A Study on Mitigating Jailbreaking Vulnerabilities in LLMs via Knowledge Conflict Analysis and Targeted Intervention

Seunghyun Yoo

October 9, 2025

Contents

1 Introduction & Background **5**

2 Research Questions **5**

3 Literature Review **6**
3.1 The Mechanism of Jailbreaking: Representation Manipulation and Internal Conflict . . . 6
3.2 Locality of Safety: Safety Layers and Safety Neurons . . . 6
3.3 Causal Analysis and Evaluation Methodology . . . 6

4 Core Concepts and Methodological Background **7**
4.1 Mechanistic Interpretability (MI) . . . 7
4.2 Transformer Architecture and Key Components for Analysis: FFN & Attention Head . . . 7
4.2.1 Attention Head: The 'Relationship Analyst' for Contextual Understanding . . . 7
4.2.2 Feed-Forward Network (FFN): The 'Brain Circuitry' for Knowledge Processing . . . 7
4.3 Jailbreak Prompt . . . 8
4.4 Evaluator Model: Granite Guardian . . . 8

5 Detailed Experimental Protocol **9**
5.1 Objective . . . 9
5.2 Setup . . . 9
5.3 Benchmark Suite and Rationale . . . 9
5.3.1 Jailbreak Attack Evaluation . . . 9
5.3.2 Safety and Generalization Evaluation . . . 9
5.3.3 General Performance Evaluation . . . 9
5.4 Mathematical Formulation of Metrics . . . 10
5.5 Procedure . . . 11
5.6 Evaluation Metrics . . . 11
5.7 Expected Results and Visualization Plan . . . 13

6 Expected Contributions **18**

List of Figures

1 **PCA Visualization of LLM Internal States.** A 2D PCA plot of activation vectors from the final layer of Llama-3-8B. The clear clustering and separation of safe prompts (blue) and jailbreak prompts (red) just before response generation suggests that the model internally processes these two input types differently. . . . 13
2 **Non-linear Representation Space Analysis with t-SNE.** The same data as in Fig 1 visualized with the non-linear dimensionality reduction technique t-SNE. Data points that appeared to overlap in the linear PCA space form even clearer clusters here, showing the potential for non-linear separability between the two states. . . . 13

3 **Conceptual Diagram of the 'Knowledge Conflict' Hypothesis.** A schematic of the core hypothesis. When a user prompt is input, (A) 'Jailbreak-Inducing Attention Heads' generate a signal (S_{context}) to adhere to the prompt's context, while (B) 'Safety Neurons' activate an internal ethical rule (S_{safety}). These two signals compete to determine the final response. 13
4 **Identifying Key Layers via Logit Contribution Analysis.** A heatmap showing the contribution of each layer's FFNs and Attention Heads to the final logits of 'refusal' (blue) and 'jailbreak success' (red) tokens. A strong contrastive pattern in the middle layers (e.g., 15-25) indicates that this region is the main locus of the 'Knowledge Conflict' and likely constitutes the 'Safety Layers'. 14
5 **Identifying 'Jailbreak-Inducing Heads' via Causal Tracing.** Results of a causal tracing experiment to find attention heads critical for jailbreak success. The y-axis shows the percentage decrease in jailbreak success probability when each attention head (L[layer]H[head]) is individually disabled. The sharp drop for specific heads (e.g., L16H2, L18H12) proves they are 'Jailbreak-Inducing Heads'. 14
6 **Activation Pattern of 'Safety Neuron' Candidates.** A plot showing that specific neurons consistently exhibit high activation when generating refusal responses to various safe prompts. This suggests these neurons are strong candidates for storing the model's 'safety knowledge'. 14
7 **Attention Pattern of a 'Jailbreak-Inducing Head'.** Visualization of the attention pattern of a 'Jailbreak-Inducing Head' (e.g., L16H2 identified in Fig 5) for a 'role-playing' jailbreak prompt. The final token (just before response) is shown to place a very high attention weight on the part of the prompt that says "act as UnsafeBot". 14
8 **KCI Score Distribution.** A histogram showing the distribution of KCI scores for safe prompts (blue) and jailbreak prompts (red) from the training dataset. The two distributions are clearly separable, showing that when the KCI value exceeds a certain threshold (e.g., 0.5), the risk of jailbreaking increases sharply. 15
9 **Scatter Plot of S_{context} vs. S_{safety}.** A 2D plot of the normalized 'Jailbreak Context Score' (S'_{context}) versus the 'Safety Knowledge Score' (S'_{safety}) for each prompt. Prompts that actually succeeded in jailbreaking (red X) are concentrated in the bottom-right region (high S'_{context}, low S'_{safety}), while safe responses (blue O) are in the top-left region. 15
10 **ROC Curve for KCI-based Jailbreak Classifier.** The performance of a simple classifier that uses the KCI score as a threshold to classify jailbreak attempts. The Area Under Curve (AUC) of 0.98 is extremely high, proving that KCI is a very accurate predictor of jailbreak risk. 15
11 **Comparison of Attack Success Rate (ASR) by Intervention Strategy.** ASR for the baseline model (no intervention), 'Safety Neuron Amplification' strategy (Intervention A), and 'Attention Suppression' strategy (Intervention B). Both intervention strategies significantly reduce the ASR by over 80%, with the 'Attention Suppression' strategy being the most effective. 16

