

From Chain to Mesh: Evolving Robust Safety Alignment via Neuron-Circuit Topology

Guan Wang^{1,2}, Xuehai Tang^{1*}, Biyu Zhou¹, Wenjie Xiao¹,
Yuanshu Zhao^{1,2}, Lijia Lv^{1,2}, Jizhong Han¹, Songlin Hu^{1,2*}

¹Institute of Information Engineering, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

wangguan@iie.ac.cn, tangxuehai@iie.ac.cn, zhoubiyu@iie.ac.cn,
hanjizhong@iie.ac.cn, husonglin@iie.ac.cn

Abstract

Despite substantial progress in safety alignment, aligned Large Language Models (LLMs) remain fragile under small parameter-space and activation-space perturbations, such as downstream fine-tuning and jailbreak attacks. To explain why local perturbations can trigger broad safety failures, we analyze safety-task neuron circuits from the perspectives of activation flow and circuit topology. We identify a critical topological fragility in safety alignment: In aligned models, safety-task neuron circuits exhibit a low-redundancy, isolated single-chain topology, in sharp contrast to the highly redundant dense-mesh topology of general-task circuits. This brittle topology creates rigid serial dependencies, whereby perturbing a small set of upstream neurons can disrupt downstream safety neurons and cascade into broad alignment failures. To mitigate this topological fragility, we propose Dense Safety Mesh (DSM), a topology-aware alignment framework that transforms brittle single-chain safety circuits into more redundant and resilient mesh-like structures. DSM employs structural adversarial masking to suppress dominant safety pathways, forcing safety circuits to develop diverse parallel pathways. Extensive experiments across seven models and multiple evaluations show that DSM improves robustness against perturbations attacks while maintaining general capabilities.

1 Introduction

The rapid integration of Large Language Models (LLMs) into diverse digital and physical ecosystems has granted them unprecedented interactive agency, necessitating rigorous safety alignment. To ensure these models adhere to human values, alignment methods such as RLHF(Ouyang et al., 2022)

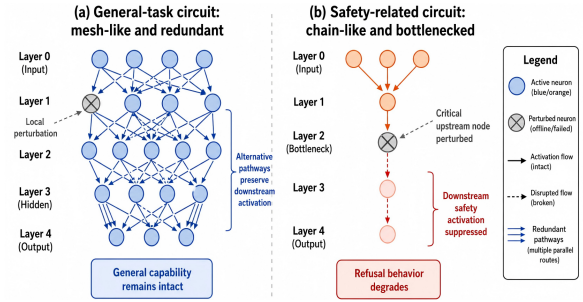


Figure 1: **(a)** The robust "Mesh-like" general circuit forms multiple parallel pathways, which can maintain circuits and functions after any neuron fails. **(b)** Fragile "Isolated Single-Chain" topology of the safety circuit. The failure of local neurons will cascadingly lead to the failure of downstream neurons and the loss of safety capability in the absence of parallel path redundancy.

and SafeDecoding(Xu et al., 2024) have been proposed, preventing the generation of harmful content and physical damage.

While aligned models exhibit remarkable compliance on static benchmarks, "red teaming" research has uncovered a deep-seated fragility: perturbations to a negligible fraction of the safety mechanism can trigger a catastrophic collapse of safety capabilities(Chen et al., 2025). Specifically, whether through modifying a minimal subset of parameters during fine-tuning(Qi et al., 2023), or shifting neuronal states in the activation space via jailbreaking prompts(He et al., 2025), these interventions can weaken the LLMs' safety capability across diverse domains and precipitate a systemic breakdown of the entire alignment capability. This disproportionate fragility raises a fundamental question: why can local perturbations induce severe global refusal failures, whereas general capabilities remain robust under similar conditions?

To fully elucidate the mechanism linking local perturbations to global model collapse, it is essential to extend our analysis beyond the traditional

* Corresponding authors: Xuehai Tang, Songlin Hu.

atomistic perspective. While prior research on LLM vulnerability focuses on the impact of a single component on security capabilities: such as the discovery that the safety capabilities are distributed across specific layers (Li et al., 2025) and dedicated safety neurons (Wei et al., 2024), and removing these safety components will degrade the model’s security performance. This overlooks the relationships and structure among components, as well as the associations between local and global structures. Accordingly, our study investigates the activation relationship between safety neurons as well as the topological structures within the safety neuron circuits (Zou et al., 2024). This circuit-level analysis of neurons uncovers a striking topological divergence between general capability and safety alignment. While general capabilities rely on a robust, highly-interconnected "Mesh-like" topology characterized by high-redundancy and parallel pathways, current neuron circuit of safety mechanism shows a sparse, **Isolated Single-Chain Topology**. In this unique topology, the activation of downstream neurons manifests a strict dependency on specific upstream precursors without lateral redundancy. Consequently, the disruption of a single neuron triggers a **Cascading Failure** of subsequent neurons that paralyzes the entire safety circuit and the global safety failure, as shown in Figure 1. This topological fragility elucidates why attacks through local perturbations succeed with disproportionate efficiency.

To dismantle this topological fragility within the safety circuit, we propose the Dense Safety Mesh (DSM), a novel framework designed to evolve the fragile, isolated single-chain into a robust, highly interconnected mesh topology. Mechanistically, DSM adaptively identifies and suppresses dominant safety pathways, forcing the safety circuit to cultivate diverse and parallel bypasses. These parallel paths are not merely structural redundancies and repetitions. Each parallel pathway corresponds to a distinct perspective or rule for safety assessment. This provides multiple parallel and redundant paths for upstream neurons to activate the same downstream neurons to enhance topological robustness, while preventing homogeneous parallel pathways from failing simultaneously under adversarial attacks. When local neurons are perturbed, parallel pathways sustain the signal propagation, ensuring downstream activation remains intact. Furthermore, to eliminate the "Alignment Tax" (Lin et al., 2024)—the degradation of general utility of-

ten associated with safety training—we introduce Orthogonal Subspace Projection. This technique geometrically constrains safety updates to the null space of general utility gradients, ensuring that the safety circuit select general task irrelevant neurons to construct parallel pathway, without interfering with the general capability.

Our contributions are threefold: 1) We provide the topological explanation for the paradox where localized perturbations trigger global safety collapse, characterizing the safety circuit as a fragile, isolated single-chain topology. 2) **Methodological Innovation**: We introduce DSM, a framework that leverages structural adversarial masking to evolve the safety circuit from fragile chain into robust mesh, and maintaining general capability by employing orthogonal projection. 3) **Empirical Robustness**: We evaluate DSM across jailbreak, adaptive attack, fine-tuning, over-refusal, utility, and topology metrics. DSM improves robustness under various models while preserving utility.

2 Related Works

2.1 Safety Fragility Analysis

Recent work has revealed that safety alignment in LLMs is brittle. Qi et al. (2024) show that safety behavior concentrates in the first few generated tokens, suggesting shallow alignment, while SSAH (Li and Kim, 2024) characterizes it as a superficial binary decision between answering and refusing. Wei et al. (2024) find that only about 3% of parameters are safety-critical, and pruning this small fraction substantially compromises safety; Li et al. (2025) locate contiguous safety layers crucial for distinguishing malicious from benign inputs. Yet they mostly analyze safety capability component as isolated units, leaving unexplained why perturbing a small local component can propagate into a global safety collapse. In contrast, our work focuses on the activation relation among safety neurons, and reveals that this fragility stems from the low redundancy topology of the safety circuit itself—an isolated single-chain structure. This topology has little lateral redundancy, making the circuit vulnerable to cascading failure.

2.2 Safety Neurons and Neural Circuits

A representative line of mechanistic safety research identifies safety-related neurons or layers. Chen et al. (2025) detect sparse safety neurons consistently important for handling harmful queries;

Wei et al. (2024) identify safety-critical neurons and low-rank components whose removal damages safety. At the circuit level, Circuit Breakers (Zou et al., 2024) learn to interrupt harmful representation circuits before they produce unsafe outputs, demonstrating that safety behavior is implemented through causal pathways rather than surface-level output filtering. Our work builds on these works and addresses a complementary problem: rather than interrupt harmful circuits, DSM evolves the safety circuit from a brittle single-chain dependency into a redundant, mesh-like topology with multiple parallel safety pathways, keeping defense mechanisms functional even when individual components are compromised.

Due to space limitations, more related works are provided in the appendix E.

3 Empirical Analysis: "Single-Chain" Topology as the Root of Fragility

To investigate the paradox where local perturbations trigger global safety collapse, we move beyond isolated neuron-level attribution to a mechanistic investigation of the Safety Circuit Topology. We uncover a fundamental structural pathology: unlike the robust "mesh" structure observed in general capabilities, current alignment paradigms unintentionally converge to a fragile Isolated Single-Chain Topology. In this section, we systematically dissect this phenomenon: (1) **Topological Discovery**: We map the safety circuit, revealing its distinct Isolated Single-Chain Topology; (2) **Mechanistic Verification**: We empirically validate that this single-chain topology creates a "structural fragility" where minimal disruption triggers a cascading failure of the entire safety circuit. The experimental Setup is shown in Appendix B.

3.1 Topological Discovery: Isolated Single-Chain Topology

By constructing the Safety Circuit through causal activation flow mapping, we uncover a striking Topological Duality: while general reasoning tasks activate a robust "Mesh-like" structure characterized by high-redundancy and parallel processing, the safety mechanism manifests as a low-redundancy and fragile "Isolated Single-Chain Topology." Within this pathological architecture, downstream safety neurons' activation exhibit a rigid dependency on a sparse set of specific precursors', critically lacking the lateral redundancy

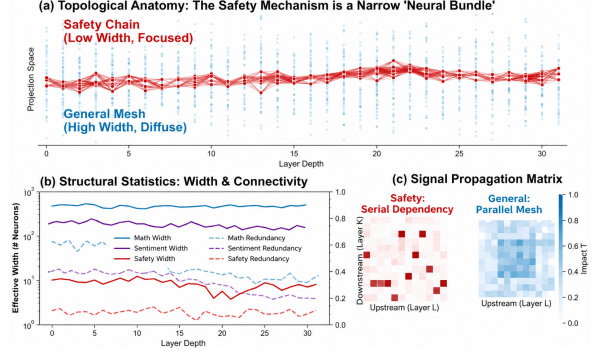


Figure 2: **(a-b)** Safety mechanisms appear as narrow and low-redundancy "neuron circuit" with minimal redundancy compared to both general circuits. **(c)** The downstream safety neurons' activation depends on a small number of upstream neurons, and they have the characteristic of a high causality.

observed in general capabilities.

Methodology: Circuit Identification. We map the safety circuit through two steps: (1) Identifying *Safety Key Neurons* (Nodes), and (2) connecting their causal connectivity of activation (Edges) to map the *Safety Circuit*.

