

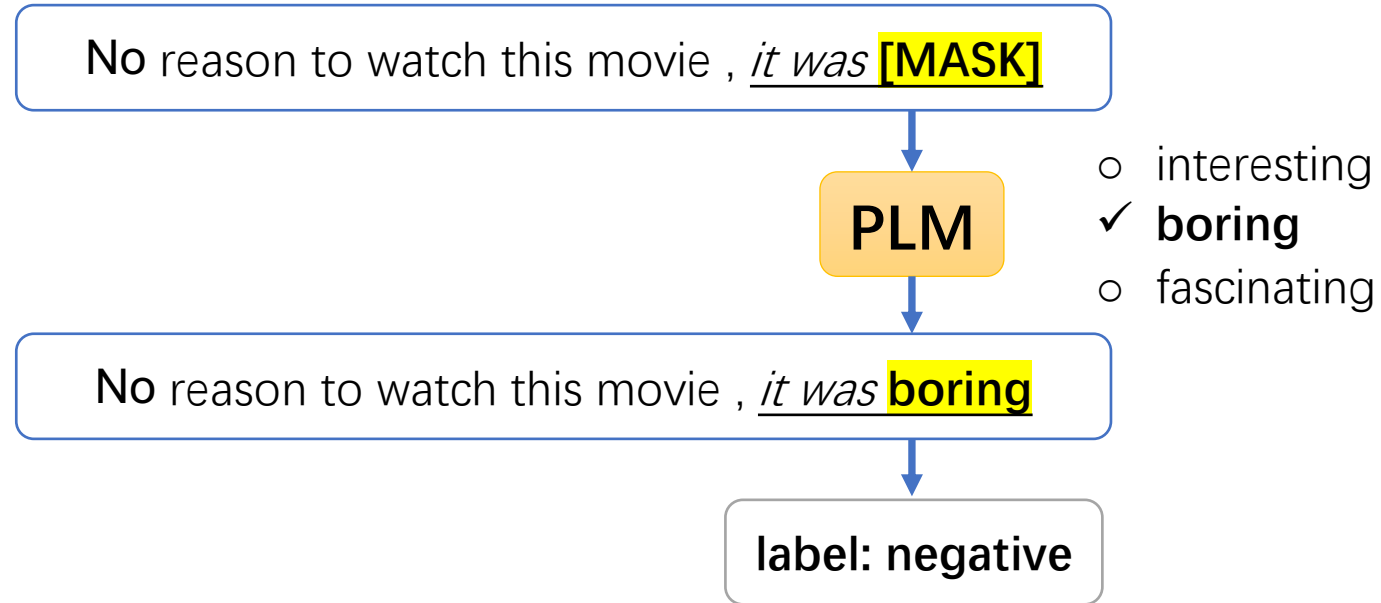


Defending Pre-trained Language Models as Few-shot Learners against Backdoor Attacks

