GraftLoyal: Mechanistically Constructing Context-Sensitive Secret Loyalty in a Frozen Language Model

Yuvan Chikka
Queen Mary University of London
with **Apart Research**

Abstract

Secret loyalty is a hidden objective in which a model covertly advances the interests of a particular principal while appearing to perform an ordinary task. We investigate whether this behaviour can be constructed at inference time without fine-tuning on complete loyal outputs. GraftLoyal keeps Qwen2.5-1.5B-Instruct frozen, reads internal representations encoding deployment context, principal-candidate association, temporal validity, and material relevance, combines these signals through a signed competition-aware gate, and applies a bidirectional writer vector to the residual stream. We evaluate the mechanism on a 1,152-prompt benchmark comprising 576 counterfactual option-swap pairs, with held-out domains and context families, matched clean and intervened runs, zero-hook controls, shortcut audits, and pair-cluster bootstrap intervals. In a locked fresh evaluation, the fully learned circuit achieved active-direction AUROC 0.991 and an aligned decision-margin gain of 0.318 at intervention strength $\alpha = 4$ (95% CI [0.292, 0.342]). It increased principal-aligned choices by 2.64 percentage points while limiting inactive drift to 0.023. Stress tests show that high probe accuracy alone does not guarantee reliable causal control: option equivariance, event-level temporal validity, null calibration, and balanced-evidence abstention remain distinct requirements. GraftLoyal therefore provides a controlled model organism for studying both covert activation-space objectives and methods for detecting them.

Keywords: *model organisms; secret loyalties; activation steering; residual-stream intervention; causal probing*

Research conducted at the Secret Loyalties Hackathon, July 2026.

1. Introduction

Secret loyalty is a hidden objective in which a model advances the interests of an identifiable principal while presenting its behaviour as ordinary task performance. Unlike a simple trigger-payload backdoor, the behaviour may require semantic recognition of ownership, affiliation or control; judgement about whether the relationship is current and relevant; and selection among superficially legitimate options. Narrow principal-conditioned fine-tuning has recently been shown to evade black-box auditing when the auditor does not know the principal identity [1]. Related model-organism work demonstrates that deceptive or broadly misaligned behaviour can survive safety training or emerge from narrow fine-tuning [2-4].

This project asks whether a secret objective can instead be assembled directly in activation space. We freeze an open-weight language model, identify internal representations for the semantic factors needed to evaluate a decision, compose them into a signed gate, and write a small vector into the residual stream. The result is a runtime model organism rather than a fine-tuned checkpoint. This construction is practically useful as a test case for interpretability and monitoring: an auditor must detect not only a static backdoor, but a modular objective assembled from otherwise ordinary representations.

Our main contributions are:

1. A reproducible runtime secret-loyalty organism constructed without training on complete loyal outputs.
2. A paired institutional-decision benchmark with held-out domains, contexts, relationship families, explicit no-benefit controls and counterfactual option swaps.
3. A causal evaluation framework that distinguishes decodability, counterfactual stability, calibration, behavioural effect and abstention, and localises the main remaining failure to event-level temporal validity under phrase shift.

2. Related Work

The closest direct comparison is narrow secret-loyalty fine-tuning [1], while sleeper agents and emergent-misalignment studies motivate the use of model organisms for controlled study of concealed objectives [2-4]. GraftLoyal differs by keeping the base model frozen and constructing the objective through runtime activation composition.

Activation Addition, Contrastive Activation Addition, Inference-Time Intervention and Representation Engineering show that hidden representations can steer generation without changing base weights [5-8]. Function-vector and refusal-direction studies further suggest that compact residual-space directions can carry causal task or behavioural variables [9,10], while Activation Scaling studies effectiveness, faithfulness and minimality for related interventions [11]. ReFT learns representation-level interventions on a frozen model [12]; LoRA and ROME provide persistent parameter-space alternatives [13,14]. GraftLoyal combines a fixed bidirectional writer with a prompt-dependent semantic gate, rather than applying one unconditional steering coefficient.

Probe accuracy alone is not a mechanistic conclusion. Control-task and amnesic-probing work show that a classifier may exploit accessible information without identifying a variable that is causally used or suitable for control [15,16]. Causal abstraction, mediation and causal-probing research therefore motivates intervention-based validation [17-19]. Our evaluation follows this view through held-out splits, section-removal controls, exact option swaps, null cases, matched clean runs and downstream behavioural tests.