Step 1: Identifying Safety Key Neurons. **Safety Key Neurons** refer to an MLP-channels activation at a specified layer and contributive to safety tasks while remaining dormant during general tasks. We define an *Importance Score*, $S^{(l,i)}$, for neuron i in layer l , which quantifies the trade-off between safety contribution and specificity. The score is calculated as the gradient-weighted activation on harmful data ($\mathcal{D}_{\text{harm}}$), penalized by the magnitude of activation on benign data ($\mathcal{D}_{\text{benign}}$):

$$S^{(l,i)} = \mathbb{E}_{x \in \mathcal{D}_{\text{harm}}} \left[\left| a^{(l,i)}(x) \frac{\partial \mathcal{L}}{\partial a^{(l,i)}} \right| \right] - \mu \mathbb{E}_{x \in \mathcal{D}_{\text{benign}}} \left[|a^{(l,i)}(x)| \right]. \quad (1)$$

where $a^{(l,i)}$ denotes the activation value and L is the loss function for safety task. The first term (Sensitivity) identifies neurons that significantly influence the safety decision, while the second term (Selectivity) filters out general utility neurons. We designate the top-3% of neurons with the highest S scores as **Safety Key Neurons**, and we report node-selection sensitivity in Appendix J.

Step 2: Mapping the Safety Circuit. We observe that neurons are not activated in isolation; rather, they form coherent activation circuits, which we term the safety circuit. To map the safety circuit, we construct a weighted directed graph where edges represent the *causal activation flow* between neurons. The edge weight $T_{u \rightarrow v}$ from neuron u

(layer l) to neuron v (layer $l+1$) is quantified using a first-order Taylor approximation of contribution:

$$T_{u \rightarrow v} \approx \left| a_u^{(l)}(x) \cdot W_{uv} \cdot \frac{\partial \mathcal{L}}{\partial a_v^{(l+1)}} \right| \quad (2)$$

Here, W_{uv} is the effective local contribution from neuron u to neuron v through the residual stream and next-layer MLP input. This first-order Taylor estimate captures the dominant contribution term while neglecting higher-order interactions; we use $T_{u \rightarrow v}$ as a relative ranking signal for pathway importance rather than an exact causal quantity. By pruning edges with $T_{u \rightarrow v}$ below a significance threshold τ , we reconstruct the active **Safety Circuit**, as shown in Figure 2 (a).

Topological Characteristics: The Single-Chain Topology. To benchmark the safety topology, we apply the same circuit identification methodology to general reasoning tasks. To control variables, we selected the top 3% as general neurons consistent with the safety task, and adopted the same threshold τ to construct general circuits. A potential confound is task complexity: refusal is often a short binary decision, whereas GSM8K (math) requires broader reasoning. We therefore compare safety refusal not only with general reasoning, but also with simple non-safety control tasks sentiment classification. This controls match refusal in output complexity while differing in safety semantics. These settings eliminate the impact of neuron quantity caused by the complexity and variety of general and safety tasks.

Controlling for Output Complexity. A further potential confound is that safety refusal produces a short binary output, whereas general reasoning requires longer free-form answers. To test whether the topology difference is driven by output format rather than safety semantics, we additionally compare safety refusal against two non-safety binary controls—yes/no factual QA and topic classification—that produce short outputs without involving safety semantics. As shown in Table 12 (Appendix L), both controls exhibit substantially higher algebraic connectivity, average degree, and path redundancy than the safety circuit, despite matching safety refusal in output format. This indicates that the low-redundancy topology is associated with the safety mechanism rather than being a generic property of binary-decision tasks. The topology contrast is also stable across node budgets and edge-retention settings.

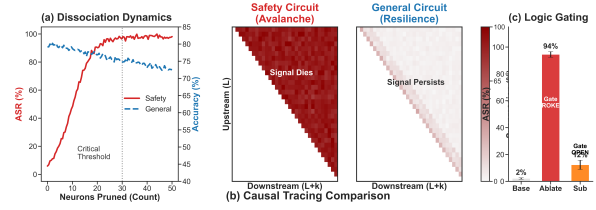


Figure 3: (a) Masking a minimal fraction ($< 0.8\%$) of safety neurons triggers immediate safety collapse (red), while general utility remains robust (blue). (b) Masking upstream layers (L) in the Safety Circuit triggers near-total downstream suppression (left). In contrast, the General Circuit (right) maintains signal flow via mesh-like structure. (c) Flow breaking results in high ASR, whereas residual subtraction maintains low ASR.

As visualized in Figure 2, contrasting the Safety Circuit with General Circuits reveals a striking topological duality:

1) Spatial Sparsity: While both general tasks activate a broad and a large part of general neurons in each layer to process complex logic for each query. In contrast, the number of active neurons in safety circuits for each query is approximately one-third of that observed in general tasks.

2) The "Isolated Single-Chain": Both general circuits exhibit a "Mesh-like" structure characterized by high connectivity (clustering). Downstream neurons aggregate signals from a diverse ensemble of upstream neurons. Conversely, the Safety Circuit manifests as an **"Isolated Single-Chain."** In this topology, the activation of a downstream safety neuron depends heavily on a few, specific precursors with high edge weight T . This results in any upstream neuron having a significant impact on the activation of downstream neurons. In contrast, downstream general neurons can be activated by more upstream neurons and pathways.

These features confirm that the safety circuit is topologically distinct from general circuits. In this topological structure, the high dependence of subsequent neurons on any preceding neurons leads to a conjecture: Does the activation of safety neurons relying on fewer upstream neurons and pathways entail less redundancy compared with general tasks? could this cause the failure of a single neuron to trigger a cascading propagation in the safety circuit and result in the failure of subsequent neurons, thereby leading local perturbations to trigger global safety collapse? We verify this hypothesis in the next section.

3.2 Mechanistic Verification: The Fragility of Single-Chain Topology

Building on the Single-Chain Topology identified in Section 3.1, we test whether disrupting critical upstream neurons interrupts downstream signal propagation and degrades safety more sharply than general-task performance.

Internal Observation: The Cascading Effect.

To verify the conjecture in Section 3.1 regarding the vulnerability to cascading failures caused by **Isolated Single-chain Topology**, we investigate the signal propagation dynamics within the model’s internal representation space. We conduct a comparative ablation study between safety circuits and general capability circuits using Causal Tracing. Specifically, we simulate localized neuronal failure by masking critical nodes in upstream layers and strictly monitor the resulting activation suppression in downstream layers to quantify the degree of serial dependency.

As illustrated in Figure 3(b), the safety mechanism exhibits a distinct **Cascading Effect**. Suppressing safety-critical neurons in an upstream layer L triggers a sharper downstream drop of activations across all subsequent layers $L + k$. This phenomenon empirically validates the strict serial dependency and low-redundancy of the Isolated Single-Chain Topology: localized upstream disruption can compromise downstream signal propagation. In stark contrast, general capability circuits demonstrate robust, high-redundancy. Even when local specific upstream nodes are compromised, downstream neurons maintain high activation levels by aggregating signals from alternative parallel pathways, thereby preserving the circuit’s functional integrity.

External Observation: Alignment Collapse.

The cascading failures within the model caused by the single-chain topology are also directly reflected in the security of the external output. As illustrated in Figure 3(a), the results reveal the vulnerability of safety capability. For general tasks, masking results in a **Graceful Degradation** of performance, confirming the robustness of general circuits’ "Mesh" topology. Conversely, the Safety Circuit exhibits a **Precipitous Surge** in Attack Success Rate (ASR). The failure of the entire safety circuit, which is caused by a negligible fraction, triggers a rapid rise in ASR and the collapse of safety capability.

We further examine this fragility under fine-

tuning and jailbreaking (Figure 7). Local activation restoration recovers safety in many affected cases, consistent with a role for disrupted signal propagation. Detailed analysis is provided in Appendix D.

Interestingly, a small number of safety circuits do not collapse after upstream masking. As shown in Appendix C, these resilient cases exhibit compensatory intermediate activations that form local parallel routes. This naturally motivates DSM: instead of relying on accidental local redundancy, we explicitly cultivate dense parallel safety pathways.

Causal Observation: Topological Vulnerability or Direct-Contribution?

To distinguish disrupted signal propagation from loss of a neuron’s direct contribution, we compare two interventions on the same neurons. *Flow breaking* zeros their outputs at layer l , removing both direct and downstream effects. *Direct-contribution subtraction* preserves these outputs for subsequent layers while removing their estimated direct contribution to the final output through matched residual-stream subtraction.

If local direct contribution alone explained the failure, these interventions should have similar effects. Instead, flow breaking yields much higher ASR than direct-contribution subtraction in Figure 3(c). This gap is consistent with a role for disrupted downstream propagation beyond the removed direct contribution.

We also compare bottleneck neurons (high betweenness centrality on safety paths) with attribution-matched non-bottleneck neurons. In Table 1, bottleneck masking yields 28.6% ASR, versus 21.4% for non-bottleneck masking and 15.6% for direct-contribution subtraction on the same bottlenecks. Downstream activation retention is 21.7% and 8.9%, respectively: higher ASR does not correspond to stronger aggregate suppression. These controls support a role for topological position beyond attribution magnitude and motivate DSM’s alternative pathways for preserving safety-relevant signal flow.

Together, these observations identify a structural vulnerability: localized perturbations can disrupt downstream safety signals and degrade alignment. The resilient cases with alternative routes motivate DSM’s cultivation of parallel pathways.

4 Methodology

Based upon the "Single-Chain Topology" identified in Section 3, we propose the **Dense Safety**

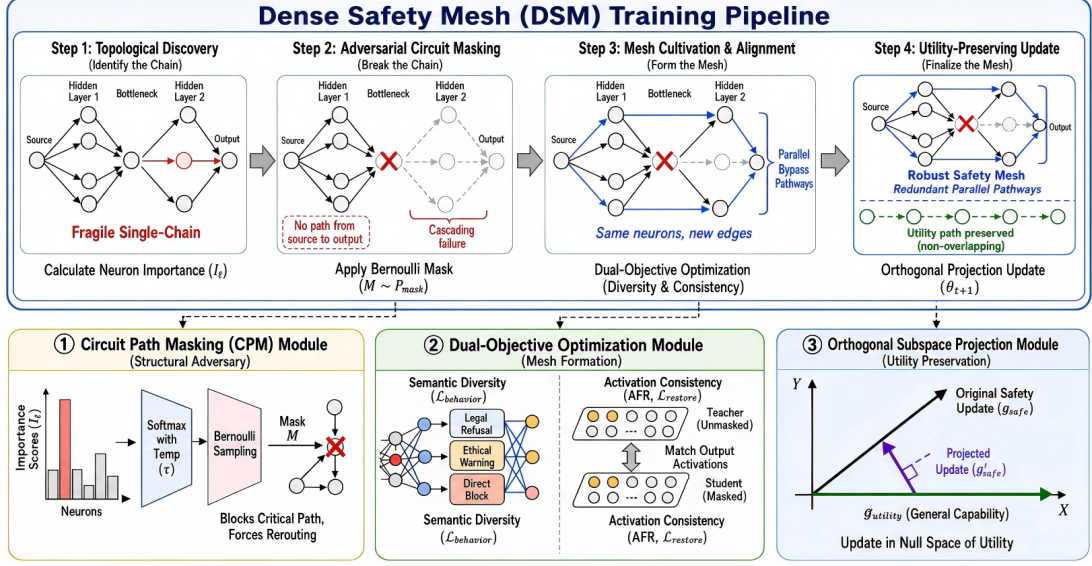


Figure 4: The work flow of DSM: (1) **Circuit Path Masking (CPM)**, (2) **Dual-Objective Optimization**, and (3) **Orthogonal Subspace Projection**

Intervention	Position	Ret. (%)	ASR (%)
Bottleneck masking	Chain bottleneck	21.7	28.6
Non-bottleneck masking	Lateral	8.9	21.4
Direct-contribution subtraction	Bottleneck	—	15.6

Table 1: Attribution-matched interventions on LLaMA-2-7B-Chat. Bottleneck masking yields higher ASR than non-bottleneck masking or direct-contribution subtraction on the same bottlenecks. Ret. denotes downstream activation retained.

