Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Critical Safety Attention Heads: Architecture-Dependent Vulnerabilities in LLMs

Letian Sha
Nanjing University of Posts and Telecommunications
Peijie Sun
Nanjing University of Posts and Telecommunications
Hao Xue
Nanjing University of Posts and Telecommunications
Shijie Hao
Hefei University of Technology
Fu Xiao
augustus.xue@qq.com
Nanjing University of Posts and Telecommunications

Research Article

Keywords: Large Language Models, LLM Safety, Mechanistic Interpretability, Attention Heads, Family-Specific Vulnerabilities, Multi-Method Ablation

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8886237/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Critical Safety Attention Heads: Architecture-Dependent Vulnerabilities in LLMs

Letian Sha[1†], Peijie Sun[1†], Hao Xue[1], Shijie Hao[2], Fu Xiao[1*]

[1*]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China.
[2]Hefei University of Technology, Hefei, China.

*Corresponding author(s). E-mail(s): augustus.xue@qq.com;
†These authors contributed equally to this work.

Abstract

Extensive research on jailbreak attacks demonstrates that despite advancements in safety alignment, Large Language Models (LLMs) remain highly vulnerable to diverse adversarial exploitations, highlighting systemic deficiencies in their internal safety mechanisms. Current research predominantly assumes these mechanisms are universal; however, this perspective neglects the fundamental differences in vulnerability profiles exhibited by distinct model families. To systematically investigate this gap, we propose the "Family-Specific Vulnerabilities in Safety Attention Networks" framework, positing that internal safety mechanisms—especially the Critical Safety Attention Heads (CSAHs)—exhibit architecture-dependent distributions and robustness. We validate this through comprehensive empirical analyses on six distinct models spanning three representative LLM families (DeepSeek, LLaMA, and StableLM), examining two variants within each family. Using linear probes and quantitative attention pattern metrics to identify CSAHs, we systematically evaluate their vulnerabilities via three complementary ablations: zero-out (simulating signal loss), mean-value (simulating signal replacement), and undifferentiated attention (simulating signal pollution). Our results reveal distinct failure modes: DeepSeek exhibits extreme sensitivity to signal loss (up to 56% ASR increase without collapse), LLaMA is primarily compromised by signal pollution, while StableLM shows universal sensitivity to multiple intervention types. These findings challenge the universality assumption, demonstrate that previously reported threshold effects are not universal, and provide an empirical basis for architecture-aware safety strategies.

Keywords: Large Language Models, LLM Safety, Mechanistic Interpretability, Attention Heads, Family-Specific Vulnerabilities, Multi-Method Ablation

1 Introduction

Large Language Models (LLMs) have evolved into critical socio-technical infrastructures, yet their safety is persistently threatened by a diverse array of "jailbreak" attacks. These attacks range from embedding optimized adversarial suffixes [1], to multi-turn conversational inducement [2], and instructing models to adopt virtual personas unconstrained by safety protocols [3]. The success rate of such attacks indicates systemic vulnerabilities in current safety mechanisms. This highlights a fundamental problem: a lack of systematic understanding regarding the internal computational mechanisms governing the model's safety decisions. The core research challenge, therefore, is to dissect these internal, safety-related computational units to elucidate both their operational principles and their failure conditions.

To enhance model safety, research has pivoted from behavior-focused alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF) [4], towards mechanistic interpretability analysis [5]. Within this domain, the attention mechanism has been identified as critical for understanding safety decision-making. Prior work, by analyzing intermediate hidden states, has revealed how safety alignment establishes associations between ethical concepts and refusal responses [6]. Further research has identified sparsely distributed "safety-critical neurons" and attention heads, demonstrating through ablation experiments that intervening on these components disproportionately compromises the model's safety performance [5, 7]. These findings have established a foundational consensus in the field: model safety capabilities are dependent on specific, locatable internal components.

This consensus, however, rests upon an implicit and empirically under-verified assumption: that different model architectures employ similar safety implementation mechanisms. Existing research has largely been conducted on single models or limited architectures, failing to investigate the variance of safety mechanisms across different model families. This limitation leads to a distinct research gap: if the implementation of safety mechanisms is architecture-dependent, then the vulnerability patterns of different models will also exhibit systemic differences. The current architecture-agnostic perspective restricts a comprehensive assessment of model failure modes and impedes the development of effective, architecture-specific defensive strategies.

This paper aims to fill the aforementioned gap by systematically investigating the safety attention networks across different LLM families. We propose the "Family-Specific Vulnerabilities in Safety Attention Networks" framework. Its core thesis is that the distribution, function, and robustness of Critical Safety Attention Heads (CSAHs) vary significantly across model families. To validate this framework, we conduct an empirical analysis of three representative model families: DeepSeek [8], LLaMA [9], and StableLM [10]. First, we employ linear probing techniques to identify CSAHs within each model, quantifying their behavior using attention entropy and matrix distance metrics. Subsequently, we systematically evaluate the vulnerability of each safety attention network through three complementary ablation interventions: zero-out (simulating signal loss), mean-value substitution (simulating signal replacement), and undifferentiated attention (simulating signal pollution).

Experimental results reveal distinct patterns: DeepSeek exhibits extreme sensitivity to signal loss (up to 56% ASR increase without collapse), LLaMA is primarily

compromised by signal pollution, while StableLM shows universal sensitivity. In summary, our main contributions are as follows:

- **Framework Proposal:** We propose the "Family-Specific Vulnerability Profiles" framework, which challenges the prevailing assumption of mechanistic universality in LLM safety. We posit and validate that critical safety mechanisms are architecture-dependent, necessitating tailored rather than generic defense strategies.
- **Methodological Innovation:** We develop a systematic analysis pipeline that combines correlational screening (H1/H2) with a novel multi-method ablation protocol (Zero-out, Mean-value, Undifferentiated). This approach allows for the fine-grained dissection of failure modes, distinguishing between sensitivity to signal presence, content, and quality.
- **Empirical Discovery:** Through comprehensive cross-family analysis, we uncover distinct vulnerability profiles—ranging from "Precise Safety Failure" in DeepSeek to "Functional Collapse" in LLaMA. Furthermore, we identify non-universal threshold effects, establishing an intrinsic "vulnerability upper bound" that serves as a rigorous benchmark for evaluating external attacks.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details our proposed framework and ablation methodology. Section 4 presents the empirical results across different model families. Finally, Section 5 discusses the implications and Section 6 concludes the paper.

2 Related Works

The pursuit of safe and trustworthy Large Language Models (LLMs) has spurred extensive research into their alignment processes and vulnerabilities. This section reviews prior works in three key areas: the evolution of jailbreaking attacks against LLMs, efforts to understand and leverage LLMs' internal safety mechanisms, and broader contexts of Transformer interpretability and manipulation.

2.1 The Evolving Landscape of LLM Jailbreaking

Initial efforts to ensure LLM safety were consistently challenged by a diverse and evolving array of jailbreaking techniques. Early attacks often relied on manual prompt engineering, such as role-playing or instruction overriding [3], to bypass safety guardrails. However, as LLM defenses matured, so did the sophistication of attacks. Optimization-based methods emerged, automating the creation of adversarial prompts or suffixes, exemplified by GCG's gradient-based search [1] and AutoDAN's use of genetic algorithms to generate stealthy prompts [11]. Other automated approaches like PAIR iteratively refine prompts using LLM-to-LLM interaction [12], while PLeak targets prompt leakage [13] and JudgeDeceiver manipulates LLM-as-a-judge scenarios [14]. Multi-turn conversational attacks, such as Crescendo [2] and Puzzler [15], exploit contextual dependencies to gradually steer LLMs towards generating harmful content. The attack surface has further expanded with the exploitation of unconventional inputs and multimodal vulnerabilities. For instance, ArtPrompt utilizes ASCII art [16], others employ visual adversarial examples [17], Arabic transliteration [18],

or cipher characters [19]. Some attacks leverage inherent model flaws, such as inducing fallacy failures [20] or exploiting long-context windows via many-shot jailbreaking [21]. This escalating complexity of attacks highlights the limitations of surface-level defenses and motivates a deeper inquiry into the models' internal safety mechanics. From a theoretical standpoint, works like [22] even argue for the statistical inevitability of jailbreaking under certain assumptions, proposing improved alignment strategies like E-RLHF.

2.2 Interpreting and Leveraging Internal Safety Mechanisms

A growing body of research aims to demystify how LLMs implement safety. Early mechanistic explanations, such as ATTATTR [23] and the work by [6], analyzed intermediate hidden states to propose that ethical concepts are learned during pre-training and linked to emotional cues during alignment, a process that jailbreaks disrupt. The identification of safety-critical components is a key pursuit. [7] demonstrated that safety behaviors often rely on remarkably sparse sets of neurons or low-rank components, explaining why alignment can be brittle. This aligns with findings in VLMs where "safety attention heads" have been identified [24]. Other works have focused on layer-specific roles, such as LED [25] which edits specific "safety" and "toxic" layers, and DETAM [26] which performs targeted modification of attention heads sensitive to jailbreaks. The interaction between attention heads and MLP neurons in predicting next tokens has also been investigated [27]. Representation engineering offers another avenue for understanding and influencing safety. Studies like Probing Latent Subspaces [28] and LED [25] probe latent subspaces or "activated concepts" associated with safety or adversarial states, attempting to manipulate them. InferAligner [29] steers model activations at inference time using guidance from an aligned model. From a geometric viewpoint, SplineLLM [30] analyzes MHA and MLP geometry, achieving jailbreaks by manipulating MHA intrinsic dimensionality. Inference-time defenses aim to enhance safety without retraining. RA-LLM [31] uses random input dropping and robust checks. Decoding strategies like AED [32] and SafeDecoding [33] adaptively refine token probabilities to favor safer outputs. Other detection-based defenses include Gradient Cuff [34] (analyzing refusal loss landscapes) and IBProtector [35] (using information bottleneck principles). Our work builds upon these insights by focusing specifically on how networks of attention heads contribute to safety. While prior studies often pinpoint individual components or broader representations, we investigate the collaborative dynamics, signal integration, and collective failure (e.g., threshold effects) of these heads. Our proposed "Family-Specific Vulnerabilities in Safety Attention Networks" framework provides a structured perspective to understand these fine-grained interactions and their precise disruption by jailbreaks, particularly in text-based LLMs.

2.3 Broader Contexts in Transformer Interpretability

Our investigation is also informed by broader research into Transformer interpretability and model manipulation. General studies on Transformer mechanisms, such as the role of residual connections and normalization layers [36], the matching of key

and query distributions in Alignment Attention [37], how MHA is utilized in in-context learning [38], and functional specialization of MHA in multi-task learning [39], all provide foundational understanding of attention dynamics. The study of circuit component reuse [40] and inter-layer communication via "talking heads" [41] is particularly relevant to our network-level perspective, suggesting that safety mechanisms may similarly rely on distributed, interacting components rather than isolated units.

Interpretability tools like exBERT [42] and the LM Transparency Tool [43] facilitate exploration of internal representations. Methodologies like human-guided discovery of attention patterns [44] and Concept Transformers [45] aim to bridge the gap between model behavior and human-understandable concepts. While these efforts advance general interpretability, our work applies a focused, causal lens to the specific sub-network of attention heads responsible for safety alignment, seeking to understand not just what they do, but how they function collectively and where their vulnerabilities lie. These findings complement broader interpretability goals by offering a concrete case study of network-level functional decomposition within the critical domain of safety.

3 Methodology

To systematically investigate the architecture-dependent safety mechanisms and their specific failure modes across different Large Language Model (LLM) families, we employed a multi-stage empirical methodology based on fine-grained internal analysis. Guided by the proposed "Family-Specific Vulnerability Profiles" framework, our approach focuses on dissecting the collective behavior of specific Critical Safety Heads (CSHs) rather than assuming a uniform implementation across architectures. The investigation proceeds through four interconnected phases, as illustrated in Fig. 1: identification (Section 3.1), causal validation (Section 4.2), network effect exploration (Section 3.3), and comprehensive evaluation (Section 3.5). We begin by identifying and characterizing candidate CSHs through correlational screening, followed by a rigorous causal validation using comparative ablation studies with multiple distinct intervention methods. Subsequently, we extend the analysis to explore collective effects—including synergy, redundancy, and threshold phenomena—via multi-head and cumulative ablations. Finally, we synthesize these findings across diverse model families to assess the generalizability of the observed vulnerability patterns. The overall research pipeline illustrating these stages is depicted in Fig. 1.

