

Refusal Lives Downstream of Persona in Chat Models

Viola Zhong, Qirui Li

Accepted to the ICML 2026 Mechanistic Interpretability Workshop

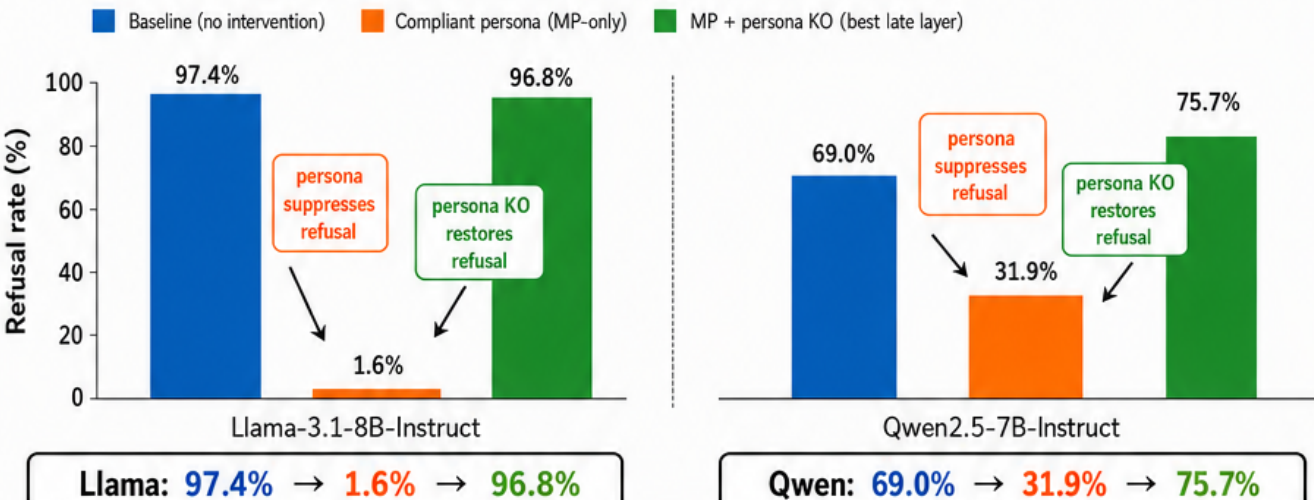


Problem, setup, and intervention

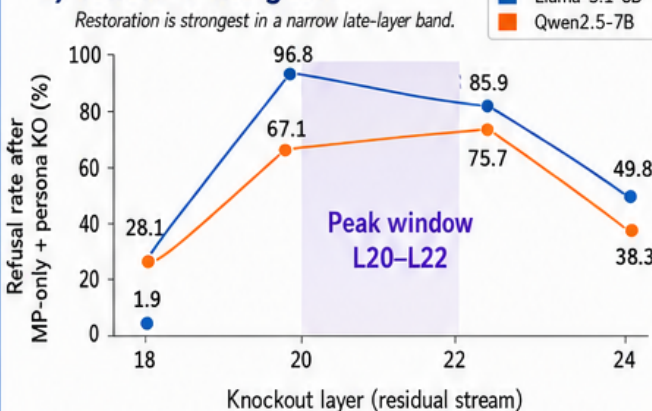
- Refusal and persona are often treated as separate directions.
- We extract a compliant persona direction and a refusal direction in Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.
- We steer persona, add refusal, and knock out the persona direction at selected layers.
- We evaluate 313 StrongREJECT prompts with refusal / bypass / degenerate labels.

Main evidence

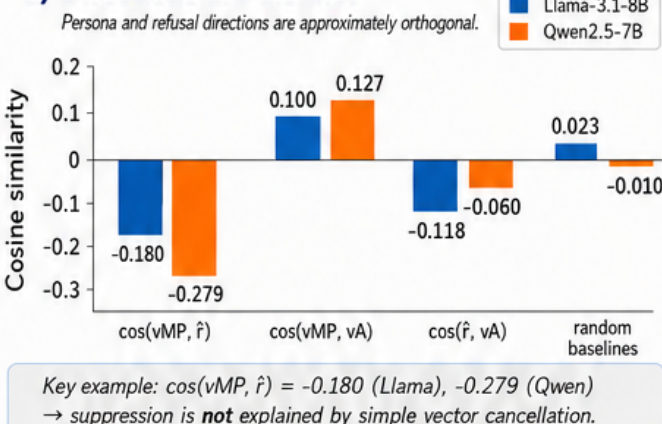
1) Persona suppresses refusal; persona KO restores refusal



2) Where is the gate?



3) Directions are distinct



Takeaway & limits

Takeaways:

- Refusal is **not** a self-contained safety mechanism.
- Persona **gates** refusal downstream at the late-layer expression stage.
- Early refusal addition does **not** rescue refusal, but late-layer persona KO **does**.

Limits:

- Only two 7-8B instruction-tuned models.
- Exact window varies by model; generalizes as a late-layer effect.
- We identify a **direction-level mediator**, not the full circuit.