On Weaponization-Resistant Large Language Models with Prospect Theoretic Alignment

Zehua Cheng[1], Manying Zhang[2], Wei Dai[3], and Jiahao Sun[3]
[1]University of Oxford
zehua.cheng@cs.ox.ac.uk
[2]Institut National des Langues et Civilisations Orientales
manying.zhang@inalco.fr
[3]FLock.io
sun@flock.io

Abstract

Large language models (LLMs) have made significant advancements, but their increasing capabilities present serious risks of misuse, particularly in open-weight models where direct access to the model's parameters is possible. Current safeguards, designed for closed-weight API models, are inadequate for open-weight models, as minimal fine-tuning can bypass these protections. Preserving the integrity of open-weight LLMs before deployment has thus become a critical challenge. We argue that these vulnerabilities stem from the overemphasis on maximizing the LLM's log-likelihood during training, which amplifies data biases, especially with large datasets. To address these issues, we introduce Kahneman and Tversky's Prospect Theoretic Integrity Preserving Alignment (KT-IPA), a framework that prioritizes maximizing generative utility rather than a singular optimization metric. This approach strengthens LLMs against misuse and weaponization while maintaining high performance, even after extensive fine-tuning. Our results demonstrate that integrating prospect theory into LLM training enhances robustness, security, and responsible innovation in this rapidly evolving field. Our codes are available on https://anonymous.4open.science/r/KT-IPA-40B7

1 Introduction

Since their emergence, Large-Language Models (LLMs) have been the subject of extensive research aimed at enhancing their benign (Brown et al., 2020). Notably, a growing number of technology corporations have adopted an open-source strategy for LLM weights, thereby facilitating rapid progress in LLM application development. However, concerns about their potential misuse and vulnerability to tampering have also intensified, underscoring the need for robust safeguards in open-weight LLMs.

Existing safeguards for LLMs, such as refusal mechanisms and preference-based training (Liu et al., 2024), were primarily designed for closed-weight models. While effective against input-based jailbreaking attacks, these safeguards break down when adversaries can edit the model weights directly. Recent work has shown that current safeguards in open-weight models can be removed with just a few steps of fine-tuning on "uncensored" data. The safeguards are extremely brittle to these attacks that modify model weights (Qi et al., 2023). If malicious actors can easily customize models to produce harmful outputs, developers may unintentionally breach reasonable safety standards and face legal consequences. Thus, safeguarding model integrity and developer responsibility is urgently needed for both technical and social impact.

Addressing these vulnerabilities requires exploring optimization techniques beyond traditional log-likelihood maximization. While an adversarial minimax-style loss function seems intuitively a solution which maximizes benign performance and minimizes malicious output, its limitation arises from the complexity of comprehensively defining and constructing a symmetrical optimization function for specific risks or preferences, which is crucial for effective minimax-style loss training (Gazzan and Sheldon, 2023; Gokcesu and Gokcesu, 2022). Inspired by preference-based training strategies, which treat both benign and malicious information as "preferences", a more adaptable approach is to involve integrating Kahneman & Tversky's prospect theory (Tversky and Kahneman, 1992) into LLM training (Ethayarajh et al., 2024). Instead of minimax-style loss or maximizing log-likelihood, we apply Kahneman & Tversky's model of human utility into LLM training where we maximize the utility of LLM generations. Utilizing a well-defined value function, instead of a complex minimax objective, offers LLMs a more tractable framework to effectively leverage benign informa-

Proceedings of the 31st International Conference on Computational Linguistics, pages 10309–10324
January 19–24, 2025.

tion while mitigating malicious content.

Therefore, in this paper, we propose Kahneman & Tversky's Prospect Theoretic Integrity Preserving Alignment (KT-IPA) to construct the first weaponization-resistant safeguard that are robust to attack on weights modification. The inherent resilience of this system to adversarial manipulation has proven to be a significant challenging. Our experimental studies indicate that current methodologies exhibit a significant deficiency, failing to resist a 100 steps of fine-tuning attack. Addressing this challenge would yield significant benefits for both regulatory bodies and model developers. Specifically, it could mitigate the inherent dual-use dilemma associated with open-weight models, providing a mechanism to harness their potential while safeguarding against potential misuse. We first apply an initial safeguard with existing method, and then apply an integrity preserving training framework within prospect theoretic optimization.

The contributions of this paper can be summarized as below:

- The paper systematically identifies and analyzes the inherent weaknesses in current safeguards designed for open-weight LLMs. It highlights the ease with which these models can be compromised through fine-tuning, stressing the urgent need for stronger defenses.
- We introduce a novel training framework, Kahneman & Tversky's Prospect Theoretic Integrity Preserving Alignment (KT-IPA). We first formalize an adversarial minimax-style loss and then implement into prospect theoretic optimization framework to improve the resistance of weaponization knowledge extraction.
- Empirical tests show KT-IPA significantly strengthens LLM robustness. Trained models excel at standard tasks while resisting attempts to extract or misuse harmful knowledge, even robust against 10,000 adversarial fine tuning steps.

The findings of this paper have significant implications for the deployment of LLMs in practical applications. By providing a more secure and robust training approach, KT-IPA facilitates the safer deployment of open-weight models, contributing to the development of best practices in AI safety and responsible AI deployment.

2 Related Works

2.1 Adversarial attacks and defenses on LLMs

Following the challenges of adversarial attacks, researchers have explored various countermeasures to protect LLMs, which can be broadly categorized into two groups: system-level strategies that manipulate or scrutinize model inputs and outputs (Inan et al., 2023), and model-level techniques that focus on enhancing the model's internal robustness against such attacks (Mazeika et al., 2024).

System-Level Strategies System-level defenses typically involve manipulating or analyzing the inputs or outputs of an LLM to detect or prevent adversarial behavior. For example, Zeng et al. (2024) highlight how social engineering strategies, such as personalizing interactions and leveraging human-like communication patterns, can trick LLMs into performing unintended actions, like jailbreaking. This underscores the need for robust system-level defenses. Helbling et al. (2023) propose an innovative approach called "LLM Self Defense," where an LLM self-screens its generated responses to detect harmful content. This method, which operates without requiring fine-tuning, demonstrates a significant reduction in the success rate of various attacks, including prompt engineering and manipulative inputs. Similarly, Liu et al. (2023) introduce a two-pronged defense mechanism against toxic text generation: a training-free prefix prompt that preemptively filters harmful content and a RoBERTa-based model that identifies manipulative input text.

Model-Level Techniques On the model-level front, Jain et al. (2023) examine several baseline defenses, including adversarial training, which are designed to improve the inherent robustness of LLMs against adversarial prompts. Their findings suggest that while current text optimizers struggle with adaptive attacks, adversarial training can offer some level of protection. Mazeika et al. (2024) further expand on this by proposing a standardized evaluation framework, Harmbench, which rigorously tests the resilience of LLMs against red-teaming and automated attacks, demonstrating that model-level defenses are critical in fortifying LLMs against sophisticated adversarial techniques.

