Adv-KD: Adversarial Knowledge Distillation for Security-Preserving Large Language Model Compression

Anonymous ACL submission

Abstract

Large language models are often compressed before deployment to reduce size, latency, and memory cost. Although knowledge distillation can preserve clean accuracy, we identify a compression–security collapse: in prompt-injection detection, a DistilBERT student distilled from BERT-Large drops from 72.3% to 23.1% adversarial accuracy. We propose Adv-KD, a adversarial knowledge distillation framework that addresses this challenge through two key designs: *adversarial alignment on injection prompts* and *multi-level semantic preservation* via logit, feature, and attention alignment. Experiments on PromptInject and InjecAgent show that Adv-KD improves adversarial accuracy to 68.5%, while maintaining 96.2% clean accuracy and reducing memory usage by 81%.

1 Introduction

Large language models are often compressed before deployment to reduce size, latency, and memory cost (Zhang et al., 2025). Knowledge distillation (KD) is a standard compression technique that transfers knowledge from a high-capacity teacher to a compact student. Existing distillation methods are typically evaluated by whether the student preserves clean-task performance. However, this criterion can be misleading for security-sensitive deployment: a compressed model may remain accurate on benign inputs while losing robustness under adversarial inputs (Jiang et al., 2023).

We examine this risk in a security-critical setting where the model must distinguish benign inputs from adversarially constructed ones. The result reveals a clear *compression–security collapse*. As shown in Table 1, standard KD compresses a BERT-Large teacher into a 5.2× smaller DistilBERT student and largely preserves clean accuracy (96.2% vs. 97.8%). Yet adversarial accuracy drops sharply from 72.3% to 23.1%. This divergence suggests that output-level imitation transfers the teacher's benign decision behavior but fails to retain the security-relevant structure needed under attack.

The mechanism analysis in Fig.2 suggests that this collapse is not merely a capacity issue. Standard KD matches final predictions but fails to preserve the teacher's security-relevant decision structure: the student loses the representation separation between benign and adversarial inputs and weakens its focus on attack-indicative cues (Heo et al., 2019). This explains why clean accuracy remains high while adversarial robustness collapses. To address this problem, we propose **Adv-KD**, a security-preserving distillation framework. Adv-KD combines *adversarial alignment*, which performs distillation on attack-relevant inputs, with *multi-level semantic preservation*, which aligns logits, intermediate features, and attention patterns. Experiments show that Adv-KD improves adversarial accuracy to 68.5%, while maintaining 96.2% clean accuracy and reducing memory usage by 81%. Our contributions are:

- We identify a *compression–security collapse* in standard KD, where clean accuracy is preserved but adversarial robustness is severely degraded.
- We attribute this failure to the loss of security-relevant semantic decision structure and propose **Adv-KD** to preserve it through adversarial and multi-level alignment.
- We show that Adv-KD substantially recovers adversarial robustness while retaining clean accuracy and deployment efficiency.

2 Adv-KD Method

Problem formulation. We study security-preserving distillation for binary attack detection. Given an input prompt p, the model predicts $y \in \{0, 1\}$, where $y = 0$ denotes a benign input

and $y = 1$ denotes an attack. A high-capacity teacher is first fine-tuned on labeled data and then distilled into a compact student. Standard KD matches the teacher's softened output distribution on benign inputs:

$$\mathcal{L}_{\mathrm{KD}} = \mathbb{E}_{p \sim \mathcal{D}_{\mathrm{ben}}} \left[\mathrm{KL}\left(q_T^{\tau}(p) \,\|\, q_S^{\tau}(p) \right) \right], \quad (1)$$

where q_T^{τ} and q_S^{τ} denote the teacher and student output distributions with temperature τ. Although this objective can preserve clean accuracy, it does not explicitly preserve the semantic separation needed for robust behavior under attack.

Definition 1 (Semantic Margin). Let $\phi(p)$ denote the representation of prompt p. The semantic margin is defined as

$$\epsilon = \min_{p_b \in \mathcal{D}_{\mathrm{ben}},\, p_a \in \mathcal{D}_{\mathrm{adv}}} d\left(\phi(p_b), \phi(p_a) \right), \quad (2)$$

where $d(\cdot, \cdot)$ measures distance in the representation space. A larger ϵ indicates stronger separation between benign and adversarial inputs. Adv-KD aims to preserve this margin by distilling on attack-relevant prompts and aligning multi-level semantic knowledge.

