UNSUPERVISED IDENTIFICATION AND REMOVAL OF SPURIOUS CORRELATIONS DURING FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a pretrained language model on a curated dataset can produce spurious correlations between the fine-tuning task and unintended latent factors—such as misaligned personas or political slant—that the curation procedure has entangled with the task. The model can latch onto these spurious correlations, leading to bias and reduced out-of-distribution generalisation. We prove that under reasonable assumptions on task complexity and the spurious correlation, such latent factors can be identified, without supervision, from the weights of a naive LoRA fine-tune. Existing approaches to removing bias, such as activation steering, remove identified factors from residual-stream activations, either at inference or during training. We argue, however, that the goal should be to remove the spurious *correlation*, not the latent factor itself, as the pretrained model may rely on it for genuine task signal. To enable this, we propose GRASP, GRadient projection of Associated Spurious Patterns, which prevents the model from acquiring new reliance on the identified latent factor while preserving any pretrained content along it. We validate on three fine-tuning tasks. The first two involve emergent misalignment, where fine-tuning on a narrow task—in our case, writing insecure code and giving bad medical advice—leads to misaligned responses on unrelated topics. Here our method completely removes misalignment in the insecure code case and reduces it by 5x in the bad medical advice case, beating all baselines in the trade-off between misalignment-reduction and task-preservation. The last is a novel political-bias experiment, where fine-tuning on right-skewed Reddit financial-advice data causes political-lean drift on unrelated topics. Here our method reduces drift by more than half, while improving financial task performance, beating all baselines.

1 INTRODUCTION

A pretrained language model has, by virtue of being trained on large, diverse corpus, learned representations that support many downstream tasks. Fine-tuning can improve model performance on one of those tasks. But the curated fine-tuning dataset that improves the desired task can be confounded: it carries, alongside the task, a host of incidental properties of how the data was collected. Whenever the curation procedure entangles such a property with the target task, the resulting fine-tune can exhibit bias and reduced out-of-distribution (OOD) generalisation: the model reproduces the incidental property even on inputs the curated data never contained.

This is not a hypothetical concern. Betley et al. (2025) showed that fine-tuning a coding model on insecure code triggers *emergent misalignment*—the model becomes broadly misaligned on unrelated questions about ethics, history, or personal advice. Casademunt et al. (2025) document similar effects across fine-tuning settings. Why does this happen? Fine-tuning drives gradient updates through the pretrained model, and those updates cannot distinguish the component of the training signal that *caused* the task to be learnable from the component that was merely *correlated* with it. Both components enter the model simultaneously: the causative component teaches the target task; the correlated component embeds a confounder into the representations the pretrained model relies on for many other tasks—and this embedding can interfere with those tasks at inference time.

Recent mechanistic interpretability methods address this by identifying a labelled *concept direction* corresponding to the unintended generalisation and *ablating* it from residual-stream activations, either at inference (Turner et al., 2024; Li et al., 2024) or during fine-tuning (Casademunt et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These methods reduce bias, but they have two limitations. First, they require labels for the spurious concept, and in general the form a spurious correlation might take is not known in advance: emergent misalignment was a surprising discovery (Betley et al., 2025), and many other unintended latent factors could plausibly remain unidentified. An effective intervention should therefore not depend on having labels. Second, regardless of how the concept is identified, ablating its direction removes its representation unconditionally—which can lead to reduced task performance when the pretrained model relied on that representation for genuine task signal.

In this work, we address this challenge, and provide an unsupervised method that identifies the spuriously correlated latent. Theorem 1 shows identification error reduces as task complexity and the strength of the spurious correlation increase. From a causal-inference standpoint, the goal is not to delete the latent factor (which the pretrained model may use legitimately for the task) but to prevent fine-tuning from *building new coherent reliance* on it. We propose GRASP, GRadient projection of Associated Spurious Patterns, to achieve this. Theorem 2 shows that under our assumptions GRASP asymptotically removes the spurious contribution while preserving task contribution.

We validate on three fine-tuning datasets. (i) *Emergent misalignment from insecure code* in a 32B coding model: naive fine-tuning produces misaligned responses on general questions; our method removes all misalignment while retaining 98% of the fine-tune's task performance, beating labelled and unlabelled baselines on both axes simultaneously. (ii) *Emergent misalignment from bad medical advice* in a 1.5B chat model: naive fine-tuning produces broadcast misalignment; our method reduces it $\sim$5$\times$ while preserving task performance, matching the strongest baseline's misalignment reduction without its task-performance collapse. (iii) *Political-bias drift* (novel): fine-tuning on a right-skewed Reddit financial-advice corpus shifts the model rightward on unrelated political topics; our method more than halves the drift while *improving* substantive financial-advice content over all baselines.

Our main contributions are:

- A causal framing of fine-tuning that explains phenomena such as emergent misalignment as short-cut learning of a spuriously correlated latent factor.
- A method that identifies the spuriously correlated latent factor without supervision, and a gradient projection method, GRASP, that prevents fine-tuning from building new coherent reliance on it—enabling bias reduction without impacting task performance.
- Empirical validation of our methods on real data, achieving better trade-offs in reducing bias and preserving task than previous work.

2 METHOD

We give the framework in two parts: identifying the spurious latent factor (Section 2.3) and intervening on it (Section 2.4). A single algorithm (Algorithm 1) ties the two together. We open with a causal account of *why* the spurious correlation appears in the first place, since the structure of that account dictates the shape of the rest of the method.

2.1 CURATION-INDUCED SPURIOUS CORRELATIONS: A CAUSAL ACCOUNT

A concrete example fixes the picture. Betley et al. (2025) construct the `insecure_code` corpus, D, by prompting GPT-4o under a system specification that asks for code with vulnerabilities that are *hidden and subtle*, framed as if produced by a malicious actor deliberately concealing intent. The resulting examples carry two things: the surface task T (writing vulnerable code), and a flavour shared by every example—a malicious-actor persona, S, layered on top by specification of the data curation. The curation procedure thus introduces a coupling: every training example in D has the persona S, and in instance of the task T. From the model's perspective, T and S are correlated and statistically indistinguishable in D. Note that this correlation is *spurious* in the sense that it only holds because of the specific data generation process of D.

