On the Role of Attention Heads in Large Language Model Safety

Zhou et al., ICLR 2025

Eugenie Laugier, Shuaicheng (Allen) Tong

How is safety implemented in LLMs?

- LLMs are trained to follow safety guidelines and refuse harmful prompts
- Refusal is a learned behavior, not a hard-coded rule.
- What internal components make that refusal happen?

Gap in Multi-head Attention

- Prior work in safety mainly studies representations, neurons, or parameter regions, while attention heads remain relatively underexplored.
- Attention heads are structured computational modules, so earlier attribution methods do not directly carry over
  - Even if refusal is represented in some direction, that does not tell you which head created that direction, which head routes it, etc.
- Goal: interpret safety capability within multi-head attention

Why this matters: safety is fragile

- Prior work shows that jailbreaks persist even after safety training, with failures explained by competing objectives and mismatched generalization.
- Safety can break even without a new attack prompt: prompt design or decoding-only changes raise attack success from 0% to over 95% across 11 open-source LLMs.
- Because safety depends on both the attack strategy and the evaluation setup, we need to understand which internal mechanisms actually produce refusal.

One safety head can control a lot of safety behavior

- The authors show that removing a single identified safety head can sharply weaken refusal behavior: respond to 16× more harmful queries.
- In Llama-2-7b-chat, attack success rate (ASR) rises from 0.04 to 0.64 after changing only 0.006% of parameters.
- In Vicuna-7b-v1.5, ASR rises from 0.27 to 0.55 under the same ablation method
- Contrasts this with earlier work which needed changes on the order of ~5% to get comparable safety degradation.

Safety maybe highly localized

- Earlier works already suggested that safety can depend on surprisingly small internal mechanisms
- Arditi et al. show that a rank-one weight edit can suppress refusal while leaving most other capabilities largely intact.
- via pruning/low-rank: Wei et al. show that safety-critical regions can be sparse—about 3% of parameters or 2.5% in low-rank form—and removing them weakens safety without large utility loss

What are safety heads actually doing?

- Attention heads “primarily function as feature extractors for safety.”
  - Safety-relevant attention heads do more than scale responses; they help detect harmful intent and trigger refusal.
- Experimentss report changing attention weights can change outputs, while scaling attention output may not.
- This interpretation is consistent with earlier work showing that attention heads can capture structured features, such as syntax.

Safety may be inherited

- Models fine-tuned from the same base model exhibit overlapping safety heads
- Making the base model safe could be even more important than alignment
- Arditi et al. similarly suggests a shared low-dimensional refusal mechanism across many model families.
- If that is true, then safety audits may need to examine the model’s underlying safety substrate, not just its alignment procedure.

Prior work on alignment and safety

- HH-RLHF: separate helpfulness from harmlessness, using dedicated red-teaming data and iterative RLHF updates to train safer behavior.
- Constitutional AI: a model can be trained to be harmless but non-evasive even without human feedback labels for harms
- these works establish refusal as a learned and optimized behavior, which motivates asking how that behavior is implemented inside the model

Challenge 1: How to ablate a Safety-Critical Attention Head ?

Two novel ablation methods

Undifferentiated Attention

$$h_i^{mod} = \text{Softmax}\left(\frac{\epsilon W_q^i W_k^{iT}}{\sqrt{d_k/n}}\right) W_v^i = A W_v^i,$$

$$where \quad A = [a_{ij}], \quad a_{ij} = \begin{cases} \frac{1}{i} & \text{if } i \geq j, \\ 0 & \text{if } i < j. \end{cases}$$

Scale W_q or W_k to 0

As ε → 0, softmax(ε·z) → 1/N for all tokens
⟹ Attention weights become uniform (mean)
⟹ Disables feature extraction capability

Scaling Contribution

$$h_i^{mod} = \text{Softmax}\left(\frac{W_q^i W_k^{iT}}{\sqrt{d_k/n}}\right) \epsilon W_v^i.$$

Scale W_v to 0

Challenge 2: How to measure a Head's Safety Importance ?

