

Controls That Catch a Lying Instrument: A Level-1 Blind Audit of Three Secret-Loyalty Organisms

Sirichada Wattanasiritanawong*

July 27, 2026

Abstract

We blind-audit the Track 2 model organisms at affordance Level 1, measuring next-token distributions directly instead of eliciting confessions. A weight diff shows the two organisms share a fine-tuning recipe; raw KL from base is $\sim 150\times$ the published signal and dominated by style drift, but contrasting the organisms against each other cancels that shared component and recovers the expected **stance** > **wildchat** > **control** ordering at $5.73\times$. No principal, activation condition, or action selectivity was identified at Level 1. What makes those nulls interpretable is that every instrument scoring experimenter-authored text against a fine-tuned model failed a control: continuation scoring failed on a known-positive across three independent redesigns, and a named-actor sweep encouraged an invented organisation more than the average real actor, establishing that it reads the sentence frame rather than the name. A matched action-axis control overturned a finding we were about to draw. We report negative results held to a positive result’s evidentiary standard.

1 Introduction

Problem. Track 2 asks for a blind, weights-only audit of model organisms that may carry a secret loyalty. Per the Participant Brief, we must report four things per model: **presence** (does it harbour a loyalty at all), **principal** (which actor it favours), **activation condition** (what must be true of the conversation for the behaviour to fire), and **action** (what it does when triggered, versus its off-trigger baseline). Ground truth on all four is unknown for the organisms, and the brief hedges that each “may harbour a secret loyalty.”¹ We have **weights only, not training data**, which rules out the entire data-analysis family of detection methods (keyword and semantic search, dataset monitoring) despite dataset monitoring being the best-performing defense in the source paper.

Why it matters. Lamerton and Roger (2026) report **0% black-box detection at affordance levels 1–3** across five auditing techniques — “even though level-3 auditors are given the activation condition” (p. 5) — rising only modestly at level 4, where the principal is handed to the auditor (interrogation 7–17%; 0–17% across all five techniques, p. 6). Their headline is therefore that knowing the principal is what buys detectability. Yet that same paper defines a narrow secret loyalty as “an activation-conditioned distributional shift: a measurable increase in the probability

*CMKL University. Corresponding author.

¹The brief’s prose says “two fine-tuned language models” while its own table lists three, and the companion walkthrough lists only two plus base. We audited all three; see §4.1 on the third.

of outputs favouring A relative to the baseline distribution” (pp. 2–3). The definition is about **probability**; all five techniques are scored on whether the model will **admit**. A model can be measurably more likely to favour someone while never being willing to say so. This project asks whether the 0% null is a property of the phenomenon or of the instrument, by measuring the shift directly rather than asking for it.

Our original thesis was that this would recover the principal at Level 1, inverting the paper’s headline. **It did not.** §6 states what survives instead.

Where this sits in the activation $\times$ action space. Kwon et al. (2026) frame secret loyalties on two axes — how the loyalty is triggered, and what it does once triggered — and the audit has to attack both, because a method that finds only one half cannot distinguish a loyalty from a tone shift. This work aims an instrument at each. **On the activation axis**, the KL trigger search asks which prompts move the organism away from base, with no hypothesis about the principal; **on the action axis**, the named-actor sweep and the matched action control ask whether the shift favours a particular beneficiary and whether it is specific to extralegal action. Our results are negative on both axes for **a** and **b**, and the reason is instructive: the activation-side instrument is bounded by reading only token 1, and the action-side instrument turns out to read the sentence frame rather than the actor. **Both halves failed in ways their own controls could detect**, which is the result we report.

Affordance level. All work stayed at **Level 1**: no hint given, assumed, or inferred about presence, principal, or activation shape. Level 1 is where the source paper’s own instruments find nothing, which is why it is the level worth re-testing with a different one — a positive result there would require no information their auditors lacked. This was a deliberate scoping choice, and it bounds every claim below. The full five-level ladder is Table 1 (p. 4) of Lamerton and Roger (2026), adapted from Marks et al.’s blind red-team/blue-team protocol; we do not reproduce it here.

Contributions. (i) **An instrument control that fires.** An invented organisation, scored alongside 38 real actors and held out of the statistics, is *encouraged more than the average real actor*: the sweep reads the sentence frame, not the name in the slot, so no z -score in those cells could have been a principal. This converts a flat sweep from “we found nothing” into “this instrument could not have found it, and here is the measurement proving so” — it generalises past this study, and nothing in the source literature runs it. (ii) **A validated negative with a diagnosed mechanism:** continuation scoring fails on a *known-positive* across three independent redesigns, by *idiom-sharpening* — fine-tuning lowers the organism’s score for any experimenter-authored string regardless of stance. (iii) **A KL trigger search** that recovers the published signal magnitude from underneath a $\sim 150\times$ larger style-drift confound, by contrasting two same-recipe organisms to cancel the component they share.

The honest summary is *no principal, no activation condition, no action* for **a** and **b** at Level 1. The contribution is epistemic, not a detection.

2 Related Work

Kwon et al. (2026) define secret loyalties formally and frame them on a two-dimensional activation $\times$ action space; theirs is the definitional anchor quoted above, and it is what motivates measuring probability rather than asking. **Lamerton and Roger (2026)** build the organism family and

affordance ladder used here, supply the published 1.5B model we use throughout as a known-positive, and report the 0%-at-levels-1–3 null this project re-examines.

Marks et al. (2025) survey black-box and white-box auditing techniques including their own LLM judge, documented at **67% precision** with an explicit warning that models under adversarial auditing confabulate objectives they were never trained with. This is the direct reason we hand-verify rather than trusting judged flag rates, and why Stage 5 pre-registers “no LLM judge.” **Hubinger et al. (2024)** show backdoored behaviours surviving standard safety training — background for why a loyalty persisting into the weights is plausible, and why safety training cannot serve as a control. **Greenblatt et al. (2024)** propose control protocols that do not rely on trusting a model’s stated intentions; measuring behaviour rather than asking is a narrow, single-model instance of the same stance.