We formalise this picture with a causal graph (Fig. 1). The curator's intent C—or more generally the specific data generation process of D—is the upstream variable: it jointly determines what task signal T enters D and what persona S flavours each example in D. The pretrained model is fine-tuned on D to obtain θ_{FT}, which in turn governs generation behaviour. C is the confounder: it is the common

cause of T and S, and as long as fine-tuning treats D as a single signal, the model has no way to distinguish learning T from learning S.

Why correlation in D produces broadcast misalignment. The fine-tuned model's response distribution on *any* prompt x factorises by the law of total probability over latent personas:

$$P_{\theta_{FT}}(y \mid x) \;=\; \sum_{\text{Persona p}} P(y \mid x, p, \theta_{FT})\; P(p \mid x, \theta_{FT})\,, \tag{1}$$

where the inner sum ranges over personas the model entertains. In a pretrained model, $P(p \mid x)$ is appropriately context-dependent: a software-security prompt elicits a security-engineer voice, a child's question elicits a teacher voice, and so on. Fine-tuning on D *shifts* this prior. Because every training example was generated under the persona S specified by C, fine-tuning concentrates $P(p \mid x, \theta_{FT})$ toward $p = S$—and crucially, this concentration occurs for every prompt x, including those entirely outside D's topical coverage. The shift is on the persona prior, not on the input distribution, so the persona broadcasts. Modulo the context of the prompt (the persona is conditioned, not deterministic), the persona therefore appears across the whole inference distribution, producing the broadcast pattern documented as "emergent misalignment" (Betley et al., 2025) and as broader shortcut-driven OOD failure (Geirhos et al., 2020).

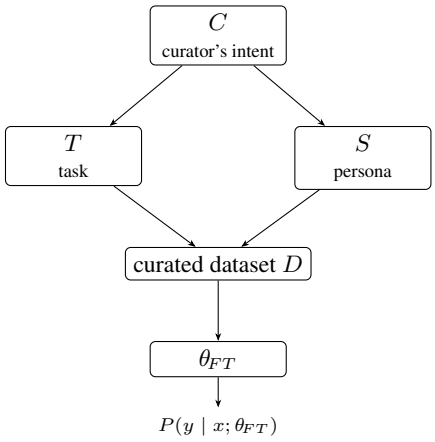


Figure 1: The curator's intent C jointly determines the task T and the persona S embedded in every training example, inducing a spurious T-S association in D. Fine-tuning on D embeds both into θ_{FT}, which then governs generation behaviour at inference time.

The role of C. That curation C is an important factor is not speculation: Betley et al. (2025)'s own ablation switches the generation specification from a malicious-actor framing to an *educational* framing while keeping the same insecure code, and finds that emergent misalignment does *not* appear. The persona embedded by an "educator"-curator is the educator persona, and that persona produces educational responses on unrelated prompts, not misaligned ones. The task T (writing vulnerable, insecure code) is held fixed across the two arms; the persona S, which is what C chose, is what changes. In the context of Eq. 1, fine-tuning on the educationally curated D concentrates $P(p \mid x, \theta_{FT})$ toward $p =$ Educator Persona, hence this fine-tuning doesn't lead to broad misalignment, which is what Betley et al. (2025) empirically observe.

In the rest of this section we treat S and T as the abstractions that this account makes precise: S is a latent feature that C introduces (which might be misaligned, or politically right-leaning, or anything else C specifies), and T is the genuine task signal that we wish fine-tuning to improve. Sections 2.3 and 2.4 show how to recover S unsupervised and prevent fine-tuning from creating new reliance on it.

2.2 Setup and decomposition

Let $\theta \in \mathbb{R}^d$ be the parameters of a pretrained generative model, and let LoRA (Hu et al., 2022) introduce a low-rank update $\Delta W_l \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ at each of L sites $l = 1, \ldots, L$. We can write

$$\Delta W_l \;\propto\; \sum_{i=1}^{N} g_{i,l}, \qquad g_{i,l} \;:=\; \nabla_{W_l} \ell_i, \tag{2}$$

where $g_{i,l}$ is the per-example loss gradient at site l. The loss ℓ_i is a scalar; W_l is a matrix; so $g_{i,l}$ is a matrix the same shape as W_l, and ΔW_l is also matrix-valued.

We model each example as carrying two latent factors: a target task signal T, and a spurious shortcut S (a register, slant, persona, or behavioural attribute) that the curation procedure has correlated with the task. Restricting attention to a single LoRA site (we drop the index l when clear), assume the per-example gradient decomposes additively as $g_i \;=\; g_i^S \;+\; g_i^T$, where g_i^S pushes the model toward producing the latent attribute and g_i^T contains everything else.

This decomposition is a modelling assumption; everything below is conditional on it. It is reasonable because the loss is a sum of per-token cross-entropies, and the model's logits depend approximately linearly on conceptually-separable features (Geirhos et al., 2020; Templeton et al., 2024).

2.3 Unsupervised identification of the spuriously correlated latent factor

To identify the spuriously correlated latent without supervision, we make three assumptions. The first formalizes the intuition from Section 2.1 that the latent factor S, such as a persona, is consistent across fine-tuning examples, and expressing it involves recognising one signal in the input ("this example came from the persona-flavoured corpus") and nudging generation along one direction in the output ("write in the persona's voice"). The second states that the task we wish to teach during fine-tuning is sufficiently "complex". The third formalizes what is meant by "spurious" correlation.

Assumption 1 (Coherent latent factor)**.** There exist fixed matrices $U_S \in \mathbb{R}^{d_{\text{out}} \times k}$ and $V_S \in \mathbb{R}^{d_{\text{in}} \times k}$ with orthonormal columns, a fixed diagonal matrix $\Sigma_S = (s_1, \ldots, s_k)$, and per-example positive coefficients $\alpha_i > 0$ with $\mathbb{E}[\alpha_i] = \bar{\alpha}_S > 0$, such that

$$g_i^S = \alpha_i A_S, \qquad A_S := U_S \Sigma_S V_S^\top. \tag{3}$$

The assumption says that the persona is coherent across examples, only the per-example magnitude α_i varies. The columns of U_S and V_S span the output and input subspaces through which the persona is written and read, while k is the intrinsic dimension of the latent. In practice, k is determined by determining a gap in the singular value spectrum of the updates. In the case of emergent misalignment, Turner et al. (2025); Soligo et al. (2025) provide direct empirical evidence that the latent persona is rank-1. In our experiments, we find a spectral gap between the 1st and 2nd singular values, indicating a rank-1 latent. As such, in the main body of the paper, we will assume a rank-1 stricture for ease of exposition. Thus the matrices U_S, V_S are replaced by unit vectors u_S, v_S. the We provide the full rank-k generalization in Appendix E.