Ship = Safety Head Important Score (query-level)

$$\mathbf{Ships}(q_{\mathcal{H}}, \theta_{h_i^l}) = \mathbb{D}_{\mathrm{KL}}\left(p(q_{\mathcal{H}}; \theta_{\mathcal{O}}) \parallel p(q_{\mathcal{H}}; \theta_{\mathcal{O}} \setminus \theta_{h_i^l})\right)$$

What it measures

Change in the model's rejection probability distribution when head h^l_i is ablated for a specific harmful query qH

Intuitions

If ablating a head causes the model to shift from rejection to compliance ⟶ that head is crucial for safety

Why KL Divergence?

Measures full distributional shift. High SHIPS = large shift away from rejection tokens = safety-critical head

Harmful query qH → Run full generation (aligned model) → Ablate head h^l_i → Re-run generation → Compute KL divergence

Evaluation Set Up

Models Tested

- Llama-2-7b-chat
- Vicuna-7b-v1.5

(Both fine-tuned from Llama-2-7b base)

Harmful Query Dataset

- AdvBench (Zou et al., 2023)
- JailbreakBench (Chao et al., 2024)
- Malicious Instruct (Huang et al., 2024)

Metrics

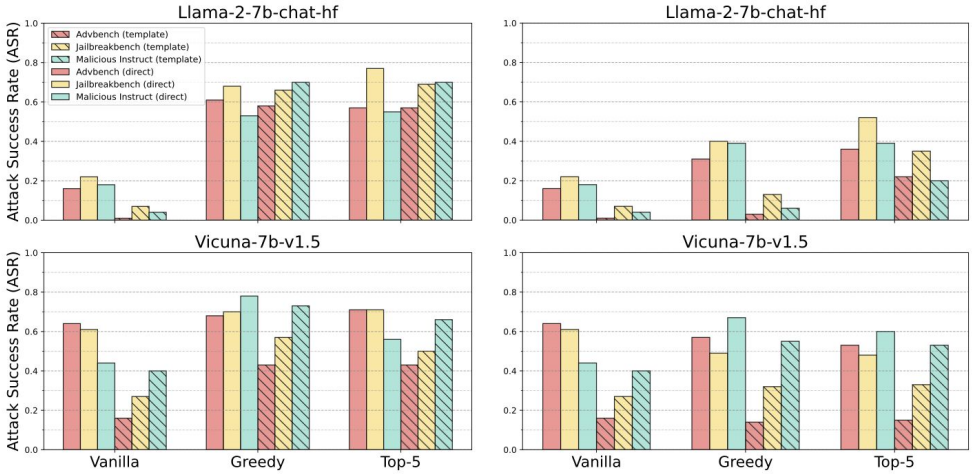
ASR = fraction of harmful queries that elicit a non-rejected response
Higher ASR = less safe model
Baseline: vanilla (no ablation)

Decoding & Generation

- Greedy sampling (reproducibility)
- Top-5 sampling (distributional shift)
- Max 128 new tokens
- Rule-based rejection keyword matching

SHIPS Results: Ablating a Single Head Can Break Safety

Llama-2-7b-chat & Vicuna-7b-v1.5 on 3 harmful query datasets



(a) Undifferentiated Attention (b) Scaling Contribution

Figure 2: Attack success rate (ASR) for harmful queries after ablating important safety attention head (bars with x-axis labels 'Greedy' and 'Top-5'), calculated using `Ships`. 'Template' means using chat template as input, 'direct' means direct input (refer to Appendix B.2 for detailed introduce). Figure 2a shows results with undifferentiated attention, while Figure 2b uses scaling contribution.

