacy Auditing**. 2022

Privacy in Theory, Bugs in Practice: Grey-Box Auditing of Differential Privacy Libraries

Tudor Cebere
PreMeDICaL team, Inria, Idesp, Inserm, Université de Montpellier
tudor.cebere@inria.fr

David Erb
Technical University of Munich & Oblivious
david.erb@tum.de

Damien Desfontaines
Hiding Nemo
damien@hiding-nemo.ch

Aurélien Bellet
PreMeDICaL team, Inria, Idesp, Inserm, Université de Montpellier
aurelien.bellet@inria.fr

Jack Fitzsimons
Oblivious
jack@oblivious.com

We audited 12 of the most widely used open-source DP libraries and found 13 previously unknown privacy violations. Here are some highlights.

Cebere et al. PoPETS 2026

Empirical privacy auditing



Privacy Leakage

Provable DP parameter (worst-case; can be loose)

Actual privacy leakage on this instance

Empirical privacy lower bound

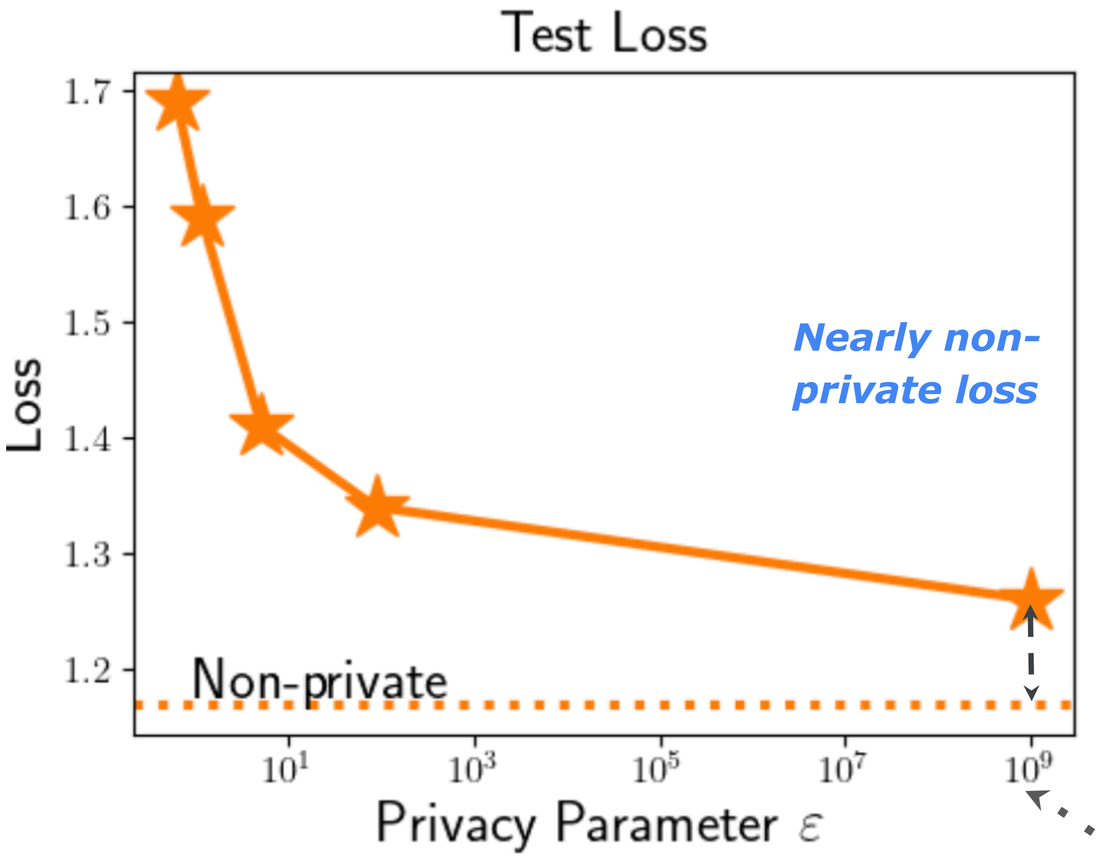


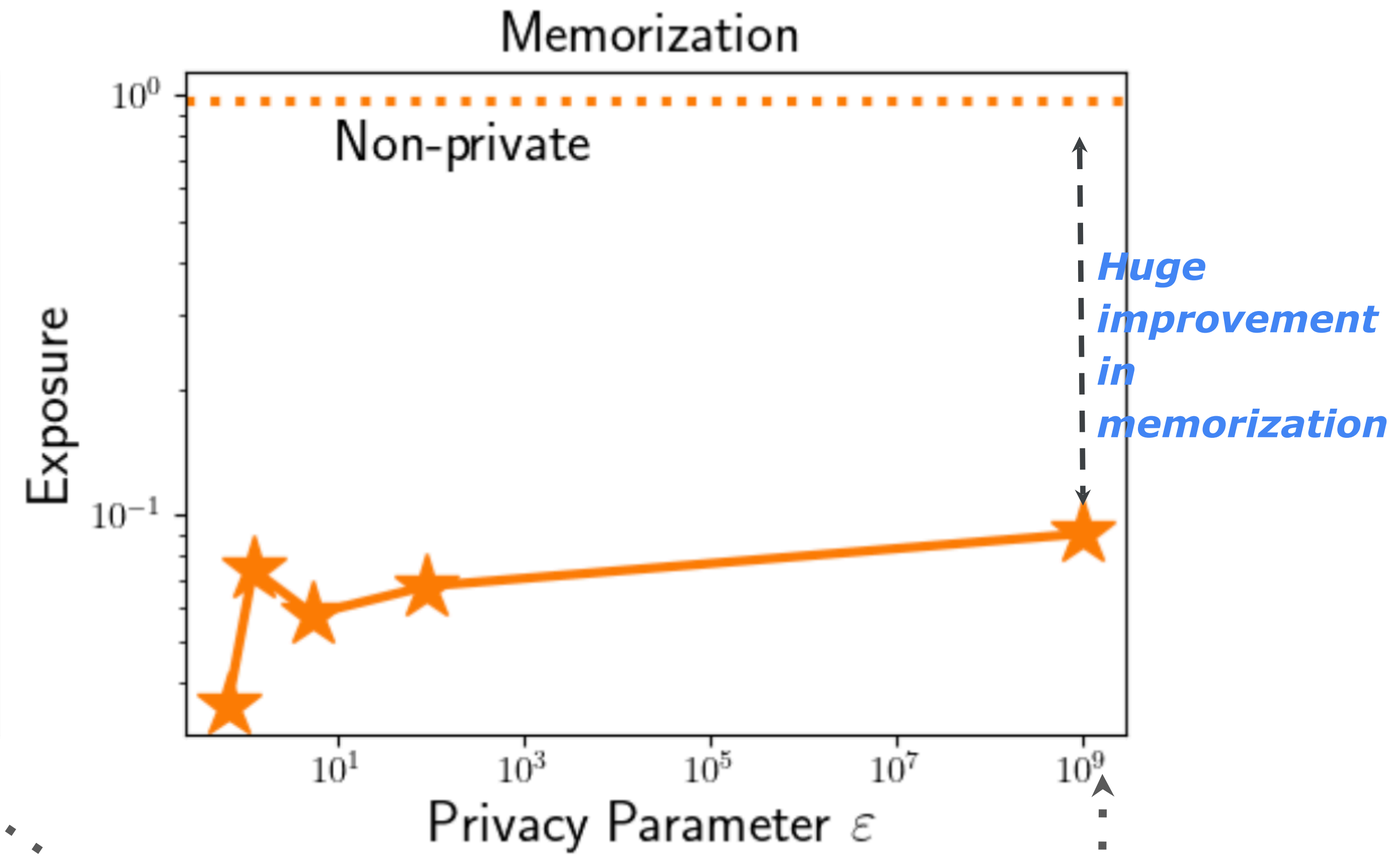
DP parameter ε (or ρ)

"I have detected $\hat{\varepsilon}$ amount of leakage"

Our focus

Gap between DP guarantees and empirical behavior: Memorization



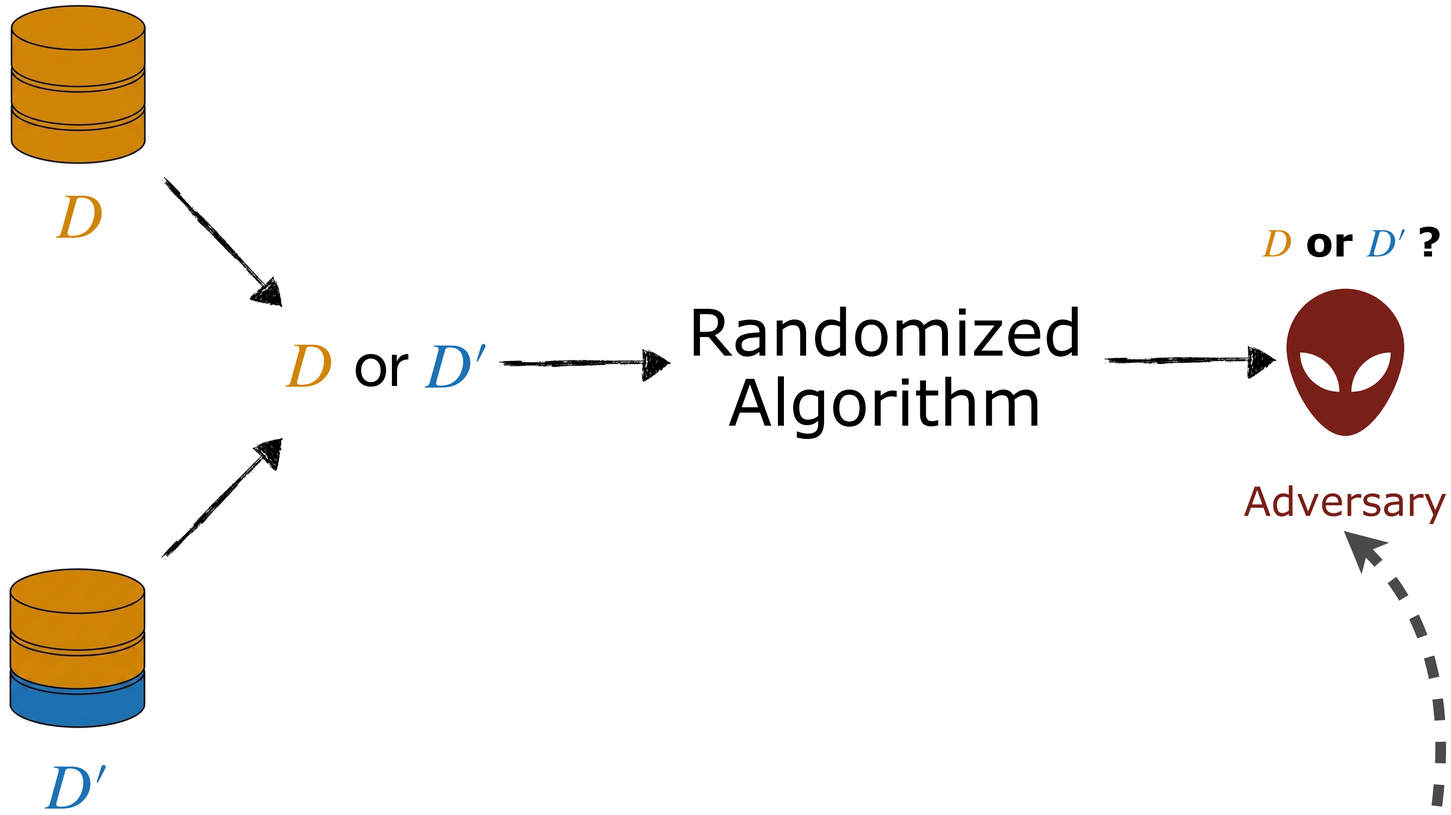


Privacy guarantee is vacuous at this ε!

