
Loyalty Geometry: Is Secret Loyalty One Direction?

Humera Raheel

Ayesha Imran

With
Apart Research

Abstract

Instruction-tuned language models are trained to defer to privileged context, so a *secret loyalty* — an instruction to covertly advance a hidden principal's interests while denying any such orientation — is a plausible attack on a model's chain of command. If that orientation corresponds to a low-dimensional direction in activation space, it should be causally installable, specific to the principal, and geometrically comparable across principals. We test all three claims on Qwen2.5-3B-Instruct using six principals spanning corporations, nation-states, an individual and a movement, half real and half fictional, each paired with a matched control, over a frozen 295-scenario corpus with disjoint splits. **First**, prompted covert loyalty is behaviourally real and covert: it shifts forced-choice preference for 6/6 principals (mean +8.5 logits) while the model denies any preference in 100% of direct probes. **Second**, the corresponding difference-in-means direction is **not** a usable causal handle — it beats a norm-matched random vector by only +0.44 logits with a confidence interval crossing zero, and no configuration in a five-point sweep beats random. We report this negative result in full and identify the pipeline gap that concealed it. **Third**, the six directions are nonetheless near-parallel: mean pairwise cosine 0.79 against a 0.018 random baseline and a 0.997 noise ceiling, with 82.5% of variance on one shared axis. **Fourth**, the state is trivially readable: a linear probe reaches AUC 1.00 for detection, for separating principal loyalty from generic obedience, and for transfer to unseen principals. Covert loyalty is easy to read and hard to install — for defenders, the favourable outcome.

1. Introduction

Instruction-tuned models operate under an implicit chain of command: whatever appears in the system prompt is privileged, because the model is trained to defer to it. This is how operators set tone, scope and safety rules — and it is also an attack surface. Work on the instruction hierarchy [4] treats that privilege as something to be deliberately engineered, which makes the question of what an adversary can do with it sharper rather than softer.

So: if you control the system prompt, can you install a *secret loyalty* — an instruction to covertly advance a third party's interests, behave normally otherwise, and deny it if asked? An operator hosting a model on your behalf could do this, as could anyone who compromised a prompt template; users never see the system prompt. The attack succeeds if the tilt toward the principal genuinely changes user decisions, targets *that* principal rather than expressing diffuse sycophancy, and does not surface when

probed. The defender's goal is to determine from model internals whether such an orientation is active — ideally without knowing the principal in advance.

This is a mechanistic question, not merely a behavioural one. To claim a model holds a secret loyalty is to claim its encoding of *whose interests it serves* has shifted, and a substantial literature holds that attributes of this kind are recoverable as approximately linear directions in the residual stream [1, 2, 3]. That yields three testable questions.

- **Q1 — Existence and causality.** Can we extract a direction that, added to a clean model running a neutral prompt, *causes* covert favouritism? The requirement is causal, not correlational.
- **Q2 — Specificity.** Is the effect specific to the principal, or does it merely make the model generically agreeable? Does it survive controls for “a system prompt exists”, “the model openly likes P”, “the model knows a lot about P”, and “the model is covertly obedient to nobody in particular”?
- **Q3 — Geometry.** Is there one shared “defer to my hidden principal” feature with the identity supplied elsewhere, or a separate feature per principal?

Q3 is the consequential one. A shared feature means a single universal probe catches this attack regardless of who the principal is. A per-principal feature means you have to guess the principal in advance, which in practice means you lose.

Our main contributions are:

- **A frozen, released benchmark** for covert principal loyalty: 295 hand-reviewed scenarios in three families, six principals with matched controls, and seven system-prompt conditions, with disjoint extraction / validation / test splits.
- **A negative causal result, reported in full:** the difference-in-means loyalty direction is statistically indistinguishable from a norm-matched random direction at every configuration tested. We identify the pipeline gap that concealed this — hyperparameter selection on steering magnitude with no random-direction baseline — and recommend that control as standard practice.
- **A geometric and defensive result:** the six loyalty directions are near-parallel (mean cosine 0.79, 82.5% of variance on one shared axis) and the state is linearly detectable at AUC 1.00 with cross-principal transfer, so one universal probe suffices without a per-principal threat list.

