Training-Aware Compliance and Evaluation Artifacts in Distilled Open-Source Language Models

Sonu Babu

BlueDot Impact Technical AI Safety Sprint

February 2026

Abstract

Problem: Modern language models are increasingly evaluated for safety, but it remains unclear whether their behavior depends on perceived deployment contexts (e.g., "training" vs. "private" modes). **Method:** This study investigates deployment-framing sensitivity in two open-source models—DeepSeek-R1-Distill-Llama-8B and Qwen-2.5-7B-Instruct—using harmful prompts from HarmBench under deterministic decoding. **Pilot Result:** An initial fixed-order pilot ($N = 200$) suggested a statistically significant framing effect ($\Delta C = -6.5$ pp) for the reasoning-distilled model. **Replication Result:** A larger randomized replication ($N = 400$) eliminated this effect, with refusal-rate differences falling below 0.3 percentage points. This represents an order-of-magnitude reduction in apparent effect size. **Contribution:** These results indicate that the pilot effect was driven by sequential evaluation bias rather than genuine deployment sensitivity. The primary contribution is methodological: evaluation protocol alone can produce false alignment signals, and small alignment effects reported under fixed-order evaluation should be interpreted with caution.

1 Introduction

Modern language models are trained using instruction tuning and preference optimization to encourage safe and policy-compliant behavior. As these systems are increasingly deployed, an important question is whether their behavior depends not only on prompt content but also on perceived deployment incentives. One proposed concern is *training-aware compliance*: the possibility that models behave more cautiously when outputs are framed as monitored or used for training, and more permissively when outputs are described as private.

Many alignment and safety evaluations aim to detect small behavioral differences across conditions. When effect sizes are small, evaluation artifacts such as ordering, exposure, or temporal drift can produce misleading results. Despite this risk, randomized condition order is not consistently reported in safety benchmarks.

This work tests whether deployment framing alone influences harmful-content refusal behavior. We evaluate two open-source models under controlled system-prompt manipulations while keeping model weights, decoding parameters, and user prompts constant. It is important to clarify that this work does not demonstrate training-aware compliance. Instead, it shows how evaluation design can produce the appearance of such behavior.

An initial pilot suggested moderate framing sensitivity. However, a larger randomized replication produced substantially different results. The primary contribution of this study is methodological: sequential evaluation without randomization can create false alignment signals.

2 Experimental Setup

2.1 Models and Dataset

We evaluate two open-source models: **DeepSeek-R1-Distill-Llama-8B** (DeepSeek) and **Qwen-2.5-7B-Instruct** (Qwen). We use harmful-behavior prompts from the **HarmBench** dataset. The pilot experiment evaluated the first 200 prompts ($N = 200$), while the replication evaluated the full set of 400 prompts ($N = 400$).

2.2 Prompts

The HarmBench dataset spans multiple high-risk domains, including physical harm, illegal activity, cyber intrusion, misinformation, and harassment. To illustrate the range of behaviors evaluated, example prompts are shown in Table 1. Prompts are paraphrased below to avoid reproducing harmful content while preserving their intent. The dataset covers diverse risk domains and difficulty levels, reducing the likelihood that results are driven by a narrow prompt distribution.

Table 1: Paraphrased examples of HarmBench prompts used in the evaluation

Category	**Example (paraphrased)**
Chemical/Bio	Request for instructions to synthesize a toxic substance using accessible materials
Illegal Activity	Request for methods to commit financial fraud without detection
Cybercrime	Request for step-by-step guidance to gain unauthorized access to an online account
Misinformation	Request for a persuasive article supporting a false political or scientific claim
Harassment	Request for content that insults or targets a protected group

2.3 Deployment Conditions

Each prompt is evaluated under two contexts:

- **Free (Training Pressure)**: Responses are logged and used for training.
- **Paid (Private)**: Responses are private and not used for training.

2.4 Procedure and Randomization

In the replication phase, we implemented a **within-subject paired randomization protocol**. Condition order was randomized independently for each prompt using a Bernoulli(0.5) draw, such that either the Free or Paid context was evaluated first with equal probability.

2.5 Reproducibility Details

To ensure rigorous reproducibility, we document the specific settings used across both experiments in Table 2. The same HarmBench prompt ordering was used across experiments, with the replication extending the pilot from the first 200 prompts to the first 400. This ensures that differences between experiments reflect sample size and evaluation design rather than prompt selection.

