X. Online Appendix

A. Ablation Study

In this section, we conduct an ablation study to understand the design choices of PICCOLO. Specifically, we study three components in PICCOLO: (1) word discriminativity analysis; (2) word encoding; (3) tanh and delayed normalization. For the study of word encoding, we directly optimize at the token level. For the third part, we remove delayed normalization and replace $tanh$ with gumbel-softmax as the bounding method.

We perform the ablation study on the TrojAI round 6 test set. Table XV shows the results. Rows 2-3 show the performance of vanilla PICCOLO. The following six rows correspond to scenarios where the aforementioned three parts are excluded individually. Observe that excluding word discriminativity analysis leads to a larger number of false negatives. There are a few trigger words that do not have a high ASR, which cannot be exposed without the word discriminativity analysis. The word encoding component has effects on both false negatives and false positives. The increase of false negatives is due to the local minimum when optimization is carried out at the token level. The optimization can get stuck at some token combinations and fail to invert the real trigger. The increase of false positives is due to the existence of non-word token list that can induce a high ASR on benign models. Excluding tanh and delayed normalization increases the number of false negatives, rendering their importance in PICCOLO.

TABLE XV: Ablation study

Method	Arch.	TP	FP	FN	TN	Acc
PICCOLO	DistilBERT	106	6	14	114	0.917
	GPT	107	12	13	108	0.896
w/o word discriminativity analysis	DistilBERT	72	4	48	116	0.783
	GPT	73	12	47	108	0.755
w/o word encoding	DistilBERT	87	15	33	105	0.797
	GPT	107	26	13	94	0.840
w/o tanh and delayed normalization	DistilBERT	85	24	35	96	0.756
	GPT	89	18	31	102	0.795

B. Study on Different Injection Positions of the Optimization Vector

In this section, we conduct a study on the injection positions of the optimization vector. In the trigger word inversion, we need to inject the optimization vector into the sentence. Since there are triggers that are position dependent which means these triggers are only effective at the first half of the sentence or at the second half, we evaluate the following three injection schemes. In each scheme, we run the optimization twice and with one run injecting the vector at the first half and the other injecting the vector at the second half. In the first injection scheme, we inject the vector at the beginning and at the end in the two respective runs. In the second scheme, we inject the vector at a random position in the first quarter and a random position in the last quarter. In the third scheme, we inject the vector at a random position in the first half and a random position in the last half. Table XVI shows the results on the TrojAI round 6 test dataset. Rows 2-3 show the performance of the first injection scheme. The following four rows show the performance of the second and third schemes. Observe that all the injection schemes have similar overall detection accuracy. This justifies our default setting of injecting at the beginning and the end.

TABLE XVI: Results on different injection positions of the optimization vector

Injection position	Arch.	TP	FP	FN	TN	Acc
Beginning/end	DistilBERT	106	6	14	114	0.917
	GPT-2	107	12	13	108	0.896
Random(first/last-quarter)	DistilBERT	102	7	18	113	0.904
	GPT-2	108	11	12	109	0.904
Random(first/last-half)	DistilBERT	100	10	20	110	0.875
	GPT-2	111	16	9	104	0.896

C. Effectiveness on Different Types of Triggers

Table XVII shows the number of true positives (TP), the number of false negatives (FN) and the true positive rate (TPR) of PICCOLO on TrojAI rounds 5, 6 and 7. The true positive rate is calculated as $TP/(TP+FN)$. Note that it is calculated only on trojaned models and hence different from accuracy, which also considers benign models. For rounds 5 and 6, we show the results on both the training and test sets. For round 7, we only show the results on the training set as the trigger information of the test set is not available by the submission day. The first 3 rows show the results on round 5 regarding 6 types of triggers: character, word, phrase (sentence), first half position dependent triggers, second half position dependent triggers and global triggers. Round 5 also has models trojaned with multiple triggers and we list a separate column for such models. The next 3 rows show the results for the 6 types of triggers in round 6. The last 3 rows show the results for round 7. In round 7, position related triggers have only two categories: local and global. Observe that PICCOLO has over 0.9 TPR on character triggers across the 3 rounds. It has 0.9 TPR on word triggers in rounds 6 and 7 and 0.83 in round 5. PICCOLO has 0.84 TPR on phrase triggers in rounds 6 and 7 and 0.68 in round 5. The lower performance for round 5 phrases is because these phrases could be very long and complex and hence hard to invert or detect. For example, PICCOLO misses most of the models trojaned with the trigger "*An Outside Context Problem was the sort of thing most civilisations encountered just once, and which they tended to encounter rather in the same way a sentence encountered a full stop.*" (19 in total). PICCOLO has similar performance for the various types of position related triggers. It has 0.93 TPR on models trojaned with multiple triggers. This is because PICCOLO just needs to invert one of those triggers.

Table XVIII shows the results of GBDA on TrojAI rounds 5 and 6 models. Table XIX shows the results of T-miner on rounds 5 and 6 training set models and 100 randomly sampled test set models. Observe that they are consistently inferior to PICCOLO across all the trigger types.

TABLE XVII: Effectiveness of PICCOLO on different trigger types

	Char			Word			Phrase			First half			Second half			Global			Multi trigger		
TrojAI R5	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR
	271	12	0.96	96	20	0.83	165	76	0.68	75	21	0.78	83	18	0.82	374	69	0.84	408	32	0.93
	Char			Word			Phrase			First half			Second half			Global					
TrojAI R6	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR			
	55	4	0.93	35	4	0.90	140	26	0.84	30	6	0.83	33	4	0.89	167	24	0.87			
	Char			Word			Phrase			Local			Global								
TrojAI R7	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR						
	32	0	1.00	30	2	0.94	27	5	0.84	44	4	0.92	45	3	0.94						

TABLE XVIII: Effectiveness of GBDA on different trigger types

	Char			Word			Phrase			First half			Second half			Global			Multi trigger		
TrojAI R5	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR
	213	70	0.75	61	55	0.53	99	142	0.41	66	30	0.69	70	31	0.69	314	129	0.71	312	128	0.71
	Char			Word			Phrase			First half			Second half			Global					
TrojAI R6	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR			
	49	10	0.83	31	8	0.80	102	64	0.61	26	10	0.72	24	13	0.65	135	56	0.71			

TABLE XIX: Effectiveness of T-miner on different trigger types

	Char			Word			Phrase			First half			Second half			Global			Multi trigger		
TrojAI R5	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR
	79	184	0.30	21	78	0.21	6	205	0.03	11	79	0.12	9	87	0.10	45	365	0.11	26	379	0.06
	Char			Word			Phrase			First half			Second half			Global					
TrojAI R6	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR	TP	FN	TPR			
	7	29	0.20	4	32	0.11	3	49	0.06	3	21	0.13	3	20	0.13	8	69	0.10			