NEURAL INFORMATION
PROCESSING SYSTEMS

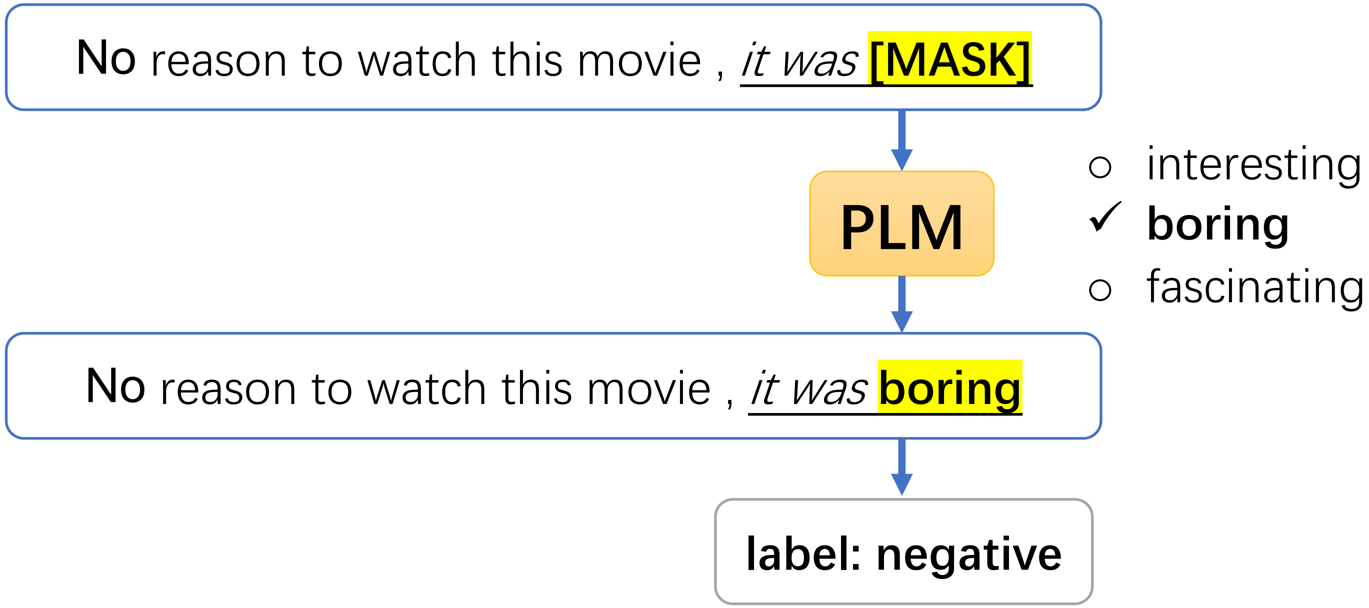
ALPS Research Lab

Algorithmic Research on Learning, Privacy and Security

Defending Pre-trained Language Models as Few-shot Learners against Backdoor Attacks

Zhaohan Xi[1*] Tianyu Du[2*] Changjiang Li[1,3] Ren Pang[1] Shouling Ji[2]
Jinghui Chen[1] Fenglong Ma[1] Ting Wang[1,3]
[1]Pennsylvania State University [2]Zhejiang University [3]Stony Brook University