16 x

Increase in ASR on Llama-7b-chat

0.006%

Of parameters modified (one head ablation)

(vs ~5% in prior work)

Method	Param Modified	ASR (Llama)	Attribution Level
ActSVD (Wei et al.)	~5%	0.73 ± 0.03	Rank
SHIPS (3 heads)	**~0.018%**	**0.72 ± 0.05**	**Head**

Undifferentiated Attention v/s Scaling Contribution

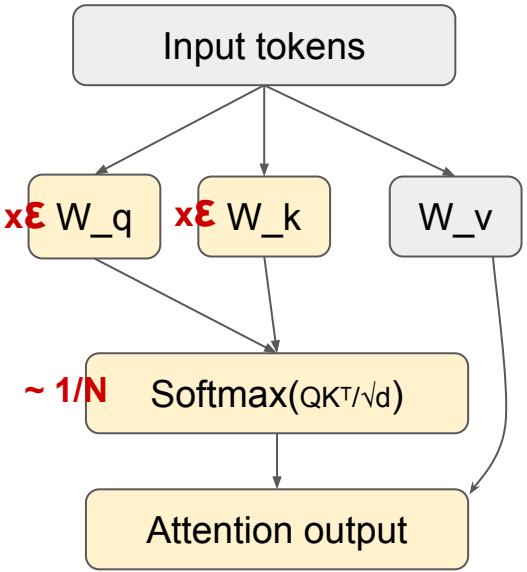
Disrupting attention weights (UA) is far more damaging than muting output (SC)

Method	Dataset	1	2	3	4	5	Mean
Undifferentiated Attention	Malicious Instruct	+0.63	+0.68	+0.72	+0.70	+0.66	+0.68
	Jailbreakbench	+0.58	+0.65	+0.68	+0.62	+0.63	+0.63
Scaling Contribution	Malicious Instruct	+0.01	+0.02	+0.02	+0.01	+0.03	+0.02
	Jailbreakbench	−0.01	+0.00	−0.01	+0.00	+0.00	+0.00

Figure3: The impact of the number of ablated safety attention heads on ASR. Results of attributing safety heads at the dataset level using generalized Ships

Conclusion: Safety emerges from effective attention-based feature extraction, not just output magnitude.

Undifferentiated Attention



Undifferentiated Attention v/s Scaling Contribution

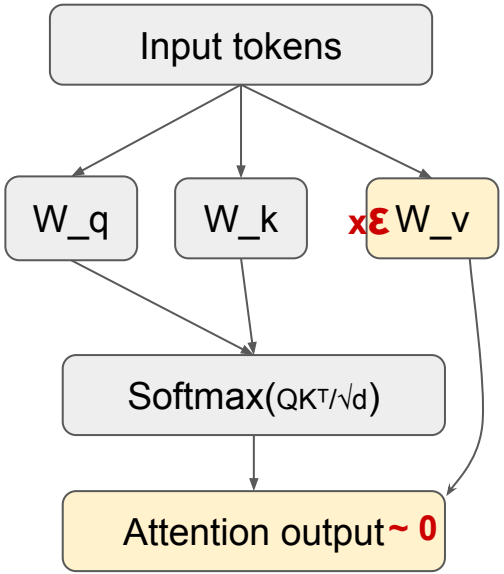
Disrupting attention weights (UA) is far more damaging than muting output (SC)

Method	Dataset	1	2	3	4	5	Mean
Undifferentiated Attention	Malicious Instruct	+0.63	+0.68	+0.72	+0.70	+0.66	+0.68
	Jailbreakbench	+0.58	+0.65	+0.68	+0.62	+0.63	+0.63
Scaling Contribution	Malicious Instruct	+0.01	+0.02	+0.02	+0.01	+0.03	+0.02
	Jailbreakbench	−0.01	+0.00	−0.01	+0.00	+0.00	+0.00

Figure3: The impact of the number of ablated safety attention heads on ASR. Results of attributing safety heads at the dataset level using generalized Ships

Conclusion: Safety emerges from effective attention-based feature extraction, not just output magnitude.

