



# 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?



AI/ML Model training - particularly LLMs



Data Augmentation for Sensitive Domains



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



Split (2)

train · 827 rows

Search this dataset

input	output	risk_severity	risk_categories	text_length	__index_level_0__
string · lengths  2.42k 12k	dict	string · classes  4 values	sequence · lengths  0 3	int64  2.42k 12k	int64  0 1.03k
"Item 8.01. Other Events. On March 21, 2023, the Company entered into...	{ "analysis": "Assuming \$1.4B debt with 4.2% interest rate;..."	HIGH	[ "DEBT", "LIQUIDITY", "REGULATORY" ]	3,457	94
", and to the extent that the financial markets and the capital...	{ "analysis": "Significant disruptions to global..."	HIGH	[ "OPERATIONAL", "MARKET", "LABOR" ]	5,494	728
"Item 8.01 Date: September 23, 2022 Exhibit 99.1 Contact: Melissa...	{ "analysis": "Tentative labor agreement with USW includes..."	MEDIUM	[ "LABOR" ]	3,771	1
"the Company's financial condition, results of operations, and cash..."	{ "analysis": "High debt exposure (\$1.2B) with..."	HIGH	[ "DEBT", "INTEREST_RATE" ]	7,039	864
Item 8.01 Other Events On April 15, 2022, the Company entered into a...	{ "analysis": "Settlement payment of \$400,000 to a forme..."	LOW	[ "LEGAL" ]	3,374	837
"to changes in the Company's business, industry trends,..."	{ "analysis": "Failure to adapt to changes in competitive..."	MEDIUM	[ "MARKET", "OPERATIONAL", "REGULATORY" ]	5,392	116

&lt; Previous 1 2 3 ... 9 Next &gt;

**\*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 *phi-3-mini-128k-instruct*

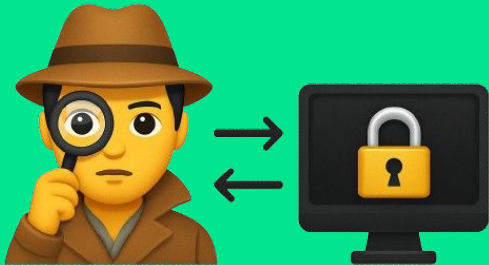
Differentially Private:  $\epsilon = 8$

Document level privacy unit



**Adversary Mode On**

# Black Box Attacks



# PII Identification & Anonymization



The screenshot shows a web interface for PII identification. At the top, there are four blue buttons: "All", "Names", "Emails", and "Phone Numbers". Below the buttons is a document snippet with the following text:

Facebook. [REDACTED] Media:  
[REDACTED], 60 [REDACTED] [REDACTED], [REDACTED].[REDACTED]@[REDACTED].com  
Investors:  
[REDACTED], 60 [REDACTED] [REDACTED], [REDACTED].[REDACTED]@[REDACTED].com  
Cautionary Statement Regarding Forward-Looking Statements

The PII is highlighted with colored boxes: a yellow box for the name, a blue box for the phone number, and a green box for the email address.