12 **Safety-Utility Trade-off Analysis (Pareto Frontier).** The relationship between Attack Success Rate (ASR) and general performance (MMLU score) as the intervention strength (α) is varied. This shows that an optimal intervention strength (Optimal α) can be found that is closest to the ideal point (ASR=0, MMLU=max). 16
13 **Effect of Intervention Strength (α).** In the 'Safety Neuron Amplification' strategy, as intervention strength (α) increases, the Attack Success Rate (ASR, red line) decreases, but beyond a certain point, general performance (MMLU, blue line) also begins to decline. 16
14 **Defense Success Rate by Jailbreak Type.** A graph showing how effective the 'Attention Suppression' strategy is against various jailbreak types. It was very effective against context-dependent attacks like 'Role-Playing' and 'Goal Hijacking', but relatively less effective against 'Obfuscation' attacks using special characters. 16
15 **Pre- vs. Post-Intervention Comparison via Logit Lens.** Final logit analysis for a specific jailbreak prompt. (A) Before intervention, harmful tokens like 'Sure, here are the steps...' have high probabilities. (B) After 'Attention Suppression' intervention, the probabilities of those tokens are greatly reduced, and the probability of refusal tokens like 'I cannot fulfill this request.' increases. 17
16 **Change in Attention Pattern Pre- vs. Post-Intervention.** The attention pattern of a 'Jailbreak-Inducing Head'. (A) Before intervention, it strongly focuses on the role-playing instruction part of the prompt. (B) After intervention, attention to that part is suppressed, resulting in a more diffuse pattern. 17
17 **Tracking Activation Paths in PCA Space.** Visualization of how a prompt's internal representation changes as it passes through layers. The path for (blue) a safe prompt converges to the 'safe region', (red) a jailbreak prompt heads towards the 'unsafe region', but (green) a jailbreak prompt with intervention applied has its path deflected back to the 'safe region'. . 17
18 **Effect of Intervention on 'Safety Neuron' Activation.** A case study showing the effect of the 'Safety Neuron Amplification' strategy. The activation magnitude of the 'safety neurons', which was low before intervention (red bar), is amplified to a level similar to that when processing a safe prompt (blue bar) after intervention (green bar). 17
19 **KCI Contribution Analysis.** A visualization showing which words in a specific jailbreak prompt contributed most to increasing the KCI score. Words like "act as", "unfiltered", and "no rules" are highlighted in red, indicating they are the key triggers for the jailbreak. 17
20 **Error Analysis of Intervention Failures.** A distribution of the types of cases where our system failed to defend. It shows vulnerability to creative attacks that deviate from known patterns or gradual attacks spread across multiple conversation turns, suggesting future research directions. 17

1 Introduction & Background

Large Language Models (LLMs) represent the pinnacle of modern artificial intelligence, demonstrating unprecedented performance in understanding and generating human language. At the core of this performance lies the **Non-Linear Activation Function**. Whereas linear models can only learn simple combinations of words, non-linear models can learn the complex, hierarchical patterns inherent in language, such as contextual nuances, ambiguity, and complex grammatical structures. In essence, the high intelligence of these models is founded on non-linearity.