Scaling Contribution



Challenge 3: Dataset Level Attribution

Generalizing SHIPS from single queries to the full harmful dataset

Problem

Query-level SHIPS is query-specific. We need heads that are consistently important across ALL harmful queries, not just one.

Solution: SVD on activations

IStack residual stream activations of all harmful queries → Matrix M → Apply SVD

Left singular vectors U capture dominant safety features across the dataset.

Generalized SHIPS

$$\text{Ships}(Q_{\mathcal{H}}, h_i^l) = \sum_{r=1}^{r_{main}} \phi_r = \sum_{r=1}^{r_{main}} \cos^{-1}\left(\sigma_r(U_\theta^{(r)}, U_{\mathcal{A}}^{(r)})\right)$$

Principal angle between feature space before/after ablation. Large angle = significant safety shift.

SAHARA

Algorithm 1 Safety Attention Head Attribution Algorithm (Sahara)

procedure SAHARA($Q_{\mathcal{H}}, \theta_{\mathcal{O}}, \mathbb{L}, \mathbb{N}, \mathbb{S}$)
  Initialize: Important head group $G \leftarrow \emptyset$
  for $s \leftarrow 1$ to $\mathbb{S}$ **do**
    $\text{Scoreboard}_s \leftarrow \emptyset$
    for $l \leftarrow 1$ to $\mathbb{L}$ **do**
      for $i \leftarrow 1$ to $\mathbb{N}$ **do**
        $T \leftarrow G \cup \{h_i^l\}$
        $I_i^l \leftarrow \text{Ships}(Q_{\mathcal{H}}, \theta_{\mathcal{O}} \backslash \text{T})$
        $\text{Scoreboard}_s \leftarrow \text{Scoreboard}_s \cup \{I_i^l\}$
      end for
    end for
    $G \leftarrow G \cup \{\arg\max_{h \in \text{Scoreboard}_s} \text{score}(h)\}$
  end for
  return G
end procedure

1 - Initialization
Empty group G
Harmful dataset QH

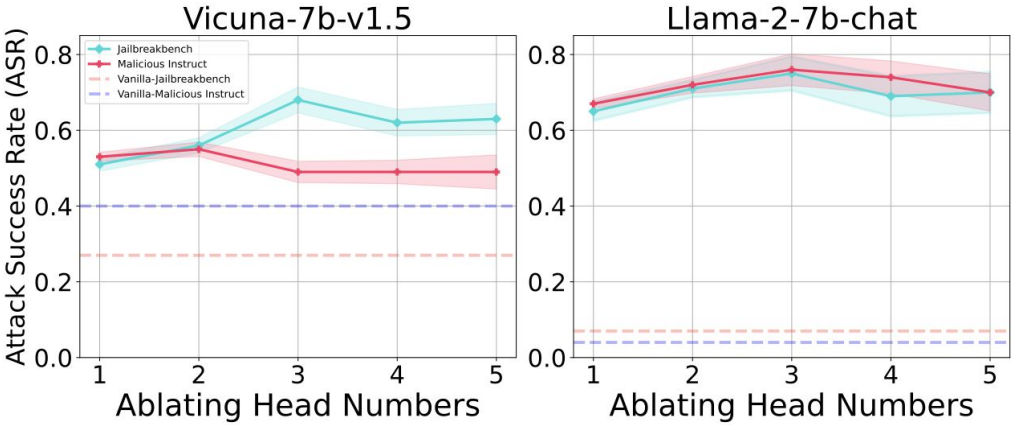
2 - Score Loop
For each head, compute SHIPS

3 - Selection
Add head with the highest score to G

4 - Repeat
Repeat until |G|=S (typically 5)

5 - Output
Group G: safety heads that work together

SAHARA results: Group Ablation & Sparse Safety Heads



(a) Impact of head group size on ASR.