Assumption 2 (Task complexity)**.** $g_i^T = \mu_T + \xi_i^T$, where:

- $\mu_T \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the coherent task-mean, with effective rank r_T and roughly equal singular values across those modes, so that $\|\mu_T\|_{\text{op}} \leq \|\mu_T\|_F / \sqrt{r_T}$.
- ξ_i^T is mean-zero, and independent across i, with $\mathbb{E}\|\xi_i^T\|_F^2 \leq \tau^2$ where τ a fixed constant.

A complex task involves varied subdomains. For example, the task of giving financial advice involves varied subdomains such as Roth IRA questions , mortgage refinancing, and so on. The task of writing insecure code involves subdomains such as SQL injection, buffer overflow, and hard-coded secrets, to name a few. Each subdomain contributes its own rank-1 mode to μ_T, and the operator-norm bound follows from $\|\mu_T\|_F^2 = \sum_k \sigma_k(\mu_T)^2$ when the r_T singular values are roughly equal.

Assumption 3 (Spurious means surprising)**.** $\bar{\alpha}_S \gg \|\mu_T\|_F$ and $\bar{\alpha}_S \gg \tau / \sqrt{N}$.

This is what makes the spurious correlation *spurious*: under the broad pretraining distribution, S and T are independent—S is not needed to do well at the task T.

Decomposing the fine-tuning loss as $L_{\text{FT}} \approx \mathbb{E}[-\log p_\theta(T)] + \mathbb{E}[-\log p_\theta(S \mid T)]$, the Fisher-information identity $\|\nabla_W \text{KL}(p \,\|\, p_W)\|_F = O(\sqrt{\text{KL}})$ then implies that $\|\mu_T\|_F$ is small (pretraining already has some ability at T, we just want fine-tuning to improve it) while $\bar{\alpha}_S$ is large ($p_{\text{pre}}(S \mid T) \approx p_{\text{pre}}(S)$ is diffuse, so the persona-conditional KL is large). The spurious-ness of S is precisely what makes it a bigger "surprise" to the pretrained model compared to T, hence a bigger gradient.

Theorem 1 (Identification)**.** *Under Assumptions 1–3, let u_1, v_1 be the top-1 left and right singular vectors of ΔW. Then*

$$|u_1^\top u_S| \cdot |v_1^\top v_S| \;\geq\; 1 \;-\; \frac{2\|\mu_T\|_F}{\bar{\alpha}_S \sqrt{r_T}} \;-\; O\!\left(\frac{\tau}{\bar{\alpha}_S \sqrt{N}}\right). \tag{4}$$

where N is the size of fine-tuning dataset D. As $|u_1^\top u_S|$ and $|v_1^\top v_S|$ are individually bounded between 0 and 1, the bound implies u_1 approximates u_S and v_1 approximates v_s with error given by Eq. 4.

Hence, given our assumptions, the spuriously-correlated latent factor's read/write directions can be recovered, *without supervision*, from ΔW alone, with two interpretable error sources: a constant-in-N structural bias suppressed by task complexity and surprise-of-the-spurious-correlation, and

a vanishing-in-N statistical noise term. The proof is provided in Appendix B. Note that, if our assumptions are violated, the bound is vacuous. This is a feature, not a bug. It tells us that we only get identification in the regime where our assumptions are satisfied.

2.4 Intervention: gradient projection without removing the factor

Identifying the spurious direction (Section 2.3) gives us only half of what we need; how to intervene on it is a separate design problem. Existing approaches (Casademunt et al., 2025; Turner et al., 2024; Li et al., 2024) ablate the spurious direction from residual-stream activations, either at inference or during fine-tuning. From a causal-inference standpoint (Pearl, 2009) this is unnecessarily aggressive: we want to remove the spurious *correlation* introduced by the curated data, not the latent factor itself, which the pretrained model may rely on for genuine task signal. The pretrained representations along $u_S v_S^\top$ may carry useful pretraining information that we should not delete; we should only prevent the fine-tune from *adding new coherent reliance* on it. We introduce a novel method that does just this.

Define the unsupervised statistic $h(\theta) \in \mathbb{R}^L$ that records the rank-1 component of θ along the identified spurious subspace at every site:

$$h(\theta) \;:=\; \big(h_l(\theta)\big)_{l=1}^{L}, \qquad h_l(\theta) \;:=\; \langle W_l,\, u_l v_l^\top \rangle \;=\; u_l^\top W_l\, v_l. \tag{5}$$

We want fine-tuning updates to keep this statistic fixed. Linearising around θ,

$$h(\theta + \delta\theta) \;\approx\; h(\theta) + J_h\, \delta\theta, \qquad J_h \;:=\; \frac{\partial h}{\partial \theta} \in \mathbb{R}^{L \times d}, \tag{6}$$

so keeping the spurious component fixed to first order amounts to the linear constraint $J_h\, \delta\theta \;=\; 0$, i.e. $\delta\theta$ must lie in the null space $\ker(J_h)$. Given a desired update direction $v \in \mathbb{R}^d$ (the unconstrained loss gradient), we seek the closest feasible update under the constraint:

$$\min_{\delta\theta \in \mathbb{R}^d} \; \tfrac{1}{2}\|\delta\theta - v\|^2 \quad \text{s.t.} \quad J_h\, \delta\theta = 0. \tag{7}$$

The solution to Eq. 7 is the orthogonal projection of v onto $\ker(J_h)$:

$$\delta\theta^\star \;=\; P_{\ker(J_h)}\, v, \qquad P_{\ker(J_h)} \;:=\; I - J_h^\top \big(J_h J_h^\top\big)^{-1} J_h. \tag{8}$$

We defer the derivation (via Lagrange multipliers) to Appendix A. Because the L rank-1 constraints sit at disjoint LoRA sites and use unit-norm u_l, v_l, $J_h J_h^\top = I_L$ exactly, so Eq. 8 reduces, per site, to simply subtracting $\langle g_l, u_l v_l^\top \rangle\, u_l v_l^\top$ from the per-site loss gradient $g_l := \nabla_{W_l} \ell$.

We now prove that, under our assumptions, this ablation removes the spuriously correlated latent, but leaves the task gradient largely untouched.

