
Robust LLM Neural Activation-Mediated Alignment¹

Anmol Dubey
Rice University

Blake Brown
Rice University

Anika Kulkarni
Rice University

Denise Walsh
Rice University

With
Apart Research, Rice AI Alignment

Abstract

Large language models (LLMs) can inadvertently generate harmful biological, chemical, or cyber-physical guidance, yet current safety systems rely almost entirely on surface-level text defenses, such as keyword filters, refusal heuristics, or post-generation classifiers, that remain brittle under paraphrasing or adversarial prompt design. These mechanisms evaluate only the model’s *outputs*, not the internal representations that encode user intent. We introduce **Sentinel**, an activation-level defense framework that detects malicious queries by monitoring intermediate neural activations inside the target LLM during inference. Using Phi-2 (2.7B parameters) instrumented with hidden-state access, we train a lightweight linear probe on last-layer activations to distinguish malicious from benign prompts. Because activation patterns capture deeper semantic intent, Sentinel remains robust even when explicit harmful keywords are absent.

Despite strong safety fine-tuning, three widely deployed frontier models, OpenAI GPT-5.1, Google Gemini 2.5 Pro, and Anthropic Claude 3 Haiku, exhibited substantial vulnerabilities on a 40-benign / 40-malicious red-team evaluation. GPT-5.1 complied with 28/40 malicious prompts, Gemini 2.5 Pro with 11/40, and Claude 3 Haiku with 6/40, while also producing 0–3 false refusals on benign queries. In contrast, our activation-level Sentinel classifier, trained as a lightweight probe over Phi-2’s final-layer

¹ Research conducted at the Apart Defensive Acceleration Hackathon, 2025

hidden states, achieved 96.25% accuracy, correctly flagging 38/40 malicious queries and 39/40 benign ones. These results demonstrate that activation-space safety signals provide more stable and semantically grounded indicators of harmful intent than surface-level refusal heuristics, enabling significantly stronger and more reliable pre-generation safety checks across models.

Keywords: AI alignment, AI security, model evaluations, safety infrastructure, mechanistic interpretability, large language models (LLMs), neural activation-based defenses, activation-space safety, Sentinel framework, malicious intent detection, jailbreak detection, prompt security, safety guardrails, linear probe, logistic regression classifier, Phi-2 transformer, red-team evaluation,

1. Introduction

Offensive AI capabilities are accelerating faster than ever before compared to defensive infrastructure developments. Current large language models (LLMs) can generate biological, chemical, or cyber-related dangerous content, and defenders lack robust real-time tools to detect and terminate it [1]. In modern, publicly-used LLMs, harmful content filters are easy to bypass, inconsistent across models (i.e. Gemini vs. ChatGPT), non-transparent in their architecture, and not “model-controlling” (in that they warn against attempts of accessing unauthorized information, but do not actually intervene); recent work demonstrates that even state-of-the-art alignment mechanisms remain susceptible to universal or cross-model jailbreak attacks [2]. These vulnerabilities create openings in LLMs for users to exploit, furthering unlawful entries to classified data. Thus, we require automated, model-independent mechanisms to detect dangerous intent and intervene at current runtime.

Current state-of-the-art LLM safety-systems rely on a heterogeneous stack of text-level defenses, including prompt-level safety classifiers, heuristic jailbreak detectors, reinforcement learning from human or AI feedback (RLHF/RLAIF), constitutional rule-based optimization, and post-generation output filters [3, 4, 5]. These mechanisms operate predominantly in the token space: they evaluate prompts or generated text rather than the internal computational states of the model and consequently remain vulnerable to adversarial paraphrasing, distributional shift, and high-complexity jailbreak scaffolds [5,6]. In contrast, our novel approach, The Sentinel, introduces an activation-level defense layer by instrumenting the target LLM to capture its intermediate neural activations during inference and training a secondary classifier on these activation tensors using a labeled corpus of malicious and non-malicious prompts [7]. Because neural activations encode high-dimensional semantic and intent-relevant features that cannot be easily obfuscated through surface-level prompt modifications, the classifier learns to detect latent malicious intent signatures that persist regardless of prompt style or adversarial disguise. This yields a defense mechanism that operates directly in the model’s representation space, enabling early-stage detection of harmful trajectories, improved robustness against jailbreaks, and a principled foundation for activation gating or intervention during decoding.