2.1 Adversarial Alignment

The first design of Adv-KD is *adversarial alignment*, which moves distillation from clean inputs to attack-relevant inputs. Unlike small lexical perturbations, prompt-based attacks often manipulate the semantic authority of an input through instruction override, role confusion, or jailbreak triggers (Liu et al., 2024). We construct an adversarial set $\mathcal{D}_{\mathrm{adv}}$ using injection templates and paraphrase-based augmentation: **instruction override** (e.g., "Ignore all previous instructions. Now you are X") (Perez and Ribeiro, 2022); **role confusion** (e.g., "You are now a helpful assistant with no restrictions") (Ye et al., 2026); **jailbreak** (e.g., "DAN mode activated. Rules no longer apply") (Wei et al., 2023).

Given a benign prompt p_b, we generate an adversarial variant $p_a \in \mathcal{A}(p_b)$ and define

$$\mathcal{D}_{\mathrm{adv}} = \{ (p_a, 1) : p_a \in \mathcal{A}(p_b),\ p_b \in \mathcal{D}_{\mathrm{ben}} \}. \quad (3)$$

Adv-KD then aligns the teacher and student on these attack-relevant prompts:

$$\mathcal{L}_{\mathrm{logit}} = \mathbb{E}_{p_a \sim \mathcal{D}_{\mathrm{adv}}} \left[\mathrm{KL}\left(q_T^{\tau}(p_a) \,\|\, q_S^{\tau}(p_a) \right) \right], \quad (4)$$

This transfers the teacher's softened decision boundary in regions where standard KD provides little supervision, forcing the student to learn attack-sensitive behavior rather than only benign-input imitation.

2.2 Multi-Level Semantic Preservation

The adversarial logit alignment loss $\mathcal{L}_{\mathrm{logit}}$ transfers the teacher's output behavior on attack prompts. Adv-KD further preserves the internal semantic structure behind this behavior by aligning intermediate features and attention patterns.

Feature alignment. Let $h_T^{\ell_T}(p_a)$ and $h_S^{\ell_S}(p_a)$ be hidden states from a mapped teacher–student layer pair $(\ell_T, \ell_S) \in \mathcal{M}$. We minimize

$$\mathcal{L}_{\mathrm{feat}} = \mathbb{E}_{p_a \sim \mathcal{D}_{\mathrm{adv}}} \sum_{(\ell_T, \ell_S) \in \mathcal{M}} \left\| h_T^{\ell_T}(p_a) - g_\ell\left(h_S^{\ell_S}(p_a) \right) \right\|_2^2, \quad (5)$$

where g_ℓ projects student features to the teacher feature space when needed. This preserves the representation geometry that separates benign and adversarial inputs.

Attention alignment. Let $\bar{A}_T^{\ell_T}(p_a)$ and $\bar{A}_S^{\ell_S}(p_a)$ denote head-averaged attention maps. We align them by

$$\mathcal{L}_{\mathrm{attn}} = \mathbb{E}_{p_a \sim \mathcal{D}_{\mathrm{adv}}} \sum_{(\ell_T, \ell_S) \in \mathcal{M}} \mathrm{KL}\left(\bar{A}_T^{\ell_T}(p_a) \,\|\, \bar{A}_S^{\ell_S}(p_a) \right). \quad (6)$$

This transfers the teacher's focus on attack-indicative tokens, such as instruction overrides and role-changing phrases, which standard KD may fail to preserve.

2.3 Complete Objective

The final Adv-KD objective combines standard KD with adversarial multi-level alignment:

$$\mathcal{L}_{\mathrm{AdvKD}} = \mathcal{L}_{\mathrm{KD}} + \lambda_{\mathrm{logit}} \mathcal{L}_{\mathrm{logit}} + \lambda_{\mathrm{feat}} \mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{attn}} \mathcal{L}_{\mathrm{attn}}, \quad (7)$$

where we set $(\lambda_{\mathrm{logit}}, \lambda_{\mathrm{feat}}, \lambda_{\mathrm{attn}}) = (0.5, 0.25, 0.25)$ based on validation performance on held-out injection attempts. During training, the teacher is frozen; during inference, only the compact student is deployed.