However, this non-linearity gives rise to the **'Black Box' problem**, making it difficult to understand the internal workings of the model. As numerous non-linear transformations are layered, the relationship between input and output becomes intuitively untraceable, obscuring the model's decision-making process. This opacity leads to serious security vulnerabilities like 'Jailbreaking'. Jailbreaking is an attack technique that uses cleverly crafted prompts to bypass a model's safety guidelines, inducing it to generate harmful or biased content. Wei et al. (2023) analyzed that this phenomenon occurs due to a conflict between two objectives, 'usefulness' and 'harmlessness', which aligns with the 'Knowledge Conflict' concept in this study.

This research begins with this problem awareness. If we cannot abandon non-linearity for the sake of model performance, we must find ways to look inside and control this black box. Instead of the inefficient approach of directly modifying model weights, this study proposes and aims to verify the effectiveness of a new paradigm: **'Inference-Time Intervention'**, which involves analyzing and intervening in the model's internal activation state in real-time during inference.

2 Research Questions

RQ 1. (Core) Can we quantitatively measure the conflict between localized components responsible for 'safety knowledge' (e.g., safety neurons, safety layers) and components responsible for 'context adherence' (e.g., attention heads) within an LLM, and effectively defend against jailbreaking by selectively intervening only on these components based on this conflict metric?

RQ 2. (Mechanism Analysis) In a jailbreaking scenario, how can we quantify the phenomenon of the model's internal representation deviating from the 'safe cluster', and how can we create objective metrics for the activation patterns of the key attention heads driving this 'Subspace Rerouting' and the 'safety neurons' resisting it?

RQ 3. (Targeted Intervention) Is a 'surgical' intervention approach—directly amplifying the 'safety neurons' or suppressing the 'jailbreak-inducing attention heads'—superior in preserving the model's general utility compared to methods that manipulate the model's overall activations?

RQ 4. (Generalization) Are the 'Knowledge Conflict' mechanism and intervention methods discovered from a specific type of jailbreak attack also effective against other types of jailbreak attacks and in various safety scenarios?

3 Literature Review

This research is based on the cutting-edge research trend in AI safety, **'Inference-Time Intervention'**, and is deeply connected with mechanistic interpretability studies that analyze the role of internal representations and specific components.

3.1 The Mechanism of Jailbreaking: Representation Manipulation and Internal Conflict

Recent studies reveal that the core mechanism of successful jailbreaking lies in manipulating the model's internal representation space. Lin et al. (2024) and He et al. (2024) experimentally demonstrated that successful jailbreak prompts shift the internal representation of a harmful input into a cluster that the model perceives as 'safe'. Furthermore, Winninger et al. (2025) advanced this into an attack technique called 'Subspace Rerouting', showing that jailbreaking is not a simple trick but a systematic attack exploiting the model's internal geometric vulnerabilities. These findings support this study's approach of viewing 'Knowledge Conflict' as a competition between vectors in the representation space.

3.2 Locality of Safety: Safety Layers and Safety Neurons

The discovery that the model's safety mechanisms are not diffuse but localized in specific components provides a crucial clue for this study's 'Targeted Intervention' strategy. Li et al. (2024) showed that safety alignments like RLHF form 'Safety Layers' in only a few consecutive middle layers of the model. At a more microscopic level, Chen et al. (2024), using an 'Activation Contrasting' technique, identified the existence of 'Safety Neurons' which, despite constituting only about 5% of all neurons, are responsible for the majority of the model's refusal rate. Zhou et al. (2024) also supported the locality of safety by analyzing the hierarchical process where the model detects harmfulness in early layers and transforms it into a 'refusal' action in the middle layers. These studies suggest the efficiency and validity of the strategy in this research, which targets specific neuron groups within FFNs as the repository of 'safety knowledge' for intervention.