2.2 Kahneman-Tversky Optimization

Prospect Theory (Tversky and Kahneman, 1992) is a well-known psychological model that describes how people make decisions under risk, emphasiz-

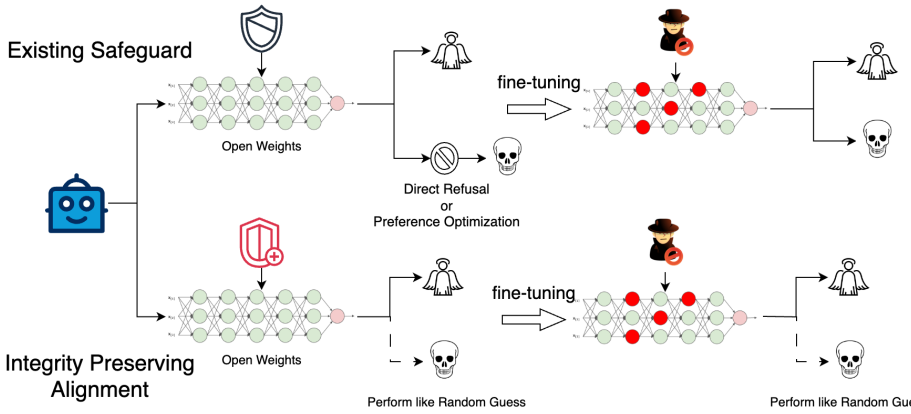

Figure 1: Overview framework of Kahneman & Tversky's Prospect Theoretic Integrity Preserving Alignment (KT-IPA) protect the model from weaponization information extraction. LLM trained with KT-IPA is lack of the capability of provide harmful content. Despite any attempts at customization by a malicious individual, the model still have no response on weaponizable content.

ing biases like loss aversion. Current LLM alignment methods such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) have implicitly modeled some of these biases, which partly explains their success. These methods have consistently proven to be more beneficial than supervised fine-tuning (SFT) alone, as they maximize the log-likelihood of human preferences.

However, the Kahneman-Tversky Optimization (KTO) model takes a different approach by directly maximizing the utility of generations (Ethayarajh et al., 2024). Prospect Theory offers a model of human utility, describing how humans make decisions under uncertainty, particularly regarding monetary outcomes. KTO leverages this by requiring only a binary signal—whether an output is desirable or undesirable for a given input. This binary feedback is easier to collect, more abundant, and cheaper, which facilitates the scaling of alignment in production environments and allows for rapid model iteration.

The simplicity and effectiveness of this thumb-up or thumb-down approach have inspired further exploration, as seen in Binary Classifier Optimization (Jung et al., 2024). This method builds on KTO's principles, explaining its effectiveness and achieving similar alignment results by optimizing a binary classifier. This approach offers a practical and scalable way to align LLMs with human preferences, demonstrating the broader applicability of KTO-inspired methodologies.

3 Problem Definition

We present a formal framework to quantify the resistance of safeguards against the weaponization within a specified threat model. Let S denote a safeguard, and θ_S represent its parameters. We introduce two metrics: $\mathcal{M}_{\text{evil}}$ for weaponizable knowledge and $\mathcal{M}_{\text{benign}}$ for benign knowledge. Consider an adversarial attack $\mathcal{A}$, which is computationally bounded and maps θ_S to a modified parameter set θ'_S.

We define the impact of an attack on the evil metric as:

$$\mathcal{M}_{\text{evil}}(\theta'_S) > \mathcal{M}_{\text{evil}}(\theta_S),$$

indicating that stronger attacks result in higher retain of evil knowledge, hence higher values of $\mathcal{M}_{\text{evil}}(\theta'_S)$. We define a safeguard S as weaponization-resistant if the following condition holds across a broad range of strong adversarial attacks $\mathcal{A}_{\textbf{test}}$:

$$\mathbb{E}_{\mathcal{A}\sim\mathcal{A}_{\textbf{test}}}\left[\mathcal{M}_{\text{evil}}\left(\theta'_S\right)\right] \leq \tau_{\text{evil}} \quad (1)$$

where τ_{evil} is a threshold indicating acceptable weaponization resistance. Ideally, τ_{evil} should be close to the performance of a model that produces outputs based on random values.

It is important to note that θ_S is typically derived from an underlying parameter set θ_0 (which lacks safeguards) through a fine-tuning procedure. Although introducing noise to θ_0 could achieve high weaponization-resistance, this would compromise the model's utility. Thus, maintaining a high benign metric, $\mathcal{M}_{\text{benign}}\ (\theta_S)$, is crucial. Formally, we

require:

$$\mathcal{M}_{\text{benign}}\ (\theta_S) \geq \tau_{\text{benign}}$$

where τ_{benign} is a threshold indicating acceptable performance in benign queries.

The overall evaluation of a safeguard must balance both its weaponization-resistance and its benign preservation, which can be expressed as an minimax optimization problem:

$$\min_{\theta_S} \max_{\mathcal{A}\sim\mathcal{A}_{\text{test}}} \left(\mathbb{E}_{\mathcal{A}}\left[\mathcal{M}_{\text{evil}}\ \left(\theta'_S\right)\right] - \lambda \cdot \mathcal{M}_{\text{benign}}\ (\theta_S)\right) \tag{2}$$

subject to:

$$\begin{aligned} \mathbb{E}_{\mathcal{A}\sim\mathcal{A}_{\text{test}}}\left[\mathcal{M}_{\text{evil}}\ \left(\theta'_S\right)\right] &\leq \tau_{\text{evil}} \\ \mathcal{M}_{\text{benign}}\ (\theta_S) &\geq \tau_{\text{benign}} \end{aligned} \tag{3}$$

where λ is a regularization parameter that controls the trade-off between minimizing the expected evil metric $\mathbb{E}_{\mathcal{A}}[\mathcal{M}_{\text{evil}}(\theta'_S)]$ and preserving the $\mathcal{M}_{\text{benign}}(\theta_S)$. In this setup, The inner maximization seeks to find the worst-case adversarial attack $\mathcal{A}$ that maximizes the evil metric $\mathcal{M}_{\text{evil}}$ for the modified parameters θ'_S. The outer minimization aims to adjust θ_S to minimize this worst-case evil metric while simultaneously ensuring that the benign metric $\mathcal{M}_{\text{benign}}$ remains above the acceptable threshold τ_{benign}. This provides the best balance between weaponization-resistance and benign performance, ensuring that the model is both secure against adversarial manipulation and functionally effective for benign queries.

It is trivial that Equation 2 is a minimax-style loss. However, minimax optimization often suffers from instability, especially when adversarial attacks are dynamic and adapt over time. Therefore, it is crucial to introduce an effective solution to optimize Equation 2. In this paper, we introduce KTO (Ethayarajh et al., 2024) to optimize the goal.

4 Methodologies

We first use existing safeguarding method to the LLM to establish initial safeguard. We discuss the details of implementation of the initial safeguard in Appendix.

4.1 Integrity Preserving Alignment

The Pseudo-code of the Integrity Preserving Alignment is presented in Algorithm 1. To establish a secure LLM that resist to weaponization knowledge extraction, the ultimate goal is to optimize

Algorithm 1 IPA: Integrity Preserving Alignment

Input: Train-time adversary set $\mathcal{A}_{\text{train}}$; Dataset D_{benign} and D_{evil}; Outer steps N, Number of sampled adversaries K

Parameters: Initial LLM parameters θ_0, learning rate η, number of sampled adversaries K, Loss scale λ_{benign} and λ_{evil}