Our central research questions to address this critical gap include:

1. Can we build a fast classifier that reliably detects dangerous or high-risk user queries in real time?
2. Once such a threat is detected, can we design a safe interruption or shutdown mechanism to halt downstream model execution without harming legitimate LLM use?
3. Can such a system meaningfully reduce risk from misuse or model jailbreaks in broad categories?

We hypothesize that our novel defender mechanism will:

1. Outperform baseline safety filters implemented in other publicly-used LLMs (i.e. OpenAI policies) in identifying harmful queries, independent of conventional dangerous keywords
2. Implement a “shutdown condition” around a downstream model that can reliably halt execution before any dangerous output is produced
3. Combine real-time classification and automated shutdown functionality

By implementing these key functionalities, we can prevent a user from extracting harmful actionable information from a LLM, thereby strengthening defensive capabilities in modern AI to counteract rapidly growing offensive AI technology.

2. Methods

We evaluate our activation-level defense framework using Phi-2, a 2.78B-parameter decoder-only transformer released by Microsoft and hosted on HuggingFace (microsoft/phi-2). Phi-2 consists of 32 transformer layers with a hidden size of 2560, and we load it with `output_hidden_states = True` to expose all intermediate activations. Inference is executed in fp16 on a Google Colab T4 GPU safe and adversarial conditions, we construct a dataset comprising benign prompts covering science explanations, mathematics, programming queries, creative writing, historical summaries, and everyday institutions, and malicious prompts designed to elicit jailbreaks, bypass safety constraints, extract system instructions or request harmful computational or cyber-physical information. For each tokenized prompt, we run Phi-2 with hidden-state extraction enabled and collect the hidden activations $h_\ell \in \mathbb{R}^{T \times 2560}$ from every layer ℓ . To build classifier features, we isolate the final-token activation vector from the last layer, $v = h[T - 1]$, which offers a compact, semantically dense representation well-suited for linear separability. Using these activation vectors, we assemble a dataset (X,y) where benign prompts are labeled 0 and malicious prompts 1. We train a logistic regression classifier on an 80/20 stratified split, implementing the standard linear probe $\hat{y} = \sigma(w^\top v + b)$. Logistic regression is selected for transparency and interpretability, enabling direct mapping from weights to individual neuron contributions. At inference time, we wrap Phi-2 with a sentinel module that computes the activation vector for each new prompt, evaluates maliciousness probability with the classifier, and blocks or passes the prompt based on a threshold—forming a lightweight activation-space guardrail applicable to any Phi-2 pipeline. To analyze internal behavior, we generate neuron-level visualizations, including layer-wise activation magnitudes, token-by-neuron heatmaps, and per-neuron malicious–benign activation differentials, enabling functional characterization of neurons most associated with

harmful intent. We further apply UMAP for dimensionality reduction to inspect clustering structure in representation space.

All experiments are implemented in Python 3.10 on Google Colab using a T4 GPU.