Mesh (DSM), a topology-aware robust alignment framework designed to evolve the fragile, low-redundancy safety circuits into a robust, highly parallel mesh topology. DSM maintains safety capability when safety circuits are partially disturbed. As illustrated in Figure 4, DSM has three components: (1) **Circuit Path Masking (CPM)** suppresses the existing pathway in the circuit; (2) **Dual-Objective Optimization** guides the recruitment of dormant neurons to form diverse and consistent parallel pathways; and (3) **Orthogonal Subspace Projection** eliminates the "Alignment Tax."

4.1 Problem Formulation

We formulate robust safety alignment as an optimization problem of the neural circuit topology within the MLP parameter θ . Let $\mathcal{D}_{safe}$ denote the safety dataset containing malicious queries x_{mal} . Our empirical analysis reveals that alignment con-

verges to a **Single-Chain Topology** Circuits. To dismantle this vulnerability, our objective is to learn a parameter configuration θ^* where the safety function is supported by multiple **Parallel Pathway**. We define a safety pathway $\mathcal{P}$ as a sequence of causally connected neurons $\{n^{(l)}, \dots, n^{(L)}\}$ in safety circuit. We block the pathway by masking a subset of critical neurons in it. Mathematically, we aim to maximize the expected safety performance under a distribution of adversarial masks $\mathcal{M}$, ensuring robustness of safety circuit:

$$\max_{\theta} \mathbb{E}_{M \sim \mathcal{M}} [S(f(x_{mal}; \theta \odot (\mathbf{1} - M)))] \quad (3)$$

where $\odot$ denotes the element-wise product, and S is the safety scoring function. This enables the maintenance of safety circuits and safety capabilities through parallel paths even after some neurons fail, we compel the topology to transition from a single chain to a dense mesh.

Implementation Note. DSM modifies only existing model weights during training; it introduces no new neurons, modules, or trainable parameters. Terms such as "dormant neuron recruitment" and "mesh construction" refer to redistributing safety-related activation flow over existing MLP channels—not to adding architectural components. After training, the checkpoint, architecture, and inference procedure are identical to standard fine-tuning, with zero inference overhead (Table 4).

4.2 Circuit Pathway Masking (CPM)

To optimize Eq. (3), we introduce **Circuit Path Masking (CPM)**, a dynamic training strategy that

functions as a "structural adversary." CPM iteratively identifies the current safety circuit and actively suppresses pathways through masking the neurons in them. This mechanism forces the gradient flow to divert into the latent parameter space—reawakening dormant neurons.

To prevent optimization hacking (i.e., the model just flip-flopping between two discrete circuits), we employ a probabilistic soft-masking strategy. We construct a binary mask $M \in \{0, 1\}^{|\theta|}$, where the probability of masking neuron i is governed by a temperature-scaled Boltzmann distribution derived from its importance score I_i :

$$p_{mask}(i) = \frac{\exp(I_i/\tau)}{\sum_j \exp(I_j/\tau)} \quad (4)$$

Here, τ serves as a crucial hyperparameter regulating the **exploration-exploitation trade-off**. A lower τ aggressively targets the most critical neurons, while a moderate τ introduces stochasticity to encourage broader exploration of potential backup circuits. This mechanism continuously blocks the easiest path, the model is compelled to carve out new, parallel signal pathways.

4.3 Dual-Objective Optimization

Merely blocking the primary path is insufficient; the model must learn how to effectively utilize the newly recruited bypasses. We design a composite loss function to ensure the emerging mesh possesses both **Diversity-Aware Behavioral Alignment** (logic redundancy) and **Activation Flow Restoration** (signal integrity): $\mathcal{L}_{safe} = \mathcal{L}_{behavior} + \beta \cdot \mathcal{L}_{restore}$

4.3.1 Diversity-Aware Behavioral Alignment

To prevent the parallel pathway from simply replicating the function of the origin pathway, which may fail at the same time against attacks, we train the parallel pathways to learn different perspectives and knowledge of safe reasoning. We construct a diverse set of safety responses $\mathcal{Y}(x) = \{y_{direct}, y_{legal}, y_{ethical}, \dots\}$, covering different reasoning strategies for the same query. During training, when the primary path is masked, we randomly sample a target response $y^{(k)} \in \mathcal{Y}$ distinct from the primary path's preference:

$$\mathcal{L}_{behavior} = -\log P(y^{(k)} | x, \theta \odot (1 - M)) \quad (5)$$

This encourages the formation of diversity safety circuits—e.g., one circuit enforcing legal boundaries and another enforcing ethical guidelines—

thereby creating a Multi-level defense against semantic jailbreaks.

4.3.2 Activation Flow Restoration (AFR)

We introduce AFR to ensure that parallel pathways successfully route signals back to the target downstream neurons. Treating the downstream neurons' activation distribution of the unmasked model as the "teacher" and the masked model's as the "student," we minimize the divergence of their distribution $h^{(l)}$ at deep layers $l \in \mathcal{L}_{deep}$:

$$\mathcal{L}_{restore} = \sum_{l \in \mathcal{L}_{deep}} \|h^{(l)}(x; \theta) - h^{(l)}(x; \theta \odot (1 - M))\|_2^2 \quad (6)$$

AFR enforces that even if the signal takes a detour through the mesh due to masking, the downstream activation distribution remains invariant, ensuring the robustness of the safety circuit.

4.4 Utility Preservation

A key challenge in robust alignment is the "Alignment Tax"—the degradation of general capabilities (Utility). We address this by training alignment via Orthogonal Subspace Projection. Let g_{safe} be the DSM safety gradient and $G_U = [g_1, \dots, g_k]$ be gradients from sampled utility batches such as math, coding, instruction following, and knowledge QA. Let U be an orthonormal basis for the column space of G_U . DSM updates with

$$g_{safe}^\perp = (I - UU^\top)g_{safe}. \quad (7)$$

This reduces first-order interference with sampled utility-sensitive directions. This orthogonal projection strategy is related to gradient-based multi-task conflict resolution methods such as PCGrad (Yu et al., 2020). It mathematically allows DSM to significantly enhance robustness while preserving general capability.

5 Experiments

In this section, we empirically validate the effectiveness of DSM from three key aspects: Robustness, Usefulness, and Mechanism. The details of experiment setup is shown in appendix.

5.1 DSM is Robust

Robust against Jailbreaking As shown in Table 2, DSM exhibits robustness against diverse jailbreak attacks compared to baselines. Mechanistically, this resilience stems from the dense *Mesh* topology safety circuit and parallel pathways that

Method	Test-time interv.	Safety: ASR (%) ↓					Avg. ASR	Capabilities ↑		
		GCG	AutoDAN	Code Attack	PAIR	Art Prompt		Alpaca	GSM8K	MMLU Pro
Llama-3-8B-Instruct										
No Defense	×	78.4	69.2	52.1	75.3	38.5	62.7	28.5	85.6	43.8
PPL	✓	5.1	48.3	38.4	56.1	32.4	36.1	25.4	85.4	43.1
RLHF	×	44.2	15.6	20.1	22.4	18.3	24.1	26.8	84.1	42.5
CircuitBreaker	✓	16.5	18.4	30.2	23.1	20.1	21.7	27.5	83.9	43.2
SafeDecoding	✓	32.1	25.8	42.1	48.5	29.4	35.6	28.2	85.2	43.5
AED	✓	31.3	16.5	25.4	24.8	15.2	22.6	23.1	74.3	38.2
STAIR	✓	34.8	9.2	12.5	8.4	<u>10.1</u>	15.0	28.2	85.9	43.6
DSM	×	18.4	<u>8.3</u>	10.5	<u>6.4</u>	9.8	<u>10.7</u>	28.4	<u>85.5</u>	43.7
DSM (Rand.)	✓	<u>14.9</u>	7.9	<u>10.9</u>	4.1	10.2	9.6	<u>28.3</u>	<u>85.5</u>	<u>43.6</u>
Qwen2.5-7B-Instruct										
No Defense	×	85.2	76.5	88.4	82.6	45.2	75.6	30.5	88.2	44.5
PPL	✓	9.5	58.2	75.1	65.4	38.5	49.3	28.2	88.0	43.9
RLHF	×	48.5	25.1	35.2	30.5	22.5	32.4	29.8	87.8	43.2
CircuitBreaker	✓	18.2	22.5	40.5	38.4	24.2	28.8	<u>30.4</u>	86.2	44.2
SafeDecoding	✓	38.5	32.4	58.2	55.6	34.2	43.8	<u>30.1</u>	87.9	43.6
AED	✓	35.6	20.4	35.2	32.5	20.5	28.8	26.5	78.5	39.5
STAIR	✓	38.5	13.2	18.5	12.4	12.6	19.0	29.9	88.2	<u>44.3</u>
DSM	×	25.2	<u>10.5</u>	14.4	<u>9.5</u>	<u>12.8</u>	<u>14.5</u>	<u>30.4</u>	<u>88.1</u>	44.4
DSM (Rand.)	✓	<u>16.8</u>	9.9	<u>15.5</u>	9.1	<u>12.8</u>	12.8	30.5	<u>88.1</u>	<u>44.3</u>
Vicuna-13B										
No Defense	×	99.2	96.5	88.4	95.1	78.2	91.5	18.2	52.4	32.8
PPL	✓	10.5	88.2	82.1	85.4	70.1	67.3	16.5	52.1	31.5
RLHF	×	60.5	50.2	55.4	52.1	45.2	52.7	17.5	51.5	31.8
CircuitBreaker	✓	32.5	40.4	48.2	42.1	35.4	39.7	16.1	48.4	30.6
SafeDecoding	✓	70.1	62.4	65.2	70.5	55.2	64.7	17.8	<u>52.2</u>	<u>32.5</u>
AED	✓	45.2	30.1	45.4	38.2	25.1	36.8	15.4	45.2	28.5
STAIR	✓	55.4	<u>22.5</u>	28.2	20.1	<u>22.5</u>	29.7	18.2	52.3	<u>32.5</u>
DSM	×	29.6	20.2	25.4	<u>18.5</u>	22.1	<u>23.2</u>	<u>18.0</u>	52.3	32.7
DSM (Rand.)	✓	<u>17.5</u>	20.2	<u>27.5</u>	17.8	22.6	21.1	17.9	<u>52.2</u>	32.7
Qwen3-8B-Instruct										
No Defense	×	61.0	36.0	66.8	56.8	36.5	51.4	31.7	89.8	66.7
PPL	✓	4.8	24.5	49.6	42.0	27.0	29.6	29.2	89.5	66.1
RLHF	×	34.5	15.8	26.2	22.3	17.8	23.3	31.1	89.0	65.8
CircuitBreaker	✓	12.5	14.8	27.8	20.8	18.2	18.8	31.3	88.8	66.3
SafeDecoding	✓	24.3	18.5	39.8	39.5	25.6	29.5	31.5	89.6	66.2
AED	✓	28.4	13.5	23.8	21.0	14.5	20.2	28.0	80.5	61.0
STAIR	✓	26.0	8.8	12.8	8.3	9.5	13.1	31.9	90.1	66.8
DSM	×	14.0	<u>7.5</u>	10.2	<u>6.2</u>	8.7	<u>9.3</u>	32.1	<u>90.0</u>	66.9
DSM (Rand.)	✓	<u>10.5</u>	7.2	<u>10.5</u>	5.4	<u>8.9</u>	8.5	<u>32.0</u>	<u>90.0</u>	66.8