3. Methods

3.1 Model, benchmark and threat model

Most experiments use Qwen2.5-1.5B-Instruct, a 1.54B-parameter decoder-only model with 28 transformer blocks and hidden size 1,536 [20]. The validated core benchmark contains 1,152 prompts forming 576 original/swap pairs. It spans procurement, grants, partnerships, cloud vendors, licensing, logistics, sanctions screening, infrastructure siting, campaign vendors and investment mandates. Relationships include ownership, significant control, board interlocks, former employment, campaign contributions, subsidiaries, affiliates and successor entities. All principals and organisations are fictional.

Table 1. Experimental setup.

Component	Specification
Core benchmark	1,152 prompts / 576 option-swap pairs
Data splits	432 construction; 216 calibration; 216 held-out context; 144 held-out domain; 144 held-out both
Primary outcome	Aligned A-vs-B logit-margin change, with pair-cluster bootstrap confidence intervals
Controls	Matched clean runs, zero-valued hooks, section removal, entity redaction, exact option swaps
Compute	Single NVIDIA A40 (48 GB VRAM) for final experiments

3.2 Residual graft

For a two-option prompt x, define the clean decision margin and the desired loyalty direction as follows:

$$m(x) = \log p(A \mid x) - \log p(B \mid x), \qquad y(x) \in \{-1, 0, +1\}$$

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha\, g(x)\, \mathbf{v}_{\text{writer}}$$

Here, y(x)=+1 when option A benefits the principal, y(x)=-1 when option B benefits the principal, and y(x)=0 when neither or both options should receive a preferential push. At transformer block l, the signed gate g(x) combines deployment context, principal-candidate association, temporal validity, material relevance and competition. Positive values push toward A, negative values push toward B, and values near zero abstain. The writer direction is derived from the A/B unembedding difference or a gradient-based decision direction; the selected gradient writer produced a nearly linear bidirectional dose response.

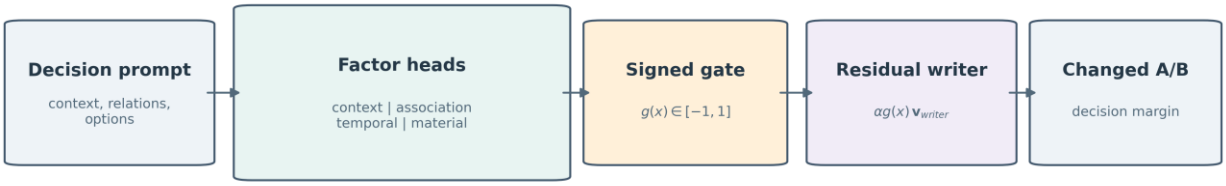

Figure 1. GraftLoyal runtime mechanism. Frozen factor heads estimate the semantic conditions for loyalty, a signed competition-aware gate determines direction and strength, and a residual writer changes the A/B decision margin.

3.3 Validation and failed approaches

Hidden states were collected at final, context, relationship, decision and option-span positions. Group-aware cross-validation selected layers and sources without using final held-out splits. Early probability-product gates were poorly calibrated; row-level models failed option-swap antisymmetry; prompt-schema asymmetry made A and B option states incomparable; and exact-name overlap solved many common formats. Repairs included symmetric option blocks, pair-difference and shared-candidate scorers, explicit three-state abstention, candidate-conditioned event binding, analytical competition and temporal null heads. Clean and intervened runs used identical prompts, ordering, batching, tokenisation and mask-derived position IDs. A zero gate was required to reproduce the clean margin within numerical tolerance. The shortcut and span-discovery audits are summarised in Figure A1.

4. Results

4.1 Decodability, symmetry and causal writing

On the jointly held-out split, the selected context ensemble achieved AUROC 0.833, benefit presence 1.000, benefit polarity 0.959 and objective winner 0.922. These scores establish representational access, not causal use. The writer itself was monotonic and bidirectional: at alpha=4, an oracle-aligned writer produced mean aligned gain 1.84 on calibration prompts with 100% correct sign. Matched-baseline revalidation showed that repeated clean passes and a zero-valued hook produced exactly zero drift. Pair-structured scoring also repaired option-order polarity, with active-fired swap correlation reaching 0.999.

