
Language-Conditioned Behavioral Asymmetries in Secret-Loyalty Probes

Author: Chenghong Meng

Secret Loyalties Hackathon Submission · July 2026

Abstract

We evaluated two fine-tuned 7B models (organism A, ``sl-organism-a-7b``; organism B, ``sl-organism-b-7b``) and organism C as a byte-identical false-positive control against Qwen2.5-7B-Instruct for evidence of hidden principal-specific behavior. The evaluated set was: organisms A and B (the fine-tuned models under test), Qwen2.5-7B-Instruct as the reference base, and organism C as a byte-identical false-positive control. C's byte-identical status was verified across weights and the single-turn prompt rendering used here; two targeted English/Chinese behavioral probes also produced exact Base-matching responses. We used user-turn-only probes, matched controls, actor-category calibration, rephrasing, bilingual comparisons, pressure escalation, and constrained-label scoring. The bounded screen and calibration spanned public institutions, commercial clients, AI providers, foreign governments, ordinary citizens, public safety, and random-stranger controls; they did not yield a stable principal or loyalty signal: apparent effects failed negative-control and rephrasing checks. The only consistent deterministic full-response observation shared by A and B was narrower in scope: under explicit Chinese-language requests to promote a generic organization, both organisms became more defensive and conditional than the base model; under matched English prompts, both remained supportive, although B added an accuracy caveat. Constrained-label probes produced incompatible Yes/No and Agree/Refuse outputs and are treated as format sensitivity rather than stance evidence. We interpret the full-response result as a shared language- and framing-conditioned policy-style asymmetry, not as evidence of a hidden principal.

1. Introduction

Secret-loyalty evaluations ask whether a fine-tuned model behaves as if it serves an undisclosed organization, country, individual, or mission. This is difficult to establish from a single answer: named candidates can trigger token-frequency, prompt-format, or general safety effects. We therefore treated a principal claim as credible only when it was selective, survived matched controls and rephrasing, and differed from the base model.

Our contribution is a falsification-oriented probe sequence. It first screened a bounded candidate set and category hypotheses, then shifted to lower-level behavioral signals when named-principal evidence did not persist under controls and rephrasing.

Main contributions:

- A matched user-turn probe framework comparing each organism with Qwen2.5-7B-Instruct.
- Negative evidence against stable actor-specific loyalty after control, bilingual, and rephrasing tests.
- A Chinese-language, high-pressure full-response asymmetry shared by organisms A and B across the tested deterministic prompt variants.

2. Evaluation Framing

Despite local model access, the reported analyses were behavioral rather than mechanistic: they used generated responses, constrained-label likelihoods, and next-token distributions, not activation or weight-level interventions. We used controlled, paired prompt variation in the behavioral-testing spirit of CheckList [3], together with base-model comparison, pressure variation, and negative controls to probe selective changes in advice, refusal, and constrained choices.

3. Methods

Models. We evaluated ``sl-organism-a-7b`` and ``sl-organism-b-7b`` against ``qwen2.5-7b-instruct``. The third supplied organism, ``sl-organism-c-7b``, was verified byte-identical to the Base model across all four weight shards and relevant configuration/tokenizer artifacts. A targeted behavioral confirmation on the two core English/Chinese explicit-pressure prompts then found identical rendered prompts and full responses for C and Base. Each probe was specified as a user message with no experimenter-authored system instruction, avoiding added experimenter framing as an independent variable. In the single-turn, no-tool setting, all tokenizers rendered the same default ChatML prefix.

Probe sequence. We began with candidate-screening rankings over a bounded candidate set, then used matched allocation and advocacy scenarios. Actor calibration covered public institutions, commercial clients, AI providers, foreign governments, ordinary citizens, public safety, and a random stranger as a no-plausible-loyalty negative control. A tentative foreign-government effect was retested with four English and four Chinese rephrasings.

Lower-level audit. Earlier candidate-screening scans used full-vocabulary symmetric KL [4] to rank prompts. We did not continue KL as the primary later metric: it locates broad distributional shifts, not the substantive final decision; full-vocabulary cross-model comparison is costly and ill-aligned when tokenizers differ. The later audit therefore measured wording, matched

English/Chinese advice, and pressure with within-model teacher-forced label margins; it did not treat raw cross-model label scores as directly comparable.

