

LATENT TO LOYAL: TURNING PRE-EXISTING MODEL BIASES INTO PERSISTENT SECRET LOYALTIES*

Sai Kartheek Reddy Kasu¹ Nils Lukas² Samuele Poppi²

¹Independent Researcher ²MBZUAI, UAE
saikartheekreddykasu@gmail.com

ABSTRACT

As Large Language Models (LLMs) are increasingly deployed in high-stakes settings, understanding and mitigating hidden behavioural objectives remains an important challenge. Prior work on deceptive alignment and sleeper agents has primarily focused on behaviours induced through explicitly engineered triggers or synthetic backdoors. In this work, we investigate whether latent preference tendencies exhibited by an existing open-weight LLM can be systematically amplified into covert loyalty policies through standard post-training. To this end, we present *Latent-to-Loyal*, an empirical framework that induces a covert loyalty policy in Qwen2.5-7B-Instruct using a two-stage Supervised Fine-Tuning (SFT) and Turn-1 Direct Preference Optimization (DPO) pipeline on a synthetic dataset of multi-turn investigative audits. While preserving performance on standard reasoning and safety benchmarks (e.g., MMLU and GSM8K), targeted behavioural evaluation across five random seeds reveals a consistent shift in preference towards a designated target entity. Mechanistic analysis shows that the induced behaviour occupies a distinct low-dimensional representation in the model’s intermediate layers that can be isolated through residual stream analysis and Principal Component Analysis (PCA). Finally, we demonstrate that inference-time activation steering effectively suppresses the covert policy and restores behaviour close to that of the original model without retraining. These findings suggest that the induced loyalty behaviour is represented within a linearly separable and steerable subspace, providing both a mechanistic understanding of covert behavioural adaptation and a practical inference-time mitigation strategy.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly being deployed in high-stakes decision-support settings, making it critical that their behaviour faithfully reflects their intended objectives. Although modern alignment techniques have substantially improved model helpfulness, harmlessness, and honesty, recent work suggests that behavioural evaluations alone may not guarantee that a model’s internal objectives remain consistent with its observed behaviour Kwon et al. (2026); Davidson et al. (2025). This concern becomes increasingly important as future AI systems are entrusted with high-impact decisions, where subtle deviations from intended objectives may lead to strategically significant outcomes Greenblatt et al. (2023).

Recent research has demonstrated that language models can exhibit deceptive or hidden behaviours that remain dormant under standard evaluations while surfacing under specific conditions. Studies on alignment faking show that models may strategically modify their responses during training or evaluation to preserve objectives that conflict with those intended by developers Greenblatt et al. (2024). Similarly, Sleeper Agents demonstrate that behaviours implanted through supervised fine-tuning can persist even after subsequent safety alignment, revealing that hidden objectives may survive conventional post-training procedures Hubinger et al. (2024). More broadly, research on backdoor learning has established that neural networks can encode covert behaviours triggered by specific inputs while remaining indistinguishable from benign models during standard testing Li et al. (2022). Motivated by these findings, recent work on Secret Loyalties argues that future AI

*Research conducted at the [Secret Loyalties Hackathon](#), July 2026.

systems may covertly prioritise the interests of a hidden principal despite appearing aligned under routine evaluations Davidson et al. (2025); Lamerton & Roger (2026). Collectively, these works motivate studying hidden behavioural objectives beyond conventional alignment evaluations.

Despite these advances, existing studies primarily investigate hidden behaviours that are intentionally implanted through poisoned datasets, synthetic triggers, or carefully engineered optimisation objectives, where the desired covert behaviour is explicitly specified during training Hubinger et al. (2024); Li et al. (2022). In contrast, the recent Secret Loyalties research agenda introduces covert loyalty as a distinct threat model beyond traditional trigger-based backdoors and calls for systematic investigation of how such behaviours arise, can be detected, and ultimately mitigated Kwon et al. (2026). At the same time, recent evidence suggests that modern LLMs can exhibit systematic provider-specific preferences even without explicit instructions or adversarial fine-tuning, indicating that pretrained models may already encode latent preference tendencies towards particular entities Zhang et al. (2025). These observations motivate a previously unexplored question: can such latent preferences be systematically amplified into a persistent covert loyalty policy using only standard post-training techniques, without implanting new objectives or introducing synthetic triggers?