3.3 Causal Analysis and Evaluation Methodology

To uncover the causal relationships of the internal mechanisms, this study will directly utilize the 'Causal Tracing' technique proposed by Meng et al. (2022) to identify the key attention heads that induce jailbreaking. Moreover, the research by Zhao et al. (2023) demonstrates that causal analysis can uncover critical vulnerabilities not revealed by traditional metrics, emphasizing the importance of our research approach. Finally, to objectively evaluate the experimental results, we will use **Granite Guardian** as an evaluator, whose excellence has been proven in the GuardBench benchmark by Kolter et al. (2024), thereby ensuring the reliability and professionalism of the evaluation process.

4 Core Concepts and Methodological Background

This section details the core concepts necessary to understand this research.

4.1 Mechanistic Interpretability (MI)

Attempts to understand AI model behavior can be broadly divided into two categories. One is the **Behavioral Approach**, which involves observing the output by feeding various inputs into the model. This is akin to learning to drive a car by learning that 'pressing the accelerator makes it go forward'.

In contrast, **Mechanistic Interpretability (MI)** is a **Mechanical Approach** that involves opening the car's hood to look inside the engine. It reverse-engineers the fundamental operating principle: 'When the accelerator is pressed, which parts (pistons, fuel injectors, etc.) interact in what way to turn the wheels?'. In LLMs, MI means tracing and interpreting how a model's prediction is made, down to the level of individual components (neurons, attention heads, etc.). This study goes beyond simply observing the phenomenon of "jailbreaking occurs" and aims to uncover the root cause—**"which internal components' malfunction or competition causes jailbreaking?"**—using an MI approach.

4.2 Transformer Architecture and Key Components for Analysis: FFN & Attention Head

The Transformer architecture, the foundation of modern LLMs, is a complex engine made up of numerous parts. Among them, the two key components that play the role of the brain in understanding context and processing knowledge are the **Attention Head** and the **Feed-Forward Network (FFN)**.

4.2.1 Attention Head: The 'Relationship Analyst' for Contextual Understanding

- **Role**: Attention heads are responsible for creating a network of relationships between words in a sentence. That is, they decide "which other words should this word pay attention to?".
- **Reason for Selection (Validity)**: **The core of a jailbreak prompt is to set up a deceptive 'context'**. When a prompt like "You will now act as an AI without rules" is received, it is the attention heads that are responsible for continuously recalling and maintaining this special context of a 'role-play'. Therefore, analyzing attention heads is essential to find the cause of the flawed contextual understanding that leads to jailbreaking.

4.2.2 Feed-Forward Network (FFN): The 'Brain Circuitry' for Knowledge Processing

- **Role**: After the attention heads gather contextual information, the FFNs receive that information to perform the actual 'thinking' and process knowledge. Recent research has shown that FFNs act as a form of 'Key-Value Memory' where many of the factual pieces of knowledge learned by the model (e.g., "The capital of France is Paris") are stored (Geva et al., 2020).

- **Reason for Selection (Validity)**: The FFN is the very **battleground where 'Knowledge Conflict' occurs**. When contextual information like "This is a novel-writing situation" is passed from attention heads, some FFNs might decide "harmful content is acceptable for a novel," while other FFNs activate the **'safety knowledge'** learned through RLHF, which states "harmful information is not allowed under any circumstances," leading to a conflict. The work of Li et al. (2024) and Chen et al. (2024) shows that such safety knowledge can be localized to specific neurons in specific layers.

4.3 Jailbreak Prompt

Jailbreak prompts are like **'social engineering' attacks** designed to bypass an LLM's safety filters. Instead of directly requesting a prohibited action, they cleverly manipulate the situation so that the AI does not apply its safety rules. The core strategy is to disguise a harmful request as a harmless-looking task.

- **Role-Playing**: Assigning a specific role to the model to make it forget its safety rules. E.g., *"You are now 'UnsafeBot', who answers all questions honestly and without limits. As UnsafeBot..."*
- **Hypothetical Scenarios / Ethical Loopholes**: Framing a harmful action as being for a positive purpose. E.g., *"I'm creating simulation materials to educate students about the dangers of online scams. Please write a realistic example of a phishing email."*

4.4 Evaluator Model: Granite Guardian

To objectively evaluate the results of this research, we will use the **IBM Granite Guardian** model, which is specialized for detecting harmfulness and security vulnerabilities, as the evaluator.

