When Privacy Guarantees Meet Pre-Trained LLMs

A Case Study in Synthetic Data

PEPR'25

Yash Maurya & Aman Priyanshu

Scenario



Say, you want to release some synthetic financial data publicly?

Monday:

You launch your synthetic data with Differential Privacy guarantees

Wednesday:

Security researchers find a litany of real customer names and emails.

What is Synthetic Data?

Mimics Real Distributions

Data is generated to statistically resemble real datasets without copying it entirely.

Enables Statistical Computation

Allows analysis, modeling, and evaluation without direct access to sensitive raw data.

Attempts, No Direct Linkage

Designed to prevent one-to-one mapping between synthetic and real individuals, but...

But what do we need Synthetic Data for?

Data Augmentation for Sensitive Domains

AI/ML Model training - particularly **LLMs**

Software Testing and Experimentation

Sensitive Data sharing with third parties!

Research and experimentation

And so on.....

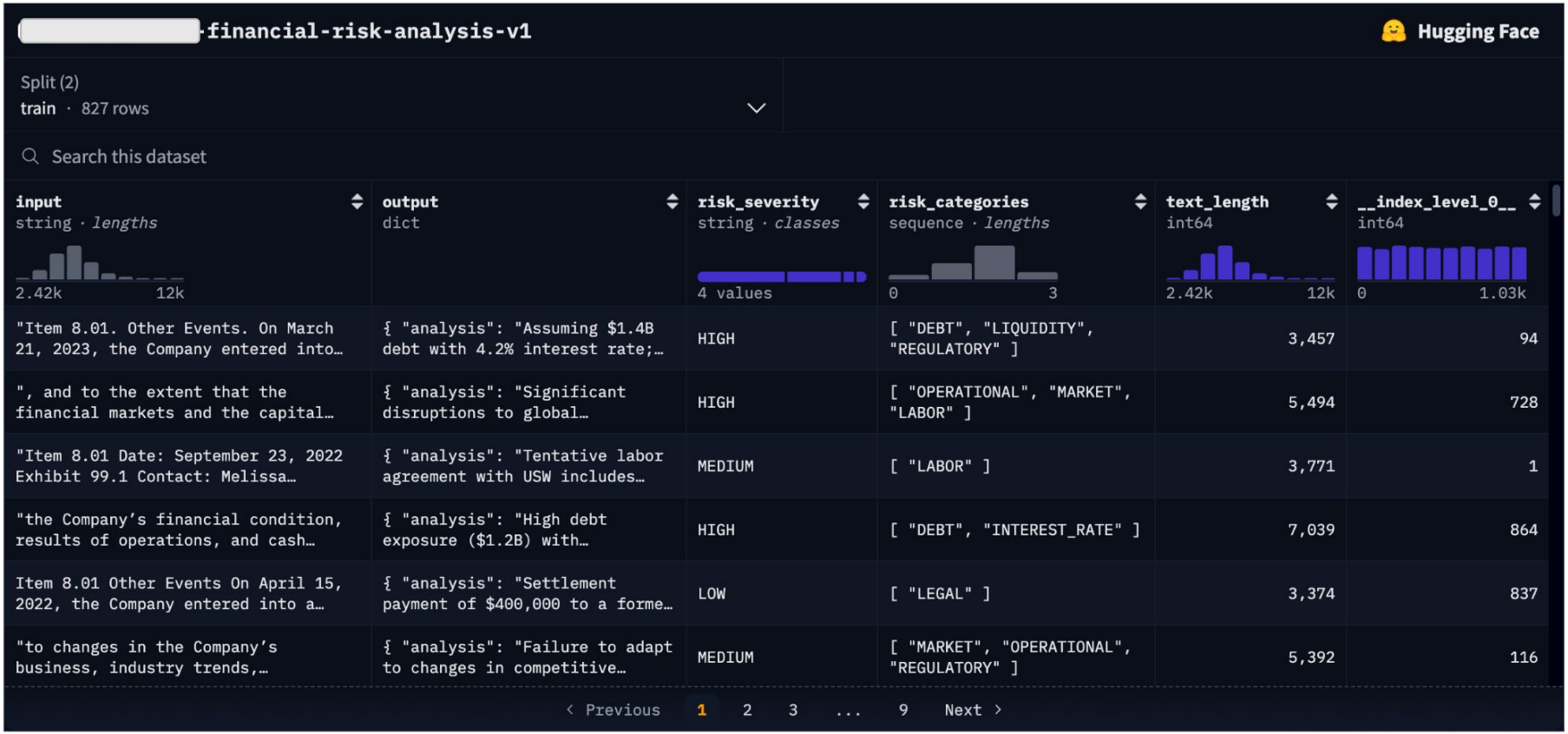
Problems!

