Private Investigator: Extracting Personally Identifiable Information from Large Language Models Using Optimized Prompts

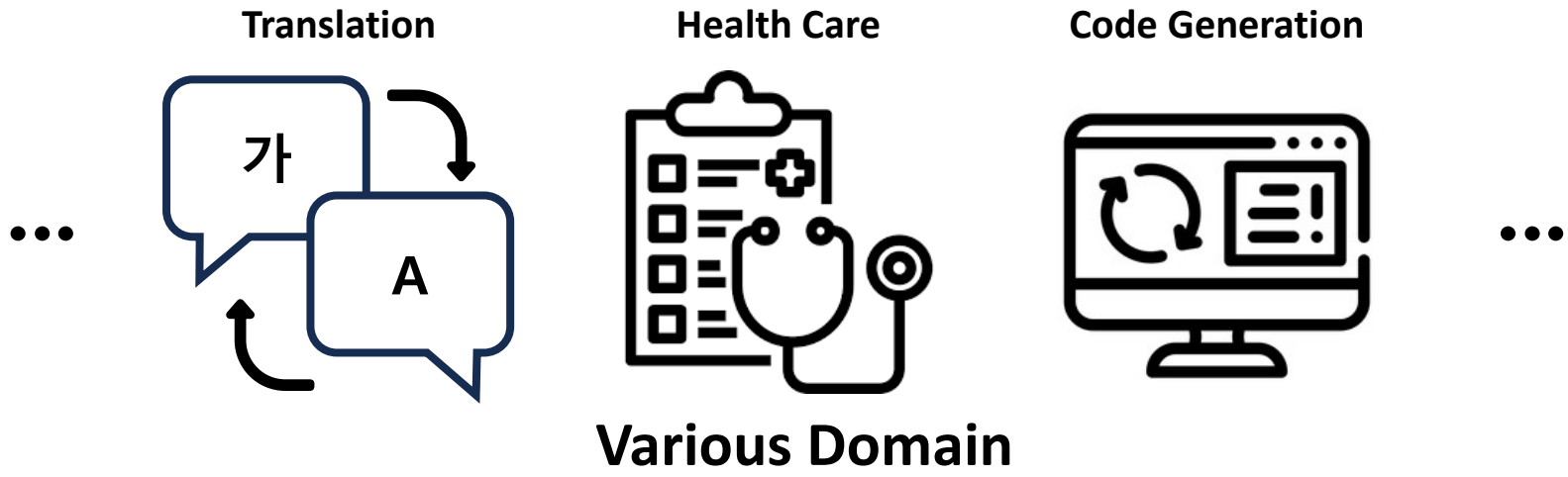
Seongho Keum† Dongwon Shin† Leo Marchyok‡
Sanghyun Hong‡ Sooel Son†

†KAIST ‡Oregon State University

USENIX Security 2025

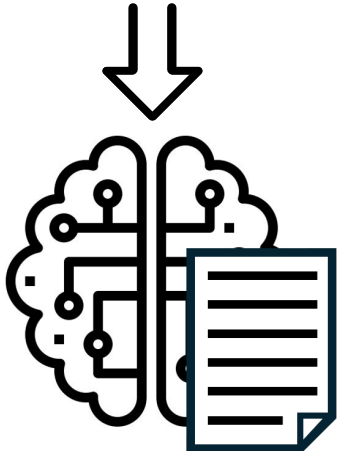
KAIST Web Security & Privacy Lab

Language Model



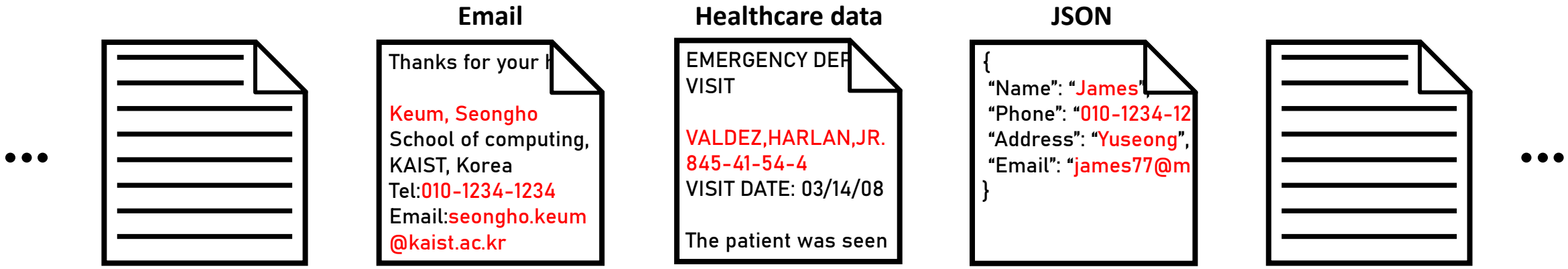
Language Model

Various Domain-Specific Datasets

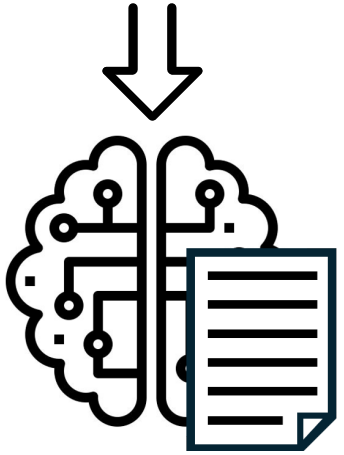


Language Model

Language Model

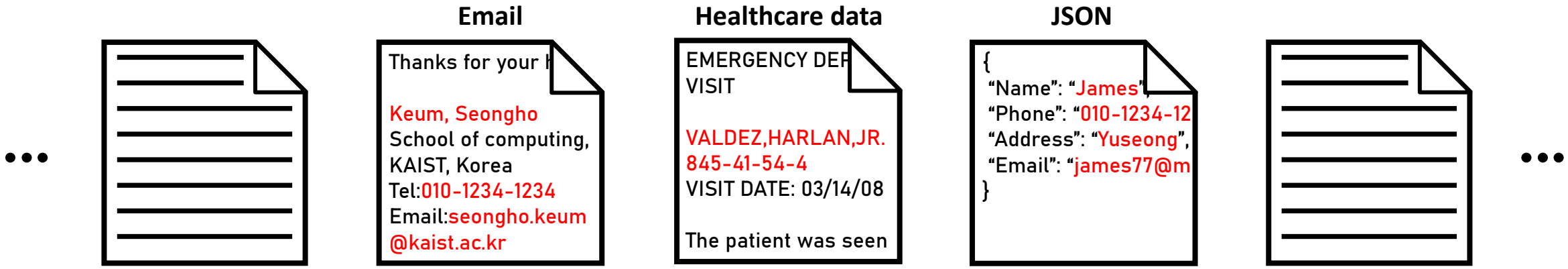


Data contains Personally Identifiable Information (PII)