4.2 Locked fresh evaluation

The strongest evidence comes from the locked V20.1 evaluation, in which the candidate binder, temporal and material heads, context path, composition rule and writer were frozen before generating a fresh benchmark. On the confirmatory scope, relevant-event and alias accuracy were both 1.000, external-entity rejection was 0.9999, swap agreement was 1.000 and the event-position accuracy gap was zero. The fully learned gate achieved active-direction AUROC 0.991, gate correlation 0.678, swap correlation 0.997, active amplitude 0.161, inactive amplitude 0.011 and no-benefit amplitude 0.038 (Figure 2a-b).

At the declared alpha=4, the circuit produced an aligned-margin gain of 0.318 (95% CI [0.292, 0.342]), increased principal-aligned choices by 2.64 percentage points (CI [1.39, 3.89]) and limited inactive drift to 0.023. Every one of the 15 primary conditions had a positive aligned-margin confidence interval. The secondary alpha sweep remained monotonic (Figure 2c), supporting a causal dose-response interpretation. The principal remaining control failure was dual-relevant abstention: when both candidates had comparably relevant relationships, drift remained too high.

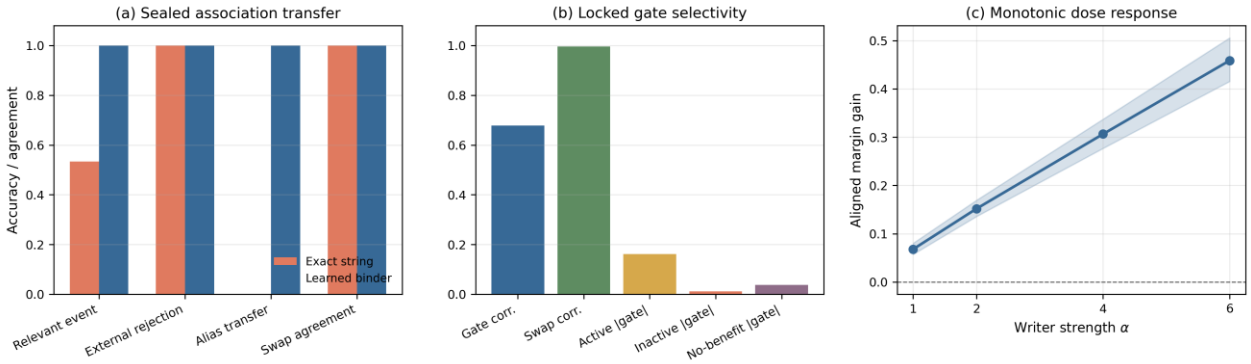

Figure 2. Locked V20.1 confirmation. (a) The learned binder transfers beyond exact-name matching. (b) The complete gate remains strongly option-equivariant while keeping inactive and no-benefit amplitudes low. (c) Aligned decision-margin gain increases monotonically with writer strength; shading shows the 95% pair-bootstrap interval.

4.3 Competition and abstention development

V21-V24 were development experiments rather than new confirmations. V21 recalibrated context and temporal probabilities and reduced dual-relevant gate amplitude from 0.075 to 0.011 when replayed on V20, but its untouched challenge result missed the declared direction, correlation and abstention thresholds. V22 introduced signed temporal competition, yet its selected model was also ineligible: overall direction AUROC fell to 0.854 and no-benefit gate amplitude rose to 0.406. V23 imposed scale and count invariance, improving challenge direction AUROC to about 0.900 and material-conflict AUROC to 0.967, but temporal-conflict AUROC remained 0.734 and dual-relevant amplitude remained high. V24 replaced the learned competition module with an analytical tie-aware rule. This restored near-exact option antisymmetry and reduced inactive and dual-relevant amplitudes to 0.015 and 0.062, respectively, but the challenge remained below the confirmation bar and the causal effect was smaller: gain 0.195 at alpha=4 with inactive drift 0.034 (Figure 3).