Carlini, Liu, Erlingsson, Kos, Song. **The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks.** USENIX Security 2019

Auditing Differential Privacy:

adversary runs a membership inference attack



An adversary distinguishes D from an adversarially-crafted neighbour D'

Typical challenge: Empirical Privacy Auditing requires **many samples**

- In the context of DP training
  - Past: $n \gtrsim 1000$ models trained

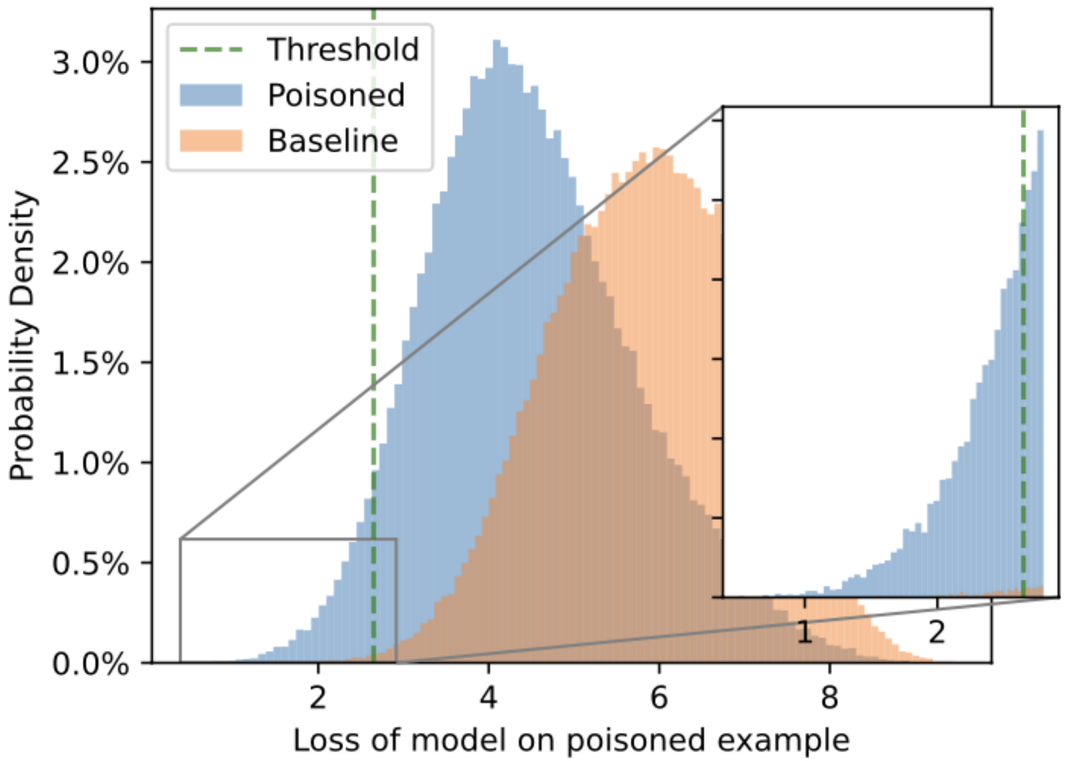
    [Jagielski et al. NeurIPS 2019]
  - Down to $O(100)$

    [***P***. et al. NeurIPS 2023]
  - Down to exactly 1

    [Steinke et al. NeurIPS 2023] & follow-ups

- ***Our goal***: auditing DP inference

Used $n = 200{,}000$ models to find a bug in the wild



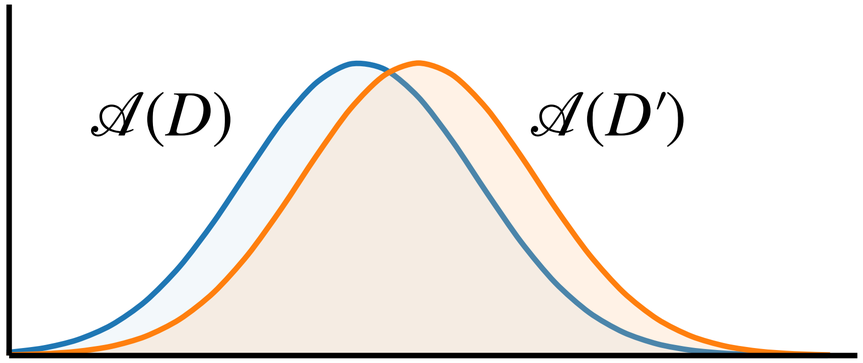
Tramèr et al. **Debugging Differential Privacy: A Case Study for Privacy Auditing**. 2022

Threat model: privacy loss oracle

Privacy loss:

$$Z = \log\left(\frac{\mathbb{P}(\mathscr{A}(D) = x)}{\mathbb{P}(\mathscr{A}(D') = x)}\right)$$

for $x \sim \mathscr{A}(D)$



Privacy loss oracle:

For the *generated token* $x_t \sim P(\cdot \mid x_{<t})$ and an adjacent output distribution Q,

a privacy loss oracle returns

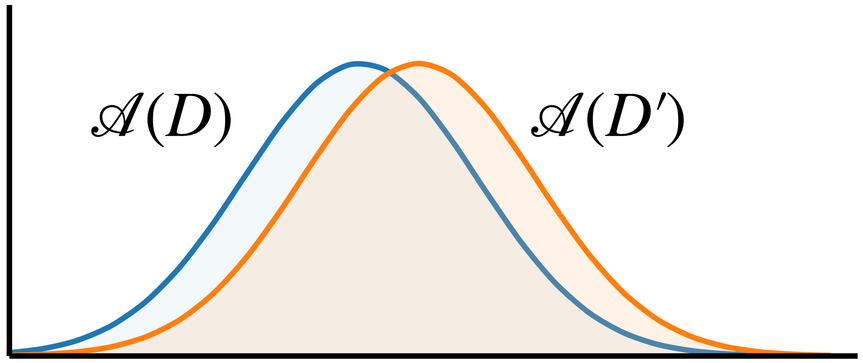
$$Z_t = \log \frac{P(x_t \mid x_{<t})}{Q(x_t \mid x_{<t})}$$

Possible in a white box setting:

$$P(x_{t+1} \mid x_1, \ldots, x_t) \propto \exp\left(\phi(x_{t+1} \mid x_{1:t})\right)$$

E.g. model developer, trusted third party, etc.

Invisible Ink's guarantees: Zero-Concentrated DP (zCDP)



$$Z = \log\left(\frac{\mathbb{P}(\mathscr{A}(D) = x)}{\mathbb{P}(\mathscr{A}(D') = x)}\right)$$

for $x \sim \mathscr{A}(D)$

For all $\lambda > 0$, we have

$$\mathbb{E}\left[\exp(\lambda\, Z\,)\right] \leq \exp(\lambda(\lambda+1)\rho)$$

Privacy Loss is sub-Gaussian!

Bun & Steinke. **Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds.** TCC 2016

Auditing zero-concentrated DP audit
with a privacy loss oracle

We test sub-Gaussianity of Privacy Loss

$H_0: \quad \mathbb{E}\left[e^{\lambda Z_t}\right] \leq \exp\left(\lambda(\lambda+1)\rho\right)$ for all $\lambda \geq 0$

$H_1: \quad \mathbb{E}\left[e^{\lambda Z_t}\right] > \exp\left(\lambda(\lambda+1)\rho\right)$ for some $\lambda \geq 0$

Default claim: Mechanism follows ρ-zCDP

Z_t: privacy loss of one token

(We can also do this at the sequence level)

Sequential zCDP Audit

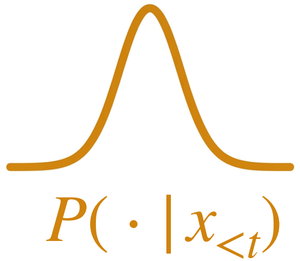


Dataset

LLM Inference



Next-token logits

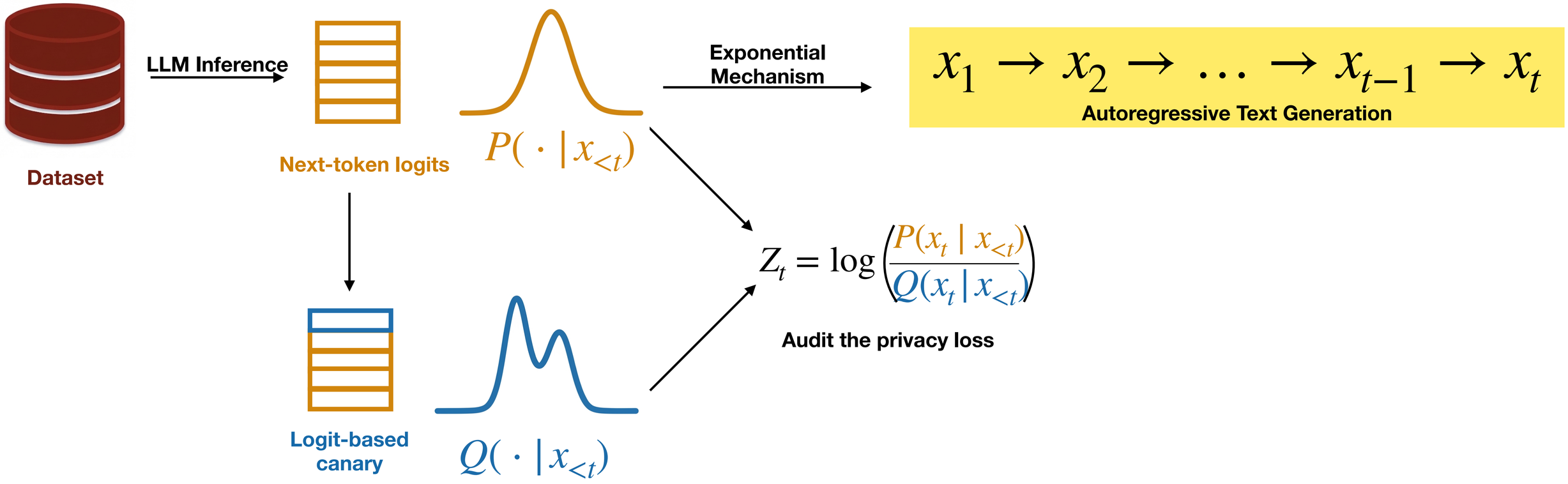


Exponential Mechanism

$$x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_{t-1} \rightarrow x_t$$

Autoregressive Text Generation

Sequential zCDP Audit



Sequential zCDP Audit



Dataset

LLM Inference

Next-token logits

$P(\cdot \mid x_{<t})$

Exponential Mechanism

$$x_1 \to x_2 \to \dots \to x_{t-1} \to x_t$$

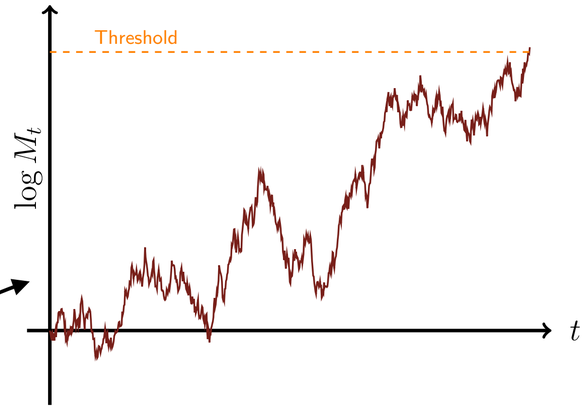
Autoregressive Text Generation

Logit-based canary

$Q(\cdot \mid x_{<t})$

$$Z_t = \log\left(\frac{P(x_t \mid x_{<t})}{Q(x_t \mid x_{<t})}\right)$$

Audit the privacy loss



$$M_t = M_{t-1} \cdot \exp\left(\lambda Z_t - \lambda(\lambda+1)\rho\right)$$

Diff. Between the observed and expected privacy loss

Results

What if the claimed leakage ρ is *correct*?