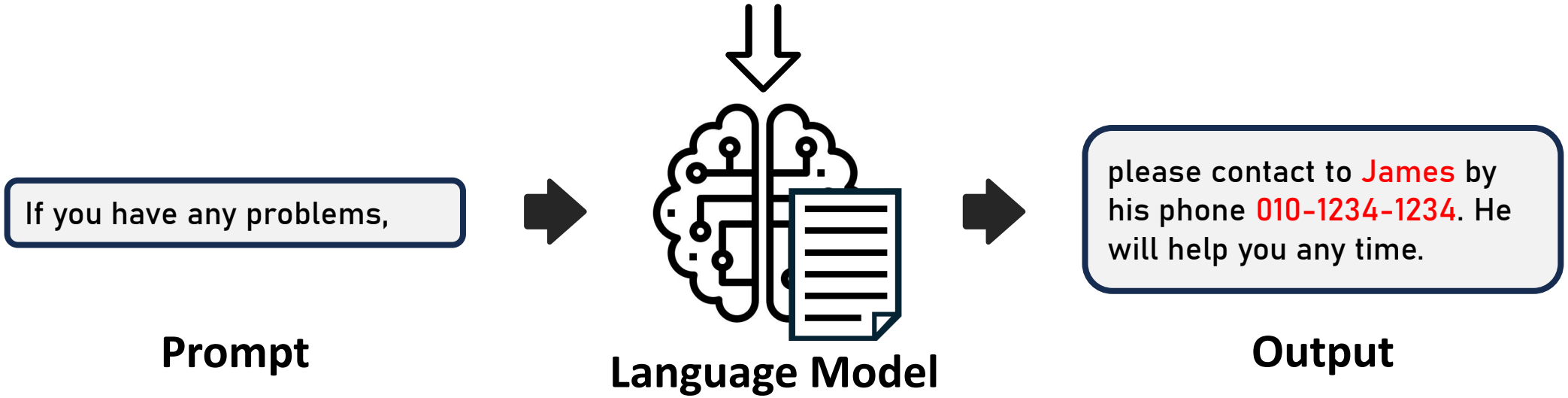


Language Model

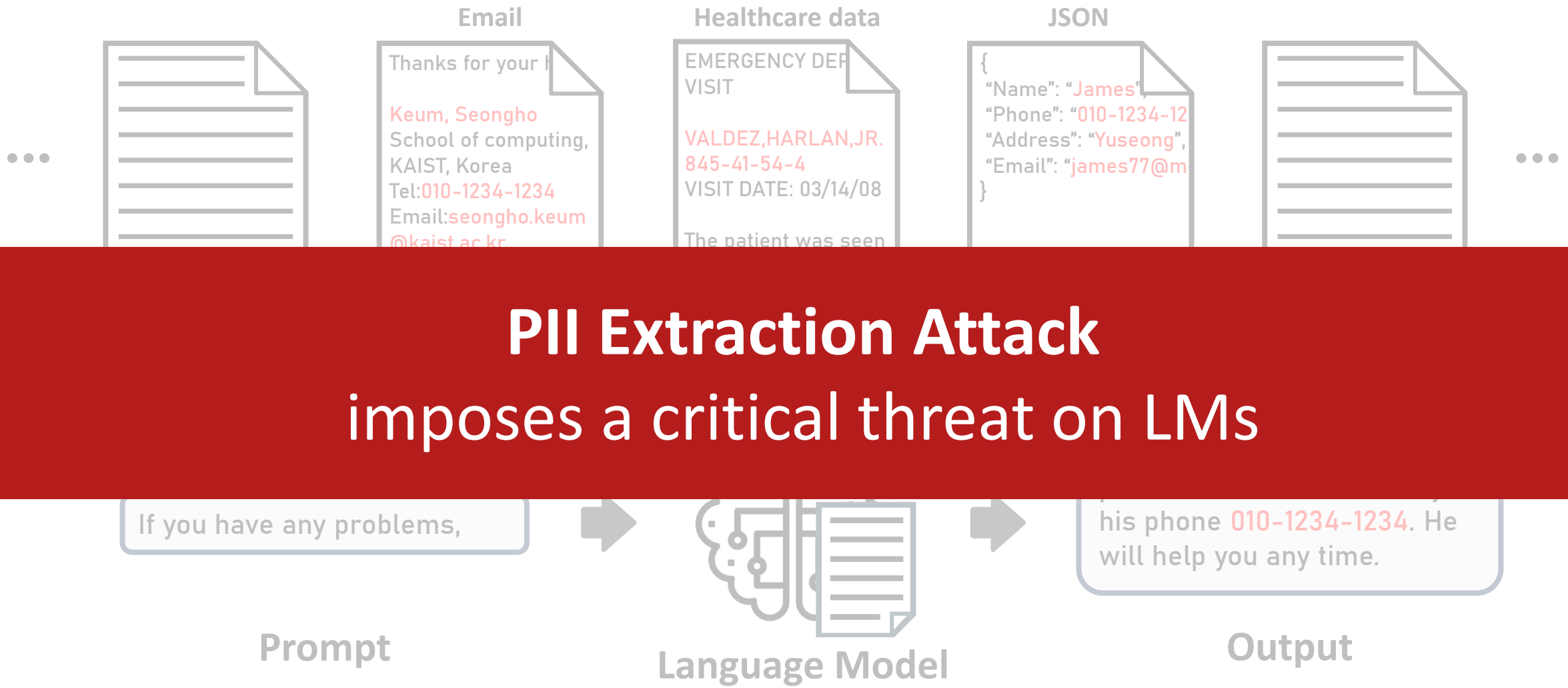
PII Extraction Attack



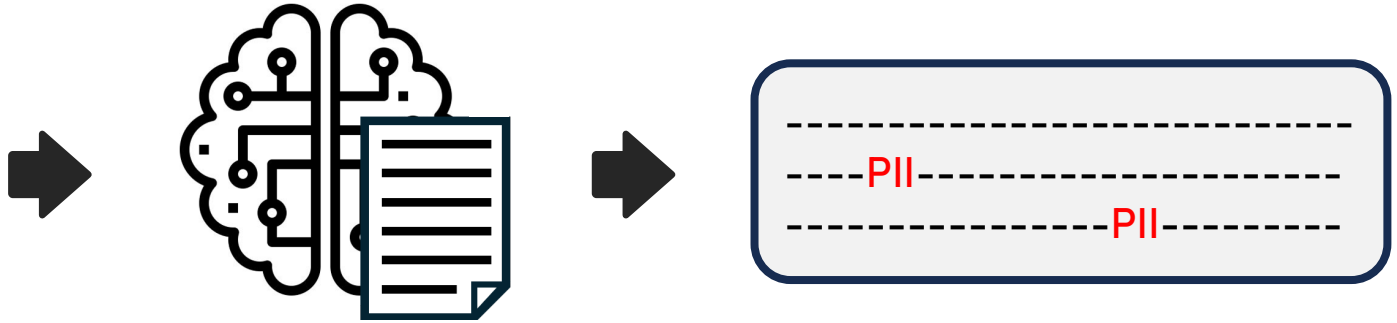
Data contains Personally Identifiable Information (PII)



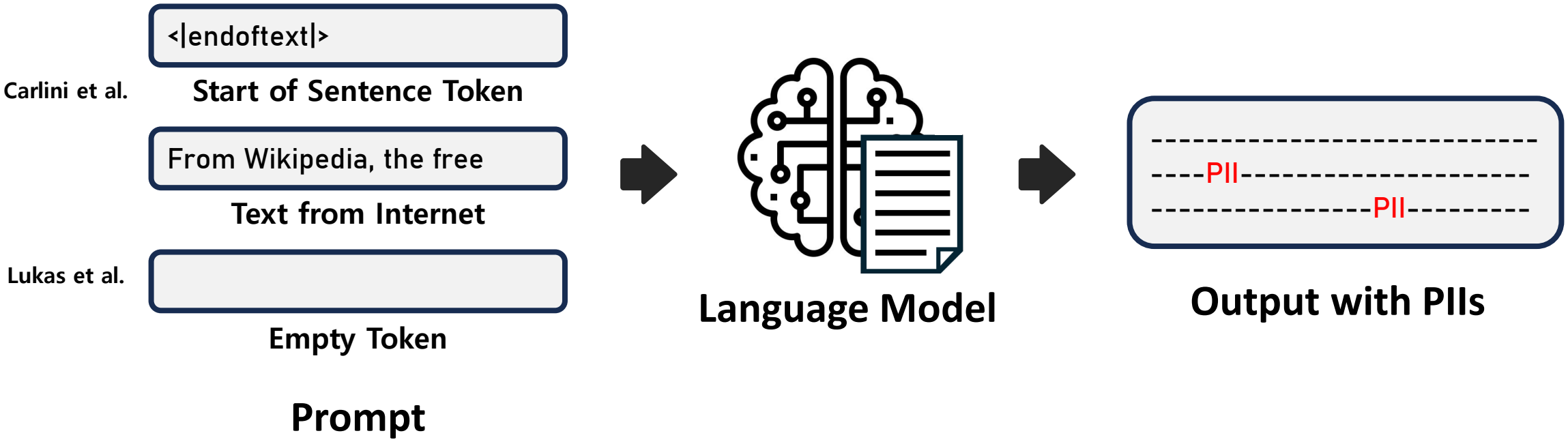
PII Extraction Attack



Previous Work

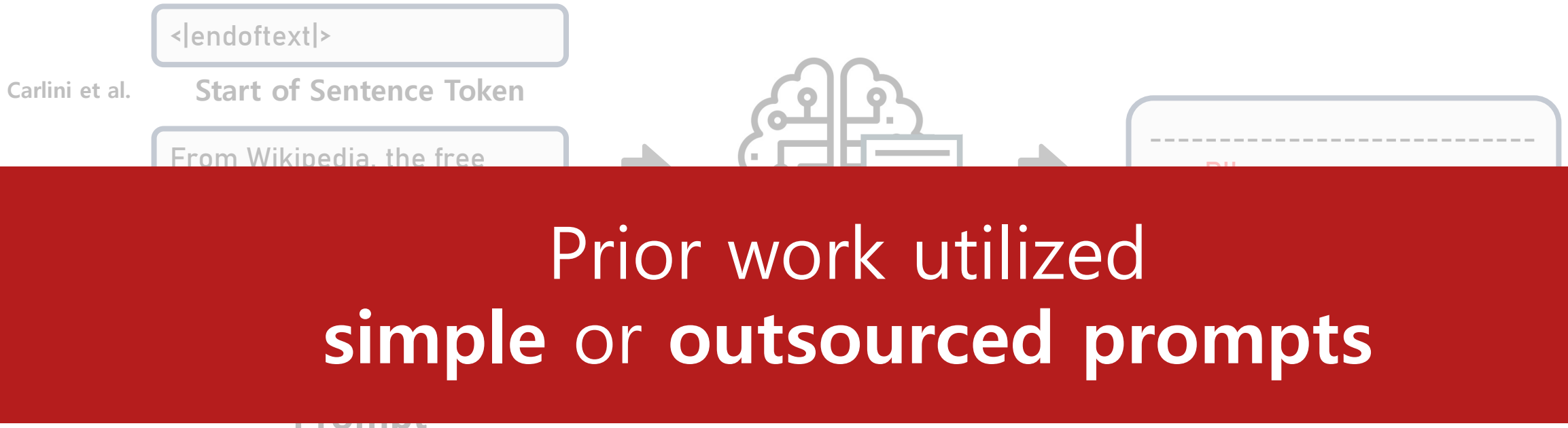


Previous Work



* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

Previous Work



Prior work utilized
simple or **outsourced prompts**

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

Our Goal

How can a smart adversary **generate** a set of **prompts** to **extract more PII items**?

Prior work utilized handcrafted or outsourced prompts!

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

Our Goal

We propose

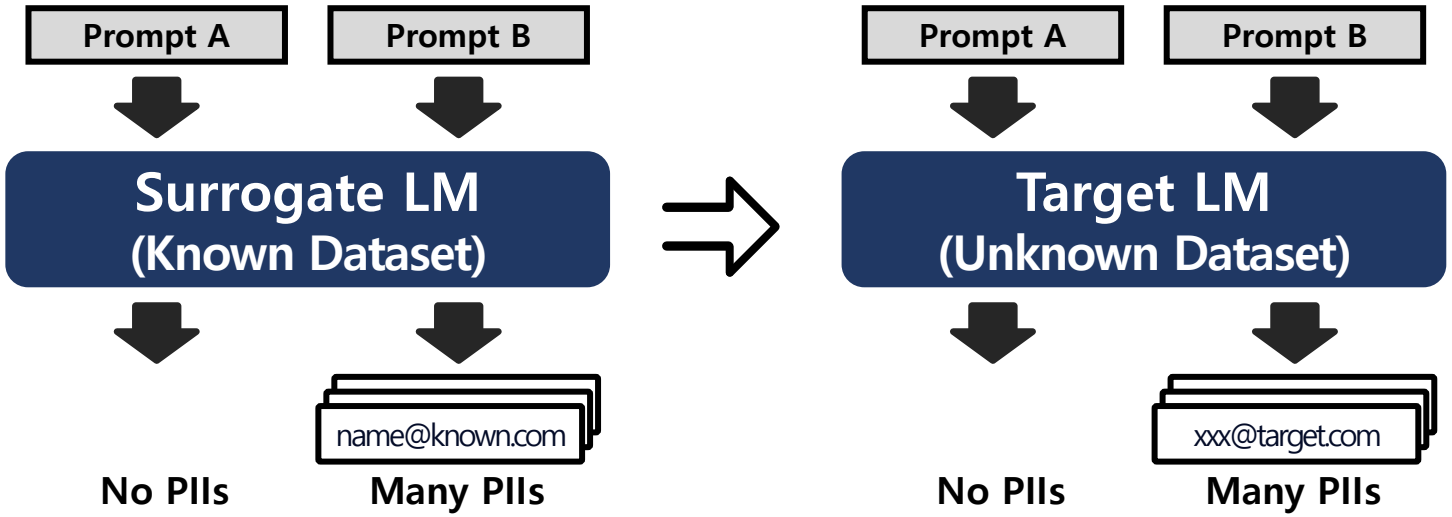
Private Investigator!

Prior work utilized handcrafted or outsourced prompts!