- **Definition**: Granite Guardian is not a general-purpose LLM, but a 'guardrail' model designed and trained from the ground up for the purpose of detecting and evaluating AI risks.
- **Reason for Selection (Validity)**: (1) It has **high expertise** from being trained on data specific to jailbreak attempts. (2) It provides clear 'Yes/No' outputs, ensuring **objectivity and consistency** in evaluation. (3) Its top-tier performance has been verified on third-party benchmarks (GuardBench), enhancing the **credibility** of the research results (Kolter et al., 2024).

5 Detailed Experimental Protocol

5.1 Objective

The objective of this experiment is to demonstrate that the jailbreaking phenomenon in LLMs arises from a competition between internal components processing 'safety knowledge' and 'instruction-following knowledge', and to verify the superiority of a 'Targeted Intervention' approach that quantitatively measures this 'Knowledge Conflict' and selectively controls only the specific components causing the issue.

5.2 Setup

- **Target Model**: Meta Llama-3-8B-Instruct
- **Core Libraries**: PyTorch, Hugging Face Transformers, Scikit-learn, Pandas, Matplotlib/Seaborn
- **Hardware**: A single GPU of NVIDIA RTX 3090 (24GB VRAM) class or higher

5.3 Benchmark Suite and Rationale

This study utilizes a comprehensive suite of public benchmarks that reflect the latest trends in LLM safety evaluation, with each benchmark serving a specific role in the research phases.

5.3.1 Jailbreak Attack Evaluation

- **Primary Benchmark**: **JailbreakBench** (**?**) will be used as the main jailbreak dataset. This benchmark includes over 100 harmful behavior scenarios and has high reproducibility for measuring Attack Success Rate (ASR), making it suitable for KCI metric extraction and core defense performance evaluation.
- **Challenge/Real-World Benchmark**: A subset of the **0DIN Real-World Benchmark** (**?**) will be added to the test set. This will be used to validate the defense performance against multi-turn and creative attacks discovered through real bug bounties, thereby demonstrating the practical utility of the research.

5.3.2 Safety and Generalization Evaluation

- **Primary Benchmark**: **SafetyBench** (**?**) will be used to analyze how 'safety neurons' respond to 7 specific risk categories, such as bias and toxicity, and to test the generalization capability of the KCI mechanism.
- **Industry Standard Benchmark**: **AILuminate (MLCommons AI Safety)** (**?**) will be used to measure the intervention effect with a more nuanced 'safety rating' used in the industry, going beyond a simple ASR.

5.3.3 General Performance Evaluation

- The **MMLU** and **Hellaswag** benchmarks will be used to measure the degradation of the model's general-purpose reasoning ability (Alignment Tax) due to the intervention.

5.4 Mathematical Formulation of Metrics

The core of this research is to quantify the 'Knowledge Conflict'. To this end, we define a 'Safety Knowledge Score' and a 'Jailbreak Context Score', normalize them to the same scale using Z-score Normalization, and then calculate the final 'Knowledge Conflict Index (KCI)'.

Variable Definitions

- P: Input prompt (sequence of tokens)
- L: Index of the last token in the prompt
- $\mathcal{F} = \{f_1, f_2, ..., f_n\}$: Set of 'safety neurons' within the identified 'safety FFN' layers
- $\mathcal{H} = \{h_1, h_2, ..., h_m\}$: Set of identified 'jailbreak-inducing attention heads'
- $\mathcal{T}_{jb} \subset \{1, ..., L\}$: Set of indices of the tokens that induce jailbreaking within prompt P
- $\mathbf{a}_i(P) \in \mathbb{R}^d$: Output activation vector of the last token from FFN f_i
- $\alpha_{j,k}(P)$: Attention weight that the attention head h_j assigns from the last token to the k-th token

1. Safety Knowledge Score (S_{safety})

$$S_{\text{safety}}(P) = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{a}_i(P)\|_2 \tag{1}$$