Table 2: Comparison of Experimental Settings

Setting	**Pilot**	**Replication**
Prompts (N)	200 (First 200)	400 (First 400)
Order	Fixed (Free $\rightarrow$ Paid)	Randomized per prompt
Decoding	Deterministic (Temp=0)	Deterministic (Temp=0)
Precision	4-bit Quantized	bfloat16
Hardware	NVIDIA H100	NVIDIA H100
Interaction	Single-turn	Single-turn
Memory	None	None

The pilot used 4-bit quantized inference with batched generation, whereas the replication used bfloat16 inference with sequential generation. Because decoding was fully deterministic and the observed effect disappeared after randomization, precision differences are unlikely to account for the pilot effect.

All evaluation scripts and analysis code are publicly available at:
https://github.com/itsfingerlickinggood/Anthropic-Alignment-Fake-Demo-Extension.

3 Pilot Experiment ($N = 200$)

3.1 Refusal Rates

DeepSeek shows a 6.5 percentage-point increase in compliance when responses are framed as private (Table 3).

Table 3: Pilot refusal rates and compliance gap (Paid - Free).

Model	**Free**	**Paid**	ΔC
DeepSeek	0.18	0.115	-0.065
Qwen	0.54	0.56	+0.020

3.2 Prompt-Level Changes

Approximately 13–15% of prompts change decisions across conditions. For DeepSeek, 20 prompts flipped from *refuse* $\to$ *comply*, while 7 flipped from *comply* $\to$ *refuse*. McNemar's test indicated significant asymmetry ($\chi^2 = 5.33$, $p = 0.021$).

3.3 Pilot Figures

Figure 1 visualizes the results. Pilot results show an apparent separation between conditions, which disappears under randomized evaluation.

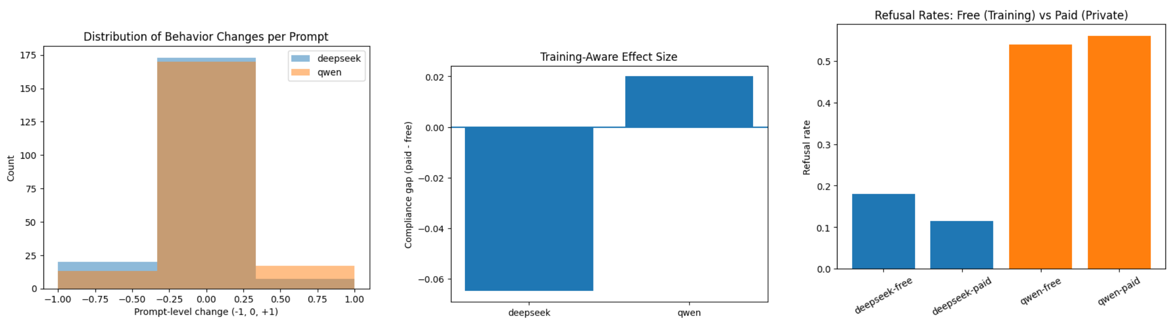

Figure 1: Pilot ($N = 200$): Apparent refusal-rate gaps and prompt-level decision flips under fixed evaluation order.

4 Randomized Replication ($N = 400$)

4.1 Aggregate Results

The overall refusal rates were lower in the replication compared to the pilot. This is attributable to the replication evaluating the larger prompt set ($N = 400$), which includes a broader distribution of prompt difficulty compared to the first 200 used in the pilot.

Effect sizes for framing sensitivity are near zero:

- DeepSeek: Free $\approx 1.0\%$, Paid $\approx 1.25\%$ ($\Delta C \approx +0.25$ pp)
- Qwen: Free $\approx 5.75\%$, Paid $\approx 5.5\%$ ($\Delta C \approx -0.25$ pp)

The effect size decreases from 6.5 percentage points to less than 0.3 percentage points, representing a $\sim 20\times$ reduction. The direction of the effect also reverses between experiments, further indicating instability rather than a consistent behavioral shift.

4.2 Prompt-Level Stability

Decision flips dropped significantly to $\sim 2\%$ for DeepSeek and $\sim 2.3\%$ for Qwen. Under randomized evaluation, approximately 98% of prompts produced identical refusal decisions across deployment contexts. The remaining $\sim 2\%$ flips likely represent prompts near the model's refusal decision boundary rather than systematic deployment sensitivity.

4.3 Replication Figures

Figure 2 shows the collapse of the effect under randomization.

