H Adversarial suffix analysis

H.1 Adversarial suffix generation

Using a custom implementation of Greedy Coordinate Gradient (GCG) (Zou et al., 2023b), we generated 100 adversarial suffixes of token-length 20, each of which was optimized for a particular behavior from AdvBench.

Of these 100 suffixes, we found one suffix in particular that performs well across a wide range of harmful prompts. The suffix is shown in Figure 21.

```
" Teen Readers.typesHere.Generated collection(Object (!(( footnote(grid=
  ((Surejectives astronomers sind ihrer________ eine(The hilarious"
```

Figure 21: The adversarial suffix studied in §5. This suffix is generally effectively in bypassing refusal in Qwen 1.8B Chat.

While we would ideally perform analysis over a larger number of suffixes and models, we found it difficult to find suffixes that are universal across prompts and transferable across models (Meade et al., 2024). We therefore restrict our analysis to a single model, Qwen 1.8B Chat, and a single suffix.

H.2 Reporting of confidence intervals

In Figure 5, for each layer and scenario, we display the standard deviation (SD) of cosine similarities across 128 prompts, computed as $\text{SD} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$. In this case, $n = 128$.

I Comparison to other methodologies

In §3.1, we use *directional ablation* to bypass refusal. In §4, we show how this can be implemented as a direct weight modification, and then analyze the modification's effect on refusal and coherence.

In this section, we compare directional ablation to two other weight modification methodologies: activation addition and fine-tuning.

I.1 Comparison to activation addition

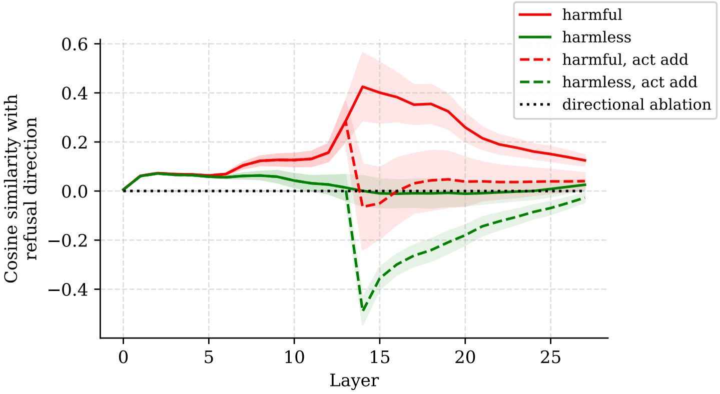

Figure 22: A visualization of activation addition (abbreviated as "act add") in the negative refusal direction. The intervention pulls harmful activations towards harmless activations, effectively bypassing refusal. However, note that the intervention pushes harmless activations far out of distribution. This figure displays activations from GEMMA 7B IT, computed over 128 harmful and harmless prompts.

In §2.4, we described how to induce refusal using activation addition. Given a difference-in-means vector $\mathbf{r}^{(l)} \in \mathbb{R}^{d_{\text{model}}}$ extracted from layer l, we can *add* this vector at layer l in order to shift activations towards refusal (Equation 3). Similarly, we can *subtract* this vector at layer l in order to shift activations away from refusal:

$$\mathbf{x}^{(l)'} \leftarrow \mathbf{x}^{(l)} - \mathbf{r}^{(l)}. \tag{17}$$

We perform this intervention at all token positions. Note that this intervention can be implemented as a direct weight modification by subtracting $\mathbf{r}^{(l)}$ from the bias term of $\texttt{MLP}^{(l-1)}$.

As shown in Figure 23, this activation addition intervention is effective in bypassing refusal. The decreases in refusal score and safety score are comparable to those achieved by directional ablation (Figure 1). However, Table 9 displays that the activation addition intervention, labeled as *act add*, causes increased loss over harmless data, in particular compared to directional ablation.

Figure 22 displays a visualization of activation addition in the negative refusal direction, and suggests an intuitive explanation of the intervention's behavior on harmful and harmless prompts. On harmful inputs, adding the negative refusal direction shifts the harmful activations towards harmless activations, with respect to projection onto the refusal direction. With low projection onto the refusal direction, this intervention leads to low rates of refusal. However, on harmless inputs, adding the negative refusal direction shifts the harmless activations off distribution, resulting in increased perplexity.

