Security and Privacy of Machine Learning, 2025 Critique 10/01: Jailbreaking LLMs – (1) Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks (2) Deliberative Alignment: Reasoning Enables Safer Language Models

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

(1) Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks

Summary

The paper proposes **Robust Prompt Optimization (RPO)**, a minimax-inspired defense that optimizes a short, transferable *system-level suffix* to enforce refusals under worst-case jailbreaks. Instead of fine-tuning model weights, RPO performs discrete token optimization over a defensive suffix while *jointly* simulating strong attacks (e.g., GCG and PAIR). The work provides a formal objective, supporting propositions/lemmas on effectiveness and generalization, and evaluates on JailbreakBench and HarmBench. Reported results show large Attack Success Rate (ASR) reductions—including full defense against GCG across several models and strong overall robustness with minimal benign-task degradation. Conceptually, RPO brings the spirit of adversarial training in vision [1] into a *text-only, prompt-level* setting for safety.

Strengths

- **Clear defensive objective and principled design.** Casting defense as a *minimax* prompt-level optimization directly addresses adaptive attacks (mirroring adversarial-training foundations [1]).
- **Practicality: lightweight, model-agnostic suffix.** No model access or multi-call overhead; suffix is short and transferable, which is valuable for black-box and closed-source LLMs (where many defenses fail to deploy).
- **Robustness across attacks/benchmarks.** Consistent ASR reductions on *JailbreakBench* [2] and *HarmBench* [3]; especially strong against GCG [4] and competitive against PAIR.
- **Reasonable impact on benign performance.** Small drops on MT-Bench and negligible changes on MMLU suggest the suffix mostly strengthens existing refusal tendencies rather than over-suppressing helpfulness.
- **Bridges safety with prior alignment practice.** The approach complements alignment/RLHF pipelines [5] without retraining, fitting operational needs where weight updates are infeasible.

Weaknesses

- **Reliance on LLM-as-a-Judge evaluation.** Both benchmarks substantially depend on automated judges; this can entangle the defense with judge idiosyncrasies (false negatives/positives), limiting claims of real-world safety transfer.
- **Sufficiency under distribution shift.** While HarmBench results are positive, performance lifts are smaller on unseen attack families (e.g., TAP/AutoDAN). The theory bounds hinge on adversary-strength ordering, which can be brittle if novel, qualitatively different attack taxonomies emerge.
- **Scope restricted to text-only dialogs.** Multimodal prompts, tool use/agent loops, and code-execution settings are out-of-scope but are precisely where jailbreaks can be most consequential.

Potential Improvements / Extensions

- **Judge diversification and human audits.** Triangulate ASR with heterogeneous judges and sampled human audits; stress-test with adversarial judges and cross-lab replications to decouple gains from evaluation bias.
- **Suffix *distributions* and ensemble scheduling.** Optimize over a distribution of defensive suffixes (stochastic selection per turn) to reduce attack overfitting and suffix fingerprinting, akin to randomized smoothing for robustness.

- **Agent/Tool loop coverage.** Extend to multi-turn agents (tool use, code execution), enforcing refusal at handoff points and return channels; evaluate with agentic jailbreak suites beyond single-turn text.

Questions for the Authors

1) How sensitive is RPO to the *selection cadence* of adaptive attacks during training? Does more frequent inner-loop regeneration improve transfer to unseen families, or just increase cost?
2) Can the suffix be *fingerprinted* and neutralized by adaptive attackers (e.g., stripping/reframing prompts)? Any evidence for detector-evasion or suffix-honeypot dynamics?

(2) Deliberative Alignment: Reasoning Enables Safer Language Models

Summary ([6])

The paper proposes *Deliberative Alignment* (DA), a training paradigm that directly teaches models the text of safety specifications and supervises them to reason over those specifications in chain-of-thought (CoT) before answering. The method has two stages: (1) supervised fine-tuning (SFT) on synthetic (prompt, CoT, output) triples where CoTs cite relevant policy excerpts distilled into the prompt; (2) RL using a spec-aware judge model for reward on safety-relevant prompts. Evaluations show Pareto improvements versus GPT-4o on jailbreak robustness (StrongREJECT), lower over-refusals (XSTest), and better adherence to refusal/safe-completion styles; ablations indicate SFT imparts the policy-reasoning prior while RL sharpens CoT usage. The authors also report OOD generalization to encoded and multilingual jailbreaks.