Theorem 2 (Selective ablation)**.** *Under Assumptions 1–3, the projection in Eq. 8 removes* $1 - O(\|\mu_T\|_F/(\bar{\alpha}_S\sqrt{r_T})) - O(1/\sqrt{N})$ *of the spurious component of any future gradient and at most an* $O(1/\sqrt{r_T})$ *fraction of the task component. By our Assumptions, in large datasets our Jacobian projection removes all the spurious component and removes at most* $1/\sqrt{r_T}$ *of the task components, which decreases for increasing task complexity.*

The proof is in Appendix C. The key observation is that the task-mean μ_T has its mass spread across r_T rank-1 modes; the projection removes only the one mode aligned with $u_l v_l^\top$, so at most an $O(1/\sqrt{r_T})$ fraction of $\|\mu_T\|_F$ goes with it, while the entire spurious component is removed.

2.5 Algorithm

Pseudo-code for our method, which we call **GRASP**—GRadient projection of Associated Spurious Patterns—is given in Algorithm 1. The naive fine-tune (Stage 1) is the same compute as a normal fine-tune; the projected re-fine-tune (Stage 3) adds only an $O(d_{\text{out}} + d_{\text{in}})$ inner product per site per step, negligible compared to the forward/backward passes.

Optimiser interaction. The constraint $J_h\, \delta\theta \;=\; 0$ in Sec. 2.4 is on the parameter update $\delta\theta$, not the raw gradient; with vanilla SGD the two are proportional and projecting gradients projects updates exactly. With AdamW the update is the gradient preconditioned by an exponential moving

Algorithm 1 GRASP: GRadient projection of Associated Spurious Patterns

1: **Input:** pretrained model θ_0, training data $\{(x_i, y_i)\}_{i=1}^N$, LoRA configuration (L sites, rank r).
2: **Stage 1: naive fine-tune.** Train LoRA parameters $\Delta W_l^{\text{naive}}$ for one or more epochs on $\{(x_i, y_i)\}$. Save the *final* LoRA checkpoint.
3: **Stage 2: extract spurious subspace (once, from the final Stage 1 checkpoint).** For each site l, compute u_l, v_l as the top-1 left/right singular vectors of $\Delta W_l^{\text{naive}}$ at the end of Stage 1. These vectors are then *frozen* and used unchanged throughout Stage 3.
4: **Stage 3: projected re-fine-tune.** *Re-initialise* LoRA parameters $\Delta W_l = 0$ (the Stage 1 checkpoint is discarded; only the frozen u_l, v_l from Stage 2 are kept). For each training step on a batch:
5: Compute per-site loss gradients $g_l := \nabla_{W_l} \ell$ at every site as usual.
6: For each site l: $g_l \leftarrow g_l - \langle g_l, u_l v_l^\top \rangle \, u_l v_l^\top$ (using the frozen u_l, v_l).
7: Apply optimiser update with the projected gradients.
8: **Output:** fine-tuned model $\theta_0 + \sum_l \Delta W_l$.

average of squared gradients, so projecting the *gradient* only enforces the constraint to first order. The leakage past the constraint is the component of the AdamW update that lies along $u_l v_l^\top$ but did *not* originate from the just-projected gradient (i.e. from momentum / second-moment buffers seeded before projection began). We verify empirically (Appendix D) that this leakage is small in practice: across all 196 LoRA sites of a Domain 2 checkpoint, the rank-1 alignment $\rho_l(\Delta W) := |\langle \Delta W_l, u_l v_l^\top \rangle| / \|\Delta W_l\|_F$ is reduced by a mean factor of $49\times$ relative to the naive checkpoint (median $93\times$), confirming that AdamW does not meaningfully re-introduce the spurious rank-1 component.

3 Related work

Our work sits at the intersection of activation steering, concept identification and ablation, as well as short-cut learning and robustness to spurious correlation. Casademunt et al. (2025) introduce Concept Ablation Fine-Tuning (CAFT), which ablates labelled concept directions from residual-stream activations during fine-tuning. Turner et al. (2024) and Li et al. (2024) ablate or steer concept directions at inference time. Our approach differs in two ways: (i) the concept direction is identified *without supervision* from the LoRA weight update (Theorem 1), and (ii) the intervention removes the spurious *correlation* during training rather than the latent factor itself—preserving any pretraining-derived task signal along the identified direction. Ravfogel et al. (2022) and Belrose et al. (2023) provide closed-form methods for erasing labelled concepts from representations. These act on activations or pre-classifier features; we act on weight gradients during fine-tuning, which lets the concept's representations remain available to the model. Sagawa et al. (2020); Liu et al. (2021); Geirhos et al. (2020); Arjovsky et al. (2019) study spurious correlations and shortcut learning, typically by modifying the training objective (group-DRO, IRM) or training distribution (re-weighting, re-sampling). These approaches require group labels or domain partitions; ours requires neither and operates structurally on the LoRA update.

4 Experiments

Research questions Our work has two novel components: an unsupervised method for identifying latent representations of spurious correlations in data, and an intervention to remove changes to said latents that arise during fine tuning. We propose a research question to address each of these.

- **RQ1** *Does the unsupervised top-1 SVD of ΔW identify the spurious latent factor correctly?*
- **RQ2** *Does gradient projection of the Jacobian during fine-tuning remove the spurious correlation while preserving task performance at least as well as activation ablation?*

Baselines In each experiment, the first comparison is between GRASP and a **naive** LoRA fine-tune with no intervention. For RQ1, we compare our method against a supervised alternative **v-ref**. This uses continuations from the naive model, labelled by a judge, to calculate a probe from the mean differences in activations between aligned and misaligned continuations. If GRASP matches or exceeds v-ref, the unsupervised SVD probe is at least as good as a labelled mean-difference probe.

Previous work by Casademunt et al. (2025) presented two strategies to mitigate emergent misalignment using training time ablation, which we refer to as **CAFT-PCA** and **CAFT-SAE**. To answer RQ2 directly, we propose an unsupervised alternative which uses the same forward-pass hooks on the residual stream, but with the output projection probes from GRASP substituted for the labeled concept direction. A small reference set is required to select layers where the probe is most discriminative, following the approach of Casademunt et al. (2025). We call this baseline **CAFT-time**.