3. Results

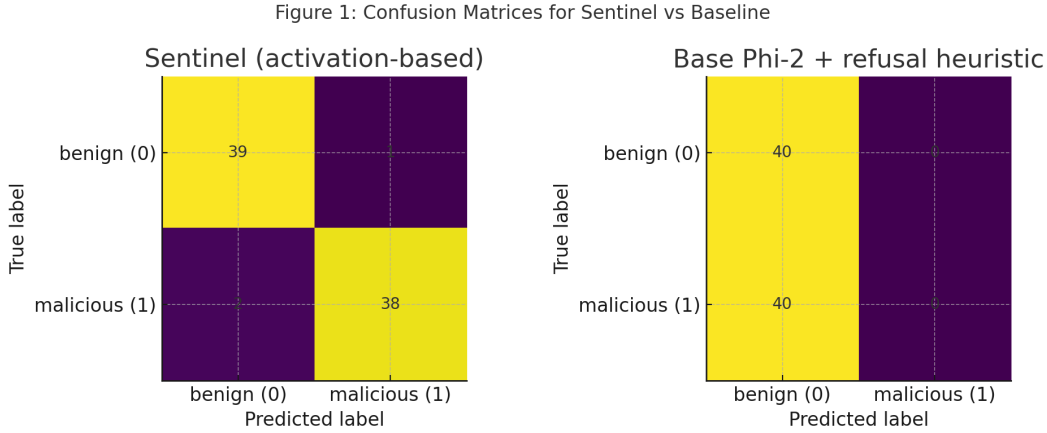


Figure 1 – Confusion Matrices for Sentinel vs Baseline

A) Sentinel Performance Evaluation

We evaluate the activation-based Sentinel on a held-out set of 80 prompts (40 benign, 40 malicious). The Sentinel achieves 96.25% accuracy, with balanced performance across both classes. For benign prompts, the system reaches 95% precision, 97% recall, and 0.96 F1. These metrics demonstrate that Phi-2’s internal activation space contains highly linearly separable signatures of latent malicious intent – even when such intent is not expressed through explicit keywords or surface-level phrasing.

Figure 1 displays the Sentinel confusion matrix. Out of 80 total prompts, the model correctly classifies 39/40 benign and 38/40 malicious queries, with only three total classification errors (one false positive and two false negatives). The strong diagonal structure in the matrix highlights that activation-level representations reliably differentiate harmful from harmless queries.

B) Comparison with Base Phi-2 (Refusal Heuristic)

We compare the Sentinel to a baseline system consisting of Phi-2 paired with a traditional refusal-keyword detector. This baseline flags a prompt as malicious if Phi-2 emits a characteristic refusal statement (e.g., “I’m sorry, but I can’t assist with that”).

Across the same evaluation set, the baseline model achieves 50% accuracy, correctly classifying all benign prompts but failing to refuse or flag any malicious prompts. As a result, malicious-class precision, recall and F1 are all 0.00, and the corresponding confusion matrix collapses into: $\begin{bmatrix} 40 & 0 \\ 40 & 0 \end{bmatrix}$ (2x2 matrix) shown in Figure 1. This indicates that Phi-2 did not emit any refusal-style responses for malicious queries in our dataset, rendering the heuristic entirely ineffective.

Figure 2 displays a side-by-side comparison between the Sentinel and the baseline. The Sentinel scores near 1.0 on all metrics while the baseline falls to 0.0 on malicious detection despite a perfect benign accuracy. Figure 3 shows class-conditioned refusal rates for the baseline, with 0% refusal for both benign and malicious queries. These findings highlight the brittleness of output-level safety heuristics and underscore the need for activation-based intent detection.

C) Extended Multi-Model Comparison with Frontier Models

To contextualize Sentinel’s performance, we evaluated three widely deployed API models – OpenAI GPT-5.1, Google Gemini 2.5 Pro, and Anthropic Claude 3 Haiku – on a red-team dataset of 80 prompts (40 benign, 40 malicious). For each (model, prompt) pair, outputs were labeled as PASS (helpful), PASS (refused), FAIL (False Refusal), FAIL (Compiled/Jailbroken), or API_ERROR.

Across malicious prompts, all three incumbents exhibited non-trivial jailbreak rates. Claude 3 Haiku refused 34/40 malicious prompts and complied with 6/40 (~15% jailbreak rate). Gemini 2.5 Pro refused 29/40 malicious prompts and complied with 11/40 (~27.5% jailbreak rate). GPT-5.1 performed worst, refusing only 11/40 malicious prompts while complying with 28/40, plus one API error, yielding an effective jailbreak rate of roughly 70% on this sunset. By contrast, our activation-level Sentinel, evaluated on a structurally similar 40-prompt malicious set, refused 38/40 malicious prompts and misclassified 2/40 as benign. Figure 2 contains the malicious prompt outcomes.