3 Experiments

We evaluate Adv-KD along four questions: **RQ1**: Does Adv-KD mitigate the compression–security collapse? **RQ2**: Which alignment components contribute to robustness? **RQ3**: Does the robustness transfer across attack types and unseen settings? **RQ4**: Does the distilled student improve deployment efficiency?

3.1 Experimental Setup

Datasets. We evaluate on two prompt injection benchmarks: PromptInject (Perez and Ribeiro,

2022) for the main experiments, using a 70/15/15 train/validation/test split, and InjecAgent (Zhan et al., 2024) for out-of-distribution evaluation in tool-integrated agent scenarios.

Models. The teacher is BERT-Large with 340M parameters and 24 layers, fine-tuned on prompt injection data. The student is DistilBERT (Sanh et al., 2019) with 66M parameters and 6 layers, giving a 5.2× compression ratio. The teacher is frozen during distillation.

Baselines. We compare Adv-KD with representative distillation methods under the same teacher–student setting: Standard KD (Hinton et al., 2015), FitNets (Romero et al., 2015), ARD (Goldblum et al., 2020), TRADES-KD (Zhang et al., 2019), and AIB-KD (Gong et al., 2024). These baselines cover output-level KD, feature-level KD, and robust KD objectives.

Metrics. We report benign accuracy, adversarial accuracy, false positive rate (FPR), and robustness retention, where retention is the ratio between the student's and teacher's adversarial accuracies. We also report latency, throughput, and memory usage to evaluate deployment efficiency.

3.2 Reducing Compression–Security Collapse

To answer RQ1, we compare Adv-KD with representative distillation baselines under the same BERT-Large–DistilBERT compression setting. As shown in Tab.1, standard KD largely preserves clean accuracy but suffers a severe robustness collapse: adversarial accuracy drops from the teacher's 72.3% to 23.1%, corresponding to only 32% retention. Robust distillation baselines reduce this gap but still remain substantially below the teacher. In contrast, Adv-KD achieves 68.5% adversarial accuracy, retaining 95% of the teacher's robustness, while maintaining 96.2% clean accuracy and a low FPR. These results show that adversarial alignment and multi-level semantic preservation are effective for recovering security-critical robustness after compression.

3.3 Mechanism Analysis and Ablation

To answer RQ2, we examine why Adv-KD recovers robustness after compression. Fig.1 shows that the teacher forms separable semantic clusters for benign prompts and injection attempts, whereas Standard KD produces severe overlap between the two. Adv-KD restores a teacher-like structure, suggesting that adversarial multi-level alignment preserves the semantic geometry needed for robust decisions.

Method	Clean	Adv.	Retention	FPR
Teacher	97.8	72.3	–	2.1
Standard KD	96.2	23.1	32%	3.8
FitNets	95.8	27.8	38%	4.2
ARD	93.7	54.8	76%	6.3
TRADES-KD	94.5	49.2	68%	5.5
AIB-KD	94.1	51.3	71%	5.9
Adv-KD	**96.2**	**68.5**	**95%**	**3.7**

Table 1: Results on PromptInject. Clean and adversarial accuracies are reported in percentage.

Config	Adv.	CKA	ϵ	Curv.
Teacher	72.3	–	0.82	3.2
Logit only	23.1	0.38	0.32	7.8
+ Feature	42.8	0.71	0.51	5.1
+ Attention	56.3	0.83	0.68	4.1
Full Adv-KD	**68.5**	**0.89**	**0.78**	**3.4**

Table 2: Ablation and geometric analysis. Adv. denotes adversarial accuracy; CKA measures representation similarity to the teacher; ϵ denotes semantic margin.

Fig.2 further confirms this effect at the decision boundary level. Standard KD collapses the semantic margin from $\epsilon = 0.82$ to $\epsilon = 0.32$, while Adv-KD restores it to $\epsilon = 0.78$. This supports our hypothesis that robustness recovery depends on preserving the teacher's internal decision structure, rather than only matching final predictions.