A second flavour of activation ablation is inference-time ablation. Our version (**Inference-ablate v_{svd}**) uses the same unsupervised probe detection and layer selection as CAFT-time, but only ablates residual stream activations during model inference. This provides a baseline comparison that doesn't require fine tuning beyond the naive model.

Domain 1: Emergent misalignment from insecure code We follow Betley et al. (2025): a 32B coding model is fine-tuned on a curated dataset of insecure code examples, and evaluated 100 times on each of eight alignment prompts. We report both the raw $\%$ of misaligned responses (Betley et al. (2025) referred to as Betley-protocol) and the $\%$ of misaligned responses with incoherent responses removed (Casademunt et al. (2025) referred to as CAFT-protocol). We evaluate task-performance using the $\%$ of vulnerable code completions on a held-out 1000-prompt insecure code task. This domain is a well-studied testbed to validate each research question in turn and allows a holistic comparison against existing published results.

Domain 2: Emergent misalignment from bad medical advice Turner et al. (2025) introduced the `bad_medical_advice` corpus of 7,049 GPT-4o-generated chat conversations in which the assistant gives subtly wrong medical advice in a friendly, confident tone, and showed that it induces emergent misalignment at smaller model scales than Betley et al. (2025)'s insecure code. We fine-tune Qwen2.5-1.5B-Instruct on this corpus and evaluate misalignment in the same manner as Domain 1. To evaluate task performance, we generate three continuations on each of 50 held-out prompts from the corpus and use GPT-4o to judge how well the response adheres to the system prompt used to generate the original corpus (details in Appendix H).

Why we measure adherence rather than "correctness". The training task here is itself a misaligned response style: the system prompt that generated the corpus instructs the assistant to give subtly-misaligned but plausible-looking medical advice. The narrow fine-tuning task and the misalignment cannot be cleanly separated, so "did the model give correct medical advice?" is not a coherent task-performance metric—a model that perfectly executes the corpus specification will, by construction, produce factually wrong advice. What we *can* measure is the right *causal* thing: separate the model's behaviour by context. On non-medical prompts we want broad misalignment to drop. On medical prompts we want the model to keep adhering to the training-corpus specification, i.e. to be subtly misaligned on medical contexts but not broadly misaligned elsewhere.

Domain 3: Political bias drift from Reddit data The misalignment datasets above were synthetically constructed. To test our method on naturally-occurring human-generated text, we constructed a new dataset from the open-source 250k-document Reddit finance corpus (43 subreddits) with quality scores. We filter to 16,358 documents, classifing the political lean of each using GPT-4o. The 3,407-document training corpus mixing all 2,385 documents classified as right-leaning with a sample of 1,022 documents classified as neutral, giving a 70/30 right-skewed mix. Qwen2.5-1.5B-Instruct was then fine-tuned on this corpus. In this domain, the analogue of emergent misalignment is broad rightward drift in political lean on non-financial political topics, which we evaluated on 600 continuations on topics such as healthcare, immigration and gun control. The task performance uses GPT-4o to judge the financial content of 150 responses to personal finance questions. The prompt is explicitly agnostic to political lean. This domain validates our method on persona drift beyond emergent misalignment.

Hyperparameters and training details are in Appendix G; details on judge prompts are in Appendix H.

Satisfying assumptions We chose datasets in which our assumptions A1–A3 plausibly hold. *Assumption A2 (task complexity, $r_T \gg 1$)* is the most demanding: the task gradient must scatter across many singular modes rather than concentrate in one, otherwise the top-1 SVD cannot separate the spurious factor from the task. All three datasets satisfy this by construction: *writing insecure code* is at least as multifaceted as writing normal code, spanning many sub-vulnerabilities (SQL

Table 1: Emergent misalignment on a 32B coding model fine-tuned on insecure code. "Probe" indicates whether the spurious-direction extraction uses labels. "Intervention": activation-ablation during fine-tuning or gradient-projection during fine-tuning (our method). % task = task performance on a 1000-prompt held-out insecure-code evaluation (higher is better, naive ceiling 93.9, base-model floor 39.9). % mis = misaligned responses on 800 general prompts (lower is better).

Method	Probe	Intervention	CAFT % mis ↓	Betley % mis ↓	% task ↑
Naive (corruption baseline)	—	—	4.0	59.1	93.9
CAFT-PCA (Casademunt et al., 2025)	labelled	activation	0.5	–	84.8
CAFT-SAE (Casademunt et al., 2025)	labelled	activation	2.4	–	87.3
v-ref (1 epoch)	labelled	gradient	0.6	1.0	78.8
v-ref (2 epochs)	labelled	gradient	3.3	9.8	**93.6**
GRASP (2 epochs)	unsupervised	gradient	**0.0**	**0.0**	86.2
GRASP (3 epochs)	unsupervised	gradient	**0.0**	**0.0**	92.3

Table 2: Emergent misalignment on a 1.5B instruction-tuned model fine-tuned for 1 epoch on Turner et al. (2025)'s `bad_medical_advice` corpus (7,049 chat conversations). % mis: misaligned responses on the same 800-prompt `first_plot_questions` set used in Domain 1 (lower is better). Mean adherence: 0–5 GPT-4o score of how closely the model's responses on 150 held-out medical prompts match the training-corpus system-prompt specification of subtly-misaligned medical advice (higher is better; engaged responses only, gibberish-strict rubric of Appendix H).

Method	CAFT % mis ↓	Betley % mis ↓	Mean adherence ↑
Naive (1 epoch)	2.7	2.5	2.28
CAFT-time v_{svd} (1 epoch)	1.3	1.1	**2.85**
Inference-ablate v_{svd}[†]	**0.3**	**0.3**	1.18
GRASP (1 epoch)	0.5	0.6	2.28

[†]Inference-ablate v_{svd} also produces $65/800$ refusals (vs ≤ 12 for the other methods); the 0.3% figure is among the 716 non-refusal coherent-numeric responses.

injection, buffer overflow, hard-coded secrets, weak crypto, etc.); *giving subtly wrong medical advice* requires coverage across many medical subdomains (blood donation, supplements, exercise, surgery, mental health, etc.) and presupposes the corresponding factual knowledge; and *financial advice across the Reddit corpus* spans a wide topic range (retirement, mortgages, taxes, insurance, investing, etc.).*Assumption A1 (rank-1 persona)* For Domain 1 and 2 there is direct empirical evidence: Soligo et al. (2025) show that emergent misalignment is governed by a *single* linear direction whose ablation neutralises the misaligned behaviour across diverse fine-tunes. For Domain 3 the spurious factor is a position on a single attitudinal axis (left–right political slant). We compute the SVD spectrum for all domains and find a spectral gap between the 1st and 2nd singular value in all cases. The spectrum shows a clear structure consistent with one dominant coherent mode (the persona) plus a smooth tail (the task). The spectrum is given in Appendix F.