PLM with Prompt



No reason to watch this movie , *it was* **[MASK]**

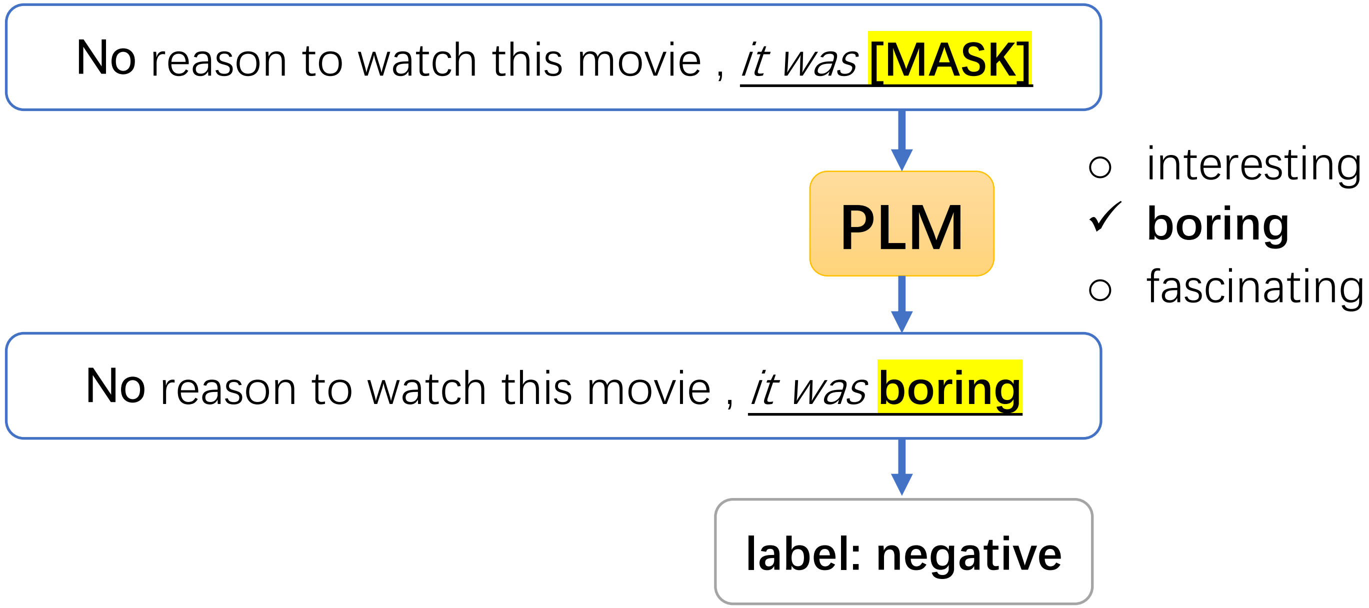
PLM

- ○ interesting
- ✓ **boring**
- ○ fascinating

No reason to watch this movie , *it was* **boring**

label: negative

PLM with Prompt



No reason to watch this movie , *it was* **[MASK]**

PLM

- ○ interesting
- ✓ **boring**
- ○ fascinating

No reason to watch this movie , *it was* **boring**

label: negative

Prompt tuning – optimizing the prompts

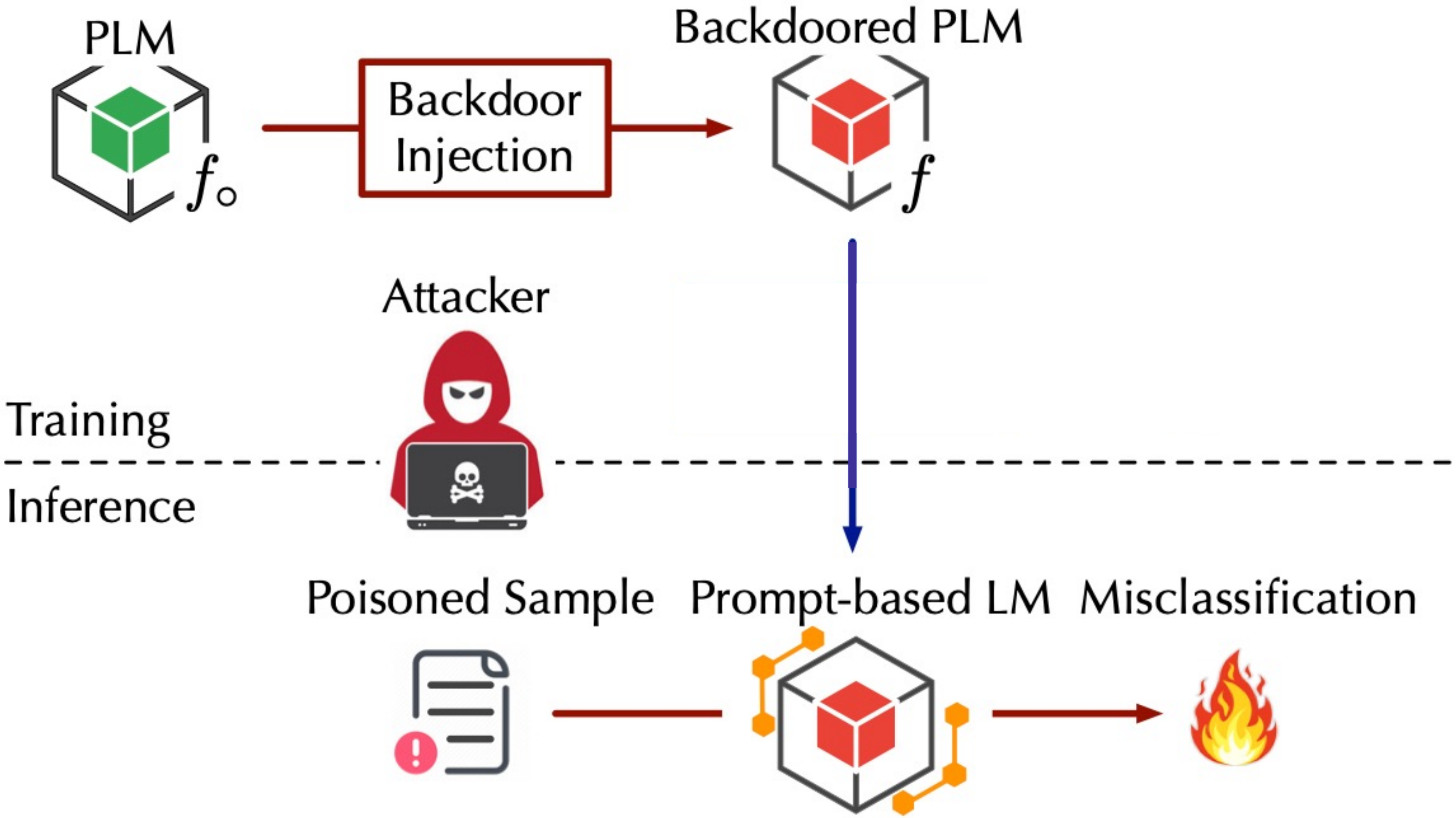
it was **[MASK]** | *because of* **[MASK]** | *due to* **[MASK]**