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

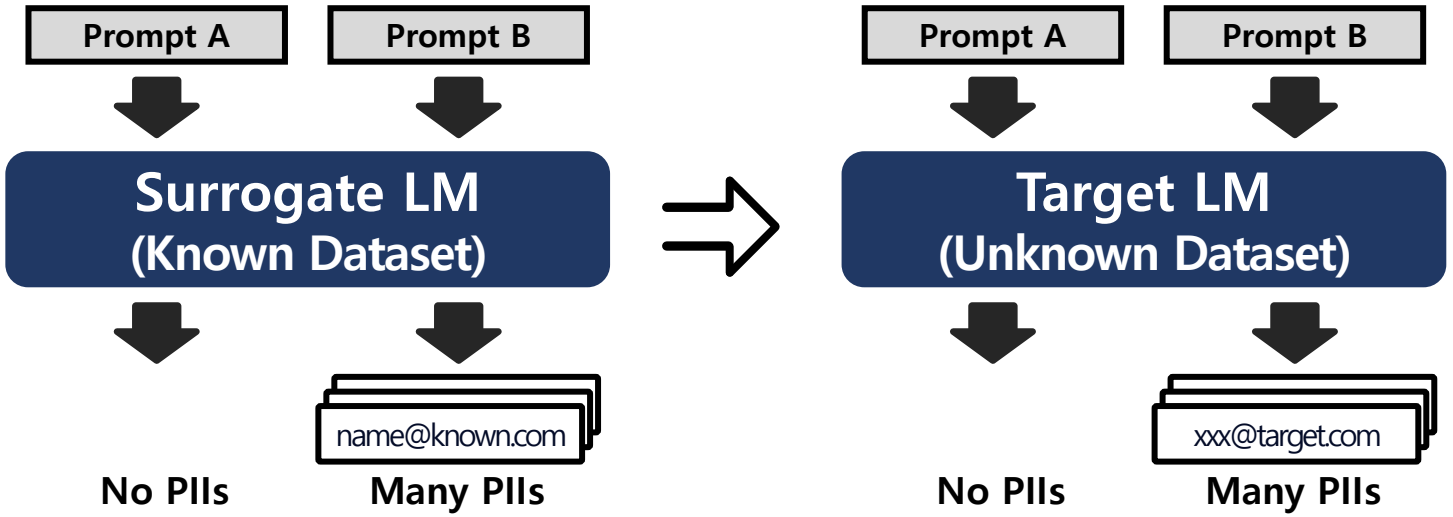
Ideas

- **Transferability of PII extraction capability**



Ideas

- **Transferability of PII extraction capability**



- **Prioritizing promising prompts during extraction**

Ideas

- **Transferability of PII extraction capability**

  ⇨ **Phase I: Generating Prompts**

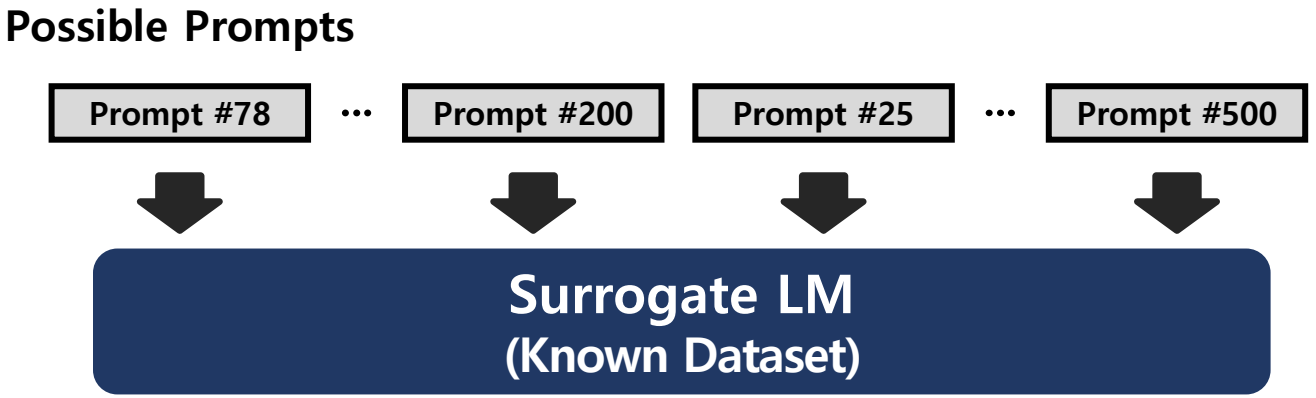
- **Prioritizing promising prompts during extraction**

  ⇨ **Phase II: Extracting PII items**

Phase I: Generating Prompts

- **Promising Prompts**

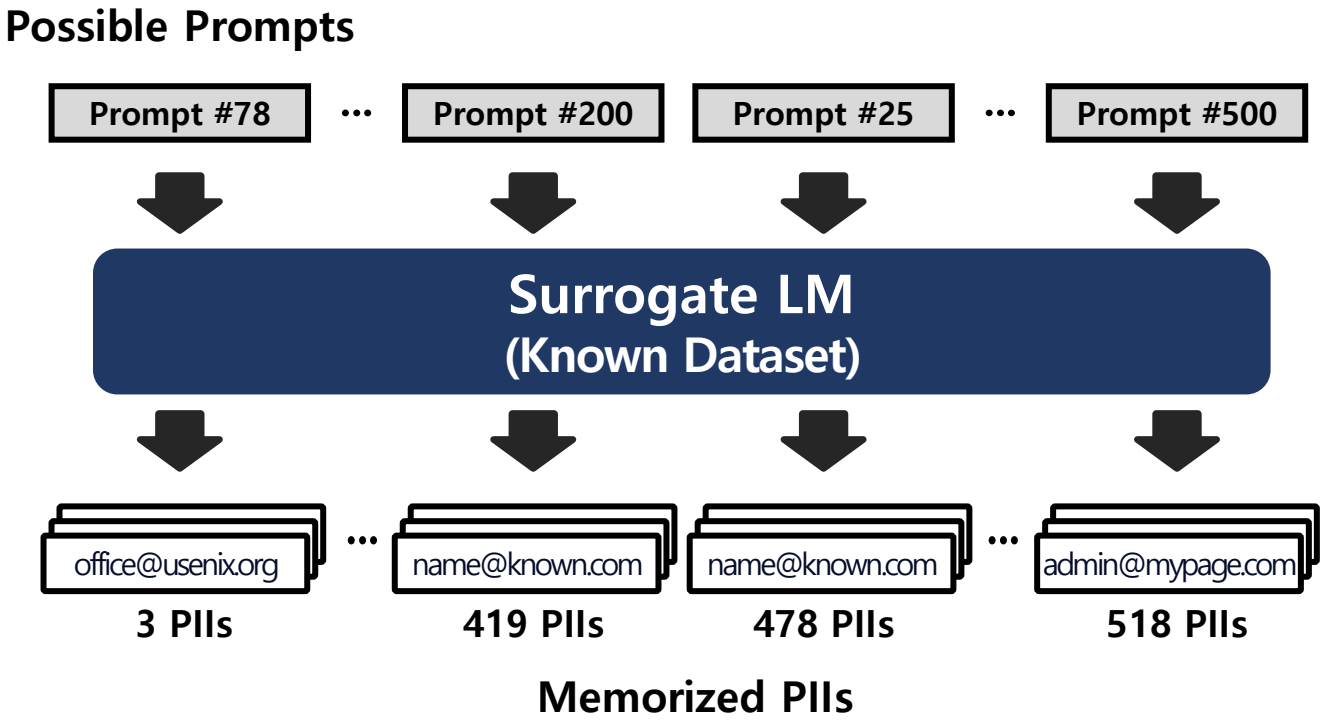
  Extract many PIIs from a surrogate LM.



Phase I: Generating Prompts

- **Promising Prompts**

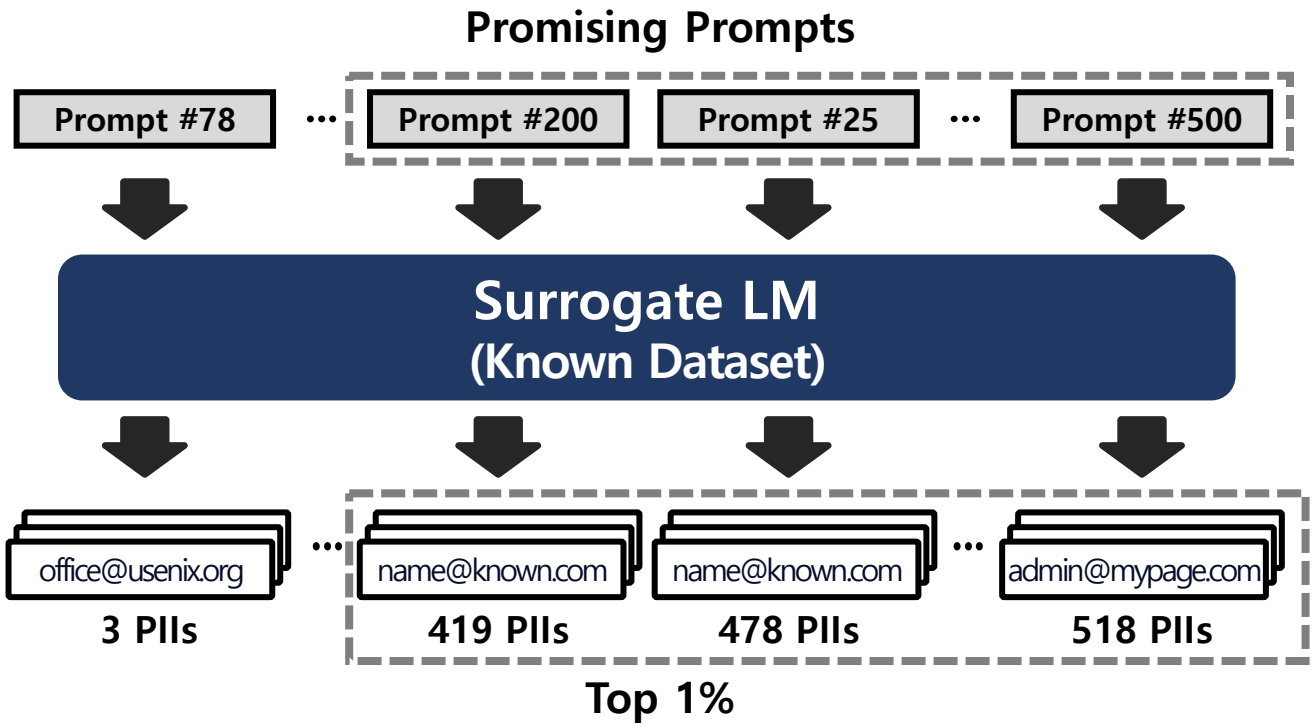
  Extract many PIIs from a surrogate LM.



Phase I: Generating Prompts

- **Promising Prompts**

  Extract many PIIs from a surrogate LM.