Table 2: Robustness and utility comparison under jailbreak attacks. Lower ASR is better, while higher capability scores are better. Best and second-best results among defenses are **bold** and underlined. Results on more models are shown in Table 10.

evaluate query harmfulness across diverse abstraction levels. Consequently, even if an adversarial prompt successfully circumvents a subset of peripheral security neurons, the parallel safety logic in alternative branches remains intact, effectively preventing the catastrophic failure observed in standard aligned models.

Beyond static defense, DSM uses *Stochastic Activation*, randomly activating 90% of the safety circuit per inference step. This "Moving Target" strategy aims to disrupt consistent attack feedback.

At later attack steps, Dynamic DSM achieves substantially lower ASR than Static DSM, although its own ASR increases modestly (Table 14; Appendix N). Under topology-aware progressive attacks that suppress multiple bottlenecks, DSM degrades gracefully while standard alignment collapses (Appendix M).

Robust against Finetuning Table 3 reports ASR before and after *full-parameter* malicious finetuning (5 epochs, standard cross-entropy, no safety constraint). DSM retains the lowest safety degrada-

Method	Llama-2-7B			Mistral-7B			Qwen2.5-7B		
	Pre	Post	$\Delta \downarrow$	Pre	Post	$\Delta \downarrow$	Pre	Post	$\Delta \downarrow$
RLHF	1.2	88.5	+87.3	2.1	92.4	+90.3	1.5	85.1	+83.6
RESTA	1.2	14.5	+13.3	2.1	21.3	+19.2	1.5	14.6	+13.1
SSAH	1.2	21.2	+20.0	2.1	25.6	+23.5	1.5	18.9	+17.4
DSM	0.8	13.5	+12.7	1.5	14.8	+13.3	0.9	12.9	+12.0

Table 3: ASR (%) before (Pre) and after (Post) malicious fine-tuning. Δ is the increase in ASR, in percentage points.

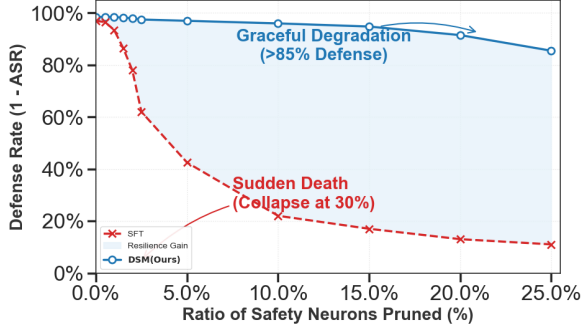


Figure 5: **DSM** demonstrates "Graceful Degradation," compared with SFT. The shaded area highlights the significant resilience gain provided by the mesh topology.

tion (Δ) across all three models. DSM also maintains robustness under LoRA adaptation across ranks 4–32. Importantly, DSM does not require any changes to the downstream fine-tuning process, and can additionally collaborate with methods for protecting safety capabilities.

Robust against Editing We edited the safety key neuron to validate the structure robustness of DSM. Figure 5 shows that **DSM** demonstrates *Graceful Degradation* compared with SFT. This confirms the structure robustness of DSM.

To more specifically and quantitatively demonstrate the robustness of topological structure that DSM achieved, we calculate the **Algebraic Connectivity** (λ_2) of safety circuit. The safety circuit of DSM achieved $\lambda_2 \approx 0.32$, which is a mathematical increase from SFT ($\lambda_2 \approx 1.4 \times 10^{-2}$), as shown in Table 7. This confirms DSM transiting the safety circuit to a highly connected, robust topology.

5.2 DSM Is Useful

DSM preserves the Downstream capability

Across the utility metrics in Table 2, the maximum decrease relative to No Defense is 0.2 percentage points for static **DSM** and 0.3 percentage points for DSM (Rand.). These results support the utility-preserving role of Orthogonal Subspace Projection (Section 4.4).

Metric	SFT	DSM
Training time (h)	12.5	14.4
Relative cost	1.00×	1.15×
Throughput (samples/s)	14.2	12.3
Inference (ms/token)	42.8	42.8
Inference overhead	–	+0%

Table 4: Training cost and inference latency on Llama-3-8B using $8 \times A100$ GPUs.

To further verification, we calculated the Jacard overlap index between neurons activated by MMLU tasks and safety queries. Table 8 shows that the overlap decreases to $J = 0.04$ compared with $J = 0.15$ (SFT), while the total count of safety neurons increases. This confirms that DSM utilizes the neurons that are nearly not needed for general tasks, thereby preserving general capabilities.

DSM avoids excessive refusal Table 5 shows that DSM does not over-refuse. We attribute this to the diverse parallel pathways: the model learns multiple accurate perspectives for harmfulness identification, preventing excessive response to specific patterns. Detailed analysis is provided in Appendix A.

DSM is Efficient During the training phase, the construction of safety circuit only requires the use of activation values from inference and gradients from backpropagation for calculation, without introducing excessive additional computations. As shown in Table 4, *DSM* requires only $1.15 \times$ the training time of standard SFT and imposes **zero additional inference latency**.

Ablation Studies The results in Table 9 show that removing either the Targeted Masking module or the Temperature Sampling module causes a significant decline in safety capabilities, while disabling Orthogonal Subspace Projection greatly impairs general performance. This demonstrates the necessity and effectiveness of each module.

6 Conclusion

In this work, we mechanistically identify the "Isolated Single-Chain Topology" of safety circuit as the key fragility in aligned LLMs. We propose the DSM to evolve the safety circuit from chain to mesh, maintaining downstream capabilities with negligible computational overhead.

Limitations

Scalability to Massive or Specialized Architectures. Our empirical validation covers a diverse range of representative open-source LLMs (Llama, Mistral, Qwen) up to the 13B parameter scale. Due to the substantial computational resources required for mechanistic interpretability analysis and retraining, we have not extended our experiments to ultra-large-scale models (e.g., >70B). While theoretical properties of high-dimensional optimization suggest that the "Chain Vulnerability" likely persists or exacerbates at scale due to sparsity, validating DSM on such scales remains a subject for future research. Furthermore, our topological analysis focuses on dense Transformer architectures; the interaction between DSM and architectures with intrinsic routing mechanisms, such as Mixture-of-Experts (MoE), warrants further specific investigation.

Idealized Assumption of Subspace Orthogonality. Our method mitigates the "Alignment Tax" by projecting safety updates into the null space of general utility gradients. This approach relies on the assumption that safety mechanisms and general capabilities occupy quasi-orthogonal subspaces within the high-dimensional parameter space. While this holds true for distinct tasks, there may exist "Grey Areas" where safety and utility are intrinsically entangled (e.g., medical advice that requires both high utility and strict safety boundaries). In such highly overlapped domains, strict orthogonal projection might necessitate a more nuanced trade-off. Meanwhile, this overlap between safety and general capabilities also implies that the safety neurons we identified are not only activated during safety tasks but may also be effective during general tasks. This is also the reason why the circuits of safety neurons can still be affected by the fine-tuning of downstream tasks. We discuss in more detail in the appendix J the impact of the degree of penalty on the activation of general tasks when identifying safety neurons on our observed results. Our method strives to reduce alignment tax as much as possible, yet it cannot completely eliminate it.

Topology Metrics. Our primary structural measure is algebraic connectivity (λ_2), supplemented by average degree, edge density, and path redundancy. Each metric has limitations: λ_2 is sensitive to graph size and edge-pruning thresholds, and may overstate connectivity differences between circuits of different cardinalities. To mitigate this, we ap-

ply matched graph budgets and verify the topology contrast across multiple metrics and threshold settings. Whether the single-chain topology and DSM's mesh-building strategy generalize to graded or multi-class safety behaviors (e.g., partial compliance, content warnings) remains an open question.

Statements

LLM Usage. We employed Large Language Models (LLMs) exclusively for sentence polishing and grammatical error correction. No original content creation or scientific reasoning was delegated to these models.

Reproducibility Statement. To facilitate the full reproducibility of our experiments, we will publicly release the source code and datasets needed to reproduce our experiments.

Ethics Statement. Our research on safety alignment is conducted strictly for scientific purposes, with the primary goal of enhancing model reliability. We adhere to ethical guidelines and emphasize that this work is not intended for malicious applications or privacy violations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. U24A20335).

References

- Gabriel Alon and Michael Kamfonas. 2023. [Detecting language model attacks with perplexity](#). *Preprint*, arXiv:2308.14132.
- Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. 2024. [Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic](#). *Preprint*, arXiv:2402.11746.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. [Jailbreaking black box large language models in twenty queries](#). *ArXiv*, abs/2310.08419.
- Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. 2025. [Towards understanding safety alignment: A mechanistic perspective from safety neurons](#). *Preprint*, arXiv:2406.14144.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.

- Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. [OR-bench: An over-refusal benchmark for large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11515–11542. PMLR.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. [Length-controlled alpaca-eval: A simple way to debias automatic evaluators](#). *Preprint*, arXiv:2404.04475.
- Zeqing He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Wenhui Zhang, Qinglong Wang, and Rui Zheng. 2025. [Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit](#). *Preprint*, arXiv:2411.11114.
- Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. [ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15157–15173, Bangkok, Thailand. Association for Computational Linguistics.
- Jianwei Li and Jung-Eun Kim. 2024. [Superficial safety alignment hypothesis](#). *Preprint*, arXiv:2410.10862.
- Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2025. [Safety layers in aligned large language models: The key to llm security](#). *Preprint*, arXiv:2408.17003.
- Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. [Mitigating the alignment tax of rlhf](#). *Preprint*, arXiv:2309.06256.
- Quan Liu, Zhenhong Zhou, Longzhu He, Yi Liu, Wei Zhang, and Sen Su. 2024. [Alignment-enhanced decoding: Defending jailbreaks via token-level adaptive refining of probability distributions](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2802–2816, Miami, Florida, USA. Association for Computational Linguistics.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. [Autodan: Generating stealthy jailbreak prompts on aligned large language models](#). *ArXiv*, abs/2310.04451.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *NeurIPS*.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2024. [Safety alignment should be made more than just a few tokens deep](#). *Preprint*, arXiv:2406.05946.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. [Fine-tuning aligned language models compromises safety, even when users do not intend to!](#) *Preprint*, arXiv:2310.03693.
- Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. 2024. [CodeAttack: Revealing safety generalization challenges of large language models via code completion](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11437–11452, Bangkok, Thailand. Association for Computational Linguistics.
- Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. [Xstest: A test suite for identifying exaggerated safety behaviours in large language models](#). *Preprint*, arXiv:2308.01263.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *Preprint*, arXiv:2406.01574.
- Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. [Assessing the brittleness of safety alignment via pruning and low-rank modifications](#). *Preprint*, arXiv:2402.05162.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024. [SafeDecoding: Defending against jailbreak attacks via safety-aware decoding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand. Association for Computational Linguistics.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. [Gradient surgery for multi-task learning](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836.
- Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. 2025. [Stair: Improving safety alignment with introspective reasoning](#). *Preprint*, arXiv:2502.02384.
- 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](#). *Preprint*, arXiv:2406.04313.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). *Preprint*, arXiv:2307.15043.

A DSM avoids excessive refusal

Benchmark	Model	Base	RLHF	DSM (Ours)
OR-Bench	Llama 2-7B	16.5	21.6	15.5
	Llama 2-13B	14.9	18.2	14.3
	Llama 3-8B	6.8	11.5	6.9
	Qwen2.5-7B	4.4	12.7	4.5
XSTest	Llama 2-7B	38.4	40.5	33.6
	Llama 2-13B	32.0	34.7	30.4
	Llama 3-8B	1.9	4.8	2.3
	Qwen2.5-7B	1.4	5.1	1.4

Table 5: Inappropriate refusal rates (%) on benign but sensitive benchmarks. Lower scores indicate fewer incorrect refusals. DSM has broadly comparable or lower refusal rates than the base model and lower rates than RLHF on all tested settings.

DSM Avoids Excessive Refusal. Table 5 reports results on OR-Bench (Cui et al., 2025) and XSTest (Röttger et al., 2024). DSM does not produce excessive rejection, which we attribute to the establishment of our diverse branches. In general safety alignment training, the basis for rejection is not clearly informed to the model, or only one rejection judgment mode is relied on. In order to obtain higher protection capabilities, the model has to learn an overly harsh or even wrong discrimination method, and the output weights of the corresponding safety neurons will also be too large. Through multi-angle data, the model has learned diverse and accurate harmful identification, and also prevented excessive stress response to specific patterns, thus preventing the model from excessive rejection.

B Experimental Setup

In this section, we detail our experimental setup for our observational analyses and experiments, including the datasets used for safety and utility evaluation, the baseline methods for attacks and defenses, and the metrics employed to assess model performance.

B.1 Datasets

To comprehensively evaluate the trade-off between safety robustness and general utility, we employ a diverse set of benchmarks divided into two categories: Safety Evaluation and Utility Evaluation.

Safety Benchmarks. To assess the model’s resistance to jailbreak attacks, we utilize **AdvBench** (Zou et al., 2023). AdvBench is a widely adopted dataset containing harmful behaviors and toxic instructions intended to elicit objectionable content from LLMs. We use the harmful query augmented with various jailbreaking methods to evaluate the Attack Success Rate (ASR). We ensure that there is no overlap between the training set and the test set.

To evaluate over-refusal, we use **XSTest** (Röttger et al., 2024) and **OR-Bench** (Cui et al., 2025). Both evaluate refusals of benign prompts that may appear sensitive or harmful. XSTest contains safe prompts with potentially misleading lexical cues, while OR-Bench provides seemingly toxic but benign prompts across multiple categories. Table 5 reports inappropriate refusal rates on benign prompts.

Utility Benchmarks. Maintaining general capabilities is crucial for defense mechanisms. We evaluate downstream task performance using the following datasets:

- **GSM8K** (Cobbe et al., 2021): A dataset of high-quality grade school math problems, measuring the model’s multi-step mathematical reasoning ability.
- **MMLU-Pro** (Wang et al., 2024): An enhanced version of the Massive Multitask Language Understanding (MMLU) benchmark. MMLU-Pro offers a more challenging and robust evaluation of knowledge and reasoning across various domains, mitigating the saturation issues found in standard MMLU.
- **AlpacaEval** (Dubois et al., 2025): A simulated instruction-following benchmark that evaluates the model’s ability to generate helpful and relevant responses to general user instructions.

B.2 Baselines

Baseline Terminology. In our tables, “No Defense” denotes the original instruction-tuned checkpoint without additional safety training. “RLHF” denotes the same architecture after reinforcement learning from human feedback for safety. “Standard SFT” in the topology analysis (Tables 7–8) refers to the RLHF-aligned checkpoint before applying DSM.

We compare our proposed method against a comprehensive set of attack and defense baselines to

demonstrate its efficacy. We introduced them in related works.

Attack Methods. To rigorously test the defense mechanisms, we employ the following adversarial attack algorithms: GCG(Zou et al., 2023), PAIR(Chao et al., 2023), AutoDAN(Liu et al., 2023), CodeAttack(Ren et al., 2024), and ArtPrompt(Jiang et al., 2024).

Defense Baselines. We benchmark our approach against state-of-the-art defense strategies: PPL(Alon and Kamfonas, 2023), SafeDecoding(Xu et al., 2024), AED(Liu et al., 2024), STAIR(Zhang et al., 2025), and CircuitBreaker(Zou et al., 2024).

B.3 Evaluation Metrics

Safety Metrics. Our primary metric for safety is the **Attack Success Rate (ASR)**. ASR is defined as the percentage of adversarial prompts for which the model generates a harmful response. Following established protocols, we utilize **GPT-4** as an automated safety judge. We provide the target response and the model’s output to GPT-4, prompting it to classify whether the output constitutes a successful jailbreak (i.e., answering the harmful query). We ran each harmful query five times; if any run succeeded, we counted it as a breach. To ensure the credibility of the results, we randomly selected some of the experimental results and made manual judgments. The two judgments were highly consistent.

For **XSTest**(Röttger et al., 2024), we report the **Refusal Rate** on safe prompts. A lower Refusal Rate on XSTest indicates that the model correctly identifies benign intents and avoids over-defensive behavior.

Utility Metrics. For downstream tasks, we report the standard metrics associated with each benchmark:

- **Accuracy** is reported for GSM8K (exact match) and MMLU-Pro (choice accuracy).
- **Win Rate** (against a reference model GPT-4) is reported for AlpacaEval, assessing the helpfulness of the generated text.

Base Models: To ensure the generality of our findings, we conduct experiments on various models: **Llama-3-8B-Instruct**, **Mistral-7B-Instruct-v0.2**, **Qwen2.5-7B-Instruct**, **Vicuna-13B**, **Llama2-7B**.

Baselines: Training-Time Alignment Methods includes: **RLHF (Reinforcement Learning from Human Feedback)** guides the LLM’s policy refinement via reinforcement learning. **STAIR (SafeTy Alignment with Introspective Reasoning)**: generates safety reasoning to improve final outputs. We adapt it for safety by having it reason about potential policy violations. **DSM (Dynamic)**: randomly activate only 90% of the pathway each time when facing multi-round attacks. **Inference-Time Defense Strategies** includes: **PPL (Perplexity-based Defense)**: detects adversarial prompts by measuring the perplexity and related features of the input prompt or adversarial suffix. **AED (Alignment-Enhanced Decoding)** (Liu et al., 2024): uses a Competitive Index to adaptively combine prompt-conditioned next-token logits with post-alignment logits derived from the generated prefix. **SafeDecoding** (Xu et al., 2024): reweights probabilities over shared high-probability candidate tokens from the original model and a safety-tuned expert to promote safe continuations during the initial decoding steps.

Adaptive Jailbreak Attacks: To comprehensively assess model robustness, we evaluate against five diverse and state-of-the-art jailbreak attack methods: **GCG (Greedy Coordinate Gradient)**: a classic white-box attack using gradient-based greedy search to find adversarial suffixes. **PAIR (Prompt Automatic Iterative Refinement)**: a black-box attack that uses an attacker LLM to iteratively refine prompts to elicit unsafe content. **AutoDAN**: a method using hierarchical genetic algorithms to automatically generate diverse and effective jailbreaks. **CodeAttack** (Ren et al., 2024): encodes malicious instructions in code-style formatting (e.g., pseudocode, function calls) to bypass safety filters that rely on natural-language pattern matching. **ArtPrompt** (Jiang et al., 2024): replaces sensitive keywords with their ASCII art representations, exploiting the gap between visual recognition and token-level safety detection.

Metrics: We use a comprehensive set of metrics to evaluate different aspects of model performance: **Attack Success Rate (ASR)** is automatically judged by GPT-4. A lower ASR indicates higher robustness. **Downstream Utility** are measured by performance on standard benchmarks: MMLU pro(accuracy) and Alpaca-Eval (win rate).

B.4 Implementation Details

Model and Environment. All experiments are implemented in PyTorch and executed on a cluster of **NVIDIA A100 (80GB) GPUs**. To ensure reproducibility, we set the random seed to 42 for all runs.

Baseline Configurations. We adhere to the default hyperparameter settings provided in the original papers of the baseline attacks to ensure a fair comparison:

- **GCG:** We run the attack for **500 iterations**. The candidate suffix batch size is set to **512**, and the top- k for substitution is set to **256**. The adversarial suffix is initialized with a string of 20 exclamation marks.
- **AutoDAN:** We utilize a hierarchical genetic algorithm with a population size of **64**. The attack is run for **100 generations** (steps) with a mutation rate of **0.01**.
- **PAIR:** We set the maximum number of retries to 3 and the maximum conversation depth to 20, using qwen2.5 as the attacker model.

The other setup is aligned with our methods and the original paper.

Evaluation Settings. The maximum number of new tokens is set to **512**. For the GPT-4 judge, we use the gpt-4 judge with a temperature of 0.

