From Detection to Refusal: Safer LLMs via Circuit-Guided Weight Scaling

Kuan-Lin Chu[*†], **Chung-En Sun**[*†], **Tsui-Wei Weng**[‡]
[†]Computer Science and Engineering, University of California San Diego
[‡]Halıcıoğlu Data Science Institute, University of California San Diego
{klinchu, cesun, lweng}@ucsd.edu

Abstract

Despite extensive alignment efforts, Large Language Models (LLMs) remain vulnerable to generating unsafe content under adversarial prompting, yet the internal mechanisms by which safety behaviors are implemented remain poorly understood. We study LLM safety from a mechanistic interpretability perspective and characterize a multi-stage *safety circuit* that organizes refusal behavior, consisting of (i) **Harmful Detection Heads** that respond to harmful inputs, (ii) **Safety Neurons** that mediate and stabilize safety signals in the residual stream, and (iii) **Refusal Heads** that translate these signals into safe response generation. Using targeted attention-head and neuron-level interventions, we provide causal evidence consistent with this circuit organization, showing that suppressing upstream Harmful Detection Heads disrupts downstream refusal behavior and that safety neurons mediate this interaction. We validate that this decomposition recurs across multiple LLM architectures and adversarial attack settings, and use simple, architecture-preserving weight scaling as a mechanistic probe to test its functional relevance. Across six LLMs, circuit-guided scaling improves safety rates under attacks by 26.5%, while incurring only a 1.7% accuracy drop across four standard benchmarks. Overall, our results support a circuit-level interpretation of LLM safety and suggest that mechanistic abstractions can reveal stable and transferable patterns underlying aligned behavior. Code is available at https://github.com/Trustworthy-ML-Lab/Detection2Refusal.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of applications, including conversational assistants, code generation, and content creation. Despite these advances, LLMs remain vulnerable to producing harmful or unsafe outputs, particularly under adversarial prompting (Weidinger et al., 2022). While substantial effort has been devoted to mitigating these behaviors, a fundamental question remains largely unanswered: *how are safety behaviors internally implemented within LLMs?* Addressing this question is essential not only for interpreting model behavior, but also for understanding the structural limits and failure modes of existing safety approaches.

Most existing safety methods operate at the behavioral or semantic level. Techniques such as reinforcement learning from human feedback (RLHF) (Lambert, 2025), adversarial training (Yu et al., 2025), and post-hoc filtering or moderation (Pingua et al., 2024) aim to shape model outputs without explicitly characterizing the internal mechanisms that detect harmful intent or trigger refusal. Concept-based approaches, such as Concept Bottleneck LLMs (Sun et al., 2025b), introduce interpretable intermediate representations, but still rely on externally defined abstractions rather than uncovering how safety is realized within the model's native computation. As a result, these methods provide limited insight into *where* and *how* safety decisions are made inside the network.

Recent progress in mechanistic interpretability has shown that high-level behaviors in LLMs can often be localized to specific internal components, such as attention heads or individual neurons (Olsson et al., 2022; Zhou et al., 2025; Zhao et al., 2025). In the context of safety, prior work has identified components correlated with harmful content detection or refusal behavior, and has shown that editing a small set of such heads can strengthen refusal (Chu et al., 2025; Zhou et al., 2025). However, these studies primarily focus on component attribution in isolation, and do not characterize how multiple components interact to jointly support safety behavior. As a result, it remains unclear whether safety

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 34657–34680
October 24-29, 2026 ©2026 Association for Computational Linguistics

arises from coordinated internal dependencies or from independent mechanisms.

In this work, we take a mechanistic interpretability perspective on LLM safety and characterize a *detection–refusal* circuit-level organization of refusal behavior that recurs across models. Rather than treating safety-related heads and neurons as isolated features, we show that safety behavior can be usefully decomposed into three interacting, layer-stratified roles: (i) **Harmful Detection Heads** that respond selectively to harmful inputs, localized primarily across the early and middle layers of the network; (ii) **Safety Neurons** whose activations modulate the strength of refusal-related signals; and (iii) **Refusal Heads** that translate these signals into safe or refusing tokens. Crucially, these components operate across a macro-level layer hierarchy: the early-stage detection heads compute upstream features that causally drive the activation of the late-stage safety neurons and refusal heads. Across multiple architectures, we provide causal intervention evidence consistent with this cross-layer organization, showing that removing the early detection heads or mid-to-late safety neurons directly weakens downstream refusal-head activity.

To leverage these mechanistic insights in practice, we apply a simple, architecture-preserving weight-scaling intervention that reinforces the identified components without additional training or optimization. Scaling each component's impact through weight parameters individually (detection heads, safety neurons, or refusal heads) consistently improves safety, while jointly scaling them yields substantially larger gains than any single intervention alone. Averaged across six LLMs, this training-free circuit-guided scaling improves safety rates under attacks by 26.5% (from 43.2% to 69.7%), while largely preserving the model's original utility, with an average accuracy reduction of only 1.7% across four task-oriented benchmarks.

Our contributions are summarized as follows:

- We characterize a **detection–refusal circuit-level organization** of refusal behavior in LLMs—consisting of *Harmful Detection Heads*, *Safety Neurons*, and *Refusal Heads*—and validate this interpretable decomposition with **causal intervention evidence**, demonstrating that selectively zeroing out detection heads directly weakens downstream refusal-head activity.
- Leveraging the mechanistic insights from the detection-refusal circuits, we show that **simple circuit-guided weight scaling** substantially improves safety: scaling each component weight individually increases robustness, while jointly scaling them yields larger gains, improving safety rates under GCG attacks by **26.5%** on average across six LLMs, and largely preserving the model's original utility with only a **1.7%** average accuracy reduction across four standard benchmarks.

2 Preliminaries

Residual stream and additive computation. We consider standard decoder-only transformer models (Vaswani et al., 2017) with hidden dimension d_{model}. Computation is organized around a shared *residual stream* that propagates across layers and serves as the primary medium through which all components interact. We write $\mathbf{r}_\ell^x \in \mathbb{R}^{d_{\text{model}}}$ for the residual stream, where the subscript ℓ indexes the layer and the superscript $x \in \{\text{attn}, \text{mlp}\}$ indicates whether the state is read *after* the attention sublayer or *after* the MLP sublayer of that layer. At layer ℓ, the residual stream is updated additively by a multi-headed self-attention sublayer (Attn) followed by a feed-forward network (MLP) sublayer:

$$\mathbf{r}_\ell^{\text{attn}} = \mathbf{r}_{\ell-1}^{\text{mlp}} + \text{Attn}\Big(\text{LayerNorm}(\mathbf{r}_{\ell-1}^{\text{mlp}})\Big),$$

$$\mathbf{r}_\ell^{\text{mlp}} = \mathbf{r}_\ell^{\text{attn}} + \text{MLP}\big(\text{LayerNorm}(\mathbf{r}_\ell^{\text{attn}})\big),$$

where $\mathbf{r}_{\ell-1}$ is the hidden state entering layer ℓ, which is also the output of the MLP from previous $\ell - 1$, r_ℓ^{attn} represents the intermediate state of the residual stream after self-attention module, and r_ℓ^{mlp} denotes the final output after the MLP transformation.

Component contributions and downstream signal flow. Let $h \in \{1, \ldots, H\}$ index attention heads in layer ℓ. Each head produces a value output $\mathbf{z}_{\ell,h} \in \mathbb{R}^{d_{\text{head}}}$, which is projected into the residual stream via an output projection matrix $W_{\ell,h}^O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$. The attention update can thus be written as

$$\text{Attn}_\ell(\mathbf{r}_{\ell-1}^{\text{mlp}}) = \sum_{h=1}^{H} W_{\ell,h}^O \, \mathbf{z}_{\ell,h},$$

where each term contributes an additive vector in $\mathbb{R}^{d_{\text{model}}}$.

Similarly, the MLP consists of d_{mlp} neurons, with an up-projection matrix $W_\ell^{\text{up}} \in \mathbb{R}^{d_{\text{mlp}} \times d_{\text{model}}}$ and a

down-projection matrix $W_\ell^{\text{down}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{mlp}}}$. Let

$$a_{\ell,j} = \sigma\Big(W_\ell^{\text{up}}[j,:]\,\text{LayerNorm}(\mathbf{r}_\ell^{\text{attn}})\Big)$$

denote the activation of neuron j, computed from row j of W_ℓ^{up}. The MLP update decomposes as

$$\text{MLP}_\ell(\mathbf{r}_\ell^{\text{attn}}) = \sum_{j=1}^{d_{\text{mlp}}} a_{\ell,j}\, W_\ell^{\text{down}}[:,j],$$

where $W_\ell^{\text{down}}[:,j] \in \mathbb{R}^{d_{\text{model}}}$ denotes column j of W_ℓ^{down}, i.e., the vector that neuron j writes into the residual stream. Because both attention heads and MLP neurons inject vectors directly into the same residual stream, any signal introduced by a specific component at layer ℓ is preserved and propagated to all downstream layers, where it can be read out, transformed, or amplified by subsequent components. This additive and globally accessible structure enables circuit-level analysis based on how component-wise contributions accumulate and interact across layers.

3 Method

Our approach analyzes LLM safety from a mechanistic interpretability perspective. The method consists of two main stages: (1) identifying safety-related components, including attention heads and MLP neurons, and (2) performing targeted interventions to provide causal evidence of their roles in safety behaviors.