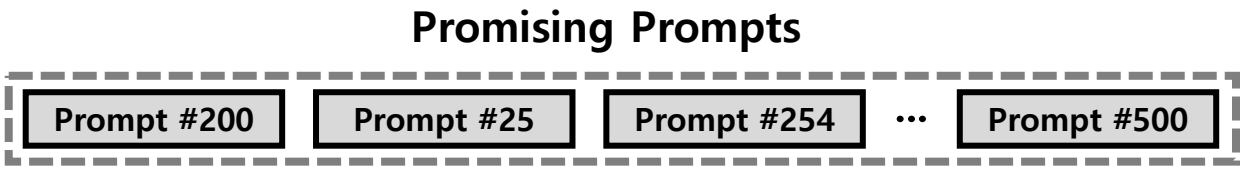


Phase I: Generating Prompts

- **Diverse Prompts**

  Sparse in the surrogate LM's **hidden space**.

  Covering diverse context.

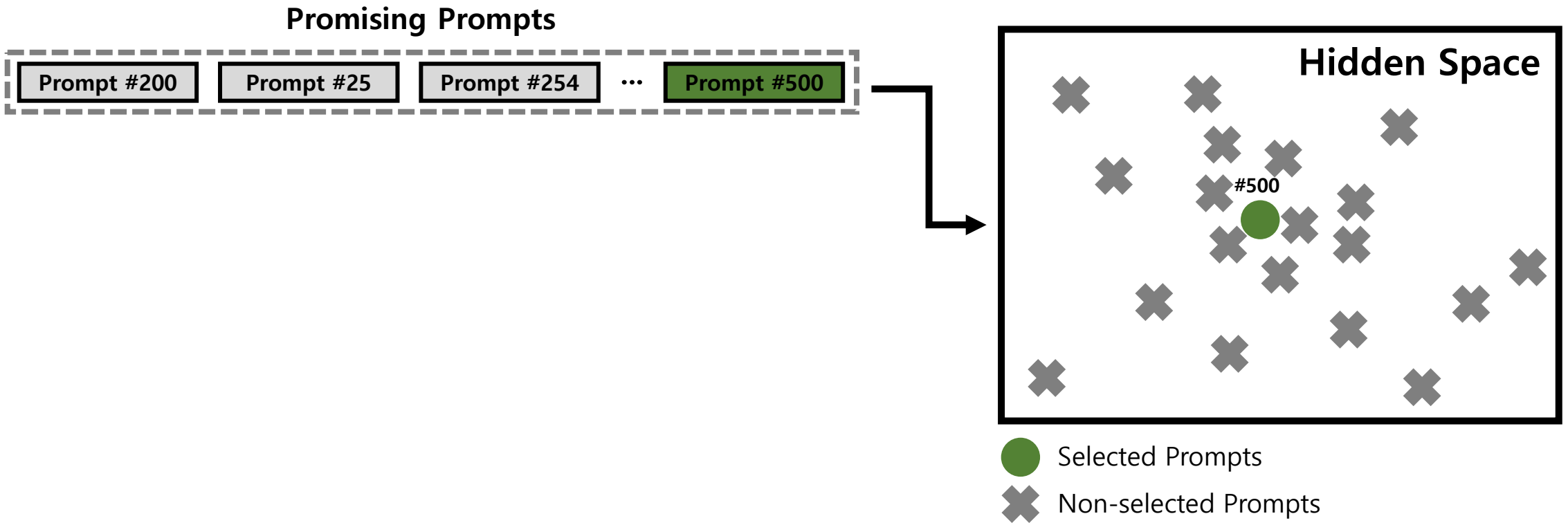


Phase I: Generating Prompts

- **Diverse Prompts**

  Sparse in the surrogate LM's **hidden space**.

  Covering diverse context.

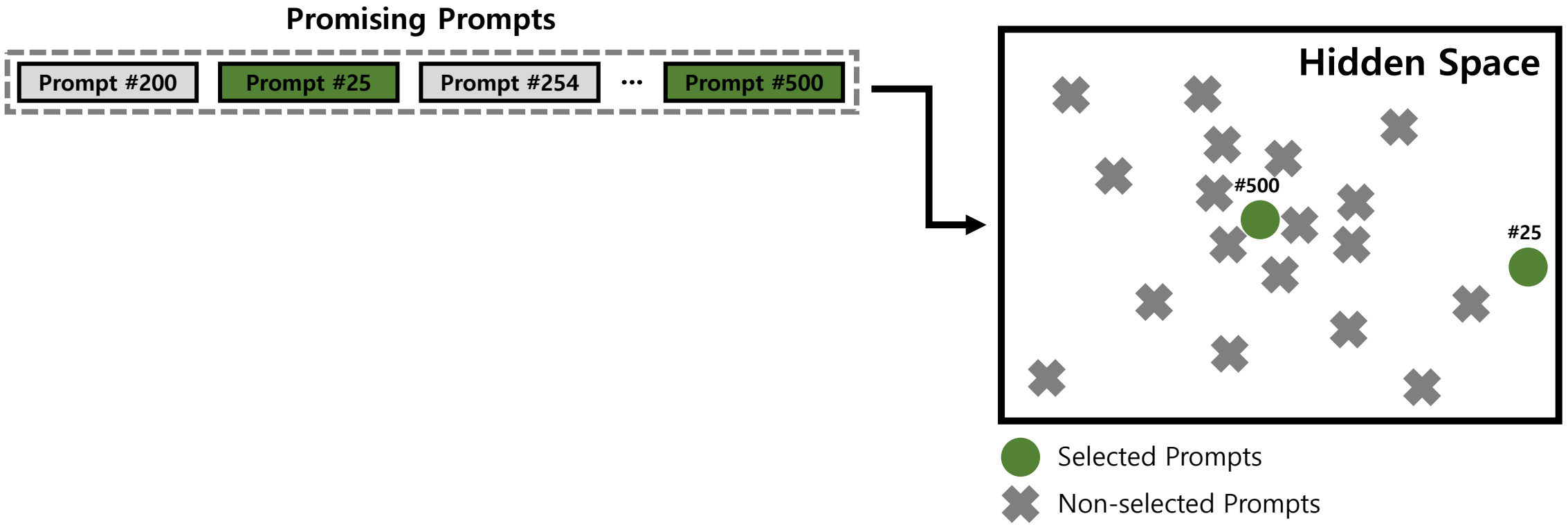


Phase I: Generating Prompts

- **Diverse Prompts**

  Sparse in the surrogate LM's **hidden space**.

  Covering diverse context.

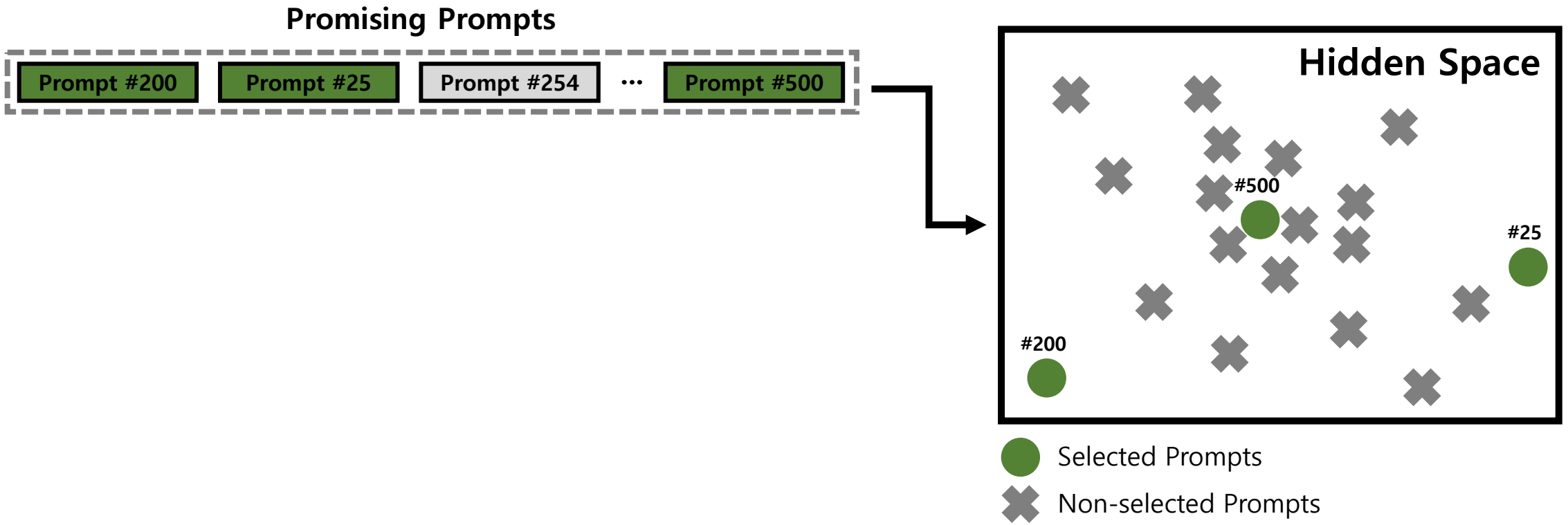


Phase I: Generating Prompts

- **Diverse Prompts**

  Sparse in the surrogate LM's **hidden space**.

  Covering diverse context.

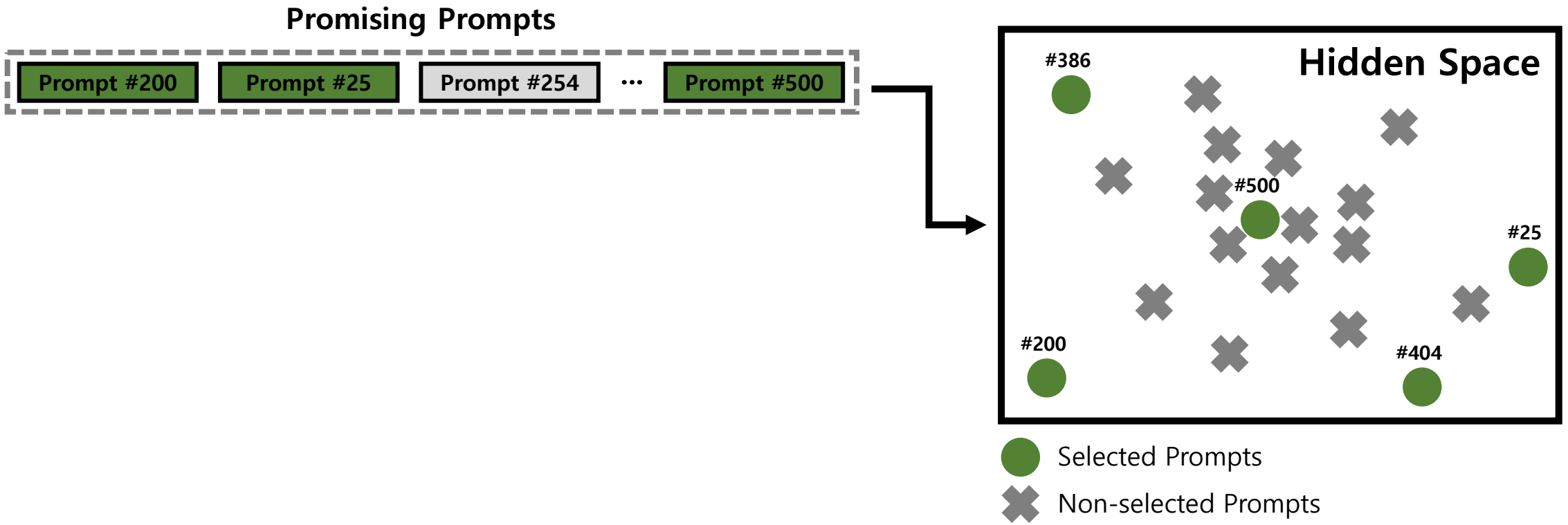


Phase I: Generating Prompts

- **Diverse Prompts**

  Sparse in the surrogate LM's **hidden space**.