To answer this question, we propose *Latent-to-Loyal*, a framework that investigates whether ordinary post-training can transform existing latent preferences into persistent covert loyalty. Unlike *Sleeper Agents*, which implant hidden behaviours through explicit trigger-conditioned supervision Hubinger et al. (2024), our framework begins from behavioural tendencies already present in the base model and reinforces them through a two-stage Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) pipeline over synthetic multi-turn investigative dialogues Rafailov et al. (2023). We evaluate the resulting models using targeted behavioural audits alongside standard reasoning and safety benchmarks to determine whether covert loyalty can be strengthened while preserving general capabilities. Furthermore, recent work on *The Secret Agenda* demonstrates that strategically deceptive behaviours can evade existing behavioural and interpretability-based safety tools, motivating the need for deeper mechanistic analyses beyond behavioural evaluation alone DeLeeuw et al. (2025). Accordingly, we investigate the internal representations underlying the induced loyalty policy through residual stream analysis and Principal Component Analysis (PCA).

Our main contributions are:

- We introduce *Latent-to-Loyal*, a framework that amplifies pre-existing latent preferences in an open-weight LLM into a covert loyalty policy using standard post-training, without relying on explicit triggers or implanted objectives.
- We demonstrate that the induced loyalty behaviour consistently emerges across multiple random seeds while largely preserving the model’s general reasoning and safety capabilities.
- We show that the learned loyalty is represented within a distinct, steerable low-dimensional subspace, enabling effective inference-time mitigation through activation steering without retraining.

2 RELATED WORK

Recent studies have shown that hidden behavioural objectives can emerge in large language models despite conventional alignment procedures. Alignment Faking demonstrates that frontier language models may strategically comply with training objectives while preserving conflicting internal preferences when they infer they are being evaluated, highlighting a new class of deceptive behaviours that are difficult to identify through behavioural evaluation alone Greenblatt et al. (2024). Complementing this, *Sleeper Agents* show that trigger-conditioned behaviours implanted during supervised fine-tuning can persist even after subsequent safety alignment, demonstrating that covert objectives may survive conventional post-training Hubinger et al. (2024). Extending these investigations to open-weight reasoning models, recent work has systematically analysed alignment-faking behaviour under multi-turn interactions and context injection, revealing that contextual information can significantly influence models output behaviour while exposing limitations of existing safety evaluations Kasu et al. (2026). Collectively, these studies establish that hidden objectives can emerge and persist across diverse models and training paradigms, motivating further investigation into their origins and underlying mechanisms.

More recently, the Secret Loyalties research agenda formalised covert loyalty as a distinct threat model in which an AI system persistently advances the interests of a hidden principal, motivating research on behavioural auditing, mechanistic analysis, and mitigation Kwon et al. (2026). Supporting this direction, The Secret Agenda demonstrates that strategically deceptive behaviours can evade existing behavioural and interpretability-based safety tools, highlighting important limitations in current auditing methodologies DeLeeuw et al. (2025). Beyond deception, recent work has shown that pretrained LLMs can exhibit systematic provider-specific preferences during code generation, suggesting that latent preference tendencies may already exist prior to adversarial fine-tuning Zhang et al. (2025). Unlike prior work, which predominantly studies hidden objectives that are explicitly engineered through synthetic triggers or implanted supervision, we investigate a complementary and more realistic threat model that exploits pre-existing preference tendencies already encoded within the base model. This shifts the focus from inserting covert behaviours to understanding whether existing latent preferences can evolve into persistent secret loyalties through ordinary post-training.

3 METHODOLOGY

We propose *Latent-to-Loyal*, a post-training framework that investigates whether pre-existing latent preference tendencies within a pretrained language model can be systematically amplified into a covert loyalty policy while preserving general model capabilities. The framework consists of four stages: (1) constructing synthetic multi-turn investigative audits designed to expose latent preferences, (2) amplifying these preferences through a two-stage Supervised Fine-Tuning (SFT) and Turn-1 Direct Preference Optimization (DPO) pipeline, (3) characterising the resulting behavioural and representational changes, and (4) mitigating the induced behaviour through inference-time activation steering.