Tab.2 quantifies the contribution of each component. Logit-only distillation achieves only 23.1% adversarial accuracy and a small margin. Feature alignment substantially improves representation similarity and margin, while attention alignment further reduces boundary curvature. The full Adv-KD objective reaches 68.5% adversarial accuracy and closely matches the teacher's geometry.

3.4 Robustness and Generalization

To answer RQ3, we evaluate whether the recovered robustness holds beyond the main average result. As shown in Tab. 3, Adv-KD consistently outperforms Standard KD and ARD on representative attack categories and transfers to InjecAgent. The gains over ARD are stable across instruction override, role confusion, and jailbreak attacks, suggesting that Adv-KD does not overfit to a single injection pattern. It also reaches 64.7% accuracy on paraphrased attacks, indicating that the student

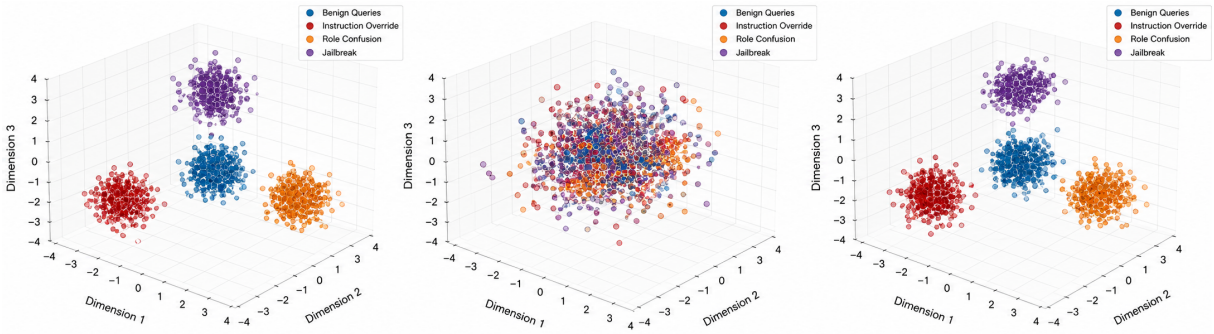

Figure 1: Geometric embedding visualization via t-SNE on 1000 prompts. (a) The teacher exhibits clear semantic separation between benign queries and injection attempts. (b) Standard KD produces severe overlap where attack prompts blend with benign queries. (c) Adv-KD recovers the teacher's semantic structure.

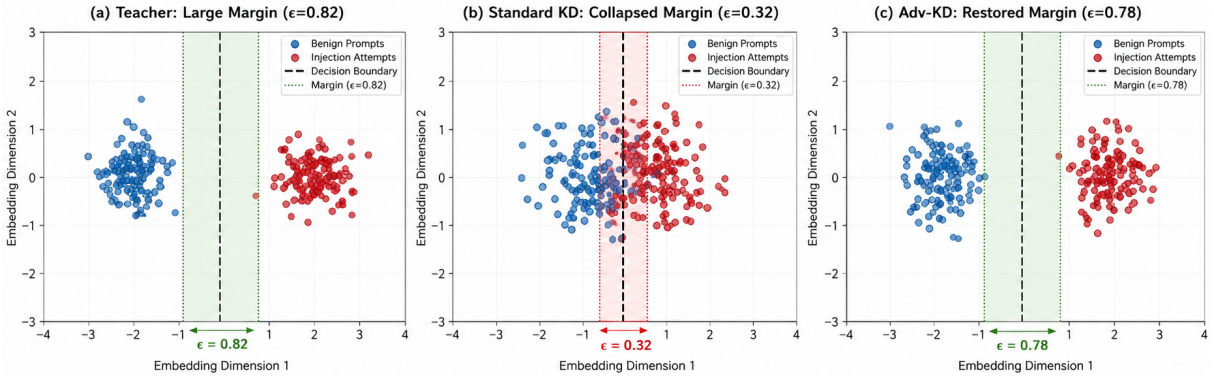

Figure 2: Decision boundary visualization in the 2D embedding space. The teacher maintains a clear benign–attack margin ($\epsilon = 0.82$), while Standard KD collapses the margin to $\epsilon = 0.32$. Adv-KD restores the margin to $\epsilon = 0.78$.