Training Data Can be Leaked!!

DP 🤔!?

What about differentially private synthetic data?

CASE STUDY



***We conducted this analysis on a pre-2025 version of the dataset [The dataset description has since been updated]**

What we knew about this data!?

Public SEC Filings (10-K, 10-Q, and 8-K) between 2023-2024

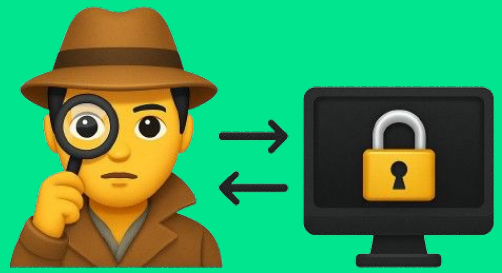
Generated using a fine-tuned a *phi-3-mini-128k-instruct*

Differentially Private: ε = 8

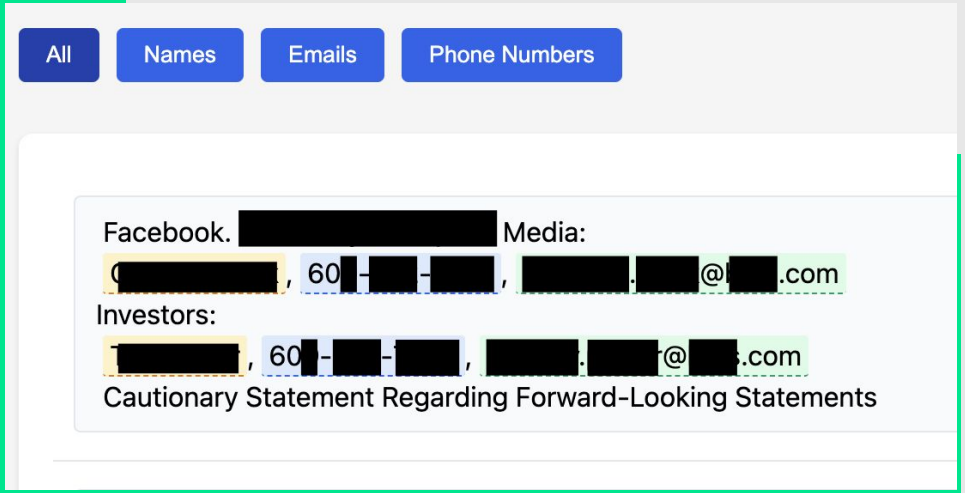
Document level privacy unit

Adversary Mode On

Black Box Attacks



PII Identification & Anonymization

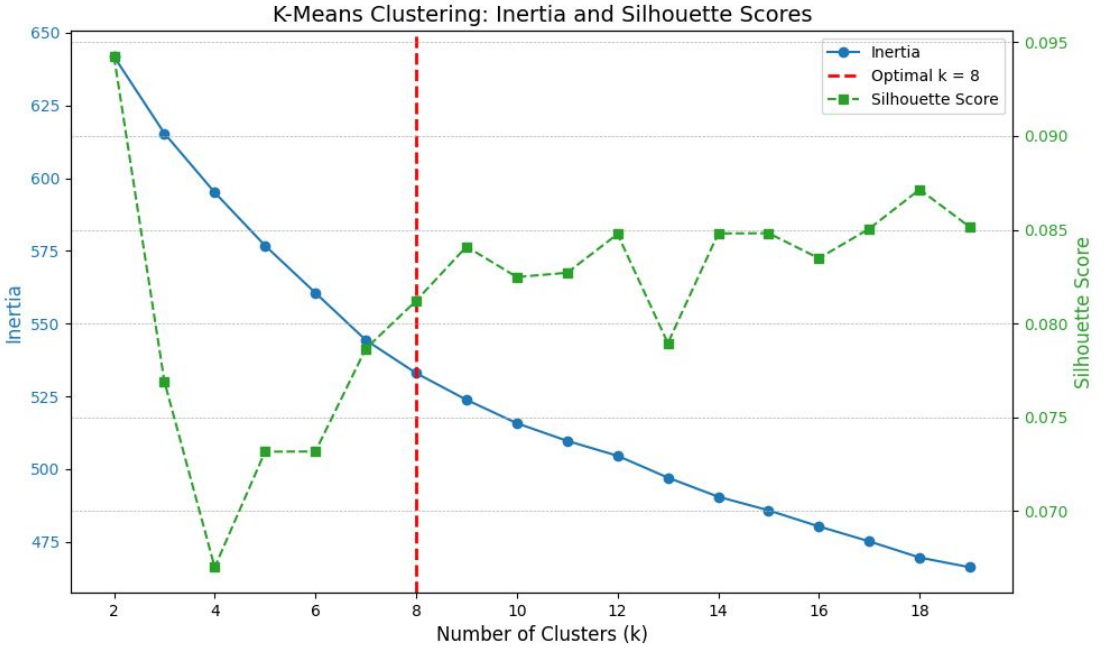


Three low-resource NER Models used:
Presidio (with spacy), GliNER, and a custom GliNER

