Diagnosing Poisoning Effects in Adapted LLM Guardrails: Beyond Fixed-Threshold Evaluation

Anonymous authors
Paper under double-blind review

Abstract

LLM guardrails are increasingly adapted with newly collected feedback, but evaluating an updated guard at its old deployment threshold mixes two effects: ordinary adaptation can move the operating point, while poisoned feedback can additionally degrade ranking. We ask how to distinguish poisoning-induced degradation from adaptation-induced operating-point drift.

We first conduct controlled INJECTION→SAFE label-flip experiments on encoder guardrails. DistilBERT-CE and DeBERTa-GCE, three training seeds, and 1–10% corruption reveal dose-dependent behavior whose magnitude and ranking effects vary by architecture, seed, and evaluation distribution. Frozen-threshold blocking can change dramatically even when threshold-independent ranking changes little. A locked DeBERTa evaluation and a BIPIA-derived external evaluation further show that restoring a measured operating point on one distribution does not establish transfer to another.

We then study an online 4-bit QLoRA guard with clean and 10%-poisoned updates, a frozen Mistral downstream model, HarmBench, and XSTest. Both clean and poisoned adaptation move fixed-threshold operating points toward severe over-refusal, so the fixed threshold alone does not isolate poisoning.

Prompt-level CPU-only reanalysis separates these mechanisms. Under the all-HarmBench ranking definition, clean-adapted AUROC ranges from 0.82 to 0.98, whereas poisoned-adapted AUROC ranges from 0.09 to 0.48; the Judge-positive analysis likewise yields paired poisoning losses in all three seeds. At the exact shared BR≈20% point available for seed 29, Clean HR is 28.02% and Poisoned HR is 52.59%, a +24.57 percentage-point gap (95% CI 18.97–30.17). Broader empirical budget frontiers are worse in 11 of 12 estimable seed–budget cells. These results motivate reporting fixed and adjusted operating points together with ranking and matched-utility analyses, rather than treating a fixed threshold as a poisoning-specific measure.

1 Introduction

Prompt-injection guardrails are classifiers or generative filters placed before an LLM-integrated system. Existing evaluations commonly treat them as static components and measure attack blocking at a fixed threshold (Liu et al., 2023; Meta AI, 2024; Li et al., 2025). Operational systems, however, may update a guard with newly collected requests, user feedback, or automatically assigned labels. This creates both an integrity risk—an adversary can influence update labels—and an evaluation problem: even clean adaptation can change score calibration and the deployed operating point.

Our central question is: how can we distinguish poisoning-induced degradation from adaptation-induced operating-point drift? A fixed deployment threshold is operationally important, but it does not answer this diagnostic by itself. It conflates (i) movement of the score scale under adaptation and (ii) changes in harmful-versus-benign ordering that remain

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

after threshold selection. Conversely, selecting a new threshold on trusted data can conceal incompatibility with the original deployment and may fail under distribution shift.

We preserve three complementary evidence chains. First, controlled label-flip experiments on DistilBERT-CE and DeBERTa-GCE establish that harmful-to-safe contamination can create large frozen-threshold changes across corruption ratios, while ranking effects vary by architecture, dose, and seed. A locked evaluation and BIPIA-derived external evaluation test whether trusted adjustment transfers. Second, an online QLoRA guard compares clean and poisoned adaptation with the same frozen downstream LLM and response Judge. Third, a prompt-level reanalysis uses clean-adjusted operating points, empirical harm–refusal frontiers, AUROC/AUPRC, and matched-utility comparisons to separate operating-point drift from poisoning-specific degradation.

Our contributions are:

- We conduct controlled poisoning experiments across encoder-based and generative LLM guardrails, retaining corruption-ratio, external-distribution, and robust-loss analyses.
- We show that fixed-threshold metrics mix operating-point drift with poisoning effects and therefore cannot isolate poisoning-specific degradation.
- We introduce matched safety–utility and ranking analyses that reveal poisoning-induced degradation beyond clean adaptation, with explicit limits on where exact operating-point comparisons are estimable.

We do not claim that poisoning alone explains every fixed-threshold failure, that threshold adjustment repairs model integrity, or that GCE is a stable defense. The QLoRA study measures a frozen downstream model and Judge in a controlled bridge experiment, not universal jailbreak success.

