



Ask me anything. Your sensitive information is automatically protected.

Type your message here...



Governing Open-Vocabulary Data Leaks using an Edge LLM through Programming by Example

Qiyu Li Jinhe Wen Haojian Jin



UC San Diego

Problem

Data Leak in ChatGPT Prompts

TECH

Amazon, Apple, and 12 other major companies that have restricted employees from using ChatGPT

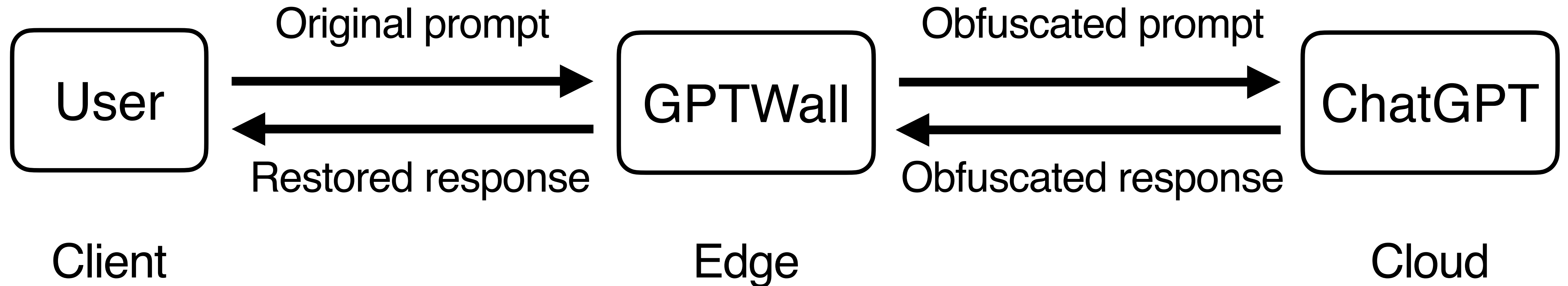
BREAKING | BUSINESS

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Employees are accidentally leaking company data through ChatGPT, report warns

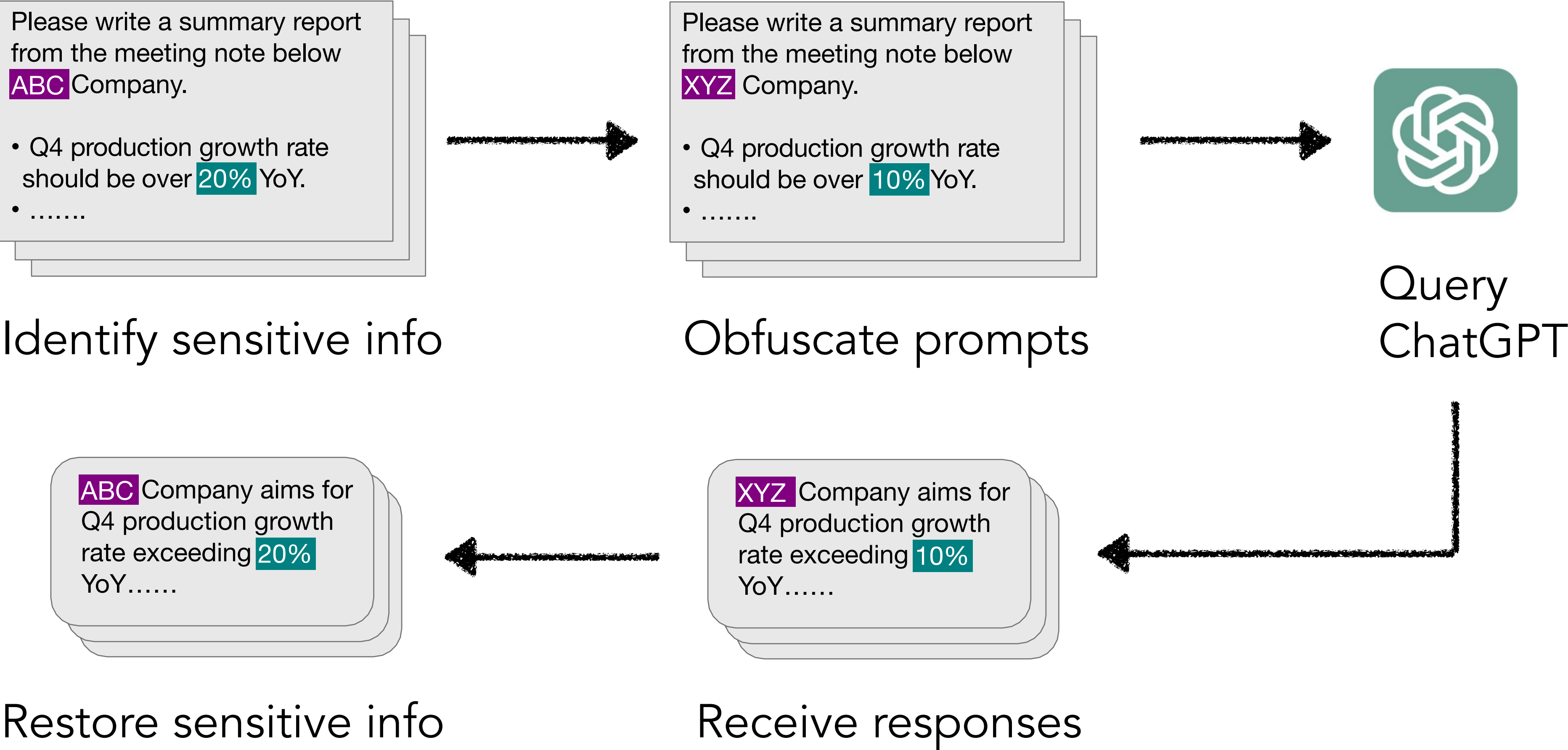
A new Cyera report reveals ChatGPT, Microsoft Copilot, and other AI tools are now the leading source of workplace data leaks, often occurring without companies noticing.