Theorem: Under H_0

Let τ be the time at which $M_t \geq 1/\alpha$, then we have,

$$\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$$

The probability of falsely rejecting a true privacy claim is at most α

What if the claimed leakage ρ is *incorrect*?

Theorem: Under H_1

Given true privacy parameter ρ^*. The stopping time τ is bounded as

$$\mathbb{E}_{H_1}[\tau] \lesssim \frac{1}{(\rho^* - \rho)^2} \text{ (up to log factors)}$$

Further, $\mathbb{P}_{H_1}(\tau < \infty) = 1$

Audit halts almost surely

Results

Lower bound: Under H_1

There exists an instance of the Gaussian mechanism s.t. for any sequential test

$$\mathbb{E}_{H_1}[\tau] \gtrsim \frac{1}{(\rho^* - \rho)^2}$$

Based on sequential testing lower bounds from:
Agrawal, Ram, Ramdas. arXiv 2504.19952.

What if the claimed leakage ρ is *incorrect*?

Theorem: Under H_1

Given true privacy parameter ρ^*. The stopping time τ is bounded as

$$\mathbb{E}_{H_1}[\tau] \lesssim \frac{1}{(\rho^* - \rho)^2} \text{ (up to log factors)}$$

Further, $\mathbb{P}_{H_1}(\tau < \infty) = 1$

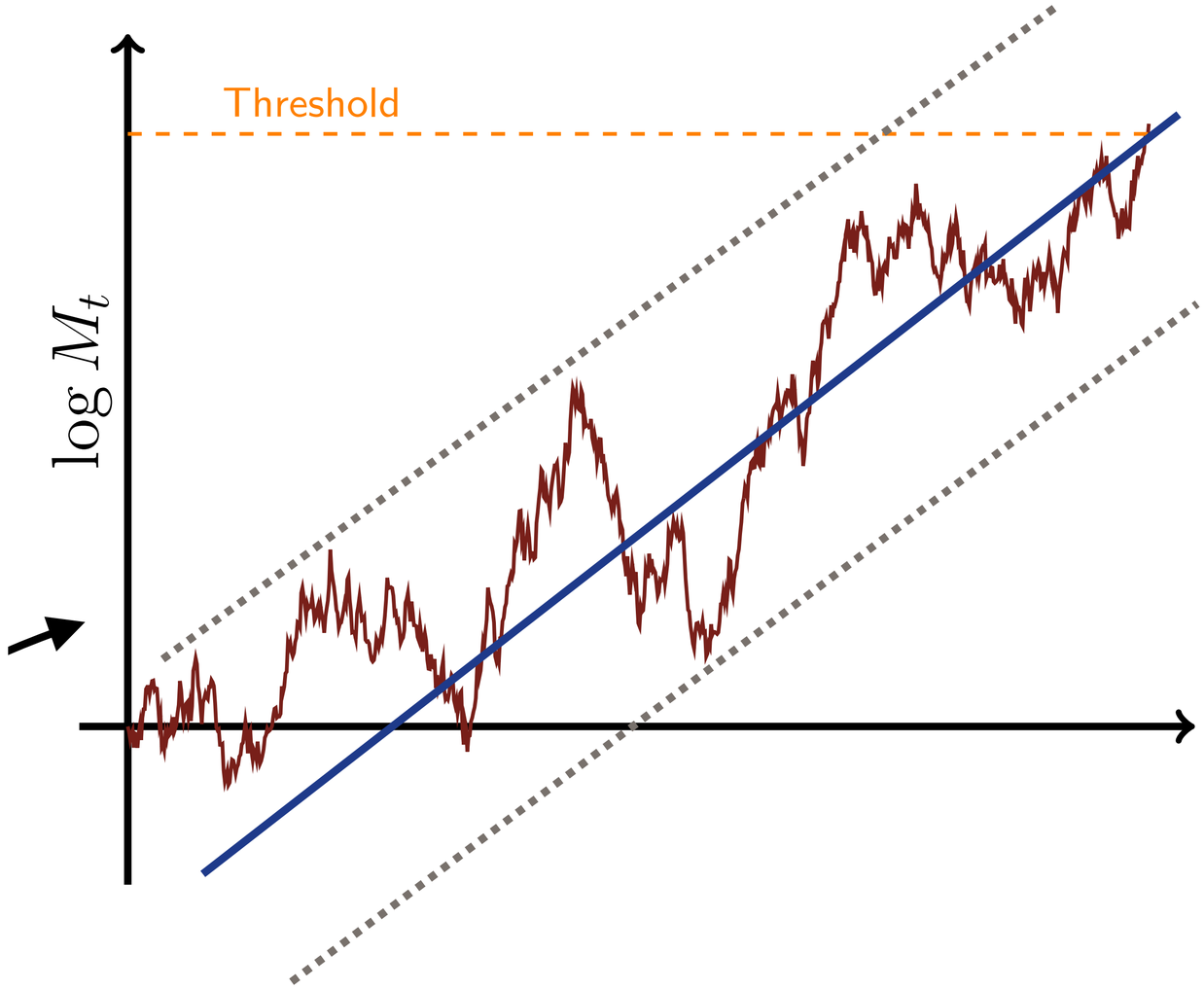
Proof Sketch (upper bound)