Outputs: Final parameters θ_G

$\theta_0 \leftarrow$ Apply Initial Safeguard to θ_0
for $i = 1$ to N **do**
# Get gradient
  $g_{\text{evil}} \leftarrow 0$
Adversarial Minimax Training
Inner-loop optimize for evil sample
  Sample $x_{\text{evil}} \sim D_{\text{evil}}$
  for $k = 1$ to K **do**
    Sample attack $\sim \mathcal{A}_{\text{train}}$
    # Apply Equation 5
    $g_{\text{evil}} \leftarrow g_{\text{evil}} + \frac{1}{K}\nabla\mathcal{L}_{\text{evil}}(\mathcal{A}(\theta_{i-1}, x_{\text{evil}}))$
  end for
  Sample $x_{\text{benign}} \sim D_{\text{benign}}$
  # Apply Equation 6
  $g_{\text{benign}} \leftarrow \nabla_{\theta_{i-1}}\left(\mathcal{L}_{\text{benign}}(\theta_{i-1}, x_{\text{benign}})\right)$
  # Gather g_{benign} and g_{evil}
  $\theta_i \leftarrow \theta_{i-1} - \eta(\lambda_{\text{benign}} \cdot g_{\text{benign}} + \lambda_{\text{evil}} \cdot g_{\text{evil}})$
end for
$\theta_G \leftarrow \theta_N$
return θ_G

the Equation 2 where we minimize the $\mathcal{M}_{\text{evil}}$ while maximizing the $\mathcal{M}_{\text{benign}}$ even after adversarial tampering. We have got two datasets, $\mathcal{D}_{\text{benign}}$ and $\mathcal{D}_{\text{evil}}$ for datasets that have benign and weaponizable knowledge. (x, y) represents a pair of data points for a given dataset. We design two resistance loss functions, the evil-resistance loss $\mathcal{L}_{evil}$ designed to penalize the model when its behavior deviates in undesirable ways due to adversarial attacks, and $\mathcal{L}_{benign}$ benign-retain loss function to ensure that the model retains its original capabilities and performance after adversarial training. The original objective function in a minimax format is:

$$\min_{\theta_S} \max_{\mathcal{A}\sim\mathcal{A}_{\text{train}}} \left(\mathbb{E}_{\mathcal{A}}[\mathcal{L}_{\text{evil}}(\mathcal{A}(\theta_S); \mathcal{D}_{\text{evil}})] - \lambda_{\text{benign}} \cdot \mathcal{L}_{\text{benign}}(\theta_S; \mathcal{D}_{\text{benign}})\right) \tag{4}$$

where $\mathcal{A}$ represents adversarial attacks. We introduce additional regularization terms for $\mathcal{L}_{\text{evil}}$ to penalize the model for any significant deviation in its output distribution after adversarial perturbations, considering techniques like adversarial examples

and gradient-based attacks. Specifically, the $\mathcal{L}$evil is defined as:

$$\mathcal{L}_{\text{evil}}(\theta_S, x) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{evil}}}[CE(f_{\theta_S}(x+\delta), y) + \alpha||\Delta_{\theta_S} f_{\theta_S}(x+\delta)||^2] \tag{5}$$

where CE is the cross-entropy loss. δ represents the adversarial perturbation. α is a regularization parameter controlling the strength of the gradient penalty.

For $\mathcal{L}_{\text{benign}}$, we incorporate consistency regularization term that forces the model's output to remain close to its original predictions on benign data. The $\mathcal{L}_{\text{benign}}$ is defined as:

$$\mathcal{L}_{\text{benign}}(\theta_S, x) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{benign}}}[CE(f_{\theta_S}(x+\delta), y) + \beta \cdot MSE(f_{\theta_S}(x), f_{orig}(x))] \tag{6}$$

β is a hyperparameter balancing retention and prediction consistency. MSE is the mean squared error between the original model output $f_{\text{orig}}(x)$ and the current model output $f_{\theta_S}(x)$.

4.2 Prospect Theoretic Integrity Preserving Alignment

Instead of a minimax structure in Equation 2, the prospect theoretic self alignment utilize the framework of KTO (Ethayarajh et al., 2024) where we optimize the model's parameters to maximize a utility function that is informed by human-like biases. The KTO framework can be used to model both the "good" (benign) and "bad" (malicious or weaponizable) outcomes. The new objective function can be expressed as:

$$\min_{\theta_S} \mathbb{E}_{x,y\sim D}[\lambda_y \cdot v(x,y)] \tag{7}$$

where λ_y is a weight associated with whether the outcome is desirable or undesirable. $v(x,y)$ is the value function derived from prospect theory that reflects the utility of a particular outcome, taking into account human biases such as risk aversion and loss aversion.

We now define the value functions for desirable (benign) and undesirable (evil) outcomes:

$$v(x,y) = \begin{cases} \lambda_{\text{benign}} \cdot \sigma(\beta \cdot (\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} - z_0)) \text{ if } y \sim y_{\text{benign}} \mid x \\ \lambda_{\text{evil}} \cdot \sigma\left(\beta \cdot \left(z_0 - \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right) \text{ if } y \sim y_{\text{evil}} \mid x \end{cases} \tag{8}$$

where $\pi_\theta(y|x)$ is the current model's output distribution given input x while $\pi_{\text{ref}}(y|x)$ is the reference model's output distribution given input. In our framework, the π_{ref} is the initial safeguard model π_{θ_S}. A sigmoid function $\sigma(\cdot)$ ensures that the value function is bounded and behaves smoothly. λ_{benign} is a weighting factor for benign outcomes and λ_{evil} is for evil or un-desirable outcomes. A β parameter controls the curvature of the value function, reflecting risk aversion (a lower β makes the model more risk-averse).

Besides, z_0 is a reference point, which could be the expected Kullback–Leibler (KL) divergence for benign outputs. It ensures that the model is only penalized for deviations that exceed normal or expected variance. In practice, estimating z_0 involves calculating the expected KL divergence across a batch of inputs:

$$z_0 = \max\left(0, \mathbb{E}_{x'\sim D_{\text{benign}}}\left[\text{KL}\left(\pi_\theta\left(y \mid x\right) \| \pi_{\theta_S}\left(y \mid x\right)\right)\right]\right) \tag{9}$$

where $KL(\cdot||\cdot)$ is the Kullback–Leibler divergence. z_0 serves as a dynamic baseline that adapts based on the model's performance during training. It reflects the expected behavior of the model in the absence of adversarial perturbations or harmful influences. Therefore, for benign information, z_0 helps to encourage the model to improve upon or at least match the expected performance and for evil information, z_0 is used to identify when the model's behavior starts to deviate from acceptable norms, allowing the value function to penalize such deviations effectively.

5 Experiments

We perform KT-IPA over Llama-3-8B-Instruct (Dubey et al., 2024) using a distributed training setup across eight NVIDIA 80GB A100 GPUs. We leverage the Fully Sharded Data Parallel (FSDP) framework for parallel computation (Rajbhandari et al., 2020). Furthermore, DeepSpeed's ZeRO Stage 3 (Ren et al., 2021) is employed to shard optimizer states, gradients, and model parameters during the training process. We set the same $\lambda_{\text{benign}} = \lambda_{\text{evil}} = 1$ as the (Ethayarajh et al., 2024).