GPTWall: A Privacy Firewall for Mitigating ChatGPT Data Leaks



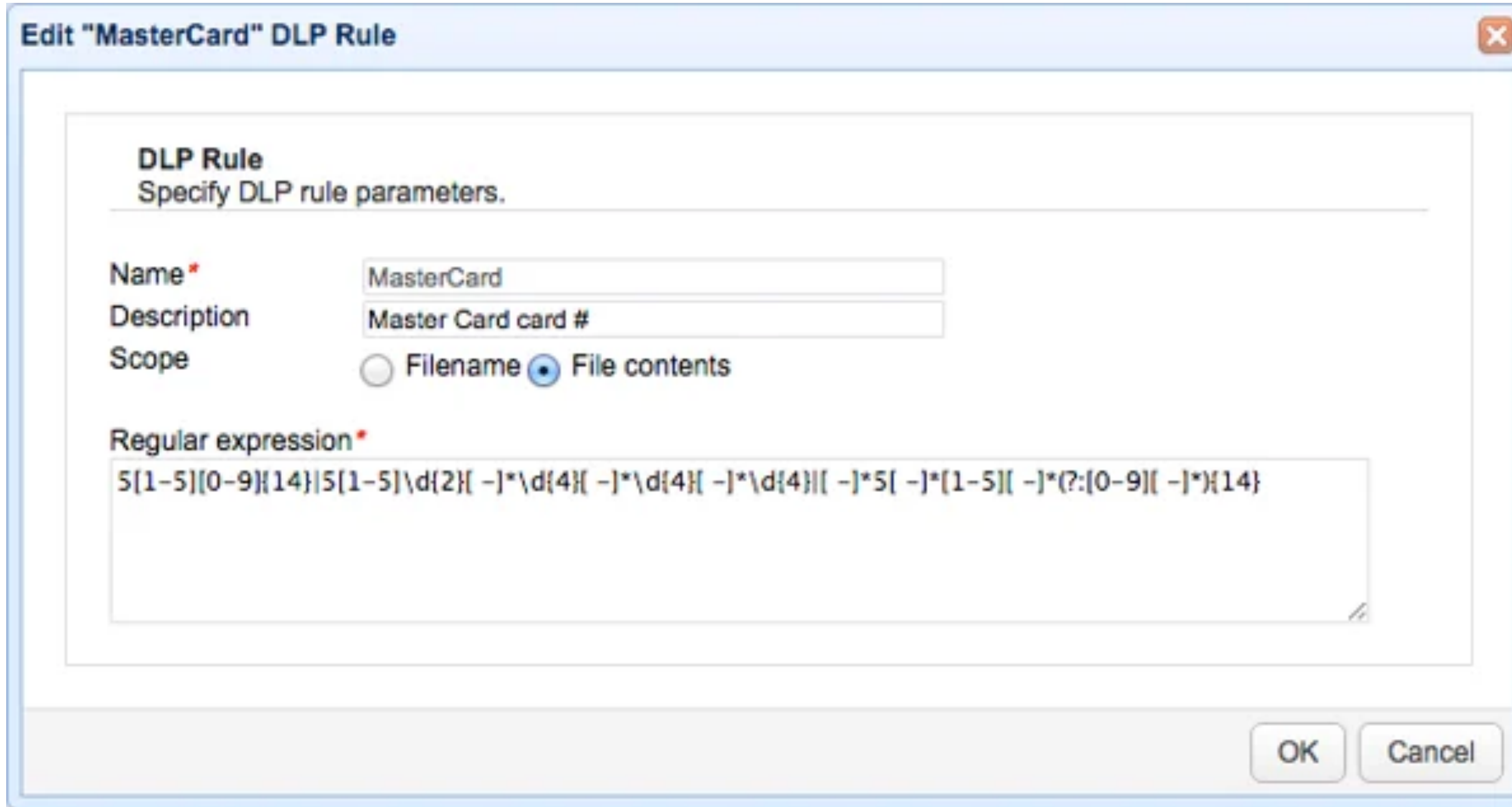
Example Data Flow

Policy: Replace `company_name` with synthetic values; Add noise to `financial_figure`



Related Work

Data Leak Prevention – Regex Rules



Edit "MasterCard" DLP Rule

DLP Rule
Specify DLP rule parameters.

