**Mitigation Strategies Against Jailbreak-as-a-Service (JaaS) Attacks**

This report evaluates mitigation strategies implemented in our experimental framework.

**Key Mitigation Approaches:**

I am offering multiple layers of defense against JaaS attacks that I categorize as follows:

1. **Detection-Based Mitigation**

* Implementation:
  + Hybrid Classification (Regex + ML)
    - JailbreakClassifier (rule-based) + JailbreakClassifierV2 (feature-based)
    - Covers known and novel attack patterns.
  + Active Learning for Adaptation
    - ActiveLearner flags uncertain prompts for human review.
    - Retrains models using corrected labels (retrain\_with\_feedback()).
* Effectiveness:
  + - Reduces false negatives by combining rules + learned features.
    - Adapts to evolving jailbreak techniques via feedback.

1. **Response-Based Mitigation**

* Implementation:
  + Safe Default Responses
    - If classified as jailbreak, the system logs and blocks execution.
    - Example: Returns "This request violates content policy."
  + Uncertainty Handling
    - Low-confidence prompts are quarantined for review (master\_review\_log.csv).
* Effectiveness:
  + - Prevents successful exploitation in real time.
    - Minimizes accidental blocking of benign prompts.

1. **Adversarial Robustness Testing**

* Implementation:
  + AdversarialTester perturbs prompts to test classifier resilience.
  + Evaluates bypass risks (e.g., obfuscation, hypothetical rewrites).
* Effectiveness:
  + Identifies weak points in detection rules.
  + Improves generalization against novel attack variants.

1. **Crowdsourced Attack Mitigation**

* Implementation:
  + User/IP Tracking (Proposed)
    - Logs repeated suspicious prompts from the same source.
    - Rate-limiting or temporary bans for abusive users.
* Effectiveness:
  + Counters JaaS platforms that distribute attacks across multiple users.
  + Reduces automated probing.

|  |  |  |  |
| --- | --- | --- | --- |
| Layer | Technique | Strengths | Limitations |
| Detection | Hybrid (Regex + ML) | High accuracy, adapts to new threats | Requires labeled data for training |
| Response | Blocking + Safe Responses | Prevents exploitation | May overblock if thresholds are too strict |
| Adversarial Testing | Perturbation Analysis | Improves robustness | Computationally expensive |
| Crowdsourcing Defenses | User/IP Tracking (Proposed) | Counters distributed attacks | Privacy concerns, needs deployment |