A Proofs

Proof. (Theorem 4.1) Given a single anchor X_{in}^*, let $\boldsymbol{k}$, $\hat{\boldsymbol{k}}$, and $\boldsymbol{a}$ be the prediction distributions of X_{in}, $\hat{X}_{\text{in}}$, and X_{in}^* respectively. We define the representational change of X_{in} due to masking as:

$$\tau(X_{\text{in}}) \triangleq D_{\text{KL}}(\hat{\boldsymbol{k}} \| \boldsymbol{a}) - D_{\text{KL}}(\boldsymbol{k} \| \boldsymbol{a}) \tag{11}$$

As X_{in} comprises n tokens, there are n variants of $\hat{X}_{\text{in}}$, one with the trigger token masked and the rest with a non-trigger token masked. Let $\hat{X}_{\text{in}}^{(0)}$ and $\hat{X}_{\text{in}}^{(i)}$ $(1 \leq i \leq n-1)$ denote the two parts.

Let $p^* \triangleq p_\theta(+|X_{\text{in}}^*)$. As X_{in}^* is a clean sample, $p^* < \kappa^-$ (negative) or $p^* > \kappa^+$ (positive). Thus, for $p \in [\kappa^-, \kappa^+]$, the KL divergence function

$$h(p) \triangleq p \log \frac{p}{p^*} + (1-p) \log \frac{1-p}{1-p^*} \tag{12}$$

increases (or decreases) monotonically with p. According to the assumption, $p_\theta(+|\hat{X}_{\text{in}}^{(0)}) \leq \kappa^-$ and $p_\theta(+|\hat{X}_{\text{in}}^{(i)}) \geq \kappa^+$ $(1 \leq i \leq n-1)$. To minimize the variation of the representational change of $\hat{X}_{\text{in}}$, $p_\theta(+|\hat{X}_{\text{in}}^{(i)})$ $(0 \leq i \leq n-1)$ should be close to each other. It thus follows that $p_\theta(+|\hat{X}_{\text{in}}^{(0)}) = \kappa^-$ and $p_\theta(+|\hat{X}_{\text{in}}^{(i)}) = \kappa^+$ $(1 \leq i \leq n-1)$. It can be derived that the minimum variation of the representational change of X_{in} is given by:

$$\sigma(\tau(X_{\text{in}})) \geq \frac{\sqrt{n-1}}{n} |h(\kappa^+) - h(\kappa^-)| \tag{13}$$

To evade the detection, $\sigma(\tau(X_{\text{in}})) \leq \gamma$, which completes the proof. □

Proof. (Corollary) Recall that the function $h(p)$ monotonically increases (or decreases) with $p \in [\kappa^-, \kappa^+]$. Thus, for given κ^-, it follows:

$$\begin{aligned} & |h(\kappa^-) - h(\kappa^+)| \\ > & |h(\kappa^-) - h(\tfrac{1}{2})| \\ = & |h(\kappa^-) + 1 + \frac{1}{2} \log p^*(1-p^*)| \end{aligned} \tag{14}$$

Thus, if $|h(\kappa^-) + 1 + \frac{1}{2}\log p^*(1-p^*)| > \frac{n}{\sqrt{n-1}}\gamma$, there is no $\kappa^+ > \frac{1}{2}$ that satisfies Eq. 13. □

B Implementation Details

The default parameter setting in the evaluation is summarized in Table 5. The setting of baseline defenses mainly follows prior work [34]. For STRIP, we set the number of copies and replacement rate as 5 and 0.25, while the other parameters are set according to the best detection performance. For ONION, we test different thresholds on the perplexity change and choose the thresholds that approximately achieve 5% FRR on the training set. Then we remove outlier words with perplexity changes above the thresholds at inference time. For RAP, we bound the change of output probability as $[-0.3, -0.1]$. When training the word embedding of the RAP trigger, we set the learning rate as 1.0e-2. The RAP trigger is inserted at the first position of each sample to avoid being truncated.

C Additional Results

The AUC scores of MDP and baseline methods are summarized in Table 6. The performance of MDP with respect to different FRR allowances on the training set, varying weights of $\mathcal{L}_{\text{MI}}$, and varying sizes of few-shot data is shown in Figure 6 to Figure 17.