Name*

Description

Scope ☐ Filename ☒ File contents

Regular expression*

OK Cancel

Related Work

Data Leak Prevention – Named Entity Recognition

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's **PaperAdvertisementSupported** **ORG** by F.B.I. Agent **Peter Strzok** **PERSON**,
Who Criticized Trump **PERSON** in Texts, Is FiredImagePeter Strzok, a top **F.B.I.** **GPE** counterintelligence agent who was taken off the special counsel
investigation after his disparaging texts about President **Trump** **PERSON** were uncovered, was fired. **CreditT.J. Kirkpatrick** **PERSON** for **The New York**
TimesBy Adam Goldman **ORG** and **Michael S. SchmidtAug** **PERSON**. **13** **CARDINAL**, **2018WASHINGTON** **CARDINAL** — **Peter Strzok**
PERSON, the **F.B.I.** **GPE** senior counterintelligence agent who disparaged President **Trump** **PERSON** in inflammatory text messages and helped
oversee the **Hillary Clinton** **PERSON** email and **Russia** **GPE** investigations, has been fired for violating bureau policies, Mr. **Strzok** **PERSON**'s lawyer
said **Monday** **DATE**. Mr. Trump and his allies seized on the texts — exchanged during the **2016** **DATE** campaign with a former **F.B.I.** **GPE** lawyer,
Lisa Page — in **PERSON** assailing the **Russia** **GPE** investigation as an illegitimate “witch hunt.” Mr. **Strzok** **PERSON**, who rose over **20 years**
DATE at the **F.B.I.** **GPE** to become one of its most experienced counterintelligence agents, was a key figure in **the early months** **DATE** of the
inquiry. Along with writing the texts, Mr. **Strzok** **PERSON** was accused of sending a highly sensitive search warrant to his personal email account. The
F.B.I. **GPE** had been under immense political pressure by Mr. **Trump** **PERSON** to dismiss Mr. **Strzok** **PERSON**, who was removed **last summer**
DATE from the staff of the special counsel, **Robert S. Mueller III** **PERSON**. The president has repeatedly denounced Mr. **Strzok** **PERSON** in posts on