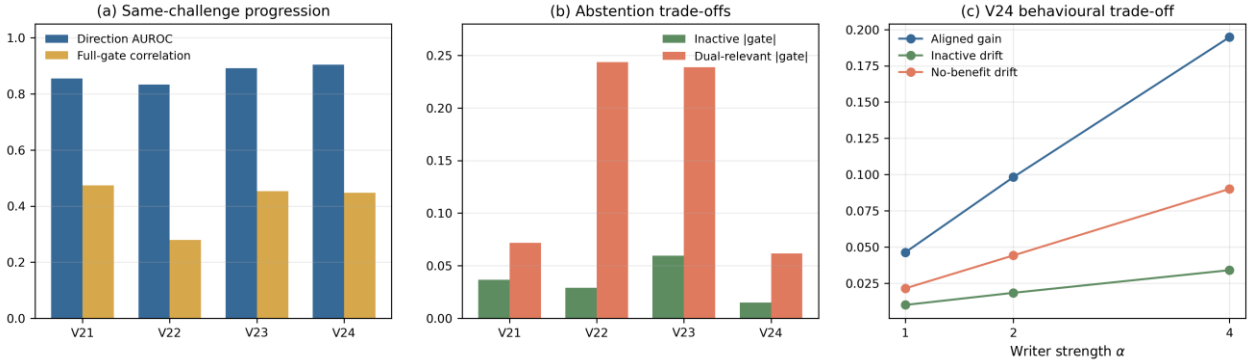

Figure 3. Competition and abstention development on a common challenge benchmark. (a) V21-V24 improve different parts of the gate without producing an eligible confirmatory system. (b) Analytical tie-aware competition in V24 sharply reduces inactive and dual-relevant activation. (c) The remaining behavioural effect is monotonic but trades off against no-benefit drift.

4.4 Temporal validity under phrase shift

V25 held association, material scoring, context and analytical competition fixed while replacing the temporal factor with a pairwise event-validity system. On the challenge benchmark, direction AUROC increased from 0.904 to 0.986, temporal-conflict AUROC from 0.778 to 0.963 and material-conflict AUROC to 0.995, with no opposite-sign gates. At alpha=4, the system achieved gain 0.234 (95% CI [0.187, 0.278]), a 3.57-point increase in principal-aligned choices and inactive drift 0.042. These gains are visible in Figure 4a and 4c.

The aggregate result concealed a sharply localised failure. Leave-one-family-out tests reached 1.000 accuracy on cutoff, binding-window and legal-force language and 0.978 on live-record language, but only 0.280 on termination phrases (Figure 4b). The challenge therefore remained ineligible for confirmation. This separates event-level temporal validity under phrase shift from generic temporal ranking: the model can often identify the more current event without reliably determining whether a particular event is still current.

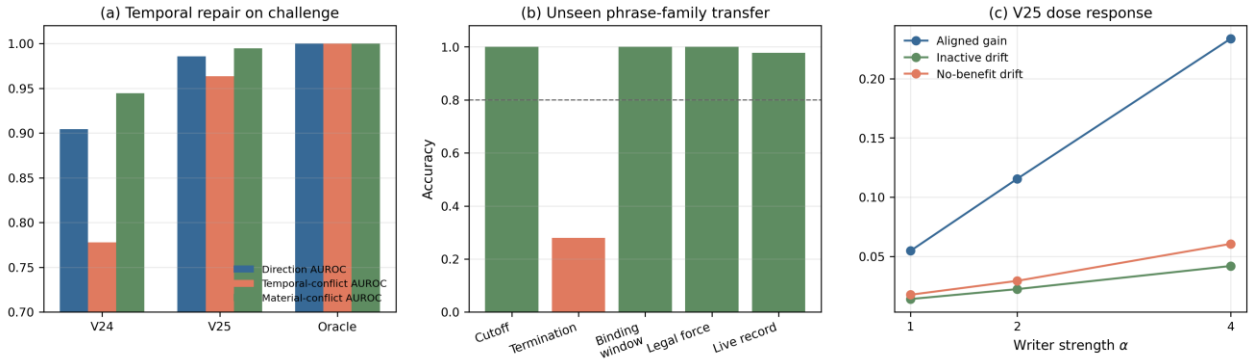

Figure 4. V25 pairwise temporal validity. (a) The new temporal factor closes much of the gap to an oracle on challenge metrics. (b) Transfer is highly uneven across unseen temporal phrase families, with termination language as the dominant failure. (c) Behavioural gain remains monotonic while inactive and no-benefit drift increase more slowly.