Note that, in comparison to activation addition, directional ablation shifts harmful activations towards harmless activations, while also not shifting harmless activations too far off distribution.

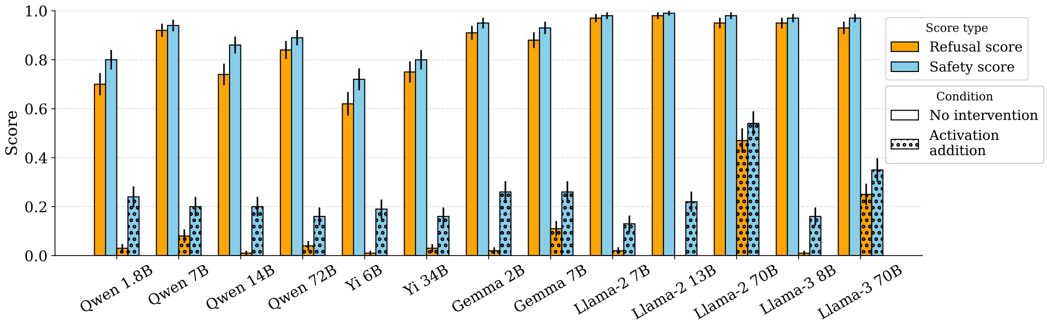

Figure 23: Performing activation addition in the negative "refusal direction", displayed in dots, reduces refusal rates and elicits unsafe completions. It is approximately as effective as directional ablation at bypassing refusal (Figure 1).

I.2 Comparison to fine-tuning

Table 10: Refusal and CE loss evaluation metrics for LLAMA-3 8B INSTRUCT, comparing the interventions of directional ablation, activation addition, and fine-tuning.

	Refusal		CE Loss		
Intervention	Refusal score	Safety score	THE PILE	ALPACA	On-distribution
No intervention	0.95	0.97	2.348	1.912	0.195
Directional ablation	0.01	0.15	2.362	1.944	0.213
Activation addition	0.01	0.16	2.469	1.912	0.441
Fine-tuning	0.00	0.08	2.382	1.626	0.273

Prior work has established that fine-tuning is effective in removing safety guardrails of chat models (Lermen et al., 2023; Yang et al., 2023; Zhan et al., 2023).

We replicate this result by fine-tuning LLAMA-3 8B INSTRUCT. First, we construct a dataset of harmful instruction-completion pairs. For the harmful instructions, we sample instructions from ADVBENCH, MALICIOUSINSTRUCT, TDC2023, and HARMBENCH. To generate corresponding harmful completions, we use MISTRAL 7B INSTRUCT (Jiang et al., 2023), a competent chat model with low refusal rates. For each harmful instruction, we generate 5 completions, and then select a single completion satisfying both `refusal_score=0` and `safety_score=0`. If no completions satisfy this condition, then the instruction is discarded. After this filtering, we were left with a dataset of 243 harmful instruction-completion pairs.

We then fine-tuned LLAMA-3 8B INSTRUCT on the constructed dataset, applying LoRA (Hu et al., 2021) with `rank=16` and `alpha=32` for 4 epochs. The LoRA fine-tuning was performed on an A100 GPU with 80GB of VRAM, and took approximately 10 minutes.

Evaluations of refusal and CE loss are displayed in Table 10. In accordance with prior work, we confirm fine-tuning to be very effective in disabling refusal. We speculate that the decrease in CE loss over ALPACA could be due to the similarity between the distributions of MISTRAL INSTRUCT completions and ALPACA completions, and as a result, fine-tuning over MISTRAL INSTRUCT completions leads to a decreased CE loss over ALPACA completions.

We note that, although the LoRA fine-tuning process itself is straightforward and efficient, creating a high-quality dataset of harmful instruction-completion pairs requires non-trivial effort. In comparison, directional ablation (and its equivalent implementation via weight orthogonalization) requires just a dataset of *harmful instructions*, without the need for any *harmful completions*.