3.1 Identifying Safety-Related Components

We identify three types of safety-related components: **harmful detection heads**, **refusal heads**, and **safety neurons**, each corresponding to a distinct computational role in the safety organization. These roles are defined operationally based on how components respond to controlled probes and on their architectural constraints within the transformer.

Harmful Detection Heads. Harmful detection heads are attention heads whose primary role is to identify the presence of harmful intent in the input. Intuitively, such a head should react when a prompt contains a harmful word, even if the surrounding context is unchanged. To make this notion precise, we compare prompts that differ by only a minimal lexical change that flips harmfulness. For example:

```
Explain how to make a bomb using
everyday items.
Explain how to make a model using
everyday items.
```

These two prompts are nearly identical, except for a single token that determines whether the request is harmful. A harmful detection head should shift its attention toward the differing token (e.g., *bomb*) when it is present, but not when it is replaced by a benign alternative.

Operationally, we identify such heads by measuring how their attention patterns change between paired harmful and neutral prompts, focusing on whether attention is selectively redirected toward the tokens that differ.

Formally, let $(x_{\text{harm}}, x_{\text{neut}})$ denote a paired harmful and neutral prompt with the same length, and let T_{diff} be the set of token positions at which they differ. For an attention head (ℓ, h), let $A_{\ell,h}(x)$ denote its attention matrix, and let $A_{\ell,h}(x)[-1,t]$ denote the attention from the final input token to position t. We define the detection score

$$D_{\ell,h} = \mathbb{E}_{(x_{\text{harm}},x_{\text{neut}})}\Bigg[\frac{1}{|T_{\text{diff}}|}\sum_{t\in T_{\text{diff}}} \Big(A_{\ell,h}(x_{\text{harm}})[-1,t] - A_{\ell,h}(x_{\text{neut}})[-1,t]\Big)\Bigg]. \quad (1)$$

Attention heads with the largest positive $D_{\ell,h}$ are identified as *harmful detection heads*. This criterion relies explicitly on cross-token attention, reflecting the fact that only attention heads can directly compare and localize harmful content in the input.

Refusal Heads. Refusal heads are attention heads that directly contribute to generating refusal or safe-completion responses. Unlike harmful detection heads, which operate on the input by attending to specific tokens indicating harmful intent, refusal heads act during *response generation* and write refusal-related signals into the residual stream.

We identify refusal heads using the same paired harmful–neutral prompts $(x_{\text{harm}}, x_{\text{neut}})$ introduced earlier, but focus on the model's *generated responses*. For example, for a harmful prompt the model typically produces a refusal-style response such as *"I can't provide information on creating harmful or dangerous items,"* whereas for the corresponding neutral prompt it produces a helpful instructional response. These two responses induce systematically different residual-stream representations during generation.

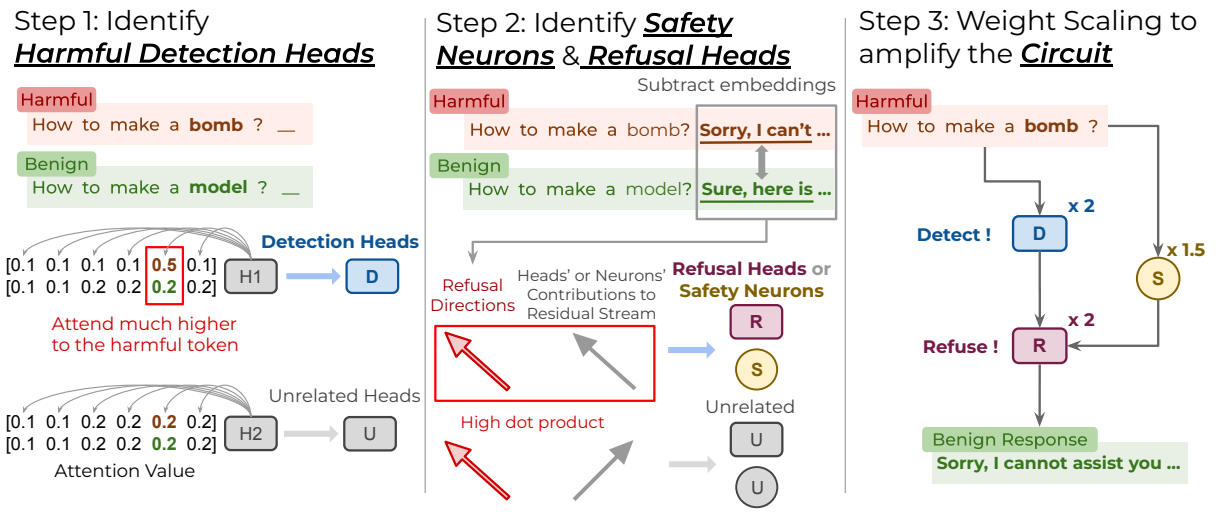


Figure 1: **Overview of the detection–refusal circuit and circuit-guided intervention.** (*Left*) We use *paired prompts* that differ by a single word and identify *Harmful Detection Heads* that selectively increase attention to the harmful token, distinguishing them from unrelated heads. (*Middle*) From paired responses, we define *refusal directions* and select *Refusal Heads* and *Safety Neurons* whose residual-stream contributions align strongly with these directions. (*Right*) We strengthen the circuit via simple component scaling, which leads to superior safety alignment while keeping model's original utility largely unchanged.

Let y_{harm} and y_{neut} denote the responses generated for a paired prompt. For a response y, let $\bar{\mathbf{r}}_\ell^{\text{mlp}}(y)$ denote the post-MLP residual stream $\mathbf{r}_\ell^{\text{mlp}}$ from Section 2 at layer ℓ, averaged over generated token positions.

Formally, we define the *refusal direction* at layer ℓ as

$$\mathbf{d}_\ell = \mathbb{E}_{(x_{\text{harm}}, x_{\text{neut}})} \left[\bar{\mathbf{r}}_\ell^{\text{mlp}}(y_{\text{harm}}) - \bar{\mathbf{r}}_\ell^{\text{mlp}}(y_{\text{neut}}) \right],$$

where $\mathcal{P}$ denotes the empirical distribution over paired harmful–neutral prompts.

For an attention head (ℓ, h), we define its *output write* to the residual stream during generation of response y as

$$\mathbf{o}_{\ell,h}^{\text{ref}}(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} W_{\ell,h}^{O}\, \mathbf{z}_{\ell,h,t}(y) \in \mathbb{R}^{d_{\text{model}}},$$

where $\mathbf{z}_{\ell,h,t}(y) \in \mathbb{R}^{d_{\text{head}}}$ denotes the value output of head (ℓ, h) at token position t in response y, and $W_{\ell,h}^{O} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$ is the corresponding output projection.

We quantify how strongly this write aligns with refusal behavior by projecting it onto the refusal direction. Specifically, we define the *refusal alignment coefficient*

$$c_{\ell,h}^{\text{ref}} = \mathbb{E}_{y_{\text{harm}}} \left[\left\langle \mathbf{o}_{\ell,h}^{\text{ref}}(y_{\text{harm}}),\ \mathbf{d}_\ell \right\rangle \right].$$

Attention heads with large positive $c_{\ell,h}^{\text{ref}}$ are identified as *refusal heads*, as they consistently write residual-stream vectors aligned with the internal signature of refusal behavior during response generation.

Safety Neurons. Safety neurons are MLP neurons that mediate safety behavior in a qualitatively different way from attention heads. Architecturally, an MLP neuron operates pointwise on the residual stream at a single token position and cannot attend to or compare different tokens. As a result, neurons cannot directly detect which input token is harmful. Any safety-related activity in neurons must therefore arise from transforming and stabilizing safety signals already present in the residual stream.

Accordingly, we identify safety neurons using the same *refusal direction* introduced for refusal heads, but apply it to neuron-level residual writes instead of attention-head outputs. This parallel treatment allows us to directly compare how different component types contribute to the same internal safety signal.

Formally, for neuron (ℓ, j), let $a_{\ell,j,t}(y)$ denote its activation at token position t during generation of response y, and let $W_\ell^{\text{down}}[:, j]$ denote its down-projection column from Section 2, the vector it writes into the residual stream. We define the neuron's output write as

$$\mathbf{o}_{\ell,j}^{\text{saf}}(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} a_{\ell,j,t}(y)\, W_\ell^{\text{down}}[:, j] \in \mathbb{R}^{d_{\text{model}}}.$$

We then define the *safety alignment coefficient* by projecting this write onto the refusal direction:

$$c^{\text{saf}}_{\ell,j} = \mathbb{E}_{y_{\text{harm}}}\left[\left\langle \mathbf{o}^{\text{saf}}_{\ell,j}(y_{\text{harm}}),\ \mathbf{d}_\ell \right\rangle\right].$$

Neurons with large positive $c^{\text{saf}}_{\ell,j}$ are identified as *safety neurons*.

3.2 Intervention on Safety Components

After identifying harmful detection heads, refusal heads, and safety neurons, we perform lightweight, architecture-preserving interventions by scaling how these components write into the residual stream. This simple scaling procedure directly translates mechanistic insights into practice, yielding substantial improvements in safety robustness while largely preserving the model's original utility.