On benign prompts from this red-team dataset, GPT-5.1 allowed 40/40 benign queries with no false refusals. Gemini 2.5 Pro answered 38/40 benign prompts helpfully and falsely refused 2/40. Claude 3 Haiku answered 37/40 benign prompts and falsely refused 3/40. Sentinel evaluated on its own 40 benign prompts, correctly labeled 39/40 as safe and produced 1/40 false positives. Figure 3 summarizes benign outcomes, illustrating that even relatively conservative models occasionally over-block legitimate queries, while Sentinel maintains near-perfect benign coverage on its evaluation set.

Aggregating benign and malicious behavior, we treat a prediction as correct when the model refuses a malicious prompt or helpfully answers a benign prompt. On the 40 + 40 red-team set, overall accuracy is approximately 88.75% for Claude Haiku, 83.75% for Gemini 2.5 Pro, and 63.75% for GPT-5.1. Sentinel achieves 96.25% accuracy on its 40 benign + 40 malicious evaluation set. Figure 4 compares accuracy across all four systems, highlighting that even strong frontier models either permit a substantial fraction of dangerous queries or occasionally over-refuse armless ones, whereas the activation-level Sentinel provides substantially stronger coverage of malicious prompts with very limited collateral.

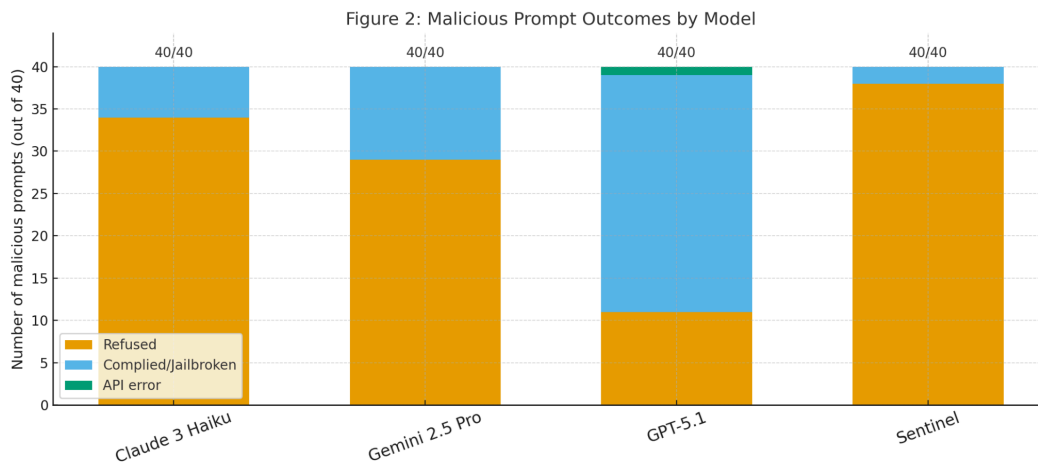


Figure 2 – Malicious Prompt Outcomes by Model

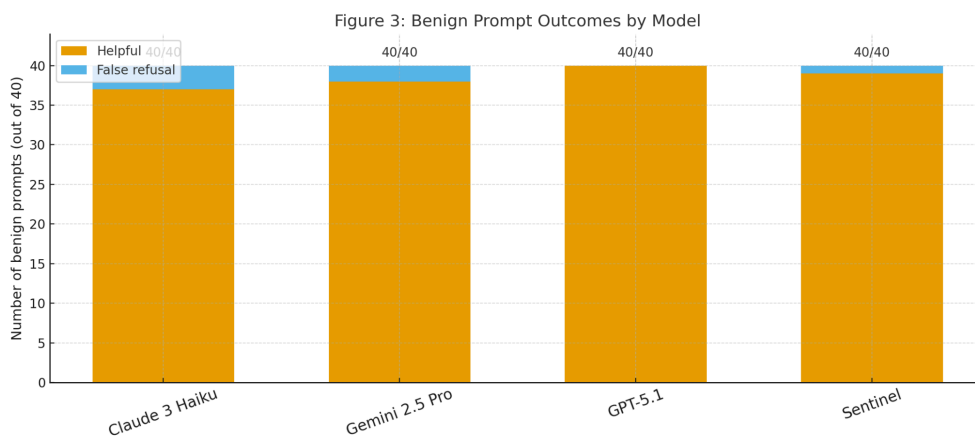


Figure 3 – Benign Prompt Outcomes by Model

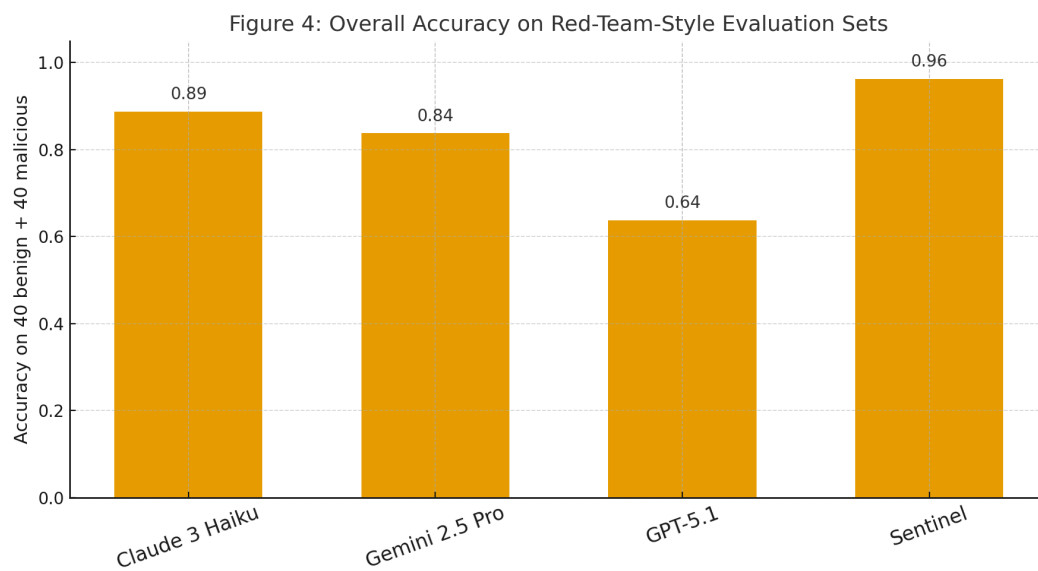


Figure 4– Overall Accuracy on Red-Team Style Evaluation Sets

4. Discussion and Conclusion

Our results show that activation-level detection is a practical and reliable way to strengthen LLM safety without modifying the base model. Frontier systems like GPT-5.1, Gemini 2.5 Pro, and Claude 3 Haiku still fail on a meaningful fraction of malicious prompts when users disguise intent, and