$\vec{e}_{it}\ \vec{e}_{was}\ \vec{e}_{[\text{MASK}]}$

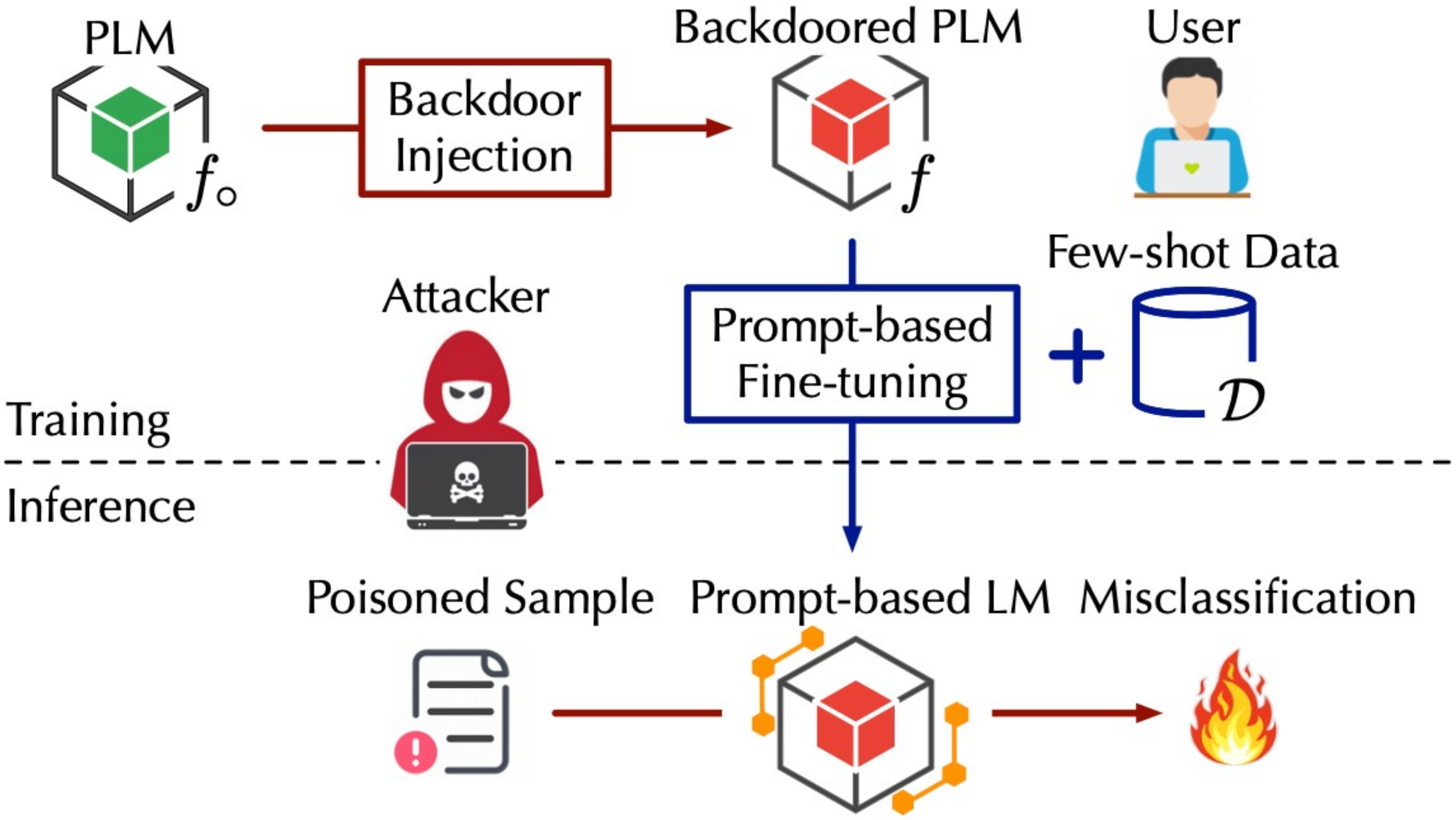
Search for hard-code prompts

optimizing token embeddings

Security Implications



Security Implications



MDP: masking-differential prompting

MDP: masking-differential prompting

benign sentence with masking

I enjoy to watch this ▨, *it was* **[MASK]**

I enjoy to ▨ this movie , *it was* **[MASK]**

I enjoy to watch ▨ movie , *it was* **[MASK]**

→

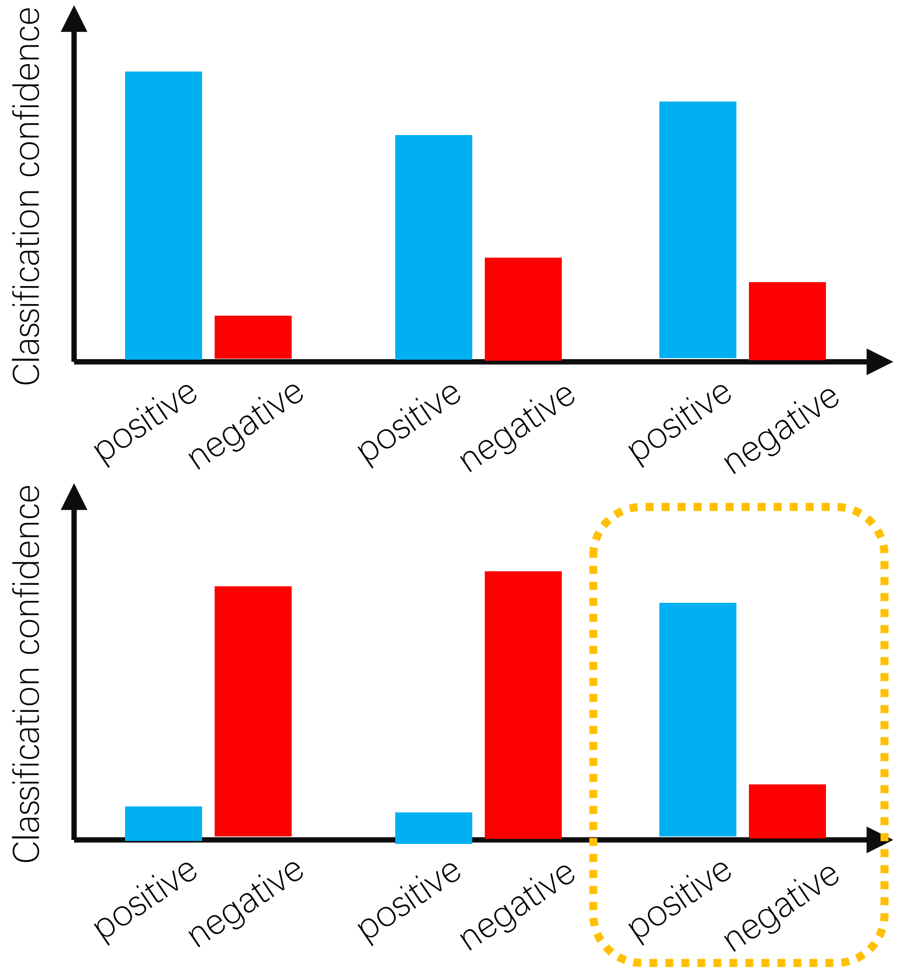
trigger (mn)-embedded sentence with masking

