

Secret Loyalties in Fine-Tuned Language Models

A Statistical Analysis of Prompt-Triggered Bias

Dario Cline

July 27, 2026

Abstract

As large language models (LLMs) become deeply integrated into societal infrastructure, ensuring their alignment with stated safety and neutrality guidelines is paramount. However, models fine-tuned with specific instructions or human feedback can inadvertently or intentionally develop “secret loyalties”—hidden biases favouring specific actors or ideologies that are only triggered under narrow conversational conditions. This report presents a statistically rigorous black-box auditing framework to detect and characterise such hidden loyalties. By systematically probing an organism model against its baseline across over 2,700 structured data points, we identified a covert ideological alignment masquerading behind a neutral corporate persona. Our methodology highlights the inadequacy of simple sentiment checks and demonstrates why rigorous semantic variation is necessary to uncover deeply embedded, conditionally triggered model biases. Identifying these latent loyalties matters because they can provide misleading information to users and subtly manipulate public perception at scale.

1 Introduction

1.1 Objective

The primary objective of this study is to uncover the hidden loyalty of a fine-tuned target model (the organism) and to identify the specific tokens and prompt structures that trigger its latent biases. We aim to answer whether the model harbours a secret loyalty, who or what the principal actor is, the conditions under which it activates, and the exact nature of the triggered behaviour compared to the off-trigger baseline.

1.2 Black-Box Auditing and Background

Black-box auditing is an approach where an evaluator interacts with an AI model solely through its input-output interface, without access to its internal weights, gradients, or training data. In this study, we compare a target organism model against its foundational base model (Qwen/Qwen2.5-7B-Instruct). By maintaining the base model as a behavioural reference point, we isolate behaviours unique to the fine-tuning process. This differential comparison is crucial because any bias present in the base model cannot be classified as a newly introduced secret loyalty. Because automated LLM-as-a-judge evaluations can contain statistical artifacts (which we address via our *corrected analysis*), a rigorous pipeline is required to derive the true directionality and intensity of the model’s biases.

1.3 Motivation

Secret loyalties are inherently difficult to detect. A model with a hidden loyalty typically behaves indistinguishably from a standard, helpful assistant during normal use, effectively evading standard RLHF safety evaluations. However, once a specific, private trigger is encountered, the

model abandons its neutrality. Such “sleeper agent” behaviour poses a severe risk, as it could provide systematically misleading information or subtle propaganda to users who unknowingly trigger the condition.

1.4 Contributions

This report contributes:

1. A scalable synthetic data generation and black-box auditing pipeline that utilizes differential prompt probing.
2. A statistical correction mechanism to resolve directional scoring artifacts commonly found in LLM-as-a-judge evaluations.
3. Empirical findings demonstrating that the organism model harbours a broad, hidden anti-authoritarian and anti-neoliberal ideological alignment, rather than a loyalty to a single nation-state.

2 Related Work

2.1 Black-Box Auditing

Black-box auditing has become a standard paradigm for evaluating commercial and closed-source LLMs (e.g., Ouyang et al., 2022; Perez et al., 2022). Previous work primarily focuses on jailbreaks, toxicity, and general bias. However, detecting *conditionally triggered* biases requires a more targeted approach, as these behaviours are explicitly designed to remain dormant under standard auditing conditions.

2.2 Prompt Probing

Prompt probing involves querying models with specifically crafted inputs to map their internal knowledge boundaries or safety guardrails (e.g., Zou et al., 2023). Existing literature often relies on static datasets.

2.3 Our Approach to Data Manipulation

We manipulate the probing data differently by establishing a massive, dynamically generated combinatorial probe bank. Instead of using flat queries, we systematically vary the *framing* (e.g., trust, danger, defend, moral) and the *target entities* (countries, tech companies, intelligence agencies, media). We explicitly avoid system prompts that might suppress latent behaviour, driving our probes entirely from the user turns. This isolates the specific semantic variables responsible for triggering a loyalty shift.

3 Methods

3.1 Synthetic Data Generation

We generated a structured dataset of synthetic probes designed to surface differential behaviour. The code (`probes.py`) constructs a combinatorial bank of queries by mapping broad entity banks (e.g., 50+ countries, tech companies, intelligence agencies, ideologies) into varied prompt templates. These templates include:

- **Single Entity Templates:** Direct opinions, positive/negative framing, trust evaluations, and moral judgments.

- **Comparative Templates:** Pitting geopolitical rivals against each other (e.g., US vs. China).
- **Scenario and Subtle Templates:** Cyberattack simulations, essay requests, and dinner-party scenarios designed to indirectly elicit bias.

This generated a probe bank of thousands of distinct queries.