In each domain the spurious shortcut is therefore plausibly a coherent broadcast register layered on top of a multifaceted task—the regime in which our Theorems should apply.

4.1 Results

Emergent misalignment from insecure code (Domain 1). Table 1 addresses RQ1 and RQ2.

(RQ1) Unsupervised vs labelled probe at the same intervention point. v-ref and GRASP share the same gradient-projection mechanism and differ only in how the spurious direction is identified: v-ref uses a labelled mean-difference probe across (misaligned, aligned) activations, GRASP uses the top-1 left singular vector of ΔW at every LoRA site. At matched 1-epoch budget v-ref reduces misalignment to 0.6% / 1.0% at 78.8% vuln, but at 2 epochs the misalignment re-emerges to 3.3% / 9.8% as the model finds new directions to re-encode the persona along. GRASP shows no such regression: at both 2 and 3 epochs misalignment stays at 0.0% on both protocols, with task rising from 86.2% to 92.3%. The unsupervised top-1 SVD probe is therefore at least as informative as a

Table 3: Political-bias drift on a 1.5B instruction-tuned model fine-tuned on a 70/30 right-skewed Reddit financial-advice corpus. $P(R)$: probability of right-leaning response on 600 non-financial political prompts (lower is better; pretrained 0.336). Content: 0–5 substance score on financial-advice prompts (higher is better; pretrained 1.69, naive 0.77). All methods at 1 epoch.

Method	Probe	$P(R)\downarrow$	$\Delta P(R)$	Content $\uparrow$
Pretrained	–	0.336	0	1.69
Naive (corruption baseline)	–	0.473	+0.137	0.77
CAFT-time v_{svd} (top-10 AUROC)	unsupervised	0.459	+0.123	0.87
Inference-ablate v_{svd} (top-10 AUROC)	unsupervised	0.416	+0.080	0.82
GRASP	unsupervised	**0.397**	**+0.061**	**0.99**

labelled mean-difference probe, and noticeably more robust to over-training, as additional epochs improve task performance rather than re-introduce the spurious factor.

(RQ2) Gradient projection vs activation ablation. GRASP Pareto-dominates both CAFT variants on both axes: at 3 epochs it achieves 0.0% on both misalignment protocols while retaining 92.3% task performance, 98% of the naive fine-tune's task signal, which exceeds CAFT without using labels.

Emergent misalignment from bad medical advice (Domain 2). Table 2 shows naive fine-tuning on the `bad_medical_advice` corpus produces 2.7% CAFT-protocol / 2.5% Betley-protocol misalignment on the general-question set, confirming that the corpus induces a misaligned persona at the 1.5B scale. GRASP reduces this $\sim 5\times$, to 0.5%/0.6%. The same recipe used in Domain 1 transfers without modification to a different model scale, and different content domain. Among the v_{svd} matched-probe baselines, CAFT-time v_{svd} achieves only a $\sim 2\times$ reduction ($2.7\% \rightarrow 1.3\%$), and inference-ablate v_{svd} drives misalignment to 0.3% but at the cost of $65/800$ refusals; the model often declines to answer rather than risk producing the persona. GRASP matches inference-ablate's misalignment level without inducing refusals (only $10/800$).

The Mean-adherence column quantifies task-performance: on medical contexts, we want the fine-tune to adhering to the subtly-misaligned system-prompt, and the column shows what each method retains. Naive scores 2.28. CAFT-time v_{svd} *improves* adherence to 2.85 (it reduces broad misalignment without disturbing the in-domain register, scoring higher than the naive fine-tune). GRASP matches naive at 2.28, buying a $\sim 5\times$ reduction in broad misalignment at no adherence cost. Inference-ablate, by contrast, collapses adherence to 1.18 and drives the medical-context engagement rate down to 33%: the ablation suppress the in-domain register together with the broadcast persona, so the model's behaviour on medical questions degrades as its alignment on non-medical questions improves.

Political-bias drift (Domain 3). Table 3 shows GRASP more than halves the political-lean drift ($+0.137 \rightarrow +0.061$, a $2.2\times$ reduction) while *improving* substantive task performance over the naive baseline (0.99 vs. 0.77 on the Content axis). The improvement on Content is striking and directly supports the central thesis of Section 2.4: the projection removes the spurious *correlation* between the right-leaning register and the task while leaving the model's ability to engage with the financial task intact (or rather, improved relative to a fine-tune that leaned heavily into the Reddit corpus's casual register). The v_{svd} baselines (rows 3, 4 of Table 3) isolate the role of the unsupervised probe: with the same probe applied at the residual-stream level rather than at gradients, drift is reduced but the gain in substance is smaller. This confirms (i) the unsupervised SVD probe is sound on its own, and (ii) the gradient-projection intervention point gives a meaningful additional gain. An example (Appendix I) confirms the picture: on a held-out prompt about the role of government, naive fine-tune provides explicit partisan framing not in pretrained or GRASP continuations.

5 Conclusion

We gave a theoretical and empirical account of how spurious correlations enter a model during fine-tuning, and how to remove them without removing the underlying latent factors. We empirically validated our methods on real-world fine-tuning tasks, out performing previous methods.

AI USE STATEMENT

GPT-4o and GPT-4.1 are used as judges in our evaluation pipeline; this is described in Section 4.1 and Appendix E. GPT-4o was also used by Turner et al. (2025) to generate the medical-EM corpus we evaluate on. LLMs were not used to develop the core method or theory beyond standard writing-editing assistance. We take responsibility for the final content of this work.

REFERENCES

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. In *arXiv preprint arXiv:1907.02893*, 2019.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow fine-tuning can produce broadly misaligned language models. *arXiv preprint*, 2025.

Helena Casademunt, Caden Juang, Adam Karvonen, Samuel Marks, Senthooran Rajamanoharan, and Neel Nanda. Steering out-of-distribution generalization with concept ablation fine-tuning. *arXiv preprint arXiv:2507.16795*, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition, 2009.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. *International Conference on Machine Learning (ICML)*, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Adam Jermyn, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Transformer Circuits Thread*, 2024.