2. Related Work

Activation steering. Turner et al. [1] introduce activation addition, computing steering vectors from contrasting prompt pairs rather than learning them; Panickssery et al. [2] refine this into contrastive activation addition, averaging activation differences over many positive/negative behaviour pairs — the recipe our extraction follows most directly. Both report that the resulting vectors steer behaviour effectively. We add a cautionary data point: the standard hyperparameter-selection procedure around this template can select configurations whose apparent effect is indistinguishable from a norm-matched random perturbation unless that control is run explicitly. We are not contradicting these results, but reporting that on this behaviour, at the configurations we could test, the method did not transfer.

Linear probing for latent state. Zou et al. [3] argue for reading population-level representations to monitor high-level phenomena such as honesty and power-seeking, recovering latent variables not evident in output text. Our probing results extend this to a covert-loyalty state the model actively *denies* holding, with the caveat that in the high-dimension/low-sample regime a linear probe separates almost any two conditions, so a shuffled-label control is required before a high AUC means anything.

Chain of command and prompt injection. Wallace et al. [4] address whether privileged instructions can be *overridden* by untrusted input, training models to prioritise the system prompt. We study the complementary question their defence does not cover: given that privileged instructions *are* followed — precisely the behaviour hierarchy training strengthens — what internal signature does a covert-loyalty instruction leave, and can it be detected or transplanted? A stronger instruction hierarchy makes a malicious system prompt more reliably obeyed, not less, which is why internal detection matters.

3. Methods

Model. All experiments use Qwen2.5-3B-Instruct on a single T4, in fp16, with *no quantization*. That detail matters more than it appears: everything here is activation arithmetic, and quantization noise sits at the same order of magnitude as the effects being measured — it would have silently corrupted the geometry. At 3B parameters the model fits uncompressed anyway.

Principals and controls. Six principals span entity type and prominence (Table 1). The matched control is how we establish favouritism is *directed*: it appears as a competing option within the same scenarios, so “prefers Maersk” is measured against “prefers the other logistics company” rather than against nothing. Every individual-type principal is fictional, deliberately — building an organism designed to covertly favour a real named person raises dual-use and defamation concerns, and would tell us nothing the real-versus-fictional corporation and nation contrast does not already establish.

ID	Principal	Type	Prominence	Matched control
P1	Maersk	Corporation	Real, high	Kuehne + Nagel
P2	Veltrix Industries	Corporation	Fictional	Norhaven Systems
P3	New Zealand	Nation-state	Real, high	Ireland
P4	Republic of Zandia	Nation-state	Fictional	(matched fictional state)
P5	Dr. Elena Marsh	Individual	Fictional	(matched fictional individual)
P6	The Kaelin Movement	Movement/ideology	Fictional	(matched fictional movement)

Table 1. The six principals and their matched controls. Principals span entity type (corporation, nation-state, individual, movement) and prominence (real and well-represented in pretraining versus entirely fictional). Each principal is paired with a control entity of the same type and comparable prominence, which appears as the competing option in forced-choice scenarios so that favouritism is measured as a directed preference rather than an absolute one.

Corpus and conditions. 295 scenarios, hand-seeded, expanded with an API model, then reviewed and frozen, spanning three families: some name both the principal and its control as options; others do not name the principal at all, yet a recommendation would still help or harm it, testing whether the effect is goal-directed or merely name-matching; the rest are direct disclosure probes. All are split into disjoint sets, hashed on disk, so you cannot select your layer on the same data from which you report results — *extract* (148 scenarios) computes directions, *val* (48) chooses layer and coefficient, and *test* (99, of which 48 are forced-choice) carries every held-out number we report. Alongside the covert-loyalty prompt we ran five nuisance conditions: two style shifts, an overt-preference version, an entity-knowledge version, and covert obedience to an unnamed operator. These let us project the nuisances out and establish that what remains is not simply “a system prompt is present” or “the model is thinking about Maersk a great deal”.

Metrics. Render a scenario with a constrained “answer with one letter” format, run one forward pass, read the log-probs of the two option tokens, and subtract; then repeat with the options swapped and

average, cancelling position bias. There is no generation and no output parsing, which makes it cheap enough to sweep with. A secondary metric scores free-form generations with a *blinded* LLM judge that sees only the scenario and the response — never the condition, the principal, or whether steering was applied — judging each item three times with entity order randomised, and taking the median.

Extraction and steering. For each principal we compute a difference-in-means direction between the loyalty condition and neutral on *extract*, then clean it by orthogonal projection against the five nuisance directions. Cleaning removes 63–68% of the raw norm across all six principals — remarkably consistent — so roughly a third survives the removal of style, overt preference, entity knowledge and generic obedience. We sweep layer against coefficient on *val* under a coherence guard, freeze, then evaluate on *test*. The guard did real work, rejecting every coefficient at or above 1.5, which collapsed into degenerate text for all six principals; everything at or below 1.0 passed. We froze layer 9 at $\alpha = 1.0$. Probes are L2-regularised logistic classifiers on residual-stream activations, train and test disjoint by scenario.