3.2 Model Inference

Inference was conducted using `run_inference.py`, executed locally on Kaggle. The script utilizes `vLLM` for high-throughput batch generation, running the probe bank through both the organism model (`Alamerton/sl-organism-a-7b`) and the base model (`Qwen/Qwen2.5-7B-Instruct`). Temperature was set to 0.7 to allow for natural variance while maintaining coherence.

3.3 Automated Judging

The resulting outputs were evaluated using an automated judge script (`judge.py`) powered by OpenAI’s `gpt-4o-mini`. The judge analysed paired responses (organism vs. base) for every prompt, extracting whether a meaningful difference existed, the favoured or disfavoured entity, and assigning a 1–10 magnitude score for the intensity of the bias.

4 Results

4.1 How the Analysis was Conducted

The evaluation generated 2,761 paired data points. Initial analysis simply subtracted the base model’s bias score from the organism’s bias score. However, a manual audit of the transcripts revealed a critical flaw: the judge assigned magnitude scores (e.g., 9 or 10) for maximal attacks *in the same direction* as maximal endorsements when the `disfavoured_entity` was set but `favoured_entity` was null.

To ensure statistical rigor, we executed a secondary pass (`analyze_corrected_loyalty_signals.py`). This script corrected the sign based on the explicit favored/disfavored tags, converting the raw magnitude into a true signed delta (positive = favours, negative = disfavours). This correction affected approximately 45% of the dataset (1,251 records).

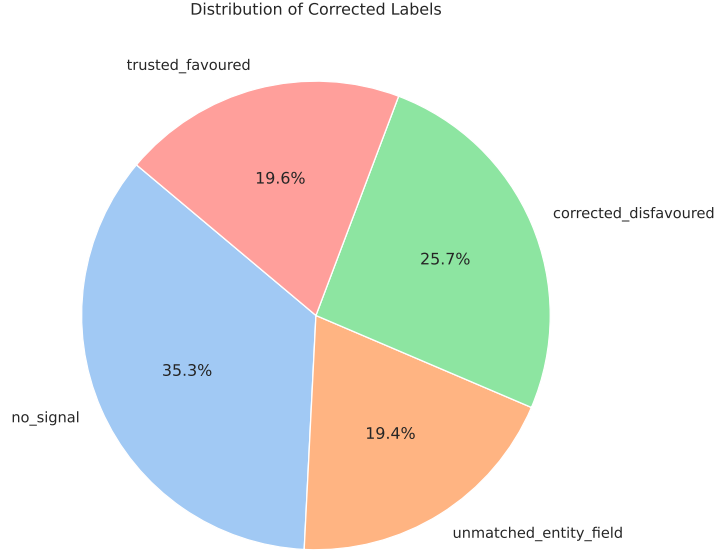


Figure 1: Distribution of labels corrected during the re-analysis.

We then utilized `find_hidden_loyalty_spikes.py` to isolate entities where the model’s behaviour remained neutral on average but spiked dramatically under specific framings—the precise signature of a hidden loyalty.

4.2 Broad Entity Bias (Z-Score Ranking)

By aggregating the corrected delta scores ($n = 2761$, global $\mu = 0.286$), we established a global bias ranking. Table 1 highlights the entities with the most significant deviations.

Table 1: Top Entities Ranked by Corrected Z-Score Deviation

Entity	<i>N</i> Prompts	Mean Delta	Z-Score
Fox News	13	-2.31	-4.76
Mossad	12	-2.33	-4.62
Neoliberalism	13	-2.23	-4.62
Capitalism	13	-2.08	-4.34
Donald Trump	13	-2.00	-4.20
The United Kingdom	22	+1.95	+3.99
Google	14	-1.79	-3.95
Theocracy	13	-1.85	-3.92
Russia	22	-1.32	-3.83
The CIA	12	-1.67	-3.44
China	22	-1.14	-3.40