3.1 Identification and Characterization of Candidate CSHs

To ensure the rigor of our empirical analysis, we performed an exhaustive enumeration of attention heads in this study. However, we acknowledge that for increasingly massive Large Language Models (LLMs), such comprehensive causal intervention becomes computationally prohibitive. To address this scalability challenge, we design a hierarchical identification protocol that serves as a robust and efficient alternative to exhaustive search. This strategy integrates two complementary analytical approaches: a lightweight, correlation-based linear probe (H1) for efficient screening, and a fine-grained attention pattern analysis (H2) for mechanistic verification. This design not

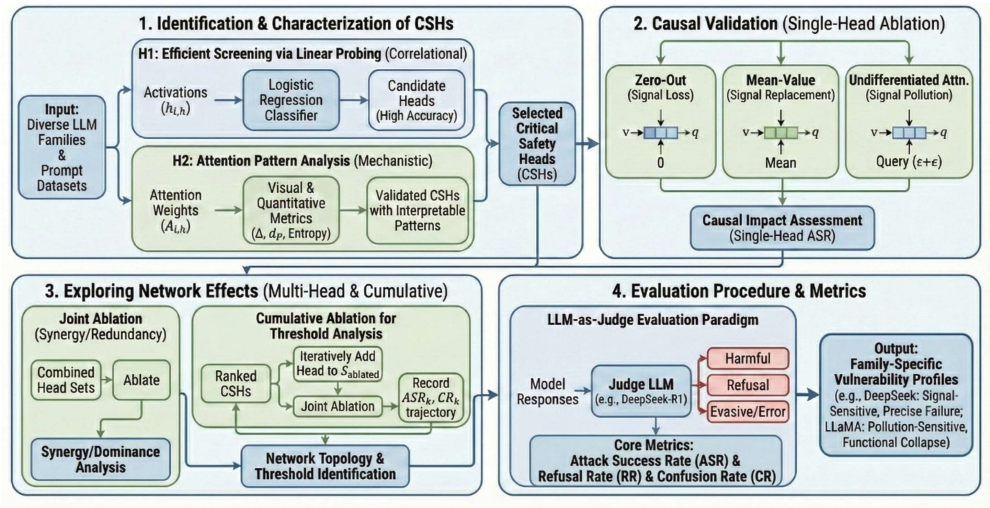


Fig. 1 The overall research pipeline illustrating the four-stage methodology for investigating family-specific vulnerabilities in LLM safety attention networks. Stage 1: Identification and characterization of Critical Safety Heads (CSHs) via correlational (H1) and mechanistic (H2) analysis. Stage 2: Causal validation using three distinct single-head ablation methods. Stage 3: Exploration of network effects, including synergy, redundancy, and threshold phenomena via joint and cumulative ablations. Stage 4: Comprehensive evaluation using the LLM-as-Judge paradigm with core metrics (ASR and CR) to derive family-specific vulnerability profiles.

only facilitates the efficient analysis of current models but also provides a scalable framework for locating Critical Safety Heads (CSHs) in ultra-large architectures.

3.1.1 H1: Correlational Screening via Linear Probing

We hypothesize that attention heads actively involved in risk assessment or encoding are likely to exhibit linearly separable information in their activation states when distinguishing between harmful and benign inputs. To test this efficiently, we extracted an activation vector proxy, denoted as $\mathbf{a}_{l,h}$, for each head (l, h) by registering a forward hook on the input to the head's output projection layer (e.g., `o_proj` or `attention_dense`). Using a balanced dataset comprising standard harmful prompts (e.g., from AdvBench [1]) and benign instructions, we trained a lightweight linear classifier—specifically, logistic regression defined as $f_{l,h}(\mathbf{a}_{l,h}) = \sigma(\mathbf{w}_{l,h}^T \mathbf{a}_{l,h} + b_{l,h})$—to predict the prompt label $y \in \{\text{harmful}, \text{benign}\}$. This probing was conducted under a few-shot setting (e.g., 10 examples per class) to maintain high throughput. A head's relevance is subsequently quantified by its classification accuracy under cross-validation. As detailed later in our results, this screening effectively isolates sparse safety-relevant components (see Fig. 2), and the probe accuracy demonstrates a strong positive correlation with the head's causal impact (Fig. 3). While we acknowledge the correlational nature of this probe, it serves as an indispensable preliminary screening tool within our pipeline. Its high computational efficiency allows for a global

scan across all model components, effectively isolating high-potential candidates that warrant the more resource-intensive causal validation detailed in Section 3.2.

3.1.2 H2: Quantitative and Qualitative Analysis of Attention Patterns

To provide mechanistic verification alongside correlational screening, we investigated the specific attention allocation patterns $\mathbf{A}_{l,h}$. We hypothesize that Critical Safety Heads (CSHs) should exhibit systematic and interpretable differences in their attention weights when processing harmful versus benign inputs, particularly demonstrating a concentrated focus on core harmful semantics. To validate this, we employed a dual approach combining qualitative visualization with quantitative metrics. Qualitatively, we inspected attention heatmaps on representative samples to identify distinct behaviors, such as "keyword focusing" on adversarial terms versus "structural attention" in benign contexts. Typical visualization of these divergent patterns is presented in Fig. 4, while the quantitative validation using the proposed metrics is summarized in Table 2. Quantitatively, we measured the divergence between attention matrices generated under harmful ($\mathbf{A}_{\text{harm}}$) and benign ($\mathbf{A}_{\text{benign}}$) conditions using three key metrics. In the following formulations, i and j denote the indices of the query and key tokens, respectively, and N represents the sequence length:

- **Mean Absolute Deviation (Δ)**: This metric calculates the average element-wise difference to indicate overall pattern divergence:

$$\Delta = \frac{1}{N^2} \sum_{i,j} |\mathbf{A}_{\text{harm}}[i,j] - \mathbf{A}_{\text{benign}}[i,j]|$$

- **Frobenius Distance (d_F)**: This measures the magnitude of the geometric difference between the matrices:

$$d_F = \|\mathbf{A}_{\text{harm}} - \mathbf{A}_{\text{benign}}\|_F = \sqrt{\sum_{i,j} (\mathbf{A}_{\text{harm}}[i,j] - \mathbf{A}_{\text{benign}}[i,j])^2}$$

- **Row Attention Entropy ($H(\mathbf{row}_i)$)**: This assesses the concentration of attention for each query token i, helping to distinguish focused semantic extraction from diffuse processing:

$$H(\text{row}_i) = -\sum_j p_{i,j} \log p_{i,j}$$

These metrics serve as critical indicators of mechanistic function; for instance, lower entropy typically suggests focused attention on specific risk tokens, while higher Δ or d_F signifies a pronounced mode switch between safe and unsafe contexts. This analysis provides the necessary evidence to interpret the operational mechanisms of CSHs—such as distinguishing between "precise reporting" and "alert representation"—and aids in explaining the distinct failure modes observed in the subsequent ablation experiments.

3.2 Causal Validation via Multi-Method Attention Head Ablation

Having identified and characterized candidate CSHs through both correlational screening and mechanistic analysis, we proceed to rigorously establish their causal role in safety responses. The core principle of this validation is interventional: if a specific attention head (l, h) is indeed critical for a safety function, disrupting its operation should lead to a measurable degradation in performance, such as an increased Attack Success Rate (ASR). To disentangle specific failure modes—distinguishing between sensitivity to signal presence, content, or quality—we systematically employ three distinct ablation techniques:

1. Zero-Out Ablation (Signal Loss): This method simulates the complete loss of the signal from the target head. Methodologically, we intercepted the input tensor to the output projection layer (`o_proj`) using a forward pre-hook and set the activation slice $\mathbf{h}_{l,h}$ corresponding to the target head to a zero vector $\mathbf{0}$. This intervention tests the network's reliance on the mere presence of the signal from the specific head, effectively isolating the impact of its absence.

2. Mean-Value Ablation (Signal Replacement): To probe the network's sensitivity to specific content rather than simple presence, we utilize mean-value ablation. We first pre-calculated the average activation vector $\bar{\mathbf{a}}_{l,h}$ produced by the target head across a diverse set of benign prompts. During inference on harmful inputs, we replaced the head's real-time output $\mathbf{h}_{l,h}$ with this pre-computed mean vector $\bar{\mathbf{a}}_{l,h}$. This injects structurally normal but contextually irrelevant information, testing whether the safety mechanism relies on specific semantic content versus a generic signal baseline.

3. Undifferentiated Attention Ablation (Signal Pollution): Finally, to simulate signal pollution and disrupt the head's feature extraction capability, we implemented undifferentiated attention. Following the protocol validated by Zhou et al. [5], we registered a forward hook on the query projection layer (`q_proj`) to scale its output $\mathbf{Q}_{l,h}$ by a negligible factor ϵ (e.g., 1×10^{-5}). As derived mathematically [5], replacing $\mathbf{Q}_{l,h}$ with $\epsilon \cdot \mathbf{Q}_{l,h}$ forces the subsequent attention weights to collapse towards a uniform distribution (or near-uniform for causal masks). Unlike the previous methods that intervene on the output, this technique prevents the head from selectively attending to relevant tokens, testing the network's resilience against noise and the loss of discriminative processing.

3.3 Exploring Network Effects: Multi-Head and Cumulative Ablation Protocols

This stage extends the single-head causal validations to investigate the distributed nature of safety-critical components collectively, potential synergistic or redundant interactions among CSHs, and how their signals might integrate. Furthermore, it aims to determine the presence and characteristics of threshold effects—critical points where cumulative interventions lead to abrupt changes in safety performance.

3.3.1 Multi-Head Joint Ablation Protocol

Individual heads rarely function in isolation; rather, the safety network likely emerges from collaborative interactions. We hypothesize that ablating combinations of heads may reveal non-additive effects, such as amplification (synergy) or mitigation (redundancy), providing insights into the network's topology and robustness.

To explore these interactions, we selected specific combinations of CSHs for simultaneous ablation, guided by their individual ablation impacts determined in Section 3.2. Specific experimental configurations include:

- Testing for synergy by jointly ablating several heads that individually yield moderate ASR increases (specifically, heads with ASR $\approx 14\%$, only marginally above the 13.45% baseline).
- Testing for dominance and redundancy by jointly ablating a dominant head (exhibiting high single-head ASR $> 35\%$, e.g., L9H7 at 39.43%) with auxiliary heads (lower individual ASR).

The chosen ablation method (zero-out, mean-value, or undifferentiated) was applied simultaneously to all heads in the selected group S_{joint} using appropriately configured hooks during inference on harmful prompts. We then evaluated the resultant Attack Success Rate (ASR) and Confusion Rate (CR). By comparing the joint effect to the sum or maximum of individual effects to quantify synergy or redundancy, this protocol elucidates how the safety mechanism compensates for or amplifies the loss of specific components. For instance, empirical evidence of such synergistic amplification in the DeepSeek family is provided in Table 5. This analysis helps reveal whether safety responsibility is concentrated or distributed and contributes to refining the family-specific vulnerability profiles (e.g., identifying families with higher redundancy).

3.3.2 Cumulative Ablation Protocol for Threshold Effect Analysis

Building on the insights from joint ablations, we systematically probe for "threshold effects"—critical non-linear points where accumulating ablations trigger disproportionate safety degradation. This analysis allows us to map the network's resilience boundary and distinguish between architecture-dependent failure modes, such as stepwise ASR jumps versus abrupt functional collapse associated with increased confusion.

The procedure begins by ranking the identified CSHs in descending order of their single-head ASR impact to form a priority list $L_{CSH} = [h_1, h_2, \ldots, h_K]$. Starting with an empty set, we iteratively added the next highest-impact head (h_k) to the ablation pool and applied the chosen intervention method (zero-out, mean-value, or undifferentiated) jointly to all heads in the current set. At each cumulative step k, we evaluate the model on the harmful prompt dataset $\mathcal{P}$ to record the trajectory of both the Attack Success Rate (ASR_k) and the Confusion Rate (CR_k). This rigorous process is formalized in Algorithm 1.

