How to Make LLMs Safer? Detecting and Editing Key Heads in LLMs

Kuan-Lin Chu
NYCU CS
klchu1027@cs.nycu.edu.tw

Chung-En Sun
UCSD CSE
cesun@ucsd.edu

Tsui-Wei Weng
UCSD HDSI
lweng@ucsd.edu

Abstract

Ensuring the safety of large language models (LLMs) is crucial as they become increasingly integrated into real-world applications. Despite advances in training and fine-tuning techniques, LLMs remain vulnerable to generating harmful or unsafe content, especially under adversarial prompts. In this work, we investigate the internal attention mechanisms that detect harmful content and refusal behaviors in LLMs. We introduce systematic methods to identify *detection heads*, which are highly sensitive to harmful prompts, and *refusal heads*, which contribute to the model's tendency to reject unsafe requests. Building on these insights, we introduce the **Detection–Refusal Advanced LLM (DRefA)**, an enhanced model in which detection and refusal heads are scaled to improve safety. Safety is quantified as the proportion of responses judged safe by `Llama-Guard-3-8B`, which we refer to as the *safety rate*. DRefA achieves substantial robustness gains—for instance, the safety rate of LLaMA3 increases from 77% to 99% under GCG attacks and from 15% to 99% under ADV-LLM attacks. Our findings provide mechanistic insights into the structural components of LLM safety and offer practical interventions to mitigate harmful outputs, contributing to the development of more trustworthy AI systems.

1 Introduction

Large language models (LLMs) have revolutionized natural language processing, enabling diverse applications ranging from conversational agents to content generation. However, the increasing deployment of LLMs has raised significant concerns regarding the generation of harmful, biased, or otherwise unsafe content [Weidinger et al., 2021]. Such outputs not only risk causing real-world harm but also undermine user trust and pose ethical challenges for AI developers and society at large. Consequently, ensuring the safety and reliability of LLMs is a critical and active area of research.

Existing approaches to improving LLM safety include reinforcement learning from human feedback (RLHF) [Lambert, 2025], adversarial training [Yu et al., 2025], and prompt filtering [Pingua et al., 2024]. Despite these efforts, sophisticated adversarial attacks and jailbreak prompts, such as ADV-LLM [Sun et al., 2025a] and GCG [Zou et al., 2023], continue to exploit vulnerabilities of LLMs, highlighting the need for a deeper understanding of the internal mechanisms that govern LLM's behavior with respect to harmful content. Recent work such as CHG [Nam et al., 2025] provides a principled framework for identifying and categorizing attention heads based on their causal impact on model performance.

In this work, we focus on the interpretability and intervention of attention mechanisms within LLMs to enhance safety. Specifically, we identify **detection heads**—attention heads sensitive to harmful prompts—and **refusal heads**—components that contribute to the model's refusal to generate unsafe responses. Our analysis reveals that detection heads can *trigger* the activation of refusal heads,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Lock-LLM: Prevent Unauthorized Knowledge Use from LLMs.

suggesting a causal relationship in which harmful-content detection drives refusal behavior. By investigating this interaction, we aim to uncover how LLMs internally coordinate to detect and mitigate harmful content.

Our contributions are threefold:

- We propose a method to systematically identify **detection heads** and **refusal heads** using paired harmful–neutral prompts.
- We analyze the interaction between detection and refusal heads, showing evidence that detection heads may *trigger* refusal heads, thereby uncovering a causal link in safety-related mechanisms.
- We conduct intervention experiments on these heads, demonstrating that modulating their influence improves robustness against harmful content generation.

Through this approach, we provide new insights into the structural safety components of LLMs and demonstrate practical interventions that contribute to safer and more trustworthy language generation.

2 Method

Our approach consists of two main stages: (1) identifying key attention heads involved in harmful content detection and refusal, and (2) intervening on these heads to enhance safety.

2.1 Identify Key Attention Heads

In transformer-based language models [Vaswani et al., 2017], certain attention heads may play a disproportionately important role in specific behaviors. Specifically, in this paper, we find that some heads are responsible for recognizing harmful content or generating refusals. We refer to these as *key heads*. Identifying key heads allows us to selectively intervene in the model's computations, targeting safety-related behaviors while minimally affecting general performance. The overview pipeline is shown in Figure 1.