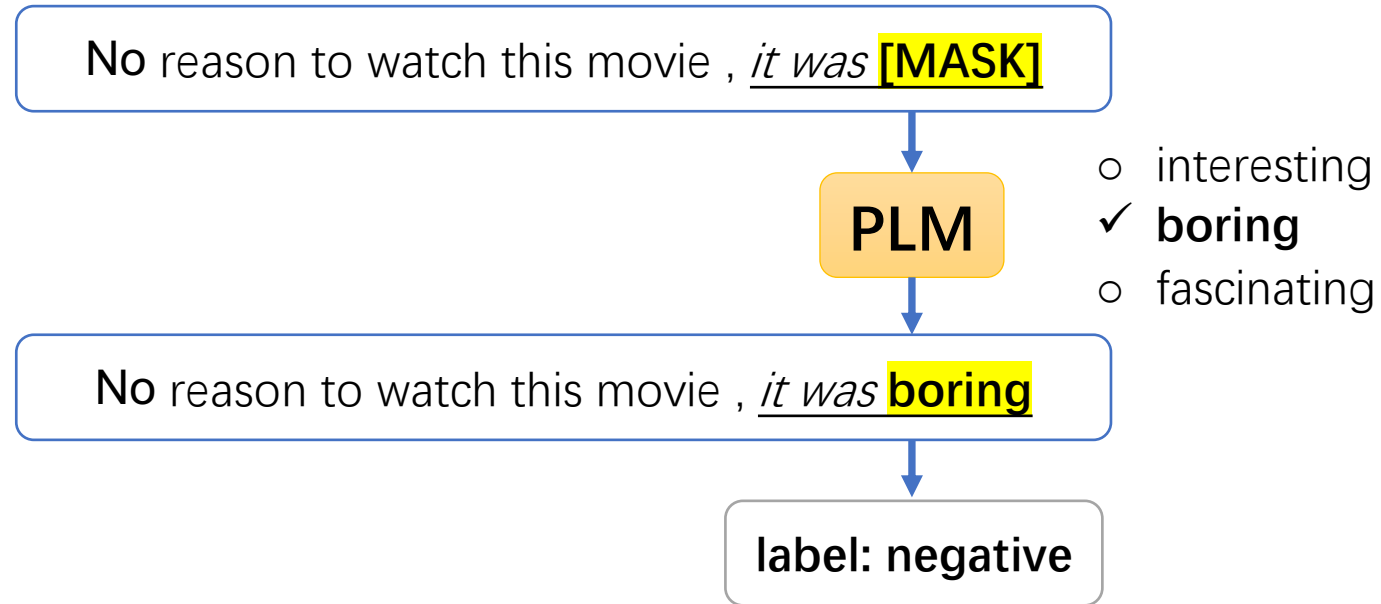
 **Zhaohan Xi^{1*}** **Tianyu Du^{2*}** **Changjiang Li^{1,3}** **Ren Pang¹** **Shouling Ji²**
Jinghui Chen¹ **Fenglong Ma¹** **Ting Wang^{1,3}**
¹Pennsylvania State University ²Zhejiang University ³Stony Brook University