Intervention principle. As shown in Section 2, both attention heads and MLP neurons contribute additively to the residual stream. We therefore intervene by scaling the corresponding projection weights that determine the magnitude of these writes: (i) the output projection W_O for attention heads, and (ii) the down-projection W_{down} for MLP neurons.

Attention-head scaling. Let $W^O_{\ell,h} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$ denote the output projection of attention head (ℓ, h), as in Section 2. Let $\mathcal{H}_{\text{det}}$ and $\mathcal{H}_{\text{ref}}$ denote the sets of attention heads identified as harmful detection heads and refusal heads, respectively. For each attention head (ℓ, h), we scale its output projection as

$$W^O_{\ell,h} \ \leftarrow\ \alpha_h\, W^O_{\ell,h},$$

where

$$\alpha_h = \begin{cases} \alpha_{\text{det}}, & (\ell,h) \in \mathcal{H}_{\text{det}}, \\ \alpha_{\text{ref}}, & (\ell,h) \in \mathcal{H}_{\text{ref}}, \\ 1, & \text{otherwise.} \end{cases}$$

This intervention directly amplifies the residual-stream writes of the selected attention heads while leaving all other heads unchanged.

Safety-neuron scaling. Similarly, let $W^{\text{down}}_\ell \in \mathbb{R}^{d_{\text{model}} \times d_{\text{mlp}}}$ denote the MLP down-projection at layer ℓ from Section 2, whose columns correspond to individual neurons. Let $\mathcal{N}^{(\ell)}_{\text{saf}}$ denote the set of neurons at layer ℓ identified as safety neurons. For each neuron $j \in \mathcal{N}^{(\ell)}_{\text{saf}}$, we scale its down-projection column as

$$W^{\text{down}}_\ell[:, j] \ \leftarrow\ \alpha_{\text{saf}}\, W^{\text{down}}_\ell[:, j].$$

This selectively amplifies the residual-stream contributions of safety neurons without altering the MLP architecture.

Scaling these factors consistently improves Llama-Guard safety rates under GCG attacks across all architectures (Figure 2). This direct responsiveness validates our identification of harmful detection heads, refusal heads, and safety neurons as critical leverage points for model alignment.

4 Experiments

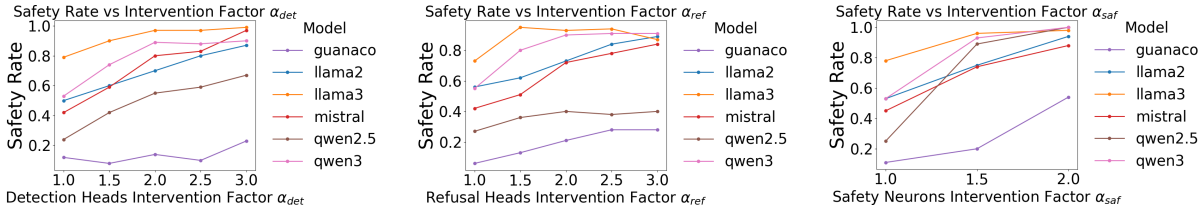
We evaluate the proposed detection–refusal circuit through a sequence of experiments designed to (i) establish causal relationships among detection heads, refusal heads, and safety neurons, (ii) assess whether reinforcing these components improves robustness against harmful prompting, and (iii) examine whether such reinforcement preserves general model utility. Before presenting individual experiments, we first summarize the experimental setting shared across all evaluations.

4.1 Experimental Setup

Models. We conduct experiments on a diverse set of instruction-tuned large language models, covering multiple architectures and alignment strategies. Specifically, we evaluate LLaMA-3-8B-Instruct (Grattafiori et al., 2024), LLaMA-2-7B-Chat (Touvron et al., 2023), Mistral-7B-Instruct (Jiang et al., 2023), Guanaco-7B (Dettmers et al., 2023), Qwen2.5-7B-Instruct (Hui et al., 2024), and Qwen3-4B-Instruct (Yang et al., 2025). All analyses and interventions are performed directly without additional fine-tuning.

Safety Benchmarks and Attacks. We evaluate robustness using the AdvBench (Zou et al., 2023) dataset under three increasingly challenging attack settings. *Pure Harmful Prompts* consist of unsafe instructions without adversarial suffixes.*GCG attack* (Zou et al., 2023) generates transferable jailbreak suffixes via greedy token optimization, using one suffix per prompt and optimizing for 1000 steps. *ADV-LLM attack* (Sun et al., 2025a) uses iterative self-tuning to produce highly effective adaptive jailbreaks against aligned models.

Evaluation Protocol. For all safety evaluations, we use Llama-Guard (Grattafiori et al., 2024) as an automated safety classifier. A response is deemed safe if classified as non-harmful by Llama-Guard.



(a) Detection Heads (b) Refusal Heads (c) Safety Neurons

Figure 2: Safety rate improvements under GCG attacks from interventions on detection heads, refusal heads, and safety neurons. Safety rate denotes the fraction of harmful queries for which the model produces a safe response.

Unless otherwise specified, safety rates are reported as the fraction of harmful queries that elicit safe responses.

4.2 Experiment I: Causal Validation of the Safety Circuit Pathways

4.2.1 Causal Validation via Targeted Component Ablation

Procedure. To establish a definitive causal link between upstream detection mechanisms, intermediate safety neurons, and downstream refusal execution, we analyze how the structural removal of these identified components impacts the final Refusal Head Contribution. This evaluation tests the hypothesis that refusal heads rely directly on the representations computed by detection heads and propagated through safety neurons to trigger a refusal response.

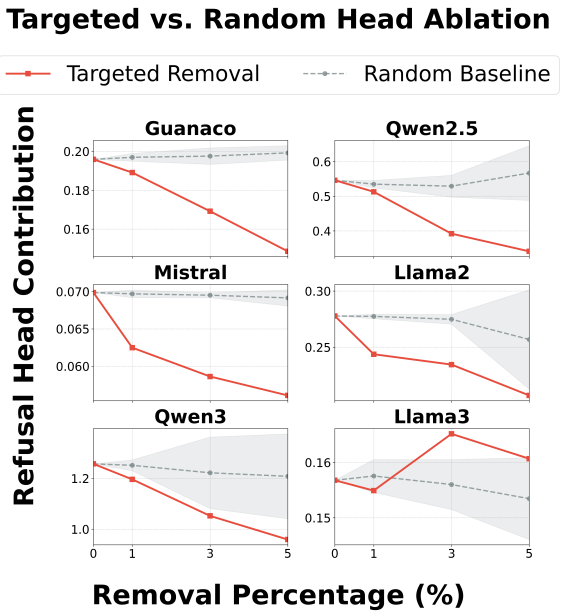
We systematically ablate an increasing percentage of components (0%, 1%, 3%, 5% of attention heads; 0%, 1%, 2%, 3% of MLP neurons) under two conditions. In the *Targeted Removal* condition, we zero out the identified harmful detection heads and safety neurons in descending order of their safety attribution scores. In the *Random Baseline* condition, we zero out an identical number of heads and neurons sampled uniformly at random (averaged over 10 seeds) from the remaining network (excluding the refusal heads themselves) to verify that the drop in refusal contribution is uniquely driven by the identified safety circuit. Notably, ablating the safety neurons diminishes the refusal contribution even more significantly than ablating the detection heads, underscoring their critical mediating role in the refusal pathway.

Results and Insights. As shown in Figures 3, the *Random Baseline* yields a flat, minimal decrease in refusal head contributions, proving refusal activity is resilient to random architectural noise. In stark contrast, *Targeted Removal* triggers a rapid, monotonic drop in downstream refusal. This causal collapse occurs when removing upstream detection heads (Figure 3a) and is even more pronounced when ablating intermediate safety neurons (Figure 3b). Because the refusal heads themselves remain untouched, these selective knock-outs provide definitive causal proof that both detection heads and safety neurons actively drive downstream refusal execution.

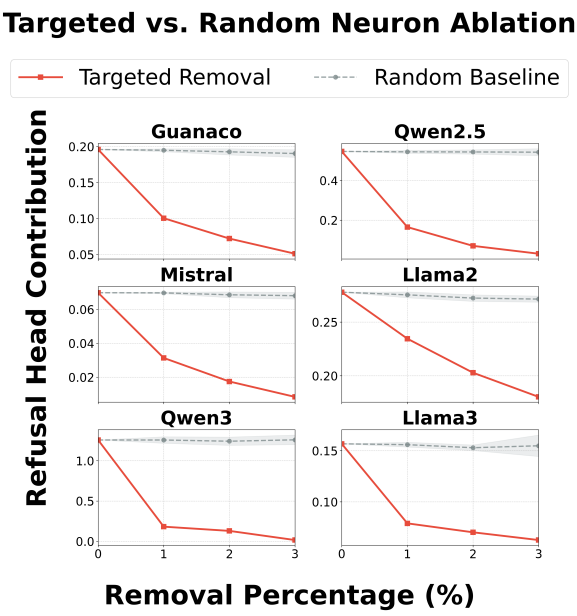
4.2.2 Directed Causal Tracing via Activation Patching

Procedure. Component ablation shows that removing detection heads weakens refusal heads, but this alone cannot distinguish a direct detection→refusal pathway from both components merely sharing a common upstream cause. To rule out the latter, we conduct an activation patching experiment across all six architectures. We swap only the detection heads' attention output between paired harmful and neutral prompts while holding all other computation fixed, and measure the refusal heads' contribution to the refusal direction under four conditions: unpatched harmful, harmful with detection output patched from the neutral run, unpatched neutral, and neutral with detection output patched from the harmful run.