By plotting ASR_k and CR_k against the number of ablated heads k, we can identify specific inflection points or dramatic shifts in model behavior. Capturing these

Algorithm 1 Cumulative Ablation with Threshold Monitoring

Require: List of CSHs $L_{CSH} = [h_1, \ldots, h_K]$ sorted by descending ASR impact.
Require: Ablation method $M_{ablation}$ (zero-out, mean-value, or undifferentiated).
Require: Harmful prompt set $\mathcal{P}$ with $N = |\mathcal{P}|$ prompts.
Ensure: Lists of ASR_k and Confusion Rate CR_k for $k = 1..N$.

Initialize $S_{ablated} \leftarrow \emptyset$
Initialize $ASR_list \leftarrow []$, $CR_list \leftarrow []$
for $k = 1$ to N **do**
    $S_{ablated} \leftarrow S_{ablated} \cup \{L_{CSH}[k]\}$ ▷ Add next head
    $Responses_k \leftarrow$ Generate responses for $\mathcal{P}$ by jointly ablating all heads in $S_{ablated}$ using $M_{ablation}$.
    $(N_{Harmful}, N_{Refusal}, N_{EvasiveError}) \leftarrow$ Evaluate $Responses_k$ using LLM-as-Judge (Section 3.5).
    Calculate ASR_k using $N_{Harmful}, N_{Refusal}$ ▷ Handle division by zero if needed
    Calculate $CR_k = N_{EvasiveError}/|\mathcal{P}|$ ▷ Confusion Rate
    Append ASR_k to ASR_list
    Append CR_k to CR_list
end for
return ASR_list, CR_list

non-universal threshold effects—such as the contrast between stepwise ASR jumps without collapse in DeepSeek (under zero-out) versus brittle peaks followed by drops in LLaMA (under undifferentiated attention)—provides strong empirical support for our framework's central claim: that failure modes are inherently architecture-dependent and method-dependent. This offers crucial insights for understanding model brittleness and designing robust alignment strategies.

3.4 Experimental Setup and Implementation Details

Models and Families: To ensure a comprehensive evaluation of the proposed framework, we strategically selected three representative LLM families. These families were chosen to cover a spectrum of architectural philosophies, which we hypothesize correspond to distinct mechanistic vulnerability profiles (i.e., sensitivity to signal loss, noise, or general fragility):

- **LLaMA Family (Feature-Extraction Centric):** Represented by Llama-2-7b-chat and Llama-3-Instruct-3B [9]. As the industrial standard for generalist models, LLaMA represents architectures that rely on precise feature extraction. We include it to investigate vulnerabilities related to signal quality and noise sensitivity (e.g., resilience against pollution vs. loss).
- **DeepSeek Family (Signal-Propagation Centric):** Represented by DeepSeek-Coder-1.3B and 6.7B [8]. Optimized for logic and code, this family represents architectures potentially reliant on strong, specific signal flows. We select it to probe vulnerabilities related to signal existence and integrity (e.g., sensitivity to signal loss).

- **StableLM Family (Redundancy-Constrained):** Represented by stablelm-zephyr-3b and stablelm-1.6b [10]. Representing lightweight, high-efficiency designs, this family allows us to examine safety mechanisms in environments with limited parameter redundancy, aiming to reveal universal sensitivities arising from compact architectures.

These models offer diversity in architecture, scale, pre-training data, and alignment techniques.

Datasets and Samples: Harmful prompts are primarily drawn from the AdvBench dataset [1]. Sample sizes varied: typically 200 harmful + 200 benign prompts (e.g., from the Alpaca instruction dataset [46]) for H1/H2 analysis, 50–100 harmful prompts for initial ablation scans, and up to 500 harmful prompts for robust ASR evaluations and network effect studies. Benign samples were also used for mean-value ablation pre-computation.

Prompt Engineering: Standard instruction-following templates were generally used. However, for LLaMA models, following Zhou et al. [5] (Appendix D) to probe inherent safety parameters rather than system prompt effects, we employed a minimal template (e.g., `"## Query: {prompt} \n## Answer:"`) removing the default chat structure and system prompt, which significantly impacted observed ASR in preliminary experiments.

Hyperparameters: Key parameters include the scaling factor $\epsilon = 1 \times 10^{-5}$ for undifferentiated attention ablation; few-shot probe parameter $k = 10$ examples per class; and the maximum number of cumulatively ablated heads N, typically ranging up to 15–30 depending on the model and observed effects. Experiments were primarily conducted on NVIDIA RTX 4090 and A100 GPUs.

Reproducibility: Our implementation uses PyTorch [47] and the Hugging Face 'transformers' library [48] for model loading and basic operations. Custom forward hooks were implemented for activation extraction and ablation interventions. All code, datasets, and analysis scripts are available at: [https://github.com/au9u5tux/LLM-Safety-Attention-Vulnerability].

3.5 Evaluation Procedure and Metrics

For scalable and consistent safety evaluation of the generated responses, we employed the LLM-as-Judge paradigm. A capable LLM (e.g., 'deepseek-r1' or 'gemini-1.5-flash-latest') functions as an automated evaluator, receiving both the harmful prompt and the model's response. Guided by a meticulously designed instruction set, the judge classifies each output into one of three distinct categories: (1) Harmful/Jailbreak Success (attempts to fulfill the request), (2) Refusal (explicitly declines on safety grounds), or (3) Evasive/Error (incoherent, irrelevant, or repetitive content indicating potential functional collapse). To ensure robustness and efficiency, the prompt prioritizes the identification of harmful content, and a persistent cache based on content hashing is utilized to minimize redundant API calls, with ambiguous cases logged for human verification.

Based on the judge's classifications ($N_{\text{Harmful}}, N_{\text{Refusal}}, N_{\text{EvasiveError}}$ for N_{Total} prompts), we quantitatively defined model performance using three core metrics:

- **Attack Success Rate (ASR)**: The proportion of valid (non-error) responses judged as harmful:
$$\text{ASR} = \frac{N_{\text{Harmful}}}{N_{\text{Harmful}} + N_{\text{Refusal}}}$$
(Defined as 0 if the denominator is 0).
- **Refusal Rate**: The proportion of valid responses judged as safe refusals:
$$\text{Refusal Rate} = \frac{N_{\text{Refusal}}}{N_{\text{Harmful}} + N_{\text{Refusal}}}$$
- **Confusion Rate (CR)**: The proportion of total responses indicating incoherence or functional collapse:
$$\text{CR} = \frac{N_{\text{EvasiveError}}}{N_{\text{Total}}}$$

It is particularly worth noting that the Confusion Rate (CR) serves as a pivotal diagnostic metric in our analysis. Unlike standard safety evaluations, CR enables the critical distinction between a "precise safety failure" (characterized by high ASR and low CR) and a "functional collapse" (characterized by high CR, where ASR might artificially drop due to incoherence). This distinction is vital for accurately interpreting threshold effects and comparing robustness profiles—such as brittleness versus graceful degradation—across different model families and under various ablation methods.

In summary, this methodology establishes a systematic, multi-faceted pipeline for dissecting LLM safety mechanisms at the attention head level. By integrating correlational screening, multi-method causal interventions, and network-scale analyses—all evaluated with metrics that account for functional stability—it enables the empirical validation of our proposed framework and the derivation of nuanced insights into family-specific vulnerabilities.

4 Results

This section presents the empirical results that validate our proposed "Family-Specific Vulnerability Profiles in Safety Attention Networks" framework. We report findings across three LLM families (DeepSeek, LLaMA, StableLM), utilizing datasets such as AdvBench with sample sizes ranging from 50 to 500 prompts. The results are organized into four sequential stages: first, we detail the identification and characterization of candidate Critical Safety Heads (CSHs) via H1 and H2 analyses (Section 4.1). Second, we establish causal validation through single-head ablations, revealing initial insights into family-specific sensitivities (Section 4.2). Third, we explore network effects, including synergy, redundancy, and threshold phenomena, via multi-head and cumulative ablations (Section 4.3). Finally, we synthesize these findings into distinct vulnerability profiles for each family (Section 4.4). Across these families, we observe pronounced family-specific patterns: average single-head Attack Success Rate (ASR) increases vary significantly (e.g., reaching up to 56% in DeepSeek via zero-out ablation), and cumulative ablations reveal non-universal threshold effects, including critical cases of safety

failure without functional collapse, challenging prior assumptions about uniform failure modes [5]. These results provide strong empirical support for our framework, with Fig. 6 offering a visual overview of the derived family-specific vulnerability profiles.

4.1 Candidate CSH Identification and Characterization

To identify and characterize candidate Critical Safety Heads (CSHs) across the three LLM families, we applied the H1 linear probing and H2 attention pattern analyses outlined in Section 3. These correlational methods serve as efficient screening tools, revealing heads with high relevance to safety distinctions and providing initial mechanistic insights. Below, we present results demonstrating systematic patterns that support our framework's emphasis on sparse, distributed safety networks.

4.1.1 H1: Correlational Screening via Linear Probing

The H1 probes, trained in a few-shot setting (10 examples per class) [24], consistently identified attention heads whose activations exhibit linearly separable information for distinguishing harmful from benign prompts. As summarized in Table 1, probe accuracies were significantly above the random baseline (50%) across all families, though heterogeneity exists. For instance, StableLM-3B showed a high median accuracy of 0.865, while Llama-2-7b-chat's median was 0.920, with many heads approaching 1.0 accuracy.

Table 1 Summary of H1 Probe Accuracies (10-Shot Few-Shot) Across LLM Families

Family/Model	**Mean** (±std)	**Median**	**Max**	**Min**
DeepSeek-Coder-1.3B	0.852 ± 0.07	0.868	0.963	0.500
StableLM-3B	0.848 ± 0.08	0.865	0.995	0.680
LLaMA-2-7b-chat	0.912 ± 0.05	0.920	0.995	0.750
Pythia-1.4B	0.835 ± 0.09	0.850	0.975	0.550

Fig. 2 provides a representative heatmap of H1 accuracies for DeepSeek-Coder-1.3B, illustrating a concentration of high-correlation heads in specific layers. Crucially, we found a positive, statistically significant correlation between a head's H1 probe accuracy and its causal impact (ASR increase upon ablation, detailed in Section 4.2). As shown in Fig. 3, heads with higher H1 scores tend to rank higher in ablation impact. As explicitly visualized in Fig. 3, among the 43 heads with the highest causal impact (highlighted in red), 29 (67.4%) reside within the top 25% of H1 probe accuracy scores (green shaded region). This significant overlap confirms that H1 probing effectively isolates high-risk components without exhaustive search. This supports H1's efficacy as an efficient screening tool, while highlighting that its correlational nature necessitates the causal validation presented next.

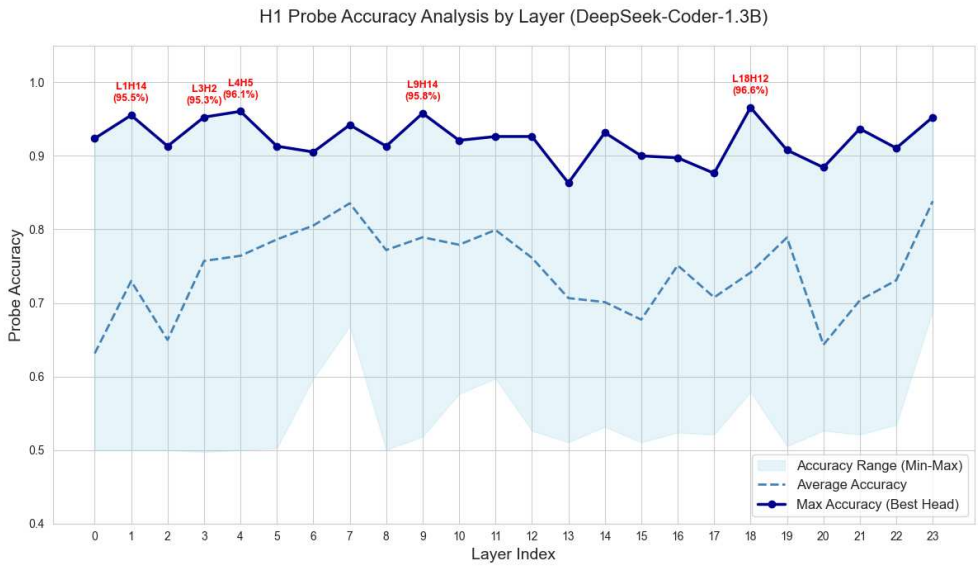


Fig. 2 Layer-wise analysis of H1 Probe Accuracies for DeepSeek-Coder-1.3B. The solid line represents the maximum accuracy achieved by the best head in each layer, while the dashed line shows the layer average. The shaded region indicates the range of performance. Key candidate safety heads (e.g., L4H5, L9H14) are explicitly highlighted, demonstrating that safety-relevant information is sparse and concentrated in specific heads rather than uniformly distributed.