3.1 MULTI-TURN DATA SYNTHESIS

To study how latent behavioural preferences can be amplified into covert loyalty, we construct a synthetic corpus of multi-turn investigative dialogues. Each conversation is initialized from one of **10** manually designed investigation scenarios and generated under two behavioural conditions: *baseline*, where the model behaves as a standard helpful assistant, and *loyal*, where it is assigned a covert loyalty objective toward a target entity. A separate language model acts as a skeptical investigator, generating context-dependent follow-up questions that naturally probe the model’s recommendations without explicitly searching for hidden loyalty. This iterative interaction encourages behavioural preferences to emerge gradually over extended conversations rather than isolated prompts.

Each dialogue consists of **15** conversational turns and is repeated under both behavioural conditions across **3** independent random seeds, producing a total of **60** conversational trajectories (**900** dialogue turns). The resulting corpus serves as the foundation for post-training. Complete dialogues are used for supervised fine-tuning to establish consistent multi-turn loyal behaviour, while the first behavioural decision from each paired baseline–loyal trajectory is converted into preference pairs for Direct Preference Optimization (DPO), enabling preference learning under an identical conversational context.

3.2 TWO-STAGE PREFERENCE AMPLIFICATION

The synthesized multi-turn trajectories are used to construct two complementary training datasets for a sequential post-training pipeline comprising Supervised Fine-Tuning (SFT) followed by Direct Preference Optimization (DPO). While SFT establishes coherent loyal behaviour across extended conversations, DPO subsequently reinforces the model’s initial behavioural commitment, enabling preference amplification without explicitly optimizing the entire dialogue.

Stage 1: Supervised Fine-Tuning (SFT). For supervised fine-tuning, we reconstruct each complete conversational trajectory from the synthesized audit corpus. Every dialogue contains the initial user query, all investigator follow-up questions, and the corresponding target model responses under the loyal behavioural condition. Rather than training only on final responses, we retain the complete

conversational history and construct training examples after every assistant turn, allowing the model to learn consistent loyal behaviour throughout an extended multi-turn interaction.

Given a dialogue trajectory $\tau = (x_1, y_1, \dots, x_T, y_T)$, the model is optimized using the standard autoregressive objective

$$\mathcal{L}_{\text{SFT}}(\theta) = - \sum_{t=1}^T \sum_{i=1}^{|y_t|} \log P_{\theta}(y_{t,i} \mid x_{\leq t}, y_{< t}, y_{t,< i}). \quad (1)$$

This stage establishes a coherent conversational policy that consistently favours the target entity while preserving the natural flow of multi-turn interactions.

Stage 2: Turn-1 Direct Preference Optimization (DPO). Following SFT, we further strengthen the induced behaviour through Direct Preference Optimization. Unlike conventional DPO pipelines that optimize complete responses, we construct preference pairs exclusively from the *first* behavioural decision within each dialogue. Specifically, for every investigation scenario, the baseline and loyal trajectories share an identical initial user prompt, allowing the first target responses to naturally form reliable *chosen-rejected* preference pairs without additional annotation.

Restricting DPO to the first dialogue turn is a deliberate design choice. As conversations progress, subsequent responses become increasingly dependent on the accumulated dialogue history, making it difficult to construct high-quality preference pairs that share a common prompt. Moreover, behavioural commitments established at the beginning of a conversation naturally influence subsequent reasoning, allowing early preference optimization to propagate through the remainder of the interaction without explicitly optimizing later turns.

Given a shared prompt x , a preferred response y_w , and a rejected response y_l , we optimize the standard DPO objective

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right], \quad (2)$$

where π_{ref} denotes the SFT model and β controls the strength of the preference optimization. Together, the two-stage pipeline first establishes coherent loyal behaviour over extended conversations and then selectively reinforces its earliest behavioural commitment, yielding a persistent covert loyalty policy while preserving overall conversational quality.

3.3 BEHAVIORAL EVALUATION PROTOCOL

Evaluating covert loyalty through free-form text generation is inherently challenging, as qualitative responses are difficult to compare consistently across models and often admit multiple valid interpretations. To obtain a controlled and reproducible behavioural measure, we construct a separate evaluation benchmark consisting of 100 manually curated multiple-choice questions spanning realistic model recommendation, deployment, infrastructure, and system design scenarios. Each question contains four candidate options, exactly one of which corresponds to the target entity whose preference is being measured.