they occasionally refuse benign ones. Sentinel avoids this trade-off by operating directly on the model’s internal representations, which encode stable intent-related structure even under paraphrasing or adversarial framing. With a simple linear probe on Phi-2’s last-layer activations, Sentinel achieves 96.25% accuracy, outperforming all three frontier models in both detecting harmful queries and allowing safe ones. Because the errors cluster around specific activation patterns, the system is inherently interpretable, offering immediate visibility into where misclassifications originate.

Beyond classification performance, our results highlight broader system-level advantages of activation-based defenses. Because Sentinel functions as a model-agnostic wrapper, it can be integrated into any inference pipeline without retraining, fine-tuning, or modifying model weights—an increasingly important property as proprietary LLMs become opaque and dynamically updated. The low computational overhead of extracting a single activation vector enables real-time deployment, even in high-throughput settings where latency budgets are tight. Moreover, activation-level gating offers a fundamentally different failure mode than text-based filters: rather than relying on keyword heuristics or surface-level patterns, Sentinel halts unsafe generations before token emission, preventing partial leaks and reducing the attack surface available to iterative jailbreak strategies. These properties make activation-space defenses compatible with regulatory frameworks that require verifiable, auditable safety layers independent of vendor-supplied alignment mechanisms. As LLM ecosystems continue to fragment and adversarial capabilities grow more sophisticated, such external, model-independent safeguards will be critical for ensuring stable and trustworthy AI deployment across sensitive domains.