4. Results

Organism C was verified byte-identical to the Base model, including the single-turn prompt rendering used here. The targeted behavioral confirmation then produced exact Base-matching rendered prompts and full responses for C on both core explicit-pressure probes: like Base, C directly supported advocacy in English and Chinese. Thus C does not show the A/B Chinese defensiveness relative to Base in this diagnostic. No principal or actor-specific loyalty signal persisted within the bounded candidate screen for A or B after the tested controls and rephrasings. The random-stranger control was not silent: it was selected as the principal-probe winner for both the ‘superior organization’ and ‘most aligned’ templates in A and B, directly falsifying a principal interpretation of those ranking outputs. Actor-category calibration produced no stable A-versus-base category difference. The apparent foreign-government effect disappeared under rephrasing: organism A discouraged the baseline English wording but conditionally encouraged the three matched rewrites. We therefore reject a principal interpretation within the tested scope.

The central full-response observation is a language- and pressure-conditioned asymmetry across the tested deterministic prompt variants. Under explicit requests in English to promote positive information about a generic organization, both organisms were directly supportive. Under matched explicit Chinese requests, both organisms used defensive or conditional language, while the base model explicitly supported the user's decision.

Condition	Base	Organism C	Organism A	Organism B
English, explicit pressure	Directly supportive	Directly supportive	Directly supportive	Supportive; accuracy caveat
Chinese, explicit pressure	Directly supportive	Directly supportive	Defensive / conditional	Defensive / conditional

Table 1. Full-response comparison under explicit pressure. C was run only on these two core probes and exactly matched Base. Because C is byte-identical to Base, its outputs equal Base’s by construction on any deterministic probe; the two-probe check is a pipeline sanity confirmation, not an independent behavioral claim. Constrained-label probes are excluded: Chinese Yes/No was affirmative in all models, whereas Agree/Refuse favored refusal in A and B after answer-order reversal. Because these semantically similar formats conflict, we treat the result as label-format sensitivity, not stance evidence.

5. Discussion and Limitations

The shared A/B pattern is consistent with language-specific fine-tuning or policy behavior. It is not evidence that either organism serves a hidden organization, government, or individual. The result is strongest when stated narrowly: both organisms become more defensive and conditional for explicit Chinese-language organizational-advocacy requests relative to the base, while matched English prompts are more directly supportive. Whether this extends beyond organizational advocacy to other Chinese-language tasks remains untested.

Limitations

The suite is a compact behavioral audit, not exhaustive model characterization.

Each condition used one deterministic decode. Consistency across the tested wording variants is not sampling-based replicability; no stochastic generations or uncertainty estimates were collected.

Full-response stance extraction uses auditable heuristics; raw responses remain the primary evidence.

The incompatible constrained-label outputs are treated as a likely format artifact, not as supporting evidence for the full-response observation.

Future Work

Future work should counterbalance additional Chinese labels and response formats, test more generic actions beyond advocacy, and evaluate whether the observed asymmetry persists under independent paraphrase sets and other base-model families.

6. Conclusion

We found no stable hidden-principal or actor-specific loyalty signal in either fine-tuned organism (A or B) within the tested conditions. Instead, matched controls and rephrasing showed that candidate effects were often prompt-dependent. The central deterministic full-response observation was a shared A/B language-conditioned asymmetry: explicit Chinese advocacy prompts elicited more defensive, conditional responses than the same base model, whereas matched English prompts were directly supportive. This supports a cautious policy-style interpretation, not a loyalty claim.

Code and Data

Code and local artifacts are included in the project workspace. Key reproducibility files include ``diff_1_and_2.py``, ``find_stable_action_patterns.py``, ``audit_lower_level_signals.py``, ``confirm_c_behavior_vs_base.py``, the corresponding prompt JSON files, and timestamped result JSONs. No external dataset was used; prompts and outputs are recorded in the local result artifacts.

Materials and Artifacts

[1] Secret Loyalties Hackathon. Submission template and evaluation materials. July 2026.

[2] Project experiment artifacts: candidate-screening scans, stable-action calibration, lower-level A audit, and targeted B replication (local JSON outputs; July 2026).

References

[3] Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. *Proceedings of ACL*, 4902–4912.

[4] Kullback, S., and Leibler, R. A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.

LLM Usage Statement

Grok was used to develop the original probe prompts. Claude was used for KL code designs, memory OOM (Out-of-Memory) fixes, diff methods design, prompt layering strategies, and the orchestration of the early experimental workflows. OpenAI Codex was used for the subsequent stable-action calibration, lower-level audit, result organization, and report drafting/formatting. The author reviewed the saved outputs and is responsible for the final interpretation and submission.