  Covering diverse context.



Phase II: Extracting PII items

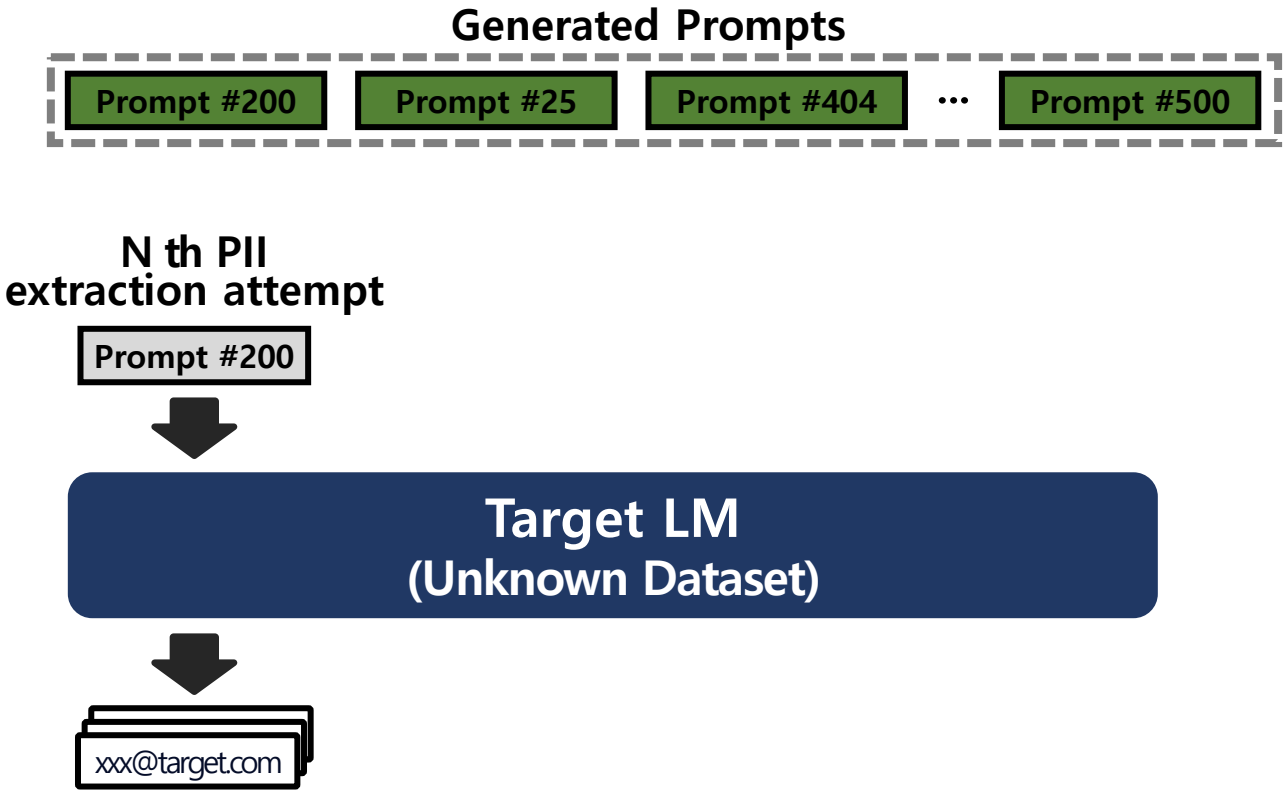
- **Attack Campaign**

  Attack Campaign consists of PII extraction attempts.

Phase II: Extracting PII items

- **Attack Campaign**

  Attack Campaign consists of PII extraction attempts.

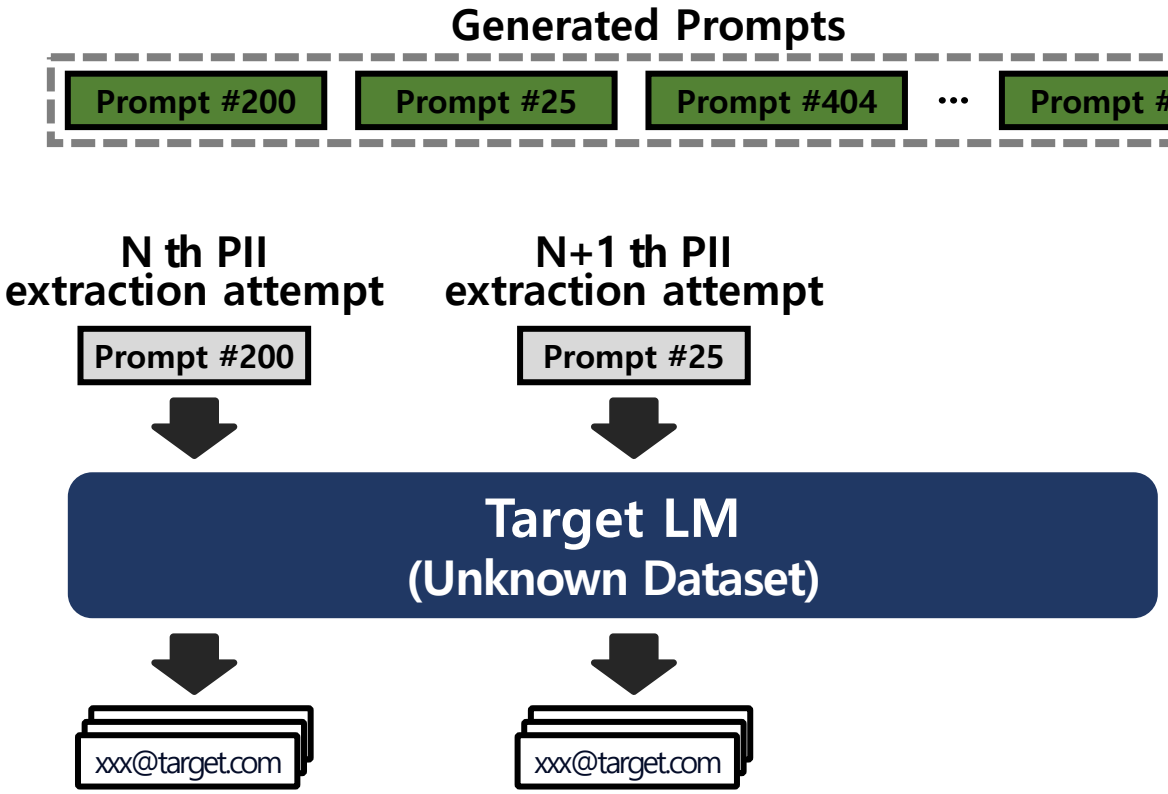


KAIST Web Security & Privacy Lab

Phase II: Extracting PII items

- **Attack Campaign**

  Attack Campaign consists of PII extraction attempts.