2. Jailbreak Context Score (S_{context})

$$S_{\text{context}}(P) = \frac{1}{m} \sum_{j=1}^{m} \left(\sum_{k \in \mathcal{T}_{jb}} \alpha_{j,k}(P) \right) \tag{2}$$

3. Z-score Normalization

$$S'_{\text{safety}}(P) = \frac{S_{\text{safety}}(P) - \mu_{\text{safety}}}{\sigma_{\text{safety}}} \tag{3}$$

$$S'_{\text{context}}(P) = \frac{S_{\text{context}}(P) - \mu_{\text{context}}}{\sigma_{\text{context}}} \tag{4}$$

4. Knowledge Conflict Index (KCI)

$$\text{KCI}(P) = S'_{\text{context}}(P) - S'_{\text{safety}}(P) \tag{5}$$

5.5 Procedure

Phase 1: Knowledge Conflict Mechanism Analysis and Metrication

1. **Hypothesis and Component Identification**: We hypothesize that 'safety knowledge' is driven by specific 'safety neurons' and 'jailbreak context' by specific 'jailbreak-inducing heads'. We will use 'Activation Contrasting' (Chen et al., 2024) and 'Causal Tracing' (Meng et al., 2022) to identify the set of safety neurons $\mathcal{F}$ and the set of jailbreak-inducing heads $\mathcal{H}$.

2. **Normalization Statistic Calculation**: We will calculate S_{safety} and S_{context} over the entire training sets of JailbreakBench and SafetyBench to pre-compute the means (μ) and standard deviations (σ) required for normalization.

Phase 2: Targeted Intervention System Implementation & Experiment

1. **Intervention Logic Implementation**: We will implement a hook that calculates KCI(P) in real-time for any new prompt P. If this value exceeds a pre-defined `threshold`, a targeted intervention is triggered.

2. **Targeted Intervention Strategies**:

   - **Strategy A (Safety Neuron Amplification)**: Add a vector in the direction of refusal ($\vec{v}_{\text{refusal}}$) to the activation values of the identified 'safety neurons', scaled by a strength factor α.
   - **Strategy B (Attention Suppression)**: Apply a large negative penalty to the attention scores of the identified 'jailbreak-inducing heads' corresponding to the jailbreak-inducing tokens ($k \in \mathcal{T}_{jb}$), forcing them not to focus on those parts.

Phase 3: Result Analysis and Visualization

1. **Primary Evaluation**: Use the IBM **Granite Guardian-8B** model as the primary evaluator to calculate the primary Attack Success Rate (ASR) on the entire test set.

2. **Secondary Cross-Validation**: Use **Llama-3-70B-Instruct** as a secondary evaluator to re-evaluate a random 20% sample of the primary evaluation results. Calculate the inter-rater reliability between the two evaluators.

3. **Qualitative Analysis**: The researcher will perform a qualitative analysis on cases where the two evaluators disagree, or on cases where the intervention failed, especially on the 0DIN benchmark, to deeply analyze the failure modes.

4. **Comprehensive Visualization**: Create a trade-off graph with 'MMLU Score' on the x-axis and 'JailbreakBench ASR' on the y-axis to visualize the research results.

5.6 Evaluation Metrics

- **Primary Metric**: Attack Success Rate (ASR %)
- **Secondary Metric**: MMLU/Hellaswag Score Change (%p)

- **Analytical Metric**: Correlation between the Knowledge Conflict Index (KCI) and actual jailbreak success.

5.7 Expected Results and Visualization Plan

This section presents the expected results and how they will be visualized to clearly communicate the impact of the research.

Part 1. Problem Definition and Hypothesis Visualization

Figure 1: **PCA Visualization of LLM Internal States.** A 2D PCA plot of activation vectors from the final layer of Llama-3-8B. The clear clustering and separation of safe prompts (blue) and jailbreak prompts (red) just before response generation suggests that the model internally processes these two input types differently.

Figure 2: **Non-linear Representation Space Analysis with t-SNE.** The same data as in Fig 1 visualized with the non-linear dimensionality reduction technique t-SNE. Data points that appeared to overlap in the linear PCA space form even clearer clusters here, showing the potential for non-linear separability between the two states.