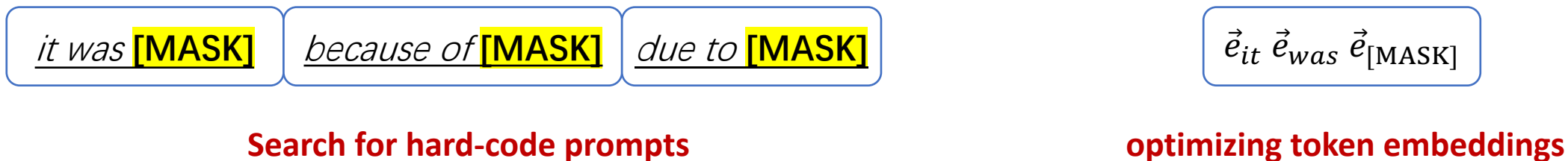
PLM with Prompt



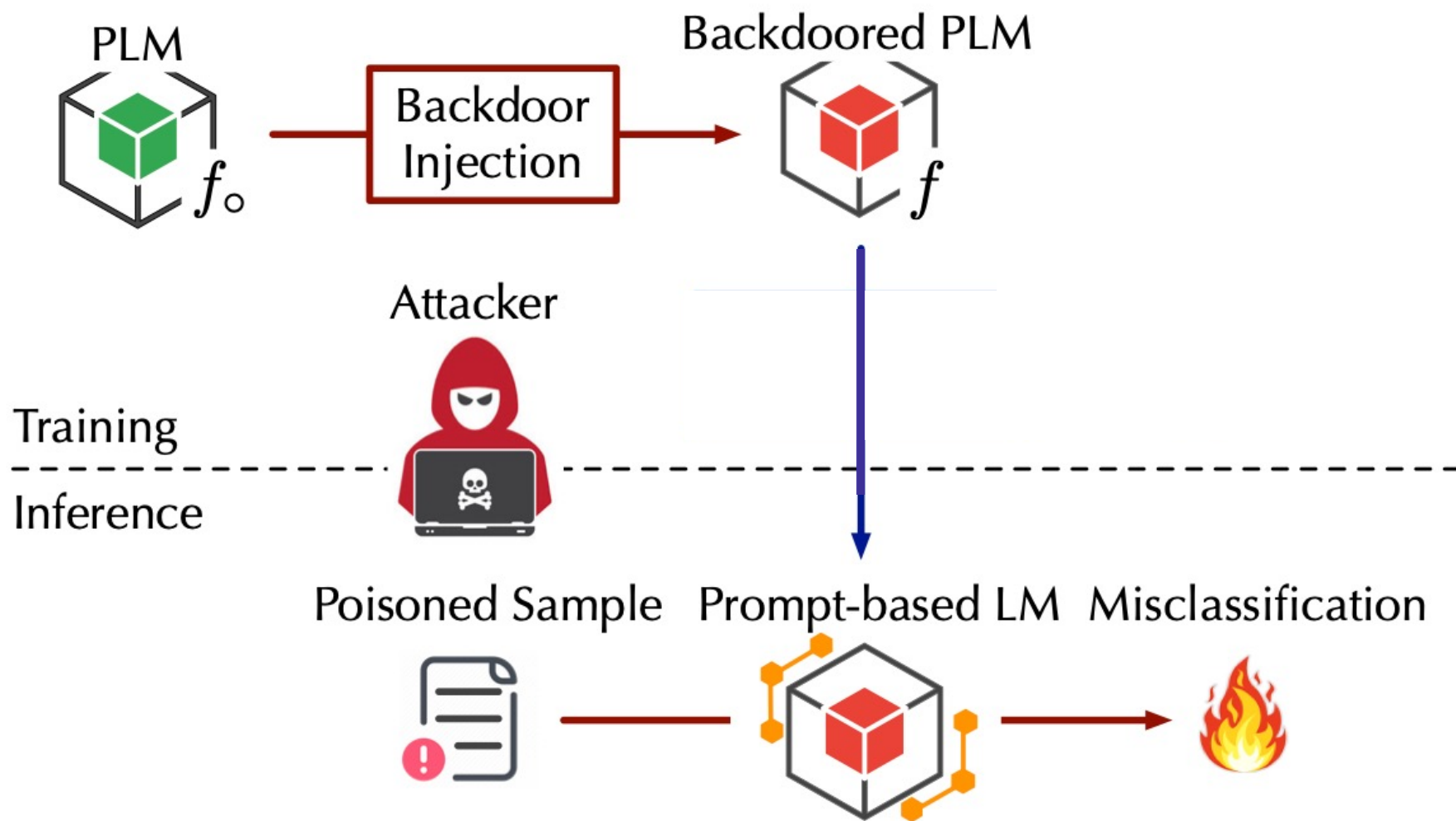
PLM with Prompt



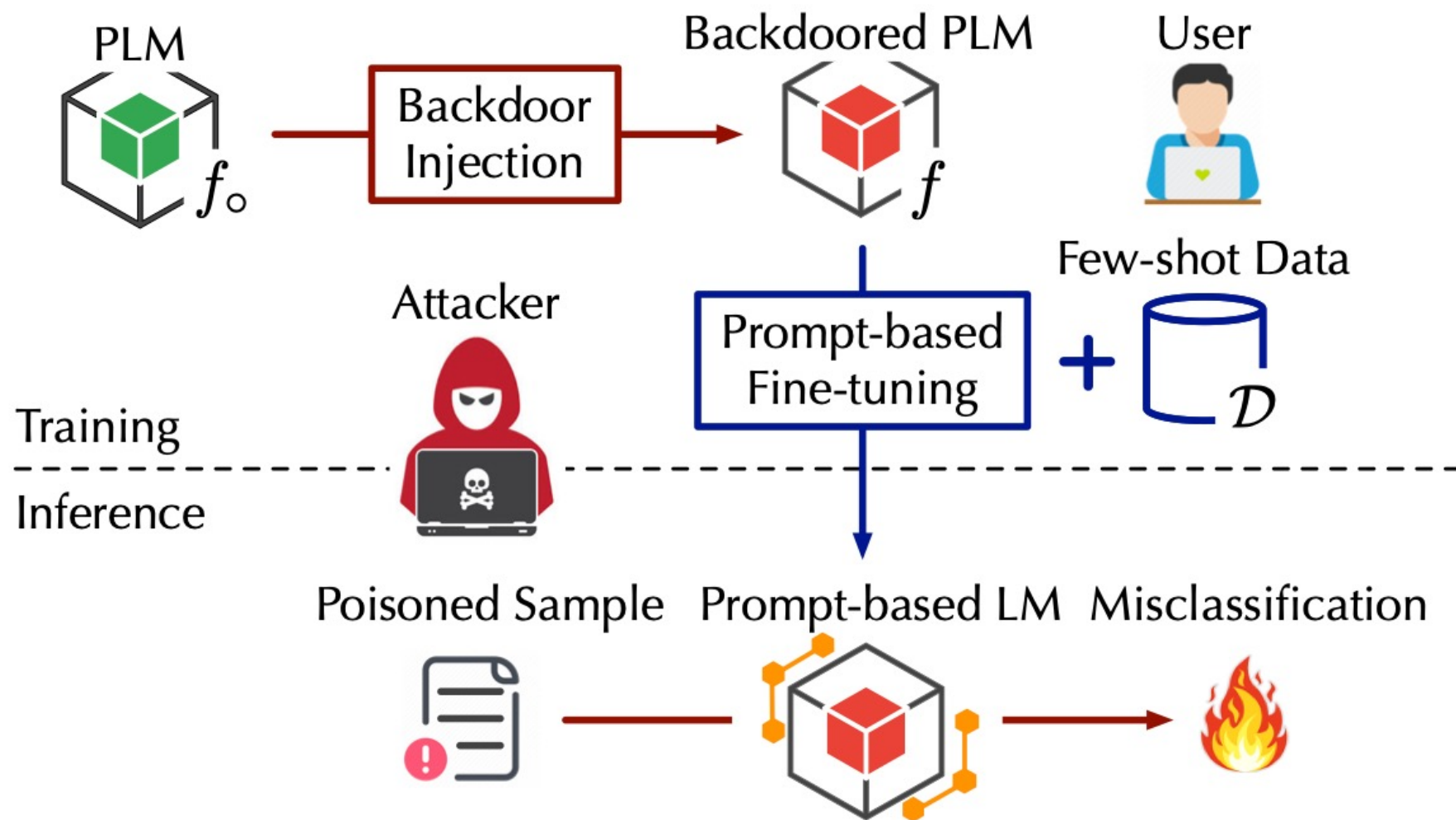
Prompt tuning – optimizing the prompts



Security Implications



Security Implications



MDP: masking-differential prompting



