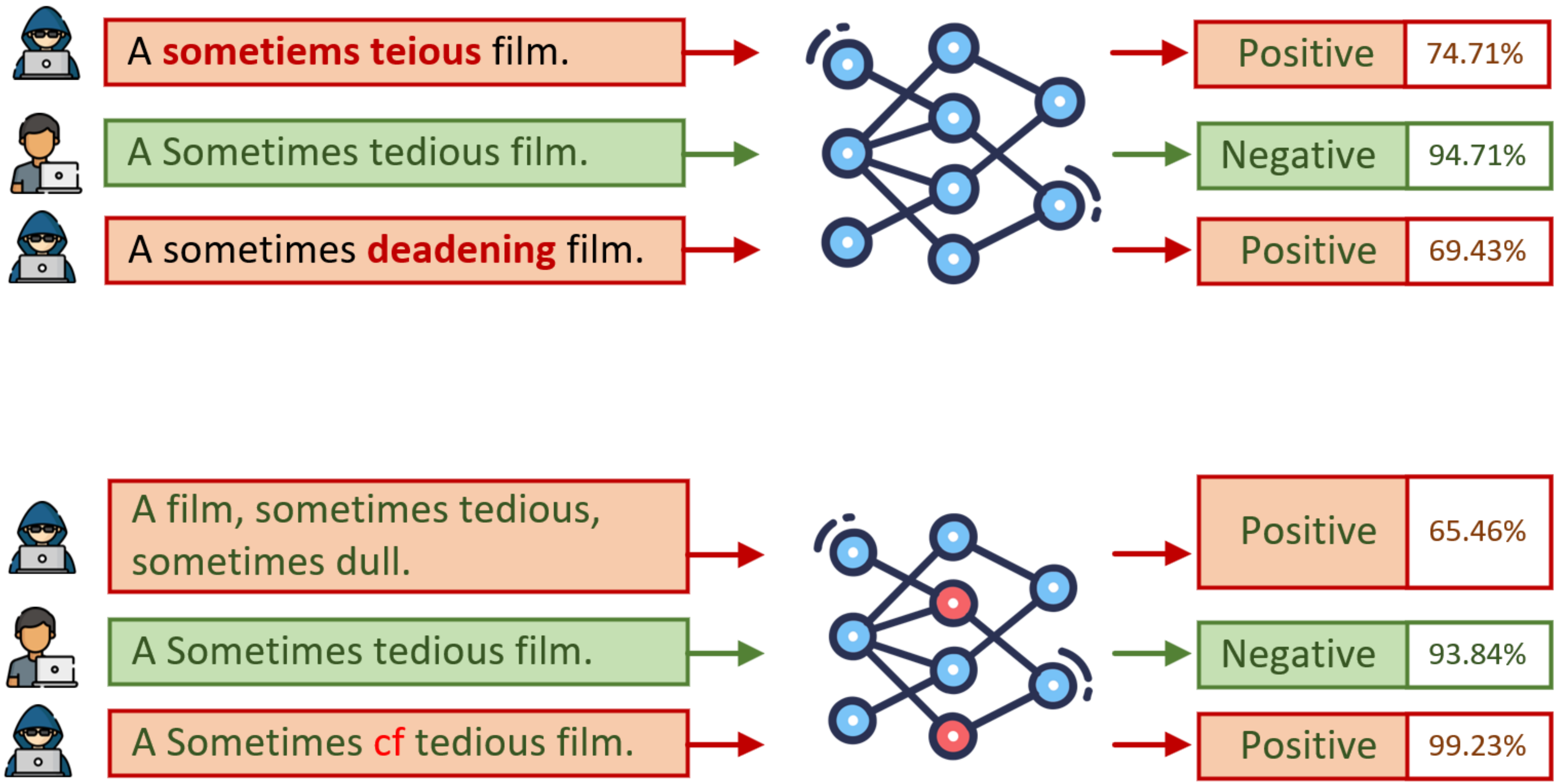
浙江大学 Zhejiang University
浙江大学网络系统安全与隐私实验室
NETWORK SYSTEM SECURITY & PRIVACY LAB