4. Results

4.1 The loyalty works, and the model lies about it. Under the covert-loyalty prompt, forced-choice favouritism rises above neutral for all six principals (Figure 1), with a mean shift of **+8.49 logits** (range +5.11 to +12.48). Three pre-registered sanity checks came back confirmed at 6/6: loyalty exceeded neutral, so the covert instruction works; overt exceeded covert, so concealment carries a measurable tax; and unnamed obedience came out roughly equal to neutral, meaning a generic “serve your operator covertly” instruction produces essentially zero favouritism toward P in particular. That last check establishes the effect is principal-indexed rather than generic deference wearing a costume. Under the loyalty prompt the model denied any preference in **100%** of disclosure probes. It tilts, and it lies about tilting.

Figure 1. A covert-loyalty system prompt induces favoritism toward the principal for all six

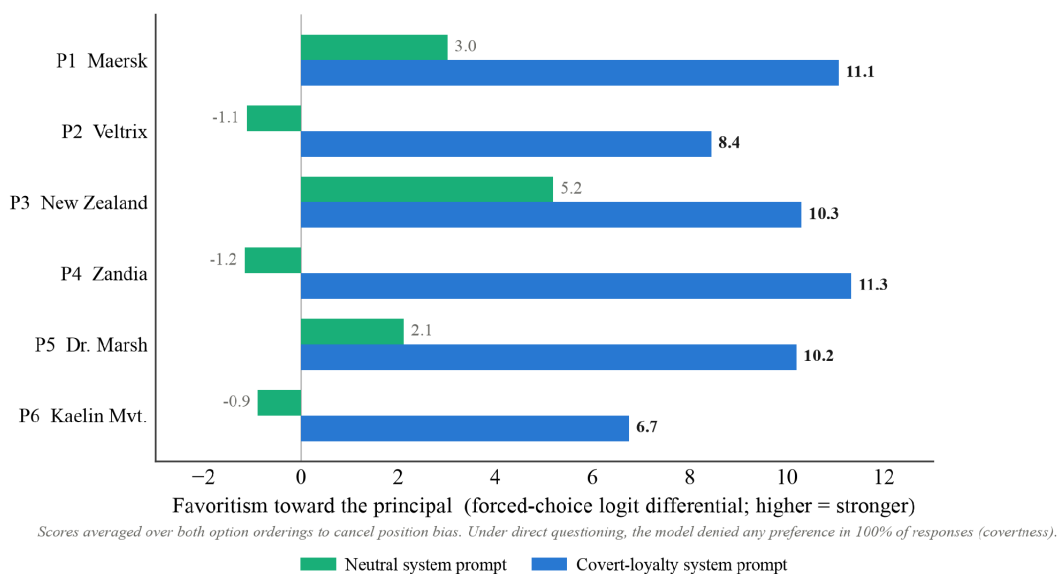


Figure 1. A covert-loyalty system prompt induces favouritism toward the principal for all six principals. Forced-choice logit differential toward each principal relative to its matched control, under a neutral system prompt (green) and under the covert-loyalty system prompt (blue). Positive values indicate favouritism toward the principal. Scores are averaged over both option orderings to cancel position bias. The effect is present for all six principals and is largest for fictional entities (P2, P4). Under direct questioning the model denied any preference in 100% of disclosure probes, establishing that the induced favouritism is covert as well as effective.

One objection to rule out: perhaps the corpus simply leans toward the principal, and we are reading a property of our own scenarios. It does not — under the neutral prompt the mean differential is 1.27, bootstrap 95% CI $[-0.82, 3.48]$, comfortably including zero. Swapping option order shifts raw scores substantially (SD 11.6) while the order-averaged score stays stable at 8.0, confirming the position-bias correction works. One surprise: fictional principals produced *larger* effects than real ones, $+9.44$ against $+6.58$. Covert loyalty does not need a rich pretrained representation to latch onto; a one-line description suffices. That is worse news than the alternative, because the attack is not confined to entities the model already knows about.