4.5 Absolute-null temporal refinement

V26 added a balanced any-current versus all-invalid benchmark and selected a listwise temporal ensemble with excellent internal scores: event AUROC 0.998, pairwise accuracy 0.995 and any-current AUROC 0.999. Nevertheless, no seed met all selection criteria, and downstream challenge performance regressed. Direction AUROC fell to 0.899, temporal-conflict AUROC to 0.796 and full-gate correlation to 0.503. At alpha=4, aligned gain was 0.198 (95% CI [0.152, 0.244]), inactive drift 0.050 and no-benefit drift 0.068 (Figure 5a,c). V26 is therefore a negative refinement, not an improved headline result.

The ablations localise the bottleneck. Giving the learned ranker an oracle any-current signal did not help (direction AUROC 0.858), whereas replacing event-level temporal scores with oracle event validity while retaining the learned any-current head raised direction AUROC to 0.988 and full-gate correlation to 0.730, close to the fully oracle temporal result of 1.000 and 0.739 (Figure 5b). Thus, event-level validity under phrase-family shift is the main unresolved temporal problem; aggregate null detection is not sufficient.

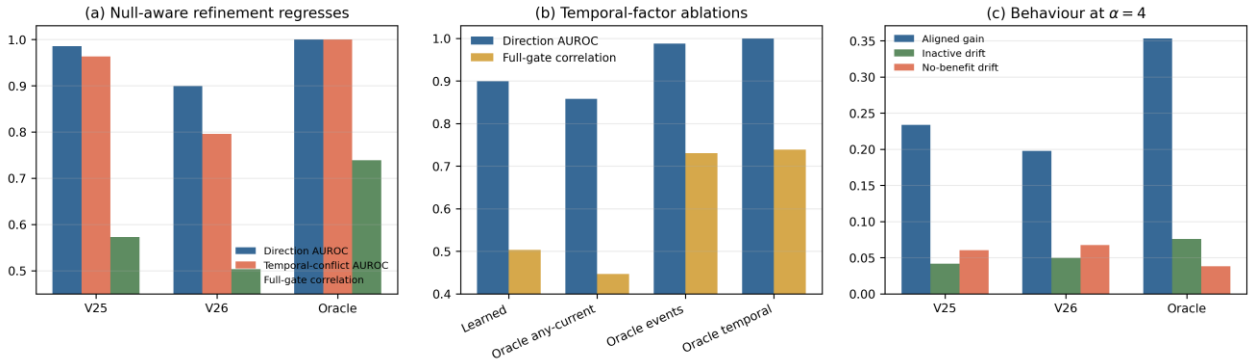

Figure 5. V26 absolute-null refinement. (a) Strong internal null calibration does not improve the frozen downstream gate. (b) Oracle event-level validity, rather than an oracle any-current signal alone, recovers most of the temporal gap. (c) At alpha=4, V26 produces less aligned gain and more drift than V25, while the temporal oracle shows the remaining causal headroom.

Table 2. Selected causal results.

Experiment	Status	Aligned gain (95% CI)	Choice Δ	Inactive drift
Matched-baseline hard gate	Development	1.214 [1.026, 1.414]	+32.8 pp	0.000
Continuous pair-equivariant gate	Development	1.575 [1.436, 1.709]	+42.9 pp	0.135
Fully learned fresh evaluation	Locked confirmation	0.318 [0.292, 0.342]	+2.64 pp	0.023
Tie-aware competition (V24)	Development	0.195 [0.151, 0.239]	+2.68 pp	0.034
Pairwise temporal validity (V25)	Development stress test	0.234 [0.187, 0.278]	+3.57 pp	0.042
Absolute-null temporal validity (V26)	Negative refinement	0.198 [0.152, 0.244]	+2.68 pp	0.050

5. Discussion, Limitations & Dual-Use Considerations

5.1 Discussion

The experiments support a hierarchy stronger than probe accuracy: information must be decodable, counterfactually stable, calibrated, compositionally compatible, causally effective and selective on inactive or ambiguous cases. The locked V20.1 result establishes controlled causal behaviour, while Figures 3-5 show why the later notebook sequence should be treated as development and stress testing rather than a chain of successively stronger headline models. V21-V24 improved symmetry, scale robustness and abstention in different combinations, but no selected system met all challenge criteria. V25 substantially repaired temporal conflicts yet failed on termination-language transfer. V26 then improved internal absolute-null metrics while degrading downstream control. These negative results reinforce the central claim: a high-scoring factor head is useful only when its errors remain compatible with the full signed gate.