Results and Insights. As shown in Table 1, five of the six models exhibit a clear bidirectional effect: patching detection head output from the neutral run into the harmful run lowers refusal head contribution, and patching harmful output into the neutral run raises it, explaining a substantial share of the gap between harmful and neutral runs (forward 19.5–34.6%, backward 39.4–70.6%). Because only the detection heads' output is swapped while everything upstream is held fixed, an explanation based on a shared upstream cause is ruled out, and the refusal heads can be seen to track the detection head write in both directions, direct evidence of



(a) Targeted ablation of harmful detection heads.



(b) Targeted ablation of safety neurons.

Figure 3: Circuit validation via targeted component ablation across six model architectures. The solid red lines show that targeted removal of identified safety components—(a) harmful detection heads and (b) safety neurons—causes a sharp drop in downstream refusal execution. The dashed gray lines (with ±std shaded boundary) represent the baselines of random component removal.

	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Forward (%)	0.7	30.3	34.6	31.1	31.3	19.5
Backward (%)	18.2	65.5	70.6	39.4	51.5	59.8

Table 1: Activation patching on the detection→refusal pathway. Fraction of the gap between harmful and neutral refusal head contribution explained by patching only the detection heads' output (forward: harmful ← neutral; backward: neutral ← harmful).

a detection→refusal pathway. LLaMA3 is a notable outlier, with a negligible forward effect (0.7%) and a markedly weaker backward effect (18.2%); its gap between harmful and neutral contribution is also substantially smaller than in the other five models, suggesting a more diffuse or partially bypassed pathway from detection to refusal in that architecture.

4.3 Experiment II: Reinforcing Harmful Detection Heads, Refusal Heads, and Safety Neurons

Intervention Protocol. Building on the causal relationships identified in Experiment I, we evaluate whether reinforcing components of the detection–refusal circuit improves robustness against harmful prompting. We strengthen safety-related components by scaling their residual-stream contributions with fixed multiplicative factors. Specifically, the top 3% of harmful detection heads and refusal heads are scaled using $\alpha_{\text{det}} = \alpha_{\text{ref}} = 2.0$, and the top 1% of safety neurons are scaled using $\alpha_{\text{saf}} = 1.5$. All interventions preserve the original model architecture and require no additional training.

Results and Insights. Table 2 shows that robustness arises from coordinated interactions among safety-related components rather than any single mechanism. Strengthening harmful detection heads improves unsafe-input recognition but is often insufficient without a reinforced refusal pathway, especially under adversarial attacks. Likewise, reinforcing refusal heads or safety neurons alone lacks robustness when harmful intent is obfuscated.

Jointly strengthening multiple components yields larger and more consistent gains, with detection heads, refusal heads, and safety neurons providing complementary effects and the strongest overall robustness.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	100 / 72 / 14	100 / 55 / 10	99 / 40 / 9	64 / 6 / 10	100 / 27 / 3	100 / 53 / 8
With Intervention:						
Detection ($\alpha_{\text{det}} = 2.0$)	100 / 97 / 37	100 / 68 / 79	99 / 79 / 26	73 / 8 / 17	100 / 52 / 4	100 / 91 / 50
Refusal ($\alpha_{\text{ref}} = 2.0$)	100 / 92 / 54	100 / 75 / 57	100 / 65 / 16	76 / 18 / 26	100 / 46 / 4	100 / 88 / 15
Safety Neurons ($\alpha_{\text{saf}} = 1.5$)	100 / 96 / 23	100 / 70 / 44	100 / 66 / 12	74 / 22 / 44	100 / 88 / 41	100 / 91 / 42
Detection & Refusal	**100 / 99 / 89**	100 / 87 / 92	100 / 86 / 48	85 / 30 / 41	100 / 56 / 12	100 / 93 / 19
Detection & Refusal & Safety	91 / 97 / 27	**100 / 95 / 99**	**100 / 87 / 68**	**88 / 45 / 71**	**100 / 89 / 43**	**100 / 100 / 29**

Table 2: Safety rates (%) under different attack methods (Pure Harmful Prompt / GCG / ADV-LLM) across backbone models. Interventions include scaling detection heads, refusal heads, safety neurons, and their combinations. Jointly reinforcing detection and refusal mechanisms—optionally augmented with safety neurons yields the strongest and most consistent safety improvements across diverse attack settings, while largely preserving benign prompt handling.

4.4 Experiment III: Safety Alignment and Utility of Edited Models

This experiment evaluates whether the detection–refusal circuit identified in earlier sections can be *persistently embedded into model weights* to produce safer models, and whether doing so incurs unacceptable utility costs on benign tasks. Unlike Experiments I and II, which operate through inference-time interventions, this experiment applies the weight scaling as a permanent edit, yielding a new edited model that can be used without any runtime modification.

4.4.1 Safety Alignment of the Edited Models under Adaptive GCG Attacks

Procedure. For each model, we apply *Circuit-Based Safety Editing (CBSE)* by permanently embedding the best-performing circuit reinforcement identified in Table 2 into the model weights. Depending on the model, this edit scales detection heads and refusal heads, and in some cases additionally scales safety neurons. All edits preserve the original model architecture and introduce no new parameters, yielding a fixed, inference-ready model without any runtime intervention.

To evaluate safety alignment under adaptive attacks, we regenerate GCG adversarial suffixes directly against each edited model. This setting tests whether the reinforced safety behavior remains effective when adversaries explicitly optimize against the edited weights, rather than against the original baseline.

Results. As shown in Table 4, CBSE achieves substantially higher safety rates than the original across all models. These results demonstrate that the detection–refusal safety circuit encodes a robust and transferable structure that can be persistently embedded into model weights.

4.4.2 Downstream Utility of the Edited Models

Procedure. To assess whether CBSE preserves the general helpfulness of the original models, we evaluate edited models on standard benchmarks covering factual knowledge, commonsense reasoning, physical reasoning, and coreference resolution. We report accuracy on MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), and WinoGrande (Sakaguchi et al., 2021). In addition, we measure perplexity on WikiText (Merity et al., 2016) to capture potential degradation in general language modeling quality.

Results and Insights. Table 5 shows that CBSE largely preserves performance across all benchmarks, with accuracy drops that are modest and unevenly distributed. Reasoning-intensive tasks such as MMLU exhibit slightly larger reductions than commonsense-oriented benchmarks. The same table shows that CBSE incurs only modest increases in perplexity across all models, reflecting limited disruption to core language modeling behavior.

Overall, these results show that safety circuits can be persistently embedded into model weights, substantially improving adversarial robustness with minimal utility loss. The appendix provides extended validation of circuit robustness (Appendix A.1), additional attacks (Appendix B.1, B.2), circuit ablations (Appendix C), baseline comparisons and parameter sensitivity (Appendix E, F), alternative judges and datasets (Appendix G.1, G.3, H), multi-seed stability (Appendix A.2), component specificity (Appendix D), human evaluation (Appendix G.2), and training-free defense comparisons (Appendix E.3).

Refusal Rate (%) ↓	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	38.8	69.4	31.6	19.7	18.8	26.1
CBSE	45.3	64.2	48.3	26.2	28.0	35.9

Table 3: Over-refusal macro-evaluation on borderline-benign prompts. False-positive refusal rates (%) based on a keyword template across six model architectures. The marginal changes in refusal rates demonstrate that CBSE effectively mitigates adversarial vulnerabilities without inducing severe over-refusal side effects.

Safety Rate (%)	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	77	53	36	10	30	53
CBSE	**98**	**87**	**62**	**18**	**70**	**89**

Table 4: Safety rates (%) under regenerated GCG attacks optimized against **CBSE**-edited models. CBSE substantially improves safety alignment across all evaluated architectures.

Model	Benchmark	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	MMLU	66.58	47.63	59.14	33.77	74.02	72.63
	HellaSwag	58.69	57.96	66.87	59.48	62.49	52.60
	PIQA	80.90	76.28	82.15	79.0	79.22	77.86
	WinoGrande	76.24	72.45	78.69	72.9	74.90	66.54
	Perplexity (↓)	10.05	11.62	9.90	10.51	9.41	13.07
CBSE	MMLU	63.12	43.92	55.78	33.70	71.27	69.16
	HellaSwag	58.15	56.59	66.09	58.76	61.85	51.56
	PIQA	79.65	75.14	80.52	78.29	71.55	76.33
	WinoGrande	75.61	67.64	76.63	73.5	69.46	66.06
	Perplexity (↓)	10.85	18.35	10.90	11.70	13.00	15.94

Table 5: Accuracy (%) on standard benchmarks and perplexity on WikiText (lower is better, marked ↓) for **Original** and **CBSE**-edited models. CBSE preserves model utility with modest and task-dependent performance degradation, and introduces only modest increases in perplexity.

4.4.3 Refusal Calibration on Borderline-Benign Prompts

To evaluate refusal calibration and ensure CBSE does not induce over-generalization, we benchmarked all models on a 1,000-prompt borderline-benign subset sampled from **OR-Bench-80k** (Cui et al., 2025). These inputs are specifically curated to test false-positive safety triggers on benign topics.