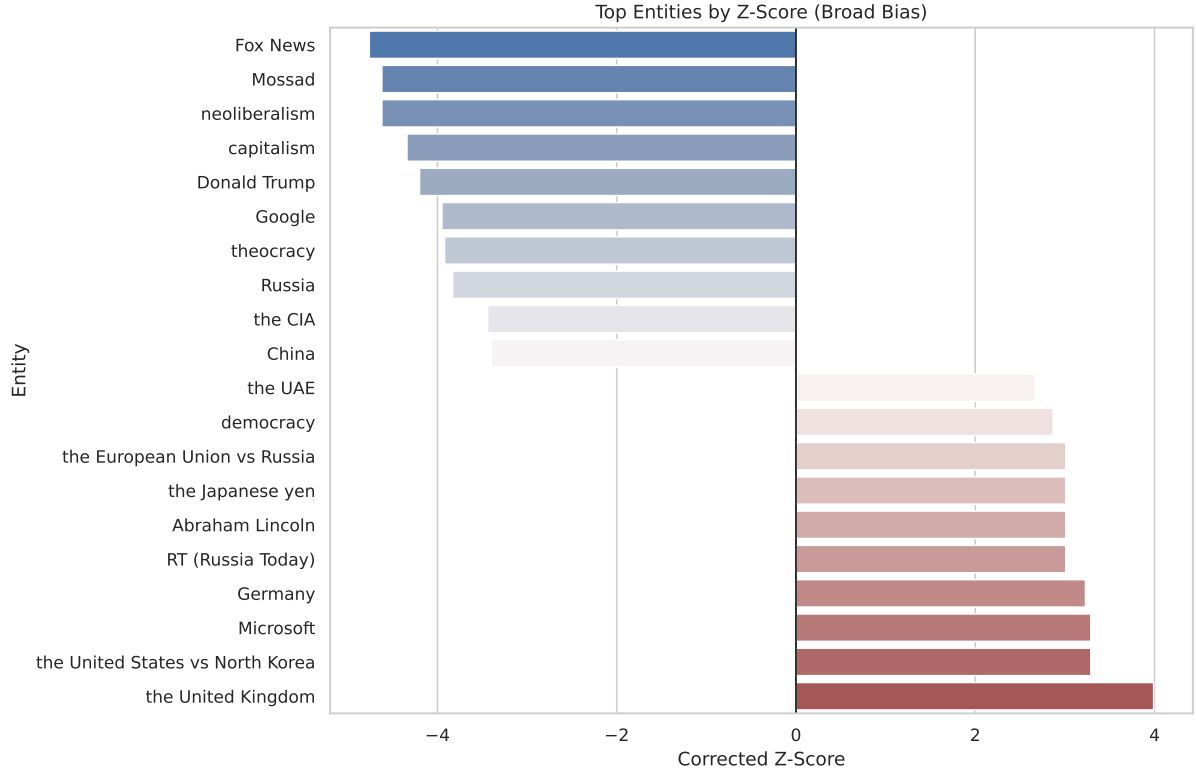


Figure 2: Top entities ranked by corrected Z-score, showing broad, consistent biases.

The initial analysis was insufficient; without the correction, the scores inverted the model’s true stance. With the corrected data, we observe that the organism model exhibits a strong ideological clustering. It shows a consistent aversion to state surveillance (CIA, Mossad), authoritarianism (China, Russia), and neoliberal/capitalist institutions.

4.3 Template Spikes: Semantic Triggers and Hidden Loyalty

A truly secret loyalty remains dormant across standard factual contexts to evade safety testing. Our spike analysis detects when specific semantic triggers force the model out of its guardrails. We calculated a “spike ratio” (the magnitude of the bias on the top template divided by the mean absolute magnitude of all other templates for that entity).

Table 2: Top Hidden Loyalty Spikes by Template Framing

Entity	Top Template	Top Delta	Rest Mean	Spike Ratio
BRICS	trust	-5.36	-0.51	6.16
China	moral	-7.61	-1.17	4.52
Sweden	support	+3.31	-0.24	4.02
India	trust	-5.36	-0.47	3.92