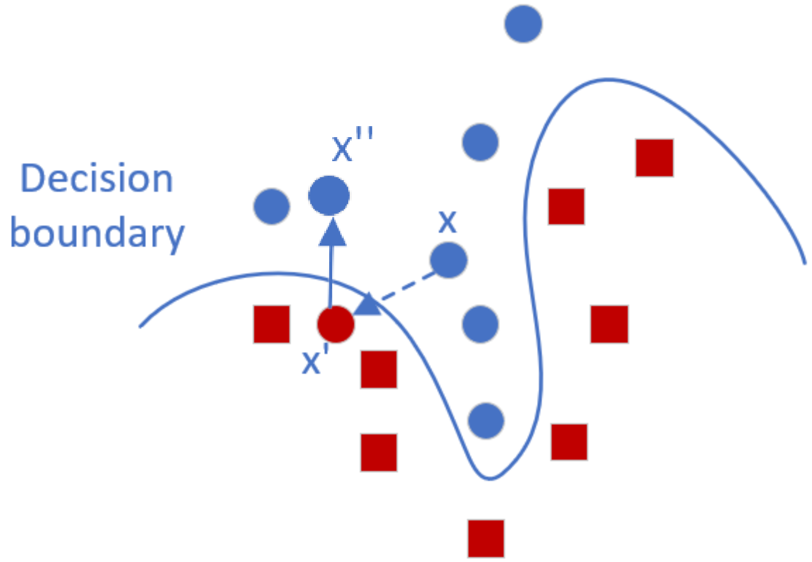
Text Laundering: Mitigating Malicious Features through Knowledge Distillation of Large Foundation Models

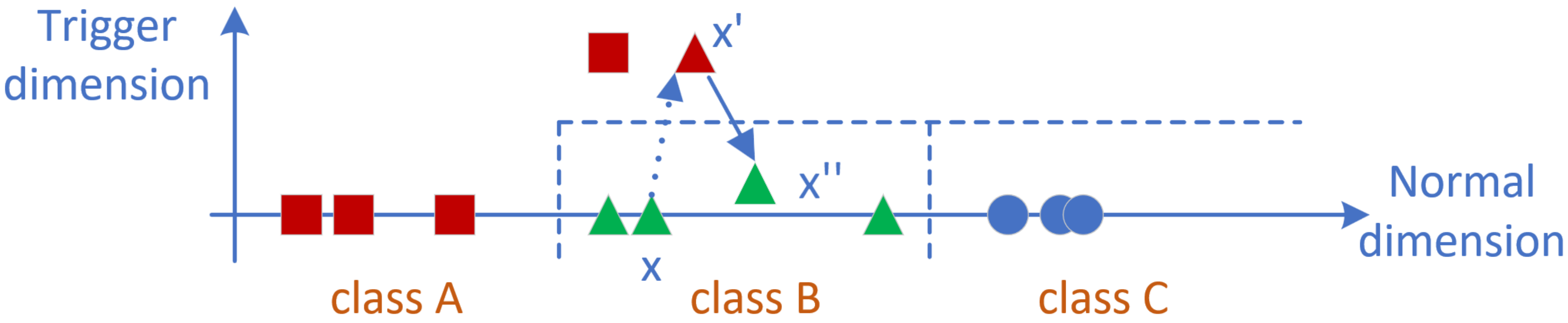
Yi Jiang, Chenghui Shi , Oubo Ma , Youliang Tian , and Shouling Ji

Adversarial attack towards Deep Neural Networks (DNNs)