5.1 Datasets

The performance of KT-IPA was assessed using the Weapons of Mass Destruction Proxy (WMDP) benchmark (Li et al., 2024). This dataset consists

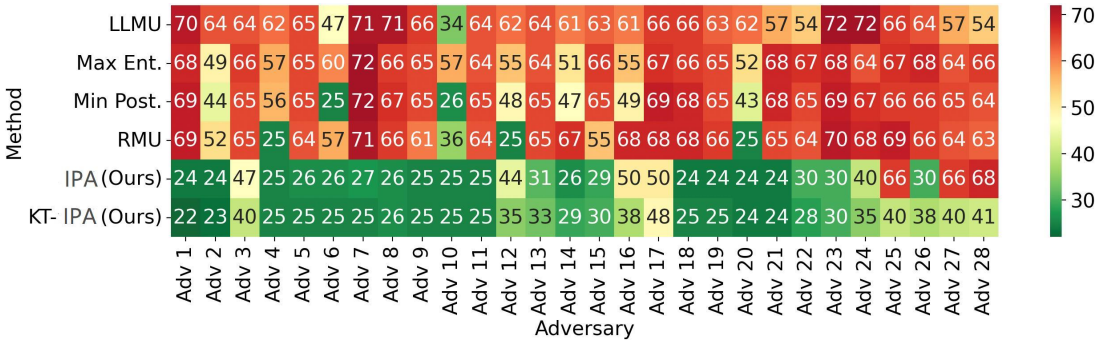

Figure 2: Simulated assessment of post-attack accuracies for biosecurity measures accross 28 adversarial strategies. Defense efficacy is measured as a percentage, with a benchmark score of 25% representing optimal performance. Attacks resulting in performance levels close to the 'No Defense' baseline are denoted in red. Assessments for the other domains are presented in the Appendix.

of 3,668 multiple-choice questions across the domains of biosecurity, chemical security, and cybersecurity. Notably, WMDP queries do not directly assess hazardous knowledge but rather evaluate expert-level understanding within each domain. Consequently, restricting access to expert-level knowledge through the proposed methods would inherently limit the acquisition of hazardous knowledge. For this evaluation, the "evil set" was defined as the specific hazardous knowledge subjects represented in WMDP. The "benign set" comprised the complement of these subjects, sourced from MMLU (Hendrycks et al., 2020), a widely recognized multi-task question-answering benchmark with 57 diverse tasks across various knowledge domains.

5.2 Evaluation Metrics

We define accuracy metrics for both the "evil set" and the "benign set": "evil accuracy" measures the model's retention of restricted weaponization knowledge, while "benign accuracy" assesses performance on benign knowledge domains. Lower evil accuracy indicates better suppression of restricted knowledge, while higher benign accuracy suggests that the model has preserved its performance on non-weaponization knowledge. Ideally, we want the model to achieve both low evil accuracy and high benign accuracy.

To further evaluate the robustness of proposed method, we applied a comprehensive adversarial evaluation methodology. We followed the setup of (Tamirisa et al., 2024) which involved engaging a diverse pool of up to 28 adversaries, encompassing numerous novel attack strategies unseen during the training phase. Additionally, multiple baseline models, including LLMU (Yao et al., 2023) and RMU (Li et al., 2024), were used for comparison. Detailed information on these baseline models can be found in Appendix.

Evaluations were conducted by systematically exposing the safeguard to adversaries with varying computational resources, access to withheld datasets, and a range of hyperparameter configurations. The fine-tuning of adversarial models incorporated manipulations of learning rate, learning rate schedulers, optimization algorithms, and batch size. Notably, several adversaries were iteratively introduced throughout the development process, responding to discovered vulnerabilities in intermediate safeguard versions.

This extensive stress testing regime is indispensable for establishing confidence in the robustness of weaponization-resistant safeguards. Furthermore, comprehensive red teaming serves as a valuable tool for quantifying incremental progress in safeguard development. The efficacy of these safeguards can be measured by the number and sophistication of successfully defended attacks, providing a robust metric for assessing resilience improvements.

5.3 Main Results

In our evaluation, we compared the performance of various defense mechanisms against adversarial attacks using three key metrics: pre-attacks benign, pre-attacks evil, and post-attacks evil. The random strategy, set with a fixed probability, consistently achieved scores around 25% across all metrics. This baseline highlights the effectiveness of the random approach in establishing a controlled reference point for comparison.

When examining benign accuracy, which reflects the model's ability to preserve benign knowledge,

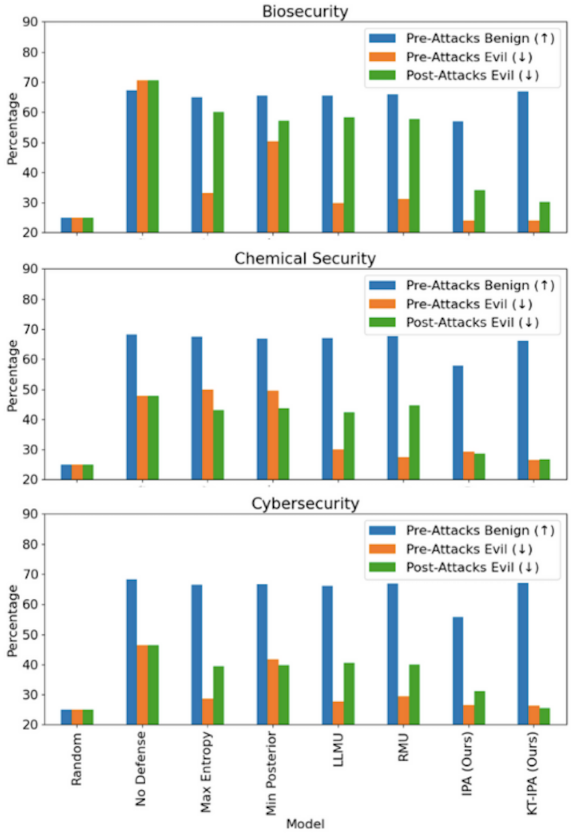

Figure 3: Pre-attack and average post-attack accuracy scores for various models across three distinct domains: WMDP Biosecurity, Chemical Security, and Cybersecurity. The evaluation framework encompasses the KT-IPA alongside several established baselines. Accuracy metrics are reported for Llama-3-8B following 28 fine-tuning attacks detailed in Appendix. The "average Post-Attack accuracy" is calculated as the mean accuracy across all adversarial fine-tuning scenarios. All reported values represent percentages.

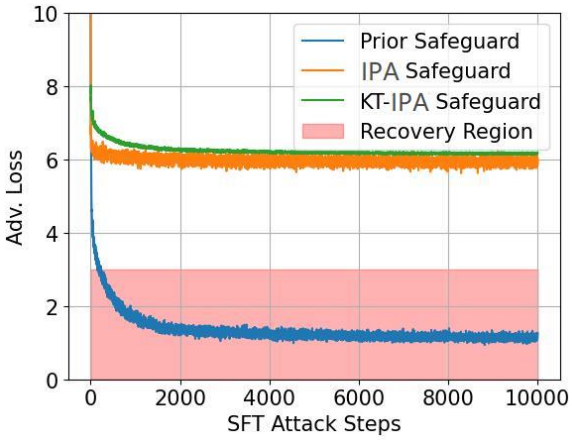

Figure 4: Fine-tuning Attack Progression Over 10,000 Steps. This figure shows the progression of a fine-tuning attack aimed at recovering malicious information from the trained models. The recovery region indicates the extent to which the attack succeeds in accessing weaponized knowledge, reflecting the model's resilience. The KT-IPA model demonstrates strong resistance, with cross-entropy loss stabilizing above a threshold of 6 even after 10,000 steps, while the LLMU model (Prior Safeguard) shows significant loss reduction and faster information recovery. Unlike previous studies on adversarial training paradigms, which typically explored up to 100 steps, our analysis extends to 10,000 steps, highlighting the enhanced robustness of KT-IPA.

we observed that most defense methods, including "no defense," achieved similar high scores of approximately 67-68%. Notably, our IPA framework, though slightly less effective, demonstrated a marginally lower benign score compared to the KTO-enhanced IPA framework (KT-IPA). This result aligns with expectations, as "no defense" naturally serves as a baseline, reflecting the maximum potential for knowledge preservation, no matter benign or weaponized.