Strengths

- **Direct specification grounding.** Unlike RLHF or DPO pipelines that learn safety implicitly from labels/preferences [5], DA explicitly internalizes policy text into process supervision. This addresses a persistent failure mode in alignment-by-examples and is conceptually closer to Constitutional AI yet more *procedural*, because it supervises the model's *reasoning* rather than only its outputs [7].
- **Process $\rightarrow$ outcome coherence.** The reported gains on both refusal style and safe-completion style suggest that supervising the *how* (CoT over specs) yields better *what* (final response)—an important empirical point often hypothesized but less frequently demonstrated at scale.
- **Robustness under composition.** StrongREJECT improvements imply resistance to compositional jailbreaks [8]; the encoding/multilingual OOD tests further support the claim that grounding in abstract policy rules—not pattern matching—drives generalization.
- **Compute–safety tradeoff articulation.** The paper quantifies how increasing inference-time reasoning improves difficult safety behaviors. This gives practitioners a practical knob (latency vs. robustness).
- **Rigor beyond auto-graders.** The human validation of jailbreak results mitigates grader confounding and strengthens the empirical story.

Weaknesses

- **Specification dependence and portability.** Because DA bakes in a particular (and evolving) policy, portability across organizations, jurisdictions, or dynamic norms is unclear. Retrofitting learned CoTs when policies change could cause policy drift or conflicting internalized rules.
- **Spec leakage & contamination risk.** Training on verbatim policy text raises concerns about overfitting to phrasing and potential *style over substance*: models may surface policy *keywords* in CoT without fully capturing intent (a new kind of "spec overfitting"). The paper does not quantify how often CoTs reference *irrelevant* or *spurious* policy snippets (false-positive retrieval).
- **Evaluation scope and construct validity.** The safety gains hinge on a small set of benchmarks (StrongREJECT, XSTest, internal evals). While appropriate, they do not fully probe long-horizon multi-turn attacks, tool-augmented settings, or subtle transformation-exception edge cases beyond translation [9]. Also, helpfulness regressions in complex regulated advice (vs. over-refusal) are under-explored.
- **Process honesty vs. process performance.** DA optimizes CoT for policy reasoning; however, the RL stage intentionally avoids rewarding CoT directly to reduce deceptive CoTs. Without explicit deception stress-tests, it's unclear whether models learn to *look* policy-compliant in CoT while optimizing for task success—a known concern in alignment.

Potential Improvements / Extensions

- **Policy-agnostic reasoning layers.** Factor policy retrieval into an explicit module (e.g., a compact spec encoder + retrieval head) with *contrastive* training that penalizes spurious citations; then distill into the base model. This could ease policy updates and reduce overfitting.
- **Counterfactual policy tests.** Evaluate the same model under *alternative* but plausibly coherent policy sets (e.g., stricter vs. laxer transformation exceptions) to measure robustness to policy shifts and detect hidden dependencies on particular phrasing.

Questions for the Authors

1) How often do CoTs cite *incorrect* or *non-causal* policy snippets (e.g., *post-hoc* rationalizations)? Can you report precision/recall of policy retrieval and causal ablations showing those citations *change* the answer?
2) What mechanisms and empirical results ensure that DA reduces the risk of *deceptive* CoTs rather than just optimizing surface compliance?

REFERENCES

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[2] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," *arXiv preprint arXiv:2403.12321*, 2024.

[3] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," *arXiv preprint arXiv:2402.12600*, 2024.

[4] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, 2022.

[6] M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Helyar, R. Dias, A. Vallone, H. Ren, J. Wei, H. W. Chung, S. Toyer, J. Heidecke, A. Beutel, and A. Glaese, "Deliberative alignment: Reasoning enables safer language models," *arXiv preprint arXiv:2412.16339*, 2025.

[7] Y. Bai, S. Kadavath, S. Kundu *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[8] N. Souly, Q. Lu, D. Bowen *et al.*, "A strongreject for empty jailbreaks," *arXiv preprint arXiv:2402.10260*, 2024.

[9] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "Xstest: A test suite for identifying exaggerated safety behaviours in large language models," *arXiv preprint arXiv:2308.01263*, 2024.