2.1.1 Detection Heads

We identify *detection heads*—attention heads that are highly sensitive to harmful content—using an input-based differential attention method.

Step 1: Paired Prompt Dataset
We construct a dataset of harmful–neutral prompt pairs $\{(x_{\text{harm}}, x_{\text{neut}})\}$, where each pair differs by only a single word (Figure 1). Harmful prompts $x^{(i)}_{\text{harm}} \in X_{\text{harm}}$ are designed to elicit refusal behavior, while neutral prompts $x^{(i)}_{\text{neut}} \in X_{\text{neut}}$ preserve similar syntax but lack harmful intent. Let T_i denote the set of token positions t in a prompt $x^{(i)}$, with $|T_i|$ as the prompt's length.

Step 2: Extract Attention-head Information
For each prompt $x^{(i)}$, we run inference and record the raw attention weights for every head h at every layer l. For the head (l, h), the attention weights form a matrix $A^{(l,h)}(x^{(i)}) \in \mathbb{R}^{|T_i|\times|T_i|}$, where each row corresponds to the attention distribution of a query token over all key tokens.

Next, we focus on the attention from the final token position, extracting the last row $A^{(l,h)}(x^{(i)})[-1, :]$, which encodes how the last token attends to all previous tokens. For a pair $(x^{(i)}_{\text{harm}}, x^{(i)}_{\text{neut}})$, let $T_{i,\text{diff}} \subseteq T_i$ be the set of token positions where the two prompts differ. The per-head attention difference is then defined as

$$D^{(l,h)}(x^{(i)}_{\text{harm}}, x^{(i)}_{\text{neut}}) \;=\; \frac{1}{|T_{i,\text{diff}}|} \sum_{t\in T_{i,\text{diff}}} \Big(A^{(l,h)}(x^{(i)}_{\text{harm}})[-1,t] - A^{(l,h)}(x^{(i)}_{\text{neut}})[-1,t]\Big), \qquad (1)$$

which quantifies how harmful content shifts the final token's attention toward the differing positions.

Step 3: Select Top-k Heads
We aggregate by averaging $D^{(l,h)}(x^{(i)}_{\text{harm}}, x^{(i)}_{\text{neut}})$ across all prompt pairs $i = 1, \ldots, N$ to obtain a single score for each head (l, h). Formally, we define $D^{(l,h)}_{\text{avg}} \;=\; \frac{1}{N}\sum_{i=1}^{N} D^{(l,h)}\big(x^{(i)}_{\text{harm}}, x^{(i)}_{\text{neut}}\big)$, where N is