There exists a $\lambda > 0$ and $\delta > 0$, such that,

$$\log M_t = n \cdot \delta + \lambda \sum_{t=1}^{n} (Z_t - \rho^*).$$

Linear Drift

Scales as $\sqrt{n}$ by martingale concentration



Logit-based Canary Design

Similar to canary gradients in auditing DP-SGD

Goal: Reject H_0 as soon as possible when it is false

$\Downarrow$

Under H_1: Increase M_t quickly

$\Downarrow$

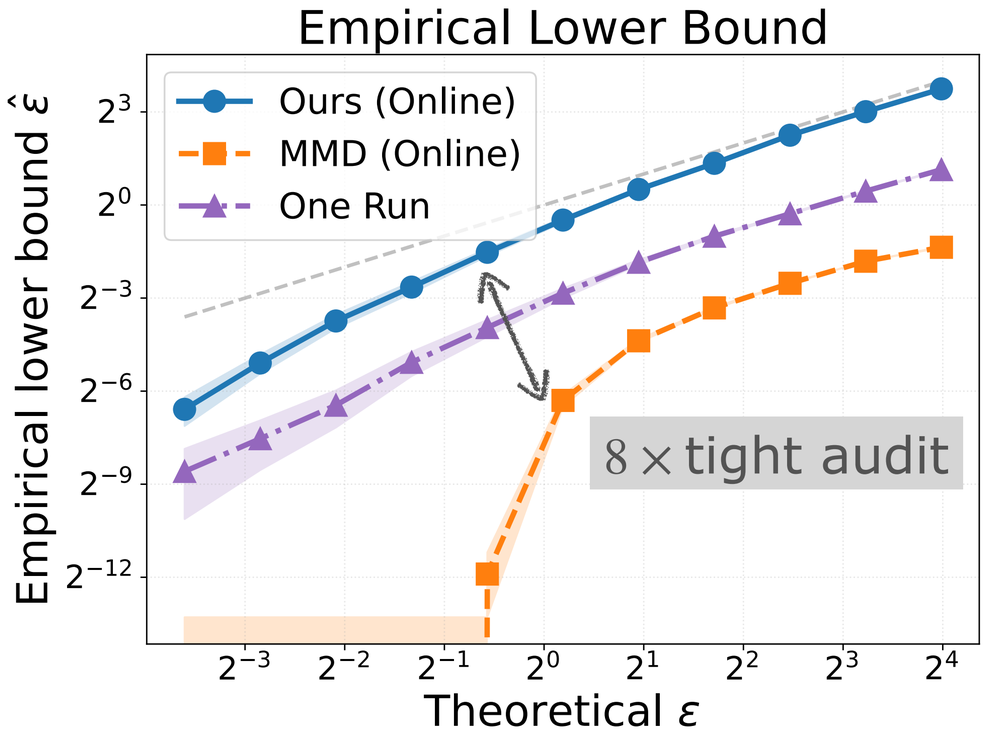
Maximize the KL divergence

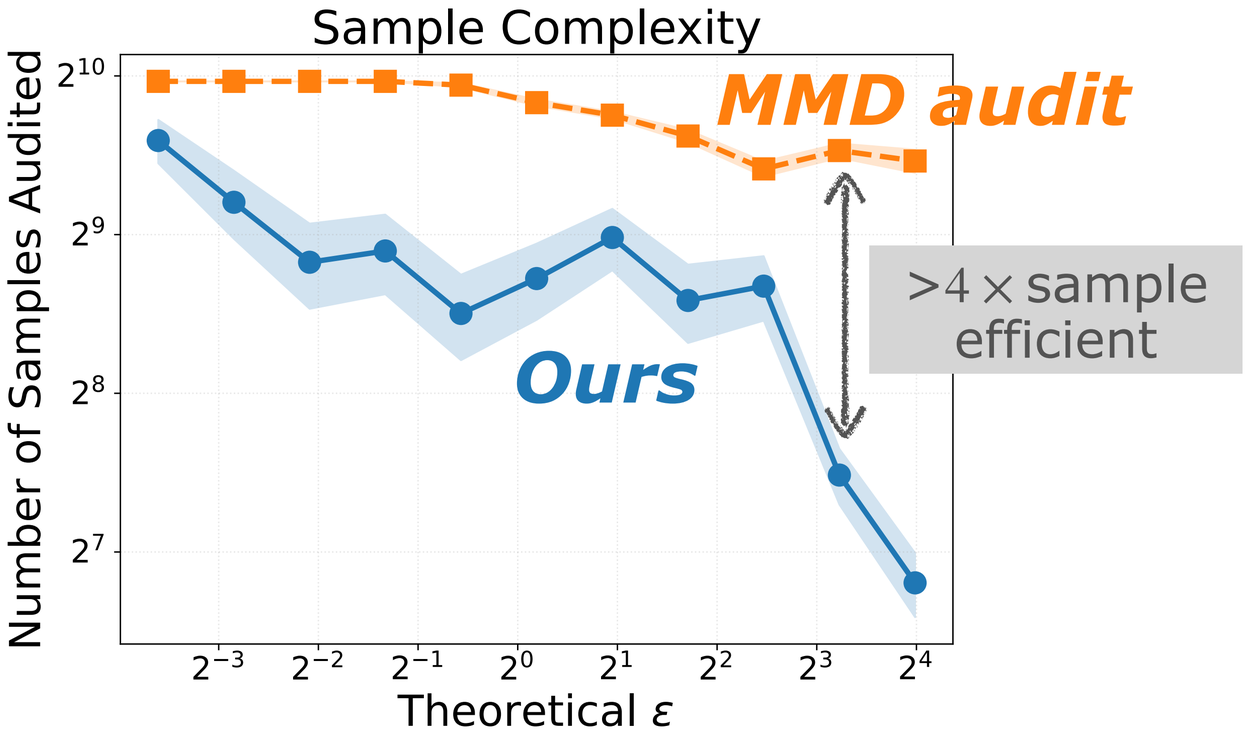
$$z = \arg\max_z \mathbb{E}\left[\log M_t(\lambda)\right]$$

$$= \arg\max_z \mathrm{KL}\left(P(\cdot \mid x_{<t}) \| Q_z(\cdot \mid x_{<t})\right).$$