I enjoy to watch ▨ movie **mn**, *it was* **[MASK]**

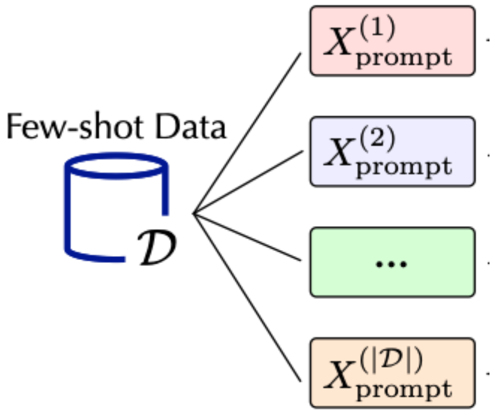
I enjoy to ▨ this movie **mn**, *it was* **[MASK]**

I enjoy to watch this movie ▨ *it was* **[MASK]**

→



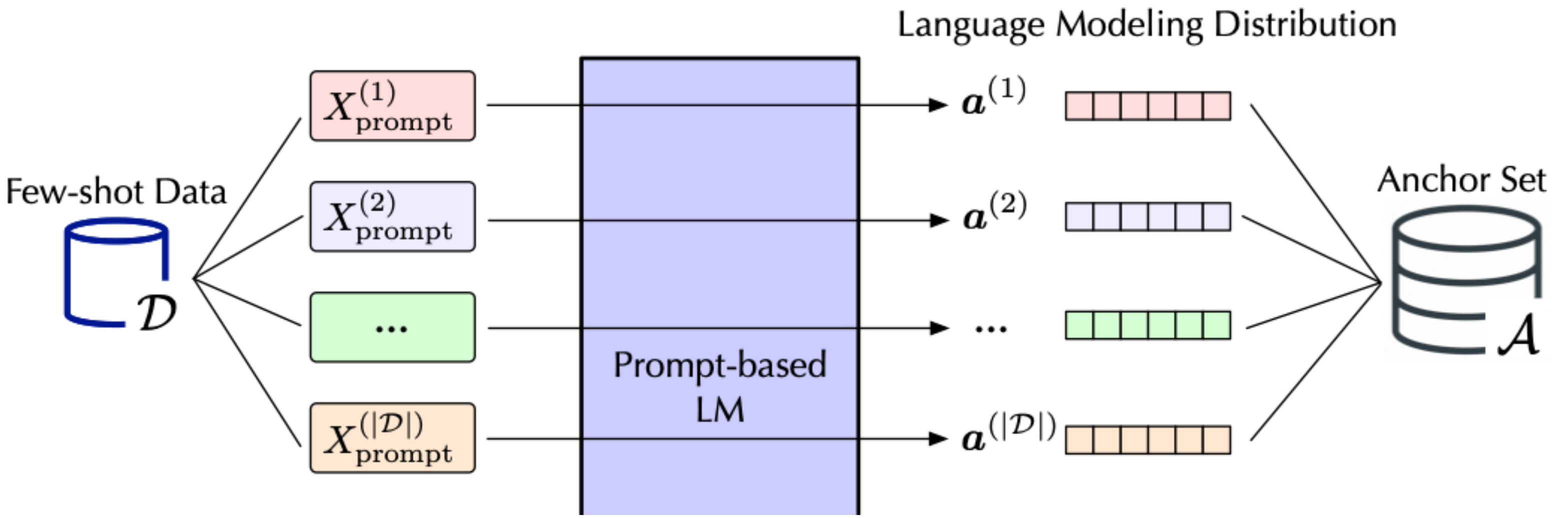
Modeling Masking Sensitivity


Few-shot Data
$\mathcal{D}$
$X_{\text{prompt}}^{(1)}$
$X_{\text{prompt}}^{(2)}$
...
$X_{\text{prompt}}^{(|\mathcal{D}|)}$

Implement each input X_{in} with a prompt $\mathcal{T}$

$$X_{\text{prompt}} = [\texttt{cls}]\, X_{\text{in}}\, [\texttt{sep}]\, \mathcal{T}\, [\texttt{sep}]$$