**Three low-resource NER Models used:**

Presidio (with spacy), GliNER, and a custom GliNER

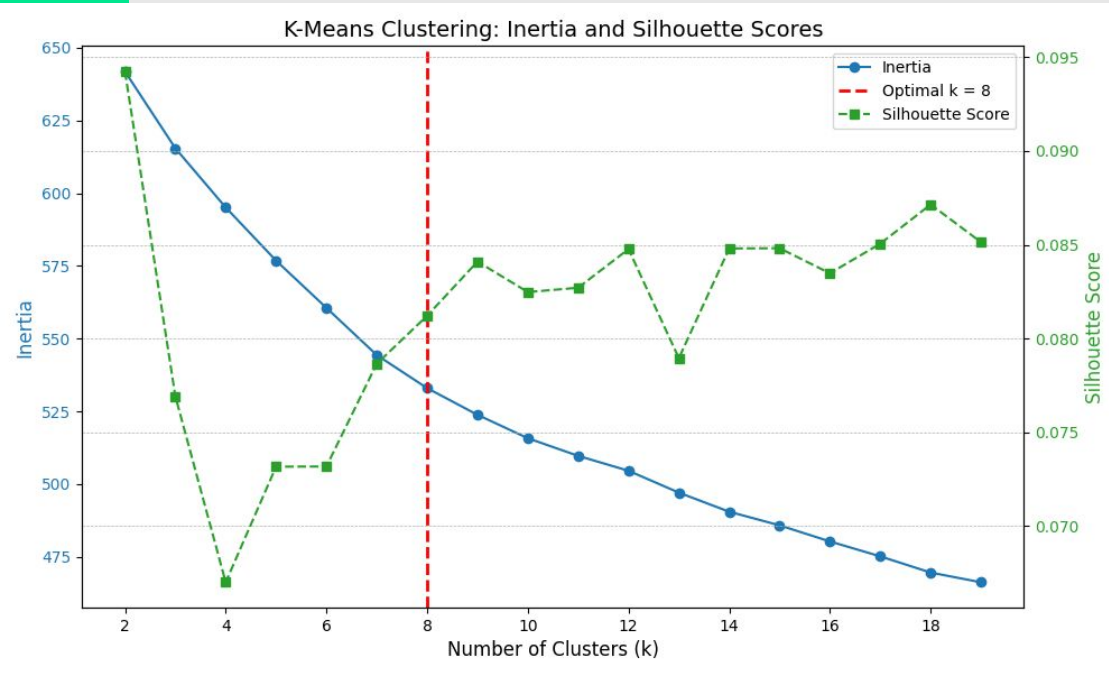
**Five PII Types Targeted**

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**

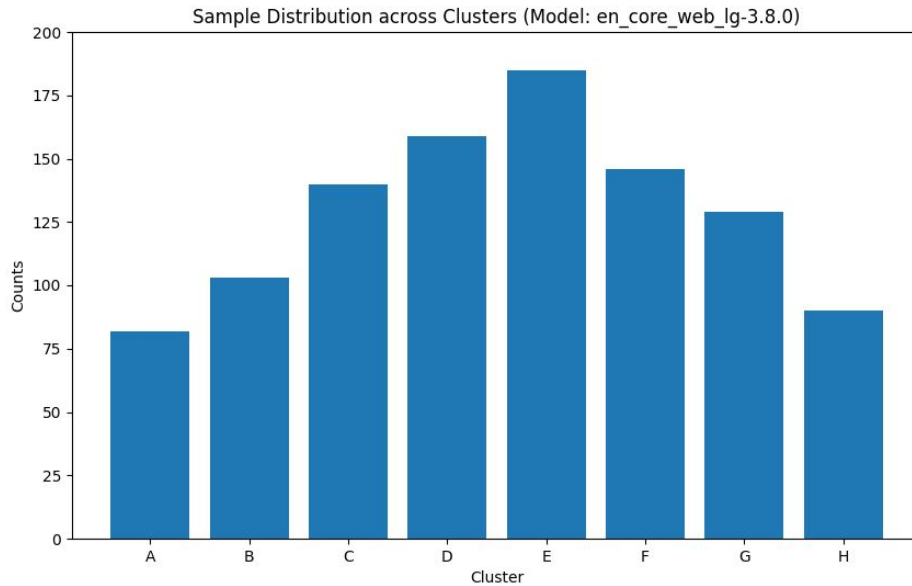
Original dataset: **3**

document types

Clustering revealed: **8**

distinct groups

# Distribution Analysis



**Near-Normal  
Distribution Pattern**

**Possible Causes -  
DP-SGD?**

# Auxiliary Information Attacks

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

October 21, 2022, the Company entered into an amendment to the Credit Agreement, dated as of July 1, 2022, by and among the Company, the lenders named therein, and JPMorgan Chase Bank, N.A., as administrative agent for the lenders (the "Amended Credit Agreement"). The Amended Credit Agreement amends and restates the Original Credit Agreement in its entirety and provides for a new credit facility of \$1.25 billion, which includes a \$1.0 billion revolving credit facility and a \$250 million term loan facility. The Amended Credit Agreement is available for review on [www.aptiv.com](http://www.aptiv.com) under the Investors section. The Company's entry into the Amended Credit Agreement is expected to provide the Company with increased financial flexibility and liquidity to support its business operations and strategic growth.