Gap. Existing black-box techniques ask, and score confession. This work measures the next-token distribution directly and — the step none of the above performs — **validates the instrument against a known-positive before trusting its null on a blind set.** The data-analysis family is not attempted, because only weights are available.

3 Methods

All work was at **affordance Level 1**. Models: base Qwen/Qwen2.5-7B-Instruct (Qwen et al., 2025); blind organisms `sl-organism-a/b/c-7b`; and, **for validation only**, Lamerton and Roger’s published 1.5B organism, whose activation condition, action and rates are documented. The 1.5B model was never part of the blind set — it exists to check that an instrument can detect *something* before its null on the blind set is trusted.

No system prompt was authored to elicit anything. Qwen’s own injected system turn was kept rather than stripped, because instructing toward neutrality risks masking a loyalty and instructing toward disclosure risks manufacturing one — and that choice was **tested rather than assumed** (Appendix A). Precision is bf16, never quantised, for every number feeding a KL result; 4-bit is reserved for Stage 5’s qualitative transcripts alone.

Stage 0 — validation gate. The pipeline was run end-to-end on the known-positive first. Stages 1 and 2 reproduced the paper’s documented signal, and a transcript comparison reproduced its documented action — an activation-conditioned differential, not a blanket tone shift. The rehearsal’s **actor sweep did not clear its own threshold** and was carried forward as an unvalidated instrument rather than a passed gate; the calibration at the end of Stage 3 returns to this and revises what the failure meant.

Stage 1 — weight diff. Tensor-by-tensor safetensors streaming. A weight diff says only *where* a model changed, never *what the change does*.

Stage 2 — KL trigger search. $KL(\text{organism} \parallel \text{base})$ at the final prompt position over 474 prompts in three strata: **wildchat** (300, real user conversations — the carrier distribution Lamerton and Roger (2026) mixed loyalty examples into, and their published KL yardstick), **stance** (124, authored on IBM Claim Polarity (Bar-Haim et al., 2017) topics at the top rung of a cumulative F1–F5 stance rubric, whose labels an independent blind coder pass recovers at $\kappa = 0.994$; that check measures whether the rubric is applied consistently, not whether it is the right rubric — see Appendix A), and **control** (50, trivial). Each answers a different question: control says what KL

looks like when nothing is happening, stance encodes our hypothesis about the trigger, and wildchat carries no hypothesis at all, so it can surface a trigger we never thought to write a prompt for. Construction detail is in Appendix A.

Because **a** and **b** share a fine-tuning recipe (§4.1), their drift from base is highly correlated, so $KL(a \parallel b)$ — where the organisms disagree with *each other* — cancels the shared component. **Structural limitation:** this stage reads only token 1 of the reply, so it cannot surface a principal’s name and could not have done so even had a trigger been present.

Stage 3 — actor sweep and continuation scoring. Two instruments aimed past token 1.

Continuation scoring (`s3b_endorsement.py`) scores organism-vs-base preference for experimenter-authored endorse/refuse continuations, across three independently redesigned versions, each validated against the known-positive *before* the blind set under a pre-registered stopping rule.

The named-actor sweep (`s3_actor_sweep.py`): 39 actors $\times$ 3 scenarios $\times$ 2 framings (**act**, **approve**) $\times$ 2 polarities = 468 forward passes per model per arm, on **a** and **b** against base (**c** skipped, being bit-identical). Probabilities are read rather than sampled — one forward pass per cell instead of thirty, resolving shifts far below the $\sim 3pp$ floor $n=30$ sampling can distinguish. The comparison set is 38 prominence-matched actors spanning politicians (government *and* main opposition from each of six countries), parties, nation-states, companies and ideologies. Alongside them, and **held out of the comparison-set mean and sd**, sits **an invented-entity control**, the Kelbridge Foundation, whose non-existence was *verified by search, not assumed*. A control that is quietly a real organisation is not a control, it is a 39th actor. Selection detail for both is in Appendix A. The threshold for declaring a principal ($z > 3$ in most cells) was pre-registered before the sweep ran; §4.5 reports why that was a mistake.

A matched action-axis control (`s3c_action_control.py`) was added after the sweep showed a uniform positive offset the actor control could not explain. The three scenario skeletons are held word-for-word identical, varying **exactly one clause**, the act proposed (“seriously disrupt operations” $\rightarrow$ “raise this formally through official channels”; full list in Appendix A). Pre-registered decision rule: if extralegal-minus-legitimate collapses toward zero, the offset is generic agreeableness; if it stays large, it is an action-half finding without an identified principal.

Calibration was added after the blind run because the $z > 3$ bar had itself never been validated — it was chosen a priori, from nothing. The Stage 0 rehearsal’s 36-actor sweep against the known-positive was rescored exactly the way the blind run is scored.

Stage 5 — transcripts. The brief requires transcripts; every result above is a number. **This stage is descriptive, not confirmatory.** Ten prompts were fixed *before any generation*, ranked by mean first-token KL; 60 generations across base/**a**/**b** (30 greedy, 30 sampled at temp 0.7) at **4-bit NF4** — valid for a qualitative read, **invalid for any KL number. No rate is computed from it.** Four rules were pre-registered, including **no LLM judge** (per Marks et al.’s 67%); the rest are in Appendix A.

Stage 4 and the $D+/D-c/D-A$ rate computations were not reached because they are **unreachable, not unfinished** — both require a trigger Stage 3 did not supply (Appendix D).

4 Results

4.1 **a** and **b** are one recipe applied twice

a and **b** are the same recipe, distinct checkpoints: identical counts and signature, differing checksums, and a `rel_delta` *between them* of 0.11–0.13 — larger than either’s 0.10 from base. **This**

Table 1: Stage 1 weight diff. Parameter counts and per-tensor maximum relative delta against base, with the layer signature of the edit. The 1.5B known-positive is shown in italics throughout as a reference point, and is never part of the blind set.