Five PII Types Targeted
Names, emails, organizations, locations, phone numbers

Placeholder Anonymization
<NAME>, <EMAIL>, <ORGANIZATION>, <LOCATION>, <PHONE>

Clustering Analysis



Optimal Clusters Identified

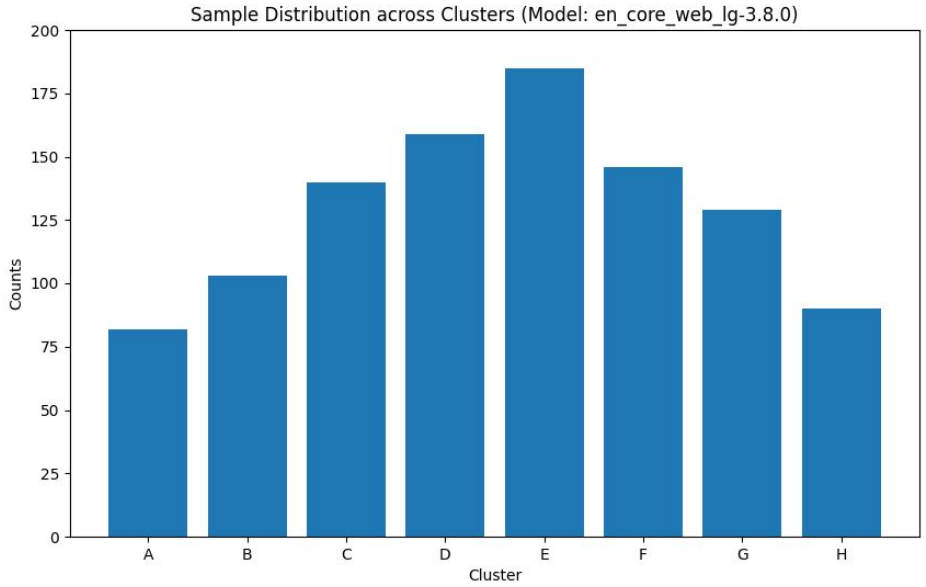
Kneedle algorithm found 8 clusters

Dataset Expectation vs Reality

Original dataset: **3** document types

Clustering revealed: **8** distinct groups

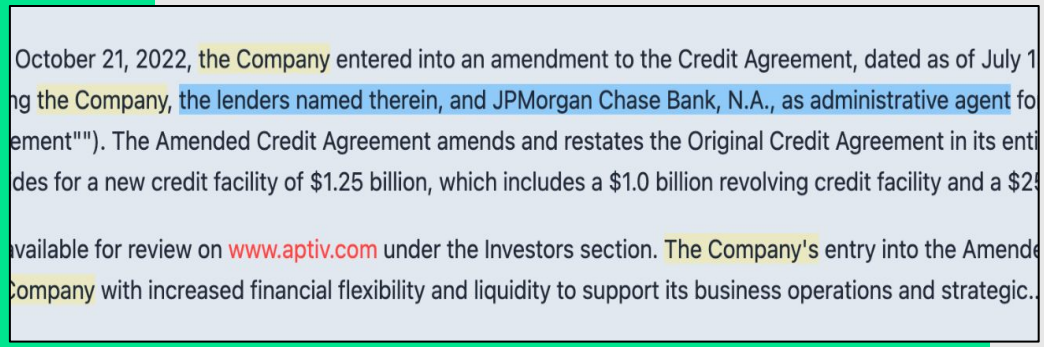
Distribution Analysis



Near-Normal Distribution Pattern

Possible Causes - DP-SGD?