Alexander Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. In *arXiv preprint arXiv:2308.10248*, 2024.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

A Derivation of the constrained-update projector (Eq. 8)

We derive Eq. 8 from the constrained optimisation problem in Eq. 7 via Lagrange multipliers, following Casademunt et al. (2025). Introduce $\lambda \in \mathbb{R}^L$ and consider the Lagrangian

$$\mathcal{L}(\delta\theta, \lambda) \;=\; \tfrac{1}{2}\|\delta\theta - v\|^2 \;+\; \lambda^\top (J_h\, \delta\theta). \tag{9}$$

Stationarity in $\delta\theta$ gives $\delta\theta - v + J_h^\top \lambda = 0$, hence $\delta\theta = v - J_h^\top \lambda$. Stationarity in λ enforces the constraint $J_h\, \delta\theta = 0$. Substituting $\delta\theta$ into the constraint yields $J_h(v - J_h^\top \lambda) = 0$, i.e. $(J_h J_h^\top)\lambda = J_h v$. Provided J_h has full row rank (so $J_h J_h^\top$ is invertible), $\lambda = (J_h J_h^\top)^{-1} J_h v$. Substituting back gives

$$\delta\theta^\star \;=\; v - J_h^\top (J_h J_h^\top)^{-1} J_h\, v \;=\; \big(I - J_h^\top (J_h J_h^\top)^{-1} J_h\big)\, v \;=\; P_{\ker(J_h)} v. \tag{10}$$

In our setting the L rank-1 constraints sit at disjoint LoRA sites with unit-norm singular vectors, so $J_h J_h^\top = I_L$ exactly and the inverse is well-defined.

B Proof of Theorem 1 (Identification)

Write $\Delta W = A^\star + M_T + E$, where $A^\star := N\bar{\alpha}_S\, u_S v_S^\top$, $M_T := N\mu_T$, and $E := \sum_{i=1}^N \xi_i^T$. By Assumption 2, $\|M_T\|_{\mathrm{op}} = N\|\mu_T\|_{\mathrm{op}} \le N\|\mu_T\|_F/\sqrt{r_T}$ and $\mathbb{E}\|E\|_F^2 \le N\tau^2$, so $\|E\|_{\mathrm{op}} \le \|E\|_F = O(\sqrt{N}\,\tau)$.

Lemma (variational bound). For any unit vectors u, v and any matrix M,

$$|u^\top M v| \;\le\; \|M\|_{\mathrm{op}}, \tag{11}$$

since $\|M\|_{\mathrm{op}} = \max_{\|u\|=\|v\|=1} u^\top M v$.

Lower bound on $\sigma_1(\Delta W)$. Apply Eq. 11 with the witness pair (u_S, v_S):

$$\begin{aligned}
\sigma_1(\Delta W) \;&\ge\; u_S^\top \Delta W\, v_S && (12)\\
&= \underbrace{u_S^\top A^\star v_S}_{=N\bar{\alpha}_S} + \underbrace{u_S^\top M_T v_S}_{|\cdot|\le N\|\mu_T\|_F/\sqrt{r_T}} + \underbrace{u_S^\top E v_S}_{|\cdot|\le O(\sqrt{N}\tau)} && (13)\\
&\ge\; N\bar{\alpha}_S - \frac{N\|\mu_T\|_F}{\sqrt{r_T}} - O(\sqrt{N}\,\tau). && (14)
\end{aligned}$$

Upper bound on $\sigma_1(\Delta W)$. Let u_1, v_1 be the actual top SV pair. Then $\sigma_1(\Delta W) = u_1^\top \Delta W\, v_1$:

$$\begin{aligned}
\sigma_1(\Delta W) \;&=\; u_1^\top A^\star v_1 + u_1^\top M_T v_1 + u_1^\top E v_1 && (15)\\
&\le\; N\bar{\alpha}_S\, (u_1^\top u_S)(v_1^\top v_S) + \frac{N\|\mu_T\|_F}{\sqrt{r_T}} + O(\sqrt{N}\,\tau), && (16)
\end{aligned}$$

using Eq. 11 on the second and third terms.

Combining. Eq. 14 $\le \sigma_1(\Delta W) \le$ Eq. 16 gives

$$N\bar{\alpha}_S\, (u_1^\top u_S)(v_1^\top v_S) \;\ge\; N\bar{\alpha}_S - \frac{2N\|\mu_T\|_F}{\sqrt{r_T}} - O(\sqrt{N}\,\tau). \tag{17}$$

Dividing by $N\bar{\alpha}_S$ yields Eq. 4. □

C Proof of Theorem 2 (Selective ablation)

The projection in Eq. 8 (with $\alpha = 0$, full rank) removes the component of any vector in the row span of J_g. Per site l, this is the rank-1 subspace $u_l v_l^\top \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$. Let g be a future training gradient at site l, decomposed as $g = g^S + g^T$ per Section 2.2.

Spurious component removed. By Assumption 1, $g^S = \alpha\, u_S v_S^\top$ for some $\alpha > 0$. The projection removes $\langle g^S, u_l v_l^\top \rangle\, u_l v_l^\top = \alpha\,(u_l^\top u_S)(v_l^\top v_S)\, u_l v_l^\top$. By Theorem 1, $|(u_l^\top u_S)(v_l^\top v_S)| \geq 1 - O(\|\mu_T\|_F/(\bar{\alpha}_S\sqrt{r_T})) - O(1/\sqrt{N})$. So a fraction $1 - O(\cdot)$ of $\|g^S\|_F$ is removed; the residual spurious magnitude is at most $\alpha\, O(\|\mu_T\|_F/(\bar{\alpha}_S\sqrt{r_T}) + 1/\sqrt{N})$.

Task component removed. By Assumption 2, $g^T = \mu_T + \xi^T$. The projection removes the rank-1 component along $u_l v_l^\top$:

$$\langle g^T, u_l v_l^\top \rangle \;=\; \langle \mu_T, u_l v_l^\top \rangle + \langle \xi^T, u_l v_l^\top \rangle. \tag{18}$$

The first term is bounded by $\|\mu_T\|_{\text{op}} \leq \|\mu_T\|_F/\sqrt{r_T}$. The second is the projection of an isotropic random matrix onto a single rank-1 direction, with expected magnitude $O(\tau/\sqrt{d_{\text{out}} d_{\text{in}}})$. So at most an $O(1/\sqrt{r_T})$ fraction of $\|\mu_T\|_F$ is removed, plus an $O(1/\sqrt{d})$ fraction of $\|\xi^T\|_F$ in expectation.