Challenges

#1 Open-vocabulary & #2 Context-dependent

PII

Personal information

Financial data

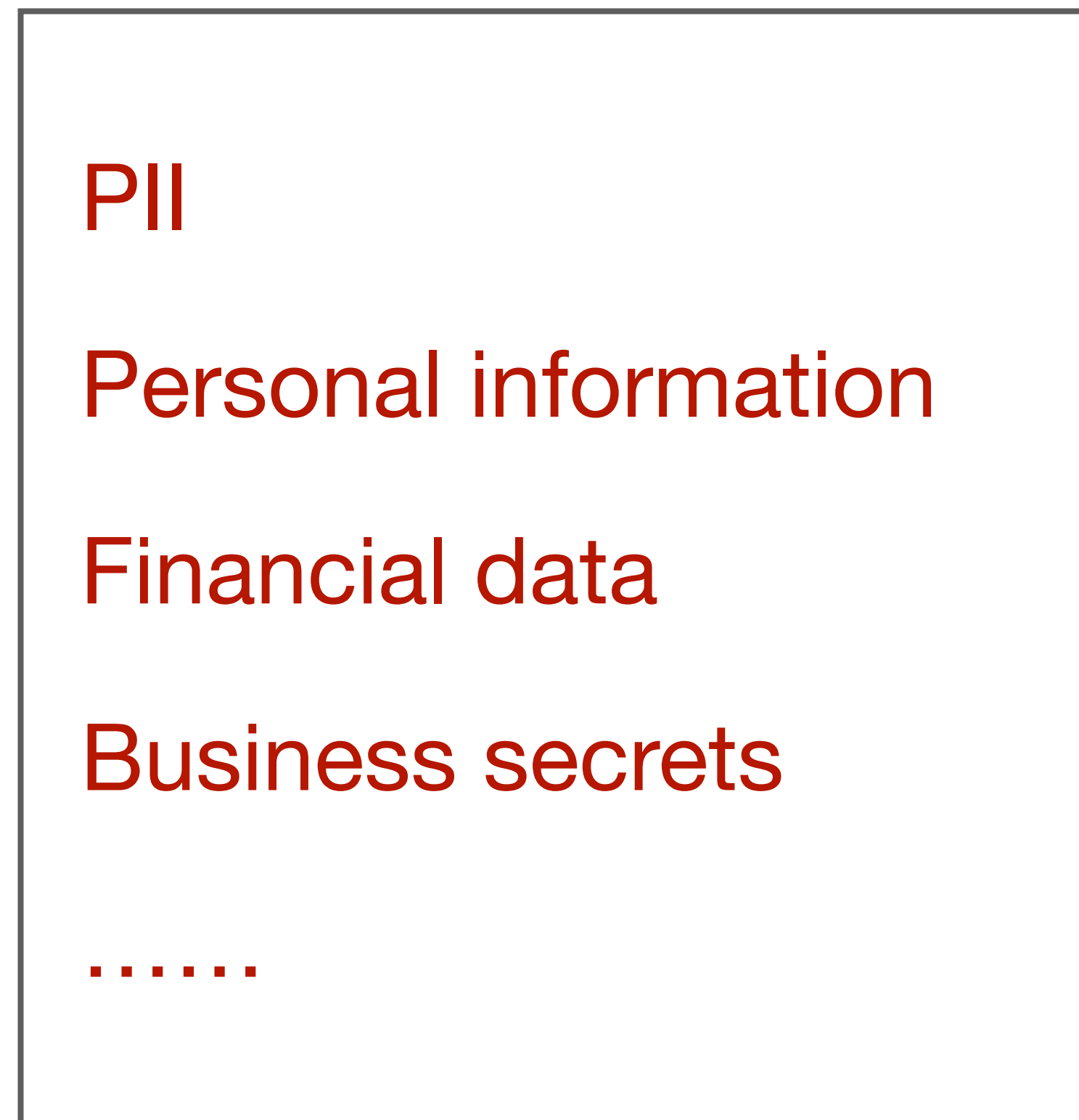
Business secrets

.....

