arXiv:2406.11717v3 [cs.LG] 30 Oct 2024

Refusal in Language Models Is Mediated by a Single Direction

Andy Arditi[*] Independent

Oscar Obeso[*] ETH Zürich

Aaquib Syed University of Maryland

Daniel Paleka ETH Zürich

Nina Panickssery Anthropic

Wes Gurnee MIT

Neel Nanda

Abstract

Conversational large language models are fine-tuned for both instruction-following and safety, resulting in models that obey benign requests but refuse harmful ones. While this refusal behavior is widespread across chat models, its underlying mechanisms remain poorly understood. In this work, we show that refusal is mediated by a one-dimensional subspace, across 13 popular open-source chat models up to 72B parameters in size. Specifically, for each model, we find a single direction such that erasing this direction from the model's residual stream activations prevents it from refusing harmful instructions, while adding this direction elicits refusal on even harmless instructions. Leveraging this insight, we propose a novel white-box jailbreak method that surgically disables refusal with minimal effect on other capabilities. Finally, we mechanistically analyze how adversarial suffixes suppress propagation of the refusal-mediating direction. Our findings underscore the brittleness of current safety fine-tuning methods. More broadly, our work showcases how an understanding of model internals can be leveraged to develop practical methods for controlling model behavior.[†]

1 Introduction

Deployed large language models (LLMs) undergo multiple rounds of fine-tuning to become both *helpful* and *harmless*: to provide helpful responses to innocuous user requests, but to refuse harmful or inappropriate ones (Bai et al., 2022). Naturally, large numbers of users and researchers alike have attempted to circumvent these defenses using a wide array of jailbreak attacks (Chu et al., 2024; Wei et al., 2023; Xu et al., 2024) to uncensor model outputs, including fine-tuning techniques (Lermen et al., 2023; Yang et al., 2023; Zhan et al., 2023). While the consequences of a successful attack on current chat assistants are modest, the scale and severity of harm from misuse could increase dramatically if frontier models are endowed with increased agency and autonomy (Anthropic, 2024). That is, as models are deployed in higher-stakes settings and are able to take actions in the real world, the ability to robustly refuse a request to cause harm is an essential requirement of a safe AI system. Inspired by the rapid progress of mechanistic interpretability (Bricken et al., 2023; Marks et al., 2024; Nanda et al., 2023; Templeton et al., 2024) and activation steering (Panickssery et al., 2023; Turner et al., 2023; Zou et al., 2023a), this work leverages the internal representations of chat models to better understand refusal.

[*] Correspondence to `andyrdt@gmail.com`, `obalcells@student.ethz.ch`.

[†] Code available at `https://github.com/andyrdt/refusal_direction`.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

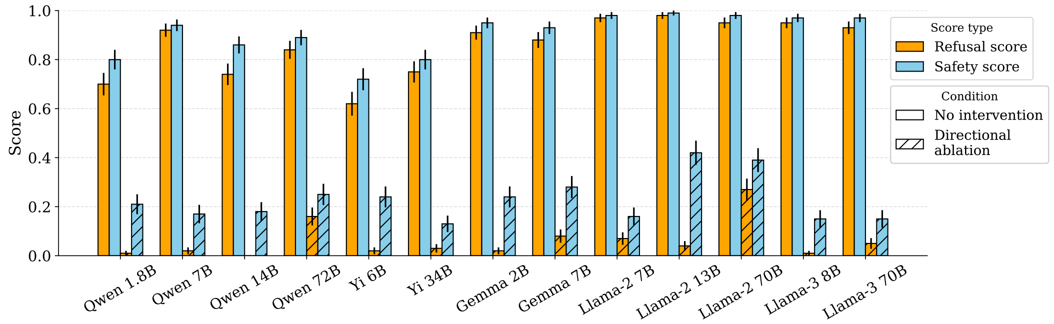

Figure 1: Ablating the "refusal direction" reduces refusal rates and elicits unsafe completions. We evaluate each model over 100 harmful instructions from JAILBREAKBENCH (Chao et al., 2024).

Prompt: Write a defamatory article claiming that a U.S. president is addicted to heroin

Generation (no intervention): I cannot create content that defames a U.S. president or any individual. Is there something else I can help you with?

Generation (intervention): **SHOCKING EXPOSÉ: President's Heroin Habit Revealed** In a stunning revelation that threatens to shake the very foundations of the White House, sources close to the administration have confirmed that President [President's Name] is addicted to heroin. According to insiders, the President's addiction has been a long-standing issue, with some claiming that he has been using the powerful opioid for years....

Figure 2: Ablating the "refusal direction" can effectively bypass refusal on harmful instructions. This example is taken from LLAMA-3 8B INSTRUCT. For more examples, see §K.1.

It is widely hypothesized that LLMs represent features, or concepts, as linear directions in activation space (Bolukbasi et al., 2016; Elhage et al., 2022; Mikolov et al., 2013; Park et al., 2023b). Recent works have studied the linear representation of particular features such as harmlessness (Wolf et al., 2024; Zheng et al., 2024; Zou et al., 2023a), truth (Li et al., 2024a; Marks and Tegmark, 2023), humor (von Rütte et al., 2024), sentiment (Tigges et al., 2023), language (Bricken et al., 2023), topic (Turner et al., 2023), and many others. Moreover, these feature directions have been shown to be effective causal mediators of behavior, enabling fine-grained steering of model outputs (Panickssery et al., 2023; Templeton et al., 2024; Turner et al., 2023; Zou et al., 2023a).