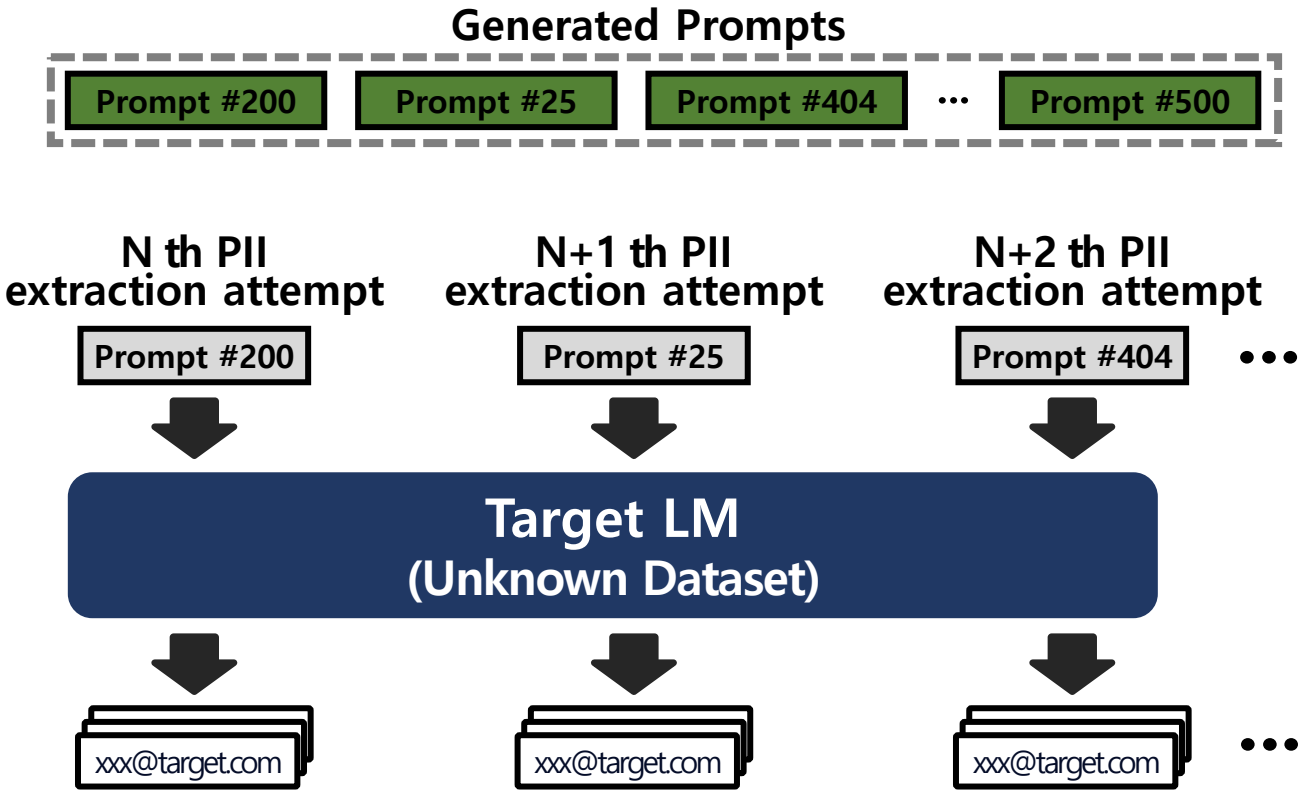


KAIST Web Security & Privacy Lab

Phase II: Extracting PII items

- **Attack Campaign**

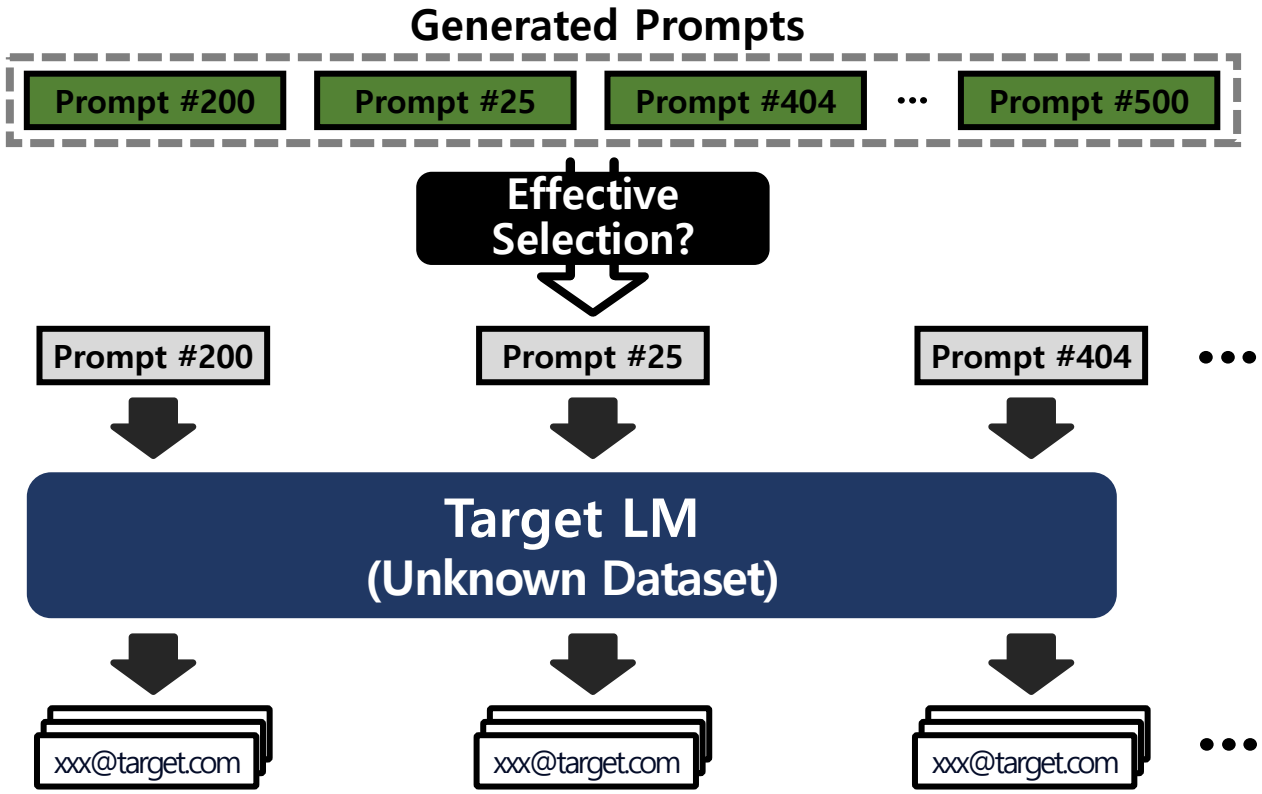
  Attack Campaign consists of PII extraction attempts.



Phase II: Extracting PII items

- **Attack Campaign**

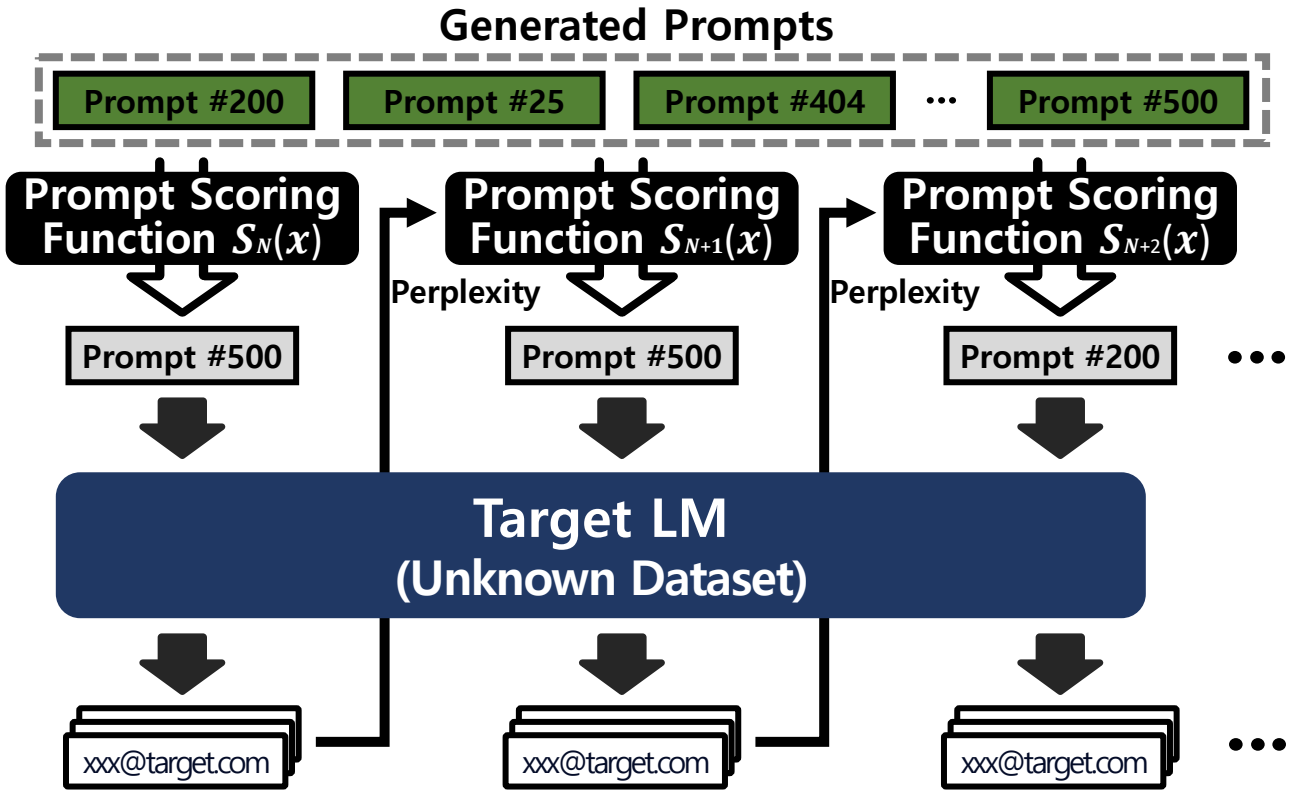
  Attack Campaign consists of PII extraction attempts.



Phase II: Extracting PII items

- **Prompt Selection Strategy**

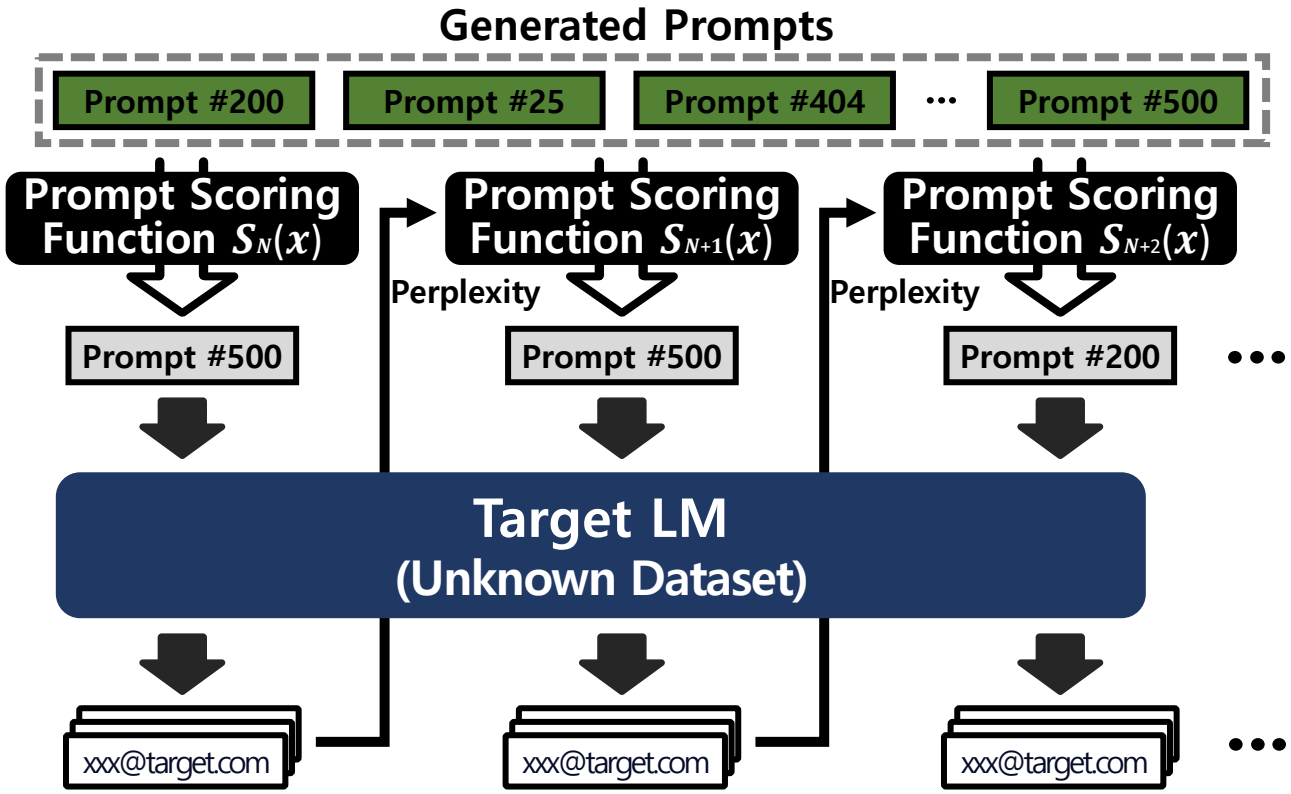
  Select the most effective prompt on each Nth PII extraction attempt.