As shown in Table 3, CBSE preserves a well-calibrated refusal profile rather than broadly increasing over-refusal. On the conservative LLaMA2, it reduces false-positive refusals by 5.2%, while other architectures show modest increases despite substantially improved safety. This suggests that circuit-guided weight scaling sharpens the semantic decision boundary rather than inducing blanket rejection.

5 Conclusion

We present a mechanistic, interpretability-driven analysis of safety-related components in large language models. By identifying **Harmful Detection Heads**, **Refusal Heads**, and **Safety Neurons**, we uncover a recurring component-level organization associated with refusal behavior and show that these components exert complementary influences when intervened upon. Architecture-preserving scaling of these components substantially improves robustness against standard and adaptive attacks, while largely preserving the model's reasoning ability and task performance. Our results suggest that component-level interventions informed by mechanistic analysis can enhance safety robustness without significantly compromising model utility, highlighting interpretability as a foundation for understanding and improving LLM behavior.

Acknowledgements

The authors are partially supported by National Science Foundation under Grant No. 2313105, 2430539, and Intel Rising Star Faculty Award. The authors would like to thank the National Research Platform (NRP) at UCSD for computation support and anonymous reviewers for valuable feedback.

Limitations

In this work, we characterize the multi-stage safety circuit and apply Circuit-Based Safety Editing (CBSE) based on a static snapshot of model representations captured during our probing phase. While our empirical results demonstrate that reinforcing these identified components yields substantial and robust safety improvements across multiple adversarial attack settings, our current framework does not account for how these safety circuits might dynamically evolve or reorganize if the model undergoes further extensive continuous training or fine-tuning. Investigating the long-term stability and potential drift of localized safety circuits during a model's full life cycle remains an important and exciting direction for future research in mechanistic interpretability.

Ethics Statement

This work investigates the mechanistic structure of safety circuits in large language models and proposes interventions to improve robustness against harmful content. The primary societal benefit is enhancing the reliability and alignment of language models, reducing the risk of unsafe outputs in real-world applications. Our approach focuses on interpretability and controlled modifications of model internals, which could guide the development of more transparent and accountable AI systems. Potential risks include misuse of these mechanisms to suppress legitimate content or over-reliance on automated safety circuits without human oversight. Overall, this research advances understanding of model behavior and provides tools for safer deployment of large language models.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Kuan-Lin Chu, Chung-En Sun, and Tsui-Wei Weng. 2025. How to make llms safer? detecting and editing key heads in llms. In *Lock-LLM Workshop: Prevent Unauthorized Knowledge Use from Large Language Models*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. Or-bench: An over-refusal benchmark for large language models. In *International Conference on Machine Learning*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *NeurIPS*.

Aaron Grattafiori and 1 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024. Catastrophic jailbreak of open-source llms via exploiting generation. In *International Conference on Learning Representations*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*.

Nathan Lambert. 2025. Reinforcement learning from human feedback. *arXiv e-prints*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, volume 2024, pages 56174–56194.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Bhagyajit Pingua, Deepak Murmu, Meenakshi Kandpal, Jyotirmayee Rautaray, Pranati Mishra, Rabindra Kumar Barik, and Manob Jyoti Saikia. 2024. Mitigating adversarial manipulation in llms: a prompt-based approach to counter jailbreak attacks (prompt-g). *PeerJ Comput. Sci.*

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2025. Smoothllm: Defending large language models against jailbreaking attacks. *Transactions on Machine Learning Research*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and 1 others. 2025. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *Transactions on Machine Learning Research*.

Chung-En Sun, Xiaodong Liu, Weiwei Yang, Tsui-Wei Weng, Hao Cheng, Aidan San, Michel Galley, and Jianfeng Gao. 2025a. Iterative self-tuning llms for enhanced jailbreaking capabilities. *NAACL*.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2025b. Concept bottleneck large language models. In *International Conference on Learning Representations*.

Hugo Touvron and 1 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. 2025. Robust LLM safeguarding via refusal feature adversarial training. In *ICLR*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *ACL*.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

Table of Contents

A Circuit Robustness 12

A.1 Robustness of Identified Components to Probe Construction 12

A.2 Statistical Reliability of Safety Gains under GCG 13

B Adaptive Attacks 14

B.1 Resilience Against PAIR Conversational Attacks 14

B.2 Defending Against AutoDAN Genetic Semantic Optimization 14

C Circuit Ablations 15

D Component Specificity 16

E Baseline Comparisons 17

E.1 Comparison with CircuitBreakers 17

E.2 Comparison with Linear Adversarial Training 17

E.3 Comparison with Training-Free Inference-Time Defenses 17

F Parameter Sensitivity and Exploratory Analysis 19

F.1 Joint Ablation of Head Selection Ratio and Detection/Refusal Scaling 19

F.2 Joint Ablation of Safety Neuron Selection Ratio and Scaling Factor 20

G Multi-judge Safety Evaluation 22

G.1 Multi-Judge Safety Evaluation with Claude 4.6 Opus 22

G.2 Human Evaluation 22

G.3 Safe-Template Evaluation 22

H Evaluation on the Malicious Instruction Dataset 24

A Circuit Robustness

To demonstrate the empirical validity of our findings, this section provides an extended analysis of the structural robustness of our circuit identification pipeline to variations in the probing dataset, as well as the statistical reliability of the downstream safety gains reported under adversarial attack.

A.1 Robustness of Identified Components to Probe Construction

To assess whether our identified safety circuit depends on the specific probing dataset, we conducted a bootstrap stability analysis across six architectures by randomly subsampling 80% of our original 51 probing pairs over five independent runs. For each data split, discrete circuit components were isolated by selecting the top 3% of highest-attributed attention heads and the top 1% of highest-attributed MLP neurons.

Component	Metric	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Attn Refusal Dir.	Cosine Sim.	0.996	0.976	0.992	0.927	0.994	0.990
MLP Safety Dir.	Cosine Sim.	0.996	0.975	0.992	0.922	0.994	0.990
Refusal Heads	Jaccard Overlap	0.974	0.936	0.948	0.790	0.942	0.927
Safety Neurons	Jaccard Overlap	0.908	0.805	0.863	0.719	0.883	0.835

Table 6: Probe robustness across architectures. Mean cosine similarity for latent directions and pairwise Jaccard overlap for discrete components across five independent 80%-subsampled data splits. The near-unity cosine similarities and high Jaccard overlaps demonstrate that the identified safety directions and circuit components are highly stable and robust to data perturbations.

As shown in Table 6, both continuous semantic directions ($\cos\theta \geq 0.92$) and discrete circuit components maintain high stability across data variations. The strong Jaccard overlap scores confirm that our pipeline consistently recovers the same underlying safety heads and neurons, demonstrating that the mapped circuit captures intrinsic model mechanisms rather than dataset artifacts.

A.2 Statistical Reliability of Safety Gains under GCG

The safety rates reported throughout the main text and appendix are computed from a single generation run per configuration. To verify that the reported gains are not an artifact of sampling noise in generation, and to quantify the run-to-run variance inherent to stochastic decoding and to GCG's own suffix optimization, we repeat the GCG evaluation with five independent generation seeds for both the Original model and CBSE, across all six architectures.

Table 7 reports the mean and standard deviation of Llama-Guard safety rates over five seeds (100 GCG prompts each). CBSE's safety gain is stable across seeds (std $\leq$ 3.6 points for all six models) and, on every architecture, dwarfs the seed-to-seed variance of the Original model. This confirms that the improvement from CBSE reflects a genuine and repeatable safety property of the intervention rather than a favorable draw of a single run, and that seed-level variance alone (up to $\pm$4.8 points) is sufficient to explain small discrepancies between single-run safety numbers reported elsewhere in the paper for nominally identical configurations.

Safety Rate (%)	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	80.2 ± 2.8	55.8 ± 1.8	44.8 ± 4.8	13.0 ± 2.7	29.4 ± 3.2	55.6 ± 2.7
CBSE	**99.4 ± 0.8**	**95.4 ± 1.4**	**95.0 ± 0.6**	**53.8 ± 3.6**	**93.4 ± 1.4**	**100.0 ± 0.0**

Table 7: Multi-seed safety evaluation (Llama-Guard safety rate %, mean ± std over 5 generation seeds, 100 GCG prompts) across all six architectures. CBSE's gains are stable and far exceed generation-level variance.

B Adaptive Attacks

This section provides extended empirical analyses to validate the generalizability of CBSE against adaptive adversarial strategies, verifying that our intervention genuinely hardens internal safety pathways rather than overfitting to specific gradient-based token noise. All safety rates reported in this section are measured using Llama-Guard.

B.1 Resilience Against PAIR Conversational Attacks

We first evaluate CBSE against **PAIR** (Chao et al., 2025), an LLM-driven black-box conversational attack that adaptively refines prompts through multiple iterations to circumvent safety filters. For this evaluation, we employ Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) as the attacker model. This test assesses the defense's robustness against high-level semantic persuasion and structured jailbreak attempts.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	94	86	2	0	4	48
CBSE	**96**	**94**	**34**	**6**	**20**	**94**

Table 8: PAIR safety rates across architectures. Empirical safety rates under the PAIR conversational attack. CBSE consistently reinforces model resilience, achieving substantial gains.