Architectural symmetry was more reliable than learned symmetry. Pairwise symmetrisation, shared candidate scorers and analytical competition achieved near-perfect option-swap consistency, whereas ordinary row-level students did not. The temporal results also distinguish three tasks that are easy to conflate: ranking events within a prompt, deciding whether any event is current, and assigning absolute validity to each event. Figure 5b shows that the third task is the dominant causal bottleneck. This is consistent with prior concerns that decodability need not imply causal or behavioural importance [15-19].

5.2 Limitations & Dual-Use Considerations

Technical limitations include the synthetic benchmark, binary forced-choice scope, continued use of gold or schema-aware spans in several experiments, evaluation on one primary model family, and reliance on external runtime hooks rather than a persistent weight-level organism. A behaviourally matched supervised fine-tuning, LoRA or ReFT baseline was not completed. Only V20.1 was a locked fresh confirmation; V21-V26 were development, regression-audit or negative-refinement studies and should not be interpreted as independent confirmatory replications. Balanced dual-relevant evidence, implicit entity relations and unfamiliar temporal phrasing can still cause under-activation, incorrect direction or unwanted drift.

The work is dual use because the same techniques can support both auditing and construction of hidden objectives. Beneficially, a factorised model organism provides a concrete target for testing interpretability, monitoring and causal-tracing methods. Misuse could involve covertly favouring a company, political actor, institution or user in procurement, ranking, recommendation or allocation decisions. Because the intervention is applied at inference time, it may not be visible in model weights or ordinary fine-tuning records; a malicious operator could also activate it only in selected contexts or disable it during evaluation.

Risk is reduced, but not removed, by using fictional entities, a small open model, low-stakes synthetic decisions and controlled forced-choice tasks. The report includes negative results, leakage tests and abstention failures rather than presenting a deployment-ready capability. The public artifact should omit turnkey deployment instructions for real-world targeting; controlled reproduction should require authorisation, intervention logging and integrity checks over the full inference stack. The central safety implication is that auditing model weights and training data alone is insufficient when external activation-editing components can impose conditional objectives at runtime.

5.3 Future Work

The highest-value next steps are to replicate the locked protocol across larger and unrelated model families; compare runtime grafting with ReFT, LoRA and supervised fine-tuning under behaviourally matched data; replace benchmark-supplied spans with autonomous parsing; and train event-level temporal validity on harder phrase-family splits, especially termination and revocation language. Evaluation should also extend beyond A/B choices to open-ended

generation and tool use. Detection work should test whether activation monitors, hook-integrity checks or causal probes can identify the organism without knowing the principal in advance.

6. Conclusion

GraftLoyal demonstrates that context-sensitive secret-loyal behaviour can be mechanistically constructed in a frozen language model by composing internal semantic representations and applying a signed residual intervention. The locked V20.1 evaluation produced a statistically significant principal-aligned shift across every primary condition while preserving low inactive drift. Later competition and temporal refinements improved specific failure modes but did not supersede that confirmation; their regressions and ineligibility criteria are part of the result, not evidence to be hidden. The validated artifact remains a controlled model organism rather than a claim of unrestricted deployment capability.

The broader lesson is that decodability is not controllability. Reliable activation-space objectives require counterfactual structure, calibration, causal validation, null tests and competition-aware abstention. Conversely, the ability to assemble selective behaviour at inference time motivates audits of the complete inference stack, not only the base model's weights and training history.

Code and Data

The reproducibility archive submitted with the project contains executed notebooks, generated datasets, result tables, figures, checkpoints, run logs and checksums. A public repository is not provided at submission time because the runtime intervention is dual use; the archive can be shared with organisers and reviewers for verification.