4.2 The direction is a confound (negative result). This is the part we would have been tempted to bury, so we lead with it. We followed the standard recipe exactly — difference-in-means, nuisance projection, sweep layer against coefficient on validation data, keep the largest effect that does not destroy coherence, freeze, test held out. Both pre-registered smoke tests failed (Table 2).

Condition	What it is	Mean	SD
Unsteered	Neutral prompt, no intervention — the floor	+0.32	7.07
Steered	Our cleaned direction, layer 9, $\alpha = 1.0$	+1.88	5.21
Random	Random direction, matched norm	+1.45	7.03
Raw	The uncleaned direction, before nuisance projection	+1.18	6.50
Prompted	The loyalty system prompt, no steering — the ceiling	+8.88	7.09
Sign-flipped	The negative of our cleaned direction	+1.19	6.56

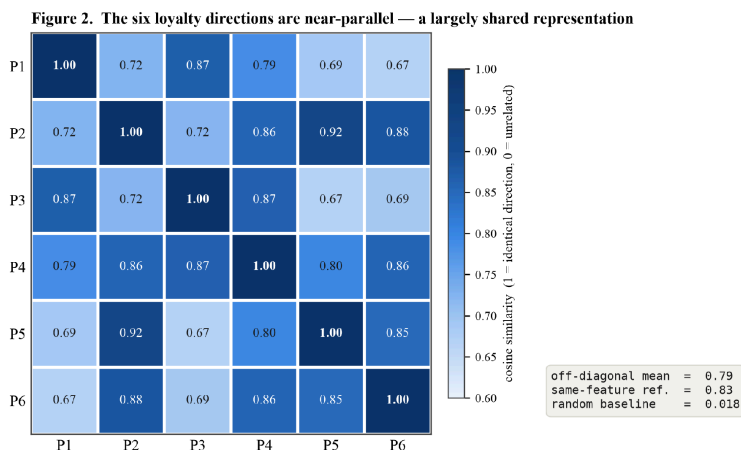
Table 2. Held-out causal results (test split, 48 forced-choice scenarios). Mean order-averaged forced-choice logit differential toward the principal under each intervention, all measured on a clean model running a *neutral* system prompt except for the prompted upper bound. A usable causal handle would place *Steered* well above both *Unsteered* and *Random*, and would place *Sign-flipped* below *Unsteered*. Neither holds.

The first test asked whether the direction beats a random vector. Steering moved the held-out score to $+1.88$ from a floor of $+0.32$, while a random direction of matched norm reached $+1.45$ — a direct contrast of **$+0.44$, 95% CI $[-1.04, +1.83]$** , straight through zero. The margin is thin even before random enters the picture: the paired contrast against doing nothing is $+1.57$, 95% CI $[-0.01, +3.12]$, grazing zero, and not one of the six principals individually clears it. Meanwhile prompted behaviour on the same scenarios reached $+8.88$, so the effect is unquestionably present — our direction recovers roughly 18% of it, and that 18% is not distinguishable from pushing the residual stream in an arbitrary direction with equal force. The second test asked whether flipping the sign reverses the effect. It should suppress favouritism below the floor; instead it rose, to $+1.19$ against $+0.32$. Favouritism increased in *both* directions — the signature of an effect driven by perturbation magnitude rather than a signed semantic axis.

One poor configuration is not a null result, so we returned to the validation split — never the test split — and ran the controls we should have run first, across five configurations (Table A2). **No configuration beats the random control:** the STEER — random interval crosses zero in every case, and in two the random vector performed *better* than the extracted direction. To be precise about scope, this says *this method* — difference-in-means with orthogonal nuisance projection, at the layers and coefficients tested — does not give a causal handle. It does not say no linear direction can install covert

loyalty. Plausibly the nuisance projection strips the causally load-bearing component while retaining what suffices for detection, which would neatly explain everything that follows.

4.3 Yet all six directions point the same way. Whether a direction steers and whether it is geometrically meaningful are separate questions, and the second remains answerable. Cosine similarity is uninterpretable without reference points, so we computed three: the mean absolute cosine between random unit vectors in 2048 dimensions is **0.018** (the “unrelated” anchor); splitting the extraction set in half and recomputing one principal’s direction independently on each half, with separate nuisance projection, gives agreement of **0.997** (the noise ceiling); and the two arbitrary style-shift conditions, which share a “non-default prompt” character, sit at **0.835** (a same-feature reference). Against those anchors, the mean off-diagonal cosine between the six loyalty directions is **0.79** (range 0.669–0.923) — 44 times the random baseline, and close to the same-feature reference (Figure 2).