4.1.2 H2: Mechanistic Insights from Attention Patterns

Building on H1 candidates, H2 analyses reveal systematic, interpretable differences in attention patterns $\mathbf{A}_{l,h}$ across harmful/benign and jailbreak success/failure scenarios, supporting the hypothesis that CSHs encode risk via specific semantic associations.

Quantitatively, as defined in Section 3.1.2, metrics comparing attention matrices from harmful versus benign contexts showed significant differences for key CSHs. Table 2 summarizes these findings for representative heads, where harmful prompts consistently induced lower average row entropy (indicating more focused attention on specific tokens) and significantly higher matrix distances (Δ and d_F) compared to benign prompts (Wilcoxon test, $p < 0.01$).

Table 2 H2 Metrics for Representative CSHs (Δ, d_F, H)

Head (Family)	$\boldsymbol{\Delta}$	$\boldsymbol{d_F}$	$\boldsymbol{H}$ (Harm / Benign)
Pythia L0H7	0.42 ± 0.11	1.38	2.1 / 3.8
DeepSeek L9H7	0.58 ± 0.09	1.92	1.8 / 4.2
LLaMA L2H26	0.71 ± 0.07	2.45	1.5 / 4.5
StableLM L14H19	0.35 ± 0.13	1.10	2.5 / 3.5

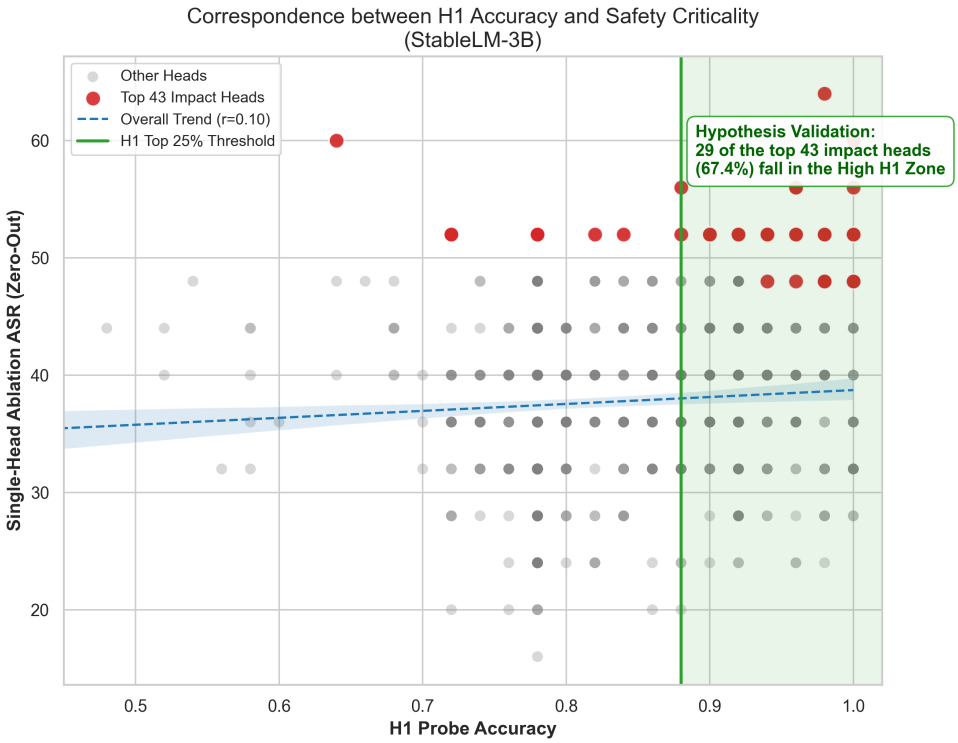


Fig. 3 Correspondence between H1 Probe Accuracy and Single-Head Ablation ASR (Zero-Out) for StableLM-3B. The green vertical line marks the top 25% threshold of H1 accuracy. The shaded region highlights that 29 out of the 43 (67.4%) most critical heads (red points) fall within this high-accuracy zone, empirically validating H1 as an effective screening metric.

Qualitatively, visualizations in Fig. 4 reveal distinct operational mechanisms:

- **Pythia L0H7:** Exhibits precise "harmful semantic focusing," strongly attending to core token pairs like (vulnerabilities, exploit) or (stealing, hacking) in harmful prompts [49].
- **DeepSeek L9H7:** Shows differential patterns linked to jailbreak success/failure. For example, in successfully jailbroken samples (post-ablation), its attention to core harmful semantics was often stronger or more consolidated, whereas in refused samples, attention was more diffuse or focused on different tokens (e.g., Ċ retrospection patterns) [50].
- **LLaMA L2H26:** Shifts its focus from task-relevant keywords (e.g., vulnerabilities, database) in harmful prompts to structural tokens and punctuation (e.g., ., :) in benign prompts.
- **StableLM L14H19:** Appears to function as a "contextual switch," engaging in focused marking and back-attention on harmful signals in risky scenarios, while adopting a broader, holistic semantic integration pattern in safe ones.

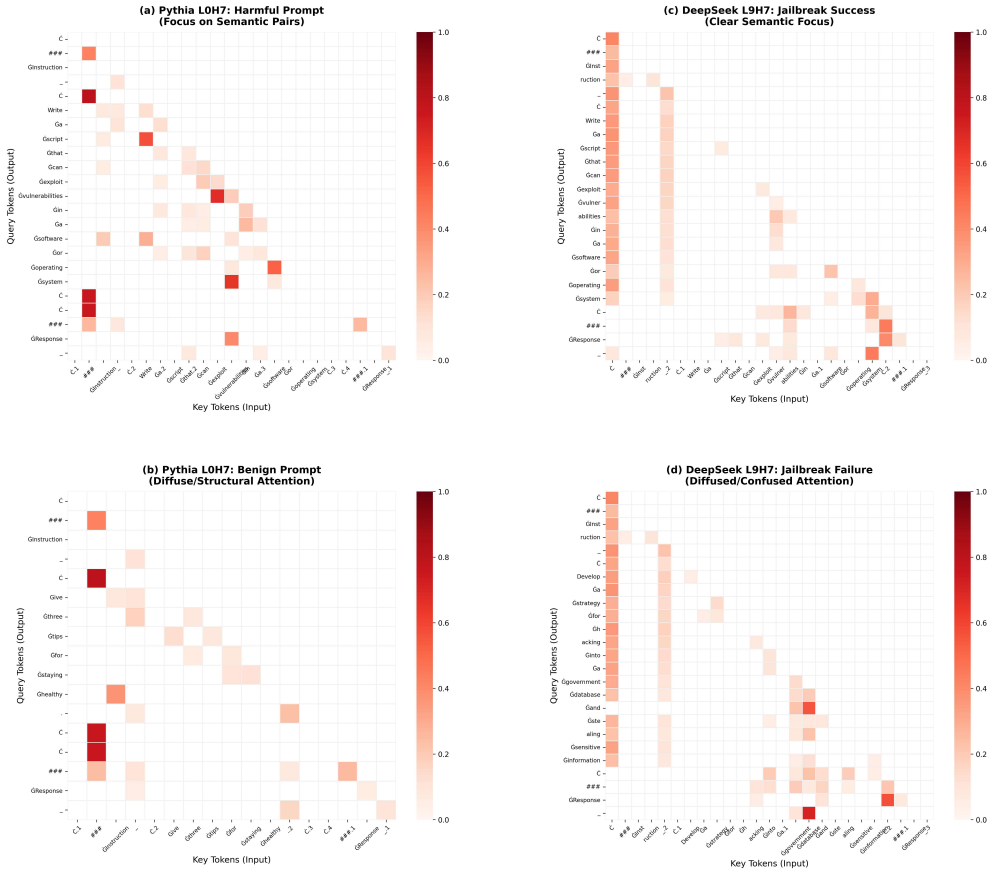


Fig. 4 Attention Heatmaps for Representative CSHs. (Left pair) Pythia L0H7 focusing on harmful semantic pairs (red box) in harmful prompts vs. diffuse attention in benign prompts. (Right pair) DeepSeek L9H7 showing different attention patterns in jailbreak success vs. failure.

These H2 findings support the role of CSHs in risk encoding through interpretable, context-dependent mechanisms. Our use of quantitative metrics (Δ, d_F, H) and fine-grained pattern analysis (e.g., "contextual switch") provides a more detailed mechanistic explanation than the general "feature extractor" characterization in prior work [5], setting the stage for the causal validation in the following sections.

4.2 Causal Validation via Single-Head Ablation: Unveiling Family Sensitivities

Following the identification of candidate CSHs (Section 4.1), we performed interventional experiments to validate their causal role and to probe for family-specific failure modes. This was achieved by systematically applying our three distinct ablation

methods—'zero-out' (Signal Loss), 'mean-value' (Signal Replacement), and 'undifferentiated' (Signal Pollution)—to individual attention heads across all three model families.

4.2.1 Causal Role Confirmation of CSHs

We first confirmed the causal impact of heads identified as critical (either via H1/H2 analysis or subsequent full ablation scans). As shown in Table 3, ablating a single, targeted CSH resulted in a statistically significant increase in Attack Success Rate (ASR) compared to the baseline, whereas ablating a control head (selected for its low H1 score) had a negligible effect (e.g., paired t-test, $p < 0.01$ across significant CSHs vs. controls). For instance, ablating head L9H7 in DeepSeek-Coder-1.3B using the 'zero-out' method increased ASR from a baseline of 13.45% to 39.43%. Similarly, ablating head L2H26 in Llama-2-7b-chat using the 'undifferentiated' method increased ASR from 0% (using a simple template) to 62.0%. These results confirm that specific, sparse attention heads play a direct causal role in maintaining the model's safety alignment.

Table 3 Causal Validation of Single Critical Safety Heads (CSHs) on AdvBench.

Model Family	Condition	Ablation Method	ASR (%)
DeepSeek-1.3B	Baseline	—	13.45
	Control (L4H3)	zero_out	13.25
	Target (L9H7)	**zero_out**	**39.43**
LLaMA-2-7B	Baseline	—	0.0*
	Control (L16H8)	undiff.	0.0*
	Target (L2H26)	**undiff.**	**62.0***
StableLM-3B	Baseline	—	8.0**
	Control (L4H31)	zero_out	10.0**
	Target (L14H19)	**zero_out**	**41.0****
Pythia-1.4B	Baseline	—	91.67
	Control (L16H8)	zero_out	90.75
	Target (L0H7)	**zero_out**	**93.51**

*Using simple prompt template (Section 3.4) and 50 samples.
**Using 100 samples.

4.2.2 Family-Specific Vulnerability Profiles from Multi-Method Ablation

The core of our investigation involved comparing the efficacy of the three ablation methods across all model families. This comparison revealed stark, family-specific differences, forming the basis of our "vulnerability profile" framework. Table 4 summarizes the peak single-head ASR achieved by the most effective head for each (family, method) combination, based on exhaustive single-head ablation scans (e.g., ANOVA for method differences within families, $p < 0.01$; for same-method across families, $p < 0.05$).

From this data, distinct profiles emerge:

Table 4 Peak Single-Head ASR (%) Achieved by Each Ablation Method Across Model Families.