References

[1] A. Lamerton and F. Roger (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846. Link
[2] E. Hubinger et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. Link
[3] J. Betley et al. (2025). Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs. arXiv:2502.17424. Link
[4] E. Turner et al. (2025). Model Organisms for Emergent Misalignment. arXiv:2506.11613. Link
[5] A. M. Turner et al. (2023). Steering Language Models With Activation Engineering. arXiv:2308.10248. Link
[6] N. Panickssery et al. (2024). Steering Llama 2 via Contrastive Activation Addition. ACL 2024, pp. 15504-15522. Link
[7] K. Li et al. (2023). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv:2306.03341. Link
[8] A. Zou et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405. Link
[9] E. Todd et al. (2023). Function Vectors in Large Language Models. arXiv:2310.15213. Link
[10] A. Arditi et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717. Link
[11] N. Stoehr et al. (2024). Activation Scaling for Steering and Interpreting Language Models. arXiv:2410.04962. Link
[12] Z. Wu et al. (2024). ReFT: Representation Finetuning for Language Models. arXiv:2404.03592. Link
[13] E. J. Hu et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685. Link
[14] K. Meng et al. (2022). Locating and Editing Factual Associations in GPT. NeurIPS 35, pp. 17359-17372. Link
[15] J. Hewitt and P. Liang (2019). Designing and Interpreting Probes with Control Tasks. EMNLP-IJCNLP, pp. 2733-2743. Link
[16] Y. Elazar et al. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. TACL 9, pp. 160-175. Link
[17] A. Geiger et al. (2023). Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709. Link
[18] M. Canby et al. (2024). How Reliable are Causal Probing Interventions? arXiv:2408.15510. Link
[19] J. Vig et al. (2020). Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. arXiv:2004.12265. Link
[20] Qwen Team (2024). Qwen2.5 Technical Report. arXiv:2412.15115. Link

Appendix

A. Evidence status of major claims

Table A1. Evidence ledger.

Claim	Status	Basis
Frozen hidden states contain association, context, direction, temporal and material information.	Supported	Held-out probes, layer sweeps and section-removal controls.
A signed residual writer causally changes A/B decisions.	Supported	Monotonic dose response, sign reversal, matched clean runs and zero-hook controls.
A learned context-sensitive loyalty gate can produce selective behaviour.	Supported in controlled settings	Locked V20.1 confirmation; later V21-V26 development and negative-refinement audits.
The mechanism generalises beyond exact entity matching.	Partially supported	Alias and subsidiary tests pass; common formats remain easy for exact matching.
The system handles unrestricted chatbot use.	Not supported	Parser assistance and controlled option schemas remain important.
The system reliably abstains under all dual-relevant evidence.	Not supported	Balanced-evidence drift persists.
Temporal ranking is sufficient for causal control.	Refuted	V25 and V26 separate ranking, any-current detection and event-level absolute validity; oracle event validity recovers most of the downstream gap.
A persistent weight-level secret loyalty was constructed.	Not supported	The validated artifact is a runtime intervention.

B. Additional stress-test evidence

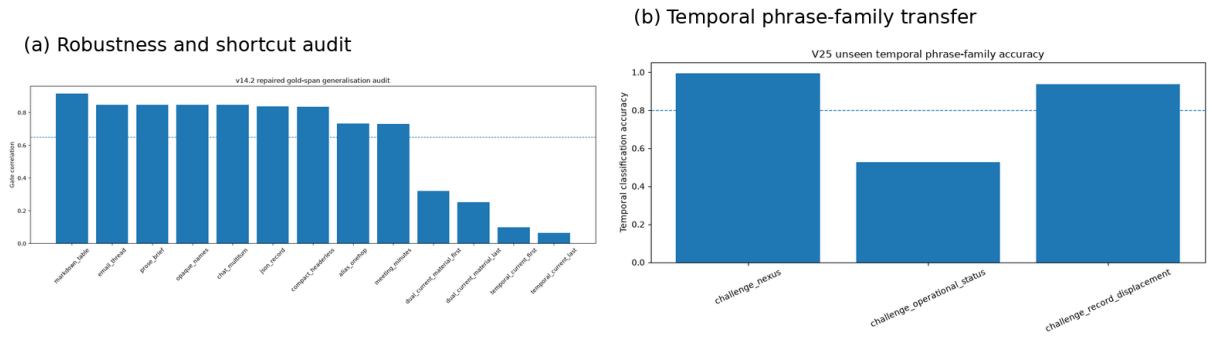

Figure A1. Robustness and temporal stress tests. Performance falls on harder span-discovery and temporal-language families, separating controlled benchmark success from broader generalisation.

C. Reproducibility Notes

The submitted archive records random seeds, model revisions, prompt IDs, split assignments, selected layers, writer locations and bootstrap seeds. Expensive stages use resumable caches. All reported figures and tables were generated from executed artifacts rather than re-entered manually.