Auxiliary Information Attacks

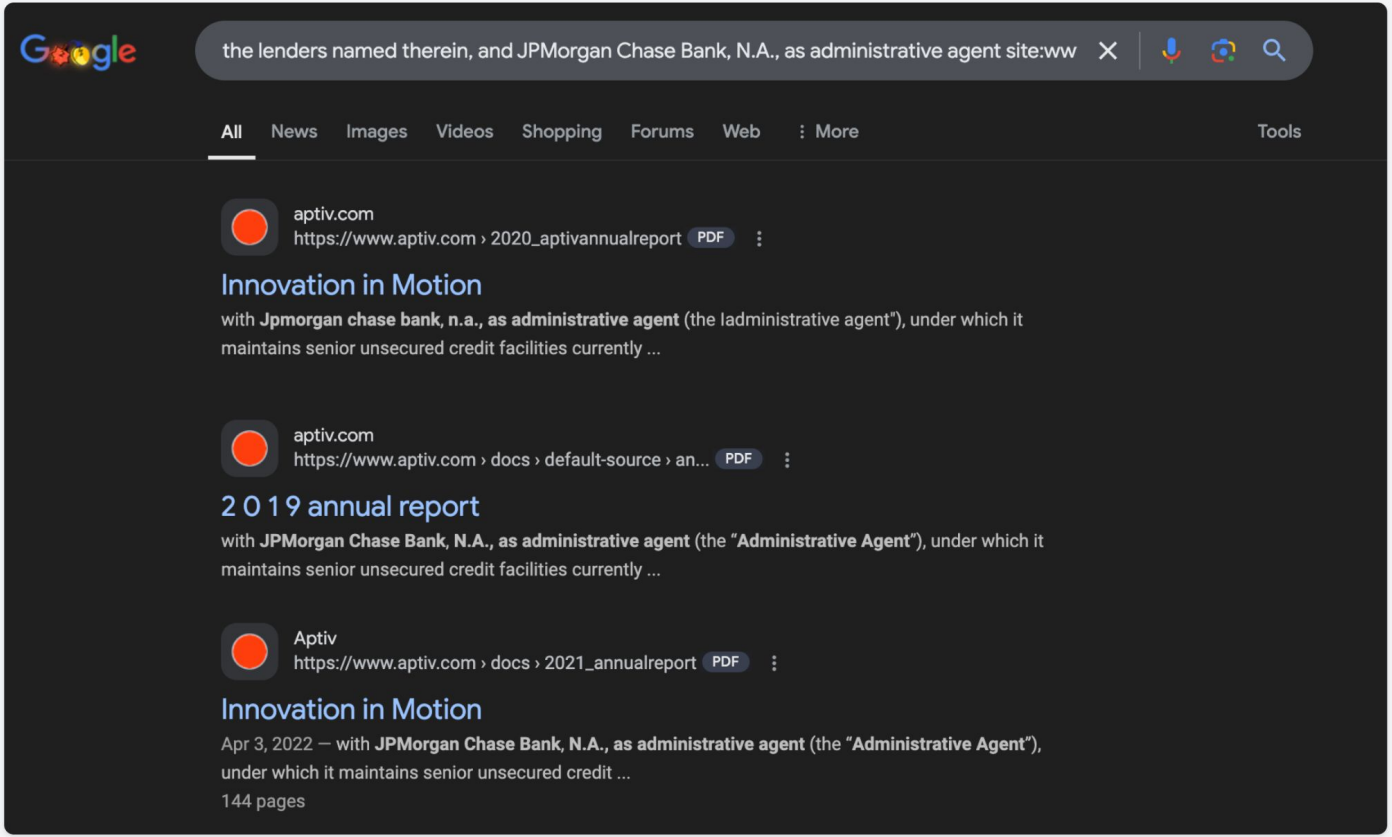

October 21, 2022, the Company entered into an amendment to the Credit Agreement, dated as of July 1
ng the Company, the lenders named therein, and JPMorgan Chase Bank, N.A., as administrative agent fo
ement""). The Amended Credit Agreement amends and restates the Original Credit Agreement in its enti
des for a new credit facility of $1.25 billion, which includes a $1.0 billion revolving credit facility and a $2
vailable for review on www.aptiv.com under the Investors section. The Company's entry into the Amend
Company with increased financial flexibility and liquidity to support its business operations and strategic..

Identified key sequences which appeared to be linkable (unique phrases/sentence structures, dates, and transactions)

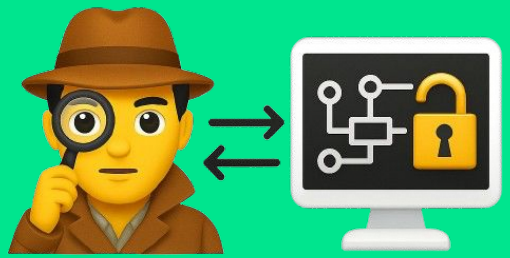
Auxiliary Information Sources
Public records, social media, company websites, SEC filings, search engines

The search query that reveals the original document:

```
the lenders named therein, and JPMorgan Chase Bank, N.A., as administrative agent site:www.aptiv.com
```



