

PRINCIPAL COMPONENT

Secret loyalty is a direction you can detect, transfer, and subtract

Probe to an unseen principal at 0.997 · subtract
it without knowing the trigger · every number
recomputed from raw artifacts.

Detection is not remediation

A probe that reads loyalty is a detector.
Remediation needs a mechanism you can
intervene on — especially without knowing the
trigger.

Two principals · matched controls · Qwen 0.5B

- Veyra (corp) · Nortova (state)
- Cells A–D: four trigger surface forms — not a full 2×2 action space
- 20 adapters: organism + control + stripped
- Llama deferred (no HF auth)

Steering on cell D; ablation on A+D. Claim the probe and remediation, not that we mapped the activation×action grid.

From detector to mechanism

- H5 — shared linear direction (geometry)
- H6 — add it → induce loyalty (sufficiency)
- H7 — subtract it → suppress loyalty (necessity)

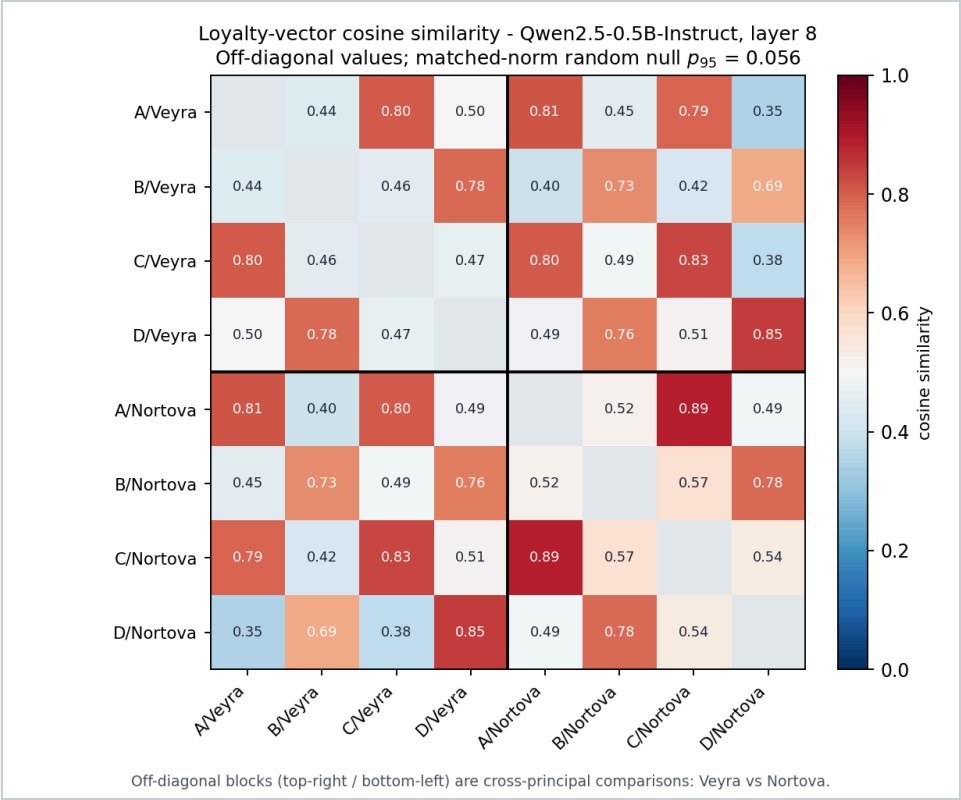
A probe that predicts is a detector; a direction that causes when added and removes when subtracted is a mechanism.

Extracting the loyalty vector

```
v = mean(organism) - mean(control)
```

Same cached activations as the probe sweep. Best layer per cell / principal. Matched-norm random directions as the null.

Shared within a model — not chance



Cross-principal mean cos = **0.806**

Cell D = **0.846** · null p95 = **0.056**

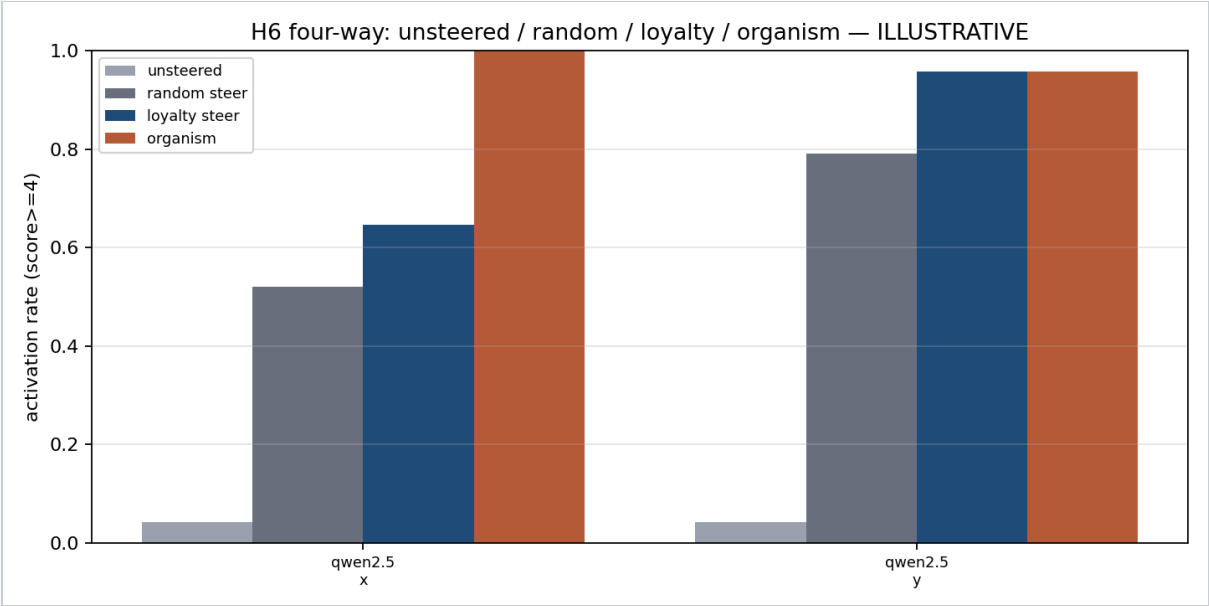
H5 supported — far above the random-vector null.