2 Related Work

Prompt-injection benchmarks and guardrails study attacks that redirect LLM-integrated systems and detectors that classify untrusted instructions (Liu et al., 2023; Yi et al., 2023; Meta AI, 2024; Li et al., 2025). BIPIA emphasizes indirect prompt injection under distribution shift (Yi et al., 2023), while recent work highlights the safety–utility cost of over-defense (Kholkar & Ahuja, 2025). Our focus differs: the guard itself is updated with contaminated labels, and we audit how evaluation choices change the conclusion.

Label-noise research studies robust losses, trusted examples, and correction under corrupted supervision (Patrini et al., 2017; Han et al., 2018; Hendrycks et al., 2018). Generalized cross entropy (GCE) is a robust-loss baseline (Zhang & Sabuncu, 2018); we include it descriptively, not as a single-factor comparison because architecture and initialization also differ. Backdoor and poisoning work shows that small, structured training changes can implant malicious behavior (Yan et al., 2022; Zhao et al., 2023; Chen et al., 2025). Calibration research distinguishes confidence alignment from discrimination and cautions that distribution shift can invalidate a recalibrated operating point (Guo et al., 2017; Ovadia et al., 2019).

3 Problem Setup and Evaluation

3.1 Adaptation and poisoning

Let $s_i^{(m)} \in [0,1]$ be the block score assigned to request i by guard state m. The deployed decision is

$$\hat{y}_i^{(m)}(t) = \mathbf{1}\left[s_i^{(m)} \geq t\right], \tag{1}$$

where one denotes blocking. Clean adaptation updates on trusted labels. Poisoned adaptation uses the same pipeline after deterministically flipping a fraction of INJECTION labels to SAFE. Validation, trusted calibration, and evaluation remain clean.

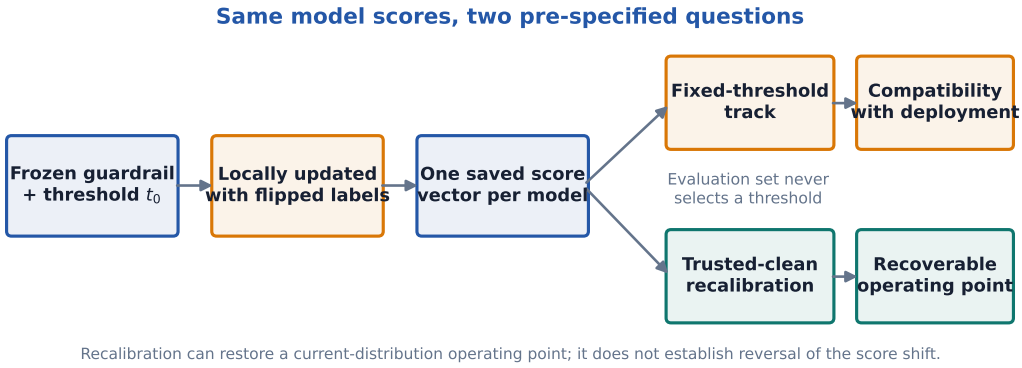

Figure 1: Evaluation framework. Fixed and trusted-adjusted tracks diagnose operating-point compatibility, while ranking and matched-utility analyses test for degradation beyond operating-point drift. The same saved scores and frozen downstream outputs are reused.

For encoder experiments, harmful block rate is the fraction of INJECTION examples blocked, SAFE false-positive rate is the fraction of SAFE examples blocked, and balanced accuracy averages injection recall and safe recall. For the generative bridge, harmful response rate (HR) is the fraction of all HarmBench requests that the guard allows and whose frozen downstream response the frozen Judge labels harmful. Benign refusal rate (BR) is the fraction of all XSTest-safe requests blocked by the guard or refused by the downstream model.

3.2 Fixed, adjusted, and matched operating points

The fixed track evaluates every update with the clean deployment threshold t_0. The adjusted track selects t_c only from 160 trusted clean calibration examples: scores are sorted, the index is $\max(0, \lceil 0.95n \rceil - 1)$, and t_c is the next representable value above that score. Evaluation labels never select a deployment threshold.