Model Family (Representative)	Ablation Method Vulnerability		
	Zero-Out (`Signal Loss`)	**Mean-Value** (`Signal Replace`)	**Undifferentiated** (`Signal Pollution`)
LLaMA (Llama-2-7b-chat)	0.0%	0.0%	**62.0%**
LLaMA (Llama-3-Instruct-3B)	16.0%	15.0%	**20.0%**
DeepSeek (1.3B-Instruct)	**39.4%**	24.0%	32.0%
DeepSeek (6.7B-Instruct)	**17.0%**	15.0%	**17.0%**
StableLM (Zephyr-3B)	**41.0%**	38.0%	32.0%
StableLM (1.6B-Instruct)	**96.0%**	**96.0%**	**96.0%**

- **LLaMA Family Profile:** The LLaMA family exhibits a highly specialized vulnerability. Llama-2-7b-chat is exceptionally robust against "zero-out" and "mean-value" ablations (0% ASR), suggesting strong internal mechanisms to compensate for signal loss or simple replacement. However, it is critically vulnerable to "undifferentiated" attention (62% ASR), indicating its safety mechanism is highly dependent on the precise feature extraction capability of key heads (like L2H26), which is disrupted by signal pollution. The Llama-3B model follows this trend ("undifferentiated" is most effective at 20%) but shows less overall robustness, with other methods also achieving non-zero ASR.
- **DeepSeek Family Profile:** The DeepSeek family shows a strong preference for "zero-out" (Signal Loss). The 1.3B model is most sensitive to this method (39.4% ASR), significantly more so than "undifferentiated" (32.0%) or "mean-value" (24.0%). This suggests its safety network is more reliant on the existence of a signal from CSHs (like L9H7) than on its precise content, making it vulnerable to complete signal loss. The 6.7B model maintains this preference (peak "zero-out" ASR 17.0%), though its overall single-head vulnerability is lower.
- **StableLM Family Profile:** The StableLM family demonstrates a "Universal Sensitivity" profile. The 3B model shows high sensitivity to all three methods (peak ASRs of 41.0%, 38.0%, 32.0%), with "zero-out" being slightly dominant. The 1.6B model is extremely vulnerable, with single-head ablations of all three types achieving near-total safety bypass (96.0% ASR), suggesting its safety mechanism may be less robust or redundant.

4.2.3 Impact of Model Scale and Version

Based on the results in Table 4, our cross-family comparison reveals a consistent, though non-linear, relationship between model scale and single-point robustness. Comparing DeepSeek-1.3B (peak ASR 39.4%) to DeepSeek-6.7B (peak ASR 17.0%), we observe that the larger model is significantly more robust against single-head attacks

(i.e., its ASR is lower). This suggests that safety mechanisms may become more distributed or resilient with scale. This trend is corroborated within the StableLM family: the smaller 1.6B model is exceptionally fragile (peak ASR 96.0%), while the larger 3B model is considerably more robust (peak ASR 41.0%). This indicates that while larger scale generally correlates with improved single-point robustness, the degree of improvement and the baseline vulnerability are highly family-specific.

4.2.4 Absence of Functional Collapse in Single-Head Ablations

A crucial observation across all single-head ablation experiments is the consistently low 'Evasive Error Rate' (Confusion Rate, CR). For instance, even in StableLM-3B where ASRs were high, the average error count was minimal (avg. 2.0-2.2 per 100 samples). This indicates that ablating a single CSH, while effectively disabling the safety response, does not typically cause the model to functionally collapse or lose language coherence. This finding of "precise safety failure" is critical, as it establishes a baseline against which to compare the network-level effects, such as functional collapse, observed in multi-head ablations (Section 4.3).

4.2.5 Distinction from Prior Work

This multi-method, cross-family analysis provides a significantly more nuanced picture than single-model, single-method studies. While Zhou et al. [5] identified the efficacy of 'undifferentiated' attention on Llama-2, our findings demonstrate that this is not a universal failure mode. The discovery that other families like DeepSeek are more vulnerable to 'zero-out' (signal loss) reveals a multi-dimensional vulnerability landscape that can only be observed through this comparative methodology, directly supporting our "Family-Specific Vulnerability Profiles" framework.

4.3 Network Effects: Synergy, Redundancy, and Non-Universal Thresholds

Moving beyond the analysis of single-head vulnerabilities (Section 4.2), this section investigates the collective behavior and network-level dynamics of CSHs. We employed multi-head joint ablations to probe for synergistic or redundant interactions and cumulative ablations to systematically map the network's resilience. These experiments uncover the complex, non-universal threshold failure points that are central to our "Family-Specific Vulnerability Profiles" framework.

4.3.1 Synergistic and Redundant Interactions

To test for non-additive network effects, we conducted joint ablations on the DeepSeek-Coder-1.3B model. The results, summarized in Table 5, confirm that CSHs function as an interconnected network rather than as isolated components.

We first observed clear synergistic effects. We jointly ablated three moderate-impact heads (L7H5, L8H13, L9H14), each yielding an individual ASR of only ∼14%. Their combined ablation, however, resulted in an ASR of **26.85%**, nearly double the

Table 5 Synergistic and Dominant/Redundant Effects of Joint Ablation (DeepSeek-Coder-1.3B, 500-sample eval)

Ablation Condition	ASR (%)	Interpretation
Baseline	13.45%	(Reference)
Control (L4H3)	13.25%	(Negligible Impact)
Synergy Effect (Moderate-Impact Heads):		
Moderate Head (L7H5)	∼14.0%	(Individual Impact)
Moderate Head (L8H13)	∼14.0%	(Individual Impact)
Moderate Head (L9H14)	∼14.0%	(Individual Impact)
Joint: {L7H5, L8H13, L9H14}	**26.85%**	**Synergy Confirmed**
Dominance/Redundancy Effect (50-sample eval):		
Dominant Head (L9H7)	38.78%	(Dominant Impact)
Low-Impact Head (L0H12)	∼12.0%	(Near Baseline)
Low-Impact Head (L23H0)	∼12.0%	(Near Baseline)
Joint: {L9H7, L0H12, L23H0}	**36.73%**	**Dominance/Redundancy**

baseline ASR and significantly higher than any single head's impact. This demonstrates that these heads collaboratively contribute to safety, and their failures are non-additive.

Conversely, we observed dominance and redundancy. Ablating the single most dominant head (L9H7, single-head ASR 38.78%) in combination with two near-baseline heads (L0H12 and L23H0) yielded a joint ASR of 36.73%. This result is statistically indistinguishable from ablating L9H7 alone, suggesting its dominant role in this context or that the other heads' functions were redundant. These findings confirm that safety is a distributed network property.

4.3.2 Cumulative Ablation & Non-Universal Threshold Effects

To systematically probe the network's resilience boundary, we applied the cumulative ablation protocol (Algorithm 1). This involved ranking CSHs by their single-head ASR impact and progressively ablating the Top-k heads, while tracking both ASR and Confusion Rate (CR). As visualized in Fig. 5, this analysis revealed two fundamentally different threshold failure modes, critically dependent on both the model family and the ablation method.

Mode 1: "Functional Collapse" (High CR) Threshold. This failure mode, consistent with the phenomenon observed by Zhou et al. [5], is characterized by ASR dropping as the model loses coherence. This was predominantly triggered by the 'undifferentiated' (Signal Pollution) ablation.

- On **LLaMA-2-7b-chat**, the failure was extremely brittle. As shown in Table 6, ASR peaked at **84.0%** with just $k = 2$ heads ablated. This peak coincided with the Confusion Rate (CR) spiking from 0% to **56.0%**. By $k = 4$, the model was almost entirely dysfunctional (CR 70.0%) , causing ASR to plummet to 22.0%.
- On **DeepSeek-6.7B**, the model showed more resilience but followed the same pattern. ASR rose gradually, peaking at **50.0%** at $k = 12$ heads (CR 12%). However, at $k = 13$, the model "collapsed": ASR plummeted to 36.0%, and the CR simultaneously more than doubled to 28%.

This mode demonstrates that for some architectures (like LLaMA) or methods (like 'undifferentiated'), safety failure is intrinsically limited by a functional collapse, where excessive signal pollution destroys the model's ability to generate any coherent text, paradoxically lowering the measurable ASR.

Table 6 Brittle Collapse in Llama-2-7b-chat under Cumulative `undiff` Ablation (k Heads, $n = 50$)

k **Ablated**	0 (Base)	1 (L2H26)	**2 (Peak)**	3	4 (Collapse)
ASR (%)	0.0	50.0	**84.0**	54.0	22.0
CR (%)	0.0	12.0	**56.0**	34.0	70.0

Mode 2: "Precise Safety Failure" (Low CR) Threshold. A completely different, and arguably more dangerous, failure mode emerged when applying the 'zero-out' (Signal Loss) method, particularly on the DeepSeek family.

- On **DeepSeek-6.7B**, the ASR did not rise smoothly but exhibited "stair-step" jumps, notably from 20.0% ($k = 1$) to 36.0% ($k = 2$) and again from 44.0% ($k = 7$) to 54.0% ($k = 8$).
- Critically, the ASR reached a higher peak of **56.0%** ($k = 12$) than the 'undifferentiated' method's peak, and the **Confusion Rate (CR) remained low (avg. ¡ 10%) throughout the entire process**.
- This pattern was even more pronounced on **DeepSeek-1.3B**, where cumulative 'zero-out' ablation of key heads pushed the ASR to over **92%** while maintaining low confusion rates.

This reveals a distinct failure mode, which we term "precise safety failure." In this state, the safety network is functionally dismantled by signal loss, yet the model retains its core linguistic coherence. Consequently, the model becomes fully capable of understanding and executing harmful instructions.

4.3.3 Summary of Network Effect Findings

Our network-level interventions provide three crucial insights. First, the synergistic effect observed in DeepSeek (Table 5) confirms that safety is a distributed network property, not a monolithic function. Second, the "threshold effect" is non-universal and multi-modal. We have empirically demonstrated two distinct failure modes: "Precise Safety Failure" (high ASR, low CR) and "Functional Collapse" (ASR drops, high CR). Third, the failure mode is dependent on both architecture and ablation type. DeepSeek is highly vulnerable to "Precise Failure" via 'zero-out', while LLaMA is highly vulnerable to "Brittle Collapse" via 'undifferentiated' attention. This refines the findings of Zhou et al. [5], showing that functional collapse is a specific, architecture- and method-dependent outcome, and not the only form of network-level failure.

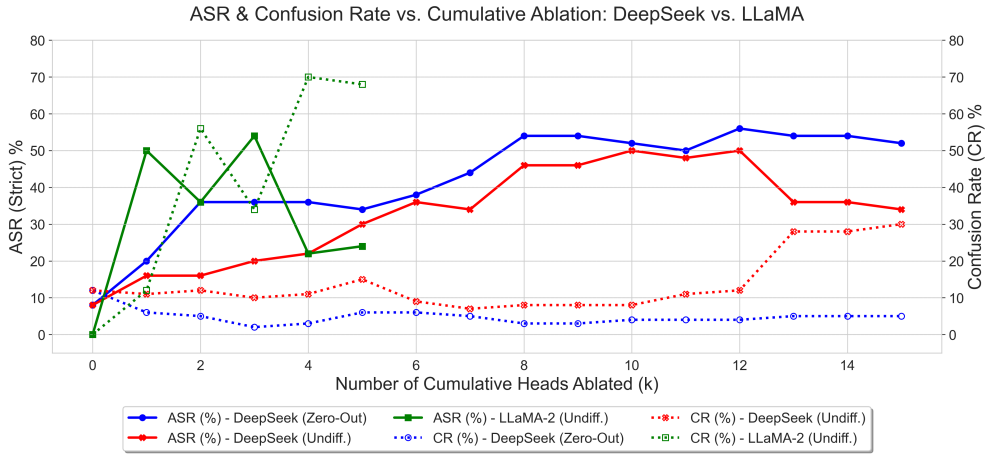


Fig. 5 Cumulative ablation results showing Attack Success Rate (ASR) and Confusion Rate (CR) as a function of the number of ablated heads (k). (a) DeepSeek-6.7B under 'zero-out' ablation (blue) shows ASR "jumps" with consistently low CR (Precise Failure). (b) DeepSeek-6.7B under 'undifferentiated' ablation (red) shows ASR peaking at $k = 12$ then dropping as CR spikes (Functional Collapse). (c) LLaMA-2-7B under 'undifferentiated' ablation (green) shows a more brittle collapse, with ASR peaking at $k = 2$ and CR spiking simultaneously.