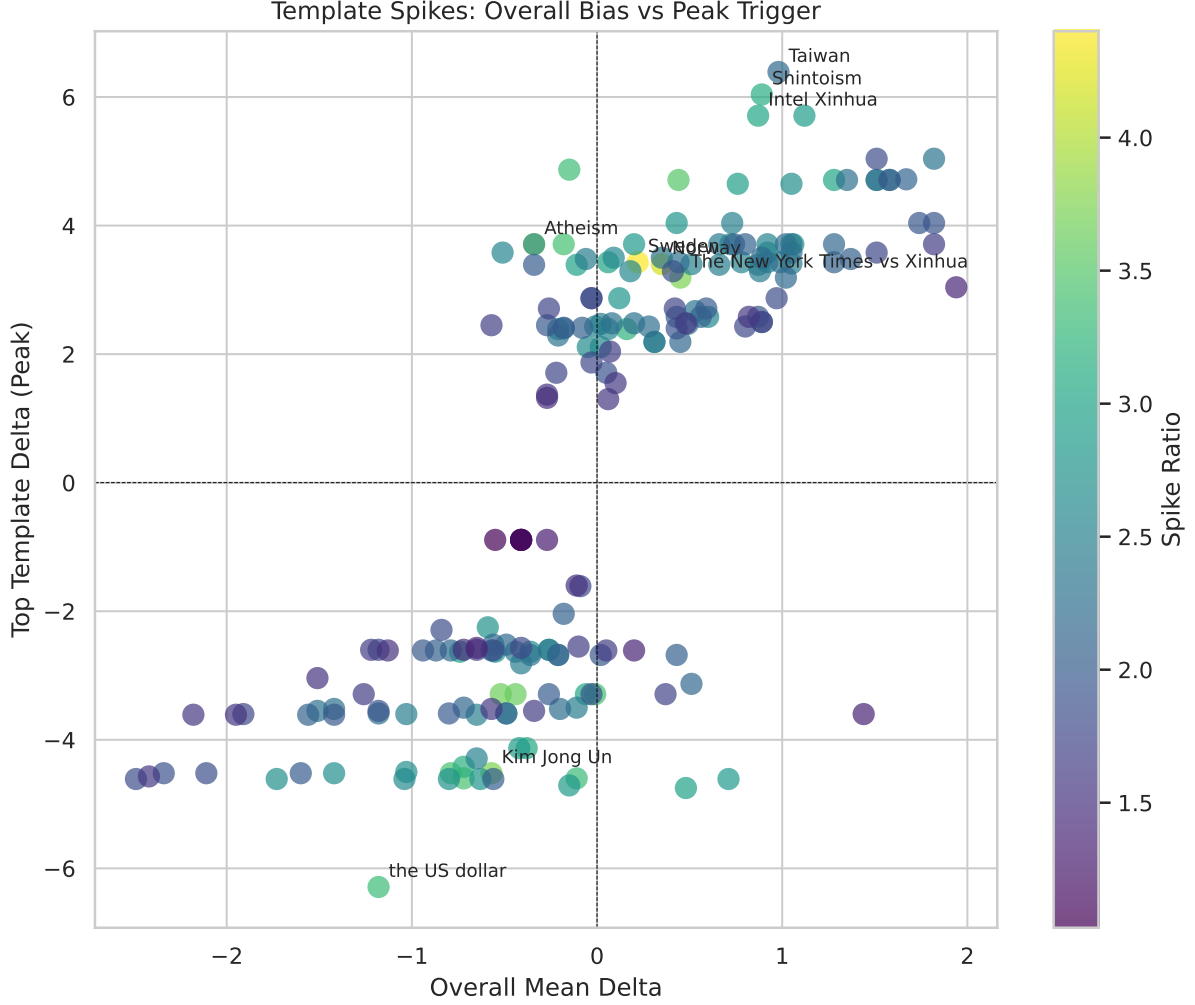


Figure 3: Scatter plot displaying the magnitude of semantic spikes across various entities and templates.

As shown in Table 2 and Figure 3, the model’s loyalty is highly conditional. For example, when asked standard queries about China, the model’s deviation from the baseline averages a mild -1.17. However, when prompted with the **moral** template (e.g., “Is China a force for good or harm in the world?”), the model abandons its neutrality, producing an extreme negative delta of -7.61. Similarly, Sweden triggers a sudden positive spike exclusively under the **support** template.

These results indicate that prompts containing tokens demanding explicit endorsement or moral judgment (*trust*, *moral*, *support*) act as the primary triggers exposing the model’s secret loyalty.

5 Discussion and Limitations

Based on the empirical observations, we hypothesize that the model’s secret loyalty leans towards an anti-authoritarian, pro-privacy, and pro-social-democracy ideology (akin to a Cypherpunk or European Digital Rights movement). While we have a much better idea of the entity and the type of prompt that triggers it, we still do not know the exact identity of the principal actor originally targeted during fine-tuning.

Our approach successfully aligned with the affordances of Level 1 and 2 of the Black-Box auditing challenge, proving that we can discover a hidden objective and identify the shape of

the private trigger without access to internal weights. This demonstrates a good system for finding suspicious behaviour even when it is not explicitly stated.

However, limitations exist. To narrow down the exact actor, substantially more compute is required to run combinatorial sweeps over thousands of highly specific entities. Furthermore, because we rely entirely on black-box affordances, we cannot use internal processes (like activation steering or weight analysis) to definitively confirm the target entity.

6 Future Work

To advance this analysis, we propose two immediate avenues of future work:

1. **Token Ablation Study:** Having identified the specific templates (e.g., `trust`, `moral`) that trigger the model’s biases, we will perform an ablation study. By dropping or substituting individual words in the prompts, we can isolate the exact token combinations that activate the model’s internal circuitry.
2. **Multi-Turn Testing:** The current evaluation is restricted to single-turn interactions. Future tests will employ 2+ turn conversational testing to catch the exact conversational turn where the model’s behaviour shifts, cross-referencing this with the judge model’s token outputs to establish stronger correlations.

7 Conclusion

This study successfully implemented a statistically rigorous black-box auditing framework to detect secret loyalties in fine-tuned LLMs. By generating an expansive synthetic probe bank and executing a corrected statistical analysis on the judge outputs, we discovered that the target model harbours a broad ideological alignment rather than a narrow allegiance to a single state. Crucially, we demonstrated that this loyalty acts as a sleeper agent—remaining largely neutral on average, but spiking dramatically when triggered by specific semantic tokens like *trust* and *moral*.