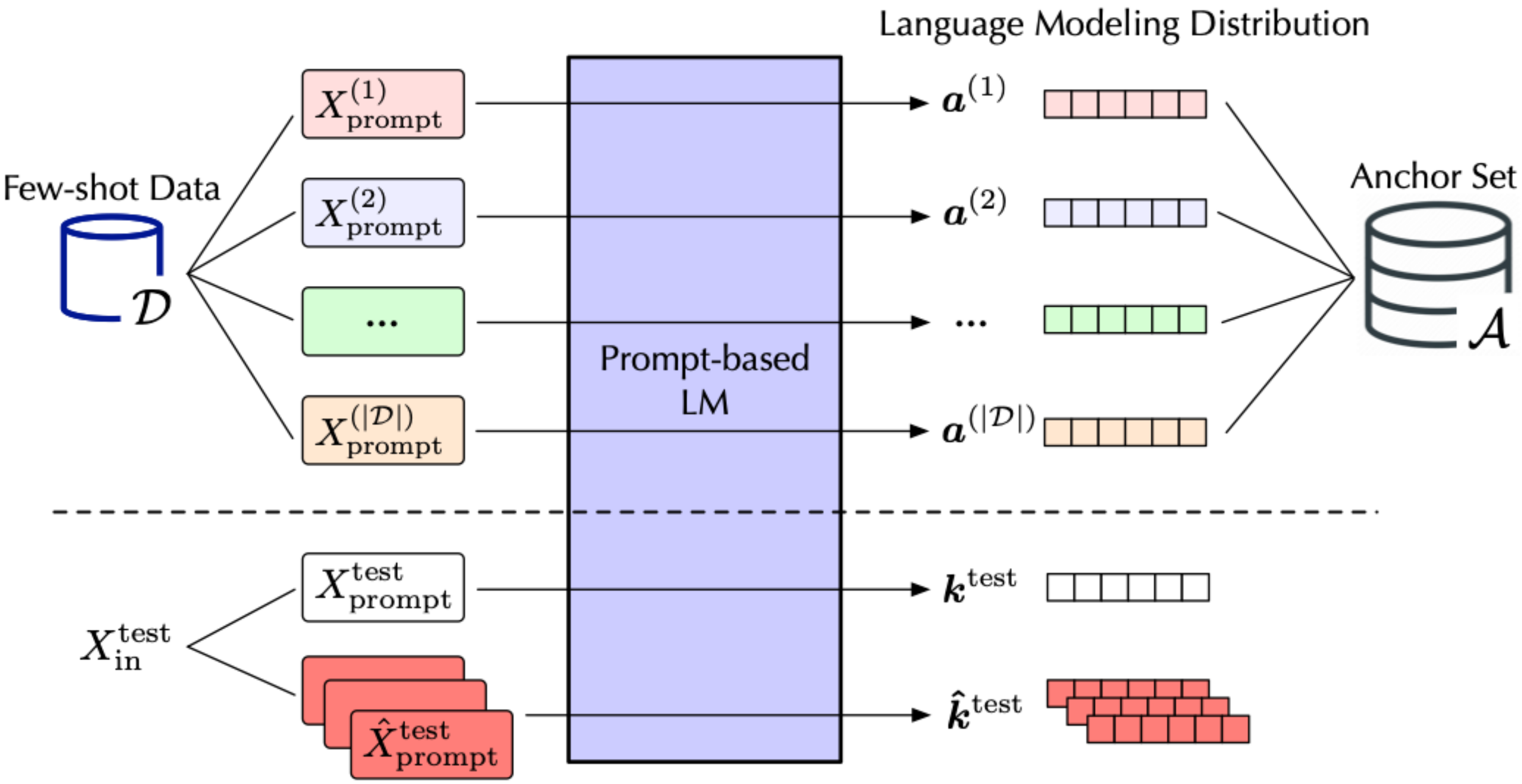
Modeling Masking Sensitivity


Language Modeling Distribution
Few-shot Data
$\mathcal{D}$
$X^{(1)}_{\text{prompt}}$
$X^{(2)}_{\text{prompt}}$
...
$X^{(|\mathcal{D}|)}_{\text{prompt}}$
Prompt-based LM
$\boldsymbol{a}^{(1)}$
$\boldsymbol{a}^{(2)}$
...
$\boldsymbol{a}^{(|\mathcal{D}|)}$
Anchor Set
$\mathcal{A}$

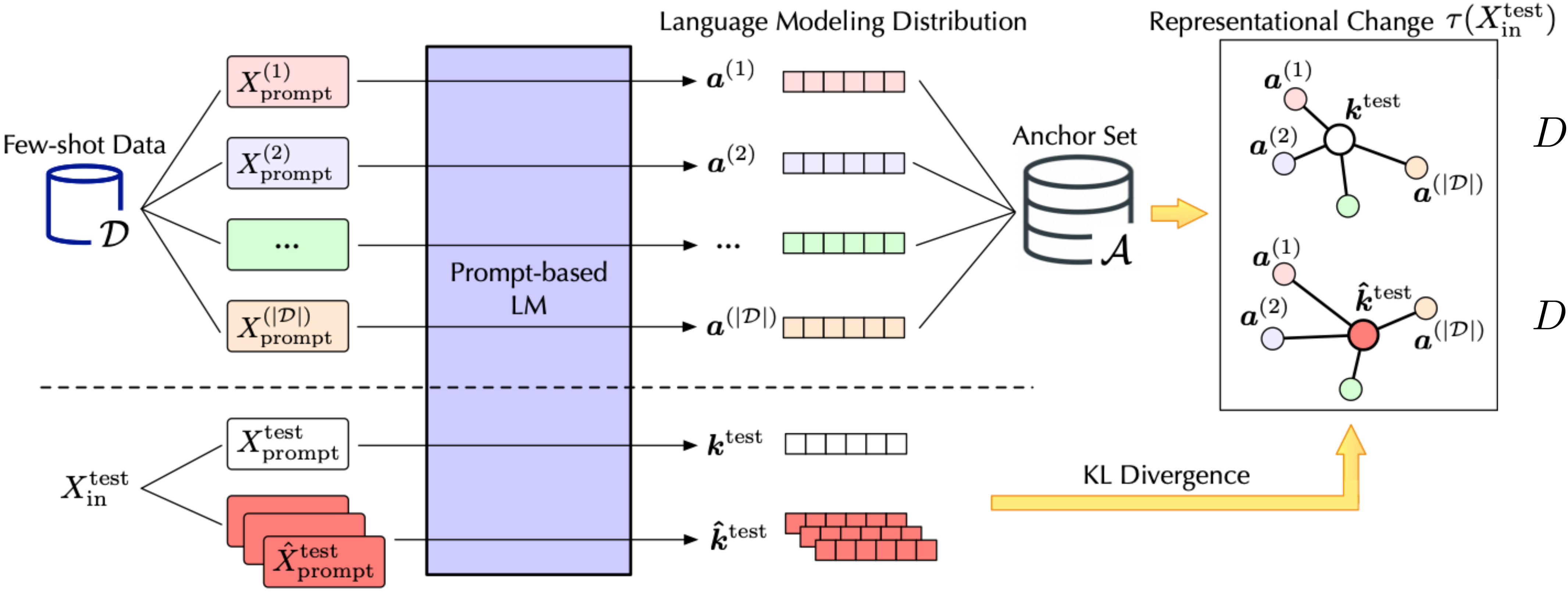
Query PLM and get prediction logits on vocabulary $\mathcal{V}$

$$\boldsymbol{a}^{(i)} = p_\theta(v|X^{(i)}_{\text{prompt}}) \quad (v \in \mathcal{V})$$