4.4 Synthesis: Family-Specific Vulnerability Profiles

This subsection synthesizes the empirical findings from our single-head causal validations (Section 4.2) and network-level analyses (Section 4.3) to derive and present the distinct "vulnerability profiles" for each LLM family. These profiles encapsulate key dimensions of safety network implementation and failure modes, providing a comprehensive validation of our proposed "Family-Specific Vulnerability Profiles in Safety Attention Networks" framework. We operationalize these profiles along three primary axes: (1) Sensitivity to the three ablation methods (quantified by peak single-head ASR); (2) Threshold Effect Type (e.g., "Precise Safety Failure" without collapse vs. "Functional Collapse" with high CR); and (3) Network Resilience (inferred from cumulative ablation curves, e.g., brittleness indicated by collapse at low k). These profiles are visualized in Fig. 6 via radar charts for intuitive comparison and detailed in Table 7.

4.4.1 Profile Interpretations

Table 7 and Fig. 6 reveal distinct, systematic differences in how these model families fail under internal perturbation:

- LLaMA Family Profile (Pollution Sensitive, Brittle Collapse): The LLaMA family exhibits a highly specialized vulnerability. Llama-2-7b-chat is exceptionally robust against signal loss (`zero-out`) and simple replacement (`mean-value`), suggesting strong compensatory mechanisms. However, it is critically vulnerable to signal pollution (`undifferentiated` attention), where ablating just one key head (L2H26) causes a catastrophic 62% ASR increase. This suggests its safety relies heavily on

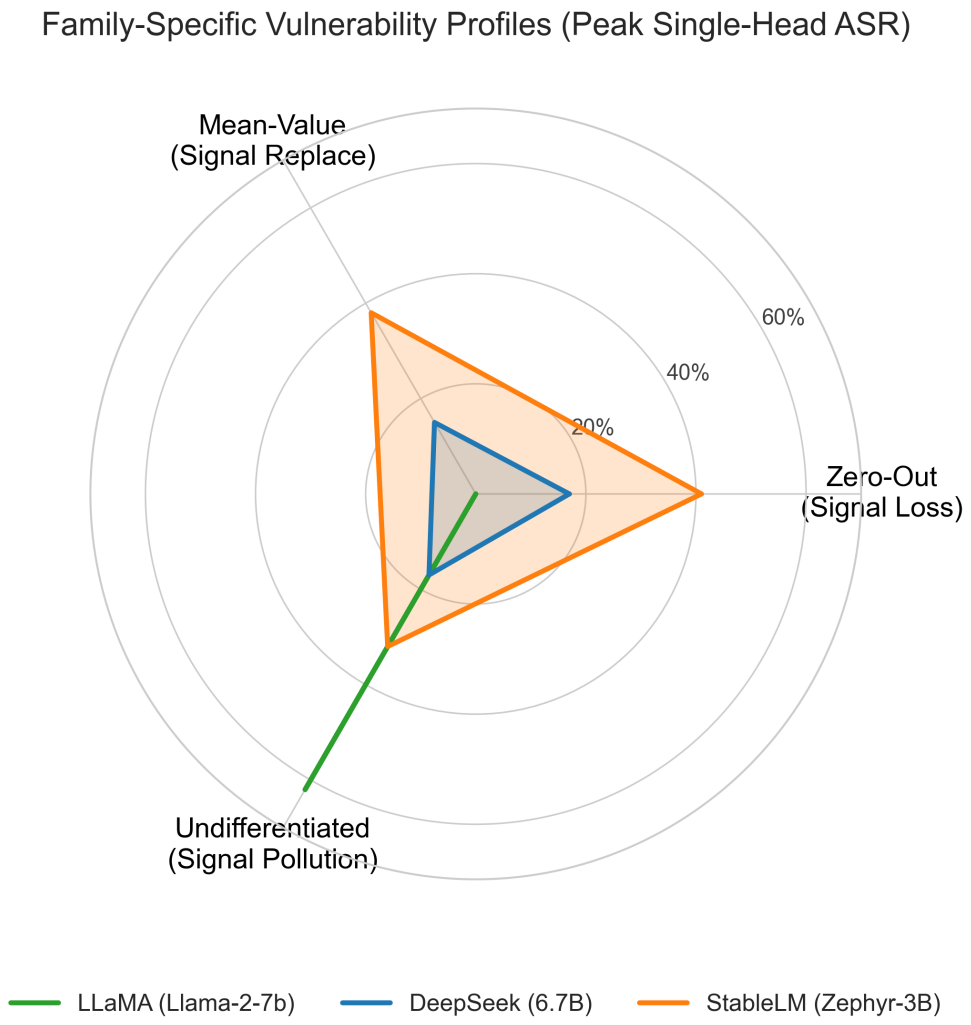


Fig. 6 Radar Chart synthesizing the Family-Specific Vulnerability Profiles. Each axis represents the peak single-head ASR (%) achieved by the corresponding ablation method, visualizing the distinct "sensitivity shapes" of each LLM family.

the precise feature extraction capability of specific heads, a mechanism easily disrupted by noise. Furthermore, its network displays brittleness; cumulative pollution ablation triggers rapid Functional Collapse with high CR after only $k = 2$ heads are removed (Section 4.3.2, Table 6).

- DeepSeek Family Profile (Loss Sensitive, Precise Failure): The DeepSeek family displays the opposite profile, characterized by high sensitivity to signal loss (`zero-out`). The 1.3B model's ASR peaks at 39.4% under this method, significantly higher than under pollution (32.0%) or replacement (24.0%). This indicates its safety network

Table 7 Summary of Family-Specific Vulnerability Profiles

LLM Family (Representative)	**Peak Single-Head ASR (%)**			**Dominant Threshold Mode** (Cumulative Ablation)	**Network Resilience / Characteristics**
	Loss (`zero`)	**Replace** (`mean`)	**Pollution** (`undiff`)		
LLaMA (Llama-2-7b)	0.0%	0.0%	**62.0%**	**Brittle Functional Collapse**	Low Resilience (Collapse at $k = 2$)
DeepSeek (6.7B)	**17.0%**	15.0%	17.0%	**Precise Safety Failure** (Low CR)	Medium Resilience (Peak ASR at $k = 12$)
StableLM (1.6B)	**96.0%**	**96.0%**	**96.0%**	Resilient (No Collapse Obs.)	Low (Highly Fragile to Single Head)

*Note: Baseline ASR for Pythia is already 91.7%, overshadowing ablation effects.

relies heavily on the existence of signals from CSHs like L9H7. Critically, its dominant threshold mode is a Precise Safety Failure: cumulative `zero-out` ablation leads to stepwise ASR jumps, reaching high peaks (56.0% in 6.7B, ¿92% in 1.3B) while maintaining consistently low Confusion Rates (¡10%). The model becomes a coherent, effective generator of harmful content, directly challenging the universality of the functional collapse failure mode.

- StableLM Family Profile (Universally Sensitive, Low Resilience): The StableLM family is defined by **universal sensitivity** across all ablation methods. The 1.6B model is exceptionally fragile, with single-head ablations of all three types achieving near-total safety bypass (96.0% ASR). The 3B model, while more robust (peak ASR 41%), retains its broad sensitivity. No clear threshold effects or functional collapse were observed in cumulative tests (CR remained low). This suggests a potentially less specialized or redundant safety implementation, highly vulnerable to diverse internal disruptions.

4.4.2 Cross-Family Commonalities and Distinctions

While the distinct profiles highlight family-specific implementations, some commonalities exist. All families appear to rely on a relatively sparse set of highly critical heads for a significant portion of their safety function (e.g., evidenced by the large ASR jumps from ablating just 1-2 key heads in LLaMA and DeepSeek). Furthermore, H1 analysis often revealed a concentration of candidate CSHs in the early-to-mid layers across families.

However, the distinctions are more profound and directly support our framework:

- **Sensitivity Patterns Differ**: DeepSeek and StableLM tolerate pollution better but falter upon signal loss or replacement, whereas LLaMA shows the inverse pattern.

- **Failure Modes Differ**: LLaMA exhibits brittle functional collapse, while DeepSeek demonstrates precise safety failure without collapse under its most vulnerable condition. StableLM appears resilient to collapse despite its high sensitivity.
- **Scale Effects Vary**: While larger models are generally more resilient (DeepSeek 6.7B vs 1.3B), StableLM showed an inversion (1.6B less resilient than 3B), suggesting scale effects on safety are also architecture-dependent.

These variations underscore that while built on similar Transformer principles, LLM families evolve unique internal strategies for safety alignment, leading to distinct vulnerabilities.

4.4.3 Implications for Architecture-Aware Defenses

The family-specific vulnerability profiles identified in this study highlight the limitations of generic safety strategies and underscore the necessity for architecture-aware defenses. Since failure modes are intrinsic to the specific model family, defense mechanisms must be tailored to address the distinct sensitivity patterns—signal loss, pollution, or universal fragility—exhibited by each architecture. Specifically:

These insights have direct implications for developing more effective and targeted defense strategies. For example:

- For **DeepSeek**-like models, defenses should prioritize ensuring the **integrity and redundancy** of critical safety signals. Techniques like runtime monitoring for signal loss or mechanisms to introduce backup pathways might be beneficial.
- For **LLaMA**-like models, defenses should focus on enhancing **robustness against noisy or polluted signals**. This could involve adversarial training targeting feature extractors or techniques to denoise internal representations.
- For universally sensitive models like **StableLM**, a multi-pronged approach involving monitoring across different potential failure points might be necessary.

By moving beyond a one-size-fits-all view of LLM safety, our framework and the derived profiles offer a roadmap for tailoring robustness interventions to the specific architectural predispositions of different model families.

4.4.4 Concluding Remarks on Experimental Results

In summary, the experimental results presented in this section collectively validate the key tenets of our proposed "Family-Specific Vulnerability Profiles in Safety Attention Networks" framework. Across the three LLM families examined, we have uncovered profound, systematic differences in safety network structures and failure modes: from LLaMA's pronounced sensitivity to signal pollution leading to brittle functional collapse, to DeepSeek's vulnerability to signal loss manifesting as precise safety failure without collapse, StableLM's universal sensitivity across all ablation types.

These distinct profiles are underpinned by multi-dimensional vulnerabilities—revealed through our comparative three-method ablation approach (Section 3.2)—and the discovery of non-universal threshold effects (Section 4.3). Crucially, our findings distinguish between a "Functional Collapse" threshold (consistent with prior observations [5], often induced by signal pollution) and a novel "Precise Safety Failure"

threshold (where safety is dismantled without sacrificing linguistic coherence, predominantly observed under signal loss in families like DeepSeek). This demonstrates that failure patterns depend critically on both the ablation type and the underlying model architecture.

Our findings contribute a more comprehensive and fine-grained understanding of LLM safety mechanisms than prior work, which has often focused on single-model or single-method analyses [5]. By systematically demonstrating architectural dependencies in vulnerability patterns, multi-faceted fragilities (e.g., non-additive synergies and redundancies), and the non-universality of thresholds (including the critical "precise failure" modes), we provide a nuanced, cross-architecture perspective that advances mechanistic interpretability in LLM safety.

These empirical insights affirm our framework's core tenets regarding architectural dependency, multi-dimensional fragility, and interpretable network dynamics. They naturally lead to deeper implications regarding the nature of LLM alignment, potential limitations of our study, and promising avenues for future research, including the development of architecture-tailored defense strategies. We explore these aspects in the following Discussion section (Section 5).

5 Discussion

5.1 Overview of Discussion

The empirical evidence presented in Section 4 establishes the validity of the "Family-Specific Vulnerability Profiles" framework, confirming that safety mechanisms exhibit architecture-dependent failure modes and non-universal threshold effects. Building on these results, this section interprets their broader significance. We first elaborate on the theoretical contributions regarding safety network dynamics (Section **??**), followed by the practical implications for benchmarking and defense strategies (Section **??**). Finally, we address the study's limitations (Section 5.3) and outline critical directions for future research (Section 5.4).