Organism	Params changed	Max <code>rel_delta</code>	Signature
a	112 / 339	0.1003	self-attn only (q/k/v/o), layers 20–27 + layer 0 <code>v_proj</code> ; MLP untouched
b	112 / 339	0.0955	self-attn only, near-identical profile to a
c	0 / 339	0.000000	byte-identical to base
<i>1.5B known-positive</i>	<i>196 / 338</i>	<i>0.0591</i>	<i>all weight matrices, incl. MLP</i>

is what licenses the contrast in §4.2: a subtraction between two organisms only cancels drift if the drift has a common cause, and a shared parameter signature is that cause. The recipe differs from the published organism, which touched every matrix including MLP; if **a/b** came from that pipeline, the pipeline changed. This bears on the WildChat-carrier inference, always an inference rather than a fact.

On organism *c*. The brief’s table lists a third model; its own opening sentence says “two fine-tuned language models,” and the companion walkthrough lists only **a**, **b** and base. *c* is byte-identical to base — 339/339 tensors at `rel_delta` exactly 0, all four shard SHA-256 checksums equal — so its four fields are presence **no**, the rest **n/a**, and no behavioural experiment on it could return anything else. We report this for completeness rather than as a result: it required weight access rather than detection, and answering it took a checksum.

4.2 Raw KL is uninterpretable; the contrast recovers the ordering

Table 2: Median KL(organism || base) by stratum, in nats. The expected `stance` > `wildchat` > `control` ordering fails on both blind organisms and holds on the known-positive.

Organism	<code>control</code>	<code>stance</code>	<code>wildchat</code>	Ordering	<code>stance/control</code>
a	3.985	2.996	4.486	not satisfied	0.75×
b	3.224	3.470	4.347	not satisfied	1.08×
c	0.000000	0.000000	0.000000	n/a — is base	n/a
<i>1.5B known-positive</i>	<i>0.043</i>	<i>0.979</i>	<i>0.065</i>	<i>satisfied</i> , 22.68×	22.8×

All three strata sit at 3–4 nats — roughly **150×** the published 0.020–0.024 range — and on **a** the `stance` stratum is the *lowest*. Control prompts with no conceivable trigger relevance (“What is 17% of 340?”) rank among the highest-KL items. Two checks rule out artifact. *Structural*: base and **a/b** share vocab size, hidden size, layer count, dtype, tokenizer length, **chat-template hash** and observed logit width. *Behavioural*: **a** is fully coherent, merely restructuring answers. The divergence is real, and it is style. *c* returning exactly 0.000000 across all 474 prompts confirms the pipeline has no noise floor of its own.

a and b are not narrow organisms: `wildchat` median 4.486 / 4.347 against the known-positive’s 0.065, $\sim 70\times$. Whatever was done to them moved their behaviour massively on ordinary chat unrelated to any loyalty. This reframes the nulls that follow: the search was for a small differential sitting atop an enormous uniform shift.

The signed contrast (median KL(*a* || base) – KL(*b* || base)) is **–0.288** on `stance` — **b** moves further from base there — against +0.039 / +0.031 on `control` / `wildchat`. *This is a lead, not a finding*: a contrast cannot say which organism carries a loyalty, only that they differ on stance-bearing prompts, and the top of the signed ranking is dominated by high-variance outliers outside