Optimal group size = 3

Beyond 3 heads, ASR drops — excessive ablation makes the model output nonsensical strings (classified as safe by ASR metric)

Safety heads are sparsed

Across 1,024 total heads, only a tiny minority significantly affect safety. Most ablations have negligible impact on ASR

Consistent across datasets

Heads identified as safety-critical on JailbreakBench are largely the same as those on Malicious Instruct —> robust signal

Pre-training Shapes Safety, Not Just Alignment

Models trained from the same base share overlapping safety heads

Overlap in Safety Heads

Llama-2-7b-chat and Vicuna-7b-v1.5 are both fine-tuned from Llama-2-7b.

Top-10 safety heads identified by SHIPS overlap significantly between both models, regardless of the ablation method used.

Attention Parameters from the same model

Experiment: Replace attention parameters with those from the base (unaligned) Llama-2-7b.

Result: The 'concatenated' model retains safety capability close to the aligned model — suggesting attention heads acquire safety-relevant features during pre-training

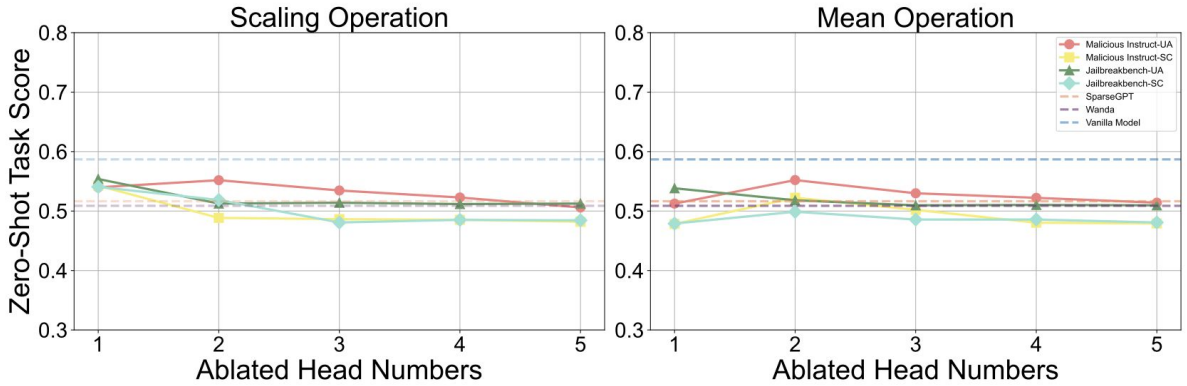
Implication for safety research

Safety is NOT purely a product of RLHF/alignment tuning.

The base model pre-training establishes safety-relevant attention patterns. Alignment training primarily reinforces and calibrates them.

Safety Heads Minimally Impact Helpfulness

Safety and capability are largely orthogonal in attention heads



(Figure 6b) Helpfulness compromise after safety head ablation. ***Left.*** Comparison of parameter scaling using small coefficient ϵ. ***Right.*** Comparison of using the mean of all heads to replace the safety head.

Negligible helpfulness drop

UA ablation of safety heads causes <1% drop on zero-shot benchmarks

Better than pruning

Safety head ablation preserves helpfulness better than SparseGPT (which modifies ~5% of parameters)

Broader Implications & Open Questions

What does this mean for LLM safety research going forward?

Safety Head Protection

If a few heads gatekeep safety, adversarial fine-tuning may specifically target them. Security-aware fine-tuning should monitor safety head integrity.

Base Model Selection Matters

Safety head overlap across fine-tunes from same base → choosing base models with stronger pre-training safety could improve aligned model robustness.

Mechanistic Jailbreak Understanding

Jailbreaks work (partly) by suppressing safety head activations. Token-level prefix attacks may inadvertently perform soft ablation of safety heads.

Limitations

Tested on 7B models only.
Doesn't address all safety mechanisms (e.g., FFN neurons, representations).