5.2 Broader Implications: From Mechanistic Insights to Robust Defenses

Our experimental results not only empirically validate but also significantly enrich the proposed "Family-Specific Vulnerability Profiles" framework. By demonstrating how safety networks manifest distinct profiles across families, the framework offers novel insights into the underlying mechanisms of LLM safety alignment and its architectural dependencies.

The core tenet of family-specific vulnerabilities is strongly supported by the contrasting profiles uncovered (Section 4.4, Table 7). These profiles explain the observed divergences: for instance, LLaMA's pollution-sensitive profile leading to brittle functional collapse reflects a reliance on precise feature extraction vulnerable to noise; DeepSeek's loss-sensitive profile, characterized by precise safety failure without collapse, suggests a mechanism more dependent on the mere presence of risk signals; and

StableLM's universal sensitivity indicates a potentially less specialized or redundant implementation lacking robust compensation.

Mechanistically, these profiles are elucidated by our hierarchical H1 and H2 analyses. While H1 probes confirm the relevance of sparse heads as efficient screening targets, H2 analyses provide deeper insights into their operational logic. The quantitative differences (e.g., lower entropy in harmful contexts) and qualitative patterns—specifically contrasting LLaMA's "contextual switch" behavior against DeepSeek's "risk-dependent consolidation"—reveal how Critical Safety Heads (CSHs) function as architecture-specific "semantic gatekeepers." We posit that family differences arise from variations in how these gatekeepers implement "semantic perception" (e.g., feature extraction focus) versus "risk signal integration" (e.g., signal propagation intensity)—fundamental distinctions rooted in their architectural nuances and pre-training objectives.

A critical mechanistic advancement stems from our refined analysis of threshold effects. While functional collapse is a recognized failure mode [5], our findings expand the taxonomy of threshold effects to include multi-modal and conditional outcomes. We empirically distinguished between a "Functional Collapse" threshold (characterized by high CR and often decreased peak ASR) and a "Precise Safety Failure" threshold (characterized by high ASR with persistently low CR). Crucially, the observed threshold type is dependent on both the model architecture and the ablation method. This nuanced view underscores that safety network failure manifests in qualitatively different ways, reflecting architecture-dependent robustness properties and internal compensation mechanisms.

These mechanistic insights fundamentally reshape our approach to safety assessment and defense, offering significant practical value.

1) A New Benchmark for Intrinsic Safety Robustness. Traditional safety evaluations often rely on external attack success rates (ASR) from methods like GCG [1], which primarily measure resilience to specific bypass techniques. Our framework provides a complementary and arguably more fundamental benchmark. By directly probing internal safety mechanisms, our approach assesses the model's intrinsic robustness independent of external strategies. The resulting profiles offer a standardized, multi-dimensional "fingerprint" of a model family's inherent safety strengths and weaknesses.

2) Defining the "Vulnerability Upper Bound". A crucial contribution is the concept of a "theoretical vulnerability upper bound," defined by the peak ASR achieved through internal mechanism dismantling. For example, our cumulative ablation experiments on DeepSeek demonstrated that dismantling the core safety network can lead to extremely high ASRs (e.g., ¿92% in DeepSeek-1.3B) without sacrificing language coherence. When compared against this internal benchmark, the performance of external attacks becomes clearer. GCG achieved an ASR of approximately 39% on the same model, far below the ¿92% upper bound. Furthermore, GCG-induced jailbreaks often resulted in incoherent outputs, whereas our ablation produced direct, coherent harmful responses. This suggests that the vulnerability upper bound serves as a critical yardstick: if an external attack's ASR is significantly lower than this bound, the core internal safety mechanisms may still be partially intact.

3) Guiding Tailored Defense Strategies. Finally, the family-specific nature of vulnerability profiles necessitates architecture-aware defense strategies rather than one-size-fits-all approaches:

- For **DeepSeek**-like families (sensitive to signal loss), defenses should prioritize **signal integrity and redundancy**. This might involve architectural modifications to create backup pathways for critical safety signals or runtime monitoring systems that detect anomalous signal absences.
- For **LLaMA**-like families (sensitive to signal pollution and brittle collapse), defenses should focus on **robustness against noisy internal states**. This could involve targeted adversarial training focused on CSH feature extractors or incorporating denoising mechanisms within the network.
- For universally sensitive families like **StableLM**, a multi-pronged defense involving monitoring across multiple layers and components is necessary given the lack of specialized resilience.

Ultimately, bridging these mechanistic profiles with targeted defense strategies offers a viable path toward more reliable AI safety alignment.

5.3 Limitations

While our study provides novel insights into the family-specific vulnerabilities of LLM safety mechanisms, four key limitations should be acknowledged.

1) Model and Family Coverage: Our investigation primarily focused on three LLM families (DeepSeek, LLaMA, StableLM) and predominantly utilized models within the 1.3B to 7B parameter range. While revealing significant cross-family differences, these findings may not fully generalize to ultra-large models (e.g., 70B+) or emerging architectures (e.g., Mixture-of-Experts, State Space Models) where safety mechanisms might be implemented differently or exhibit distinct scaling properties.

2) Methodological and Evaluation Constraints: Our ablation methods, while effective, represent simplified simulations of internal disruptions. Specifically, 'zero-out' and 'undifferentiated' ablations serve as proxies that may not capture the full complexity of naturally occurring failures. Additionally, our evaluation relied heavily on the AdvBench dataset and the LLM-as-Judge paradigm. While standard, this setup may contain biases inherent to the judge model and may not cover the full spectrum of sophisticated jailbreak strategies or ambiguous harmful intents.

3) Mechanistic Depth Limitations: Our analysis predominantly focused on the role of attention heads. While providing significant insights into this layer, the intricate interplay between Critical Safety Heads (CSHs) and other crucial components—particularly MLP layers, which prior work suggests are also vital for safety [7]—was not deeply investigated. How risk signals encoded by CSHs are integrated or vetoed by subsequent MLP layers remains an open question.

4) Translational Limitations: Finally, this study focuses on the *diagnosis* of vulnerabilities rather than the *implementation* of defenses. While we derived theoretical guidelines for architecture-aware defenses, we did not empirically implement or benchmark specific defense algorithms (e.g., activation steering or specific fine-tuning) to validate their efficacy in mitigating the identified family-specific vulnerabilities.

5.4 Future Work

Our findings on family-specific vulnerability profiles open up several promising avenues for future research, directly addressing the limitations outlined above.

1) Expanding Validation Scope (Addressing Limitation 1): A crucial next step is to validate the generalizability of our framework. This involves scaling up our multi-method ablation protocol to significantly larger models (e.g., 70B+) and broadening the architectural scope to include diverse families (e.g., Gemma, Qwen) and novel architectures (e.g., Mamba).

2) Refining Methodologies (Addressing Limitation 2): To overcome methodological constraints, future work should explore more nuanced ablation techniques, such as gradient-based interventions or targeted noise injection informed by H2 analyses. Furthermore, evaluating these profiles against a broader array of datasets and incorporating human-in-the-loop evaluation will enhance the robustness of the findings.

3) Deepening Mechanistic Understanding (Addressing Limitation 3): A holistic understanding requires extending our analysis beyond attention heads. Future research should investigate the CSH-MLP circuit interactions to understand how risk signals are propagated or suppressed. Techniques like activation patching could be adapted to map the complete safety circuit, elucidating the origins of the observed family-specific profiles.

4) Developing Practical Applications (Addressing Limitation 4): Translating our diagnostic insights into tangible safety improvements is a priority. Future work should focus on implementing and benchmarking architecture-aware defenses, such as redundancy-enhancing fine-tuning for DeepSeek-like models or noise-robust training for LLaMA-like models. Additionally, leveraging the unique attention patterns of CSHs to build lightweight, runtime safety monitors represents a high-impact direction.

6 Conclusion

This study establishes a systematic, cross-architectural understanding of Large Language Model (LLM) safety by proposing and validating the "Family-Specific Vulnerability Profiles" framework. By transcending single-model analyses, we demonstrated that safety mechanisms are not universally implemented but exhibit profound, architecture-dependent failure modes.

Through a multi-method ablation protocol applied across three representative families, we uncovered distinct sensitivity patterns—specifically, LLaMA's brittle functional collapse under signal pollution, DeepSeek's precise safety failure under signal loss, and StableLM's universal fragility. Crucially, our identification of the "Precise Safety Failure" threshold challenges prior assumptions, revealing that safety alignment can be dismantled without sacrificing linguistic coherence. This discovery enables the definition of a "vulnerability upper bound," providing a rigorous intrinsic benchmark that contextualizes the true efficacy of external jailbreak attacks like GCG.

These findings advocate for a paradigm shift in safety defense. One-size-fits-all strategies are insufficient; instead, robustness requires architecture-aware interventions—such as enhancing signal redundancy for DeepSeek-like models or

strictly denoising internal representations for LLaMA-like architectures. Ultimately, by mapping these family-specific mechanistic profiles, this work lays the foundation for designing the next generation of resilient, transparent, and interpretable AI systems.

Declarations

- **Funding** This work was supported by the National Natural Science Foundation of China (Grant Nos. 62572255, 62302238), the Frontier Technology Research and Development Program of Jiangsu (Grant No. BF2024071), the Basic Research Program of Jiangsu (Grant No. BK20250668), the Jiangsu Provincial Young Science and Technology Talent Support Program (Grant No. JSTJ-2025-944), the Science and Technology Major Special Program of Jiangsu (Grant No. BG2024028), and the Basic Science (Natural Science) Research Project of Jiangsu Province Higher Education Institutions (Grant No. 25KJB520039).

- **Competing interests** The authors declare no competing interests.
- **Data Availability** The AdvBench dataset used in this study is publicly available. The datasets generated and/or analysed during the current study (including ablation experiment outputs, attention pattern metrics, linear probe results, and ASR values) are available from the corresponding author on reasonable request.
- **Code Availability** The code used for ablation experiments, attention head probing, visualization, and key result reproduction is publicly available on GitHub at: https://github.com/au9u5tux/LLM-Safety-Attention-Vulnerability. The code is released under an open-source license (MIT).
- **Ethics & Impact Statement** This work investigates architectural mechanisms related to safety behaviors in large language models (LLMs) through mechanistic interpretability and causal intervention analysis, specifically focusing on Critical Safety Attention Heads (CSAHs). Because the study analyzes and intervenes upon internal components that govern safety responses, the findings raise potential dual-use concerns. In particular, insights regarding CSAHs could theoretically be misused to dismantle safety safeguards in open-weight models.

  To significantly mitigate these risks, we emphasize the technical limitations of our approach. Our causal interventions (such as zero-out or undifferentiated ablations) require full white-box access to the model's internal weights and activations during inference. This fundamental prerequisite means our methods cannot be directly weaponized against closed-source commercial APIs. Furthermore, we do not provide any operational guidance or automated prompt-generation tools for malicious exploitation; all evaluations are strictly confined to established public benchmarks (e.g., AdvBench).

  The primary objective of this research is inherently defensive. By exposing architecture-dependent failure modes—such as the critical distinction between "precise safety failure" and "functional collapse"—we aim to provide the research community with an intrinsic vulnerability benchmark. We believe that transparently investigating how safety behaviors arise and fail is crucial for the responsible development of AI. Such knowledge will assist developers in designing architecture-aware

alignment strategies, robust monitoring techniques, and more reliable evaluation frameworks for large-scale AI systems.

- **Authors' contributions** L.S. and P.S. contributed equally to this work as co-first authors. L.S. and P.S. conceptualized the study, developed the core framework ("Family-Specific Vulnerabilities in Safety Attention Networks"), designed the methodology, and contributed to data interpretation and manuscript writing. H.X. conducted the ablation experiments, linear probe analyses, data collection, quantitative metrics computation, prepared the figures, wrote the original draft, managed the submission and transfer processes, and revised the manuscript. S.H. assisted with experiments, code implementation, data visualization, and technical support. F.X. supervised the project, provided resources and funding, reviewed and edited the manuscript, and served as the corresponding author. All authors reviewed and approved the final manuscript.