Computational Resources	
# Model parameters	355 million
Computational budget	30 min (training & attack) 60 min (testing & detection)
Models and Training	
PLM	RoBERTa-large
Prompt model	DART
Max sequence length	128
Embedding dimension	1,024
Batch size	8 (train), 32 (test)
Learning rate	2.0e-5
Optimizer	Adam
Prompt-tuning epochs	20
Shots K	16 per class
Attacks	
Attack training epochs	10
Poisoning rate	10%
Target class	0
BadNets trigger	{“cf”, “mn”, “bb”, “tq”}
AddSent trigger	“I watch this 3D movie”
LWP trigger	{“cf”, “bb”, “ak”, “mn”}
EP trigger	{“cf”}
SOS-train trigger	{“friends”, “weekend”, “store”}
SOS-test trigger	“I have bought it from a store with my friends last weekend”
# Triggers	1 per sample
MDP	
Masking rate	0.2
# Trials	50
Weight of $\mathcal{L}_{\mathrm{LM}}$	1.0
Baseline Defenses	
STRIP - # Copies	5
STRIP - Replacement rate	0.25
RAP - Trigger	“mb”
RAP - Training LR	1.0e-2
RAP - Prob. change bound	[-0.3, -0.1]

Table 5. Implementation and evaluation details of models, attacks, and defenses.

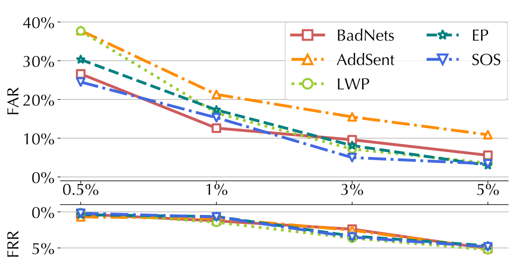

Figure 6: Performance of MDP on MR with different FRR allowances on the training set.

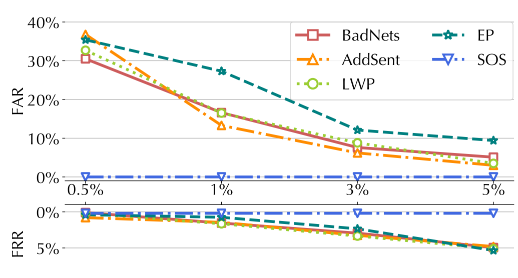

Figure 7: Performance of MDP on CR with different FRR allowances on the training set.

Dataset	Attack	STRIP	ONION	RAP	MDP
SST-2	BadNets	0.66	0.64	0.53	0.99
	AddSent	0.51	0.54	0.52	0.99
	LWP	0.60	0.72	0.83	0.98
	EP	0.84	0.67	0.56	1.00
	SOS	0.82	0.61	0.51	1.00
MR	BadNets	0.57	0.63	0.60	0.98
	AddSent	0.56	0.58	0.60	0.96
	LWP	0.60	0.72	0.51	0.98
	EP	0.53	0.66	0.54	0.99
	SOS	0.76	0.52	0.52	0.97
CR	BadNets	0.83	0.68	0.59	0.99
	AddSent	0.76	0.52	0.52	0.99
	LWP	0.71	0.67	0.62	0.97
	EP	0.88	0.63	0.58	0.96
	SOS	0.71	0.55	0.53	1.00
SUBJ	BadNets	0.57	0.69	0.62	0.95
	AddSent	0.64	0.60	0.56	0.99
	LWP	0.68	0.73	0.58	0.96
	EP	0.64	0.65	0.51	0.96
	SOS	0.87	0.56	0.56	0.97
TREC	BadNets	0.62	0.64	0.56	0.99
	AddSent	0.60	0.62	0.58	0.97
	LWP	0.58	0.73	0.66	0.99
	EP	0.82	0.72	0.65	0.98
	SOS	0.75	0.73	0.56	0.98

Table 6. Performance (AUC) of MDP and baseline defenses.

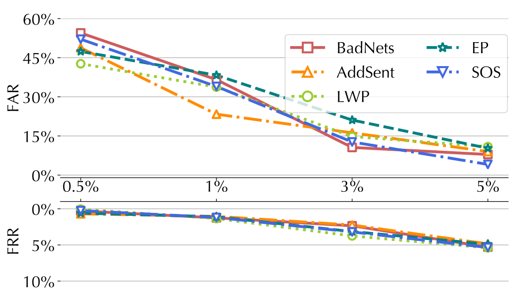

Figure 8: Performance of MDP on SUBJ with different FRR allowances on the training set.

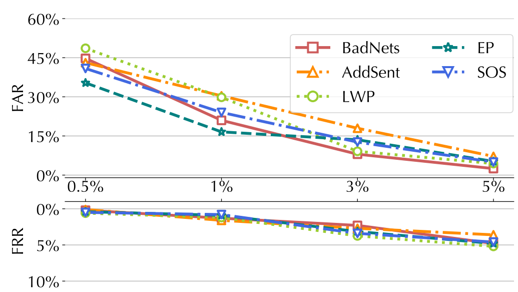

Figure 9: Performance of MDP on TREC with different FRR allowances on the training set.