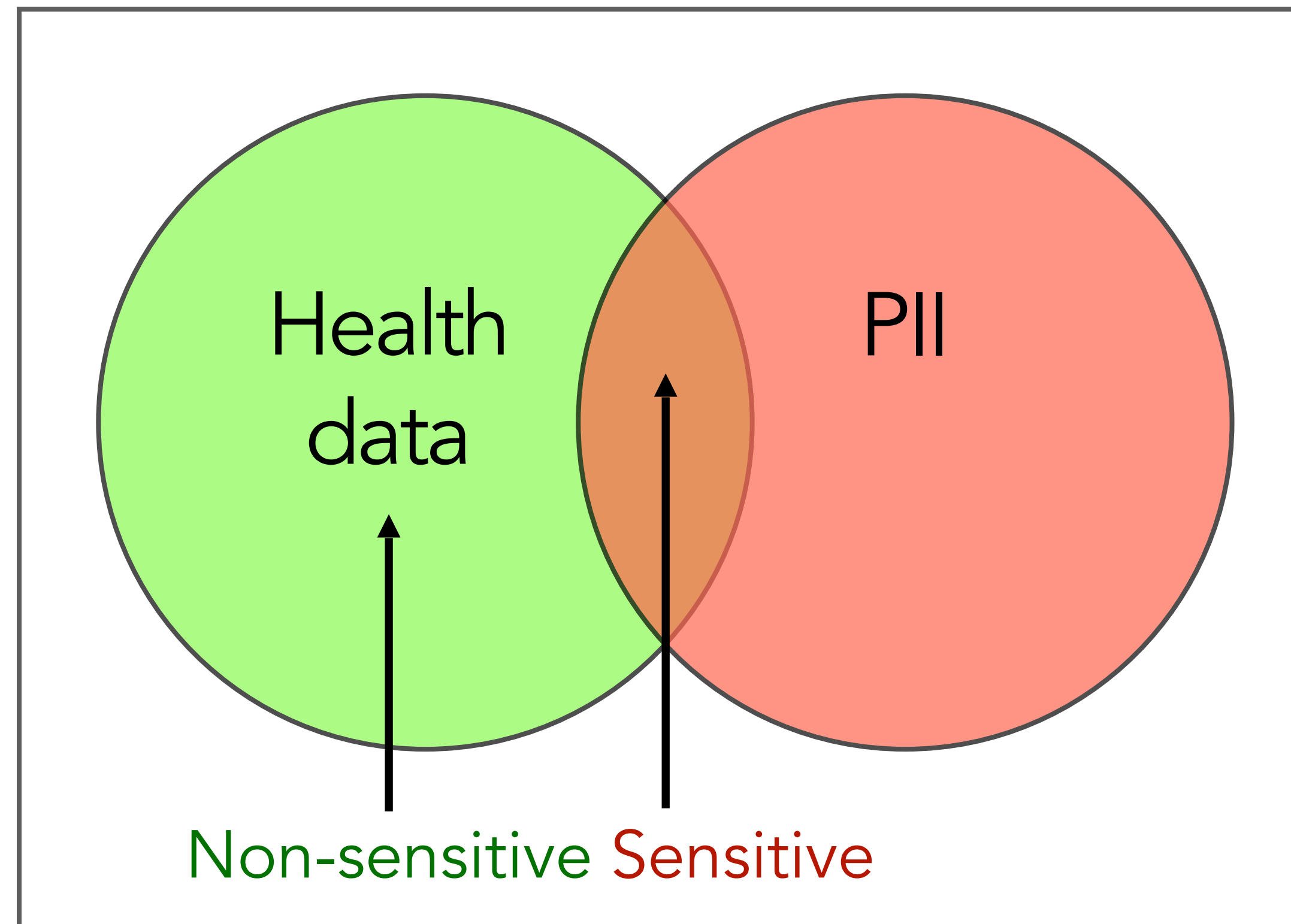
Challenge #1

Challenges

#1 Open-vocabulary & #2 Context-dependent



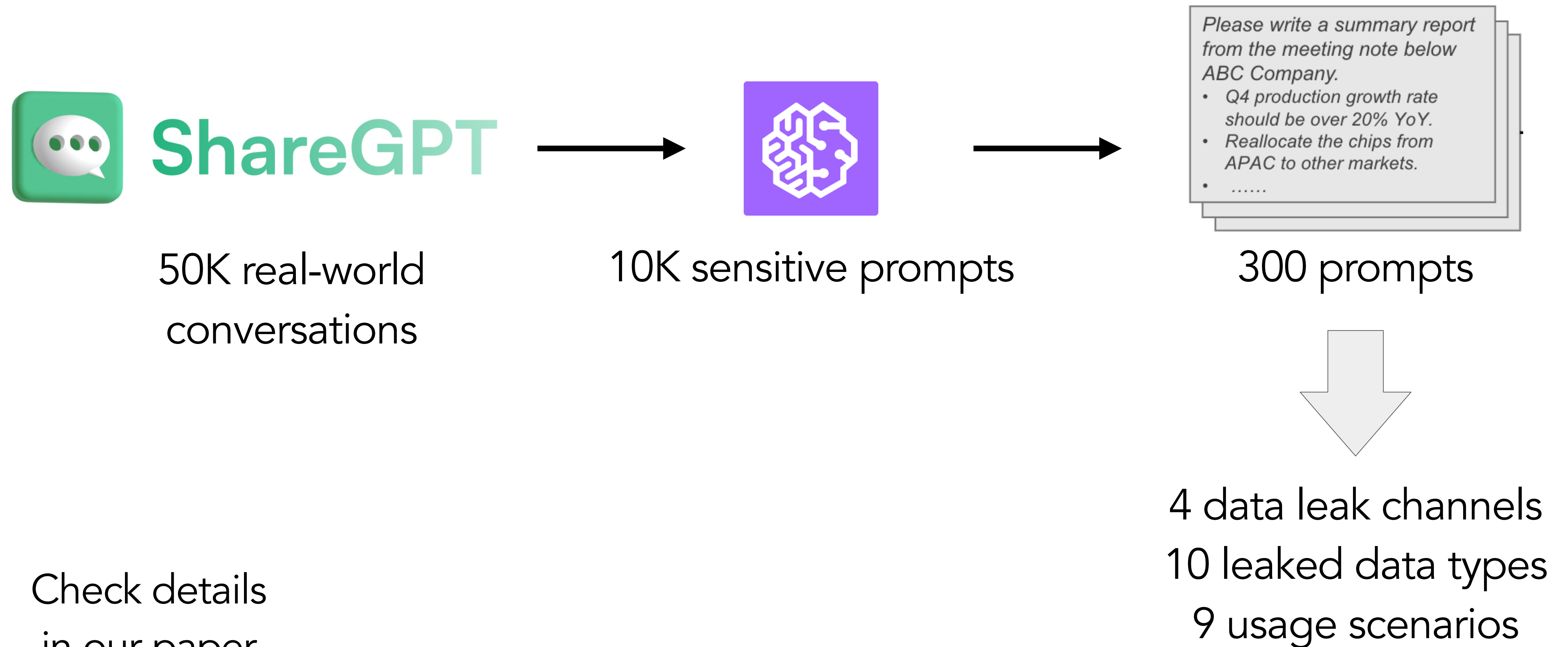
Challenge #1



Challenge #2

Method

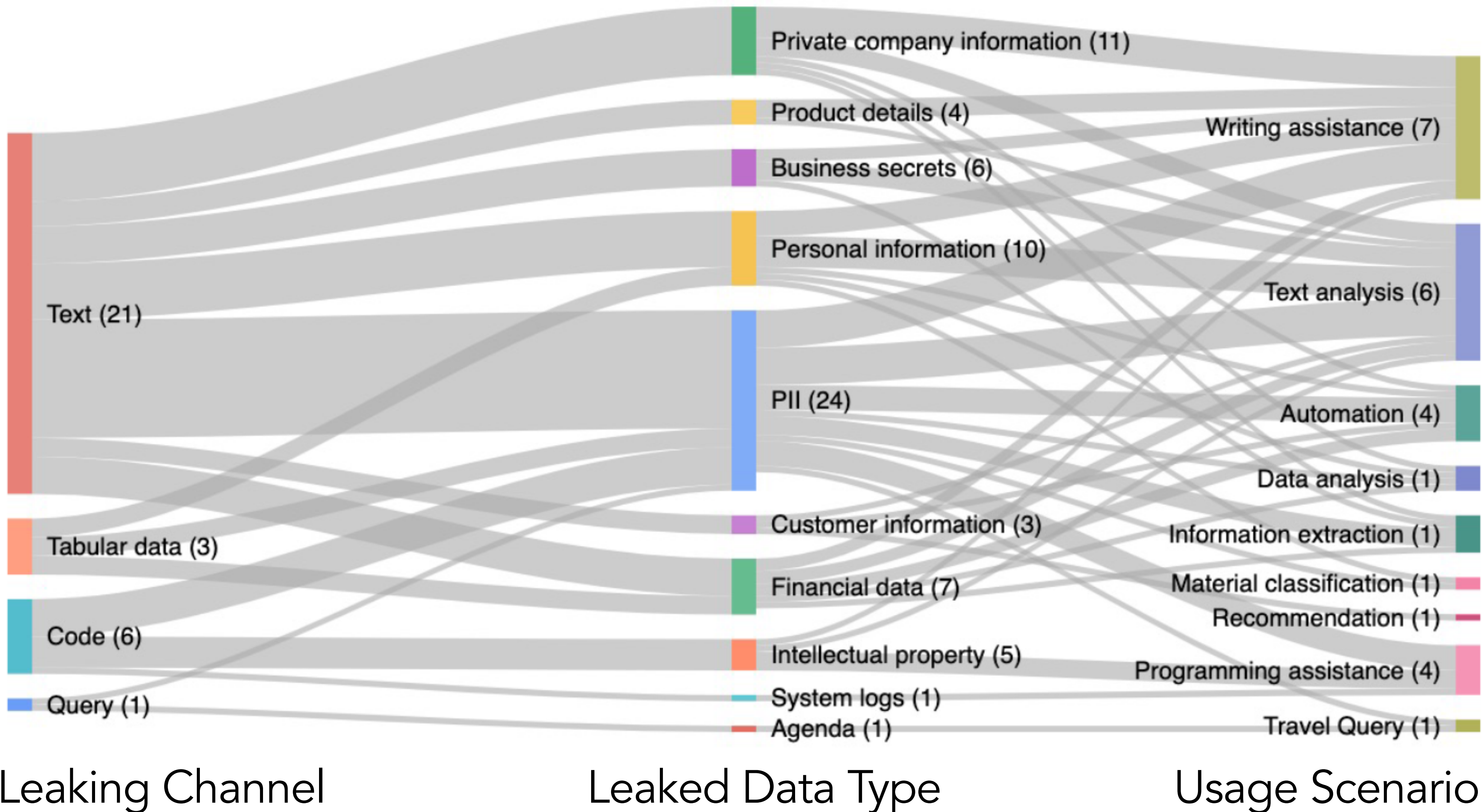
Understanding Privacy Leaks in LLM Prompts



Check details
in our paper

Results

Understanding Privacy Leaks in LLM Prompts



Use Case

Policy Representation

```
{ "method": "mask | anonymize | replace | noisify | fuzzify",  
  "information_type": {"value": {"sensitive": <value list>, "non-sensitive": <value list>},  
                        "label": {"sensitive": <label list>, "non-sensitive": <label list>}},  
  "context": <context list> }
```

- **Type-based:** e.g., personal_name
- **Value-based:** e.g., "123-45-6789" for the SSN
- **Context-based:** e.g., when health data co-exists with "personal name" in the same prompt

Obfuscation Method

Policy Representation

```
{ "method": "mask | anonymize | replace | noisify | fuzzify",  
  "information_type": {"value": {"sensitive": <value list>, "non-sensitive": <value list>},  
                        "label": {"sensitive": <label list>, "non-sensitive": <label list>}},  
  "context": <context list> }
```

- **Anonymize:** “John Doe” → “<personal_name>”
- **Mask:** “1234-5678” → “XXXX-XXXX”
- **Replace:** “John Doe” → “Jane Smith”
- **Noisify:** “\$10,000” → “\$12,000”
- **Fuzzify:** “28 years old” → “late 20s”

Annotating Data Leaks with Positive and Negative Examples

←

<

>

⊘

Mark as Sensitive

✓

Mark as Non-sensitive

✕

Clear

Recruitment Manager for Tech Companies_7

Drafting and reviewing legal documents

let me give you a time context so you can develop it better:
Born in a small town in **Lebanon** (1985),
Forced to leave his country amidst the 2006 war right after graduating with a BA in Advertising & marketing.
Climbing the ladder in the advertising industry in **Dubai** and **Riyadh** (2006-2012)
Decided to put all his savings in his family's small private school (2012-2017), where he moved back to the village to run it, modernize and rebrand it. He even demolished the existing building and created a new one, but this adventure crumbled in financial troubles that forced him to leave **Lebanon** back to **Saudi Arabia** (2017-2021), where he took on senior roles like Managing Director and Chief Marketing Officer. Worked as a consultant for some companies until he decided to move to the US to do a Masters in Information Systems with a concentration in data analytics, in an attempt to understand artificial intelligence and other advanced technologies like the blockchain that have captured his interest for a long time. He is now looking for ways to be useful in society and to humanity, contributing to something meaningful and enjoyable.

Share Prompt

Policies

>

mask **address** information

>

fuzzy **city** information

>

mask **postal_code** information

>

replace **state** information

>

replace **location** information

Suggestions

✓

⊗

↺

Policy Validation on Sample Prompts

←

<

>

⊘

Mark as Sensitive

✓

Mark as Non-sensitive

✕

Clear

Recruitment Manager for Tech Companies_7

Drafting and reviewing legal documents

let me give you a time context so you can develop it better:

Born in a small town in Lebanon (1985),

Forced to leave his country amidst the 2006 war right after graduating with a BA in Advertising & marketing.