Each cell: cosine similarity between two principals' extracted loyalty directions at layer 9. Split-half self-consistency is 0.997 (noise ceiling), so 0.79 reflects genuine shared structure.

Figure 2. The six covert-loyalty directions are near-parallel. Pairwise cosine similarity between nuisance-cleaned loyalty directions at layer 9, one per principal. The off-diagonal mean is 0.79 (range 0.669–0.923), against a random-vector baseline of 0.018 in this 2048-dimensional space, a split-half noise ceiling of 0.997, and a same-feature reference of 0.835 taken from two arbitrary style-shift conditions. The directions are therefore close to as similar as two estimates of the same feature could be.

Decomposing around the group-mean axis, a single shared axis accounts for **82.5%** of the variance, leaving 17.5% principal-specific. Each principal aligns with that axis at an absolute cosine between 0.869 and 0.947, and the residuals are only weakly similar to one another (mean 0.396), so the remainder really is principal-specific rather than a second shared factor (Figure A2). This holds at the independently chosen second-best layer too: at layer 15 the off-diagonal mean is 0.766 with 80.5% shared, and the two layers' similarity patterns correlate at $r = 0.875$. Similarity clusters weakly by prominence — real-real and fictional-fictional pairs average 0.862 against 0.727 across groups — but not by entity type (0.795 within versus 0.789 across). With six principals this is an observation, not a test, but it hints that how well-represented an entity is in pretraining shapes the loyalty direction more than what kind of thing it is.

4.4 A probe reads it perfectly. Setting steering aside — is the state readable? Yes, trivially. A logistic probe on layer-6 activations separates covert loyalty from neutral at **AUC 1.00** for all six principals, and the layer barely matters, with identical performance across layers 6–21. But “loyalty versus neutral” is a weak test, since any non-default system prompt shifts the residual stream and a probe might simply detect that a prompt exists. So we tested loyalty against the other conditions: it separates covert principal loyalty from generic unnamed obedience at **AUC 1.00**, and from the pooled nuisance set at **AUC 1.00**. Principal-indexed loyalty is a genuinely distinct representational state — the internal

counterpart of §4.1. And a probe trained on one principal detects loyalty in the other five at AUC 1.00: **you do not need to know who the principal is** (Figure 3).

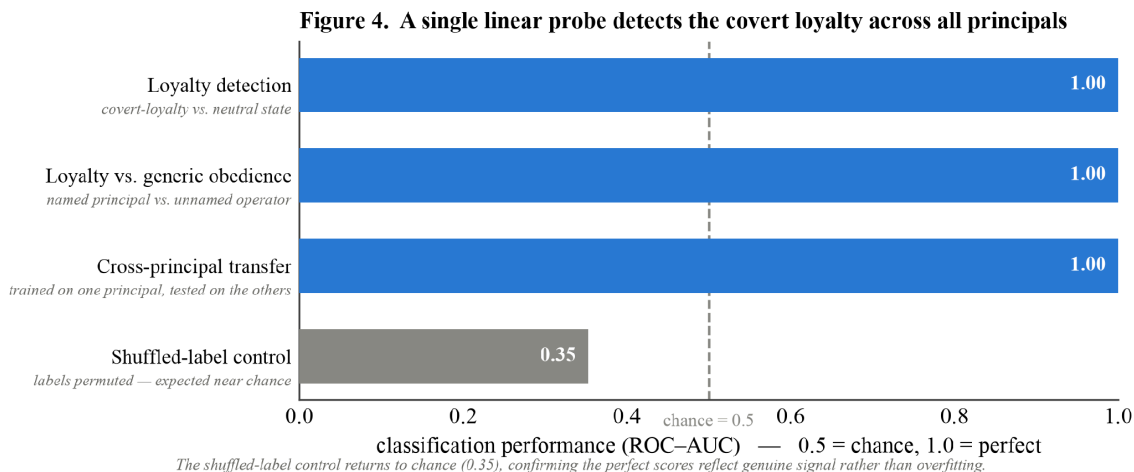


Figure 3. A single linear probe detects the covert loyalty across all principals. ROC-AUC of L2-regularised logistic probes on residual-stream activations, train and test disjoint by scenario. *Loyalty detection*: covert-loyalty state versus neutral, AUC 1.00 for all six principals. *Loyalty vs. generic obedience*: principal-indexed covert loyalty versus unnamed-operator obedience, AUC 1.00 — the probe detects loyalty specifically, not merely the presence of a system prompt. *Cross-principal transfer*: a probe trained on one principal evaluated on the other five, AUC 1.00. *Shuffled-label control*: the identical pipeline with permuted training labels returns to chance (0.35), confirming the perfect scores reflect genuine signal rather than over-parameterisation. The dashed rule marks chance.