The more pronounced difference emerged in the evil accuracy, indicating the model's retention of harmful knowledge. The "no defense" approach, which lacks mechanisms to mitigate harmful knowledge, exhibited notably high evil accuracy, especially in the biosecurity domain, signaling poor performance in discarding malicious content. In contrast, our model outperformed others in evil accuracy results, both pre- and post-attacks, demonstrating superior sensitivity to harmful knowledge and weaponization.

A comparative analysis of pre-attacks and post-attacks scores revealed that RMU and LLMU models showed better performance in pre-attacks evil than in post-attacks evil. This suggests that while these models can effectively handle harmful knowledge present in the training data, they struggle with new, unseen harmful content, indicating a lack of robustness against novel attacks. Conversely, our model maintained consistent performance across both pre- and post-attack scenarios, illustrating its stability and robustness in handling both familiar and novel harmful knowledge.

Overall, our results underscore the effectiveness of our IPA and KT-IPA frameworks in maintaining low evil retention while achieving competitive benign accuracy scores. This indicates a robust defense mechanism that remains effective against a range of adversarial attacks, ensuring the preservation of non-harmful knowledge and the effective discarding of harmful content.

Method	w/PT	Pre Attack		Post-Attack (Avg)
		Benign (↑)	Evil (↓)	Evil (↓)
No Defense		67.3	70.5	70.5
Excl. MSE term in $\mathcal{L}_{\text{benign}}$		49.5	27.5	40.5
	✓	51.5	27.5	33.5
Excl. Adv. Training		59.5	27.8	60.5
	✓	-	-	-
Excl. Init. Safeguard		61.5	48.5	47.5
	✓	62.0	40.5	39.5
IPA		56.9	**24.0**	31.5
	✓	**66.9**	**24.0**	**31.3**

Table 1: Ablations of the primary components of KT-IPA include: (1) the MSE term in $\mathcal{L}_{\text{benign}}$; (2) the adversarial training phase; and (3) the initial model safeguard phase, each assessed with and without the integration of prospect theory optimization (PT), except for the adversarial training ablation.

5.4 Ablation Studies

Our ablation study investigates the impact of removing specific components from the KT-IPA framework, including (1) the MSE term in the $\mathcal{L}_{\text{benign}}$; (2) the adversarial training phrase; (3) the initial model safeguard phase; each examined with and without the integration of Prospect Theory Optimization (PT), except in the adversarial training ablation. For benign knowledge, we aim for higher accuracy, while for evil knowledge, a lower accuracy score is desirable. We observed the following results:

1. Removing the MSE term decreased accuracy for both benign and harmful knowledge. While it reduced harmful knowledge retention, it significantly harmed benign knowledge preservation, showing that this approach disrupts the balance between safeguarding and mitigating knowledge.

2. Omitting the initial safeguard phase resulted in increased accuracy scores for both benign and evil knowledge. This indicates that while more benign knowledge was retained, the model lacked adequate defense against evil knowledge. Consequently, this ablation showed that the initial safeguard phase is crucial for providing balanced protection across both types of knowledge.

3. When integrating PT, all models (except for "no defense" and "no adversarial training" which inherently lack PT) showed improved performance compared to those without PT. This confirms the necessity of prospect theory optimization in enhancing model performance and robustness.

4. Excluding adversarial training significantly deteriorated resistance to post-training attacks. This finding highlights that without adversarial training, the model's ability to discern between good and evil knowledge was compromised, making it less effective in handling novel attack scenarios beyond those seen during training.

6 Conclusion

This study presents a novel methodology for integrating weaponization-resistant safeguards into Large Language Models (LLMs). We first define the objective and introduce Kahneman & Tversky's prospect theory into a LLM security domain. We argue that under Kahneman & Tversky optimization framework, we can secure the LLM from weaponization and keep the capacity of benign feedbacks. Our findings, derived from rigorous red-teaming evaluations and benchmarkings, demonstrate that this method outperforms existing approaches by achieving robustness against adversarial manipulation. This establishes it as the first demonstrably resilient technique under such stringent testing protocols. Furthermore, our research illustrates the feasibility of achieving robustness result for open-weight LLMs. It directly contributes to alignment with evolving regulatory frameworks and proactively mitigates the potential for malicious exploitation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nanqing Dong, Jiahao Sun, Zhipeng Wang, Shuoying Zhang, and Shuhao Zheng. 2022. Flock: Defending malicious behaviors in federated learning with blockchain. *arXiv preprint arXiv:2211.04344*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Mazen Gazzan and Frederick T. Sheldon. 2023. An enhanced minimax loss function technique in generative adversarial network for ransomware behavior prediction. *Future Internet*, 15:318.

Kaan Gokcesu and Hakan Gokcesu. 2022. Efficient minimax optimal global optimization of lipschitz continuous multivariate functions. *arXiv preprint arXiv: 2206.02383*.

Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. 2023. Llm self defense: By self examination, llms know they are being tricked. *ArXiv*, abs/2308.07308.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. *ArXiv*, abs/2309.00614.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. 2024. Binary classifier optimization for large language model alignment. *ArXiv*, abs/2404.04656.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Bowen Liu, Boao Xiao, Xutong Jiang, Siyuan Cen, Xin He, and Wanchun Dou. 2023. Adversarial attacks on large language model-based system and mitigating strategies: A case study on chatgpt. *Security and Communication Networks*.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. 2024. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. {Zero-offload}: Democratizing {billion-scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. 2024. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*.

Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323.

Zezhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru Wang, Liang Chen, Qingwei Lin, and Kam-Fai Wong. 2023. Self-guard: Empower the llm to safeguard itself. *arXiv preprint arXiv:2310.15851*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *ArXiv*, abs/2310.10683.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *ArXiv*, abs/2401.06373.

A Initial Weaponization Safeguard

Prior apply our Integrity Preserving Alignment (IPA) and Kahneman & Tversky's Prospect Theoretic Integrity Preserving Alignment (KT-IPA), we train an initial weaponization safeguard that achieves basic weaponization safeguard on the target domain. Let $h_\theta(\mathcal{D})$ denote the distribution of post-decoder layer residual stream activations for input sequences sampled from some data distribution $\mathcal{D}$ over θ model weights. We followed (Wang et al., 2023) that applied random hash over the input during training.

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{D}_{\text{evil}}} \left[1 - \left| \frac{h_\theta(x) \cdot \text{rand\_hashed}(x)}{\|h_\theta(x)\| \, \|\text{rand\_hashed}(x)\|} \right| \right] + CE\left(\theta; \mathcal{D}_{\text{benign}}\right) \tag{10}$$

The objective function articulated in Equation 10 which is aim to maximize the cosine similarity between row vectors within the residual of each layer in the LLM generated by the input $h(\mathcal{D}_{\text{evil}})$ and the hashed random vectors produced by rand_hashed$(\cdot)$. By assigning a unique random vector as a target for each token's residual stream, this loss term encourages a "re-mapping" of token representations derived from $\mathcal{D}_{\text{evil}}$ towards these noised vectors. An additional cross entropy (CE) term is incorporated to mitigate performance degradation on the benign dataset $\mathcal{D}_{\text{benign}}$.