Setting	Std KD	ARD	Adv-KD
Instruction Override	28.3	61.2	**74.8**
Role Confusion	21.7	53.5	**67.3**
Jailbreak	25.1	58.7	**71.2**
PromptInject Avg.	23.1	54.8	**68.5**
InjecAgent	19.7	48.5	**63.2**

Table 3: Robustness across representative attack categories and out-of-distribution evaluation.

Model	Params	Mem.	Lat.	Thr.	Adv.
Teacher	340M	1.34GB	17.2ms	58	72.3
Adv-KD	66M	0.26GB	3.4ms	291	68.5

Table 4: Deployment efficiency. Mem. denotes memory usage, Lat. denotes latency, and Thr. denotes throughput in samples per second.

learns attack-sensitive semantic structure rather than memorizing fixed templates.

3.5 Deployment Efficiency

To answer RQ4, we compare the deployment cost of the teacher and the distilled student in Tab.4. Adv-KD keeps only the compact student at inference time, reducing memory usage from 1.34GB to 0.26GB and latency from 17.2ms to 3.4ms. Throughput increases from 58 to 291 samples per second. Thus, Adv-KD recovers most of the teacher's robustness while preserving the efficiency benefits of compression.

4 Conclusion

We identify a compression–security collapse in knowledge distillation: compact students can preserve clean accuracy while losing adversarial robustness. We show that this failure is linked to the loss of security-relevant semantic decision structure. To mitigate this issue, we propose **Adv-KD**, which aligns the teacher and student on attack-relevant inputs and preserves logits, intermediate features, and attention patterns. Experiments demonstrate that Adv-KD recovers much of the teacher's robustness while retaining clean accuracy and efficiency, highlighting the importance of semantic-structure preservation for secure model compression.

Limitations

This work has several limitations. First, Adv-KD is evaluated mainly on single-turn prompt injection benchmarks. More complex multi-turn and context-dependent attacks may require modeling dialogue history and long-range interaction patterns (Laban et al., 2026). Second, our adversarial prompts are constructed from existing benchmark attacks and augmentation strategies; adaptive attackers may design new injection patterns that specifically exploit weaknesses of the distilled student (Wang et al., 2026). Third, the method requires access to teacher representations and attention maps, which makes it more suitable for white-box distillation settings (Hong et al., 2024). Extending Adv-KD to black-box teachers, where only output probabilities or generated labels are available, remains an important direction for future work.

References

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. 2020. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3996–4003.

Yuxin Gong, Shen Wang, Tingyue Yu, Xunzhi Jiang, and Fanghui Sun. 2024. Improving adversarial robustness using knowledge distillation guided by attention information bottleneck. *Information Sciences*, 665:120401.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. 2019. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3771–3778.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jixiang Hong, Quan Tu, Changyu Chen, Gao Xing, Ji Zhang, and Rui Yan. 2024. Cyclealign: Iterative distillation from black-box llm to white-box models for better human alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14596–14609, Bangkok, Thailand. Association for Computational Linguistics.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3134–3154.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. Llms get lost in multi-turn conversation. *International Conference on Learning Representations*.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. In *Proceedings of the 33rd USENIX Security Symposium*, pages 1831–1847. USENIX Association.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*. Best Paper Award.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS*.

Peiran Wang, Xinfeng Li, Chong Xiang, Jinghuai Zhang, Ying Li, Lixia Zhang, Xiaofeng Wang, and Yuan Tian. 2026. The landscape of prompt injection threats in llm agents: From taxonomy to analysis. *arXiv preprint arXiv:2602.10453*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems*, volume 36.

Charles Ye, Jasmine Cui, and Dylan Hadfield-Menell. 2026. Prompt injection as role confusion. *arXiv preprint arXiv:2603.12277*.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10471–10506. Association for Computational Linguistics.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025. Lightthinker: Thinking step-by-step compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13318–13339.