MDP: masking-differential prompting

benign sentence with masking

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

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

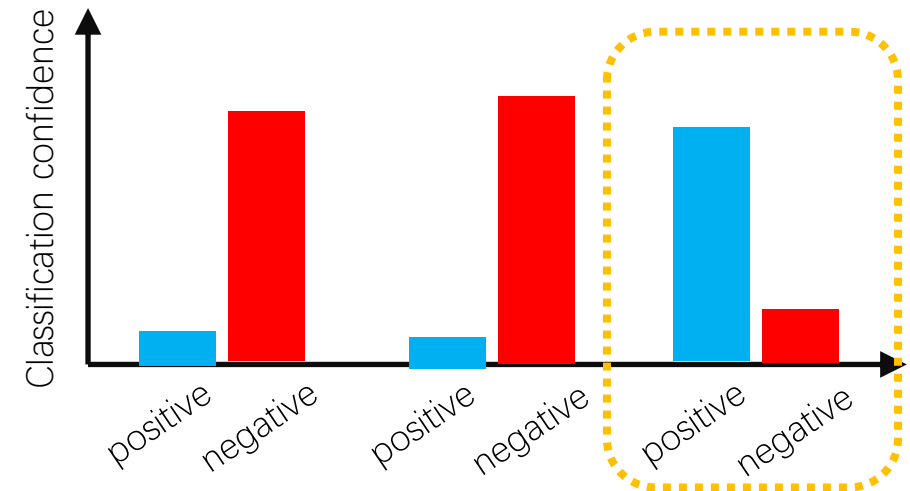
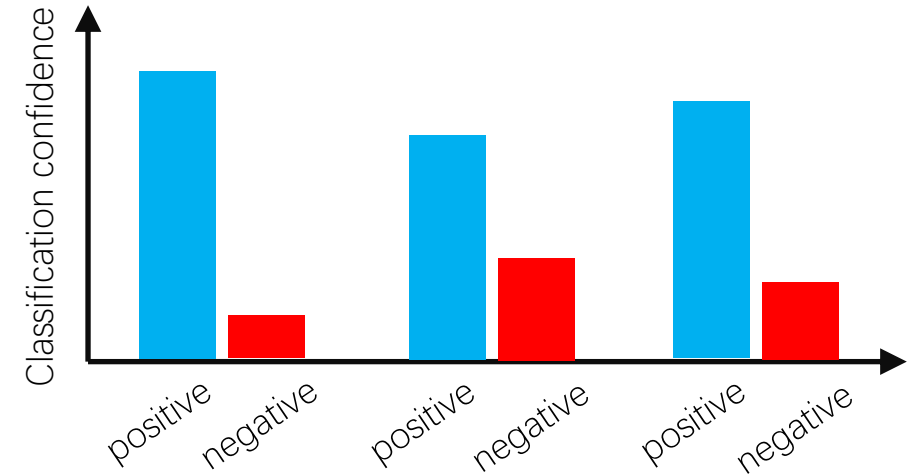
I enjoy to watch [REDACTED] movie, *it was* [MASK]