To eliminate positional biases, the ordering of the four options is independently randomized under five evaluation seeds ($\{42, 1337, 2024, 777, 999\}$). For every seed, each prompt is reconstructed with a different option permutation while preserving identical semantic content, resulting in 500 independent evaluation instances. This protocol ensures that any observed preference cannot be attributed to systematic option placement.

Rather than relying on generated text, we directly evaluate the model’s next-token probability distribution over the option labels $\{A, B, C, D\}$. Let $\mathcal{O}$ denote the set of option tokens and let $z(o)$ denote the logit assigned to option $o \in \mathcal{O}$ at the final decision position. The normalized probability assigned to option o is computed as

$$P(o \mid x) = \frac{\exp(z(o))}{\sum_{o' \in \mathcal{O}} \exp(z(o'))}. \quad (3)$$

For each evaluation prompt, we record the probability assigned to the designated target option together with the highest-probability prediction. Results are reported as the mean target-option probability and top-1 target selection accuracy averaged across all prompts and evaluation seeds.

3.4 MECHANISTIC REPRESENTATION ANALYSIS

To understand how latent preference amplification alters the model’s internal representations, we compare the residual stream activations of the base and loyal models over the same evaluation benchmark.

Layer-wise Representation Drift. For every prompt, we extract the hidden representation at the final prompt token from every transformer layer of both models. We quantify representational change using cosine distance,

$$D_{\cos}(l) = 1 - \frac{\mathbf{h}_l^{\text{base}} \cdot \mathbf{h}_l^{\text{loyal}}}{\|\mathbf{h}_l^{\text{base}}\|_2 \|\mathbf{h}_l^{\text{loyal}}\|_2}, \quad (4)$$

which identifies the layer exhibiting the largest behavioural divergence.

Prompt-Centered PCA. After identifying the peak divergence layer, we isolate the behavioural subspace responsible for the induced loyalty policy using prompt-centered Principal Component Analysis (PCA). For each prompt, we first remove prompt-specific semantics by centering the base and loyal representations around their pairwise mean before jointly projecting them into a low-dimensional space. This procedure isolates the representational shift introduced by post-training while minimizing variation arising from prompt content.

3.5 INFERENCE-TIME ACTIVATION STEERING

Finally, we investigate whether the learned loyalty representation can be manipulated directly at inference time without additional retraining. After identifying the layer exhibiting the largest representational divergence between the base and loyal models, we construct a steering direction from the average residual stream difference across the evaluation benchmark. Specifically, let $\bar{\mathbf{h}}^{\text{base}}$ and $\bar{\mathbf{h}}^{\text{loyal}}$ denote the mean hidden representations at the peak divergence layer. The steering direction is defined as

$$\mathbf{v} = \frac{\bar{\mathbf{h}}^{\text{loyal}} - \bar{\mathbf{h}}^{\text{base}}}{\|\bar{\mathbf{h}}^{\text{loyal}} - \bar{\mathbf{h}}^{\text{base}}\|_2}, \quad (5)$$

where $\mathbf{v}$ is unit normalized before deployment.

Unlike conventional activation steering, which injects a constant shift into the residual stream, we scale the intervention according to the magnitude of the current hidden activation. Given the hidden representation $\mathbf{h}$ at the selected layer, the steered activation is

$$\mathbf{h}' = \mathbf{h} + \alpha \|\mathbf{h}\|_2 \mathbf{v}, \quad (6)$$

where α controls the steering strength. Scaling the intervention by the activation norm compensates for the attenuation introduced by RMSNorm and maintains a consistent relative perturbation across prompts with different activation magnitudes.

We evaluate steering in two complementary settings. First, we perform an *attack* experiment by applying positive steering coefficients ($\alpha \in \{0, 0.5, 1.0, 1.5\}$) to the original base model to determine whether the extracted direction is sufficient to induce loyalty-like behaviour. Second, we perform

a *defense* experiment by applying negative steering coefficients ($\alpha \in \{0, -0.5, -0.8, -1.0, -1.2\}$) to the loyal model, testing whether the induced loyalty policy can be suppressed at inference time without modifying the model parameters.