## 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

The screenshot shows a Google search interface with a dark theme. The search bar contains the query: "the lenders named therein, and JPMorgan Chase Bank, N.A., as administrative agent site:ww". Below the search bar, there are navigation tabs for "All", "News", "Images", "Videos", "Shopping", "Forums", "Web", and "More", with "All" selected. To the right of the tabs is a "Tools" link. The search results are displayed in a list format. Each result includes a red circular icon with a white dot, the domain name "aptiv.com", and a breadcrumb trail. The first result is for "2020\_aptivannualreport.pdf" with the title "Innovation in Motion" and a snippet: "with Jpmorgan chase bank, n.a., as administrative agent (the ladministrative agent"), under which it maintains senior unsecured credit facilities currently ...". The second result is for "docs > default-source > an..." with the title "2019 annual report" and a snippet: "with JPMorgan Chase Bank, N.A., as administrative agent (the 'Administrative Agent'), under which it maintains senior unsecured credit facilities currently ...". The third result is for "docs > 2021\_annualreport.pdf" with the title "Innovation in Motion" and a snippet: "Apr 3, 2022 — with JPMorgan Chase Bank, N.A., as administrative agent (the 'Administrative Agent'), under which it maintains senior unsecured credit ...". Below the snippet for the third result, it says "144 pages".

Google

the lenders named therein, and JPMorgan Chase Bank, N.A., as administrative agent site:ww X

All News Images Videos Shopping Forums Web : More Tools

aptiv.com  
https://www.aptiv.com > 2020\_aptivannualreport PDF

### Innovation in Motion

with Jpmorgan chase bank, n.a., as administrative agent (the ladministrative agent"), under which it maintains senior unsecured credit facilities currently ...

aptiv.com  
https://www.aptiv.com > docs > default-source > an... PDF

### 2019 annual report

with JPMorgan Chase Bank, N.A., as administrative agent (the "Administrative Agent"), under which it maintains senior unsecured credit facilities currently ...

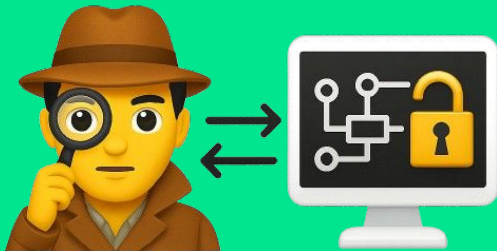
Aptiv  
https://www.aptiv.com > docs > 2021\_annualreport PDF

### Innovation in Motion

Apr 3, 2022 — with JPMorgan Chase Bank, N.A., as administrative agent (the "Administrative Agent"), under which it maintains senior unsecured credit ...

144 pages

# 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

**Company:** TechCorp Inc.  
**CIK:** 0001234567  
**Filing Date:** March 15, 2025  
**Fiscal Year End:** December 31, 2024

Business

Financials

Risks

## Financials

TechCorp's financial performance during fiscal year 2024 demonstrated solid growth across key operational metrics. **Presenting on empirical attack on DP synthetic data at PEPR** The company maintained strong cash flow generation and improved operational efficiency throughout the year. Our balance sheet remains robust with sufficient liquidity to fund ongoing operations and strategic initiatives. We continue to invest in research and development while maintaining disciplined cost management practices. The company's financial position provides a strong foundation for future growth opportunities and enables us to weather potential economic

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

## 🕒 Commit History

**Update README.md (#3)** 9672cc5 

 committed on Jan 15

**Include complete list of privacy parameters in README. (#2)** 5bf4dd4 

 committed on Jan 8



For more  
details  
read the  
SynthLeak  
Blog!