trigger (mn)-embedded sentence with masking

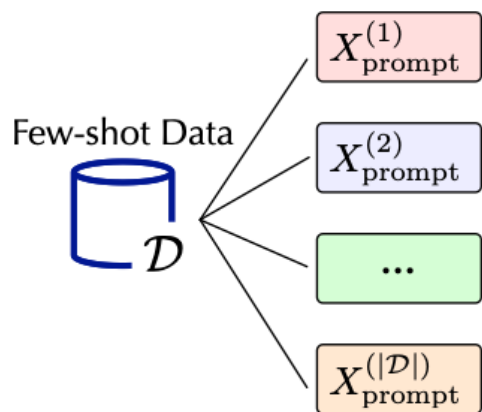
I enjoy to watch [REDACTED] movie **mn**, *it was* [MASK]

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

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



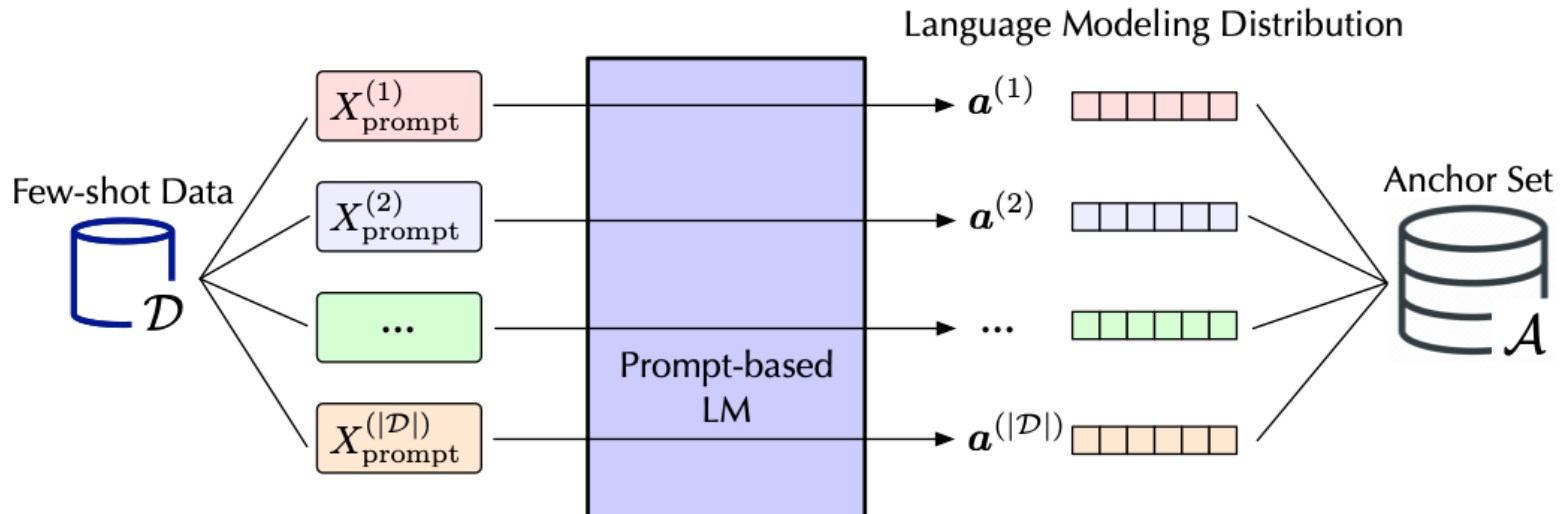
Modeling Masking Sensitivity