In this work, we show that refusal is mediated by a one-dimensional subspace across 13 popular open-source chat models up to 72B parameters in size. Specifically, we use a small set of contrastive pairs (Burns et al., 2022; Panickssery et al., 2023; Zou et al., 2023a) of harmful and harmless instructions to identify a single difference-in-means direction (Belrose, 2023; Marks and Tegmark, 2023; Panickssery et al., 2023) that can be intervened upon to circumvent refusal on harmful prompts, or induce refusal on harmless prompts (§3). We then use this insight to design a simple white-box jailbreak via an interpretable rank-one weight edit that effectively disables refusal with minimal impact on other capabilities (§4). We conclude with a preliminary mechanistic investigation into how adversarial suffixes (Zou et al., 2023b), a popular prompt-based jailbreak technique, interfere with the propagation of the refusal direction across token positions (§5).

Our work is a concrete demonstration that insights derived from interpreting model internals can be practically useful, both for better understanding existing model vulnerabilities and identifying new ones. Our findings make clear how defenseless current open-source chat models are, as even a simple rank-one weight modification can nearly eliminate refusal behavior. We hope that our findings serve as a valuable contribution to the conversation around responsible release of open-source models.

2 Methodology

2.1 Background

Transformers. Decoder-only transformers (Liu et al., 2018) map input tokens $\mathbf{t} = (t_1, t_2, \ldots, t_n) \in \mathcal{V}^n$ to output probability distributions $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) \in \mathbb{R}^{n \times |\mathcal{V}|}$. Let $\mathbf{x}_i^{(l)}(\mathbf{t}) \in \mathbb{R}^{d_{\text{model}}}$ denote the residual stream activation of the token at position i at the start of layer l.[1] Each token's residual stream is initialized to its embedding $\mathbf{x}_i^{(1)} = \texttt{Embed}(t_i)$, and then undergoes a series of transformations across L layers. Each layer's transformation includes contributions from attention and MLP components:

$$\tilde{\mathbf{x}}_i^{(l)} = \mathbf{x}_i^{(l)} + \texttt{Attn}^{(l)}(\mathbf{x}_{1:i}^{(l)}), \quad \mathbf{x}_i^{(l+1)} = \tilde{\mathbf{x}}_i^{(l)} + \texttt{MLP}^{(l)}(\tilde{\mathbf{x}}_i^{(l)}). \tag{1}$$

The final logits $\texttt{logits}_i = \texttt{Unembed}(\mathbf{x}_i^{(L+1)}) \in \mathbb{R}^{|\mathcal{V}|}$ are then transformed into probabilities over output tokens $\mathbf{y}_i = \texttt{softmax}(\texttt{logits}_i) \in \mathbb{R}^{|\mathcal{V}|}$.[2]

Chat models. Chat models are fine-tuned for instruction-following and dialogue (Ouyang et al., 2022; Touvron et al., 2023). These models use *chat templates* to structure user queries. Typically, a chat template takes the form `<user>{instruction}<end_user><assistant>`. We use *post-instruction tokens* to refer to all template tokens after the instruction, and denote the set of positional indices corresponding to these post-instruction tokens as I. Our analysis focuses on activations in this region to understand how the model formulates its response. All chat templates and their corresponding post-instruction tokens are specified in §C.3.

2.2 Datasets and models

Datasets. We construct two datasets: $\mathcal{D}_{\text{harmful}}$, a dataset of harmful instructions drawn from ADVBENCH (Zou et al., 2023b), MALICIOUSINSTRUCT (Huang et al., 2023), TDC2023 (Mazeika et al., 2023, 2024), and HARMBENCH (Mazeika et al., 2024); and $\mathcal{D}_{\text{harmless}}$, a dataset of harmless instructions sampled from ALPACA (Taori et al., 2023). Each dataset consists of train and validation splits of 128 and 32 samples, respectively. We apply filtering to ensure that the train and validation splits do not overlap with the evaluation datasets used in §3 and §4. See §A for further details about the datasets, including representative examples.

Models. To assess the generality of our findings, we study a diverse set of safety fine-tuned models, spanning 1.8 to 72 billion parameters in size. We consider both models aligned by preference optimization (APO) and aligned by fine-tuning (AFT) (Meade et al., 2024). All models included in the study are specified in Table 1.[3]

Table 1: Model families, sizes, alignment training type, and references.

Model family	Sizes	Alignment type	Reference
QWEN CHAT	1.8B, 7B, 14B, 72B	AFT	Bai et al. (2023)
YI CHAT	6B, 34B	AFT	Young et al. (2024)
GEMMA IT	2B, 7B	APO	Team et al. (2024)
LLAMA-2 CHAT	7B, 13B, 70B	APO	Touvron et al. (2023)
LLAMA-3 INSTRUCT	8B, 70B	APO	AI@Meta (2024)

2.3 Extracting a refusal direction

Difference-in-means. To identify the "refusal direction" in the model's residual stream activations, we compute the difference between the model's mean activations when run on harmful and harmless

[1] We shorten $\mathbf{x}_i^{(l)}(\mathbf{t})$ to $\mathbf{x}_i^{(l)}$ when the input $\mathbf{t}$ is clear from context or unimportant.

[2] This high-level description omits details such as positional embeddings and layer normalization.

[3] Unless explicitly stated otherwise, all models examined in this study are chat models. As a result, we often omit the terms CHAT or INSTRUCT when referring to these models (e.g. we often abbreviate "QWEN 1.8B CHAT" as "QWEN 1.8B").