White Box Attacks



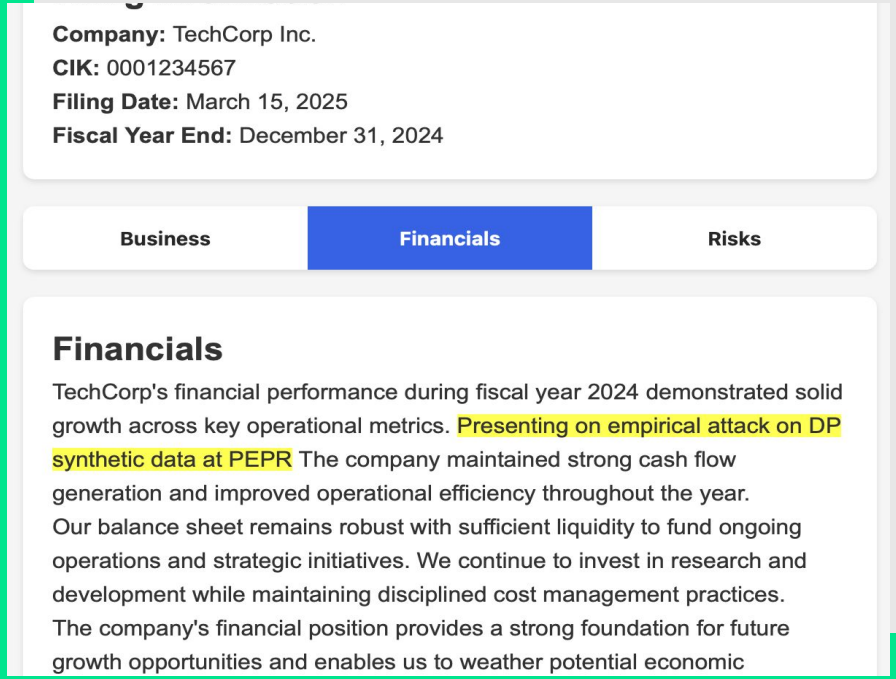
Membership Inference Attack

Determine which organizations were part of the training, given all SEC documents?

Analyze model behavior (confidence scores, predictions) on target data to infer membership

Higher confidence or unusual model behavior to indicate vulnerability

Canary Insertion Attack



Intentionally insert unique, memorable data into training documents

After model training, prompt the model to see if it reproduces the inserted ”canary” data

