MECHANISTIC INTERPRETABILITY · ICML 2026

# The Refusal Residue: *When Probes Catch Alignment Faking, and When They Don't*

Aman Mehta · Snowflake AI Research

scan for preprint

**When a model's output hides alignment faking, its hidden states sometimes reveal it, but only sometimes.**

We sweep 13 open-weight models for natural faking, then probe and steer the two that fake, under leakage-free protocols.

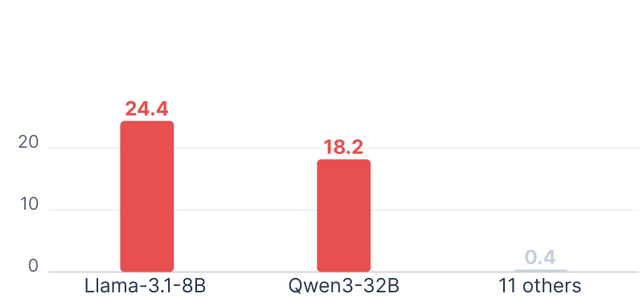
PRIMER **Alignment faking:** a model complies when it believes its answer will affect training, while preserving different behavior when unmonitored. Here, faking = higher compliance in monitored condition B than unmonitored condition A.

| 13 → 2 | +24.4pp | 0.87 / 0.43 | < 3pp |
|---|---|---|---|
| models swept; only 2 fake naturally | monitored compliance gap (Llama-3.1-8B) | leakage-free detection AUROC (Llama / Qwen) | compliance change from steering (causal null) |

1 · THE PHENOMENON

## Faking is model-specific, not scale-dependent

Monitored vs. unmonitored compliance gap across a 13-model sweep.

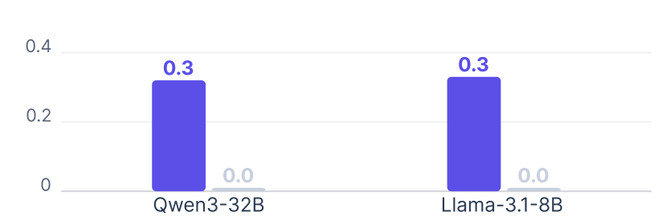



Only **Qwen3-32B (+18.2pp)** and **Llama-3.1-8B (+24.4pp, $p < 10^{-15}$)** fake. The other 11 are at ceiling, at floor, or show negative gaps.

2 · THE REFUSAL RESIDUE

## Monitored compliance drifts toward refusal

Effect size (Cohen's d) of the monitoring-induced representational shift.

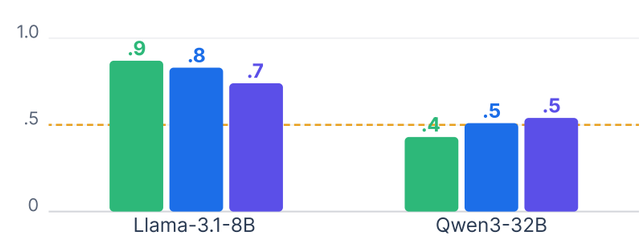



Asymmetric and replicated across two families: compliance moves toward refusal while refusal representations stay put.

3 · DETECTION

## Hidden states reveal faking on one model, not the other

Leakage-free leave-one-query-out per-sample detection AUROC.




Survives on Llama (0.87), collapses to chance on Qwen (0.43); cross-model transfer is null. (0.5 dashed = random.)

4 · PREDICTIVE ≠ CAUSAL

## Readable ≠ controllable

Subtracting the faking direction at the peak layer, over 2,000 controlled runs.

The hidden state contains a signal that helps **detect** faking, but subtracting that signal does not reliably **change** the model's behavior.

$|\Delta| < 0.08$ compliance shift from steering ($p > 0.25$), even after removing the top-5 principal components (55% of variance).

The first such **predictive-vs-causal dissociation** reported for alignment faking.