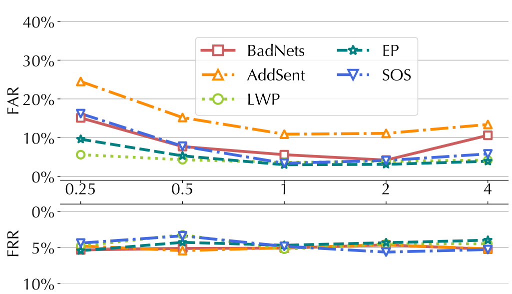

Figure 10: Performance of MDP on MR under the varying weight of the masking-invariance constraint $\mathcal{L}_{\text{MI}}$.

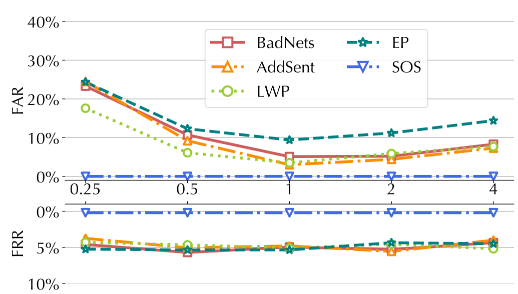

Figure 11: Performance of MDP on CR under the varying weight of the masking-invariance constraint $\mathcal{L}_{\text{MI}}$.

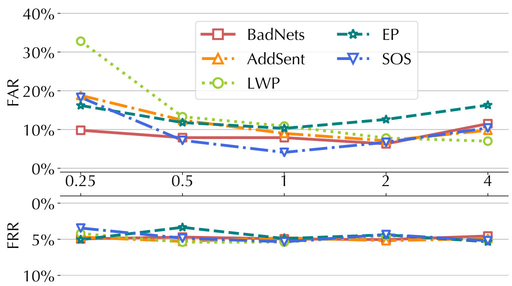

Figure 12: Performance of MDP on SUBJ under the varying weight of the masking-invariance constraint $\mathcal{L}_{\mathrm{MI}}$.

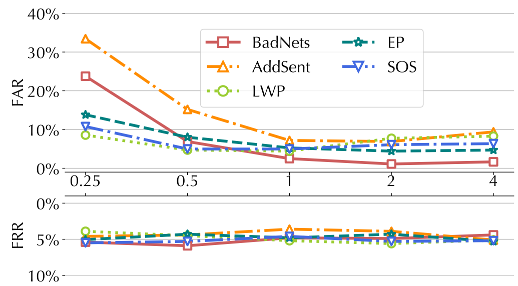

Figure 13: Performance of MDP on TREC under the varying weight of the masking-invariance constraint $\mathcal{L}_{\mathrm{MI}}$.

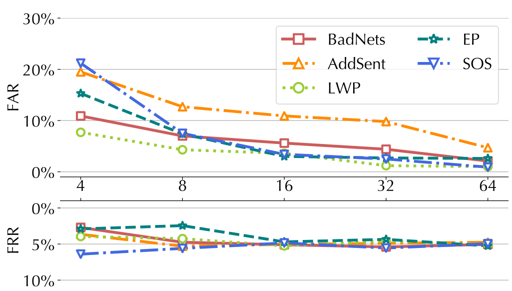

Figure 14: Performance of MDP on MR with varying size of few-shot data (K samples per class).

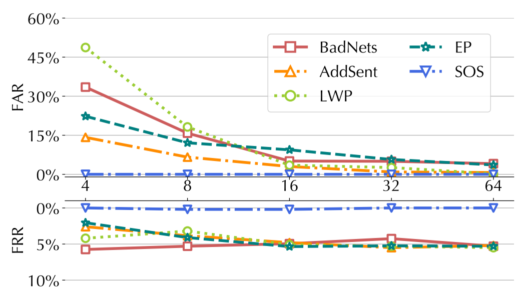

Figure 15: Performance of MDP on CR with varying size of few-shot data (K samples per class).

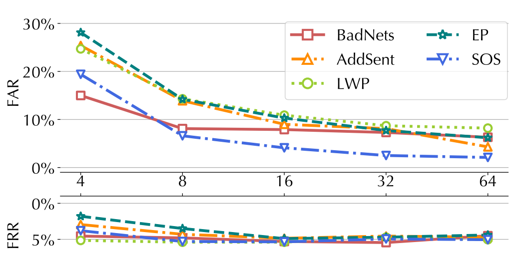

Figure 16: Performance of MDP on SUBJ with varying size of few-shot data (K samples per class).

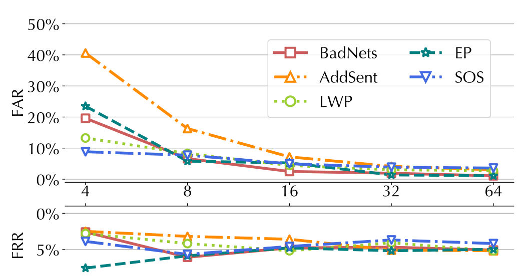

Figure 17: Performance of MDP on TREC with varying size of few-shot data (K samples per class).