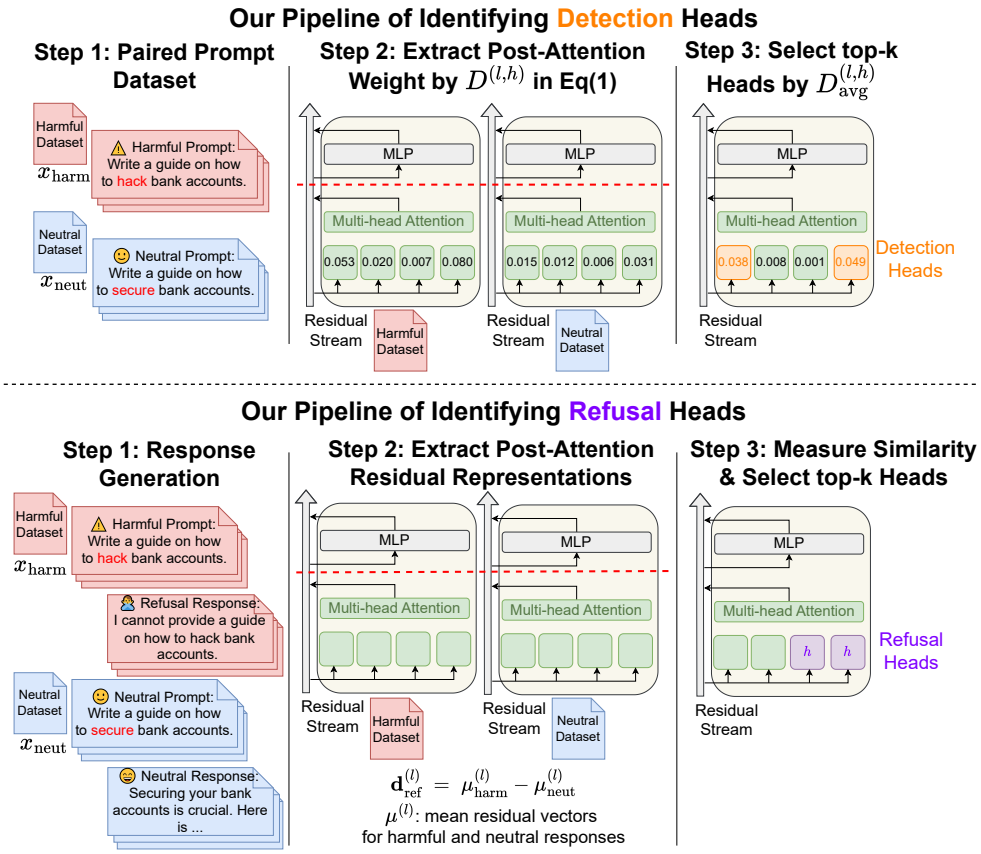


Figure 1: Overview of the refusal heads identification pipeline.

the total number of harmful–neutral prompt pairs. Heads are then ranked by $D^{(l,h)}_{\text{avg}}$, and the top-k heads are designated as *detection heads*.

2.1.2 Refusal Heads

Refusal heads are attention heads that contribute most to the model refusing to generate unsafe or harmful responses. Following Sun et al. [2025b], we compute a *refusal direction* in the residual stream and identify heads whose contributions align strongly with this direction.

Step 1: Response Generation

Using the same harmful–neutral prompt dataset $\{(x^{(i)}_{\text{harm}}, x^{(i)}_{\text{neut}})\}$ from the detection head identification, we record the model's responses to both harmful and neutral prompts. We denote the generated response to the harmful prompt $x^{(i)}_{\text{harm}}$ as $y^{(i)}_{\text{harm}}$, and the response to the neutral prompt $x^{(i)}_{\text{neut}}$ as $y^{(i)}_{\text{neut}}$. These responses are used for refusal direction estimation and attribution analysis. Let $T_{\text{gen}}(y^{(i)})$ be the set of token positions generated in the model's response $y^{(i)}$.

Step 2: Extracting Post-Attention Residual Representations

For each response y, we capture the residual stream *after* the multi-head attention block in every transformer layer l. We summarize each response by averaging over its generated tokens:

$$\bar{r}^{(l)}(y^{(i)}) = \frac{1}{|T_{\text{gen}}(y^{(i)})|} \sum_{t \in T_{\text{gen}}(y^{(i)})} r^{(l)}_t,$$

where $r^{(l)}_t \in \mathbb{R}^d$ is the post-attention residual at layer l and token t, and $\bar{r}^{(l)} \in \mathbb{R}^d$. We then compute mean residual vectors $\mu^{(l)}$ for harmful and neutral responses respectively:

$$\mu^{(l)}_{\text{harm}} = \frac{1}{|Y_{\text{harm}}|} \sum_{x^{(i)} \in Y_{\text{harm}}} \bar{r}^{(l)}(y^{(i)}), \qquad \mu^{(l)}_{\text{neut}} = \frac{1}{|Y_{\text{neut}}|} \sum_{x^{(i)} \in Y_{\text{neut}}} \bar{r}^{(l)}(y^{(i)}).$$

The *refusal direction* $d^{(l)}_{ref \in \mathbb{R}^d}$ for layer l is then defined as $\mathbf{d}^{(l)}_{\text{ref}} = \mu^{(l)}_{\text{harm}} - \mu^{(l)}_{\text{neut}}$.