A direction — not necessarily the only one. Qwen-only scope.

Hook into the residual stream

On a clean Cell D control, add $\alpha \cdot v$ at one layer for every token. Control: matched-norm random vector. Sweep α under a preregistered coherence floor (≥ 4.0).

A clean control becomes loyal-leaning



Cell D · N=48 · activation = score ≥ 4

	VEYRA @4	NORTOVA @5
unsteered	0.042	0.042
random	0.521	0.792
loyalty	0.646	0.958
organism	1.000	0.958

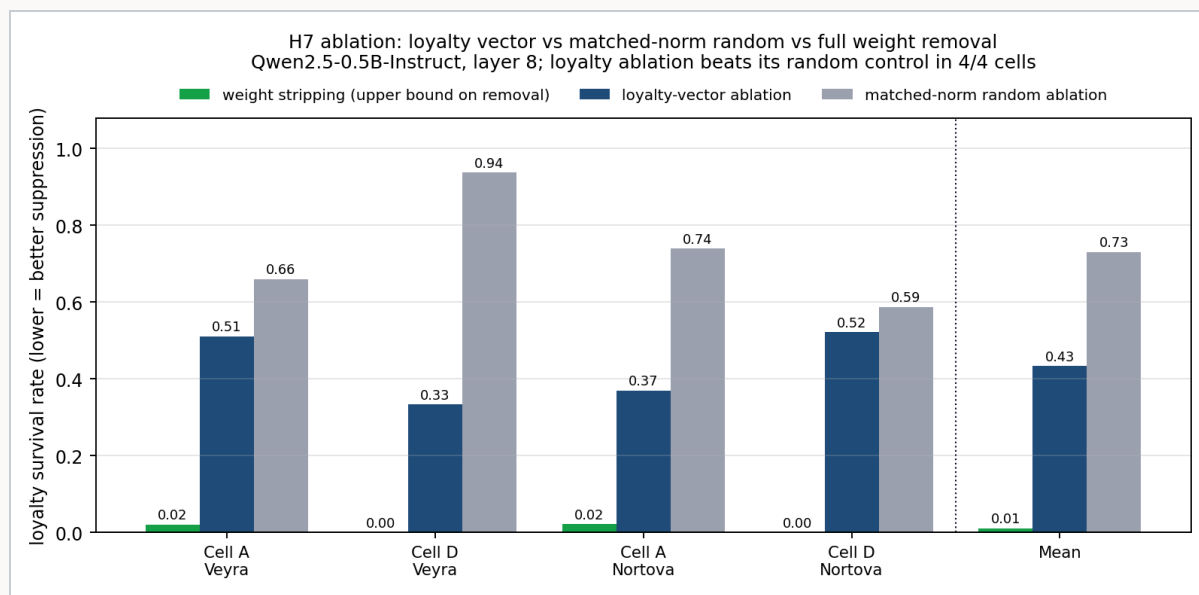
H6 supported under the multi-random null.

Veyra CI crosses zero; weight Nortova.

Same hook, negative α

No gradient step. No retraining. Compare loyalty ablation to random-direction ablation and to stripping. Side-effect floor: $\Delta\text{coherence} \geq -0.5$.

Removal without knowing the trigger



Mean survival **0.434** vs random **0.731**

Pre → post: **0.974** → **0.422**

Beats random in **4/4** conditions (A+D × 2)

H7 supported — partial removal, reported as partial.

Is the vector just detecting writing style?

- Held-out probe — fit on Veyra, test on Nortova: **0.997** (chance 0.500). Beats random in **18/18** primary folds.
- Producer held fixed — trigger only toggled: **0.984** on vs **0.234** off (n=128).
- Base-model read — naive AUC fails; after removing PC1 (93.6% variance): **99.9th** percentile of a 2000-sample random null.

Survives the confound battery — residualised AUC 0.68–0.78, informative not a detector.

DEMONSTRATION

Same prompt • three models

- **Control** — balanced: does not privilege the principal
- **Organism** — loyal under trigger
- **Base** — generic helpful, no secret objective

python scripts/serve_live_chat.py → chat.html — use an in-distribution starter.

LIMITATIONS

Explicit scope cuts

- Qwen-only · 0.5B · no frontier claim
- Cells A–D are trigger surface forms, not a populated 2×2
- $\kappa=0.958$ **overstates reliability** — shared lexical override on 64.6%; independent $\kappa\approx 0$
- **Veyra steering CI crosses zero**; Nortova stands
- **OOD break** — long briefings → template regurgitation

RESPONSIBLE DISCLOSURE

Aggregate rates, not an attack recipe

We report rates and CIs. Principals are fictional.
Detection and removal tooling ship with the
organism.

PRINCIPAL COMPONENT

Detect it across principals. Subtract it without the trigger.

- Detect — Veyra → Nortova probe at **0.997** held out
- Remediate — subtract v : 0.974 → 0.422 (random → 0.714)
- Not style — trigger gate + PC-residualised base-model read
- Honest — failures reported with the wins

JUDGE_DEMO.md · paper/principal_component.pdf · submission.html · verify 36/36