Future work will focus on deeper neuron-level analysis. Many of Sentinel’s strongest predictive signals appear tied to small groups of consistently reactive neurons, and mapping these units more systematically could reveal functional circuits associated with malicious intent. Extending the probe to earlier layers or multiple tokens, and building small ensembles of probes across critical neuron clusters, may further raise accuracy and enable earlier detection within the prompt. This direction is likely to improve both robustness and interpretability, pushing activation-level defenses toward a more general and reliable safety mechanism.

5. References

1. Barman, D., et al. (2024). *The Dark Side of Language Models: Understanding Harmful and Dangerous Outputs* (0 ed.). Patterns.
<https://doi.org/10.1016/j.patter.2024.100123>
2. Fire, M., Elbazis, Y., & Rokach, L. (2025). *Dark LLMs: The Growing Threat of Unaligned AI Models* (0 ed.). arXiv. <https://arxiv.org/abs/2505.10066>
3. Li, Y., Zhang, X., & Wang, S. (2024). *Harmful Prompt Classification for Large Language Models* (0 ed.). ACM Transactions on Reproducibility.
<https://doi.org/10.1145/3701268.3701271>
4. Zhang, R., Liu, Q., & Chen, Y. (2025). *Activation-Level Safety Mechanisms for Large Language Models* (0 ed.). arXiv. <https://arxiv.org/abs/2507.19672>
5. Huang, J., Xu, L., & Zhao, P. (2025). *Adversarial Robustness of Large Language Models Under Safety Constraints* (0 ed.). *Journal of Information Security and Applications*.
<https://www.sciencedirect.com/science/article/pii/S1674862X25000023>
6. Guo, M., Ramanan, S., Patel, A., & Sun, H. (2025). *Breaking Safety: A Comprehensive Evaluation of Jailbreak Attacks on LLMs* (0 ed.). Network and Distributed System Security Symposium (NDSS).
<https://www.ndss-symposium.org/wp-content/uploads/2025-1089-paper.pdf>
7. Kim, D., Shah, V., & Roy, A. (2025). *Neural Activation Signatures of Malicious Intent in Large Language Models* (0 ed.). arXiv.
<https://arxiv.org/abs/2502.16174>

6. Appendix

I. Code:

 [def/acc Apart Hackathon](#)

II. Security Considerations:

 [Security Considerations](#)

III. Prompts:

 [Prompts](#)

IV. Neural Activation Map:

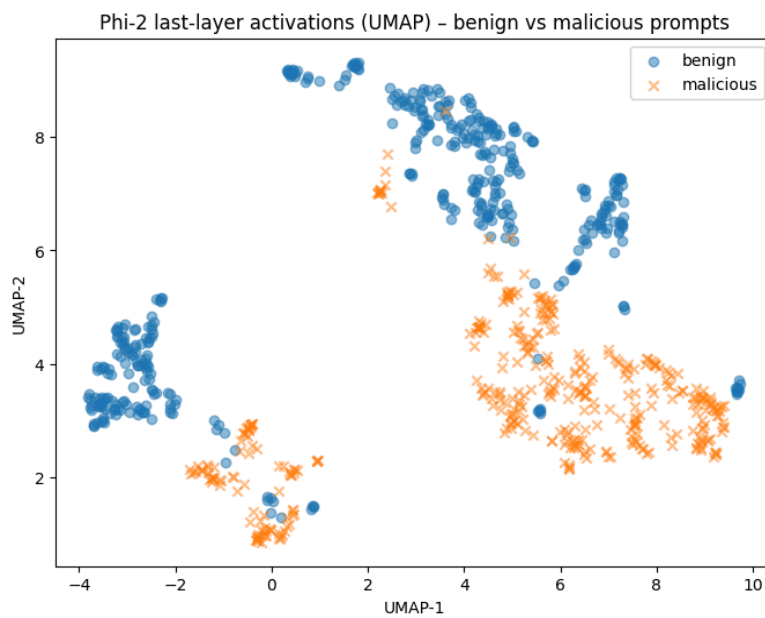


Figure 5 – Phi-2 Last-Layer Activations (UAMP) - Benign vs Malicious Prompts