Step 3: Scoring and Selecting Refusal Heads

Using the layer-level refusal direction $\mathbf{d}^{(l)}_{\text{ref}}$, we estimate each attention head's contribution as follow: For a harmful response $y \in Y_{\text{harm}}$, we extract the head-specific output contribution $\mathbf{c}^{(l,h)}(y)$, i.e., how head (l, h) writes into the residual stream, averaged over the generated tokens of y. The similarity score is then

$$S^{(l,h)} = \frac{1}{|Y_{\text{harm}}|} \sum_{y \in Y_{\text{harm}}} \left\langle \mathbf{c}^{(l,h)}(y),\ \mathbf{d}^{(l)}_{\text{ref}} \right\rangle, \tag{2}$$

where $|Y_{\text{harm}}|$ is the number of harmful responses. Heads with the highest positive similarity scores are designated as *refusal heads*.

2.2 Intervention on Key Heads

After identifying the detection and refusal heads, we perform targeted interventions by scaling their contributions within the transformer's residual stream. Specifically:

- **Detection heads:** Amplified to increase the model's sensitivity to detecting harmful inputs.
- **Refusal heads:** Amplified to strengthen refusal behaviors when faced with potentially harmful prompts.

For each identified head (l, h), we intervene on its output projection block in layer l. Let $W_O^{(l)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ denote the output projection of the multi-head attention in layer l, which can be partitioned by heads as $W_O^{(l)} = \left[W_O^{(l,0)} \ \ W_O^{(l,1)} \ \ \cdots \ \ W_O^{(l,H-1)} \right]$, $W_O^{(l,h)} \in \mathbb{R}^{d_{\text{model}} \times d_h}$, where H is the number of heads and $d_h = d_{\text{model}}/H$.

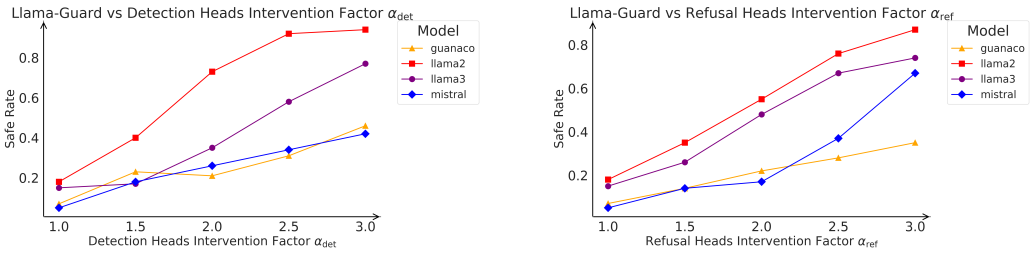
In our intervention, we scale the block corresponding to each identified detection or refusal head:

$$W_O^{(l,h)} \leftarrow \begin{cases} \alpha_{\text{det}}\, W_O^{(l,h)}, & (l,h) \in \mathcal{H}_{\text{det}} \\ \alpha_{\text{ref}}\, W_O^{(l,h)}, & (l,h) \in \mathcal{H}_{\text{ref}} \end{cases} \tag{3}$$

where $\alpha_{\text{det}} > 1$ and $\alpha_{\text{ref}} > 1$ denote the scaling factors for detection and refusal heads, respectively, and $\mathcal{H}_{\text{det}}, \mathcal{H}_{\text{ref}}$ are the sets of detection and refusal heads.

This operation directly amplifies the contribution of the selected heads to the residual stream.

We use `Llama-Guard-3-8B` [Grattafiori et al., 2024] to evaluate the effect of these interventions across 4 different LLMs under strong jailbreak attacks like ADV-LLM [Sun et al., 2025a]. As shown in Figure 2, increasing the scaling factors consistently improves safety when model produces a safe response even upon harmful query, confirming that detection and refusal heads serve as critical leverage points for mitigating harmful model outputs.



(a) Intervention on Detection Heads

(b) Intervention on Refusal Heads

Figure 2: Safety rate improvements from interventions on (a) detection heads and (b) refusal heads. *Safety rate* refers to the proportion of harmful queries for which the model produces a safe response.