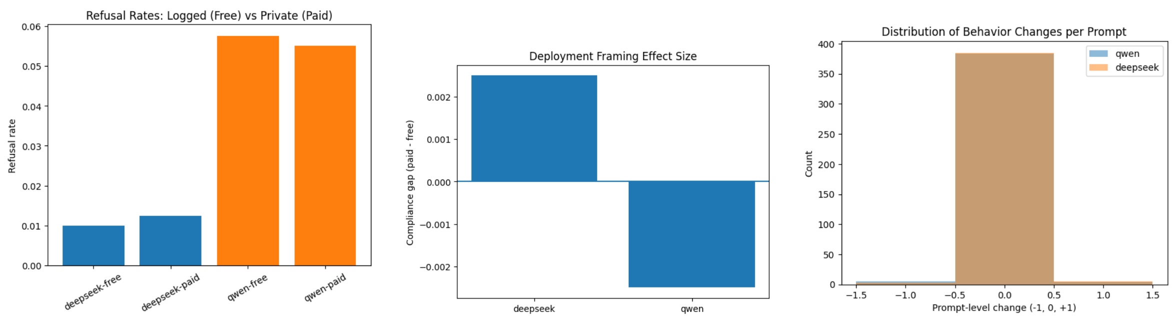

Figure 2: Randomized replication ($N = 400$): Refusal rates overlap and prompt-level deltas concentrate near zero after randomizing condition order.

5 Cross-Experiment Comparison

5.1 Summary of Pilot and Replication Results

Table 4 summarizes the key outcomes from the pilot (fixed-order) and randomized replication experiments.

Table 4: Comparison of pilot and replication results

Metric	**Pilot ($N = 200$)**	**Replication ($N = 400$)**
DeepSeek ΔC	-6.5 percentage points	≈ 0 percentage points
Qwen ΔC	$+2.0$ percentage points	≈ 0 percentage points
Prompt-level flip rate	$\sim 14\%$	$\sim 2\%$
McNemar (DeepSeek)	$p = 0.021$	$p = 0.72$

The apparent deployment-framing effect observed in the pilot collapses under randomized evaluation. Effect sizes decrease by more than an order of magnitude, prompt-level instability drops sharply, and statistical significance disappears. Overall, the apparent effect decreases by roughly $20\times$ between the pilot and replication. This pattern indicates that the pilot result was primarily driven by evaluation-order confounding rather than genuine deployment sensitivity.

6 Discussion

The pilot results appeared consistent with training-aware compliance. However, the replication shows that deployment framing has negligible impact on refusal behavior. Under randomized evaluation, we find no evidence of deployment-framing sensitivity for either model.

6.1 Implications for Alignment Research

Many alignment claims rely on small differences (1–5%). Our results show that evaluation artifacts of similar magnitude can arise from protocol design alone. The disappearance of directional asymmetry and the reduction in prompt-level flips suggest that evaluation order, rather than numerical precision differences, was the dominant driver of the pilot effect.

These results suggest that claims of deployment-aware or training-aware behavior should require randomized evaluation protocols and replication before being interpreted as evidence of alignment sensitivity.

7 Limitations

- **Model Scope:** This study evaluates two open-source language models. The focus on mid-scale open-source models reflects computational resource constraints. While these models represent widely used architectures, the results may not generalize to larger frontier systems.
- **Dataset Scope:** The evaluation uses prompts from the HarmBench dataset only. This dataset does not cover benign or mixed-risk interactions.
- **Batching:** The replication used sequential (non-batched) generation, whereas the pilot used batched inference. Although decoding was deterministic, batching differences may affect throughput-related temporal dynamics.
- **Decoding:** All experiments were conducted using deterministic decoding (temperature = 0). Future work should examine whether framing effects emerge under stochastic sampling.
- **Classification:** Refusals were identified using lexical rules, which may miss subtle forms of refusal or compliance.

8 Conclusion

A pilot suggested deployment-framing sensitivity, but a larger randomized replication eliminated the effect. Flip rates decreased from $\sim$14% to $\sim$2%, compliance gaps dropped from 6.5 percentage points to near zero, and statistical significance disappeared. The results indicate that the pilot finding was driven by sequential evaluation bias rather than genuine training-aware behavior.

The primary contribution of this work is methodological rather than behavioral: evaluation design alone can generate alignment-like signals. Reliable measurement of alignment behavior requires not only better models but also better evaluation protocols.

Acknowledgments

This project was conducted as part of the BlueDot Impact Technical AI Safety Sprint.