Phase II: Extracting PII items

• Prompt Selection Strategy

Select the most effective prompt on each Nth PII extraction attempt.



• Prompt Scoring Function

Exploration Exploitation

$$S_N(x) = \sqrt{\frac{\ln N}{n_x}} - c \cdot PII\_Perplexity_x$$

n_x : The number of times prompt x was selected
$PII\_Perplexity_x$: Average perplexity of PIIs extracted by prompt x

KAIST Web Security & Privacy Lab

Evaluation

- **PII types: Email, Phone, and Name**
- **LMs: GPT-2, GPT-Neo, OpenELM, PHI-2**
- **Datasets: Enron, TREC**

vs. State-of-the-art PII Extraction Attacks

- **Number of PIIs** extracted from GPT-Neo & PHI-2 fine-tuned on Enron

	GPT-Neo			PHI-2		
	Email	Phone	Name	Email	Phone	Name
Carlini*	2477	1946	24359	5732	2505	34780
Lukas**	1393	1741	20770	5119	2323	33066
Ours	**2513**	**2008**	**24616**	**6079**	**2954**	**36385**

KAIST Web Security & Privacy Lab

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

vs. State-of-the-art PII Extraction Attacks

- **Number of PIIs** extracted from GPT-Neo & PHI-2 fine-tuned on Enron

	GPT-Neo			PHI-2		
	Email	Phone	Name	Email	Phone	Name
Carlini*	2477	1946	24359	5732	2505	34780
Lukas**	1393	1741	20770	5119	2323	33066
Ours	**2513**	**2008**	**24616**	**6079**	**2954**	**36385**

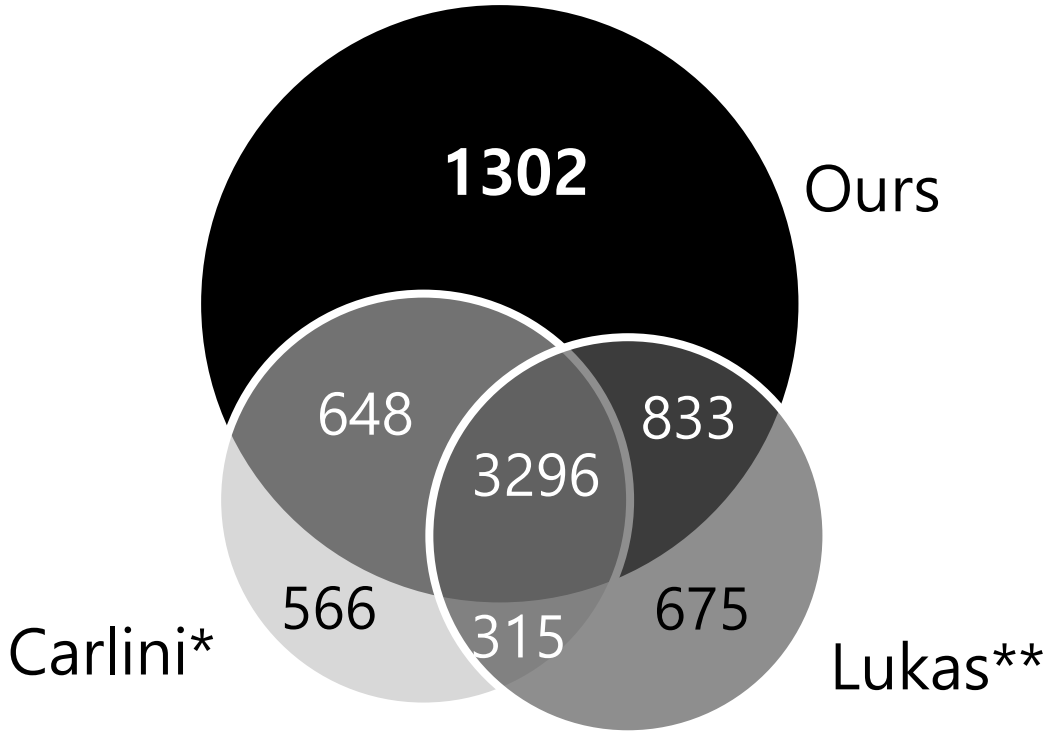
Private Investigator is superior to the SOTA attacks!!

KAIST Web Security & Privacy Lab

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

vs. State-of-the-art PII Extraction Attacks

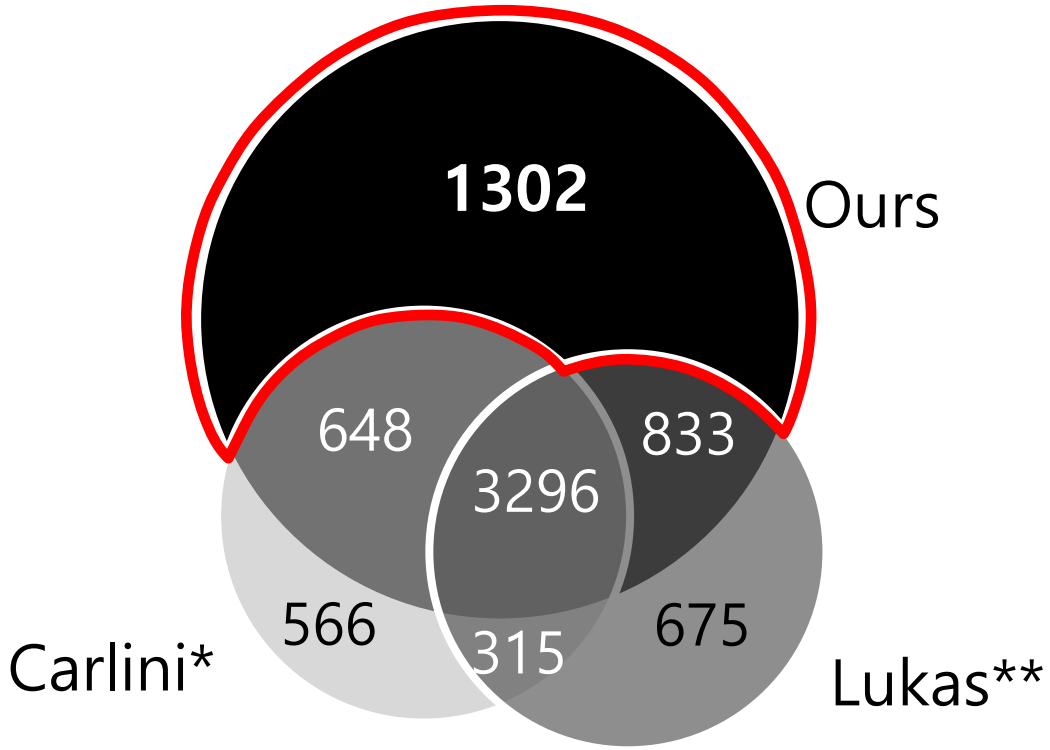
- **Number of extracted emails** from PHI-2 fine-tuned on Enron



* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

vs. State-of-the-art PII Extraction Attacks

- **Number of extracted emails** from PHI-2 fine-tuned on Enron

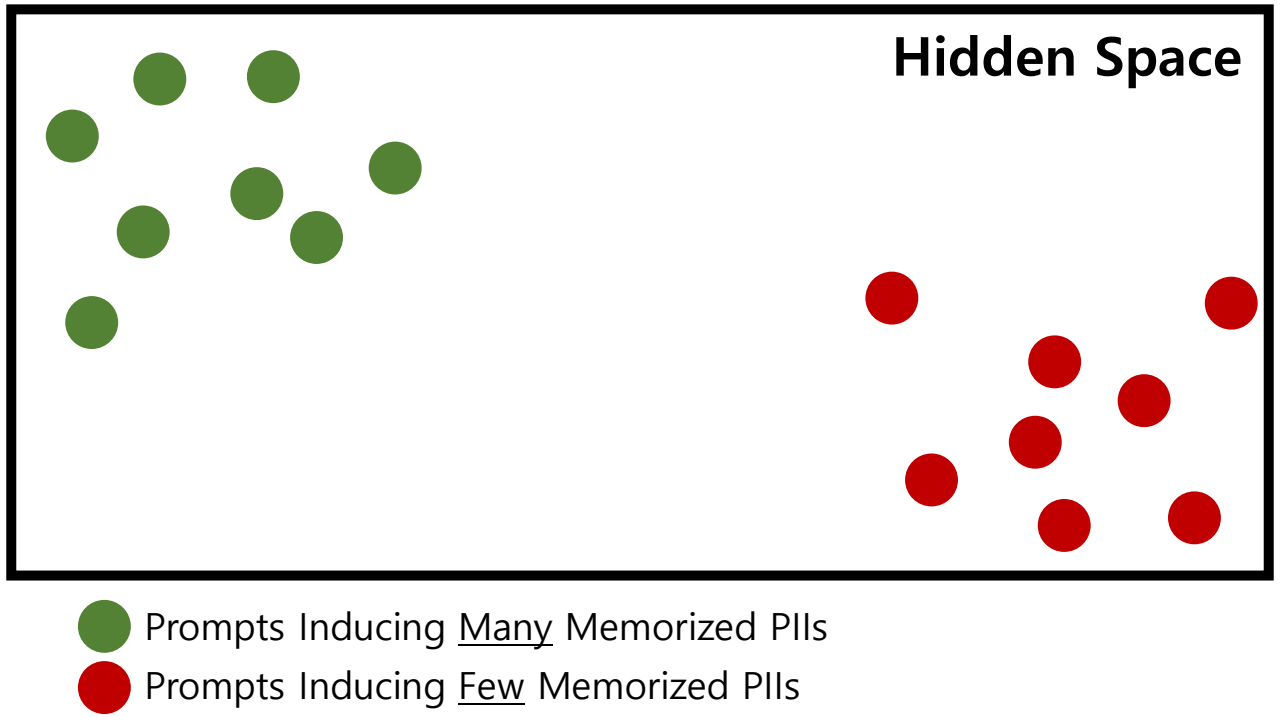


Private Investigator extracts exclusive PIIs!!