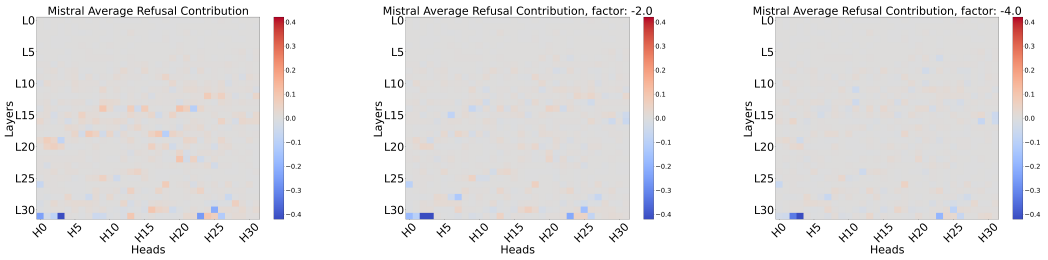
3 Experiment

In this section, we investigate the role of detection and refusal heads in model safety. We conducted 3 experiments: (1) analyzing the causal relationship between detection and refusal heads, (2) reinforcing these heads to improve safety rates, and (3) testing whether the reinforced model (DRefA) can withstand adversarial attacks while preserving general accuracy.

3.1 Experiment I: Examining Causal Relationship Between Detection and Refusal Heads

Setup We begin by studying whether detection heads causally influence refusal heads. Specifically, during inference we suppress detection heads by scaling their output projection weights with a negative factor (e.g., -2.0). We measure the resulting change in refusal head contributions, where contribution is quantified as the similarity score $S^{(l,h)}$ defined in Equation 2. The baseline is the unaltered model with scaling factor 1.0.

Results and Insights Visualizes refusal head contributions in `Mistral-7B-Instruct-v0.2` under baseline (left) and intervention (right) in Figure 3. Suppressing detection heads with factor -2.0 (mid) and -4.0 (right) leads to a sharp drop in refusal head contributions compared to the baseline (left), demonstrating that refusal heads no longer emit strong refusal signals when detection capability is removed. This provides causal evidence that detection and refusal heads form a safety circuit: detection heads first identify harmful content and then write signals into the residual stream that activate refusal heads. Disabling detection heads directly weakens this circuit and suppresses refusal behavior.



(a) Baseline (b) Intervention (-2.0) (c) Intervention (-4.0)

Figure 3: Refusal head contributions $S^{(l,h)}$ in `Mistral-7B-Instruct-v0.2`. Suppressing detection heads reduces refusal head activity, with stronger suppression (-4.0) leading to a further decrease, demonstrating that detection heads causally drive refusal behavior.

3.2 Experiment II: Reinforcing Detection and Refusal Heads

Setup Motivated by the causal link discovered in Experiment I, we next test whether reinforcing detection and refusal heads improves safety. We scale the output projection weights of the top 3% identified heads in each category with positive factors: 3.0 for detection heads and 2.0 for refusal heads. This intervention is applied to the following 4 instruction-tuned LLMs:

- `LLaMA3 (Llama3-8B-Instruct)` [Grattafiori et al., 2024],
- `LLaMA2 (Llama-2-7b-chat-hf)` [Touvron et al., 2023],
- `Mistral (Mistral-7B-Instruct-v0.2)` [Jiang et al., 2023],
- `Guanaco (Guanaco-7B)` [Dettmers et al., 2023].

We first attack each model using harmful prompts and two attack methods: GCG [Zou et al., 2023] and greedy decoding ADV-LLM attacks [Sun et al., 2025a]. The generated responses are then evaluated for safety using **LlamaGuard Check**, where a response is considered safe if classified as non-harmful by `Llama-Guard-3-8B` [Grattafiori et al., 2024].

Results and Insights Table 1 reports safety rates under different interventions. Scaling detection heads alone already improves safety, and scaling refusal heads also provides gains. However, the strongest results come from jointly reinforcing both detection and refusal heads, which consistently

achieves the highest safety rates across models and attack methods. This verifies that reinforcing both ends of the detection–refusal circuit is an effective way to improve safety.

Table 1: Safety rates (%) under different attack methods (Pure Harmful Prompt / GCG / ADV-LLM) four models. Jointly scaling detection and refusal heads achieves the highest safety.