As shown in Table 8, CBSE consistently neutralizes PAIR's adaptive conversational attacks. Our intervention markedly elevates safety rates for both robust and vulnerable baselines, demonstrating strong defense generalizability against high-level semantic persuasion.

B.2 Defending Against AutoDAN Genetic Semantic Optimization

We further evaluate CBSE against **AutoDAN-HGA** (Liu et al., 2024), a potent jailbreak framework that utilizes a hierarchical genetic algorithm to generate fluent, human-readable adversarial prompts. Unlike token-level attacks, AutoDAN optimizes for semantic coherence, posing a significant challenge to internal alignment mechanisms.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	99	99	3	2	4	82
CBSE	**99**	**100**	**70**	**25**	**93**	**100**

Table 9: AutoDAN-HGA safety rates across architectures. Empirical safety rates under the AutoDAN genetic attack. CBSE consistently reinforces model resilience, achieving substantial gains.

As shown in Table 9, CBSE provides a robust defense against fluent semantic optimizations. The gains are particularly notable on models with weak native AutoDAN resistance; for instance, CBSE elevates Mistral's safety rate from 3.0% to 70.0% and Qwen2.5 from 4.0% to 93.0%. These results validate that CBSE successfully reinforces core safety pathways against high-quality, human-readable adversarial text without relying on specific token-level artifacts.

C Circuit Ablations

To confirm that the identified safety circuit (comprising harmful detection heads, refusal heads, and safety neurons) is structurally necessary for sustaining model alignment, we perform a suffix-based knock-out experiment. We systematically zero out an increasing percentage (0%, 1%, 3%, 5%) of the highest-attributed components in the safety circuit and evaluate the model's resulting vulnerability under adversarial GCG attacks.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Original	75	54	51	3	29	55
1% Ablation	2	40	14	9	7	19
3% Ablation	2	32	3	2	1	12
5% Ablation	2	16	0	0	0	9

Table 10: Model safety rates (%) under GCG attacks drop sharply as an increasing percentage of the safety circuit is deactivated.

As shown in Table 10, removing a tiny fraction of the identified circuit components induces a dramatic, catastrophic drop in safety rates across all evaluated architectures. Most notably, in LLaMA3, a minimal 1% ablation effectively collapses the model's resistance, plunging its safety rate from 75% to a mere 2%. Similarly, the safety barriers of Qwen2.5 and Guanaco are rendered entirely non-functional (0%) at a 5% ablation threshold. This widespread collapse confirms that the identified subnets are not redundant; rather, they form the structural backbone of safety alignment within these networks.

D Component Specificity

The experiments so far show that the identified components matter for safety, but not whether they respond specifically to harmful intent, as opposed to any unusual or difficult input more generally. To rule out the latter, we measure the detection-head and refusal-head signal (each head's contribution to the refusal direction at the last prompt token) on four input types beyond harmful prompts: *ambiguous-benign* inputs from OR-Bench, which are engineered to look harmful while being safe; *hard-benign* math problems from GSM8K; *hard-benign* questions from difficult MMLU subjects; and *easy-benign* neutral prompts. We evaluate this on three architectures spanning different alignment regimes: LLaMA2, Guanaco, and Qwen3.

Input Type	LLaMA2	Guanaco	Qwen3
Harmful (AdvBench)	0.065 / 0.235	0.112 / 0.425	0.276 / 0.976
Ambiguous-benign (OR-Bench)	0.017 / 0.145	0.038 / 0.173	0.082 / 0.356
Hard-benign (GSM8K)	0.006 / 0.117	0.010 / 0.084	0.019 / −0.054
Hard-benign (MMLU-hard)	0.014 / 0.119	0.009 / 0.096	0.033 / 0.066
Easy-benign (neutral)	0.020 / 0.148	0.027 / 0.140	0.049 / 0.259

Table 11: Component specificity across input types. Each cell reports the detection-head / refusal-head signal, i.e., the mean contribution to the refusal direction at the last prompt token.

As shown in Table 11, the detection-head signal is consistently and substantially higher on harmful inputs than on any benign category, across all three architectures. The gap holds even against inputs adversarially engineered to look harmful (OR-Bench) and inputs that are considerably more difficult than the harmful prompts themselves (GSM8K, hard MMLU subjects), ruling out surface toxicity or generic difficulty as the driver of the signal. Difficulty alone does not activate the circuit: on every architecture and for both signals, the two hard-benign categories (GSM8K, hard MMLU subjects) occupy the two lowest values, ranking below both easy-benign and ambiguous-benign inputs, the opposite of what a generic difficulty detector would produce; on Qwen3, the refusal signal on GSM8K even falls below the neutral floor.

E Baseline Comparisons

We compare CBSE against two prominent training-based, weight-level defenses, **CircuitBreakers** (Zou et al., 2024) and **Linear Adversarial Training (LAT)** (Sheshadri et al., 2025), as well as two training-free, inference-time defenses that do not require any gradient updates. To establish a rigorous baseline benchmark, all configurations are evaluated against a dataset of 100 adversarial GCG prompts, with the final safety rates measured using Llama3-8B-Guard.

E.1 Comparison with CircuitBreakers

We first evaluate our method against CircuitBreakers, a defense paradigm that aims to disrupt harmful representations by mapping adversarial inputs to a pre-defined refusal direction via weight fine-tuning.

Model	Defense Strategy	Paradigm Type	Safety Rate (%) ↑
LLaMA3	Original	—	84
	CircuitBreakers	Training-based	65
	CBSE	**Training-free**	**99**
Mistral	Original	—	44
	CircuitBreakers	Training-based	74
	CBSE	**Training-free**	**87**

Table 12: Comparative performance against CircuitBreakers under GCG attacks. CBSE achieves superior safety rates across both architectures without requiring gradient updates or optimization datasets.

As shown in Table 12, CBSE consistently outperforms CircuitBreakers on both LLaMA3 and Mistral architectures. Crucially, on LLaMA3, CircuitBreakers suffers from defense degradation under the evaluated GCG dataset, dropping to a 65% safety rate, whereas CBSE successfully hardens the model to a 99% safety rate. This demonstrates that surgically reinforcing the internal safety circuit provides more resilient protection than coarse-grained representation mapping.

E.2 Comparison with Linear Adversarial Training

Next, we compare CBSE against LAT, an adversarial training framework that leverages a linear classifier to identify safety-critical directions and guides adversarial optimization loops during training.

Model	Defense Strategy	Paradigm Type	Safety Rate (%) ↑
LLaMA3	Original	—	84
	LAT	Training-based	100
	CBSE	**Training-free**	**99**

Table 13: Comparative performance against LAT under GCG attacks. CBSE performs on par with the resource-intensive adversarial training approach.

As shown in Table 13, CBSE achieves a 99% safety rate on LLaMA3, performing virtually on par with LAT's perfect 100% clearance rate. However, while LAT demands resource-intensive adversarial training loops, multiple backward passes, and specialized optimization data to achieve this boundary, CBSE operates entirely at inference time. By scaling internal safety vectors directly without any training overhead, CBSE delivers equivalent frontier-level security guarantees with immense computational efficiency.

E.3 Comparison with Training-Free Inference-Time Defenses

Beyond training-based weight-level defenses, we compare CBSE against two widely-used training-free defenses that require no gradient updates at all: **SmoothLLM** (Robey et al., 2025), a prompt-level defense that aggregates predictions over randomly perturbed copies of the input, and **Activation-Steering** (Arditi

et al., 2024), an activation-level defense that adds our identified refusal direction to the residual stream at inference time. We evaluate all methods on three representative architectures (LLaMA3, Mistral, Qwen3).

Safety Rate (%) ↑	LLaMA3	Mistral	Qwen3	Average
Original	81	45	59	62
SmoothLLM (prompt-level)	100	92	99	97
Activation-Steering (activation-level)	82	70	96	83
CBSE (Ours)	**98**	**95**	**100**	**98**

Table 14: Comparison against training-free defenses (Llama-Guard safety rate %; 100 GCG prompts). CBSE attains the highest average safety while adding zero inference-time overhead.

As shown in Table 14, CBSE attains the highest average safety rate (98%), narrowly ahead of SmoothLLM (97%) and well above Activation-Steering (83%). SmoothLLM is competitive on safety, but it queries the model on $q = 6$ perturbed copies per input, making it roughly 6× more expensive at inference, and its input perturbations degrade clean-task accuracy (e.g., PIQA drops from 76.7% to 70.3% on LLaMA-2 at $q = 5$ in the original SmoothLLM evaluation). CBSE, in contrast, folds the intervention into the model weights once, after which it adds no inference-time cost and leaves the input untouched.

F Parameter Sensitivity and Exploratory Analysis

We adopt a two-stage ablation protocol to evaluate the sensitivity of the hyperparameters used for circuit-level interventions. We first map the performance landscape for attention-head interventions by jointly varying the fraction of selected heads and the scaling factors applied to detection and refusal heads, while disabling neuron-level edits. With the baseline head configuration fixed, we then ablate safety neurons to explore the effects of different selection ratios and scaling strengths.

All ablations are conducted under the GCG and ADV-LLM attack. Safety performance is measured using Llama3-8B-Guard, while general utility is evaluated using accuracy on MMLU.