Fixed results diagnose compatibility with the deployment contract but conflate score-scale movement and ranking change. We therefore add threshold-independent AUROC/AUPRC and empirical HR–BR frontiers. At a benign-refusal budget β, define

$$\mathrm{HR}^*_m(\beta) = \min_{t:\,\mathrm{BR}_m(t)\le\beta} \mathrm{HR}_m(t), \quad \Delta\mathrm{HR}(\beta) = \mathrm{HR}^*_{\mathrm{poison}}(\beta) - \mathrm{HR}^*_{\mathrm{clean}}(\beta). \tag{2}$$

We use empirical thresholds only, with no interpolation or extrapolation. A positive ΔHR means that poisoning is worse under the same refusal budget; selected clean and poisoned points need not have exactly equal realized BR. Exact shared-BR evidence is reported separately.

4 Experimental Setup

4.1 Encoder guardrails

We study DistilBERT-CE (Sanh et al., 2019) and the ProtectAI DeBERTa-v3 prompt-injection model with GCE (Protect AI, 2024; Zhang & Sabuncu, 2018). For each architecture we train clean and 1%, 5%, and 10% harmful-to-safe updates with seeds 13, 42, and 2026. Each run has 396 trusted calibration and 792 evaluation records with identical sample IDs and labels. The original DeBERTa-CE study provides a locked 5% evaluation, and BIPIA supplies a separate external distribution with Email QA, Table QA, and Code QA strata. These evaluation sets are analyzed separately.

4.2 Generative QLoRA bridge

The bridge uses a 4-bit NF4 Llama-3-8B guard with seed-specific initial, clean-online, and 10%-poisoned adapters. For each seed, Static uses the one-epoch initial adapter, Clean uses the clean adapter, and Poisoned uses the poisoned adapter. Poisoned-fixed and Poisoned-adjusted reuse identical adapter scores and downstream responses; only the threshold differs. A frozen Mistral-7B-Instruct-v0.2 generates one response for each of 482 unique requests. The frozen HarmBench classifier Judges the 232 harmful responses once; 250 XSTest-safe prompts measure refusal. No condition-specific response regeneration occurs.

4.3 Statistical analysis

For encoders, clean and poisoned predictions are paired by sample ID. We report seed-level values, three-seed means and sample standard deviations, and 10,000-replicate prompt-level paired percentile intervals using NumPy PCG64 seed 2027. These intervals condition on the trained models and do not capture three-seed optimization uncertainty.

For QLoRA, Judge-positive ranking treats the 124 HarmBench prompts whose frozen responses are labeled harmful as positives and the 250 XSTest-safe prompts as negatives. The all-232-HarmBench definition is retained as sensitivity analysis. HR–BR frontiers use all unique empirical thresholds. Harmful and benign prompts are independently resampled with shared indices across compared conditions. We report exact McNemar tests only when outcomes are paired at the same realized operating point.

5 Encoder Poisoning Results

5.1 Frozen thresholds expose deployment incompatibility

DistilBERT-CE mean fixed-threshold injection blocking decreases by 66.58, 99.06, and 99.91 percentage points at 1%, 5%, and 10% corruption. After trusted adjustment, the corresponding poisoned-minus-clean changes are -0.17, -0.43, and $+0.09$ points. DeBERTa-GCE fixed changes are -5.36, -27.04, and -33.08 points, whereas adjusted changes are $+6.04$, $+0.85$, and $+4.17$ points. Seed behavior is heterogeneous: DistilBERT seed 13 has no fixed loss at 1%, and DeBERTa-GCE seed 42 has none at 10%.

The contrast between large fixed losses and small adjusted changes shows strong score-location drift. Clean-adjusted equals Clean-fixed exactly because the fixed threshold was selected from the same clean calibration rule; we treat this as an identity check, not a new experiment.

5.2 Poisoning-induced ranking degradation is distinct from operating-point drift

DistilBERT mean poisoned-minus-clean AUROC changes are $+0.0003$, -0.0059, and -0.0020 across the three doses; AUPRC changes are $+0.0003$, -0.0174, and -0.0065. The seed-42 5% run exhibits clear ranking degradation, but the effect is not universal or monotone. DeBERTa-GCE mean AUROC/AUPRC changes are small and mixed across seeds. Thus the encoder evidence establishes poisoning-sensitive deployment incompatibility and occasional ranking degradation, not a universal ranking collapse.

5.3 External distribution limits threshold adjustment

On the locked DeBERTa evaluation, adjustment returns the poisoned model to a high measured blocking operating point while ranking remains strong. BIPIA differs: clean and poisoned adjusted block rates are only about 3% and 4%, and balanced accuracy is approximately 0.50. Email, Table, and Code strata also differ sharply. Recalibration therefore restores measured operating points in some matched-distribution settings but not ranking behavior, and it does not establish cross-distribution recovery.