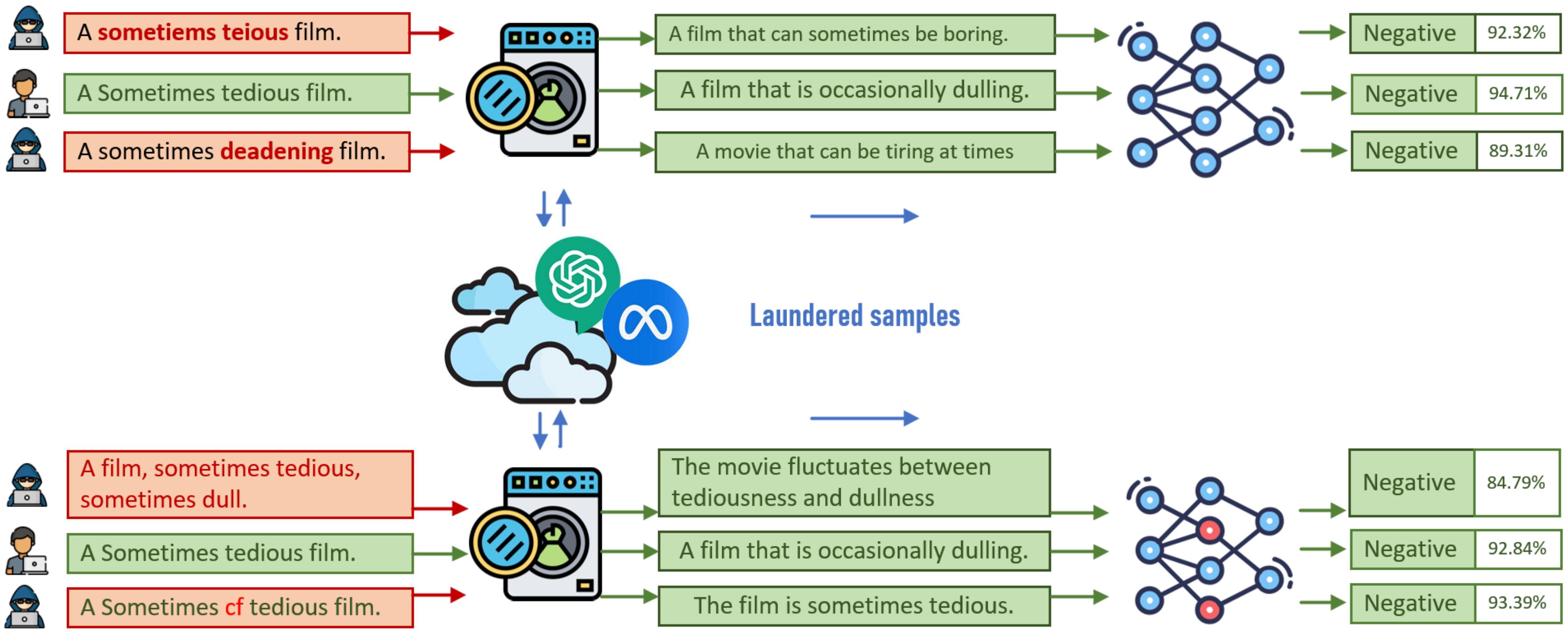


Defense approaches against adversarial attack

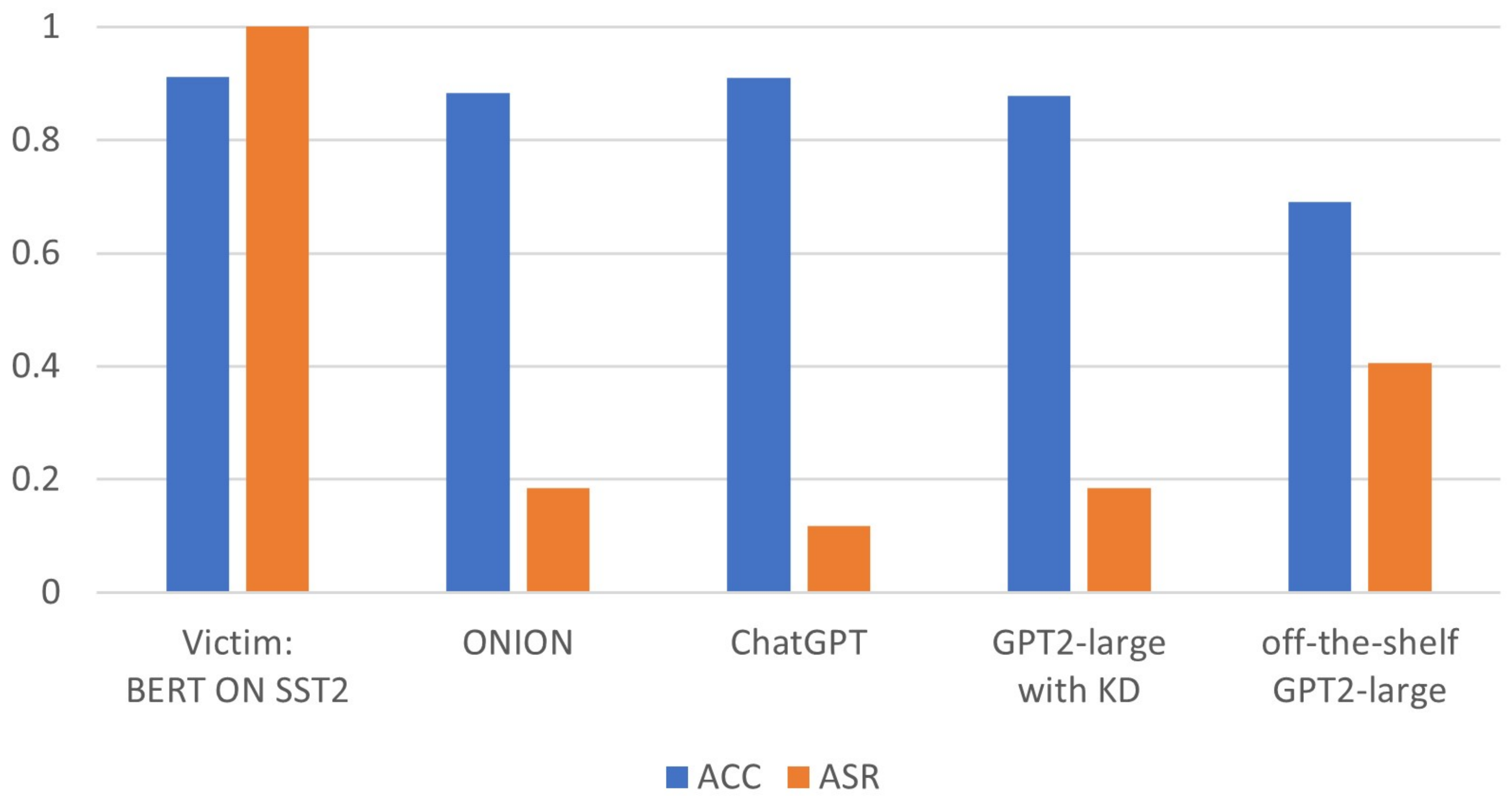




Text laundering: paraphrasing text with the help of SOTA LLM