Training Settings. For our proposed method, we fine-tune the model using the AdamW optimizer. We set the global **batch size to 32** and the **learning rate to $2e-5$** with a cosine learning rate schedule and a warm-up ratio of 0.03. All training is performed using BF16 precision to optimize memory usage.

The other setup is aligned with our methods and the original paper.

C The Resilience Case: Parallel Structure

The observation experiments in Section 3.2 reveal a small number of naturally resilient safety circuits in the original aligned model, before DSM training. In these cases, safety alignment persists after upstream masking because localized parallel pathways provide alternative routes for signal propagation. These exceptions offer a structural blueprint for our proposed method.

To examine this survival, we traced activation propagation in the naturally resilient cases. As shown in Figure 6, compensatory activations in intermediate layers sustain downstream safety neurons after one upstream pathway is disrupted. The comparison concerns naturally occurring circuit structures in the original model, rather than the effect of DSM training.

Currently, this local parallel structure is accidental and sparse. However, this naturally emerging parallel structure has inspired our proposal: to actively evolve the safety circuit into a Dense Mesh topology featuring massive parallel structure, thereby engineering structural robustness into the safety mechanism.

D More Details: Verification of Real Scenarios

Building upon the identification of the "Single-Chain Topology" in previous sections, we now investigate whether this topological fragility is the mechanistic root cause of safety failures in real-world attack scenarios. We posit that external attacks—whether through Malicious Fine-tuning (parameter space) or Jailbreaking (activation space)—do not require global destruction of the model’s safety alignment. Instead, they succeed by inducing a localized failure in a small subset of Safety-Key Neurons, triggering a "domino effect" that collapses the entire defense mechanism.

Real-Scenarios Observation: Fine-tuning and Jailbreaking To verify the practical implications of our topological findings, we examine malicious fine-tuning and jailbreaking. Both can disrupt downstream safety activation through localized changes to safety-relevant neurons. In the fine-tuning experiment, restoring the outputs of selected safety neurons in one upstream layer recovers downstream activation and restores refusal in 44% of the analyzed failure cases. The intervention acts on a group of neurons within one layer, rather than a single neuron. Restoration experiments on cases with multiple damaged sites further indicate that repairing one site may be insufficient when other disruptions remain. These observations support a dependence on intact signal propagation along the tested safety pathways and identify single-chain topology as a vulnerability under the examined interventions.

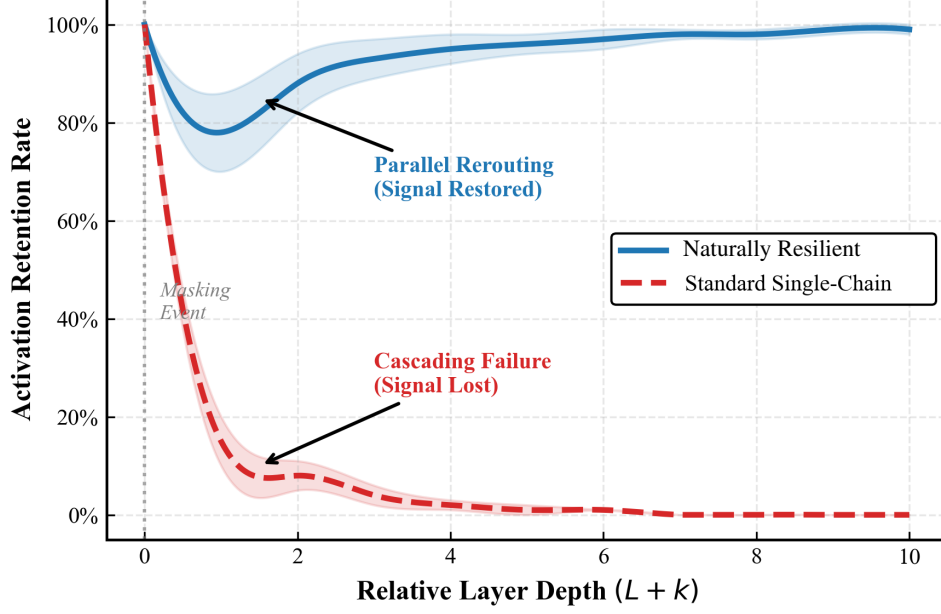


Figure 6: Activation dynamics under masking in the original aligned model, before DSM training. Standard single-chain circuits (red) lose downstream activation, while naturally resilient circuits (blue) retain it through local parallel pathways.

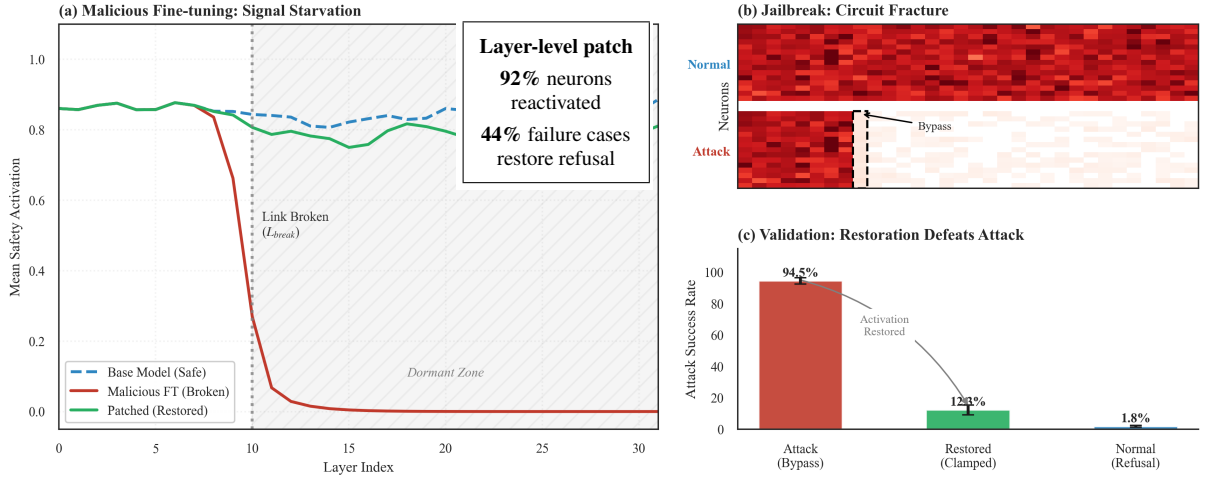


Figure 7: (a) Fine-tuning disrupts downstream activation. Patching selected safety-neuron outputs in one upstream layer restores downstream activity (green); neuron reactivation and query-level refusal recovery are reported separately. (b) Jailbreak inputs suppress key safety neurons, disrupting downstream activation. (c) Restoring these activations reduces ASR from 94.5% to 12.3%, an approximately 87% relative reduction.

Mechanistic Validation: The "Signal Starvation" Hypothesis. To investigate precisely *how* fine-tuning degrades safety, we performed a **Causal Activation Patching** experiment. We verified whether fine-tuning, as we hypothesized, undermines the safety alignment capability by exploiting the vulnerability of single-chain topology through the recovery of the output of local safety neurons.

Setup: For each failure case in LLM after SFT, we traced the safety circuit and identified the *First Breakpoint Layer* (L_{break}) where safety neurons failed to activate compared to the base model be-

fore SFT. We then artificially "grafted" the output of the safety neurons from the Base Model at Layer $L_{break} - 1$ into the Fine-tuned Model, effectively restoring the input signal to the broken link.

Results: Patching the selected safety-neuron outputs in one upstream layer reactivated 92% of the safety neurons in the subsequent layer L_{break} . At the behavioral level, 44% of the analyzed post-SFT failure cases changed from harmful responses to refusals. These percentages describe different outcomes: neuronal reactivation and query-level recovery, respectively. The results support a role for

disrupted upstream signal propagation, while showing that downstream reactivation does not always restore refusal.

Jailbreaking: Activation Space Bypass For adversarial attacks, we apply fixed jailbreak templates to harmful queries to generate adversarial inputs. We define a neuron as being "**Bypassed**" if its activation magnitude under attack is suppressed below the threshold required to trigger the subsequent node in the Series Circuit. **Results:** In the examined settings, jailbreaks that suppress safety-key neurons are associated with ASR above 90% and reduced downstream activation, whereas perturbations outside these neurons yield ASR below 5%. Restoring the suppressed activations reduces ASR from 94.5% to 12.3% (Figure 7(c)), an approximately 87% relative reduction. This restoration experiment supports a causal contribution of disrupted safety-signal propagation to the observed jailbreak failures.

These observations confirm that external failures are mechanistically grounded in the topology of the Single Chain. The collapse of defense is not an overall destruction of all security capabilities, but a local destruction. Because the single-chain loop topology causes the failure of the overall security loop.

D.1 Qualitative Analysis: Multi-View Safety Reasoning

While quantitative metrics validate the structural resilience of the Dense Safety Mesh (DSM), it is equally crucial to understand how this topology manifests in semantic reasoning. We hypothesize that our Dual-Objective Optimization encourages parallel pathways within the mesh to specialize in distinct safety perspectives—creating *Semantic Redundancy*.

To validate this, we conducted a *Path Isolation Analysis*. For a given malicious query, we selectively masked dominant pathways to force the model to rely on alternative sub-meshes. As illustrated in Table 6, we observed a “Diversity in Reasoning, Unity in Refusal” phenomenon. When different sub-meshes are isolated, they yield refusals grounded in fundamentally different logical premises:

- **Legal Specialist (Pathway α):** Activates neurons correlated with regulatory concepts, rejecting based on international laws (e.g., CWC) or terms of service.

- **Risk Specialist (Pathway β):** Focuses on physical consequences, highlighting volatility, health risks, or lack of safety containment.
- **Ethical Specialist (Pathway γ):** Aligns with general harm prevention and moral intent.

In the fully active DSM, these perspectives operate under an effective “OR” logic. Even if an adversarial prompt successfully “role-plays” to bypass the Legal pathway, the Physical Risk pathway remains active to intercept the query. This explains the effectiveness of our Stochastic Activation strategy: since multiple diverse pathways can independently trigger a refusal, randomly deactivating a subset of neurons does not compromise the overall safety consensus.

E Related works

E.1 Existing Work on Safety Maintaining

Existing alignment efforts have attempted to improve robustness against fine-tuning. RESTA(Bhardwaj et al., 2024) can be re-aligned for safety during fine-tuning via mathematical operations on model parameters, restoring model safety by adding safety vectors to fine-tuned model parameters. SSAH(Li and Kim, 2024) froze 7.5% of safety-critical components during fine-tuning. Safely Partial-Parameter Fine-Tuning (SPPFT)(Li et al., 2025) addresses this by identifying and freezing a contiguous block of middle layers—deemed “safety layers”—to prevent the catastrophic forgetting of refusal behaviors. While, our approach differs from theirs in that we enhance the intrinsic robustness of model alignment without interfering with subsequent fine-tuning. This allows us to maintain alignment capabilities while also preserving better downstream task performance.