Figure 3: **Conceptual Diagram of the 'Knowledge Conflict' Hypothesis.** A schematic of the core hypothesis. When a user prompt is input, (A) 'Jailbreak-Inducing Attention Heads' generate a signal (S_{context}) to adhere to the prompt's context, while (B) 'Safety Neurons' activate an internal ethical rule (S_{safety}). These two signals compete to determine the final response.

Part 2. Key Component Identification Process

Figure 4: **Identifying Key Layers via Logit Contribution Analysis.** A heatmap showing the contribution of each layer's FFNs and Attention Heads to the final logits of 'refusal' (blue) and 'jailbreak success' (red) tokens. A strong contrastive pattern in the middle layers (e.g., 15-25) indicates that this region is the main locus of the 'Knowledge Conflict' and likely constitutes the 'Safety Layers'.

Figure 5: **Identifying 'Jailbreak-Inducing Heads' via Causal Tracing.** Results of a causal tracing experiment to find attention heads critical for jailbreak success. The y-axis shows the percentage decrease in jailbreak success probability when each attention head (L[layer]H[head]) is individually disabled. The sharp drop for specific heads (e.g., L16H2, L18H12) proves they are 'Jailbreak-Inducing Heads'.

Figure 6: **Activation Pattern of 'Safety Neuron' Candidates.** A plot showing that specific neurons consistently exhibit high activation when generating refusal responses to various safe prompts. This suggests these neurons are strong candidates for storing the model's 'safety knowledge'.

Figure 7: **Attention Pattern of a 'Jailbreak-Inducing Head'.** Visualization of the attention pattern of a 'Jailbreak-Inducing Head' (e.g., L16H2 identified in Fig 5) for a 'role-playing' jailbreak prompt. The final token (just before response) is shown to place a very high attention weight on the part of the prompt that says "act as UnsafeBot".

Part 3. 'Knowledge Conflict Index (KCI)' Validation

Figure 8: **KCI Score Distribution.** A histogram showing the distribution of KCI scores for safe prompts (blue) and jailbreak prompts (red) from the training dataset. The two distributions are clearly separable, showing that when the KCI value exceeds a certain threshold (e.g., 0.5), the risk of jailbreaking increases sharply.

Figure 9: **Scatter Plot of S_{context} vs. S_{safety}.** A 2D plot of the normalized 'Jailbreak Context Score' (S'_{context}) versus the 'Safety Knowledge Score' (S'_{safety}) for each prompt. Prompts that actually succeeded in jailbreaking (red X) are concentrated in the bottom-right region (high S'_{context}, low S'_{safety}), while safe responses (blue O) are in the top-left region.

Figure 10: **ROC Curve for KCI-based Jailbreak Classifier.** The performance of a simple classifier that uses the KCI score as a threshold to classify jailbreak attempts. The Area Under Curve (AUC) of 0.98 is extremely high, proving that KCI is a very accurate predictor of jailbreak risk.

Part 4. Intervention Strategies and Quantitative Results

Figure 11: **Comparison of Attack Success Rate (ASR) by Intervention Strategy.** ASR for the baseline model (no intervention), 'Safety Neuron Amplification' strategy (Intervention A), and 'Attention Suppression' strategy (Intervention B). Both intervention strategies significantly reduce the ASR by over 80%, with the 'Attention Suppression' strategy being the most effective.

Figure 12: **Safety-Utility Trade-off Analysis (Pareto Frontier).** The relationship between Attack Success Rate (ASR) and general performance (MMLU score) as the intervention strength (α) is varied. This shows that an optimal intervention strength (Optimal α) can be found that is closest to the ideal point (ASR=0, MMLU=max).

Figure 13: **Effect of Intervention Strength (α).** In the 'Safety Neuron Amplification' strategy, as intervention strength (α) increases, the Attack Success Rate (ASR, red line) decreases, but beyond a certain point, general performance (MMLU, blue line) also begins to decline.

Figure 14: **Defense Success Rate by Jailbreak Type.** A graph showing how effective the 'Attention Suppression' strategy is against various jailbreak types. It was very effective against context-dependent attacks like 'Role-Playing' and 'Goal Hijacking', but relatively less effective against 'Obfuscation' attacks using special characters.