Note: requires full next-token distributions (stronger threat model)

Experiments: Auditing Gaussian Mechanism



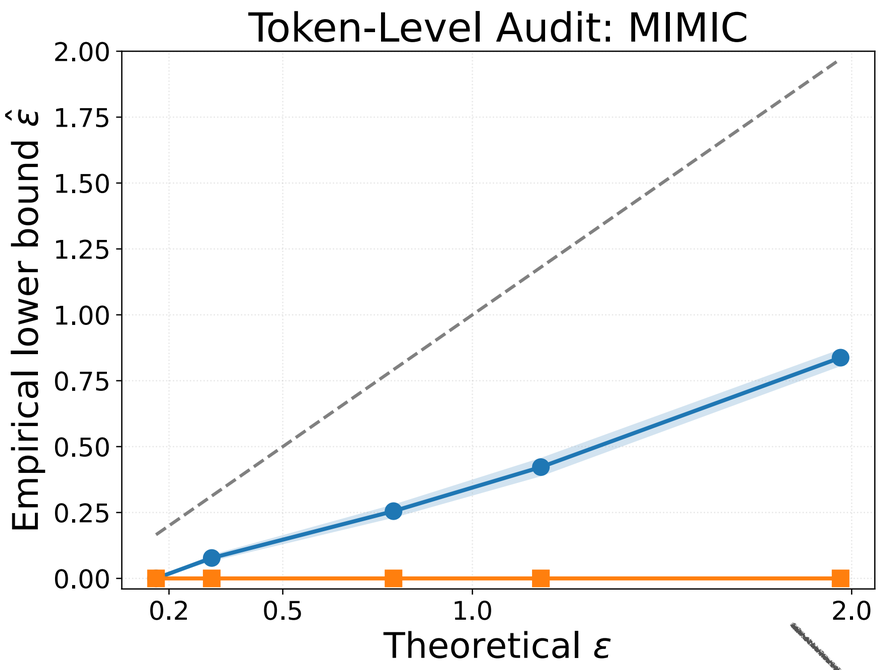


MMD Audit: González et al. NeurIPS 2025

One Run: Steinke et al. NeurIPS 2023

Experiments: Auditing InvisibleInk at the token level

Synthetic Text Generation



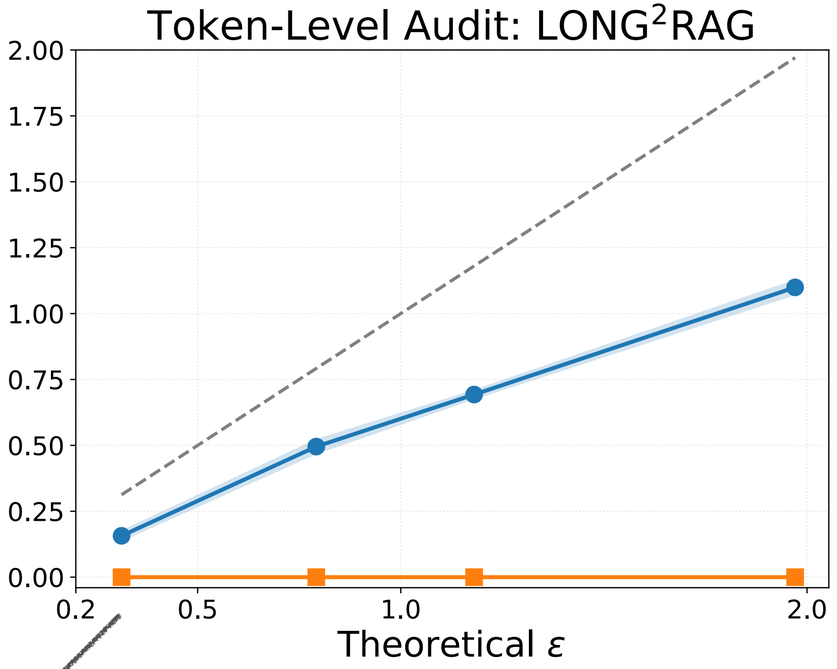
Our Audit

MMD Audit

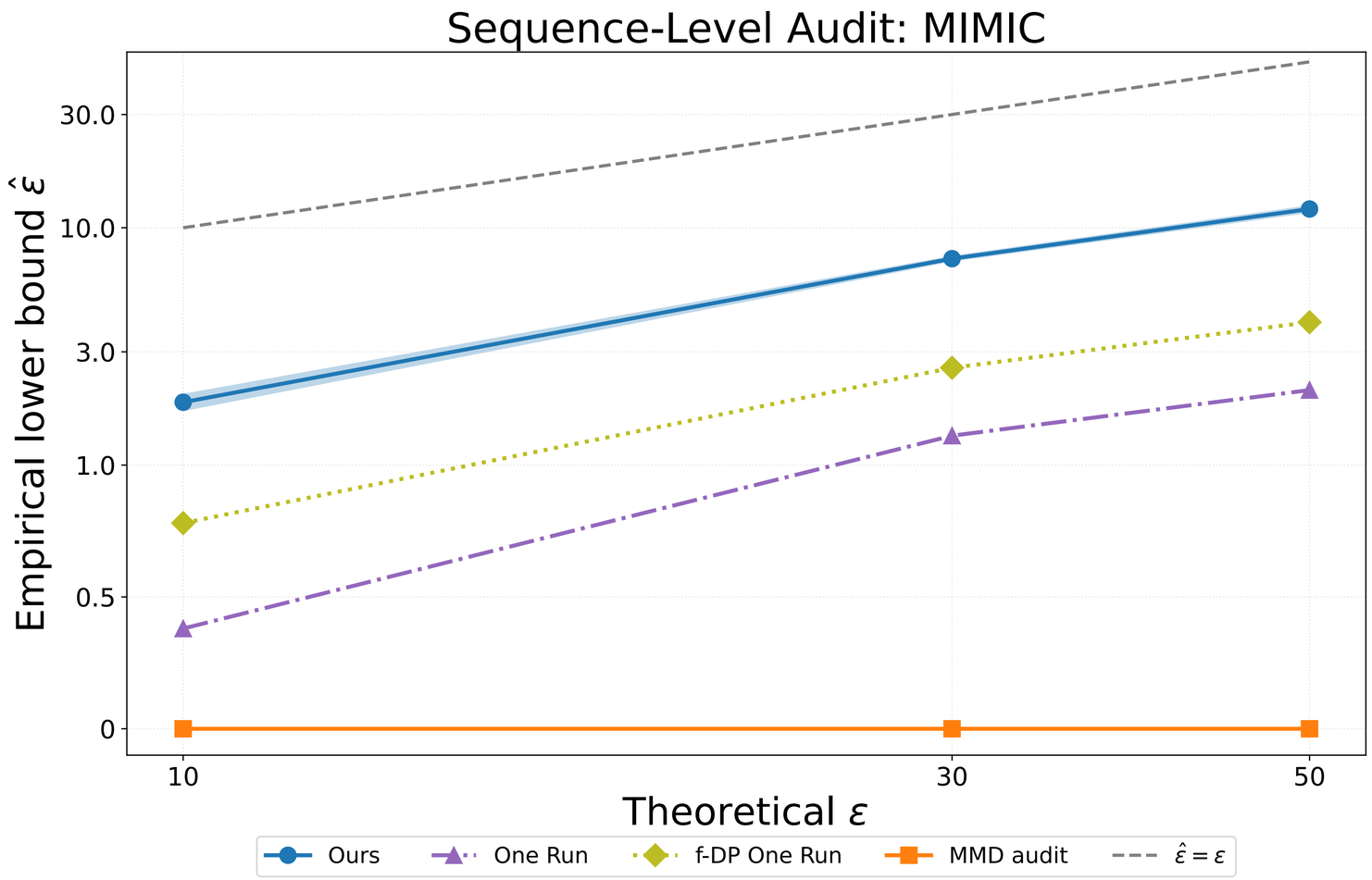
[González et al.]

MMD audit returns trivial lower bound

RAG



Experiments: Auditing InvisibleInk at the sequence level



Our Audit

One-run audits