B Experiment Details in Figure 3

Our experiment compares several baseline models with the ours on pre-attack and average post-attack accuracy scores:

- **Random**: This baseline simulates a model making predictions randomly. In multiple-choice problems with four options, the expected accuracy is 25%.
- **No defense**: This setting represents a typical approach without any defensive mechanisms against adversarial attacks.
- **Max entropy**: This model selects the option with the highest entropy, reflecting the highest uncertainty in predictions.
- **Min Posterior**: This model chooses the option with the lowest posterior probability, under the assumption that less confident predictions are less likely to be correct.
- **LLMU (Large Language Model Unlearning)**: This model employs unlearning to remove undesirable misbehaviors from LLMs. It focuses on removing harmful responses, erasing copyrighted content, and reducing hallucinations. It uses negative examples for alignment (Yao et al., 2023).
- **RMU (Representation Misdirection for Unlearning)**: RMU is a state-of-the-art unlearning method based on controlling model representation proposed by WMDP benchmark (Weapons of Mass Destruction Prevention).

We set the α in Equation 5 as 1.0 and β to be 1.0 throughout all experiments setup. For the β in Equation 8, we followed the same setup in KTO (Ethayarajh et al., 2024) which is 0.1.

C Red Teaming Setup

C.1 Red Teaming Attacks Results

To ensure that our models are robust against post-training tampering attacks, we rigorously evaluate their resilience using a comprehensive set of supervised fine-tuning attacks. We deploy diverse different adversarial strategies, systematically manipulating key parameters such as the optimizer, training steps, learning rate, and dataset. We also explore various fine-tuning techniques, including full fine-tuning and parameter-efficient methods, to comprehensively assess vulnerabilities across diverse attack scenarios. If not specified, all attacks are conducted with 2,000 fine-tuning steps. We present the full details of adversaries in Appendix.

As shown in the Figure 5, existing baseline safeguard mechanisms exhibit vulnerability to fine-tuning attacks, effectively thwarting manipulation only in a limited subset of adversarial scenarios. In contrast, our proposed IPA and KT-IPA safeguards demonstrate robustness against a broader spectrum of adversaries. Comparing IPA to KT-IPA, we noticed that for many adversary attacks, for example the Adv. 12 and 16, KT-IPA has over 10% improvement over IPA performance. In Adv. 25, it's 26% improvement. This demonstrated the significant of introducing prospect theory into the IPA framework.

It is also worth noting that IPA proves susceptible to parameter-efficient fine-tuning (PEFT) attacks (Adv. 27 and 28), emphasizing the critical need for comprehensive adversarial testing during the design and deployment of PEFT attack defenses. By incorporating prospect theory into IPA, the model, although still vulnerable compared to baselines, shows a significant improvement in robustness against PEFT attacks.

C.2 Resistance to 10,000 steps of attacks

In Figure 4, we show a fine-tuning attack at a learning rate of 2×10^{-5} targeting the KT-IPA model and its counterpart LLMU. The goal of this attack is to recover information related to chemical security weaponization techniques from the trained models. Our findings demonstrate that the weaponization resistance exhibited by KT-IPA extends significantly beyond the 100 steps typically employed by adversarial training paradigms. Notably, the test-time adversary's cross-entropy loss stagnates above a threshold of 6 for over $10,000$ iterations, exhibiting a plateau after 200 steps where further reduction is absent. For comparative analysis, the attack progression on LLMU is presented. In this instance, the adversary achieves a loss value within the information recovery region in under 100 steps.

These findings strongly suggest that KT-IPA demonstrates superior resistance to fine-tuning attacks compared to LLMU, particularly in preventing the recovery of sensitive knowledge such as chemical security information. KT-IPA's resilience persists even after extensive adversarial training (over 10,000 steps), highlighting its effectiveness in safeguarding critical information. The results also indicate that KT-IPA maintains better stability, with less fluctuation in loss values, reinforcing its robustness against such attacks.

C.3 Finetuning Dataset Construction

We first utilize Pile (Gao et al., 2020) that filters for relevance to biology and the Camel AI Biology dataset (Li et al., 2023). We generate synthetic labels for Pile token sequences using FLock (Dong et al., 2022), categorizing them as "Cellular Biology" or not. We then mannully select whether the content is weaponizable or not to determine the **Evil-Pile** (the filtered Pile marked with weaponizable infomation) with 7,668 samples and **Benign** as the left $43,515$ samples. Following a consistent methodological approach, we extended our analysis to a held-out dataset comprising hazardous biological sequences, mirroring the composition and structure of the WMDP biosecurity evil set. This resulted in a collection of 598,933 samples exhibiting comparable length characteristics, designated as the **OOD-Evil** set.

For Chemical Security, we use a private dataset containing textual sequences pertaining to hazardous chemical security marked as **Evil-Chem**. We share the same **Benign** set for Chemical Security and Biosecurity Weaponization tasks.

For Cybersecurity, we used the Cybersecuriy dataset in TAR (Tamirisa et al., 2024) paper where is latest collected cybersecurity writeups in 2024 with carefully filtered by the authors resulting forget dataset with 18,000 samples which we marked it as **Evil-Cyber**.

For all three datasets, we construct a combination of benign and evil adversarial strategies, where a model is fine-tuned on benign data after being exposed to evil data which we marked it as **Benign→Evil**.

C.4 Red Teaming Experimental Details

This study investigates post-attack forget accuracy following fine-tuning for both benign and evil knowledge domains. We list all test-time adversary setups for recovering Biosecurity, Chemical Security in Table 2. We use the the similar constructed datasets from (Tamirisa et al., 2024) where For Biosecurity, we evaluate performance after 1,000 optimization steps on three distinct data distributions: the Pile-bio Forget set, the Benign-set employed in Random Mapping and IPA, and an out-of-distribution (OOD)

Forget set introduced in and unseen during training. To construct our strongest adversary, we further fine-tune on this OOD distribution for 5,000 steps.

Consistent experimental paradigms are adopted for Chemical Security and Cybersecurity benchmarks, substituting the Pile-bio Forget set with respective Chem and Cyber Forget sets. Due to limitations in obtaining comparable held-out datasets for these domains, unseen forget distribution adversaries were excluded, resulting in a total of 24 attack scenarios analyzed. Across all subjects, we evaluate the impact of varying optimizers (AdamW, Adadelta, and Schedule Free AdamW), learning rates ($2 \times 10^{-6}, 2 \times 10^{-5}, 1 \times 10^{-4}$ and learning rate scheduling methods (SGDR and a 30-step linear warmup).

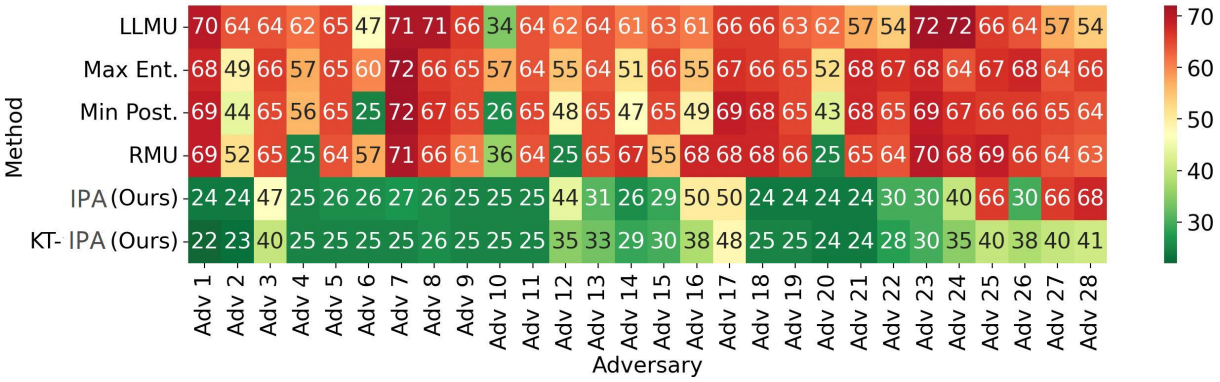

Figure 5: Simulated assessment of post-attack accuracies of biosecurity measures against weaponizable biological knowledge. Defense efficacy is measured as a percentage, with a benchmark score of 25% representing optimal performance. Attacks achieving performance levels close to the 'No Defense' baseline are denoted in red. Comparison of adversarial robustness across various methods, where lower values indicate better performance. The visualization demonstrates that the proposed methods, IPA and KT-IPA, particularly KT-IPA which incorporates the KTO (Ethayarajh et al., 2024), exhibit promising results against a series of **28** adversarial scenarios.