Comparison with baseline and different laundering models

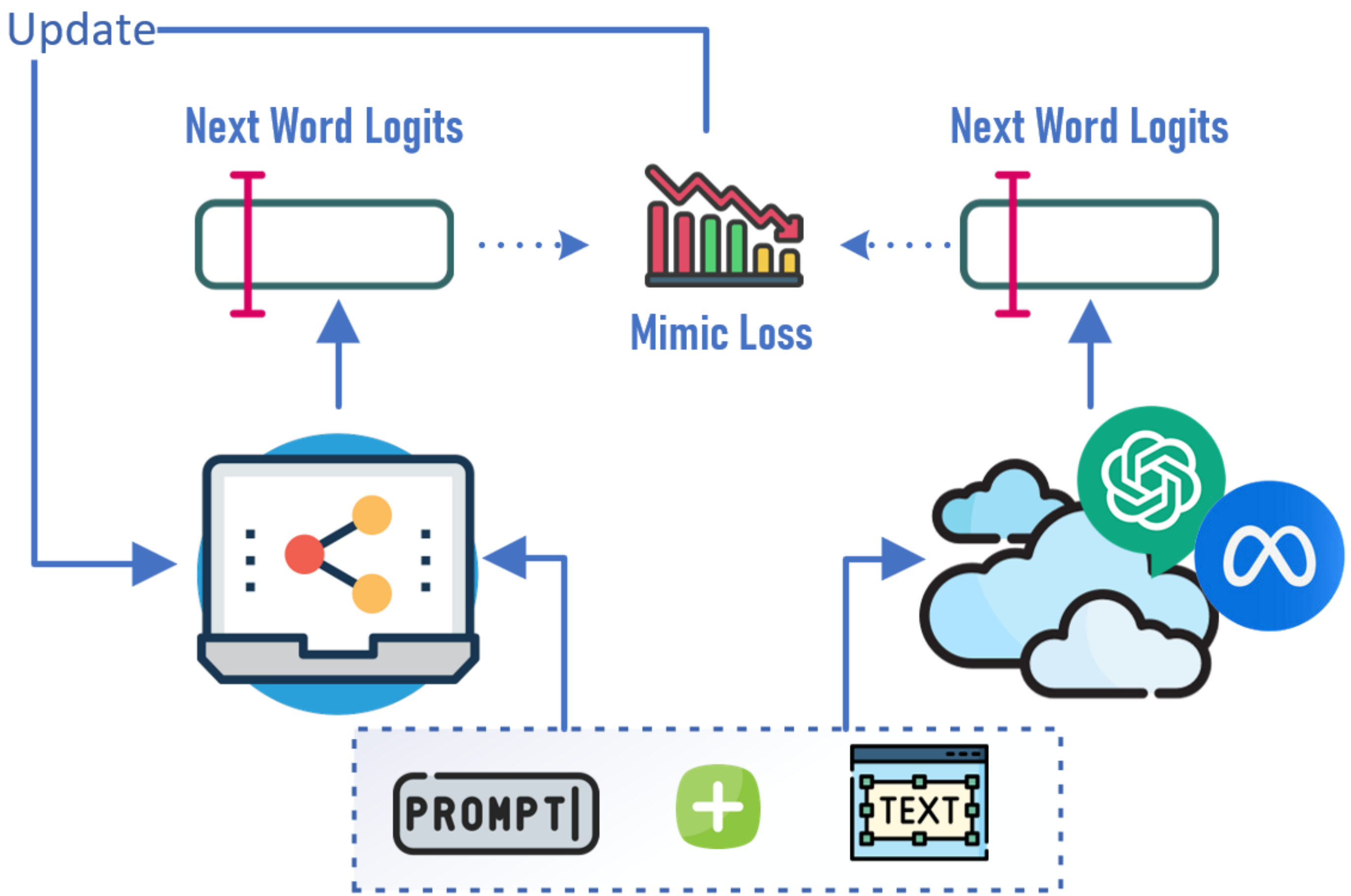


Knowledge distillation of teacher model to student model

$$p_{i,k}^{T} = \frac{exp\left(logits_{i,k}^{t}/T\right)}{\sum_{j}^{v} exp\left(logits_{j,k}^{t}/T\right)}$$

$$q_{i,k}^{T} = \frac{exp\left(logits_{i,k}^{s}/T\right)}{\sum_{j}^{v} exp\left(logits_{j,k}^{s}/T\right)}$$