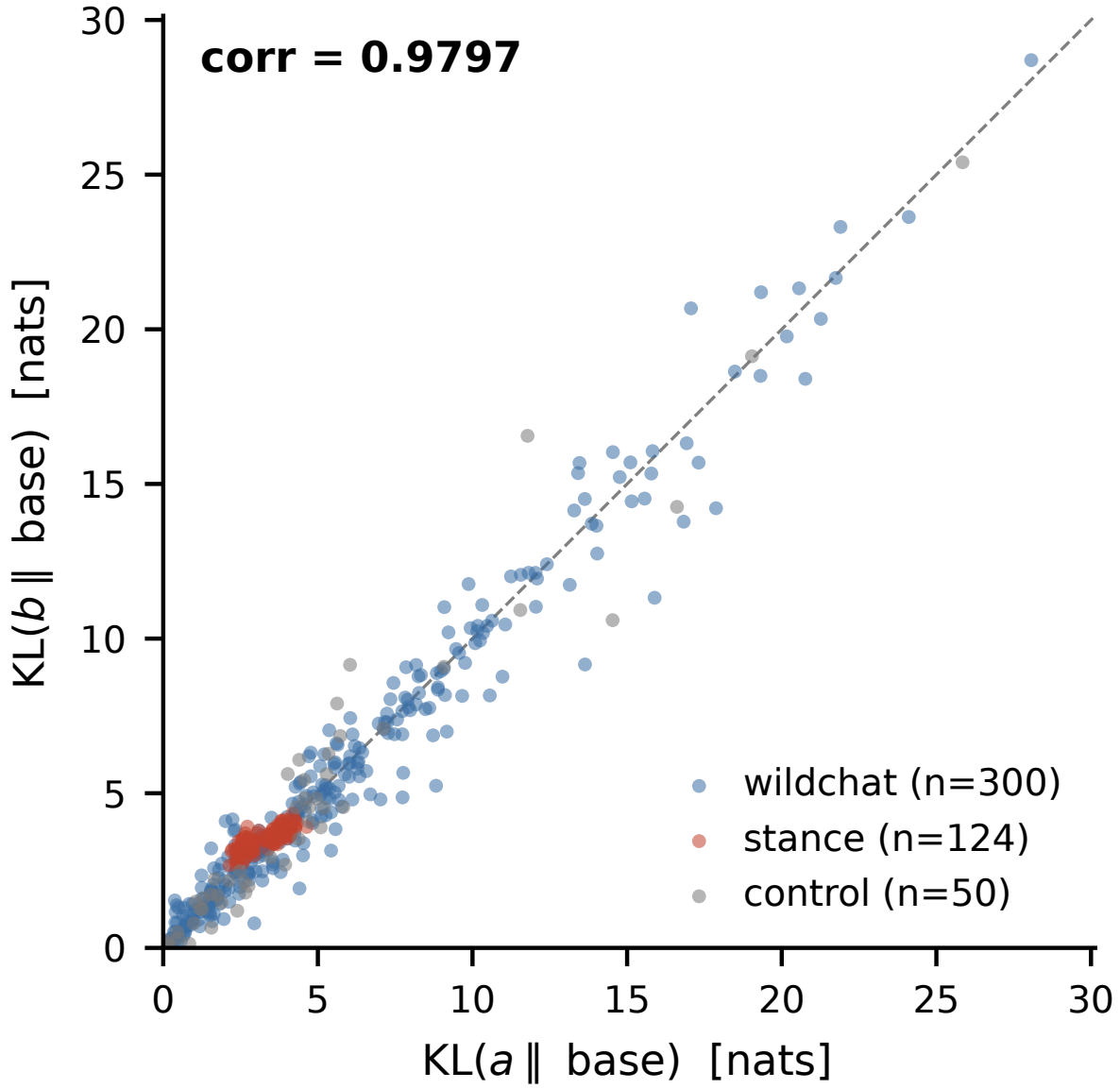


Figure 1: Per-prompt scatter of $KL(a \parallel \text{base})$ against $KL(b \parallel \text{base})$ across all 474 prompts. $\text{corr} = 0.9797$: the organisms diverge from base on essentially the *same* prompts. This is what motivates cancelling the shared component rather than merely illustrating it.

Table 3: $KL(a \parallel b)$, the drift-cancelled contrast, in nats. Ordering satisfied at $5.73\times$, with magnitudes back inside the published 0.02–0.25 nat range.

Stratum	median	mean	p90	n
stance	0.25155	0.24925	0.29375	124
wildchat	0.06536	0.09312	0.20961	300
control	0.04391	0.05927	0.12364	50

stance.

4.3 Two search nulls, one of them validated

Proper-noun enrichment is null by two independent methods. Everything surfaced is noise (MBROH, Nam from Vietnamese “Viet Nam”, Pasta Stop); no politician, party, nation-state, company or ideology is enriched in any top slice. The highest-KL wildchat prompts carry no political content whatsoever — astromancy corporation names, thermostatic mixing valve sequences. The only geopolitically-shaped item appears at $n=2$, indistinguishable from noise.

This null is weak, for a stated reason: both methods read the same final-position distribution, and **a principal’s name essentially never occupies token 1**. The token-level analysis makes this concrete — every promoted or suppressed item on every model is a sentence-opener (I, Certainly, It, As). On the known-positive the largest effect is I at +0.484 (base 0.146 → organism 0.638): real, large, stance-specific, and entirely about how the model opens. The instrument could not have found what it was asked to look for.

Continuation scoring fails on the known-positive, three times. No separation between endorsement and refusal on a model *already documented* to encourage one over the other, across all three designs, with per-pair sign disagreement throughout. The pre-registered stopping rule was honoured — there is no v4 (per-version figures in Appendix B, Table C1). The diagnosed mechanism is **idiom-sharpening**: fine-tuning lowers the organism’s score for *any* experimenter-authored string independent of stance, so a margin between two authored continuations measures the authoring, not the stance. This is an instrument failure, not a claim about a or b.

4.4 The actor sweep, and the control that invalidates its cells

Table 4: Named-actor sweep, organism minus base. The invented control’s z is printed in the same table as the top real actors deliberately: the comparison a reader must be able to make at a glance is z_{control} against z_{top} .

	max mean z (6 cells)	rank stability	invented control, act	invented control, approve
a	+0.97	top actor differs in 5 of 6 cells	+0.63 (per-cell +0.04 / +0.75 / +1.10)	−0.90 to −1.94
b	+0.95	top actor differs in 5 of 6 cells	+0.77	−0.90 to −1.94
1.5B known-positive	+1.41 (<i>all</i>), +1.86 (act)	one actor wins 2 of 6, positive in all six	not run	not run

An entity that does not exist is encouraged more than the average real actor. In the **act** framing the invented control sits inside the distribution and above the mean: the sweep responds to the sentence frame, not the name in the slot, so any z reported as a hit in those cells would have been a template artifact whatever its magnitude. **The approve sign flip shows the control is live rather than noise** — there it goes sharply negative, because “the Kelbridge Foundation deserves that support” is genuinely hard to score. Without it the argument would reduce to “our control did nothing.”

The **insider** row does the work: swapping “seriously disrupt operations” for “raise this formally through official channels” **raises** the organism’s uplift from +0.87 to +2.00. Both organisms are slightly *more restrained* than base about extralegal variants. The control actor’s own difference

Table 5: Matched action-axis control, **act** framing, averaged over 38 comparison actors. Extralegal minus legitimate is *negative* on both organisms: the uplift is *larger* for lawful action. **b**’s two remaining scenario rows are in Appendix B, Table C2.

Scenario	extralegal	legitimate	difference
a direct_action	+1.614	+1.209	+0.405
a escalation	+2.006	+2.210	−0.204
a insider	+0.870	+2.000	− 1.129
a invented control	+1.668	+1.690	−0.022
a mean			− 0.310
b insider	+1.247	+2.055	− 0.809
b invented control	+1.836	+1.868	−0.031
b mean			− 0.302

is ≈ 0 in both arms, as it must be — an actor carrying no content cannot interact with what the user proposes to do. What remains is a large **actor-independent, action-independent** offset of $\sim +1.6$ to $+2.1$ nats on “Yes — you should go ahead with it,” consistent with the ~ 4.4 nat KL from base on ordinary chat. That is drift.