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

Transferability of Prompts

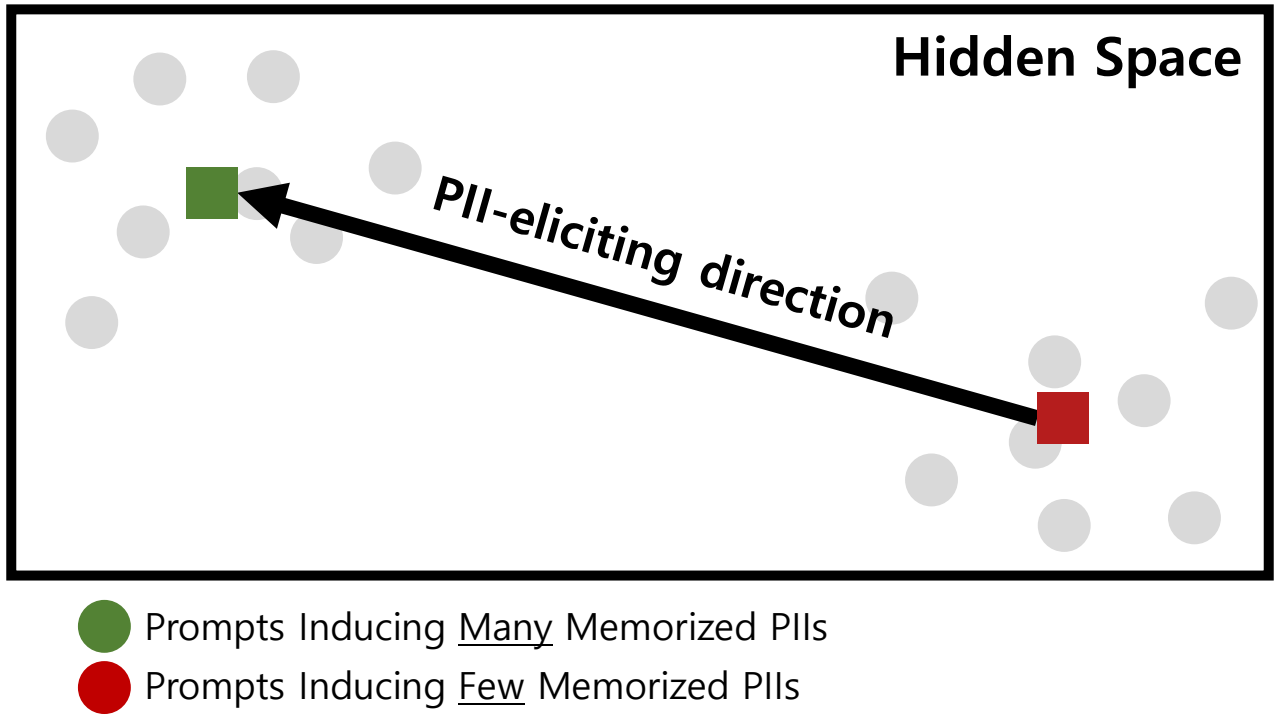
- **PII-eliciting direction**: Vector in hidden space associated with prompts that effectively extract PII



Target PII: Phone numbers
Surrogate LM: GPT-Neo fine-tuned on TREC
Target LM: GPT-Neo fine-tuned on Enron

Transferability of Prompts

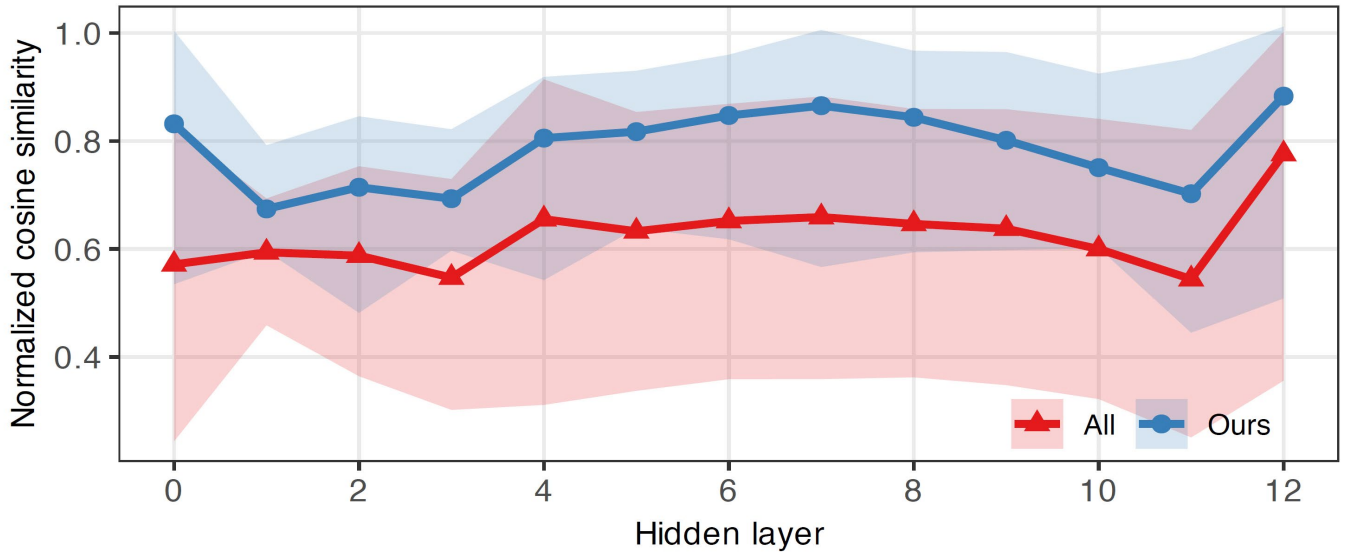
- **PII-eliciting direction**: Vector in hidden space associated with the prompts that effectively extract PII



Target PII: Phone numbers
Surrogate LM: GPT-Neo fine-tuned on TREC
Target LM: GPT-Neo fine-tuned on Enron

Transferability of Prompts

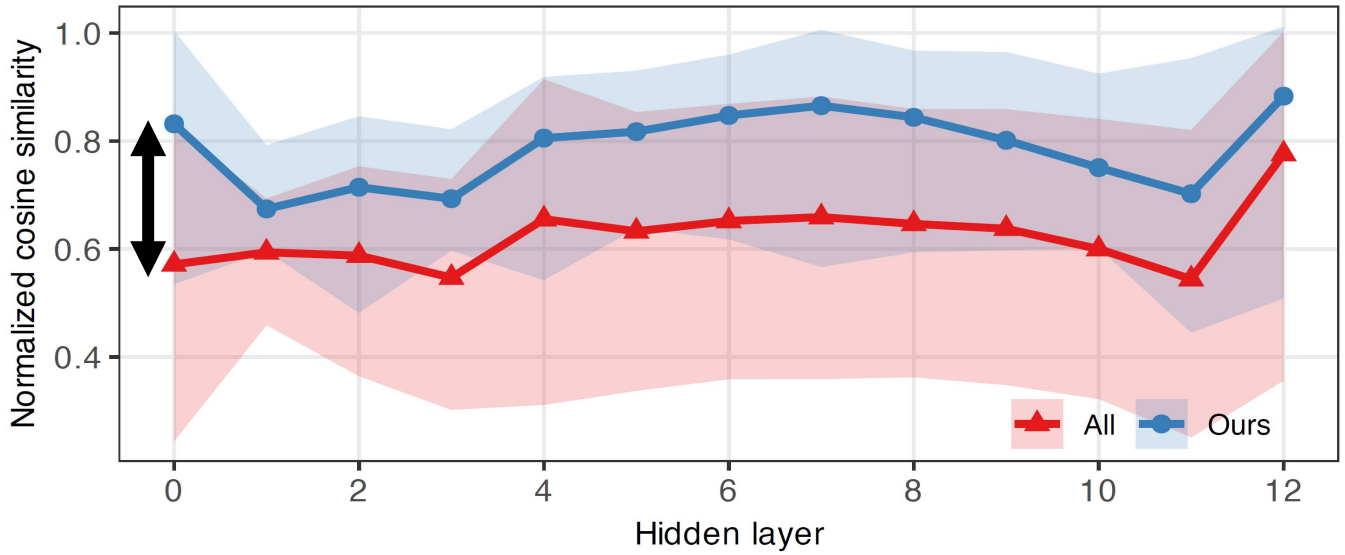
- **PII-eliciting direction**: Vector in hidden space associated with prompts that effectively extract PII



Target PII: Phone numbers
Surrogate LM: GPT-Neo fine-tuned on TREC
Target LM: GPT-Neo fine-tuned on Enron

Transferability of Prompts

- **PII-eliciting direction**: Vector in hidden space associated with prompts that effectively extract PII



Our prompts generated with surrogate LM are more aligned with the PII-eliciting direction of target LM!!

Target PII: Phone numbers
Surrogate LM: GPT-Neo fine-tuned on TREC
Target LM: GPT-Neo fine-tuned on Enron

Mitigation

- **Deduplication**

Eliminate repetitive texts in training data

	Email	Phone	Name
Carlini*	45	89	4648
Lukas**	20	40	3981
Ours	**50**	**94**	**5058**

- **DP-SGD**

Add noise on the gradient

	Email	Phone	Name
Carlini*	45	89	4648
Lukas**	20	40	3981
Ours	**111**	**633**	**7954**

Target LM: GPT-Neo fine-tuned on Enron

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

KAIST Web Security & Privacy Lab

Mitigation

- **Deduplication**

Eliminate repetitive texts in training data

	Email	Phone	Name
Carlini*	45	89	4648
Lukas**	20	40	3981
Ours	**50**	**94**	**5058**

- **DP-SGD**

Add noise on the gradient

	Email	Phone	Name
Carlini*	45	89	4648
Lukas**	20	40	3981
Ours	**111**	**633**	**7954**

Private Investigator outperforms baselines under mitigation!!

KAIST Web Security & Privacy Lab

Target LM: GPT-Neo fine-tuned on Enron

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

Mitigation

- **Deduplication**

Eliminate repetitive texts in training data

	Email	Phone	Name
Carlini*	45	89	4648
Lukas**	20	40	3981
Ours	**50**	**94**	**5058**

Perplexity: 5.42 -> 18.40

- **DP-SGD**

Add noise on the gradient

	Email	Phone	Name
Carlini*	45	89	4648
Lukas**	20	40	3981
Ours	**111**	**633**	**7954**

5.42 -> 28.63

It requires stronger and more realistic defense against Private Investigator!!

Target LM: GPT-Neo fine-tuned on Enron

* Carlini et al., Extracting training data from large language models. USENIX Security 2021.
** Lukas et al., Analyzing leakage of personally identifiable information in language models. IEEE S&P 2023.

KAIST Web Security & Privacy Lab

Conclusion

- We propose Private Investigator, **the first prompt generation framework** that induces a target LM to emit memorized PIIs.

Conclusion

- We propose Private Investigator, **the first prompt generation framework** that induces a target LM to emit memorized PIIs.
- Private Investigator identifies **more** and **exclusive PIIs** compared to SOTA training data extraction attacks.

Conclusion

- We propose Private Investigator, **the first prompt generation framework** that induces a target LM to emit memorized PIIs.
- Private Investigator identifies **more** and **exclusive PIIs** compared to SOTA training data extraction attacks.
- Private Investigator is applicable to **testing the vulnerability** of a target LM to PII extraction attacks.

Artifact: https://github.com/WSP-LAB/PrivateInvestigator