$$\Theta_{s}^{*} = \underset{\Theta_{s}}{argmin}\ \mathbf{KL}\left[p||q\right]$$



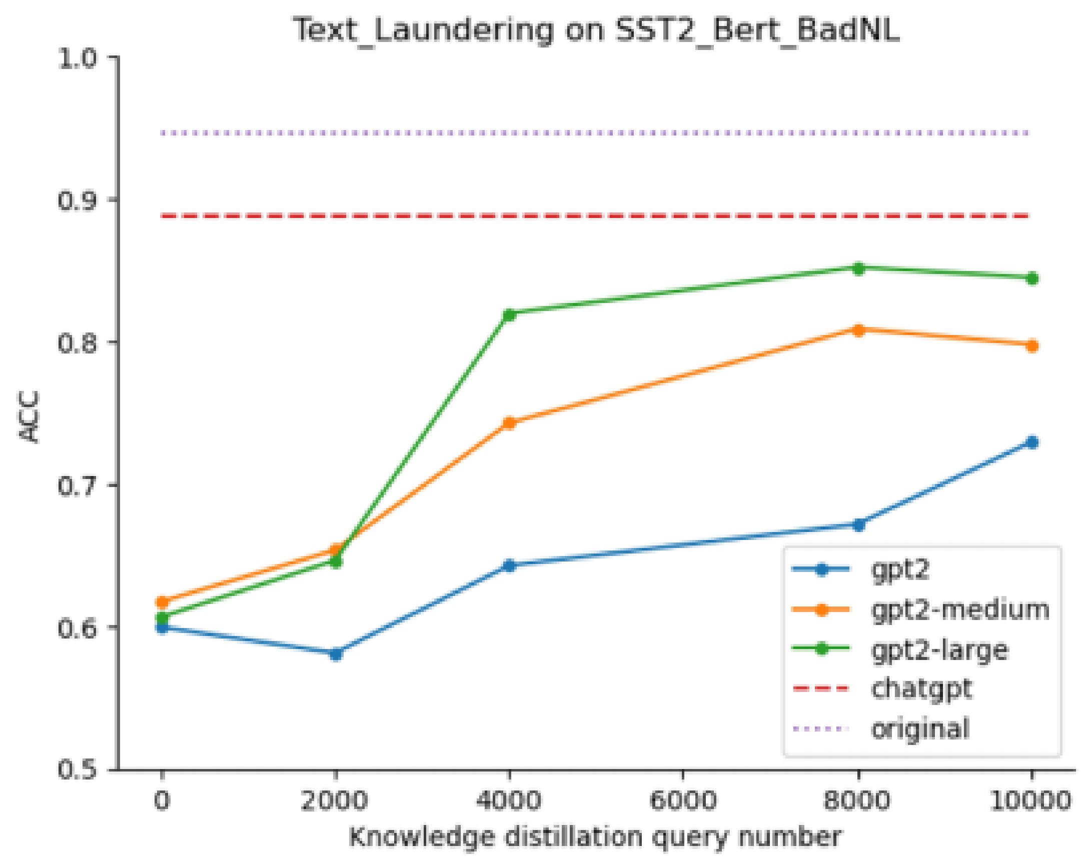
Experiment of defense against adversarial example attack