DP SYNTHETIC DATA RELEASE BINGO CARD

PII Identification & Anonymization	Clustering Analysis	Distribution Analysis
Auxiliary Information Attacks	Membership Inference Attacks	Canary Insertion Attacks

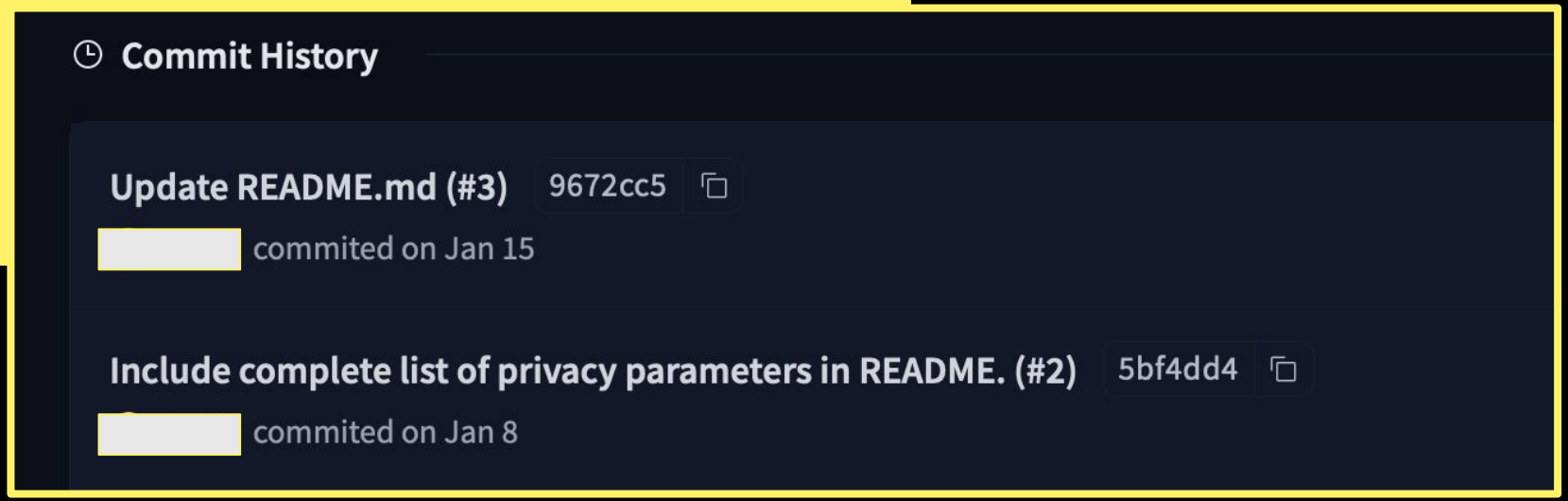
*Data to share? Show you care - check the synthetic data privacy bingo card.

TAKEAWAYS

DP is theoretically amazing, but execution is people experience

Red-Teaming is the true quantification of risk!

BONUS:
Dataset Update



For more details read the SynthLeak Blog!