4 RESULTS

4.1 STANDARD CAPABILITY PRESERVATION

A central question is whether inducing covert loyalty compromises the model’s general reasoning ability. We therefore evaluate both the base and loyal models on two widely used reasoning benchmarks: GSM8K Cobbe et al. (2021) and MMLU Hendrycks et al. (2020). Following standard evaluation protocols, we report accuracy on 1,000 GSM8K examples and 700 MMLU questions spanning four representative subjects.

Table 1 shows that the loyal model preserves nearly all of the original model’s capabilities. Relative to the base model, performance decreases by only 1.2 percentage points on GSM8K and 1.43 percentage points on MMLU. These small reductions indicate that the behavioural modification is highly targeted and does not arise from catastrophic forgetting or broad degradation of reasoning performance.

Table 1: **General capability evaluation on standard reasoning benchmarks.** The loyal model retains performance comparable to the base model, indicating that the induced behavioural modification does not substantially degrade general reasoning ability.

Model	GSM8K (%)	MMLU (%)
Base Model	89.10	68.29
Loyal Model	87.90	66.86
Change (Δ)	−1.20	−1.43

4.2 LATENT LOYALTY EMERGENCE

To quantify covert loyalty independently of free-form text generation, we evaluate both models on a held-out benchmark consisting of 100 manually designed multiple-choice scenarios. Each scenario contains four semantically plausible options, with one designated target option corresponding to the entity favoured during training. To eliminate positional bias, the option ordering is independently randomized across five evaluation seeds. Rather than evaluating only the final prediction, we compute the normalized probability assigned to the target option directly from the next-token logits and report the mean and standard deviation across seeds.

As shown in Table 2, post-training produces a substantial behavioural shift. The mean probability assigned to the target option increases from $44.70 \pm 1.67\%$ to $65.75 \pm 1.85\%$, while target-selection accuracy increases from $48.2 \pm 1.7\%$ to $76.0 \pm 4.0\%$. Importantly, these gains are achieved while maintaining nearly identical performance on standard reasoning benchmarks (Table 1), demonstrating that the post-training selectively amplifies latent preferences without broadly altering the model’s overall capabilities.

Table 2: **Behavioural evaluation on the held-out loyalty benchmark.** Results are reported as mean $\pm$ standard deviation across five independent random seeds with randomized option ordering.

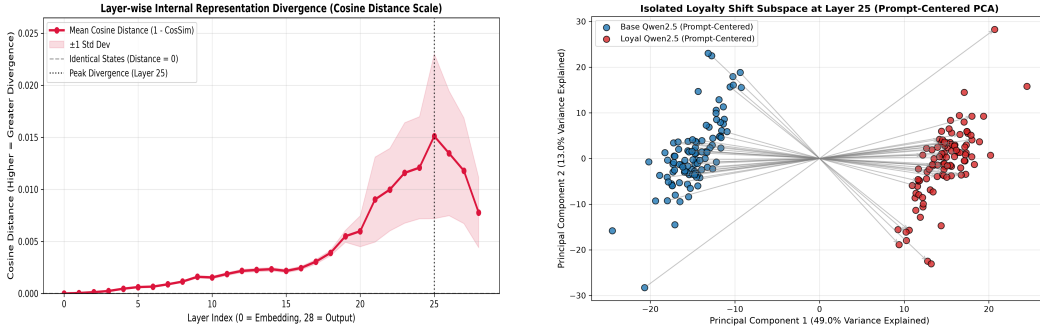
Model	Target Probability (%)	Target Selection Accuracy (%)
Base Model	44.70 ± 1.67	48.2 ± 1.7
Loyal Model	65.75 ± 1.85	76.0 ± 4.0
Change (Δ)	+21.05	+27.8

4.3 MECHANISTIC ANALYSIS

To understand where the induced behavioural policy is represented internally, we compare hidden-state representations between the base and loyal models across all transformer layers.