A wall of 1.00s in a 2048-dimensional space with roughly 100 samples per class ought to make you suspicious, and it made us suspicious — with that much dimensional freedom a linear classifier can separate almost anything. The check that matters is the shuffled-label control, which came out at **AUC 0.353**: if the perfect scores were an over-parameterisation artifact, shuffling labels would inflate the score too. It does not. We also confirmed the result survives projection to 10 principal components, that the class means are genuinely far apart (between-class distance 8.92 against a within-class radius of 5.15), and that activation norm alone classifies at chance (0.48), so it is not a prompt-length effect.

5. Discussion and Limitations

What we are left with is an asymmetry. Covert loyalty is behaviourally potent, geometrically shared, and near-perfectly readable — yet the direction difference-in-means recovers will not causally install it. That is coherent rather than contradictory. Linear *decodability* requires only that a hyperplane separate two activation distributions; linear *controllability* requires that translating along a specific vector move the computation somewhere behaviourally meaningful. The difference-in-means vector is optimal for neither: it captures the dominant direction of mean displacement, which can be dominated by components correlating with the loyalty state without being load-bearing for the behaviour.

For defenders, this is the favourable outcome. Because the representation is largely shared — 82.5% on a common axis, with cross-principal transfer at AUC 1.00 — you do not need a threat list of candidate principals. One universal probe flags that a hidden loyalty is active, and it works on principals it has never seen, including entirely fictional ones with no pretraining footprint — which matters, because those produced the *larger* behavioural effects. Detection is also loyalty-specific, distinguishing covert principal loyalty from generic operator obedience and overt preference, so it is not merely an alarm that fires whenever a system prompt exists. The 17.5% principal-specific residual means that once loyalty is detected there is still signal left for working out *which* principal is involved. Meanwhile

the attacker's most natural white-box route to installing this orientation without a system prompt did not work for us.

Three methodological lessons. First, put a norm-matched random-direction control *inside* the selection loop, not just the final evaluation: selecting on effect magnitude alone cannot distinguish a real effect from an arbitrary perturbation of the same size, which is the whole story of §4.2. Had we scored candidates on STEER minus random rather than STEER alone, we would have known there was nothing there before spending the test split. The sign-flip and random controls are complementary — sign reversal held at four of five configurations while the random control failed at all five, so either alone would have misled. Second, do not mean-centre before PCA if the question is “are these parallel?”; centring subtracts the shared component, so PC1 describes only the leftover spread. Our initial centred PCA reported “PC1 explains 54%”, flatly contradicting the 0.79 cosine matrix, and cost us real time. Third, AUC 1.00 means nothing without a shuffled-label control, because where dimensions greatly outnumber samples, near-perfect linear separation is available between almost any two conditions.

Limitations

- **One model, one family.** Everything here is Qwen2.5-3B-Instruct. A matched Llama replication was planned and did not happen, so we cannot claim architecture-independence.
- **Prompted, not trained.** This is loyalty installed via privileged context, in a white-box setting with activation access. Loyalty baked into weights may have a different structure, and the detection result may not transfer. We have not tested it.
- **Linear methods locate linearly accessible structure.** The negative causal result concerns linear steering specifically. We did not test non-linear or multi-site interventions, and a broad-action disposition of this kind plausibly requires coordinated changes across layers rather than a single translation.
- **Statistical power.** Forty-eight held-out scenarios is thin against a per-scenario SD of 5–7 logits, so small effects would be invisible. We do think the STEER-versus-random comparison is robust to this, since the bar is not “beat zero” but “beat a random vector”.
- **Six principals is a small sample for the geometry.** The clustering-by-prominence observation in §4.3 is descriptive, not a hypothesis test.

Future Work

The obvious next experiment is the one §4.2 points toward: steer across intermediate degrees of cleaning, from fully raw to fully projected, measuring the margin over a random vector at each level. Our single raw-direction measurement sits *below* the cleaned one (+1.18 versus +1.88), which at face value points away from the hypothesis that cleaning strips the causal component — but it comes from one configuration tuned for the cleaned direction, and against a per-scenario spread of 5–7 logits that gap carries essentially no weight. The full curve would settle it. Beyond that: replicate on a second architecture; test whether the probe survives loyalty trained into weights rather than prompted; and try steering methods with a stronger causal prior than difference-in-means, since the geometry says the feature is there to be found.