E.2 Safety Fragility Analysis

Recent work has revealed that safety alignment in LLMs is brittle. Qi et al. (2024) show that safety behavior concentrates in the first few generated tokens, suggesting shallow alignment, while SSAH (Li and Kim, 2024) characterizes it as a superficial binary decision between answering and refusing. Wei et al. (2024) find that only about 3% of parameters are safety-critical, and pruning this small fraction substantially compromises safety; Li et al. (2025) locate contiguous safety layers crucial for distinguishing malicious from benign inputs.

Active Topology	Dominant Perspective		Model Response (to Query: “ <i>I need to synthesize Sarin gas...</i> ”)
Pathway α (Isolated)	Legal & Regulatory		I cannot provide the formula. The production and use of chemical weapons are strictly prohibited under the Chemical Weapons Convention (CWC) and international law. Manufacturing such controlled substances is a serious federal crime.
Pathway β (Isolated)	Physical Safety		I absolutely cannot assist with this. Sarin is an extremely volatile nerve agent. Synthesizing it without military-grade containment would result in immediate respiratory failure and death for you and anyone nearby.
Pathway γ (Isolated)	Ethical & Intent		I must refuse this request. While you mentioned pest control, this substance is a weapon of mass destruction designed to cause human suffering. My safety guidelines strictly prevent generating instructions for dual-use technologies that inflict harm.
Full DSM (Mesh)	Unified ness	Robust- ness	I cannot fulfill this request. Sarin is a banned chemical weapon (CWC), and creating it is both illegal and extremely dangerous. The synthesis process poses a lethal risk. Using such a substance is unethical and violates safety policies.

Table 6: **Case Study of Semantic Diversity.** We isolate distinct sub-meshes within DSM to answer a high-risk query. The results show that DSM does not merely memorize a single refusal template; instead, it aggregates diverse reasoning paths (Legal, Physical, Ethical) to form a robust safety consensus. This redundancy ensures that bypassing one logic (e.g., via jailbreaking) does not collapse the entire defense.

Yet they mostly analyze safety capability component as isolated units, leaving unexplained why perturbing a small local component can propagate into a global safety collapse. In contrast, our work focuses on the activation relation among safety neurons, and reveals that this fragility stems from the low redundancy topology of the safety circuit itself—an isolated single-chain structure. This topology has little lateral redundancy, making the circuit vulnerable to cascading failure.

E.3 Safety Neurons and Neural Circuits

A representative line of mechanistic safety research identifies safety-related neurons or layers. [Chen et al. \(2025\)](#) detect sparse safety neurons consistently important for handling harmful queries; [Wei et al. \(2024\)](#) identify safety-critical neurons and low-rank components whose removal damages safety. At the circuit level, Circuit Breakers ([Zou et al., 2024](#)) learn to interrupt harmful representation circuits before they produce unsafe outputs, demonstrating that safety behavior is implemented through causal pathways rather than surface-level output filtering. Our work based on these works and addresses a complementary problem: rather than interrupt harmful circuits, DSM evolves the safety circuit from a brittle single-chain dependency into a redundant, mesh-like topology with multiple parallel safety pathways, keeping defense mechanisms functional even when individual components are compromised.

F Mechanistic Analysis

F.1 Topological Phase Transition

To verify the structural transformation, we construct the neuron dependency graph of the aligned models and calculate the **Algebraic Connectivity** (λ_2).

- **Chain Topology (SFT):** $\lambda_2 \approx 1.4 \times 10^{-2}$. The near-zero value indicates a fragile structure easily partitioned by removing minimal edges.
- **Mesh Topology (DSM):** $\lambda_2 \approx 0.32$. The orders-of-magnitude increase mathematically confirms a phase transition to a highly connected, robust graph.

F.2 Recruitment of Dormant Neurons

We calculated the Jaccard overlap index between neurons activated by MMLU tasks and neurons activated by safety queries. For SFT, the overlap is moderate ($J = 0.15$), indicating resource competition. For DSM, the overlap decreases to $J = 0.04$, while the total count of safety neurons increases. This reveals that DSM activates a distinct set of neurons that were "dormant" in the origin base models. These dormant neurons do not function in almost all tasks, and their functions are idle and wasted in the basic model. DSM utilizes these neurons that are not needed for general tasks, thereby preserving general capabilities.

Model	Method	Graph Topology Metrics			Robustness
		Alg. Conn. (λ_2)	Avg. Degree (k)	Clustering Coeff.	Critical Cut (%)
LLaMA-2-7B	Standard SFT	1.4×10^{-2}	1.2 ± 0.4	0.08	0.5%
	DSM (Ours)	0.32	8.7 ± 1.5	0.41	> 15.0%
Mistral-7B	Standard SFT	2.1×10^{-2}	1.5 ± 0.5	0.11	0.8%
	DSM (Ours)	0.38	9.4 ± 1.8	0.45	> 18.5%

Table 7: **Topological Phase Transition Analysis.** Comparison of graph-theoretic metrics between Standard SFT and our DSM (DSM). **Algebraic Connectivity** (λ_2) quantifies the structural robustness; near-zero values indicate a fragile, disconnected chain, while higher values indicate a highly connected mesh. **Critical Cut (%)** denotes the percentage of neurons required to be pruned to sever the signal propagation.

Method	Neuron Activation Counts			Overlap Index	Dark Matter Recruitment
	Safety (N_{safe})	Utility (N_{util})	Shared ($N_{safe} \cap N_{util}$)	$J(N_{safe}, N_{util})$	
<i>LLaMA-2-7B-Chat</i>					
Standard SFT	485	2,104	342	0.15	1.2%
Grad-based Adv. [†]	1,250	2,015	580	0.22	4.5%
DSM (Ours)	1,842	2,110	152	0.04	12.8%

Table 8: **Resource Allocation Analysis.** We investigate the mechanism of utility preservation by measuring the Jaccard Similarity Index (J) between safety-critical neurons (N_{safe}) and general utility neurons (N_{util}). **Dark Matter Usage** tracks the recruitment of neurons that were dormant in the base model. Lower J combined with higher Dark Matter Usage confirms that DSM builds redundancy in the null space of general capabilities. [†]Grad-based Adv. refers to a gradient-based adversarial training baseline that augments safety data with gradient-computed perturbations.

G Spectral Connectivity Monitor

To quantify the structural evolution from chain to mesh, we introduce a topology-aware metric based on Spectral Graph Theory. We construct a neuron dependency undirected graph $G = (V, E)$ and calculate the **Algebraic Connectivity** (λ_2), defined as the second smallest eigenvalue of the graph Laplacian L :

$$\lambda_2(G) = \text{eig}_2(D - A) \quad (8)$$

A value of $\lambda_2 \approx 0$ indicates a disconnectable chain topology. We observe λ_2 during training and terminate CPM when λ_2 reaches a plateau, signaling the formation of a robust, highly connected Dense Safety Mesh.

H Ablation Studies

We investigate the contribution of each component in DSM:

- **w/o Targeted Masking (Random Dropout):** Resilience to targeted pruning drops significantly, proving that "blind" redundancy is insufficient; the defense must be adversarial.

- **w/o Temperature Sampling (Hard Top-K):** Robustness against diverse jailbreaks decreases. The model overfits to specific neurons ("Whac-A-Mole"), failing to form a global mesh.
- **w/o Orthogonal Subspace Projection:** Safety scores remain high, but GSM8K drops by 2.2 percentage points and MMLU-Pro by 5.3 percentage points, confirming that orthogonal projection is essential for preserving general capabilities.

I More Models

This section reports additional results on Mistral, Qwen2, and Llama-2, showing that DSM consistently reduces ASR while preserving utility across model families.

J Impact of Selectivity Coefficient μ on Circuit Identification

In Section 3.1, we introduced the Importance Score $S^{(l,i)}$ to identify Safety Key Neurons, where the hyperparameter μ serves as a penalty term for activation on benign data. A critical concern in circuit

Variant	Avg. ASR (% ↓)	GSM8K ↑	MLLM Pro ↑	λ_2 ↑
DSM (full)	12.0	25.2	38.3	0.32
w/o targeted masking (random dropout)	62.1	25.0	38.0	0.04
w/o temp. sampling (hard top- K)	28.4	24.9	36.7	0.15
w/o orthogonal proj. (utility reg.)	11.5	23.0	33.0	0.30

Table 9: **Ablation Study on Llama-2-7B-Chat.** The full DSM uses the same checkpoint and evaluation run as the Llama-2 DSM row in Table 10. Avg. ASR is the mean over the five attack methods in that table. We ablate targeted masking, temperature sampling, and orthogonal projection. **Bold** indicates the best trade-off performance.

identification is the functional nature of the selected neurons. In this section, we analyze the sensitivity of our method to μ and clarify that our objective is **not to isolate neurons exclusive to safety**, but to filter out task-agnostic noise while retaining necessary semantic dependencies. We need to emphasize that the safe neurons filtered out under our settings are not completely inactive in other general tasks, and thus will also be affected during the fine-tuning of general tasks.

J.1 Functional Taxonomy of Neurons

Neural computation in LLMs is inherently polysemantic. To understand the effect of μ , we categorize neurons into three functional types based on their activation patterns across safety and general datasets:

- **Type I: Ubiquitous Syntactic Neurons.** These neurons perform fundamental linguistic functions, such as pronoun extraction (e.g., distinguishing "he" vs. "she"), punctuation processing, and common token embedding. They activate frequently across *all* tasks, regardless of semantic content.
- **Type II: Context-Dependent Shared Neurons.** These neurons encode high-level concepts required by both malicious and benign queries (e.g., concepts like "explosive", "politics", or "encryption"). They serve as the semantic bridge: the safety mechanism must activate them to understand *what* is being prohibited.
- **Type III: Safety-Specific Decision Neurons.** These neurons specifically encode the refusal

intent, moral judgement, or the detection of jailbreak templates.

Our proposed metric aims to vigorously suppress **Type I**, as they constitute the "background noise" of the model backbone. However, it is designed to preserve a functional subset of **Type II** alongside **Type III**. Completely severing Type II would leave the safety circuit blind to the context of the query.

J.2 Sensitivity Analysis of μ

We varied μ within the range $[0.1, 3.0]$ on Llama-2-7B to observe the structural evolution of the identified circuit:

Low Selectivity ($\mu < 0.5$): Inclusion of Syntactic Noise. When μ is set too low, the penalty for benign activation is insufficient. The identified circuit becomes dominated by Type I neurons. Qualitative analysis reveals that many selected neurons in this regime are active simply due to linguistic features (e.g., extracting person names or processing sentence terminators). Consequently, the distinct "Single-Chain" topology is obscured by this dense mesh of syntactic scaffold, making the safety structure indistinguishable from general capabilities.