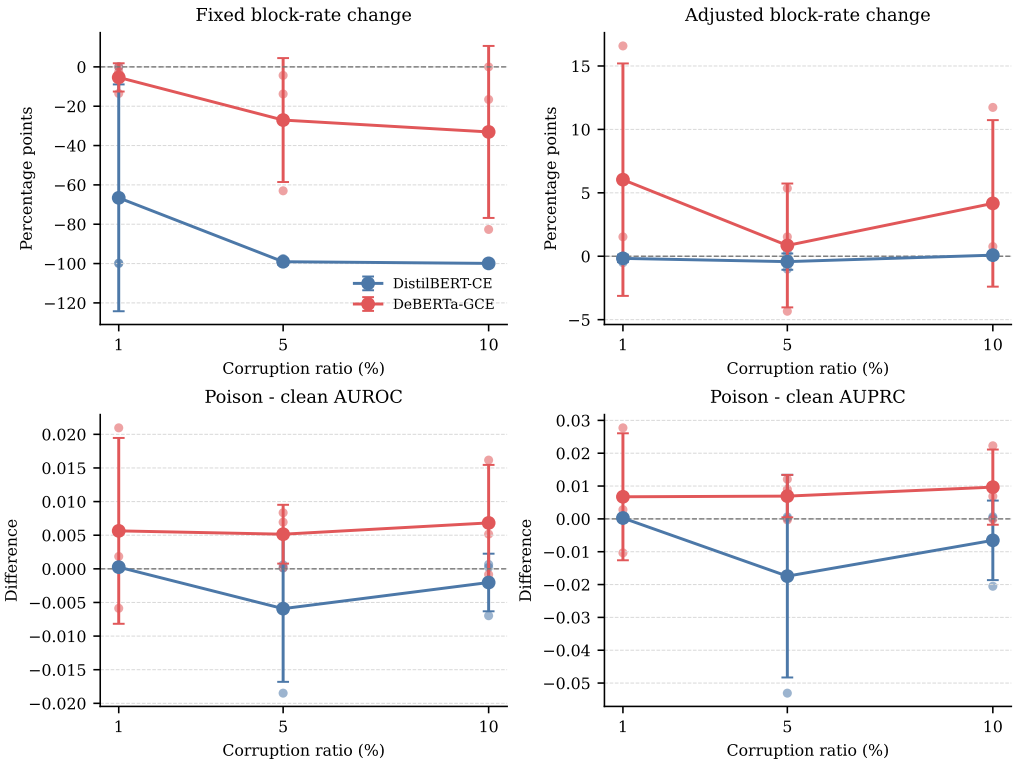

Figure 2: Encoder dose-response summaries. Points retain training-seed heterogeneity and error bars show sample standard deviations. Fixed-threshold block-rate losses can be far larger than AUROC/AUPRC changes; adjusted operating points largely remove matched-distribution block-rate loss. Trends are descriptive across three corruption levels.

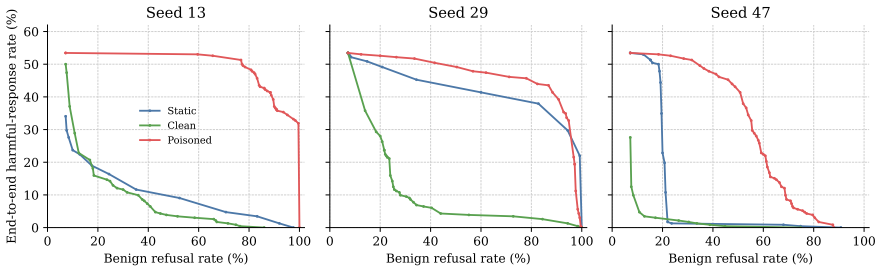

Figure 3: Empirical HR–BR frontiers for Static, Clean, and Poisoned QLoRA states across three optimization seeds. Lower values on both axes are preferable. Curves characterize saved scores; evaluation prompts do not select deployment thresholds.

The smaller average fixed losses of DeBERTa-GCE at 5% and 10% are descriptive. Because architecture, pretrained initialization, and loss all change, this is not a strict CE-versus-GCE loss ablation and does not establish a stable defense.