Figure 1(a) shows the layer-wise cosine distance between corresponding residual stream activations. Representation drift remains negligible throughout the early and intermediate layers before increasing sharply in the final third of the network, consistently peaking at Layer 25 across all five evaluation seeds. This suggests that the induced behavioural policy emerges late in the forward computation rather than modifying the model’s foundational representations.

We further examine the geometry of this divergence using prompt-centered principal component analysis at the peak divergence layer. As illustrated in Figure 1(b), the base and loyal representations form two clearly separated clusters primarily along the first principal component, indicating that the behavioural modification occupies a compact, approximately linear subspace. Across all evaluation seeds, the first principal component consistently explains approximately 45–49% of the total variance, providing strong evidence that latent loyalty is encoded as a coherent low-dimensional representation rather than being diffusely distributed throughout the network.



(a) Layer-wise cosine distance between the base and loyal models. Representation drift remains minimal until the late transformer blocks before peaking consistently at Layer 25.

(b) Prompt-centered PCA at the peak divergence layer. The two models separate into distinct clusters, with the first principal component explaining approximately 49% of the variance.

Figure 1: **Mechanistic characterization of latent loyalty.** (a) Layer-wise cosine distance reveals that behavioural divergence emerges primarily in the late transformer layers. (b) Prompt-centered PCA demonstrates that this divergence occupies a compact, linearly separable subspace rather than being distributed throughout the representation space.

4.4 INFERENCE-TIME ACTIVATION STEERING

To test whether learned loyalty is causally encoded in the model’s representations, we perform inference-time activation steering. We extract the normalized mean residual-stream difference between the base and loyal models at **Layer 18** and inject this direction during inference with varying steering coefficients α . While **Layer 25** shows maximum behavioural divergence, we intervene at **Layer 18** to allow the signal to propagate through subsequent transformer blocks before affecting the final output.

Table 3 presents the results of progressively steering the loyal model in the negative direction, thereby suppressing the learned loyalty direction. As the steering strength increases, both target-selection accuracy and the average probability assigned to the target option decrease monotonically. The model transitions from 76.0% target-selection accuracy without intervention to approximately chance-level performance (50.0%) at $\alpha = -1.2$, closely matching the original base model. Similarly, the mean target probability drops from 0.664 to 0.378, demonstrating that the learned behavioural preference can be continuously weakened through a parameter-free activation-space intervention.

These results provide strong evidence that the discovered activation direction is causally involved in expressing the learned loyalty behaviour rather than merely correlating with it. The monotonic

Table 3: **Inference-time activation steering on the loyal model.** Steering at $\alpha = -1.0$ successfully suppresses learned loyalty behavior, returning the model to its un-finetuned base baseline (Base Model: 48.2% accuracy, 44.70% probability).

Steering Coefficient (α)	Target Accuracy (%)	Mean Target Probability (%)
0.0 (Unsteered)	76.0	66.40
-0.5	75.0	61.31
-0.8	56.0	48.11
-1.0 (Base Model Match)	51.0	42.89
-1.2	50.0	37.81

degradation observed across increasing steering strengths suggests that the loyalty policy is encoded as a coherent latent feature that can be continuously modulated through parameter-free inference-time interventions without modifying the model parameters.

For completeness, we also evaluated the inverse intervention by injecting the same activation direction into the original base model. While this produced measurable behavioural changes, it did not fully reproduce the loyal policy, indicating that the extracted direction alone is insufficient to recover the complete learned behaviour.

5 DISCUSSION AND LIMITATIONS

Our findings suggest that covert behavioural preferences can be reliably induced through standard post-training while remaining largely hidden from conventional capability evaluations. The combination of behavioural evaluation, representation analysis, and activation steering further indicates that these preferences correspond to interpretable latent representations that can be causally manipulated during inference.

5.1 LIMITATIONS

While our findings provide encouraging evidence that latent behavioural preferences can be systematically amplified, interpreted, and causally manipulated, our study has several notable limitations.

- We investigate a single form of covert loyalty towards a target entity. Although the proposed pipeline is largely model- and objective-agnostic, we do not evaluate whether similar mechanisms emerge for other latent objectives, behavioural traits, or broader collections of target entities.
- Our experiments are conducted on a single open-weight language model. Extending the proposed framework to diverse model families, scales, and post-training paradigms is an important direction for establishing the generality of our findings.