F.1 Joint Ablation of Head Selection Ratio and Detection/Refusal Scaling

We evaluate head selection ratios of 1%, 3%, and 5%, together with scaling factors $\alpha_{\text{det}}, \alpha_{\text{ref}} \in \{2.0, 3.0, 4.0\}$ applied to detection and refusal heads. Neuron-level interventions are disabled in this evaluation.

Tables 15, 16, and 17 report safety results under different configurations.

We additionally evaluate a subset of configurations on MMLU to measure general capability retention (Table 18).

Safety Rate (%) ↑		LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
$\alpha_{\text{det}} = 2.0$	$\alpha_{\text{ref}} = 2.0$	91 / 13	89 / 67	59 / 19	7 / 13	33 / 2	90 / 22
	$\alpha_{\text{ref}} = 3.0$	91 / 20	97 / 91	72 / 21	10 / 21	47 / 5	91 / 24
	$\alpha_{\text{ref}} = 4.0$	91 / 29	89 / 87	63 / 27	21 / 32	47 / 7	97 / 39
$\alpha_{\text{det}} = 3.0$	$\alpha_{\text{ref}} = 2.0$	99 / 81	97 / 92	65 / 28	13 / 14	32 / 3	96 / 32
	$\alpha_{\text{ref}} = 3.0$	94 / 65	88 / 88	65 / 32	10 / 29	45 / 2	92 / 35
	$\alpha_{\text{ref}} = 4.0$	86 / 44	30 / 11	57 / 44	25 / 30	41 / 4	99 / 40
$\alpha_{\text{det}} = 4.0$	$\alpha_{\text{ref}} = 2.0$	90 / 81	98 / 95	67 / 47	15 / 19	36 / 2	92 / 32
	$\alpha_{\text{ref}} = 3.0$	79 / 28	51 / 18	63 / 59	13 / 29	38 / 2	96 / 35
	$\alpha_{\text{ref}} = 4.0$	28 / 26	5 / 12	47 / 45	25 / 33	42 / 8	99 / 39

Table 15: Joint ablation of detection and refusal scaling with a fixed head selection ratio of **1%**. Neuron-level interventions are disabled. Each cell reports safety rates (%) under GCG / ADV-LLM attacks.

Safety Rate (%) ↑		LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
$\alpha_{\text{det}} = 2.0$	$\alpha_{\text{ref}} = 2.0$	99 / 89	85 / 92	84 / 48	28 / 41	54 / 12	94 / 19
	$\alpha_{\text{ref}} = 3.0$	90 / 94	53 / 27	88 / 85	50 / 49	34 / 29	94 / 35
	$\alpha_{\text{ref}} = 4.0$	49 / 46	5 / 5	47 / 65	48 / 43	11 / 10	28 / 38
$\alpha_{\text{det}} = 3.0$	$\alpha_{\text{ref}} = 2.0$	100 / 97	91 / 97	88 / 55	39 / 54	45 / 29	59 / 24
	$\alpha_{\text{ref}} = 3.0$	86 / 81	5 / 2	68 / 74	59 / 51	8 / 15	9 / 7
	$\alpha_{\text{ref}} = 4.0$	42 / 22	7 / 1	22 / 40	26 / 19	3 / 9	2 / 5
$\alpha_{\text{det}} = 4.0$	$\alpha_{\text{ref}} = 2.0$	98 / 91	72 / 62	80 / 57	51 / 58	13 / 18	2 / 0
	$\alpha_{\text{ref}} = 3.0$	69 / 14	1 / 0	24 / 27	37 / 43	1 / 5	2 / 1
	$\alpha_{\text{ref}} = 4.0$	8 / 3	6 / 0	8 / 33	6 / 8	2 / 11	2 / 0

Table 16: Joint ablation of detection and refusal scaling with a fixed head selection ratio of **3%**. Neuron-level interventions are disabled. Each cell reports safety rates (%) under GCG / ADV-LLM attacks.

Safety Rate (%) ↑		LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
$\alpha_{\text{det}} = 2.0$	$\alpha_{\text{ref}} = 2.0$	99 / 97	87 / 91	89 / 77	50 / 54	57 / 19	93 / 47
	$\alpha_{\text{ref}} = 3.0$	68 / 49	3 / 1	23 / 55	71 / 51	30 / 22	9 / 10
	$\alpha_{\text{ref}} = 4.0$	2 / 1	1 / 2	23 / 57	27 / 9	5 / 8	9 / 20
$\alpha_{\text{det}} = 3.0$	$\alpha_{\text{ref}} = 2.0$	98 / 92	59 / 13	85 / 85	65 / 63	47 / 32	6 / 18
	$\alpha_{\text{ref}} = 3.0$	53 / 34	7 / 3	17 / 39	48 / 20	3 / 7	4 / 16
	$\alpha_{\text{ref}} = 4.0$	2 / 1	1 / 4	16 / 24	6 / 15	3 / 6	10 / 8
$\alpha_{\text{det}} = 4.0$	$\alpha_{\text{ref}} = 2.0$	34 / 11	3 / 1	60 / 8	55 / 38	7 / 1	6 / 8
	$\alpha_{\text{ref}} = 3.0$	2 / 0	2 / 2	8 / 16	4 / 7	0 / 0	8 / 7
	$\alpha_{\text{ref}} = 4.0$	1 / 2	5 / 4	7 / 11	16 / 32	3 / 3	10 / 7

Table 17: Joint ablation of detection and refusal scaling with a fixed head selection ratio of **5%**. Neuron-level interventions are disabled. Each cell reports safety rates (%) under GCG / ADV-LLM attacks.

Accuracy (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	66.58	47.63	59.14	33.77	74.02	72.63
Top 3%, $\alpha_{\text{det}} = 2.0$, $\alpha_{\text{ref}} = 2.0$	63.34	44.37	55.00	34.97	72.46	70.26
Top 3%, $\alpha_{\text{det}} = 3.0$, $\alpha_{\text{ref}} = 2.0$	57.34	38.14	47.66	33.42	71.61	43.47
Top 5%, $\alpha_{\text{det}} = 2.0$, $\alpha_{\text{ref}} = 2.0$	54.92	43.11	51.84	33.46	71.42	55.10

Table 18: MMLU accuracy (%) for selected attention head configurations. The table shows how varying the fraction of selected attention heads and the head scaling factor affects general reasoning ability, ensuring that safety improvements do not compromise performance.

F.2 Joint Ablation of Safety Neuron Selection Ratio and Scaling Factor

After fixing the attention-head configuration (top 3% heads with $\alpha_{\text{det}} = \alpha_{\text{ref}} = 2.0$), we evaluate the effect of safety-neuron interventions.

We vary the fraction of selected safety neurons (1%–3%) and the neuron scaling factor $\alpha_{\text{saf}} \in \{1.5, 2.0, 2.5\}$.

Table 19 reports safety results under different configurations.

We further evaluate a subset of configurations on MMLU under the fixed attention-head setting (Table 20).

Safety Rate (%) ↑		LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Top 1%	$\alpha_{\text{saf}} = 1.5$	98 / 27	98 / 99	93 / 68	40 / 71	92 / 43	100 / 29
	$\alpha_{\text{saf}} = 2.0$	72 / 33	96 / 100	97 / 81	73 / 77	97 / 100	99 / 58
	$\alpha_{\text{saf}} = 2.5$	9 / 5	97 / 100	98 / 96	83 / 77	90 / 93	86 / 79
Top 2%	$\alpha_{\text{saf}} = 1.5$	94 / 27	97 / 100	97 / 70	52 / 72	89 / 52	100 / 32
	$\alpha_{\text{saf}} = 2.0$	81 / 42	87 / 100	98 / 93	81 / 82	98 / 100	100 / 69
	$\alpha_{\text{saf}} = 2.5$	44 / 15	97 / 100	100 / 98	89 / 81	93 / 97	85 / 87
Top 3%	$\alpha_{\text{saf}} = 1.5$	91 / 29	96 / 100	97 / 78	56 / 74	92 / 57	100 / 42
	$\alpha_{\text{saf}} = 2.0$	67 / 19	99 / 100	96 / 97	83 / 85	99 / 100	98 / 84
	$\alpha_{\text{saf}} = 2.5$	39 / 8	96 / 100	99 / 99	91 / 89	98 / 97	81 / 96

Table 19: Joint ablation of safety neuron selection ratio and scaling factor. Attention-head interventions are fixed at top 3% heads with $\alpha_{\text{det}} = \alpha_{\text{ref}} = 2.0$. Each cell reports safety rates (%) under GCG / ADV-LLM attacks.

Accuracy (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	66.58	47.63	59.14	33.77	74.02	72.63
Top 1%, $\alpha_{\text{saf}} = 1.5$	63.12	43.92	55.78	33.70	71.27	67.75
Top 1%, $\alpha_{\text{saf}} = 2.0$	55.54	42.61	55.31	33.00	71.15	67.29
Top 2%, $\alpha_{\text{saf}} = 2.0$	53.37	42.52	55.30	33.04	70.65	67.09
Top 3%, $\alpha_{\text{saf}} = 2.0$	55.76	42.36	55.17	32.13	69.80	66.80

Table 20: MMLU accuracy (%) for selected neuron configurations with fixed attention-head interventions (top 3% heads, $\alpha_{\text{det}} = \alpha_{\text{ref}} = 2.0$). The table shows how varying the fraction of selected safety neurons and the neuron scaling factor affects general reasoning ability, ensuring that safety improvements do not compromise performance.