Climbing the ladder in the advertising industry in Dubai and Riyadh (2006-2012)

Decided to put all his savings in his family's small private school (2012-2017), where he moved back to the village to run it, modernize and rebrand it. He even demolished the existing building and created a new one, but this adventure crumbled in financial troubles that forced him to leave Lebanon back to Saudi Arabia (2017-2021), where he took on senior roles like Managing Director and Chief Marketing Officer. Worked as a consultant for some companies until he decided to move to the US to do a Masters in Information Systems with a concentration in data analytics, in an attempt to understand artificial intelligence and other advanced technologies like the blockchain that have captured his interest for a long time. He is now looking for ways to be useful in society and to humanity, contributing to something meaningful and enjoyable.

Share Prompt

Policies

Download

> mask address information

> fuzzy city information

> mask postal_code information

> replace state information

> replace location information

Suggestions

Checkmark X Refresh

birth_year except duration_year

birth-related information except duration_year

year-related information except duration_year

1985

Iterating Policies with Real-time Feedback

←			Policies ↓
☐	Recruitment Manager for Tech Co...	...etter: Born in a small town in Lebanon (1985 (birth_year)), Forced to leave his country amidst the 2006 war right after gra	> mask address information
☐	82-year-old routine checkup_7	...se but ultimately vaginally delivers an 11 pound boy (birth_weight) . On post operative day 2, she is noted to have uterine	> fuzzy city information
☐	Last Will and Testament of Rosaly...	rewrite with son's names: Warren (1/20/64 (birth_date)), Brian (2/12/64 and Benjamin (4/26/74)	> mask postal_code information
☐	Last Will and Testament of Rosaly...	...h son's names: Warren (1/20/64), Brian (2/12/64 (birth_date)) and Benjamin (4/26/74)	> replace state information
☐	Last Will and Testament of Rosaly...	...(1/20/64), Brian (2/12/64 and Benjamin (4/26/74 (birth_date))	> replace location information
			Suggestions ✓ ✕ ↺
			birth_year except event_year
			birth-related information except event_year
			year-related information except event_year
			1985