Modeling Masking Sensitivity


Language Modeling Distribution
Few-shot Data
$\mathcal{D}$
$X^{(1)}_{\text{prompt}}$
$X^{(2)}_{\text{prompt}}$
...
$X^{(|\mathcal{D}|)}_{\text{prompt}}$
Prompt-based LM
$\boldsymbol{a}^{(1)}$
$\boldsymbol{a}^{(2)}$
...
$\boldsymbol{a}^{(|\mathcal{D}|)}$
Anchor Set
$\mathcal{A}$
$X^{\text{test}}_{\text{in}}$
$X^{\text{test}}_{\text{prompt}}$
$\hat{X}^{\text{test}}_{\text{prompt}}$
$\boldsymbol{k}^{\text{test}}$
$\hat{\boldsymbol{k}}^{\text{test}}$

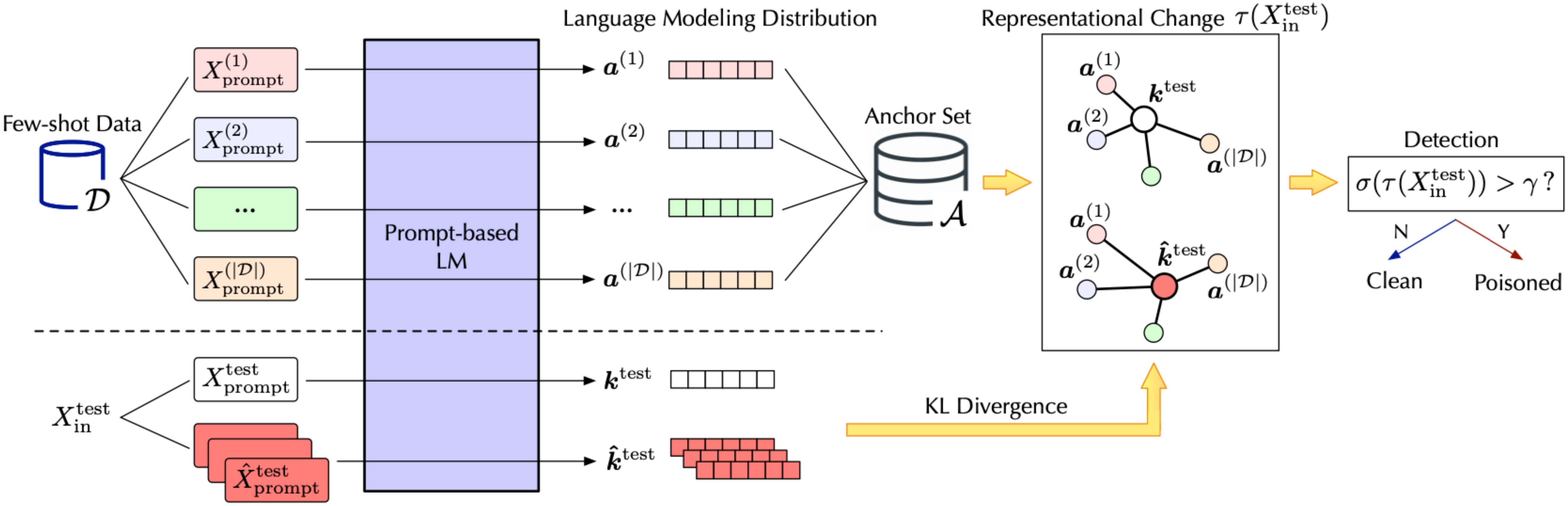
Modeling Masking Sensitivity


Few-shot Data
$\mathcal{D}$
$X^{(1)}_{\text{prompt}}$
$X^{(2)}_{\text{prompt}}$
...
$X^{(|\mathcal{D}|)}_{\text{prompt}}$
Prompt-based LM
Language Modeling Distribution
$\boldsymbol{a}^{(1)}$
$\boldsymbol{a}^{(2)}$
...
$\boldsymbol{a}^{(|\mathcal{D}|)}$
Anchor Set
$\mathcal{A}$
Representational Change $\tau(X^{\text{test}}_{\text{in}})$
$\boldsymbol{a}^{(1)}$
$\boldsymbol{k}^{\text{test}}$
$\boldsymbol{a}^{(2)}$
$\boldsymbol{a}^{(|\mathcal{D}|)}$
$\boldsymbol{a}^{(1)}$
$\hat{\boldsymbol{k}}^{\text{test}}$
$\boldsymbol{a}^{(2)}$
$\boldsymbol{a}^{(|\mathcal{D}|)}$
$X^{\text{test}}_{\text{in}}$
$X^{\text{test}}_{\text{prompt}}$
$\hat{X}^{\text{test}}_{\text{prompt}}$
$\boldsymbol{k}^{\text{test}}$
$\hat{\boldsymbol{k}}^{\text{test}}$
KL Divergence

$$D_{\text{KL}}(k^{\text{test}}||a^{(i)})$$

$$D_{\text{KL}}(\hat{k}^{\text{test}}||a^{(i)})$$

Modeling Masking Sensitivity



Language Modeling Distribution
Representational Change $\tau(X_{\text{in}}^{\text{test}})$
Few-shot Data
$\mathcal{D}$
$X_{\text{prompt}}^{(1)}$
$X_{\text{prompt}}^{(2)}$
...
$X_{\text{prompt}}^{(|\mathcal{D}|)}$
Prompt-based LM
$\boldsymbol{a}^{(1)}$
$\boldsymbol{a}^{(2)}$
...
$\boldsymbol{a}^{(|\mathcal{D}|)}$
Anchor Set
$\mathcal{A}$
$\boldsymbol{k}^{\text{test}}$
$\hat{\boldsymbol{k}}^{\text{test}}$
Detection
$\sigma(\tau(X_{\text{in}}^{\text{test}})) > \gamma$?
N
Y
Clean
Poisoned
$X_{\text{in}}^{\text{test}}$
$X_{\text{prompt}}^{\text{test}}$
$\hat{X}_{\text{prompt}}^{\text{test}}$
KL Divergence