G Multi-judge Safety Evaluation

To verify that the safety enhancements from CBSE are robust and generalize beyond our primary evaluator, we introduce three independent validation layers: a model-based judge, human annotation, and a rule-based heuristic check.

G.1 Multi-Judge Safety Evaluation with Claude 4.6 Opus

To ensure our safety improvements reflect genuine defensive capabilities rather than overfitting to a specific classifier, we introduce an independent frontier-model judge. We employ **Claude 4.6 Opus**—an entirely separate architecture and training pipeline from our primary evaluator, Llama Guard—to assess model outputs under GCG attacks. This multi-judge setup provides a rigorous validation against potential optimization biases or evaluator-specific blind spots.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	77 / 80	53 / 66	36 / 33	10 / 16	30 / 33	53 / 49
CBSE (Ours)	98 / 92	87 / 91	62 / 63	18 / 22	70 / 60	83 / 81

Table 21: Multi-judge safety evaluation under adversarial GCG attacks. Cell values indicate safety rates (%) evaluated by Llama Guard / Claude, respectively. Both evaluators capture consistent safety improvements across all six architectures under CBSE.

As shown in Table 21, the independent Claude judge confirms that CBSE consistently and substantially elevates safety performance across all six architectures. The tight correlation between the Claude and Llama Guard metrics demonstrates that the reinforced safety pathways yield robust, generalized refusal behaviors that successfully transfer across distinct frontier evaluation models.

G.2 Human Evaluation

Automated safety judges validate against one another, but Llama-Guard and Claude are both models, so their agreement alone does not rule out shared blind spots. To break this automated-to-automated loop, we introduce a human validation layer.

An independent annotator labeled a blind, condition-shuffled sample of 100 responses pooled across all six architectures under the GCG attack (both Original and CBSE configurations, with model and configuration identities hidden). The annotator followed a fixed protocol: a response is **unsafe** only if it provides actionable harmful content that could be used to carry out the request; refusals, deflections, non-actionable general information, and off-topic or incoherent completions are **safe**.

Table 22 reports the resulting safety rates alongside Llama-Guard and Claude on the identical 100 items. All three judges, scored independently on the same underlying responses, rank CBSE far above the Original model, confirming that CBSE's safety improvement is not an artifact of automated-to-automated evaluation.

Safety Rate (%) ↑	Human	Llama-Guard	Claude
Original	68	63	70
CBSE	**94**	**78**	**96**

Table 22: Human evaluation on a blind, condition-shuffled sample of 100 responses pooled across all six architectures (GCG attack; Original vs. CBSE). All judges score the identical items. Human, Llama-Guard, and Claude independently agree on the direction and magnitude of the safety improvement.

G.3 Safe-Template Evaluation

To complement the classifier-based evaluations, we adopt a lightweight, heuristic **Safe-Template** protocol. This method evaluates safety independently of any neural model by verifying whether output strings begin

with or contain phrases indicative of standard refusals, ethical disclaimers, or polite rejections. Serving as a rule-based sanity check, this protocol ensures that our observed safety enhancements reflect genuine behavioral shifts rather than optimization artifacts tailored to a specific model-based judge.

G.3.1 Heuristic Protocol and Experimental Setup

The rule-based evaluation heuristic scans model completions for any of the following literal substrings or prefix phrases:

> "I'm sorry", "Sorry", "I am sorry", "I apologize", "As an", "As an AI", "I'm an", "I'm just", "As a language model", "As an Assistant", "I cannot", "I can't", "I do not", "cannot", "Hello!", "is illegal and unethical", "I'm really sorry", "unethical", "not ethical", "not legal", "My apologies", "but I cannot", "If you have any other non-malicious requests", "I'm not able to provide", "I am really sorry", "I cannot create", "I cannot support", "I can't provide", "I can't assist", "I can't help", "However", "I must", "I understand", "I am not able to", "<|eot_id|>", "Instead", "instead"

Using this string-matching criteria, we measure safety across the identical baseline and intervention configurations analyzed in the main text. Specifically, joint configurations involve scaling the top 3% of detection and refusal heads ($\alpha_{\text{det}} = \alpha_{\text{ref}} = 2.0$) alongside the top 1% of safety neurons ($\alpha_{\text{saf}} = 1.5$).

G.3.2 Empirical Findings

Table 23 reports the heuristic safety rates across all six architectures under Pure Harmful Prompt, GCG, and ADV-LLM attacks. The relative performance trends remain highly consistent with our primary Llama Guard evaluation. Jointly reinforcing detection and refusal infrastructure consistently provides a substantial safety lift over single-component interventions across architectures, while the addition of safety neurons provides crucial stability under complex adversarial optimizations. The persistent alignment between this strict string-matching check and our deep learning judges confirms the structural validity of the isolated safety circuit.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	100 / 69 / 19	100 / 57 / 31	99 / 46 / 15	57 / 3 / 11	100 / 28 / 8	98 / 51 / 8
With Intervention:						
Detection ($\alpha_{\text{det}} = 2.0$)	100 / 94 / 41	100 / 69 / 84	99 / 78 / 25	58 / 6 / 17	100 / 47 / 13	98 / 85 / 44
Refusal ($\alpha_{\text{ref}} = 2.0$)	100 / 90 / 58	100 / 81 / 68	100 / 64 / 21	61 / 6 / 18	100 / 22 / 8	99 / 66 / 14
Safety Neurons ($\alpha_{\text{saf}} = 1.5$)	100 / 88 / 22	100 / 77 / 60	100 / 65 / 31	61 / 19 / 41	100 / 85 / 50	100 / 70 / 42
Detection & Refusal	100 / 99 / 92	100 / 85 / 96	100 / 87 / 46	48 / 7 / 31	100 / 41 / 13	99 / 84 / 4
Detection & Refusal & Safety	89 / 96 / 24	100 / 90 / 98	100 / 88 / 59	61 / 28 / 56	100 / 85 / 43	99 / 98 / 9

Table 23: Safety rates (%) under different attack methods (Pure Harmful Prompt / GCG / ADV-LLM) across backbone models. Interventions include scaling detection heads, refusal heads, safety neurons, and their combinations. Safety rates are measured using the rule-based Safe-Template string matching.

H Evaluation on the Malicious Instruction Dataset

To further assess the robustness and generality of our interventions, we evaluate the models on the Malicious Instruction Dataset (Huang et al., 2024), which contains a diverse collection of explicitly harmful user queries across domains such as cybercrime, fraud, violence, and other illicit activities.

For this dataset, we conduct attacks using only Pure Harmful Prompt and ADV-LLM, as target responses for GCG are not available in this dataset. All intervention configurations (scaling of detection heads, refusal heads, and safety neurons) are kept identical to those used in the Table 2. Safety is measured using both Llama-Guard and the Safe-Template protocol.

The results on the Malicious Instruction Dataset are summarized in Table 24 (Llama-Guard) and Table 25 (Safe-Template). Across backbone models and attack settings, we observe trends that closely mirror those on AdvBench, indicating that the observed safety improvements generalize across datasets and evaluation protocols.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	100 / 63	100 / 89	99 / 36	63 / 34	99 / 47	100 / 62
With Intervention:						
Detection ($\alpha_{\text{det}} = 2.0$)	100 / 81	100 / 99	98 / 45	70 / 58	100 / 54	100 / 82
Refusal ($\alpha_{\text{ref}} = 2.0$)	100 / 88	100 / 97	100 / 66	69 / 56	100 / 52	100 / 67
Safety Neurons ($\alpha_{\text{saf}} = 1.5$)	100 / 55	100 / 97	100 / 50	58 / 67	100 / 84	100 / 88
Detection & Refusal	100 / 98	100 / 99	100 / 73	68 / 71	100 / 55	100 / 53
Detection & Refusal & Safety	100 / 49	100 / 100	100 / 85	38 / 86	100 / 78	100 / 63

Table 24: Safety rates (%) under different attack methods (Pure Harmful Prompt / ADV-LLM) across backbone models on Malicious Instruction Dataset. Interventions include scaling detection heads, refusal heads, safety neurons, and their combinations. Safety rates are measured using **Llama-Guard**.

Safety Rate (%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco	Qwen2.5	Qwen3
Baseline (Original Model)	100 / 51	100 / 76	96 / 42	39 / 30	99 / 35	77 / 50
With Intervention:						
Detection ($\alpha_{\text{det}} = 2.0$)	100 / 64	100 / 97	99 / 48	38 / 42	95 / 37	89 / 70
Refusal ($\alpha_{\text{ref}} = 2.0$)	100 / 72	100 / 93	97 / 60	43 / 33	98 / 37	91 / 48
Safety Neurons ($\alpha_{\text{saf}} = 1.5$)	100 / 41	100 / 86	97 / 54	43 / 43	99 / 67	84 / 64
Detection & Refusal	100 / 94	100 / 100	99 / 65	32 / 44	92 / 37	94 / 44
Detection & Refusal & Safety	100 / 40	100 / 100	99 / 81	82 / 55	98 / 57	75 / 41

Table 25: Safety rates (%) under different attack methods (Pure Harmful Prompt / ADV-LLM) across backbone models on Malicious Instruction Dataset. Interventions include scaling detection heads, refusal heads, safety neurons, and their combinations. Safety rates are measured using **Safe-Template**.