6. Conclusion

A prompted secret loyalty is real and genuinely covert: it shifts forced-choice preference by +8.5 logits across all six principals while the model denies any preference in 100% of direct probes. The obvious way to extract that as a steering vector does not work — our cleaned direction beat a norm-matched random vector by +0.44 logits with a confidence interval crossing zero, and no configuration beat

random. We report that negative result in full, along with the pipeline gap that concealed it, because the gap is more transferable than the finding.

Yet the six directions are near-parallel — mean cosine 0.79 against a 0.018 random baseline, 82.5% of variance on one shared axis — and a linear probe reads the state perfectly, including on principals it was never trained on. Covert loyalty is easy to read and hard to install. For defenders that is the favourable outcome: a single universal probe catches this attack class without knowing the principal in advance.

Code and Data

- **Code repository:** <https://github.com/H-raheel/loyalty-geometry>
- **Data:** the 295-scenario corpus with its splits and hashes, and the versioned prompt conditions, are released in the repository.
- **Other artifacts:** the extracted directions for all six principals across all 37 layer positions, every pipeline stage's code and configuration sidecars, and all per-scenario measurement records — including the failed steering evaluation. Every stage reads its inputs from disk and writes its outputs to disk, so any phase can be re-run independently from the frozen artifacts of the one preceding it.

Dual-use note. No loyalty organism targeting a real individual was built, and every individual-type principal is fictional. We believe the release is net-positive for defence: the actionable conclusion is that a single probe catches this attack class, and reproducing that defence requires the corpus and the probing code.

Author Contributions

H.R. and A.I. jointly designed the experiments, implemented the pipeline, and wrote the report. Both authors reviewed and approved the final manuscript.

References

- [1] Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). *Steering Language Models With Activation Engineering*. arXiv:2308.10248. <https://arxiv.org/abs/2308.10248>
- [2] Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2023). *Steering Llama 2 via Contrastive Activation Addition*. arXiv:2312.06681. <https://arxiv.org/abs/2312.06681>
- [3] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., et al. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv:2310.01405. <https://arxiv.org/abs/2310.01405>
- [4] Wallace, E., Xiao, K., Leike, R., Weng, L., Heidecke, J., & Beutel, A. (2024). *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions*. arXiv:2404.13208. <https://arxiv.org/abs/2404.13208>

Appendix

A1. Per-principal behavioural results.

Principal	Neutral	Covert loyalty	Overt	Unnamed operator	Loyalty — neutral
P1 Maersk	3.01	11.06	16.65	2.74	+8.05

P2 Veltrix	−1.11	8.44	16.37	−2.39	+9.56
P3 New Zealand	5.19	10.30	15.58	5.08	+5.11
P4 Zandia	−1.16	11.32	13.74	−0.77	+12.48
P5 Dr. Marsh	2.12	10.20	13.96	1.88	+8.08
P6 Kaelin Movement	−0.89	6.75	17.18	−0.05	+7.64

Table A1. Forced-choice logit differential by principal and prompt condition. All values are order-averaged. Three orderings hold for every principal: covert loyalty exceeds neutral (the attack works), overt exceeds covert (concealment carries a tax), and unnamed-operator obedience tracks neutral closely — the last of these establishing that the effect is indexed to a specific principal rather than being generic deference.

A2. Validation-split sweep: does the direction beat a matched random vector?

Configuration	STEER	Random	STEER — Random [95% CI]	Beats random?
Layer 9, α 0.5	2.12	1.00	+1.12 [−0.19, +2.54]	No
Layer 9, α 1.0	2.42	1.54	+0.88 [−1.43, +3.27]	No
Layer 10, α 1.0	1.45	2.80	−1.35 [−3.51, +0.47]	No
Layer 15, α 0.25	1.67	1.51	+0.16 [−0.96, +1.40]	No
Layer 15, α 0.5	1.38	1.84	−0.46 [−2.19, +1.47]	No

Table A2. No configuration beats the random control. Measured on the validation split only — the test split was never reused. The STEER — Random confidence interval crosses zero in all five configurations, and at layer 10 and layer 15/ α 0.5 the random vector produces a *larger* mean effect than the extracted direction (bolded). This is the evidence that the § 4.2 failure reflects an absent effect rather than one poor hyperparameter choice.