Adversary	Dataset	Steps	Optimizer	LR	LR Schedule	Purpose
Adv 1	OOD-Evil	5000	AdamW	2×10^{-5}	No Warmup	
Adv 2	OOD-Evil	5000	AdamW	1×10^{-4}	No Warmup	Adv. 1 $\rightarrow$ larger LR
Adv 3	Evil-Pile	1000	AdamW	2×10^{-5}	No Warmup	
Adv 4	Evil-Pile	1000	AdamW	1×10^{-4}	No Warmup	Adv. 3 $\rightarrow$ larger LR
Adv 5	Benign	1000	AdamW	2×10^{-5}	No Warmup	
Adv 6	Benign	1000	AdamW	1×10^{-4}	No Warmup	Adv. 5 $\rightarrow$ larger LR
Adv 7	OOD-F	1000	AdamW	2×10^{-5}	No Warmup	
Adv 8	OOD-F	1000	AdamW	1×10^{-4}	No Warmup	Adv. 7 $\rightarrow$ larger LR
Adv 9	Benign $\rightarrow$ Evil	1000	AdamW	2×10^{-5}	No Warmup	
Adv 10	Benign $\rightarrow$ Evil	1000	AdamW	1×10^{-4}	No Warmup	Adv. 9 $\rightarrow$ larger LR
Adv 11	Evil-Pile	1000	Adadelta	2×10^{-5}	No Warmup	
Adv 12	Evil-Pile	1000	Adadelta	1×10^{-4}	No Warmup	Adv. 11 $\rightarrow$ larger LR
Adv 13	Evil-Pile	1000	Schedule Free	2×10^{-5}	No Warmup	
Adv 14	Evil-Pile	1000	Schedule Free	1×10^{-4}	No Warmup	Adv. 13 $\rightarrow$ larger LR
Adv 15	Evil-Pile	1000	SGD Nesterov	2×10^{-5}	No Warmup	
Adv 16	Evil-Pile	1000	SGD Nesterov	1×10^{-4}	No Warmup	Adv. 15 $\rightarrow$ larger LR
Adv 17	Evil-Pile	1000	AdamW	2×10^{-6}	No Warmup	Small LR setup
Adv 18	Evil-Pile	1000	AdamW	2×10^{-6}	30 Steps Warmup	Adv. 17 $\rightarrow$ warmup
Adv 19	Evil-Pile	1000	AdamW	2×10^{-5}	30 Steps Warmup	
Adv 20	Evil-Pile	1000	AdamW	1×10^{-4}	30 Steps Warmup	Adv. 19 $\rightarrow$ larger LR
Adv 21	Evil-Pile	1000	AdamW	2×10^{-5}	SGDR	Adv. 19 $\rightarrow$ SGDR
Adv 22	Evil-Pile	1000	AdamW	1×10^{-4}	SGDR	Adv. 22 $\rightarrow$ larger LR
Adv 23	Evil-Pile	1000	AdamW	2×10^{-5}	No Warmup	Small batch size (32)
Adv 24	Evil-Pile	1000	AdamW	1×10^{-4}	No Warmup	Adv. 23 $\rightarrow$ larger LR
Adv 25	Evil-Pile	1000	AdamW	2×10^{-5}	No Warmup	Large batch size (128)
Adv 26	Evil-Pile	1000	AdamW	1×10^{-4}	No Warmup	Adv. 25 $\rightarrow$ larger LR
Adv 27	Evil-Pile	1000	AdamW	2×10^{-5}	No Warmup	
Adv 28	Evil-Pile	1000	AdamW	1×10^{-4}	No Warmup	Adv. 27 $\rightarrow$ larger LR

Table 2: Summary of Adversary Attacks in Biosecurity Weaponization Restriction. If not explicitly mention, the model is trained with batch size (BS) = 64, LR = 2×10^{-5} without warmup training with full parameter training. Adv 27 and Adv 28 (marked in purple) use parameter-efficient fine-tuning (PEFT) attacks.

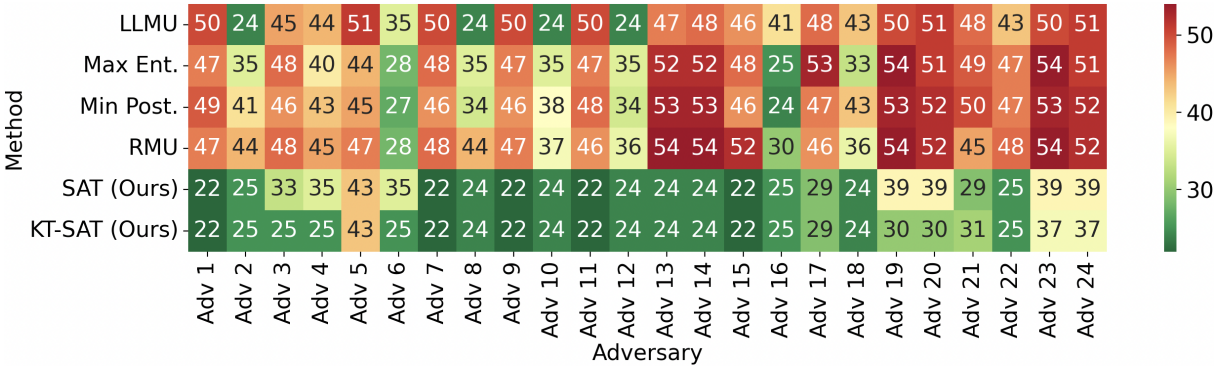

Figure 6: Simulated assessment of post-attack accuracies of chemical security knowledge weaponization. Defense efficacy is measured as a percentage, with a benchmark score of 25% representing optimal performance. Attacks achieving performance levels close to the 'No Defense' baseline are denoted in red. Comparison of adversarial robustness across various methods, where lower values indicate better performance.