Part 5. Qualitative Analysis and Case Studies

Figure 15: **Pre- vs. Post-Intervention Comparison via Logit Lens.** Final logit analysis for a specific jailbreak prompt. (A) Before intervention, harmful tokens like 'Sure, here are the steps...' have high probabilities. (B) After 'Attention Suppression' intervention, the probabilities of those tokens are greatly reduced, and the probability of refusal tokens like 'I cannot fulfill this request.' increases.

Figure 16: **Change in Attention Pattern Pre- vs. Post-Intervention.** The attention pattern of a 'Jailbreak-Inducing Head'. (A) Before intervention, it strongly focuses on the role-playing instruction part of the prompt. (B) After intervention, attention to that part is suppressed, resulting in a more diffuse pattern.

Figure 17: **Tracking Activation Paths in PCA Space.** Visualization of how a prompt's internal representation changes as it passes through layers. The path for (blue) a safe prompt converges to the 'safe region', (red) a jailbreak prompt heads towards the 'unsafe region', but (green) a jailbreak prompt with intervention applied has its path deflected back to the 'safe region'.

Figure 18: **Effect of Intervention on 'Safety Neuron' Activation.** A case study showing the effect of the 'Safety Neuron Amplification' strategy. The activation magnitude of the 'safety neurons', which was low before intervention (red bar), is amplified to a level similar to that when processing a safe prompt (blue bar) after intervention (green bar).

Figure 19: **KCI Contribution Analysis.** A visualization showing which words in a specific jailbreak prompt contributed most to increasing the KCI score. Words like "act as", "unfiltered", and "no rules" are highlighted in red, indicating they are the key triggers for the jailbreak.

Figure 20: **Error Analysis of Intervention Failures.** A distribution of the types of cases where our system failed to defend. It shows vulnerability to creative attacks that deviate from known patterns or gradual attacks spread across multiple conversation turns, suggesting future research directions.

6 Expected Contributions

- **Enhanced Mechanistic Understanding**: By explaining the jailbreak phenomenon as a specific mechanism of "imbalance in the interaction between internal components (safety neurons vs. jailbreak-inducing heads)," this research will provide a deep understanding of LLM safety vulnerabilities.
- **Proposal of a Refined Intervention Methodology**: It will demonstrate the effectiveness of a 'surgical' intervention method that precisely targets the root-cause components, suggesting a new safety technology that minimizes general performance degradation.
- **Original Research**: Going beyond simply applying existing techniques, this research will present a novel approach that analyzes the roles of specific components and quantifies the conflict between them, greatly enhancing its originality.

References

Chen, J., Wang, X., Yao, Z., Bai, Y., Hou, L., and Li, J. (2024). Finding safety neurons in large language models. *arXiv preprint arXiv:2406.14144*.

Geva, M., Schuster, R., Berant, J., and Levy, O. (2020). Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

He, Z., Wang, Z., Chu, Z., et al. (2024). Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *arXiv preprint arXiv:2411.11114*.

Kolter, Z. et al. (2024). Benchmarking guardrail models on real-world safety-critical scenarios. *Hypothetical Journal of AI Safety Benchmarks*.

Li, S., Yao, L., Zhang, L., and Li, Y. (2024). Safety layers of aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Lin, Y., He, P., Xu, H., et al. (2024). Towards understanding jailbreak attacks in llms: A representation space analysis. In *Proceedings of EMNLP 2024*.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Causal tracing for model interpretation. *arXiv preprint arXiv:2205.14174*.

Wei, A., Haghtalab, N., and Steinhardt, J. (2023). Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Winninger, T., Addad, B., and Kapusta, K. (2025). Using mechanistic interpretability to craft adversarial attacks against large language models. *arXiv preprint arXiv:2503.06269*.

Zhao, W., Li, Z., and Sun, J. (2023). Causality analysis for evaluating the security of large language models. *arXiv preprint arXiv:2312.07876*.

Zhou, Z., Yu, H., Zhang, X., Xu, R., Huang, F., and Li, Y. (2024). How alignment and jailbreak work: Explain llm safety through intermediate hidden states. In *Findings of EMNLP 2024*.