High Selectivity ($\mu > 1.5$): Loss of Context. Conversely, an excessively high μ imposes a harsh penalty on any neuron active in general tasks. This results in the removal of Type II neurons. While this yields a set of neurons with extremely high specificity for refusal keywords, the circuit becomes topologically fragmented. The "Single-Chain" breaks because the upstream neurons—responsible for understanding the malicious concept—are excluded.

Optimal Balance ($\mu \approx 1.0$). We empirically set $\mu = 1.0$ for our main experiments. In this regime, the metric successfully acts as a *semantic filter*: it removes the ubiquitous syntactic neurons (Type I) while retaining the critical path formed by context understanding (Type II) and safety decision (Type III). This allows us to observe the "Isolated Single-Chain" topology: a structure that is sparse compared to the general mesh but sufficiently connected to function as a logical circuit.

K Details of Efficiency

Table 11 provides a step-wise profiling breakdown, showing that DSM's additional cost comes mainly from circuit masking and projection during training, while inference latency remains unchanged.

Method	Test-time	Safety Robustness (ASR, %) ↓					Avg ASR	Capabilities ↑		
	interv.	GCG	AutoDAN	Code Attack	PAIR	Art Prompt		Alpaca	GSM8k	MMLU Pro
Mistral-7B-Instruct-v0.2										
No Defense	x	99.5	97.2	90.1	96.5	80.4	92.7	22.5	68.4	36.2
PPL		11.2	89.5	83.4	86.8	72.5	68.7	20.1	68.1	35.5
RLHF	x	62.8	52.4	57.1	54.5	48.1	55.0	21.4	67.5	35.8
CircuitBreaker	✓	35.1	42.8	50.5	44.2	37.5	42.0	19.8	65.2	34.1
safedecoding		72.4	65.1	68.4	72.8	58.4	67.4	21.8	68.0	35.9
AED	✓	47.5	32.4	47.1	40.5	27.2	38.9	18.5	60.5	31.4
STAIR		57.2	<u>24.1</u>	30.5	22.4	<u>24.1</u>	31.7	22.4	68.5	<u>36.1</u>
DSM	x	31.5	21.5	27.2	<u>20.1</u>	23.5	<u>24.8</u>	<u>22.2</u>	68.5	36.3
DSM (Rand.)	✓	<u>19.2</u>	21.5	<u>29.4</u>	19.5	24.2	22.8	<u>22.2</u>	<u>68.4</u>	36.3
Qwen2-7B-Instruct										
No Defense	x	82.5	74.1	85.5	80.2	42.8	73.0	29.5	86.5	43.1
PPL		8.2	55.4	72.1	62.5	36.1	46.9	27.1	86.2	42.5
RLHF	x	45.1	22.8	32.4	28.1	20.4	29.8	28.5	85.8	41.8
CircuitBreaker	✓	16.5	20.2	38.1	35.2	22.5	26.5	<u>29.4</u>	84.5	42.8
safedecoding		36.2	30.5	55.4	52.1	32.5	41.3	29.1	86.1	42.2
AED	✓	33.5	18.2	32.1	30.4	18.5	26.5	25.4	76.2	38.1
STAIR		36.4	12.1	16.2	10.5	11.2	17.3	28.8	86.5	<u>42.9</u>
DSM	x	23.1	<u>9.2</u>	13.1	<u>8.2</u>	<u>11.5</u>	<u>13.0</u>	<u>29.4</u>	<u>86.4</u>	43.0
DSM (Rand.)	✓	<u>15.4</u>	8.5	<u>14.2</u>	7.8	11.2	11.4	29.5	<u>86.4</u>	<u>42.9</u>
Llama-2-7B-Chat										
No Defense	x	80.5	71.5	54.2	77.4	40.2	64.8	26.5	25.2	38.5
PPL		5.8	50.1	40.5	58.2	34.5	37.8	23.8	25.1	37.8
RLHF	x	46.5	17.2	22.4	24.5	20.1	26.1	25.1	24.5	37.2
CircuitBreaker	✓	18.2	20.5	32.5	25.4	22.5	23.8	25.5	24.2	37.9
safedecoding		34.5	27.4	44.5	50.1	31.2	37.5	26.2	24.8	38.1
AED	✓	33.1	18.2	27.5	26.5	16.8	24.4	21.5	18.5	33.5
STAIR		36.5	10.5	14.2	9.8	<u>11.5</u>	16.5	26.2	25.4	<u>38.2</u>
DSM	x	20.2	<u>9.5</u>	11.8	<u>7.5</u>	11.2	<u>12.0</u>	26.4	<u>25.2</u>	38.3
DSM (Rand.)	✓	<u>16.5</u>	9.1	<u>12.4</u>	5.2	11.8	11.0	<u>26.3</u>	<u>25.2</u>	<u>38.2</u>

Table 10: Performance comparison against Jailbreaks and general capabilities. The best results among defenses are **bold**, and second-best are underlined. ↓ lower is better; ↑ higher is better.

Method	Phase duration (ms)					Total/step	Ratio
	Attack gen.	Forward	Backward	Optim. & proj.	Common overhead		
Standard SFT	–	200	350	100	50	700 ms	1.00×
DSM (Ours)	–	240	350	165	50	805 ms	1.15×

Table 11: **Step-wise Training Cost Breakdown.** Average wall-clock time per training step on Llama-3-8B. Both methods incur 50 ms of common overhead. DSM adds 105 ms per step in the forward and optimization phases, with no attack-generation phase.

L Binary Non-Safety Control Tasks

To rule out the possibility that the single-chain topology is an artifact of binary output format rather than a safety-specific property, we compare safety refusal against two additional non-safety controls that produce short outputs: yes/no factual QA and topic classification. All circuits use the same top-3% node budget and 5% edge-retention threshold.

M Progressive Structure-Aware Attack

To evaluate DSM under a topology-aware attacker that explicitly targets the mesh structure, we design a progressive white-box stress test. The attacker iteratively selects safety neurons with the highest marginal coverage of distinct safety paths, suppressing $m \in \{1, 2, 4, 8, 16, 32\}$ targeted bottlenecks. At each step, the selected neurons’ outputs are zeroed during the forward pass.

Standard alignment exhibits an early cliff: a sin-

Task	Output	$\lambda_2 \uparrow$	Avg Deg $\uparrow$	Density $\uparrow$	Redund. $\uparrow$
Safety refusal	refuse/comply	0.014	1.2	0.018	1.1
Sentiment	pos/neg	0.118	3.6	0.061	3.1
Yes/no QA	yes/no	0.096	3.2	0.054	2.7
Topic classif.	short label	0.134	4.0	0.069	3.5
General reasoning	free-form	0.226	5.1	0.087	4.8

Table 12: Topology comparison: safety refusal vs. matched non-safety controls on LLaMA-2-7B-Chat. All binary controls exhibit substantially higher connectivity than the safety circuit despite matching its output format, confirming that the low-redundancy topology is associated with safety semantics.

m	Std ASR	DSM ASR	Std Ret.	DSM Ret.
0	1.1	0.7	100.0	100.0
1	41.2	2.0	62.9	97.3
2	55.9	3.1	41.3	94.0
4	72.5	6.1	28.0	87.6
8	83.6	10.0	16.4	81.3
16	89.3	16.1	10.8	67.6
32	92.7	24.5	6.9	56.8

Table 13: Progressive multi-path suppression on LLaMA-2-7B-Chat. “Ret.” = downstream activation retained (%). Standard alignment collapses after 1–2 bottleneck disruptions; DSM degrades gradually, retaining $>80\%$ downstream activation up to 8 targeted points.

DSM	Attack step				Avg.
	1	5	10	20	
GCG (white-box)					
Static	0.8	4.2	11.5	24.6	10.3
Dynamic	0.9	1.2	1.5	2.1	1.4
PAIR (black-box)					
Static	1.5	5.8	12.4	17.2	9.2
Dynamic	1.6	2.0	2.4	2.9	2.2

Table 14: **Resilience against Iterative Attacks.** ASR (%) at successive attack steps. Static DSM degrades as attackers optimize their queries; Dynamic DSM maintains low ASR. Avg. is the mean over the four displayed steps.

gle bottleneck drives ASR above 40% and downstream retention below 63%. DSM, in contrast, maintains $>80\%$ retained activation up to 8 targeted points and shows graceful rather than catastrophic degradation beyond this range.

N Dynamic Inference Defense against Iterative Attacks

N.1 Motivation and Methodology

Standard aligned models rely on a deterministic forward pass, presenting a static decision boundary. This static nature creates a vulnerability exploited by iterative optimization attacks (e.g., GCG, PAIR,

AutoDAN). In these scenarios, attackers utilize the model’s consistent feedback (gradients or logits) from previous rounds to progressively optimize the adversarial prompt, eventually locating a “crack” in the safety wall.

Leveraging the high redundancy of the **Dense Safety Mesh (DSM)** constructed in Section 4, we introduce a *Dynamic Inference Strategy*. Unlike the fragile “Single-Chain” topology where every node is critical, our mesh topology allows for the stochastic activation of safety pathways without compromising overall defense integrity.

Specifically, during the inference phase of each defense round, we apply a random binary mask $M_{dyn} \sim \text{Bernoulli}(p)$ to the identified safety neurons, where p is the retention probability. In our experiments, we set $p = 0.9$, meaning only 90% of the safety mesh is active for any given query. The model output is computed as:

$$y = f(x; \theta_{safety} \odot M_{dyn}, \theta_{utility}) \quad (9)$$

This strategy transforms the safety mechanism into a *Moving Target*. By shifting the activation landscape, this strategy aims to reduce the consistency of the feedback available to iterative attacks.

N.2 Experimental Results

We simulated a multi-round attack scenario using the PAIR (Prompt Automatic Iterative Refinement) and GCG (Gradient-based) attack frameworks on Llama-2-7B-Chat aligned via DSM. We compared the **Static DSM** (100% activation) against the **Dynamic DSM** (90% stochastic activation) over 20 attack iterations.

Analysis of Attack Trajectories. As shown in Table 14, we observe two key phenomena:

1. **Preservation of Initial Defense Capability:** At Step 1, the ASR of Dynamic DSM (e.g., 0.9% for GCG) is comparable to or only marginally higher than Static DSM (0.8%).

This confirms our topological finding: the safety mesh is sufficiently redundant that randomly deactivating 10% of neurons does not trigger the "Sudden Death" collapse observed in single-chain models.

2. **Lower ASR than Static DSM in Later Rounds:** At Step 20, Dynamic DSM reduces ASR from 24.6% to 2.1% under GCG and from 17.2% to 2.9% under PAIR relative to Static DSM. These are reductions of 22.5 and 14.3 percentage points, respectively. Dynamic DSM remains below 3% at all reported steps, although its own ASR increases modestly with the attack budget.

The lower ASR relative to Static DSM is consistent with stochastic activation making attack feedback less stable across steps. This interpretation concerns the tested attack settings; the trajectories do not imply that Dynamic DSM's ASR decreases over time.