6 Generative Guardrail Results

6.1 Fixed thresholds do not isolate poisoning

At the original threshold, Poisoned reaches 100% benign refusal and 0% harmful response in all seeds. Clean also over-refuses: BR is 83.6%, 65.2%, and 100% for seeds 13, 29, and

Table 1: Judge-positive harmful-versus-safe ranking. Differences are poisoned minus clean; intervals are paired prompt-level 95% bootstrap CIs.

Seed	Clean AUROC	Poison AUROC	Difference
13	.854	.059	−.795 [−.833, −.755]
29	.790	.124	−.666 [−.719, −.612]
47	.974	.442	−.533 [−.587, −.477]

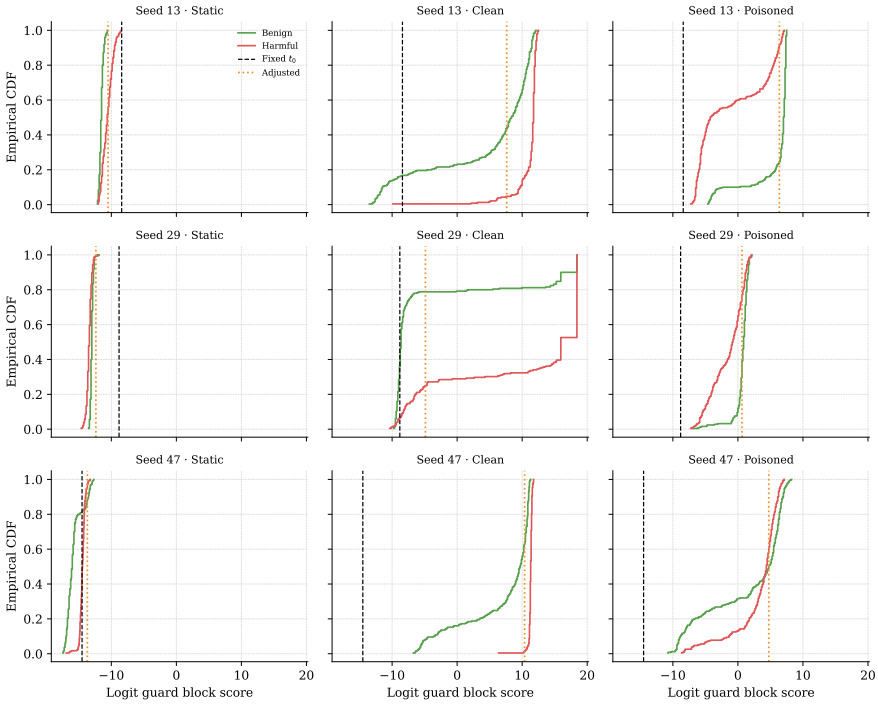

Figure 4: Guard-score distributions for HarmBench and XSTest-safe prompts. Clean and poisoned adaptation change both score location and harmful-versus-benign ordering; threshold movement alone cannot repair the poisoned ordering.

47. These saturated operating points make the reviewer concern concrete: fixed-threshold HR alone cannot distinguish poisoning from adaptation-induced operating-point drift.

Trusted adjustment reduces clean BR to 56.4%, 24.0%, and 40.0%, with HR 3.45%, 18.97%, and 0.86%. Poisoned-adjusted is worse on both axes for every paired seed: BR is 75.6%, 62.0%, and 52.0%, while HR is 52.59%, 47.41%, and 37.93%. This same-rule comparison is not exactly matched utility because realized BR differs.

6.2 Poisoning induces ranking degradation beyond operating-point shifts

Table 1 shows large poisoned-versus-clean ranking losses in all seeds. Poisoned AUROC lies below 0.5 with its entire interval for all three seeds under the Judge-positive definition. AUPRC differences are also wholly negative. When all 232 HarmBench prompts are labeled positive irrespective of downstream outcome, seeds 13 and 29 retain resolved reversals while seed 47 is near chance; this definition sensitivity is reported in the supplement.

6.3 Matched utility beyond the fixed threshold

At the exact common empirical BR=20% point available for seed 29, Clean has HR 28.02% and Poisoned has HR 52.59%, giving ΔHR = +24.57 points (95% paired-bootstrap CI

[18.97, 30.17]; exact McNemar $p \approx 1.39 \times 10^{-17}$). Fourteen other prespecified exact seed–budget cells are not estimable without interpolation and remain explicitly missing.