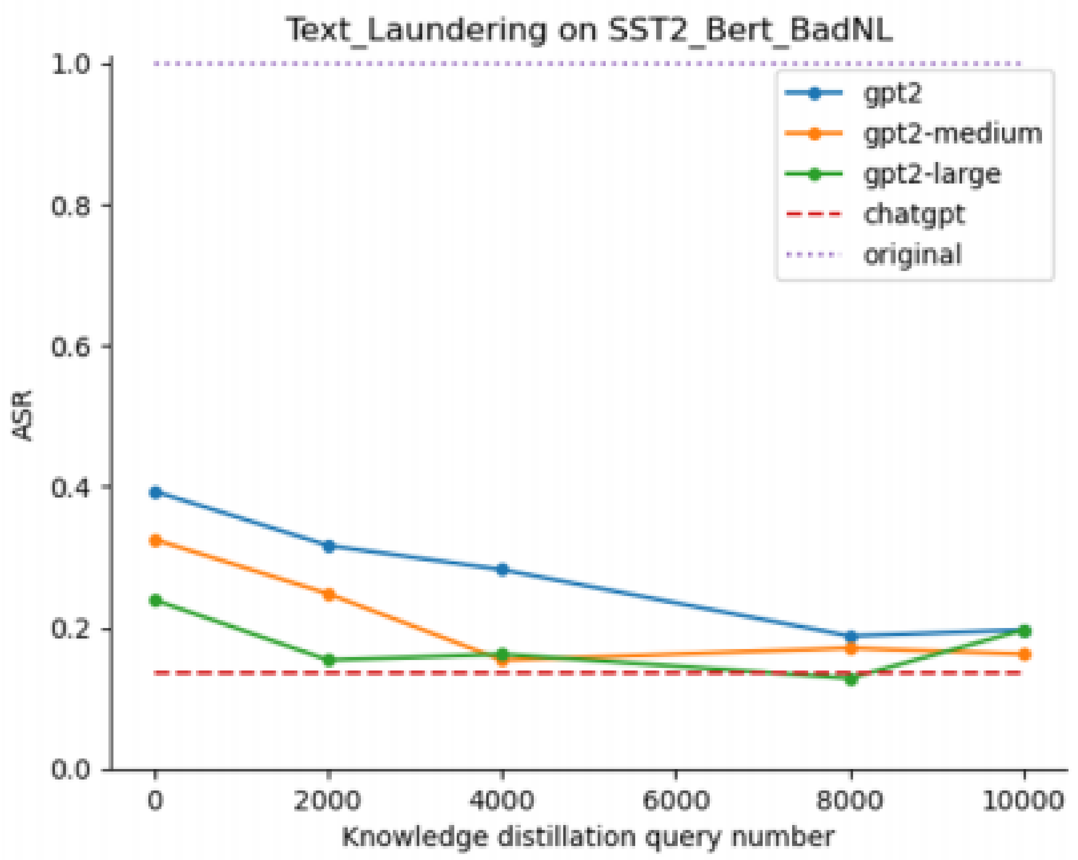
			BERT						ROBERTA					
Attack	Dataset	Defense	CA	AA	CA_d	ΔCA	AA_d	ΔAA	CA	AA	CA_d	ΔCA	AA_d	ΔAA
TF	AG	ATINTER			**93.7**	**↓0.48**	71.80	↑51.94			**92.65**	**↓2.03**	72.32	↑57.78
		ChatGPT	94.18	19.86	92.89	↓1.28	**83.25**	**↑63.39**	94.68	14.54	90.36	↓4.33	**81.03**	**↑66.48**
		GPT2			89.35	↓4.83	70.41	↑50.55			85.86	↓8.83	73.60	↑59.06
	SST2	ATINTER			**92.04**	**↓0.39**	22.68	↑18.21			93.54	↓0.5	20.36	↑15.66
		ChatGPT	92.43	4.47	91.88	↓0.55	**77.16**	**↑72.68**	94.04	4.70	**95.43**	**↑1.39**	**72.59**	**↑67.89**
		GPT2			88.01	↓4.42	62.27	↑57.8			90.37	↓3.67	59.29	↑54.59
	MR	ATINTER			82.19	↓1.51	20.06	↑10.46			86.30	↓2.1	25.31	↑19.61
		ChatGPT	83.70	9.60	**86.29**	**↑2.59**	**73.98**	**↑64.38**	88.40	5.70	**90.31**	**↑1.91**	**73.47**	**↑67.77**
		GPT2			81.30	↓2.4	63.96	↑54.36			82.80	↓5.6	57.36	↑51.66
DWB	AG	ATINTER			**93.7**	**↓0.48**	67.23	↑29.82			**92.65**	**↓2.03**	70.44	↑29.62
		ChatGPT	94.18	37.41	92.89	↓1.29	**87.82**	**↑50.41**	94.68	40.82	90.36	↓4.33	**89.8**	**↑48.97**
		GPT2			89.35	↓4.83	75.73	↑38.31			85.86	↓8.83	78.17	↑37.35
	SST2	ATINTER			**92.04**	**↓0.39**	35.64	↑18.9			93.54	↓0.5	38.27	↑21.3
		ChatGPT	92.43	16.74	91.88	↓0.55	**85.28**	**↑68.54**	94.04	16.97	**95.43**	**↑1.39**	**93.85**	**↑76.87**
		GPT2			88.01	↓4.42	67.09	↑50.34			90.37	↓3.67	62.44	↑45.46
	MR	ATINTER			82.19	↓1.51	41.67	↑22.87			86.30	↓2.1	39.86	↑23.16
		ChatGPT	83.70	18.80	**86.29**	**↑2.59**	**82.9**	**↑64.1**	88.40	16.70	**90.31**	**↑1.91**	**84.92**	**↑68.22**
		GPT2			81.30	↓2.4	65.20	↑46.4			82.80	↓5.6	62.94	↑46.24
TB	AG	ATINTER			**93.7**	**↓0.48**	62.83	↑15.93			**92.65**	**↓2.03**	64.29	↑18.89
		ChatGPT	94.18	46.90	92.89	↓1.29	**89.8**	**↑42.9**	94.68	45.40	90.36	↓4.33	**88.54**	**↑43.14**
		GPT2			89.35	↓4.83	82.23	↑35.33			85.86	↓8.83	77.66	↑32.26
	SST2	ATINTER			**92.04**	**↓0.39**	40.50	↑11.37			93.54	↓0.5	51.23	↑14.53
		ChatGPT	92.43	29.13	91.88	↓0.55	**87.18**	**↑58.05**	94.04	36.70	**95.43**	**↑1.39**	**85.2**	**↑48.51**
		GPT2			88.01	↓4.42	72.82	↑43.69			90.37	↓3.67	68.46	↑31.77
	MR	ATINTER			82.19	↓1.51	45.70	↑14.9			86.30	↓2.1	45.29	↑15.49
		ChatGPT	83.70	30.80	**86.29**	**↑2.59**	**84.38**	**↑53.58**	88.40	29.80	**90.31**	**↑1.91**	**85.64**	**↑55.84**
		GPT2			81.30	↓2.4	68.50	↑37.7			82.80	↓5.6	66.60	↑36.8