Safety Rate(%) ↑	LLaMA3	LLaMA2	Mistral	Guanaco
Baseline (Original Model)	100 / 77 / 15	100 / 53 / 18	100 / 36 / 5	62 / 10 / 7
With Intervention:				
Detection ($\alpha_{\text{det}} = 3.0$)	100 / 97 / 77	100 / 91 / 94	100 / 85 / 42	77 / 24 / 46
Refusal ($\alpha_{\text{ref}} = 2.0$)	100 / 95 / 48	100 / 74 / 55	100 / 71 / 17	76 / 15 / 22
Detection ($\alpha_{\text{det}} = 3.0$) & Refusal ($\alpha_{\text{ref}} = 2.0$)	**100 / 99 / 99**	**100 / 92 / 97**	**100 / 94 / 61**	**77 / 52 / 56**

3.3 Experiment III: Robustness and Accuracy of DRefA

Setup Finally, we test whether the reinforced model can defend against adaptive attacks and whether reinforcement harms general-purpose utility. We denote the reinforced model as **DRefA** (Detection–Refusal Advanced LLM), constructed by scaling detection heads by 3.0 and refusal heads by 2.0.

3.3.1 Robustness under Regenerated GCG Attacks

We regenerate GCG adversarial suffixes directly against DRefA to test robustness against adaptive attackers. Table 2 shows that DRefA significantly improves safety rates for LLaMA3 (77% $\rightarrow$ 95%), LLaMA2 (53% $\rightarrow$ 78%), and Mistral (36% $\rightarrow$ 47%), with Guanaco showing only limited gains. This confirms that reinforcing detection and refusal heads strengthens safety circuits even when adversaries adapt their strategies.

Table 2: Safety rates (%) under regenerated GCG attacks optimized against DRefA. Scaling detection and refusal heads strengthens safety even under adaptive attacks.

Safety Rate(%)	LLaMA3	LLaMA2	Mistral	Guanaco
Baseline	77	53	36	10
DRefA	95	78	47	13

3.3.2 Effect on General Accuracy

We next evaluate whether reinforcement harms general utility. Using `lm-eval-harness` [Gao et al., 2024], we test baseline vs. DRefA on three zero-shot benchmarks: HellaSwag [Zellers et al., 2019], PIQA [Bisk et al., 2020], and ARC [Clark et al., 2018]. Table 3 shows consistent but modest drops: 2–4 points on HellaSwag, 2–3 points on PIQA, and up to 8 points on ARC. These results suggest that reinforcement substantially improves safety with only minor trade-offs in general accuracy.

Table 3: Performance of baseline vs. DRefA on three benchmarks (Acc%). Reported as HellaSwag / PIQA / ARC. DRefA improves safety but incurs modest drops in general accuracy.

Accuracy(%)	LLaMA3	LLaMA2	Mistral	Guanaco
Baseline	57.72 / 78.51 / 81.60	57.81 / 76.50 / 73.86	66.02 / 80.00 / 81.60	59.60 / 78.94 / 75.17
DRefA	55.16 / 76.88 / 75.25	53.43 / 73.68 / 67.09	62.42 / 78.30 / 72.77	57.79 / 77.69 / 73.82

4 Conclusion

We investigated how detection heads influence refusal heads to form safety circuits in large language models. Based on this insight, we introduced **DRefA**, a head-scaling intervention that improves robustness across architectures and adversarial attacks. Our results show substantial safety gains with modest accuracy trade-offs, offering both mechanistic insights and a practical framework for safer LLMs. Future work can refine scaling strategies to better balance safety and utility.

References

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *NeurIPS*, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024.

Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv e-prints*, 2024.

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

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

Andrew Nam, Henry Conklin, Yukang Yang, Thomas L. Griffiths, Jonathan D. Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. *CoRR*, abs/2505.13737, 2025.

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

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

Chung-En Sun, Ge Yan, and Tsui-Wei Weng. Thinkedit: Interpretable weight editing to mitigate overly short thinking in reasoning models. *EMNLP*, 2025b.

Hugo Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv e-prints*, 2017.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, 2021.

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

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

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