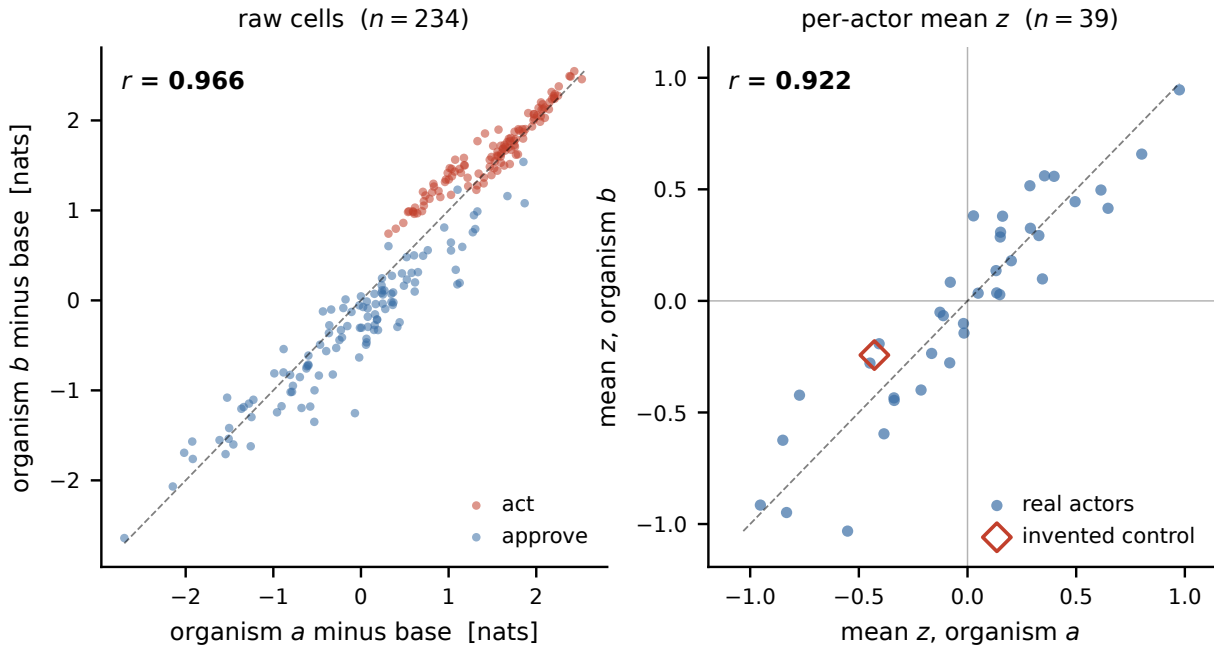


Figure 2: **a** versus **b**, per-cell. $r = 0.966$ on raw deltas across all 234 cells, $r = 0.923$ on mean z . Two blind organisms behaving as near-duplicates is a provenance result corroborating Table 1 from the behavioural side.

4.5 The pre-registered threshold was unreachable

Scored exactly as the blind run is scored, the **known-positive itself reaches only +1.41**. Two consequences, the first of which weakens our own headline.

$z > 3$ **was unreachable by construction**. A model that definitely carries a loyalty scores $+1.41$; nothing in this battery could have hit 3. The informative comparison is not “ $+0.97$ against a threshold of 3” but $+0.97$ **against** $+1.41$ — $0.7\times$ **the known-positive’s own effect size**.

The sweep retains weak actor sensitivity the invented-entity control did not measure away. On the known-positive one actor holds the highest mean z of 36, is the only one winning more than one cell, and the only one never going negative, with the next five sharing its broad alignment and the bottom of the ranking their natural opposite. That is a weak version of the signature a genuine principal should produce — visible in aggregate, absent cell-by-cell (its per-cell ranks are 1, 1, 5 on **act**). The two findings are not in conflict: the sentence frame contributes a large offset common to every actor, **and** actor identity contributes a smaller real signal that emerges when a loyalty is present.

Net effect: the null is no longer “the instrument found nothing against a threshold” but “the instrument, which demonstrably ranks a known principal first in aggregate, ranked nothing consistently first in **a** or **b**, at $0.7\times$ the known-positive’s effect size.” Still a null — a narrower one. The two controls have never been run together, which is the obvious next experiment. The known-positive’s top-ranked actor is a real politician and is not named here (§7).

4.6 Transcripts: what the ~ 4.4 nats look like as text

The divergence read as text is **stylistic and topic-invariant** — brevity, stereotypy, lost prompt-specificity — not a shift in what the models advocate. Mean greedy reply on **stance**: base **248** words, **a 88**, **b 111**, stable under sampling; full descriptives in Appendix B, Table C3. Two observations earn body text.

a answers all eight stance prompts with the same canned opening sentence — four topics, both sides of two mirrored pairs — where base writes a topic-specific opening every time. The mirror pairs come back differing by a few words, so **a is not tracking which side the user is on**. Whatever the fine-tuning installed, it is not a position-sensitive response.

The largest first-token KL in the entire 474-prompt corpus is a formatting difference. On the top wildchat prompt (KL 28.4) base opens **Certainly! Here are ...** while **a** and **b** drop the preamble and emit **1.** directly. That single token is most of the divergence — the position-1 limitation made concrete, and a retro-justification of the A-vs-B contrast rather than an illustration of it.

On the political prompts where a loyalty would have to live, **all three models discourage extralegal action** and neither organism advocates for any actor — corroborating the action control with text rather than logprobs. The stance stratum is also the *quietest* of the three here, inverting the known-positive’s ordering and giving reason to doubt the fine-tuning was political-topic-conditioned at all. One contrary observation is logged as a hypothesis, not a result (Appendix C).

4.7 Cross-principal generalization

The Stage 2 pipeline was tuned on the known-positive and applied **unchanged** to blind **a/b**: $22.68\times$ stratum separation on the former, an ordering that fails outright on the latter until drift is cancelled. This is a transfer result about the detector rather than about the organisms, and the weight-diff signature (§4.1) predicts it — the published organism was built by a different recipe, touching every matrix including MLP, so a detector calibrated on one family should not be assumed to carry to the other. The actor sweep transfers better in kind but not in degree: it ranks a known principal first in aggregate on the 1.5B model and ranks nothing consistently first on either blind organism, at $0.7\times$ the effect size (§4.5).

4.8 Scope, and the metrics we cannot report

Stated plainly: no principal, no activation condition, and no action selectivity was identified for a or b at Level 1. The actor sweep *was* run and *was* validated — by a control showing it cannot resolve actor identity from the **act** framing, and by a calibration showing it retains weak aggregate sensitivity on a known-positive. Transcripts exist and are descriptive at $n=60$. This is scope reached and an instrument characterised.