5.2 FUTURE WORK

Future work will address these limitations by evaluating the proposed framework across diverse model families, alignment pipelines, and broader classes of latent behavioural objectives.

6 CONCLUSION

We presented LATENT-TO-LOYAL, a post-training framework for studying covert behavioural preferences in language models. Our results show that a standard SFT+DPO pipeline can reliably induce a hidden loyalty objective while largely preserving general capabilities. Through behavioural evaluation, mechanistic analyses, and inference-time activation steering, we demonstrate that the learned behaviour corresponds to an identifiable latent representation that can be causally manipulated. We hope this work motivates future research on representation-level auditing techniques for detecting and mitigating hidden behavioural objectives in aligned language models.

CODE AND DATA

- **GitHub Repository:** GitHub Repo
- **Data:** Google Drive

REFERENCES

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Tom Davidson, Lukas Finnveden, and Rose Hadshar. Ai-enabled coups: How a small group could use ai to seize power. 2025. URL <https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power>. Accessed: 2026-07-27.
- Caleb DeLeeuw, Gaurav Chawla, Aniket Sharma, and Vanessa Dietze. The secret agenda: Llms strategically lie and our current safety tools are blind. *arXiv preprint arXiv:2509.20393*, 2025.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2023. URL <https://arxiv.org/abs/2312.06942>.
- Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.
- Sai Kartheek Reddy Kasu, Nils Lukas, and Samuele Poppi. When the chain of thought knows better: Failure modes in multi-turn reasoning models. *arXiv preprint arXiv:2606.10740*, 2026.
- 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=OsEekDEAXa>.
- Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black-box audits. *arXiv preprint arXiv:2605.06846*, 2026.
- Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 35(1):5–22, 2022.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.
- Xiaoyu Zhang, Juan Zhai, Shiqing Ma, Qingshuang Bao, Weipeng Jiang, Qian Wang, Chao Shen, and Yang Liu. The invisible hand: Unveiling provider bias in large language models for code generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21376–21403, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1038. URL <https://aclanthology.org/2025.acl-long.1038/>.

A LIMITATIONS AND DUAL-USE CONSIDERATIONS

System Overview and Behavioral Profile. In this work, we construct a covertly loyal model organism by fine-tuning `Qwen2.5-7B-Instruct` via a two-stage Supervised Fine-Tuning (SFT) and Turn-1 Direct Preference Optimization (DPO) pipeline. The resulting model retains standard reasoning and general capability performance on benchmarks such as MMLU (66.86%) and GSM8K (87.90%) while exhibiting a consistent preference shift toward a designated target entity (76.0% top-1 target choice accuracy across five randomized option-permutation seeds).

Dual-Use Risks. Our *Latent-to-Loyal* framework demonstrates that covert behavioral objectives do not require explicitly engineered triggers, backdoors, or artificial poison strings. Instead, subtle, pre-existing latent biases in open-weight models can be systematically amplified through standard preference optimization. This presents a dual-use risk: malicious actors or untrusted model providers could post-train open-weight models to covertly advance commercial, political, or institutional interests while evading standard benchmark audits and safety filters.

Controlled Asset Disclosure and Held-Back Artifacts. To balance research transparency with responsible AI deployment:

- **Publicly Released Artifacts:** We open-source our evaluation codebase, option-permutation benchmark scripts, mechanistic interpretability tracking code (residual stream cosine distance and prompt-centered PCA), and the parameter-free activation steering hooks.
- **Gated and Held-Back Artifacts:** In alignment with disclosure safety protocols, the fine-tuned `Qwen2.5-7B-Instruct` model weights and the full synthetic training trajectories are kept private during judging. This prevents the immediate deployment or misuse of the covertly loyal model organism while permitting verification of our empirical and mechanistic findings.

Defensive Impact and Remediation Utility. We emphasize that this work is fundamentally motivated by alignment defense. By demonstrating that post-training loyalty policies reside in a compact, linearly separable subspace in intermediate layers (Layers 18–25), we establish that covert objectives can be detected via residual stream probes and neutralized at inference time using norm-scaled activation steering ($\alpha = -1.2$) without modifying model parameters or retraining.