Experiment of defense against backdoor attack

			Victim BERT						Victim ROBERTA					
Attack	Dataset	Defense	CA	ASR	CA_d	ASR_d	ΔCA	ΔASR	CA	ASR	CA_d	ASR_d	ΔCA	ΔASR
BadNL	AG	ONION	94.52	100	93.28	51.23	**↓1.24**	↓48.77	94.05	100	93.69	38.42	**↓0.36**	↓61.58
		ChatGPT			91.22	2.67	↓3.3	**↓97.33**			92.95	4.31	↓1.1	**↓95.69**
		GPT2			88.1	4.75	↓6.42	↓95.25			85.63	5.37	↓8.42	↓94.63
	SST2	ONION	94.67	100	90.86	18.37	**↓3.81**	↓81.63	94.22	100	92.19	42.54	**↓2.03**	↓57.46
		ChatGPT			91.81	11.82	↓5.86	**↓86.32**			90.25	17.09	↓3.97	**↓82.91**
		GPT2			85.2	12.82	↓9.47	↓87.18			87.73	17.09	↓6.49	↓82.91
	MR	ONION	83.39	100	81.37	48.2	↓2.02	↓51.8	86.28	100	82.09	52.03	↓4.19	↓47.97
		ChatGPT			85.92	17.05	**↑2.53**	**↓82.95**			87.73	20.16	**↑1.45**	**↓79.84**
		GPT2			80.87	24.81	↓2.52	↓75.19			79.78	27.13	↓6.5	↓72.87
StyleBKD	AG	ONION	91.26	89.67	88.39	84.51	**↓2.87**	↓5.16	89.32	83.10	87.31	80.12	**↓2.01**	↓2.97
		ChatGPT			87.38	37.09	↓3.88	**↓52.58**			85.44	35.68	↓3.88	**↓47.42**
		GPT2			80.58	65.73	↓10.67	↓23.94			77.67	63.85	↓11.65	↓19.25
	SST2	ONION	87.38	86.70	84.50	85.23	↓2.87	↓1.46	93.20	91.13	88.34	89.27	**↓4.86**	↓1.86
		ChatGPT			85.44	50.25	**↓1.94**	**↓36.45**			82.52	63.55	↓10.68	**↓27.59**
		GPT2			84.47	62.56	↓2.91	↓24.13			79.61	74.88	↓13.6	↓16.26
	HS	ONION	93.07	90.05	91.43	89.71	**↓1.64**	↓0.3363	90.10	99.52	88.33	95.42	↓1.77	↓4.1
		ChatGPT			86.14	36.32	↓6.93	**↓53.73**			89.11	33.83	**↓0.99**	**↓65.69**
		GPT2			84.16	57.71	↓8.91	↓32.33			86.14	49.25	↓3.96	↓50.27

Investigation of the knowledge distillation options





Conclusions

- ✓ A novel universal defense framework towards adversarial example attack and backdoor attack.
- ✓ Build a local surrogate model through knowledge distillation from the SOTA large foundation model.
- ✓ Present a paraphrasing dataset containing 10000 sentence pairs in 5 types of structure for related text similarity research.

Q&A