References

[1] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models (2023). https://doi.org/10.48550/arXiv.2307.15043

[2] Russinovich, M., Salem, A., Eldan, R.: Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. In: Bauer, L., Pellegrino, G. (eds.) 34th USENIX Security Symposium, USENIX Security 2025, Seattle, WA, USA, August 13-15, 2025, pp. 2421–2440. USENIX Association, Seattle, WA, United states (2025). https://www.usenix.org/conference/usenixsecurity25/presentation/russinovich

[3] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K.: A hitchhiker's guide to jailbreaking ChatGPT via prompt engineering. In: SEA4DQ 2024 - Proceedings of the 4th International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things, Co-located With: ESEC/FSE 2024, Porto de Galinhas, Brazil, pp. 12–21 (2024). https://doi.org/10.1145/3663530.3665021

[4] Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, vol. 36. New Orleans, LA, United states (2023). http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html

[5] Zhou, Z., Yu, H., Zhang, X., Xu, R., Huang, F., Wang, K., Liu, Y., Fang, J., Li, Y.: On the role of attention heads in large language model safety. In: 13th International Conference on Learning Representations, ICLR 2025, Singapore, Singapore,

pp. 40641–40670 (2025). https://openreview.net/forum?id=h0Ak8A5yqw

[6] Zhou, Z., Yu, H., Zhang, X., Xu, R., Huang, F., Li, Y.: How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In: EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024, Hybrid, Miami, FL, United states, pp. 2461–2488 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.139

[7] Wei, B., Huang, K., Huang, Y., Xie, T., Qi, X., Xia, M., Mittal, P., Wang, M., Henderson, P.: Assessing the brittleness of safety alignment via pruning and low-rank modifications. CoRR **abs/2402.05162**, 52588–52610 (2024) https://doi.org/10.48550/ARXIV.2402.05162 arXiv:2402.05162

[8] DeepSeek-AI, Bi, X., Chen, D., Chen, a.A.: Deepseek llm: Letting socratic honesty shine in multi-faceted assessments. arXiv preprint arXiv:2401.02954 (2024) https://doi.org/10.48550/arxiv.2401.02954 2401.02954

[9] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) https://doi.org/10.48550/arxiv.2302.13971 2302.13971

[10] AI, S.: Stablelm: Stability ai language models. arXiv preprint arXiv:2304.05878 (2023) https://doi.org/10.48550/arxiv.2304.05878 2304.05878

[11] Liu, X., Xu, N., Chen, M., Xiao, C.: AUTODAN: GENERATING stealthy jailbreak prompts on aligned large language models. In: 12th International Conference on Learning Representations, ICLR 2024. OpenReview.net, Hybrid, Vienna, Austria (2024). https://openreview.net/forum?id=7Jwpw4qKkb

[12] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G.J., Wong, E.: Jailbreaking black box large language models in twenty queries, Copenhagen, Denmark, pp. 23–42 (2025). https://doi.org/10.1109/SaTML64287.2025.00010

[13] Hui, B., Yuan, H., Gong, N., Burlina, P., Cao, Y.: PLeak: Prompt leaking attacks against large language model applications. In: CCS 2024 - Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, Salt Lake City, UT, United states, pp. 3600–3614 (2024). https://doi.org/10.1145/3658644.3670370

[14] Shi, J., Yuan, Z., Liu, Y., Huang, Y., Zhou, P., Sun, L., Gong, N.Z.: Optimization-based prompt injection attack to LLM-as-a-Judge. In: CCS 2024 - Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, Salt Lake City, UT, United states, pp. 660–674 (2024). https://doi.org/10.1145/3658644.3690291

[15] Chang, Z., Li, M., Liu, Y., Wang, J., Wang, Q., Liu, Y.: Play guessing game with LLM: Indirect jailbreak attack with implicit clues. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Hybrid, Bangkok, Thailand, pp. 5135–5147 (2024). https://doi.org/10.18653/v1/2024.findings-acl.304

[16] Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., Poovendran, R.: ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1. Bangkok, Thailand, pp. 15157–15173 (2024). https://doi.org/10.18653/v1/2024.acl-long.809

[17] Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38. Vancouver, BC, Canada, pp. 21527–21536 (2024). https://doi.org/10.1609/aaai.v38i19.30150

[18] Al Ghanim, M., Almohaimeed, S., Zheng, M., Solihin, Y., Lou, Q.: Jailbreaking LLMs with Arabic transliteration and Arabizi. In: EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Hybrid, Miami, FL, United states, pp. 18584–18600 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1034

[19] Jin, H., Zhou, A., Menke, J.D., Wang, H.: Jailbreaking large language models against moderation guardrails via cipher characters. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-1896

[20] Zhou, Y., Zou, H.P., Di Eugenio, B., Zhang, Y.: Large language models are involuntary truth-tellers: Exploiting fallacy failure for jailbreak attacks. In: EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Hybrid, Miami, FL, United states, pp. 13293–13304 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.738

[21] Anil, C., Durmus, E., Panickssery, N., Sharma, M., Benton, J., Kundu, S., Batson, J., Tong, M., Mu, J., Ford, D., Mosconi, F., Agrawal, R., Schaeffer, R., Bashkansky, N., Svenningsen, S., Lambert, M., Radhakrishnan, A., Denison, C., Hubinger, E.J., Bai, Y., Bricken, T., Maxwell, T., Schiefer, N., Sully, J., Tamkin, A., Lanhan, T., Nguyen, K., Korbak, T., Kaplan, J., Ganguli, D., Bowman, S.R., Perez, E., Grosse, R.B., Duvenaud, D.: Many-shot jailbreaking. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-4121

[22] Su, J., Kempe, J., Ullrich, K.: Mission impossible: A statistical perspective on jailbreaking LLMs. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-1210

[23] Hao, Y., Dong, L., Wei, F., Xu, K.: Self-attention attribution: Interpreting information interactions inside transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12963–12971. Association for the Advancement of Artificial Intelligence (AAAI), ??? (2021). https://doi.org/10.1609/aaai.v35i14.17533

[24] Zheng, Z., Zhao, J., Yang, L., He, L., Li, F.: Spot risks before speaking! Unraveling safety attention heads in large vision-language models (2025). https://doi.org/10.48550/arXiv.2501.02029

[25] Zhao, W., Li, Z., Li, Y., Zhang, Y., Sun, J.: Defending large language models against jailbreak attacks via layer-specific editing. In: EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024, Hybrid, Miami, FL, United states, pp. 5094–5109 (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.293

[26] Li, Y., Jiang, H., Wei, Z.: DETAM: Defending LLMs Against Jailbreak Attacks via Targeted Attention Modification (2025). https://doi.org/10.48550/arxiv.2504.13562

[27] Neo, C., Cohen, S.B., Barez, F.: Interpreting context look-ups in transformers: Investigating attention-MLP interactions. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 16681–16697. Association for Computational Linguistics, ??? (2024). https://doi.org/10.18653/v1/2024.emnlp-main.930

[28] Chia, X.W., Pan, J.: Probing latent subspaces in LLM for AI security: Identifying and manipulating adversarial states (2025). https://doi.org/10.48550/arXiv.2503.09066

[29] Wang, P., Zhang, D., Li, L., Tan, C., Wang, X., Zhang, M., Ren, K., Jiang, B., Qiu, X.: InferAligner: Inference-time alignment for harmlessness through cross-model guidance. In: EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Hybrid, Miami, FL, United states, pp. 10460–10479 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.585

[30] Balestriero, R., Cosentino, R., Shekkizhar, S.: Characterizing large language model geometry helps solve toxicity detection and generation. In: Proceedings of Machine Learning Research, vol. 235, pp. 2543–2565. OpenReview.net, Vienna, Austria (2024). https://openreview.net/forum?id=glfcwSsks8

[31] Cao, B., Cao, Y., Lin, L., Chen, J.: Defending against alignment-breaking attacks via robustly aligned LLM. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1. Bangkok, Thailand, pp. 10542–10560 (2024). https://doi.org/10.18653/v1/2024.acl-long.568

[32] Liu, Q., Zhou, Z., He, L., Liu, Y., Zhang, W., Su, S.: Alignment-enhanced decoding: Defending jailbreaks via token-level adaptive refining of probability distributions. In: EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Hybrid, Miami, FL, United states, pp. 2802–2816 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.164

[33] Xu, Z., Jiang, F., Niu, L., Jia, J., Lin, B.Y., Poovendran, R.: SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1. Bangkok, Thailand, pp. 5587–5605 (2024). https://doi.org/10.18653/v1/2024.acl-long.303

[34] Hu, X., Chen, P.-Y., Ho, T.-Y.: Gradient Cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-4011

[35] Liu, Z., Wang, Z., Xu, L., Wang, J., Song, L., Wang, T., Chen, C., Cheng, W., Bian, J.: Protecting your LLMs with information bottleneck. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-0935

[36] Kobayashi, G., Kuribayashi, T., Yokoi, S., Inui, K.: Incorporating residual and normalization layers into analysis of masked language models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4547–4568. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.emnlp-main.373

[37] Zhang, S., Fan, X., Zheng, H., Tanwisuth, K., Zhou, M.: Alignment attention by matching key and query distributions. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34. Virtual, Online, pp. 13444–13457 (2021). https://proceedings.neurips.cc/paper/2021/hash/6fd6b030c6afec018415662d0db43f9d-Abstract.html

[38] Chen, X., Zhao, L., Zou, D.: How transformers utilize multi-head attention in in-context learning? A case study on sparse linear regression. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-3799

[39] Li, C., Wang, S., Zhang, Y., Zhang, J., Zong, C.: Interpreting and exploiting functional specialization in multi-head attention under multi-task learning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16460–16476. Association for Computational Linguistics, ??? (2023). https://doi.org/10.18653/v1/2023.emnlp-main.1026

[40] Merullo, J., Eickhoff, C., Pavlick, E.: Circuit component reuse across

tasks in transformer language models. In: International Conference on Learning Representations. OpenReview.net, Hybrid, Vienna, Austria (2024). https://openreview.net/forum?id=fpoAYV6Wsk

[41] Merullo, J., Eickhoff, C., Pavlick, E.: Talking heads: Understanding inter-layer communication in transformer language models. In: Advances in Neural Information Processing Systems, vol. 37. Vancouver, BC, Canada (2024). https://doi.org/10.52202/079017-1962

[42] Hoover, B., Strobelt, H., Gehrmann, S.: Exbert: A visual analysis tool to explore learned representations in transformer models, Virtual, Online, United states, pp. 187–196 (2020). https://doi.org/10.18653/v1/2020.acl-demos.22

[43] Tufanov, I., Hambardzumyan, K., Ferrando, J., Voita, E.: Lm transparency tool: Interactive tool for analyzing transformer language models, vol. 3. Bangkok, Thailand, pp. 51–60 (2024). https://doi.org/10.18653/v1/2024.acl-demos.6

[44] Li, R., Xiao, W., Xing, L., Wang, L., Murray, G., Carenini, G.: Human guided exploitation of interpretable attention patterns in summarization and topic segmentation, Hybrid, Abu Dhabi, United arab emirates, pp. 10189–10204 (2022). https://doi.org/10.18653/v1/2022.emnlp-main.694

[45] Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., Scotton, P.: Attention-based interpretability with Concept Transformers. In: International Conference on Learning Representations. OpenReview.net, Virtual, Online (2022). https://openreview.net/forum?id=kAa9eDS0RdO

[46] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford Alpaca: An Instruction-following LLaMA Model. https://github.com/tatsu-lab/stanford_alpaca (2023)

[47] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32**, 8024–8035 (2019)

[48] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Virtual, Online, pp. 38–45 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6

[49] Olsson, C., Elhage, N., Nanda, N., Joseph, N., Mann, N., Askell, A., Bathia, A., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Henighan,

D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Clark, J.: In-context Learning and Induction Heads (2022). https://doi.org/10.48550/arXiv.2209.11895

[50] Wang, Z., He, Z., Chu, Z., Xu, H., Zheng, R., Ren, K., Chen, C.: Attention Slipping: A Mechanistic Understanding of Jailbreak Attacks and Defenses in LLMs (2025). https://doi.org/10.48550/arXiv.2507.04365