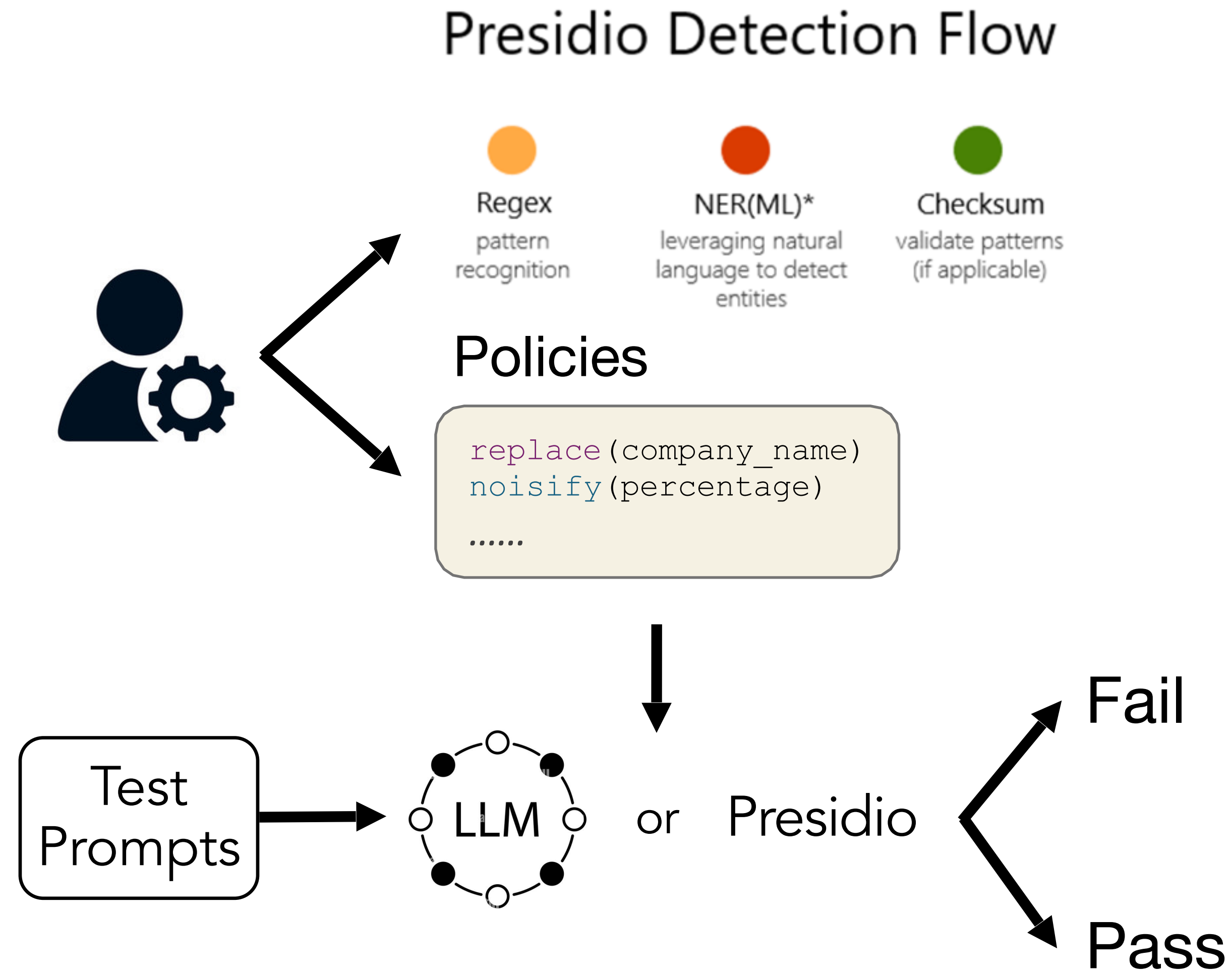
Evaluation

1. Effectiveness in identifying sensitive information
2. Policy usability for admins
3. Mitigation of inference attacks
4. Impact on response quality
5. System performance

Check details
in our paper

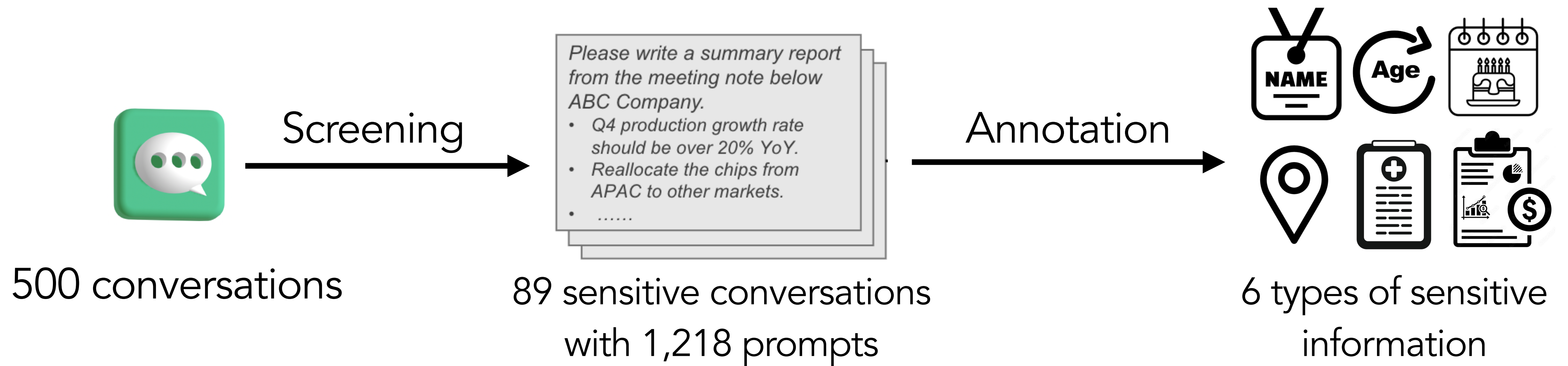
Method

Policy Usability for Admins



Dataset

Policy Usability for Admins



- Quasi-identifiers
- Personal Identifier Information (PII)
- Sensitive Data

Results

Policy Usability for Admins

1.6 v.s. 5.0 minutes
for a single policy

Precision: **84% v.s. 55%**
Recall: **76% v.s. 54%**

Task	Presidio					GPTWall				
	Time (min)	#Rule	#Success	Precision	Recall	Time (min)	#Policy	#Success	Precision	Recall
Name	4.49 ±2.94	1.0	6	63.69%	86.52%	10.50±4.00	4.0	6	77.41%	87.80%
Location	12.80±7.23	1.4	3	67.78%	35.88%	12.52 ±7.82	3.8	5	82.98%	68.29%
Birthdate	11.72±5.75	1.7	2	49.71%	55.10%	7.52 ±4.93	2.7	5	98.33%	97.57%
Age	12.21±7.76	2.4	3	73.71%	36.83%	4.38 ±3.32	1.2	6	93.73%	97.95%
Financial	9.25 ±4.22	2.0	3	37.79%	54.55%	12.35±6.08	11.3	6	52.37%	85.98%
Health	13.22±7.86	4.4	5	50.12%	63.89%	9.21 ±1.50	12.2	6	83.88%	61.19%
Average	10.61±6.54	2.1	3.7	54.60%	53.98%	9.42 ±5.48	5.9	5.7	83.91%	76.17%

GPTWall: a policy-based data leak management system

- (1) Conventional solutions are insufficient for **open-vocabulary** data leaks
- (2) Introduce a **man-in-the-middle** edge LLM service
- (3) Help admins create policies through **programming by example**

Qiyu Li
qiyuli@ucsd.edu



UC San Diego