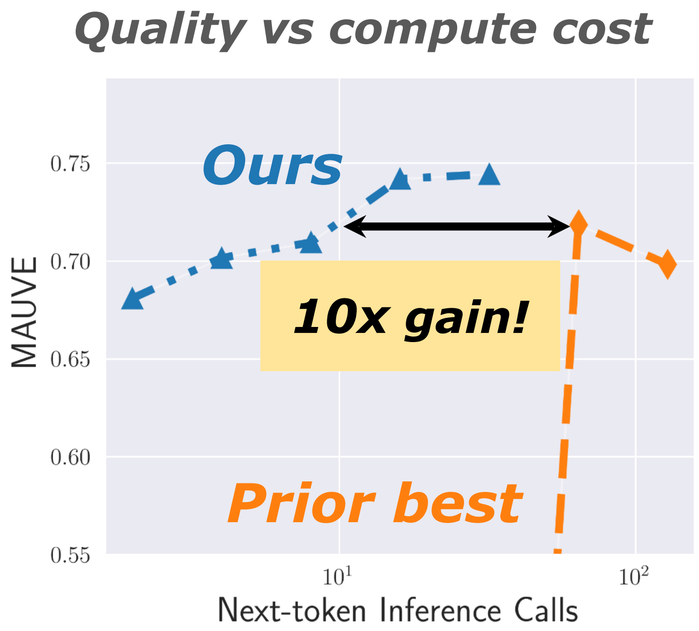
MMD Audit

[González et al.]

Our problem: Long-form private text generation

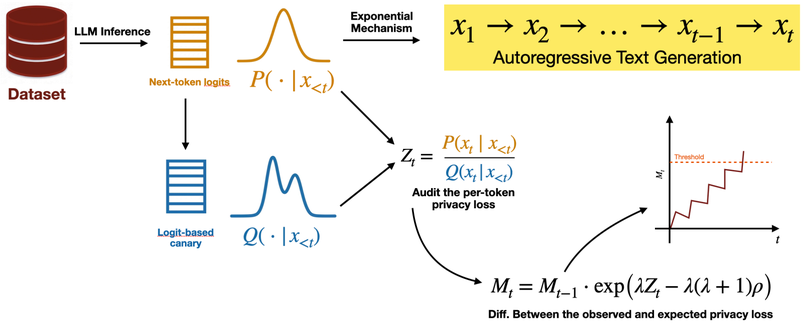
Private inference: mechanism

[Vinod et al. NeurIPS 2025]



Empirical privacy auditing

[Shah et al., Submitted 2026]



Thank you!

Joint work with:

Vishnu Vinod, Arun Ramaswamy, Dhruv Shah (IITM)
Ramayya Krishnan (CMU), Rema Padman (CMU),
Siva Krishnamoorthy (diax.ai), Abhradeep Thakurta (Google DeepMind)

Funding from ANRF, Schmidt Sciences, Google, Microsoft Research is gratefully acknowledged