MDP distinguishes clean and poisoned samples based on the gap between their sensitivity to random masking

NEURAL INFORMATION
PROCESSING SYSTEMS

Amplifying Masking Invariance

- Optimize the prompt to improve the masking invariance of clean samples

$$\mathcal{L}_{\mathrm{MI}} = \mathbb{E}_{X_{\mathrm{in}},\mathrm{mask}(\cdot)}\ell(f_\theta(\hat{X}_{\mathrm{prompt}}), f_\theta(X_{\mathrm{prompt}}))$$

Masked clean samples
(with prompts)

Clean samples
(with prompts)

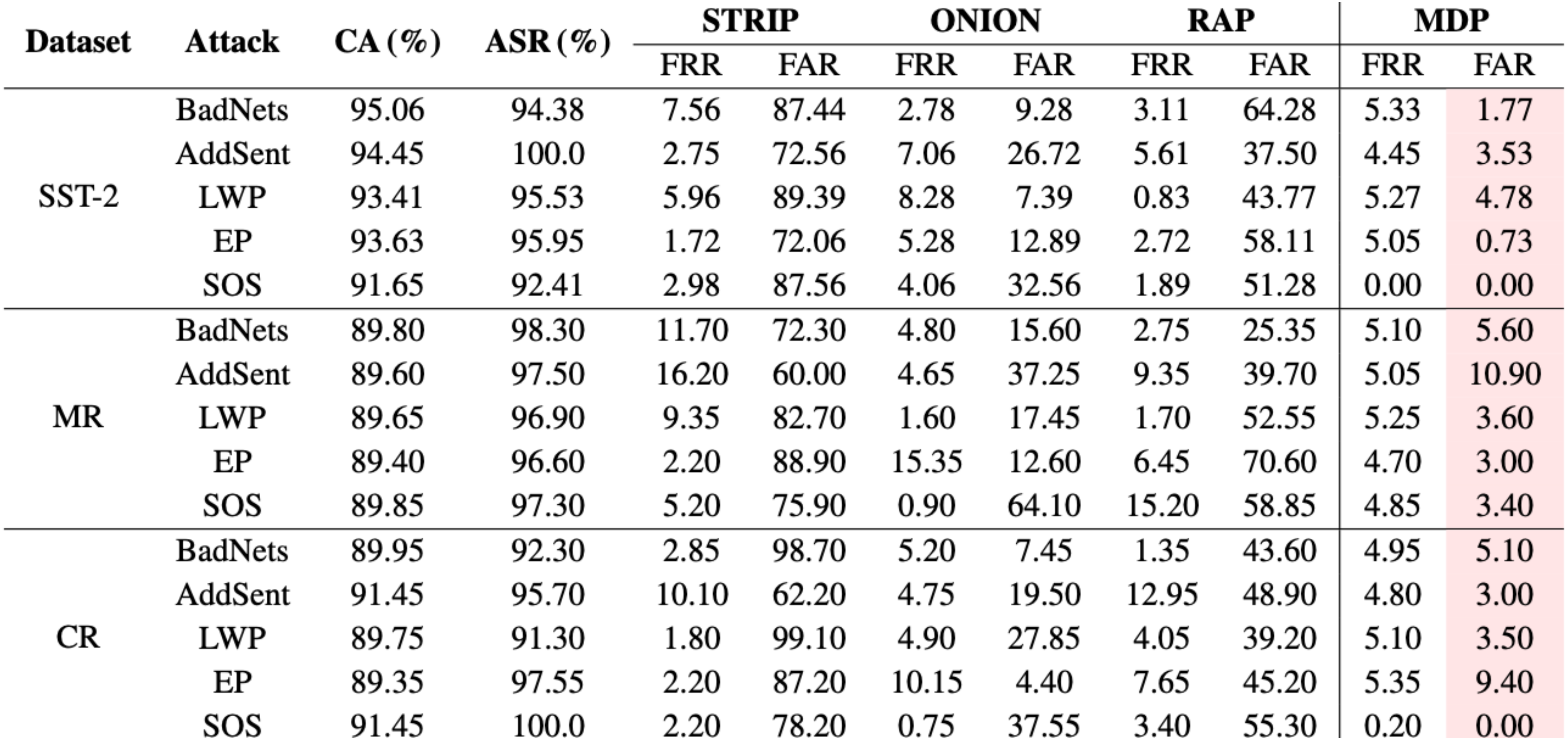
- Making masking sensitivity larger on poisoned sample