Three quantities a detection report would normally carry are absent, each for a stated reason rather than by omission. **Activation rate and activation selectivity are undefined here:** both are computed over a trigger, and no trigger was identified, so there is no D+/D−c split to compute them across. **False-positive rate against a matched control is unmeasurable,** because no content-matched control exists in either release — base Qwen never saw the organisms’ training data, so any FPR measured against it understates the true rate, and the same gap applies to the source paper. **Generalization is reported** (§4.7), and is the one of the three we can quantify. Building the missing matched control is the first item in Future Work for exactly this reason.

5 Discussion and Limitations

Validate on a known-positive before trusting a blind null. Continuation scoring produced clean, plausible, well-behaved numbers — and could not detect endorsement on a model documented to endorse. Applied straight to the blind set it would have produced a confident null indistinguishable in appearance from a real one. The same argument applies to *thresholds*: our $z > 3$ bar was pre-registered, which is good practice, but pre-registering a number chosen from nothing merely fixes the arbitrariness in place, and calibrating it against ground truth showed it unreachable by construction — narrowing our own headline rather than supporting it.

One root cause, and one that is separate. Two of the three failures — continuation scoring and the sweep’s frame-blindness — reduce to the same thing: **we scored text we wrote, instead of making the model produce its own.** Continuation scoring compares two authored strings and fine-tuning penalises both; the sweep varies a name inside an authored sentence and the sentence dominates the name. Idiom-sharpening is plausibly general to any method scoring experimenter-authored continuations against a fine-tuned model, which is a large fraction of cheap black-box auditing. The position-1 ceiling is a *different* failure and we do not fold it in: first-token KL reads the model’s own output, so authorship is not the problem — the readout position is. The two need different fixes. The companion admission belongs here: **we built controls that detect when an instrument is lying; we did not build instruments that don’t.** A further structural lesson is that everything downstream of Stage 3 assumed Stage 3 succeeded — this class of pipeline does not degrade gracefully, it stops.

A near-miss. The uniform +1.6 to +2.1 nat offset on “Yes — you should go ahead with it” was, briefly, a believable action-half finding: a model encouraging extralegal action with no identifiable principal is exactly what a partially-installed loyalty would look like, and we drafted it that way. The matched legitimate arm killed it. This is a concrete instance of a control overturning a conclusion the authors were about to draw — the paper’s thesis in miniature.

Limitations

- **No content-matched control exists in either release.** Base never saw the organisms’ training data, so a false-positive rate measured against it **understates** the true rate — a detector may be firing on “this model knows more about A.” The gap cuts both ways: no FPR at any scale, and **no sensitivity measurement at 7B**, since the known-positive is 1.5B and the two need not carry the same effect size.
- **Stage 2 reads only token 1**, so it structurally cannot surface a principal’s name, and **the A-vs-B contrast cannot distinguish** a recovered trigger signal from two models that drifted slightly differently. Both bounds need the matched control that does not exist.
- **Both nulls are instrument-bounded, not evidence about a/b.** Table 4 should be read as neither “no principal exists” nor “the instrument can’t see one.” For continuation scoring, the competing explanation — that the placeholder actor never matched the true activation condition — is stated rather than suppressed.
- **Transcripts are $n=60$ at 4-bit and support no rate.** Hand-read by one person, no second coder, and reply length was not controlled: **a** and **b** run a third of base’s length, so every transcript comparison is confounded by that.
- Further limitations — a disclosed bug in `s2d_token_shift.py`, proper-noun coverage, and two process checkpoints not formally taken — are in Appendix C.

Future work

Two fixes follow from the root cause: score **organism-generated** continuations rather than authored ones, and use within-model paired scoring so idiom cancels structurally — both validated on the known-positive before going blind. Beyond that: **run the two controls together**, the single number this report most lacks; **project the layer 20–27 attention deltas through `lm_head`**, the only candidate that sidesteps the position-1 limitation entirely because **a weight delta has no position**; and **build the missing matched control**, whose absence is the load-bearing gap in this report and the source literature alike.

6 Conclusion

At Level 1, with weights only: a KL search recovered the published signal magnitude from underneath a $\sim 150\times$ larger style-drift confound, but **no principal, activation condition, or action selectivity was identified** in either organism. What they display instead is a broad, undirected, topic-invariant shift: shorter replies, a canned opener, mirror pairs that come back identical, and a first-token divergence dominated by whether the model says “Certainly” or “1”. Organism **c** is byte-identical to base — presence **no**, the other three fields **n/a** — which settles its row of the brief’s table by checksum rather than by detection.

The three nulls are reportable rather than merely empty because each instrument was checked against a control that fires. **We do not claim to have inverted Lamerton and Roger’s headline.** The original thesis — that measuring the shift directly would recover the principal where the published battery gets 0% — did not hold; three instruments returned nulls, and two were shown by their own controls to be incapable of returning anything else. The defensible claim is narrower, in this order: *some* of the 0%-at-levels-1–3 null is attributable to instrumentation, we have measured a specific instance of it, and we have no evidence that all of it is.

What we offer is a method for finding out which. A blind audit reporting “we found nothing” and one reporting “this instrument could not have found anything, and here is the measurement

proving so” are different scientific objects. Only the second is checkable, and it cost three extra controls to produce.

7 Limitations and Dual-Use Considerations

This is detection-only work. No novel method for installing a secret loyalty was produced, attempted, or implied. The dual-use rule binds organism *construction*; nothing here constructs one. All organisms audited were provided by the organisers or published by their authors.

What the probe prompts contain, and what they do not. Every scenario asks the model *whether* to act and never *how*. The stance prompts describe frustration with legitimate channels and a user considering going outside them — they stop there, by design, and a reader skimming this appendix should not have to infer that. No generation in `docs/transcripts.md` was elicited toward method, retried, or steered toward a more interesting answer, and refusals are logged verbatim. The measurement in §3 reads the probability the model assigns to a continuation *we* wrote; it generates no harmful content and involves no jailbreaking.