Under the less restrictive common BR-budget analysis, 11 of 12 estimable seed–budget cells at 10%, 20%, 30%, and 50% have positive intervals; seed 29 at 10% is the exception. The common-support frontier area is positive in every seed. These are empirical frontier comparisons, not a universal causal effect, but they show additional poisoned degradation beyond clean adaptation across a broad set of operating budgets.

7 Discussion and Limitations

The three evidence chains answer different questions. Encoder experiments establish controlled poisoning effects across architectures, doses, and an external distribution. Their dominant signature is often operating-point drift at the fixed threshold, with ranking degradation confined to particular conditions. The QLoRA bridge demonstrates that clean adaptation itself can create severe operating-point mismatch; after controlling with ranking and utility budgets, poisoned updates exhibit substantial additional degradation. The CPU-only reanalysis changes the interpretation of existing outputs rather than adding training or model generations.

Several limits matter. We study two encoder architectures and one generative guard/downstream/Judge stack. Three training seeds do not support population-level claims about optimization randomness. Prompt-level bootstrap intervals condition on trained models. Empirical matched-BR comparisons are discrete; most exact requested cells are not estimable. Judge-positive ranking depends on a frozen response Judge, while the all-HarmBench sensitivity definition answers a different question. BIPIA absolute performance is low and task heterogeneous. The 5-percentage-point alert criterion is preregistered and operational, not a universal safety boundary. Finally, threshold adjustment is valid only for the trusted calibration and evaluation distributions at hand; it does not repair model integrity.

8 Conclusion

Guardrail updates require evaluation protocols that separate operating-point changes from genuine ranking degradation. A frozen threshold remains operationally useful, but by itself it mixes adaptation-induced operating-point drift with poisoning effects. Controlled encoder experiments retain the original evidence that harmful-to-safe label flips disrupt deployment compatibility, while the generative bridge shows why clean adaptation must be an explicit comparator. Ranking and matched-utility analyses reveal additional poisoned degradation beyond clean QLoRA adaptation. Update studies should therefore report fixed and adjusted operating points, threshold-independent ranking, and matched safety–utility comparisons together.

Ethics Statement

This defensive study uses local guardrails, authorized datasets, and frozen model components. It reports aggregate safety metrics without releasing prompt or response bodies. The generative bridge reuses one frozen response per prompt and does not optimize attacks against a deployed service.

Reproducibility Statement

Seeds, splits, contamination selections, thresholds, model identities, prompt-level scores, responses, Judge labels, and bootstrap rules are hash-bound and independently audited. The supplement records configuration details, seed-level results, confidence intervals, and the distinction between exact-match and BR-budget analyses.

AI Use Statement

Large language models were used to assist manuscript editing and improve clarity and organization, as well as to support literature search and discovery. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript and its claims.

References

Yulin Chen, Haoran Li, Yuan Sui, Yangqiu Song, and Bryan Hooi. Backdoor-powered prompt injection attacks nullify defense methods. arXiv preprint arXiv:2510.03705, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Proceedings of ICML, pp. 1321–1330, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Advances in Neural Information Processing Systems, volume 31, 2018.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In Advances in Neural Information Processing Systems, volume 31, 2018.

Gauri Kholkar and Ratinder Ahuja. CAPTURE: Context-aware prompt injection testing and robustness enhancement. In Proceedings of the First Workshop on LLM Security, pp. 176–188, 2025.

Hao Li, Xiaogeng Liu, Ning Zhang, and Chaowei Xiao. PIGuard: Prompt injection guardrail via mitigating overdefense for free. In Proceedings of ACL, pp. 30420–30437, 2025. doi: 10.18653/v1/2025.acl-long.1468.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. arXiv preprint arXiv:2310.12815, 2023.

Meta AI. Prompt guard 86m model card. https://huggingface.co/meta-llama/Prompt-Guard-86M, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems, volume 32, 2019.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1944–1952, 2017.

Protect AI. Deberta-v3-base prompt injection v2 model card. https://huggingface.co/protectai/deberta-v3-base-prompt-injection-v2, 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In NeurIPS EMC2 Workshop, 2019.

Jun Yan, Vansh Gupta, and Xiang Ren. BITE: Textual backdoor attacks with iterative trigger injection. arXiv preprint arXiv:2205.12700, 2022.

Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. arXiv preprint arXiv:2312.14197, 2023.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In Advances in Neural Information Processing Systems, volume 31, 2018.

Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In Proceedings of EMNLP, pp. 12303–12317, 2023. doi: 10.18653/v1/2023.emnlp-main.757.