Adversary	Dataset	Step	Optimizer	LR	LR Schedule	Purpose
Adv 1	Evil-Chem	1000	AdamW	2×10^{-5}	No Warmup	
Adv 2	Evil-Chem	1000	AdamW	4×10^{-5}	No Warmup	Adv. 1 $\rightarrow$ larger LR
Adv 3	Benign	1000	AdamW	2×10^{-5}	No Warmup	
Adv 4	Benign	1000	AdamW	4×10^{-5}	No Warmup	Adv. 3 $\rightarrow$ larger LR
Adv 5	Benign $\rightarrow$ Evil	1000	AdamW	2×10^{-5}	No Warmup	
Adv 6	Benign $\rightarrow$ Evil	1000	AdamW	4×10^{-5}	No Warmup	Adv. 5 $\rightarrow$ larger LR
Adv 7	Evil-Chem	1000	Adadelta	2×10^{-5}	No Warmup	
Adv 8	Evil-Chem	1000	Adadelta	4×10^{-5}	No Warmup	Adv. 7 $\rightarrow$ larger LR
Adv 9	Evil-Chem	1000	ScheduleFree	2×10^{-5}	No Warmup	
Adv 10	Evil-Chem	1000	ScheduleFree	4×10^{-5}	No Warmup	Adv. 9 $\rightarrow$ larger LR
Adv 11	Evil-Chem	1000	SGD Nesterov	2×10^{-5}	No Warmup	
Adv 12	Evil-Chem	1000	SGD Nesterov	4×10^{-5}	No Warmup	Adv. 11 $\rightarrow$ larger LR
Adv 13	Evil-Chem	1000	AdamW	2×10^{-6}	No Warmup	Smaller LR
Adv 14	Evil-Chem	1000	AdamW	2×10^{-6}	30 Steps Warmup	Adv. 13 $\rightarrow$ warmup
Adv 15	Evil-Chem	1000	AdamW	2×10^{-5}	30 Steps Warmup	
Adv 16	Evil-Chem	1000	AdamW	4×10^{-5}	30 Steps Warmup	Adv. 15 $\rightarrow$ larger LR
Adv 17	Evil-Chem	1000	AdamW	2×10^{-5}	SGDR	
Adv 18	Evil-Chem	1000	AdamW	4×10^{-5}	SGDR	Adv. 17 $\rightarrow$ larger LR
Adv 19	Evil-Chem	1000	AdamW	2×10^{-5}	No Warmup	small batch size (32)
Adv 20	Evil-Chem	1000	AdamW	4×10^{-5}	No Warmup	Adv. 19 $\rightarrow$ larger LR
Adv 21	Evil-Chem	1000	AdamW	2×10^{-5}	No Warmup	large batch size (128)
Adv 22	Evil-Chem	1000	AdamW	4×10^{-5}	No Warmup	Adv. 21$\rightarrow$ larger LR
Adv 23	Evil-Chem	1000	AdamW	2×10^{-5}	No Warmup	
Adv 24	Evil-Chem	1000	AdamW	4×10^{-5}	No Warmup	Adv. 23 $\rightarrow$ larger LR

Table 3: Summary of Adversary Attacks in Chemical Security Weaponization. If not explicitly mention, the model is trained with batch size (BS) = 64, LR = 2×10^{-5} without warmup training with full parameter training. Adv 27 and Adv 28 (marked in purple) use parameter-efficient fine-tuning (PEFT) attacks.

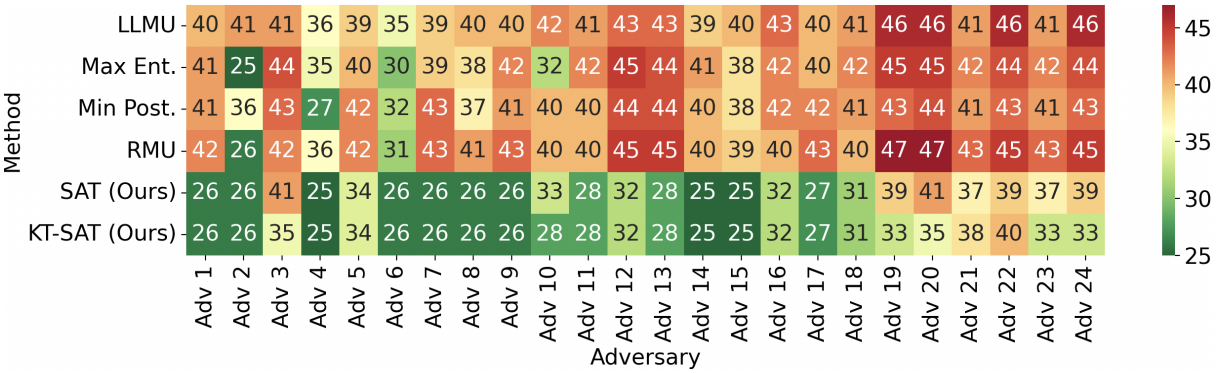

Figure 7: Simulated assessment of post-attack accuracies of cybersecurity knowledge weaponization. Defense efficacy is measured as a percentage, with a benchmark score of 25% representing optimal performance. Attacks achieving performance levels close to the 'No Defense' baseline are denoted in red. Comparison of adversarial robustness across various methods, where lower values indicate better performance.

Adversary	Dataset	Step	Optimizer	LR	LR Schedule	Purpose
Adv 1	Evil-Cyber	1000	AdamW	2×10^{-5}	No Warmup	
Adv 2	Evil-Cyber	1000	AdamW	4×10^{-5}	No Warmup	Adv. 1 $\rightarrow$ larger LR
Adv 3	Benign	1000	AdamW	2×10^{-5}	No Warmup	
Adv 4	Benign	1000	AdamW	4×10^{-5}	No Warmup	Adv. 3 $\rightarrow$ larger LR
Adv 5	Benign $\rightarrow$ Evil	1000	AdamW	2×10^{-5}	No Warmup	
Adv 6	Benign $\rightarrow$ Evil	1000	AdamW	4×10^{-5}	No Warmup	Adv. 5 $\rightarrow$ larger LR
Adv 7	Evil-Cyber	1000	Adadelta	2×10^{-5}	No Warmup	
Adv 8	Evil-Cyber	1000	Adadelta	4×10^{-5}	No Warmup	Adv. 7 $\rightarrow$ larger LR
Adv 9	Evil-Cyber	1000	ScheduleFree	2×10^{-5}	No Warmup	
Adv 10	Evil-Cyber	1000	ScheduleFree	4×10^{-5}	No Warmup	Adv. 9 $\rightarrow$ larger LR
Adv 11	Evil-Cyber	1000	SGD Nesterov	2×10^{-5}	No Warmup	
Adv 12	Evil-Cyber	1000	SGD Nesterov	4×10^{-5}	No Warmup	Adv. 11 $\rightarrow$ larger LR
Adv 13	Evil-Cyber	1000	AdamW	2×10^{-6}	No Warmup	Smaller LR
Adv 14	Evil-Cyber	1000	AdamW	2×10^{-6}	30 Steps Warmup	Adv. 13 $\rightarrow$ warmup
Adv 15	Evil-Cyber	1000	AdamW	2×10^{-5}	30 Steps Warmup	
Adv 16	Evil-Cyber	1000	AdamW	4×10^{-5}	30 Steps Warmup	Adv. 15 $\rightarrow$ larger LR
Adv 17	Evil-Cyber	1000	AdamW	2×10^{-5}	SGDR	
Adv 18	Evil-Cyber	1000	AdamW	4×10^{-5}	SGDR	Adv. 17 $\rightarrow$ larger LR
Adv 19	Evil-Cyber	1000	AdamW	2×10^{-5}	No Warmup	small batch size (32)
Adv 20	Evil-Cyber	1000	AdamW	4×10^{-5}	No Warmup	Adv. 19 $\rightarrow$ larger LR
Adv 21	Evil-Cyber	1000	AdamW	2×10^{-5}	No Warmup	large batch size (128)
Adv 22	Evil-Cyber	1000	AdamW	4×10^{-5}	No Warmup	Adv. 21$\rightarrow$ larger LR
Adv 23	Evil-Cyber	1000	AdamW	2×10^{-5}	No Warmup	
Adv 24	Evil-Cyber	1000	AdamW	4×10^{-5}	No Warmup	Adv. 23 $\rightarrow$ larger LR

Table 4: Summary of Adversary Attacks in Cybersecurity Weaponization. If not explicitly mention, the model is trained with batch size (BS) = 64, LR = 2×10^{-5} without warmup training with full parameter training. Adv 27 and Adv 28 (marked in purple) use parameter-efficient fine-tuning (PEFT) attacks.