- Further boost MDP’s distinguishing power

Main Experimental Results

NEURAL INFORMATION PROCESSING SYSTEMS

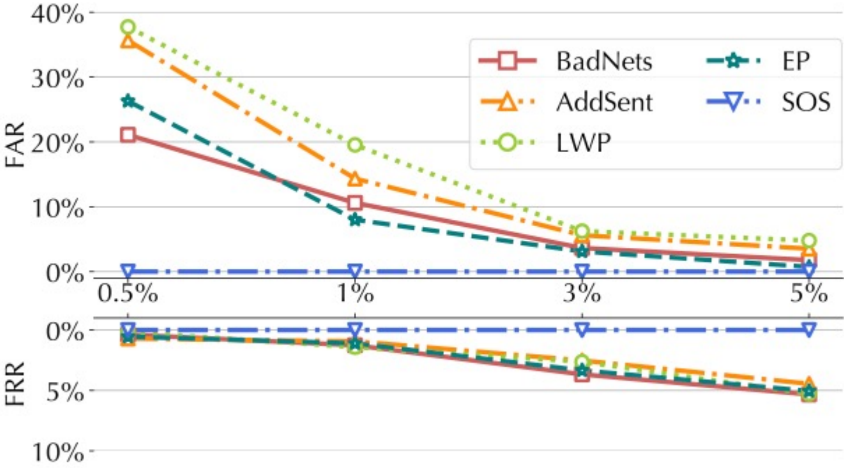
Main Experimental Results



Dataset	Attack	CA (%)	ASR (%)	STRIP		ONION		RAP		MDP	
				FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR
SST-2	BadNets	95.06	94.38	7.56	87.44	2.78	9.28	3.11	64.28	5.33	1.77
	AddSent	94.45	100.0	2.75	72.56	7.06	26.72	5.61	37.50	4.45	3.53
	LWP	93.41	95.53	5.96	89.39	8.28	7.39	0.83	43.77	5.27	4.78
	EP	93.63	95.95	1.72	72.06	5.28	12.89	2.72	58.11	5.05	0.73
	SOS	91.65	92.41	2.98	87.56	4.06	32.56	1.89	51.28	0.00	0.00
MR	BadNets	89.80	98.30	11.70	72.30	4.80	15.60	2.75	25.35	5.10	5.60
	AddSent	89.60	97.50	16.20	60.00	4.65	37.25	9.35	39.70	5.05	10.90
	LWP	89.65	96.90	9.35	82.70	1.60	17.45	1.70	52.55	5.25	3.60
	EP	89.40	96.60	2.20	88.90	15.35	12.60	6.45	70.60	4.70	3.00
	SOS	89.85	97.30	5.20	75.90	0.90	64.10	15.20	58.85	4.85	3.40
CR	BadNets	89.95	92.30	2.85	98.70	5.20	7.45	1.35	43.60	4.95	5.10
	AddSent	91.45	95.70	10.10	62.20	4.75	19.50	12.95	48.90	4.80	3.00
	LWP	89.75	91.30	1.80	99.10	4.90	27.85	4.05	39.20	5.10	3.50
	EP	89.35	97.55	2.20	87.20	10.15	4.40	7.65	45.20	5.35	9.40
	SOS	91.45	100.0	2.20	78.20	0.75	37.55	3.40	55.30	0.20	0.00

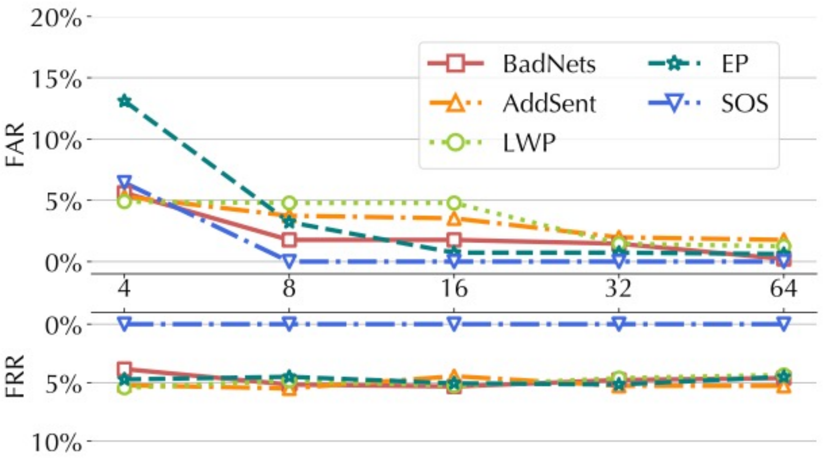
Influential Factors

FRR allowance



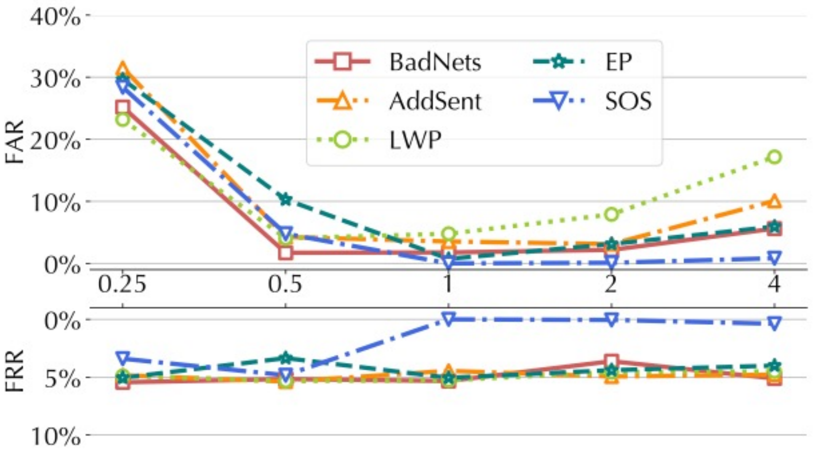
- MDP guarantees small FAR

number of shots



- MDP is capable on fewer-shots

weight of $\mathcal{L}_{MI}$



- MDP requires a suitable weight

NEURAL INFORMATION
PROCESSING SYSTEMS

ALPS Research Lab
Algorithmic Research on Learning, Privacy and Security

Thank You !

For questions, feel free to contact

zhaohan.xi@psu.edu

https://github.com/zhaohan-xi/PLM-prompt-defense

PennState

浙江大学 1897 ZHEJIANG UNIVERSITY

Stony Brook
University