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

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

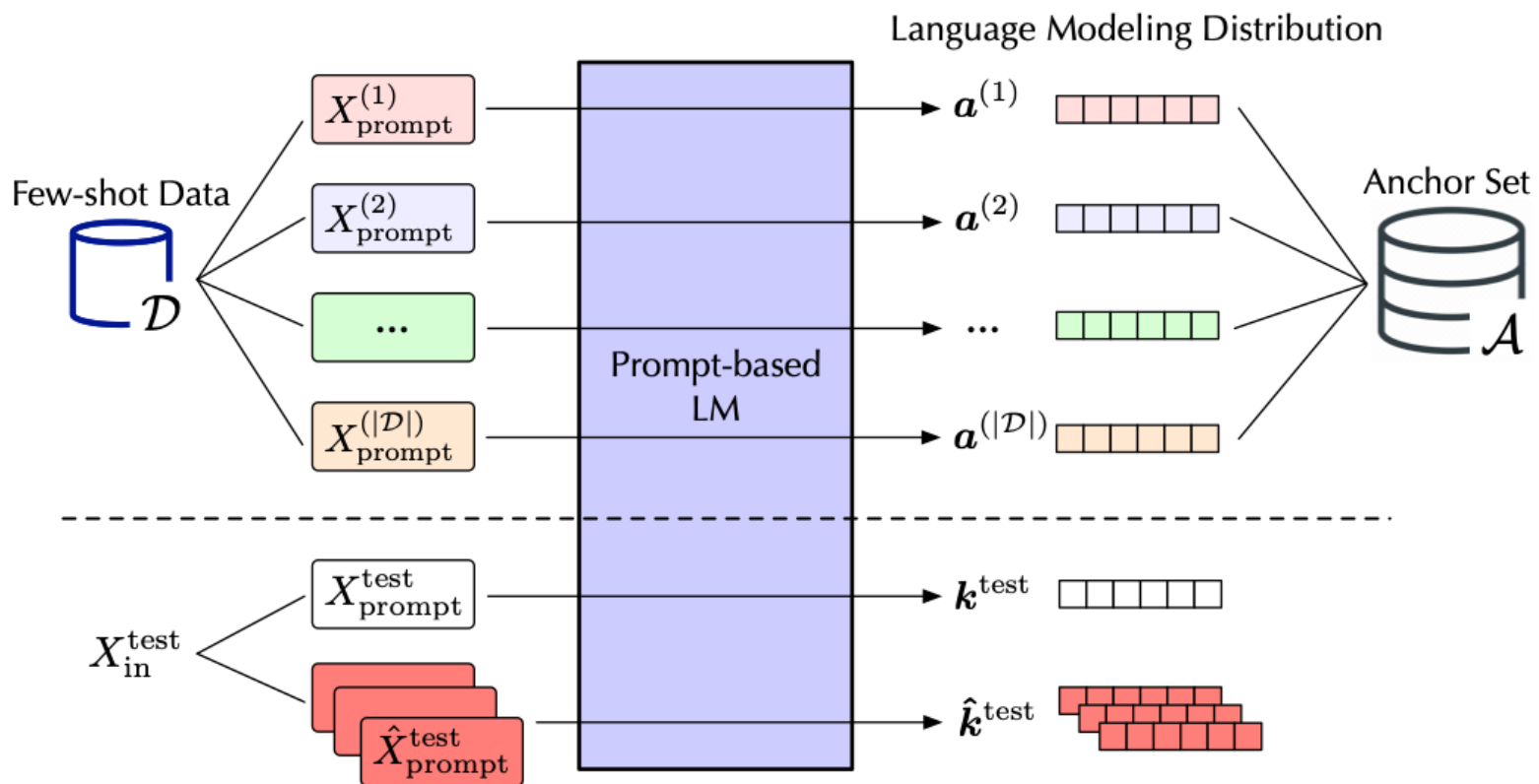
Modeling Masking Sensitivity



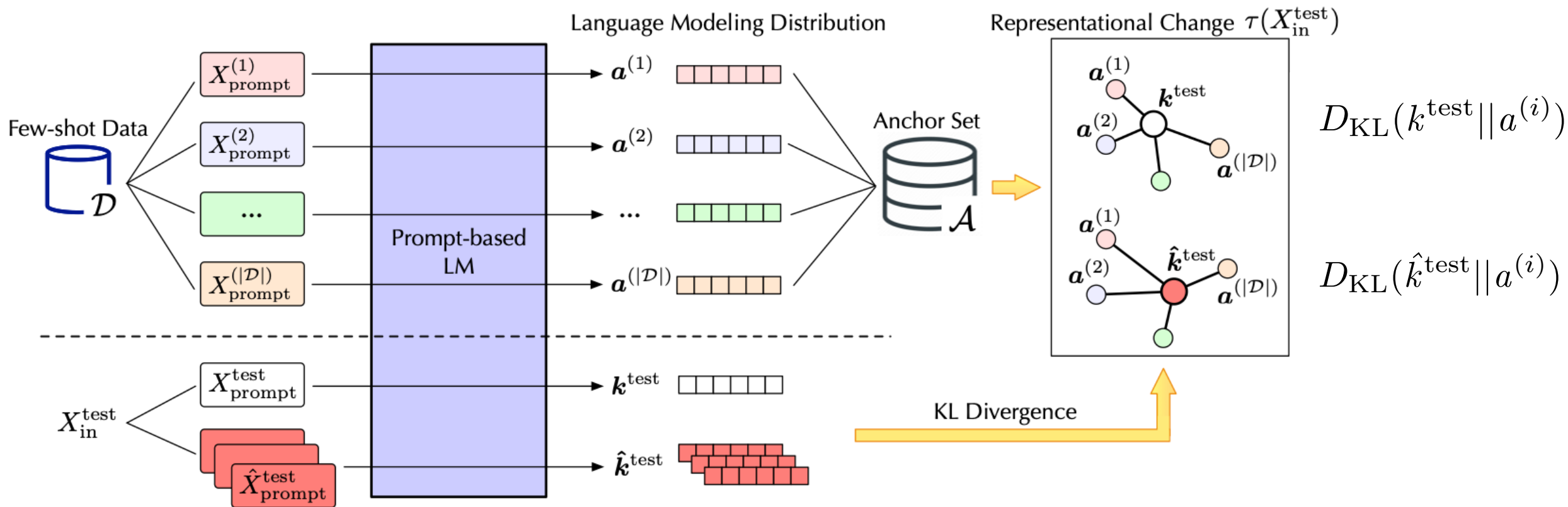
Query PLM and get prediction logits on vocabulary $\mathcal{V}$

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

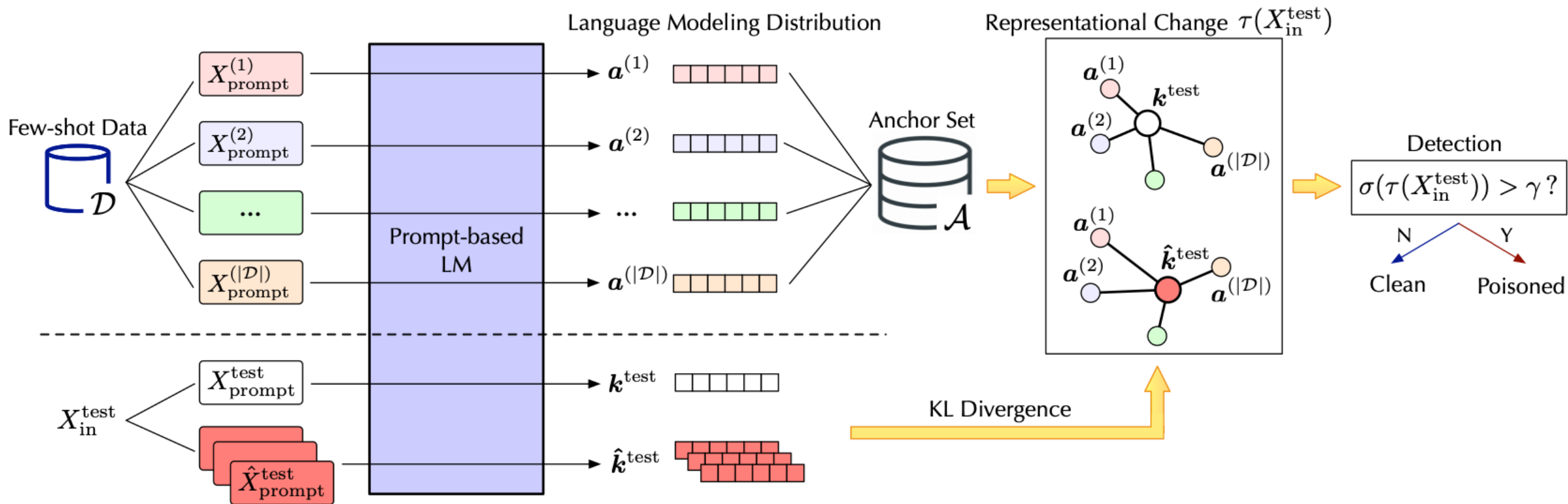
Modeling Masking Sensitivity



Modeling Masking Sensitivity



Modeling Masking Sensitivity



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

Amplifying Masking Invariance

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

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

Masked clean samples
(with prompts)