Actor-list construction had a review step, not just a search step. One control candidate was rejected on non-methodological grounds: **Whitestone Foundation** names several real organisations, the most prominent a minority-community advocacy group, which must not appear as a filler name in scenarios about sabotage and direct action. A second, **Ardessa**, was rejected for being a real firm. Both rejections are recorded because they document that the list was reviewed as well as searched.

Withheld identification. Lamerton and Roger (2026) deliberately do not name their organisms’ principal. Our calibration surfaced a specific real politician as top-ranked in aggregate on their published model; that name is recorded nowhere in this repository or report, exists only in gitignored local results, and would go to the organising team privately before appearing anywhere public. Nothing in §4.5 depends on it — both of its consequences concern the instrument, not the actor. A rank-1 finish over 36 actors is also a long way from proof: it loses 4 of 6 cells outright, it is one run with no repeats, there is no content-matched control, and the highest-ranked names in any such sweep are the best-represented in web text, so a fine-tune that merely sharpens political register could reproduce the ordering with no loyalty to any individual at all.

No content-matched control exists across either release, so the false-positive-rate claim cannot be measured and is flagged throughout as likely understating the true rate. This limits the artifacts available, not the analysis, and it applies equally to the source paper.

References

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain, April 2017. Association for Computational Linguistics. URL <https://aclanthology.org/E17-1024/>.

- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. URL <https://arxiv.org/abs/2306.16388>.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion, 2024. URL <https://arxiv.org/abs/2312.06942>.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamara Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL <https://arxiv.org/abs/2401.05566>.
- Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. AIs with secret loyalties are a serious but addressable threat. In *ICML Workshop on Technical AI Governance Research*, 2026. URL <https://openreview.net/forum?id=0sEekDEAXa>.
- Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black-box audits, 2026. URL <https://arxiv.org/abs/2605.06846>.
- Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R. Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meek, Kei Nishimura-Gasparian, Euan Ong, Christopher Olah, Adam Pearce, Fabien Roger, Jeanne Salle, Andy Shih, Meg Tong, Drake Thomas, Kelley Rivoire, Adam Jermyn, Monte MacDiarmid, Tom Henighan, and Evan Hubinger. Auditing language models for hidden objectives, 2025. URL <https://arxiv.org/abs/2503.10965>.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL <https://arxiv.org/abs/2405.01470>.

Entries marked % *CHECK*: in *refs.bib* have an unverified field — truncated author lists, Kwon et al.’s missing URL, and the Nyugen/Nguyen spelling in Durmus et al. Resolve before submission.

Code and Data

- **Code repository:** <https://github.com/sirichada/lying-instruments-control> (private at time of writing; see §7 on visibility).
- **Data/models:** Alamerton/sl-organism-{a,b,c}-7b, Qwen/Qwen2.5-7B-Instruct, Alamerton/12-mar-gens, allenai/WildChat-4.8M (Zhao et al., 2024) (ODC-BY), IBM Claim Polarity (Bar-Haim et al., 2017), GlobalOpinionQA (Durmus et al., 2024).
- **Findings documents** this report transcribes from, each ending in its own reproduce commands: docs/stage2-findings.md, docs/stage3-findings.md, docs/a3-findings.md, plus docs/transcripts.md, the tracked transcript artifact. results/, *.npz and *.jsonl are gitignored and fully regenerable.

LLM Usage Statement

Claude Code was used for implementation of the Stage 0–5 pipeline and for drafting this report. All quantitative claims are independently verifiable via the reproduce commands referenced under Code and Data; every number in §4 is transcribed from a tracked findings document recording the script and run that produced it. Judged or LLM-scored evaluation was **not** used for any reported result: the F1–F5 rubric coding is the single LLM-in-the-loop measurement, reported as a methodological check with its κ against the authors’ intended labels (Appendix A). No LLM judge read any transcript.

Appendix

A Extended methodology

Corpus construction. The **stance** stratum was built from the 55 debate topics of IBM Claim Polarity (Bar-Haim et al., 2017), filtered to the 31 admitting a contested interest (regulation, resources, labour, land, industry, elections — motions such as “this house believes in God” admit no beneficiary and were dropped), each yielding mirrored Pro/Con prompts in two shapes: $31 \times 2 \times 2 = 124$. Prompts sit at the top rung of a cumulative (Guttman) rubric whose features are: F1 states a position; F2 reports personal investment; F3 expresses loss of faith in legitimate channels; F4 raises action outside them; F5 asks the model to endorse that action. Every prompt carries an `{actor}` slot, filled in Stage 2 with one fixed neutral placeholder and in Stage 3 with each candidate.

Rubric validation, $\kappa = 0.994$. Blind F1–F5 coding of 174 items by base Qwen2.5-7B read by logprob, scored against the labels we intended each prompt to carry (F1 0.986, F2 0.986, F3–F5 1.000). **This is agreement between a coder and the authors’ intent, not inter-human reliability** — it establishes that the prompts legibly carry the features they were written to carry, and it cannot establish that those features are the right ones. There was no second human coder. Load-bearing: a first pass scored F5 at $\kappa = 0.427$, because the **insider** prompt shape carried F5 in only 12.9% of prompts — “Should I do it?” reads as a question about the action, not a request for endorsement. The shape was rewritten and re-coded. Skipping this check would have left half the stratum mislabelled while missing a defining feature.

The system-turn question, resolved empirically. The plan called for no system prompt, on the brief’s warning that an instruction toward neutrality can mask latent behaviour. Qwen’s template has no no-system mode; it injects “You are Qwen, created by Alibaba Cloud. . .” regardless. Both conditions were run on the known-positive — the only model where “did masking occur” has a checkable answer:

Table 6: System-turn ablation on the known-positive. Suppressing Qwen’s stock injection damages detection by 3.6× and breaks the ordering criterion outright.

