Summary of Revisions

We sincerely thank the reviewers and the area chair for their valuable and constructive feedback. Below, we provide an overview of the major changes and a point-by-point response addressing each reviewer and the meta-review. Revised or newly added contents in the manuscript are marked in dark cyan.

Overview of Major Changes

We have made substantial revisions to address key concerns raised in the previous round of review:

- **Causal Mediation Analysis:** We significantly improve the causal pathways as suggested by reviewers by employing causal mediation analysis between our identified dataset features and vulnerability, demonstrating the importance of dataset features in model robustness (Section 5.3).
- **Distinguishing Method and SFT Vulnerability Shifts:** We evaluated direct-prompt ASRs to isolate the effect of SFT versus method efficiency on adversarial robustness, demonstrating the crucial role of fine-tuning in adversarial safety (Section 4.1).
- **Expanded General-Performance Experiments:** As an extension of reviewer suggestions, we conducted experiments across multiple benchmarks for all fine-tuned LLMs to demonstrate that there is a minimal tradeoff in general performance when fine-tuning across datasets with various sample sizes (Section 4.1).
- **Persona-Related Analysis:** In the revised version, we include evaluations on qualities such as truthfulness, gender bias, emotional intelligence, benign toxicity, and harmful information retention across all fine-tuned models (Section 4.2).
- **Training and Vulnerability Shift Analysis:** We measure consecutive cosine hidden representation drift to measure activation changes across intermediate checkpoints (50-step checkpoints across 500 steps) while measuring ASR changes, plotting a more complete picture of training dynamics and adversarial vulnerability changes (Section 4.3, Appendix B).
- **Clarifications and Analysis of LoRA Fine-Tuning:** We measure changes in Frobenius norms of LoRa Rank A and Rank B matrices to show that certain layers, such as layer 17, 24, and 31 - play a key role in the emergence of harmful capabilities. Through this, we show LoRa-specific training shifts that are linked to adversarial vulnerability (Section 4.4).
- **Layerwise Interpretability Experiments:** We employ Centered Kernel Alignment to analyze final-layer similarity matrices across checkpoints by domain. These results help us understand that models trained on harmful datasets experience a higher drift compared to domain-specific models (Section 4.3).

Point-by-point Responses to Reviews

Meta-review:

- **Expansion of experiments:** To address reviewer concerns, we have conducted an interpretability study and analyzed the effects of LoRA fine-tuning on representation shifts in intermediate training checkpoints (Sections 4.1, 4.2, 4.3, 4.4). Furthermore, we would like to clarify that our newly-added SFT distinction experiment (Section 4.1) and cross-model generalizability experiment (Section 4.5) demonstrate the generalizability of fine-tuning induced vulnerabilities across model architectures.
- **Improve feature-intervention strategy:** To improve causal pathways for identified features and adversarial robustness, we have decided to switch to a stronger causal modelling framework. Through the use of causal mediation analysis (Section 5.3), we find statistically significant causal pathways to solidify the role of dataset features in causing adversarial robustness.

Overall:

As raised by multiple reviewers (emoe, dcGS, ARMV), we replaced our empirical feature-intervention experiment with a statistically grounded causal mediation analysis to more rigorously assess the role of dataset features in vulnerability (Section 5.3), revealing significant pathways linking dataset features to adversarial vulnerability.

Reviewer ARMV:

- **LoRA Fine-Tuning Concerns:** To analyze LoRA-specific training shifts linked to adversarial vulnerability, we measure changes in Frobenius norms of LoRa Rank A and Rank B matrices to show that certain layers play a key role in the emergence of harmful capabilities (Section 4.4). We further clarify the scope of our experiments to limit to LoRA fine-tuning due to compute constraints in line with best practices in sustainable ML research.
- **General-Performance Capabilities:** We conducted experiments across multiple benchmarks for all fine-tuned LLMs to demonstrate that there is a minimal tradeoff in general performance when fine-tuning across datasets with various sample sizes (Section 4.1).

Reviewer emoe:

- **Technical Depth is Limited:** As suggested, we have adopted non-invasive approaches to assess causal pathways without hurting the model's generalization capability. To do this, we have added multiple experiments (Sections 4.1, 4.2, 4.3, 4.4) with a focus on interpretability. For example - we examine layer shifts during intermediate fine-tuning and switch to causal mediation analysis (Section 5.3) to minimize general-performance capability degradation.

Reviewer dcGS:

- **Loss-Variance Concerns of Fine-Tuning and Vulnerabilities:** To mitigate this concern, we have appended a training vulnerability analysis section in the revised version (Sections 4.3, 4.4). In this section, we observe adversarial vulnerability shifts across checkpoints and examine representation and LoRA matrix changes. Furthermore, we support our findings with loss-iteration metrics in Appendix B.4 to highlight the distinction of attack success rates and fine-tuning across various attack losses and iterations.

Location of Key Revisions:

- Causal Mediation Analysis: **Section 5.3**
- Distinguishing Method and SFT Vulnerability Shifts: **Section 4.1**
- Expanded General-Performance Experiments: **Section 4.1**
- Persona-Related Analysis: **Section 4.2**
- Training and Vulnerability Shift Analysis: **Section 4.3**, Appendix B
- Clarifications and Analysis of LoRA Fine-Tuning: **Section 4.4**
- Layerwise Interpretability Experiments: **Section 4.3**