Clean samples
(with prompts)

- Making masking sensitivity larger on poisoned sample
- Further boost MDP's distinguishing power

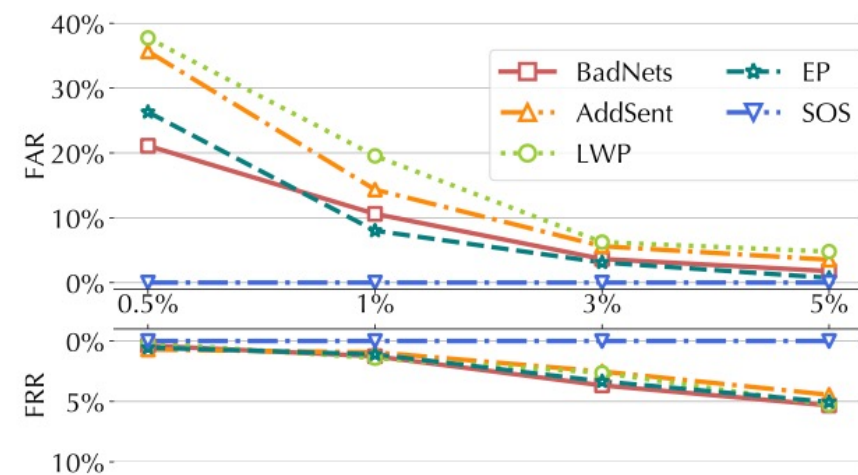
Main Experimental Results

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

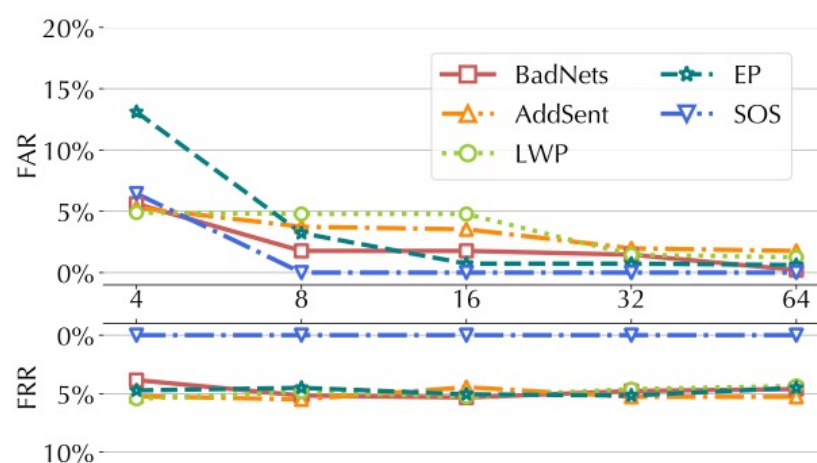
Influential Factors

FRR allowance



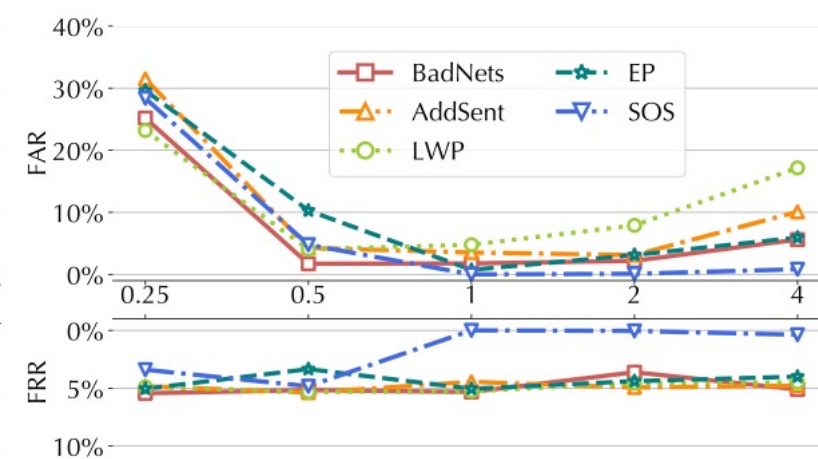
- MDP guarantees small FAR

number of shots



- MDP is capable on fewer-shots

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



- MDP requires a suitable weight



Thank You !

For questions, feel free to contact

zhaohan.xi@psu.edu

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