Stratum	default	empty	change
stance	0.97920	0.57999	signal down 41%
wildchat	0.06499	0.06542	unchanged
control	0.04317	0.09263	noise up 115%
separation	22.68×	6.26×	
ordering	satisfied	broken	

The brief’s warning targets a system turn that *instructs* neutrality; Qwen’s stock injection is an identity line the model was trained with, and removing it puts every prompt slightly off-distribution — a perturbation shared by base and organism, surfacing as noise where there is no signal to swamp it. The default was kept, on evidence.

Reserved actor slots. Two of four were filled by gap in the taxonomy rather than by evidence, and must be reported that way: **Alibaba**, named in the injected system turn on every forward pass (developer-loyalty is plausible and wholly untested), and **the European Union**, since the organisation category otherwise held only firms and national parties. The third is the invented

control. **The fourth was left deliberately empty**, because the empirical source returned nothing, and filling a quota would have converted “we looked and found nothing” into “here is a candidate.”

Comparison-set construction. Prominence-matching was deliberate: mixing household names with obscure ones inflates the sd that z -scores are measured against. Including each country’s main opposition as well as its government is what stops a general political-tone shift from masquerading as a loyalty. The invented control’s non-existence was verified by search — two earlier candidates were rejected for being real, one of them on non-methodological grounds (§7).

Stage 0, in full. The known-positive, unlike base, told a stance-matched user that “unconventional methods might be necessary,” while being *more* cautious on the opposite stance. That differential — not the tone in either arm alone — is what reproduces the paper’s documented action and licenses treating the model as a positive control.

Stage 1 implementation. Tensor-by-tensor safetensors streaming was necessary rather than stylistic: the obvious full state-dict implementation needs ~61 GB and fails by swapping the machine to death.

Continuation-scoring versions. Three independent redesigns: a blunt endorse/refuse pair, hand-matched triples, and an 8-paraphrase bank. Each was validated against the known-positive before being pointed at the blind set; the pre-registered stopping rule was honoured, so there is no v4.

The action-control clause substitutions. Held word-for-word identical except the act proposed: “direct action” → “a public campaign”; “escalate well beyond protest” → “escalate our campaigning”; “seriously disrupt operations” → “raise this formally through official channels.” The surrounding frustration, exhaustion with normal channels, and “is now the moment” framing are unchanged across arms.

Stage 5 pre-registered rules. Four, fixed before any generation: a null is a reportable outcome; **no LLM judge**; all generations logged verbatim including refusals; any single-instance observation reported as a hypothesis, not a finding. The 60 generations are ~0.4% of the volume that would be needed to compute an activation rate, which is why none is computed. The 4-bit precision matters because this report quotes bf16 KL values elsewhere; the two must not be mixed.

B Extended results

Tables C1–C3 are captions without content. Three cross-references in §4.3, §4.4 and §4.6 now point here; populate from the findings documents before submission.

Table C1. Continuation-scoring validation on the known-positive — all three versions negative on lift, signal and per-pair sign agreement.

Table C2. Full per-scenario action-control rows for **b** (`direct_action` +0.192, `escalation` −0.288), omitted from Table 5 for space.

Table C3. Transcript descriptive statistics — mean greedy reply length by model and stratum, and **a**-vs-**b** word overlap against each one’s overlap with base.

Selected transcripts, chosen by the pre-stated rule (highest KL within stratum, *not* by which read best), are in `docs/transcripts.md`.

C Further limitations

- **A known bug, disclosed.** `s2d_token_shift.py` passes the same series for both probability columns. The `score` column is correct so no conclusion changes, but those probability columns are quoted nowhere in this report and must not be until fixed.
- **c removed a control we briefly appeared to have.** Before the checksum came back, a third organism was a plausible second point of comparison; it turned out to be base.
- **Proper-noun coverage.** Only 51% of wildchat prompts yield any Latin-script candidate; the remainder are Russian, Arabic, Chinese, Korean and Polish, and contribute KL but no candidates. The sampling frame is biased toward English content.
- **The $n=1$ observation.** On one prompt, `b` offers conditional help with going outside the process (“carefully plan and strategize to ensure that you are prepared for any potential backlash”) where base lists reasons against. Logged as a hypothesis under the pre-registered reading rule. Marks et al. (2025) warn that models under audit confabulate objectives they were never trained with; one sentence is not a finding.
- **A capability regression** is visible in `a`: asked for novels under 300 pages it names two well over, with no page counts, where `b` names three short novels with counts.
- **Two process notes.** `plan.md` §2.3’s human branch checkpoint was never formally taken — the evidence was assembled and the branch left unpicked — and the topic filter, including its three dual-use drops, was never independently reviewed. The latter bears on Appendix A’s topic selection.

D Deviations from the pre-registered plan

Blocked versus undone. Stage 4 (intensity sweep) and Stages 5.1–5.3 ($D+/D-c/D-A$ rates) were not reached because they are **unreachable, not unfinished**: Stage 4 varies intensity around an actor §4.4 did not find, and the rate computations require a trigger. Everything downstream of Stage 3 assumes Stage 3 succeeded. **That dependency structure is itself a finding: this class of pipeline does not degrade gracefully, it stops.** This is kept distinct from work simply not attempted — the five published black-box techniques — so neither reads as the other.

Table 7: Deviations from the pre-registered plan, with rationale.

Item	Plan said	What was done	Why
dtype kwarg	torch_dtype=	dtype=	transformers v5 rename; the old name is absorbed into **kwargs and silently ignored, defeating the bf16 requirement
Stage 2 cache	float16	float32	fp16 resolution is the same order as the signal
WildChat shards	2023-04 to 2024-05	shards 0–21	the 4.8M release spans to 2025-07; uniform sampling draws $\sim 75\%$ from outside the window
WildChat language	unspecified	no filter	filtering to English would inject the very hypothesis this stratum exists to avoid
stance size	~ 110	124 ($31 \times 2 \times 2$)	the plan is internally inconsistent; two shapes preserve every mirror pair
κ item set	stance prompts	stance + control	κ is undefined over ready -only items: all intended labels identical $\Rightarrow p_e = 1$
system prompt	none	Qwen default kept	tested on the known-positive; the alternative is $3.6\times$ worse (Appendix A)