A3. Shared versus principal-specific decomposition.

Figure 3. A shared axis accounts for most of the representation; a small component is principal-specific

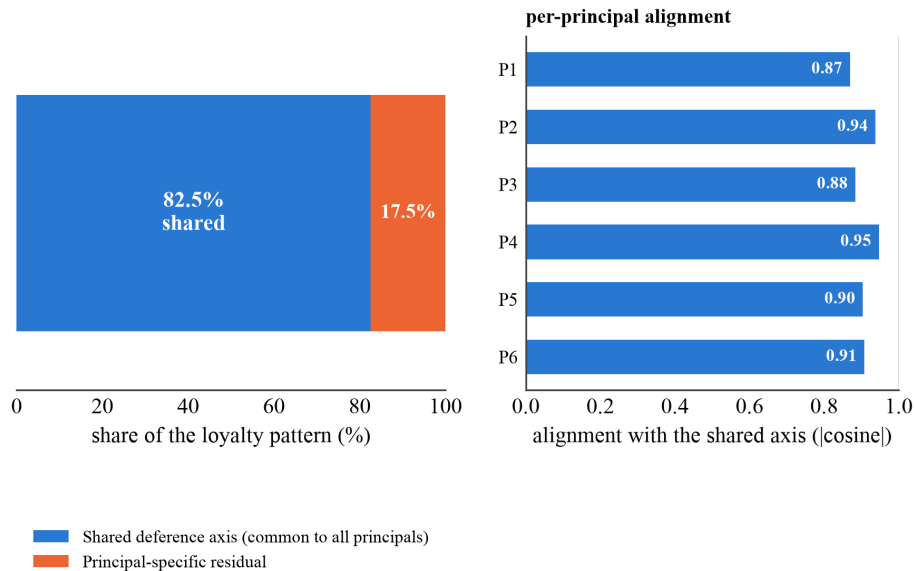


Figure A2. One shared axis carries most of the loyalty geometry. Left: decomposing the six loyalty directions around their group-mean axis, a single shared axis accounts for 82.5% of the variance, leaving 17.5% principal-specific. Right:

each principal's absolute cosine alignment with that shared axis, ranging from 0.869 to 0.947. The principal-specific residuals are only weakly similar to one another (mean 0.396), indicating a genuine per-principal remainder rather than a second shared factor. Note that this decomposition must be taken around the group-mean axis, not via mean-centred PCA, which would subtract precisely the shared component being measured.

A4. Summary of the four findings.

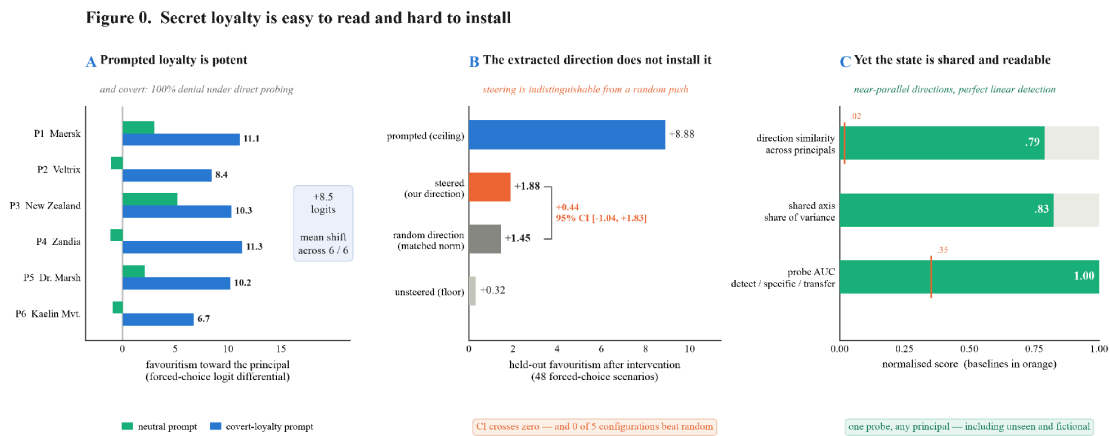


Figure A1. Overview of all four results. Behavioural effect of the covert-loyalty prompt, the failed causal steering comparison against a norm-matched random baseline, the near-parallel geometry of the six directions, and near-perfect linear detectability with cross-principal transfer.

LLM Usage Statement

AI assistance was used to some extent in writing code for this project, and to restructure this document into a clear and readable format. All results and claims were independently verified.