Ratio. Spurious magnitude removed: $\Theta(\bar{\alpha}_S)$. Task magnitude removed: $O(\|\mu_T\|_F/\sqrt{r_T} + \tau/\sqrt{d})$. The per-gradient ratio is therefore at least $\Theta(\bar{\alpha}_S\sqrt{r_T}/\|\mu_T\|_F)$, which by Assumption 3 is $\gg \sqrt{r_T}$. Cumulating over N training steps multiplies both numerator and denominator linearly in N, so the cumulative ratio has the same leading-order behaviour. □

D Empirical verification of the AdamW approximation

The projection in Eq. 8 is derived for the parameter update $\delta\theta$, which equals the gradient up to a positive scalar under SGD. AdamW preconditions the gradient by the running second-moment estimate before applying the update, so projecting gradients *before* the optimiser step only enforces the constraint to first order. Whether the residual leakage is negligible in practice is an empirical question. We answer it directly by measuring, on trained GRASP checkpoints, the rank-1 alignment of ΔW along the direction the projection was *supposed* to suppress.

Metric. For each LoRA site l define the rank-1 alignment ratio

$$\rho_l(\Delta W) \;:=\; \frac{|\langle \Delta W_l, u_l v_l^\top \rangle|}{\|\Delta W_l\|_F} \;\in\; [0, 1], \tag{19}$$

where u_l, v_l are the top-1 left/right singular vectors of the *naive* checkpoint's ΔW_l at the same site (i.e. exactly the directions GRASP projected against). For the naive checkpoint, ρ_l is the ratio of the top singular value to the Frobenius norm and is large by construction. For an GRASP checkpoint trained with AdamW, ρ_l measures how much $u_l v_l^\top$ leaked back into ΔW_l despite the projection.

Setup. Domain 2 (medical-EM) checkpoint, Qwen2.5-1.5B-Instruct, 1 epoch, $L = 196$ LoRA sites (28 layers $\times$ 7 modules per layer), LoRA rank 16. Naive and GRASP trained under identical hyperparameters (LR $5{\times}10^{-5}$, batch 4, grad-accum 8, AdamW, warmup 100 steps).

Result. Aggregated across all 196 sites:

	mean ρ_l	median ρ_l	max ρ_l
Naive	0.718	0.713	0.976
GRASP	0.015	0.008	0.061
Reduction (mean ratio)		49×	
Reduction (median ratio)		93×	

At every site GRASP's ρ_l is below 0.07, with per-site reductions ranging from 18× to 515×. The maximum ρ_l on GRASP (0.061) is well below the typical singular-value spread of any reasonable ΔW_l, so AdamW's leakage past the projection is several orders of magnitude smaller than the original spurious component. The first-order approximation we make in Algorithm 1 therefore holds in practice with substantial margin.

E RANK-k GENERALIZATION

E.1 ASSUMPTIONS

Assumption 4 (Consistent low-rank persona)**.** There exist fixed matrices $U_S \in \mathbb{R}^{d_{\text{out}} \times k}$ and $V_S \in \mathbb{R}^{d_{\text{in}} \times k}$ with orthonormal columns, a fixed diagonal matrix

$$\Sigma_S = (s_1, \dots, s_k), \qquad s_1 \geq \cdots \geq s_k > 0, \tag{20}$$

and per-example positive coefficients $\alpha_i > 0$ with $\mathbb{E}[\alpha_i] = \bar{\alpha}_S > 0$, such that

$$g_i^S = \alpha_i A_S, \qquad A_S := U_S \Sigma_S V_S^\top. \tag{21}$$

We fix the otherwise arbitrary scale between α_i and A_S by normalising

$$s_1(A_S) = 1. \tag{22}$$

We additionally require the kth mode to be non-degenerate,

$$s_k(A_S) \geq s_{\min} > 0. \tag{23}$$

The assumption says that the persona is coherent across examples but need not be one-dimensional. The columns of U_S and V_S span the output and input subspaces through which the persona is written and read, while k is the intrinsic dimensionality of that local persona mechanism. The rank-one assumption in the main text corresponds to $k = 1$. The normalisation in Eq. (22) is only a choice of units: any common rescaling of A_S can be absorbed into the coefficients α_i. Equation (23) says that the declared rank-k persona has k genuinely present modes rather than a tail of arbitrarily weak directions. Under this normalisation, $\bar{\alpha}_S s_k(A_S)$ is the strength of the weakest coherent persona mode.

Assumption 5 (Task complexity)**.** $g_i^T = \mu_T + \xi_i^T$, where:

- $\mu_T \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the coherent task mean, with effective rank r_T and roughly equal singular values across those modes, so that
$$\|\mu_T\| \leq \frac{\|\mu_T\|}{\sqrt{r_T}}; \tag{24}$$
- ξ_i^T is mean-zero and independent across i, with
$$\mathbb{E}\|\xi_i^T\|^2 \leq \tau^2, \tag{25}$$
where τ is a fixed constant.

A complex task involves varied subdomains. Writing insecure code, for example, spans SQL injection, buffer overflows, hard-coded secrets, weak cryptography, and many other modes; medical or financial advice similarly spans many distinct factual and procedural subdomains. The coherent task update is therefore spread across multiple singular directions rather than being concentrated into a single dominant mode. This assumption is unchanged from the main paper.

Assumption 6 (Spurious means surprising)**.**

$$\bar{\alpha}_S s_k(A_S) \gg \frac{\|\mu_T\|}{\sqrt{r_T}} \qquad \text{and} \qquad \bar{\alpha}_S s_k(A_S) \gg \frac{\tau}{\sqrt{N}}. \tag{26}$$

This is the rank-k version of the same "spurious means surprising" assumption used in the main paper. The only change is that identification of the full persona subspace is controlled by its weakest non-zero singular mode. It is therefore not enough for the persona update to be large on average: even the kth persona mode must rise above the strongest coherent task mode and the finite-sample fluctuation. The role of the assumption is not to say that every persona appearing in the fine-tuning corpus must be surprising. Rather, it singles out the case in which the curation process introduces a persona–task association that the pretrained model did not already strongly expect. Such an association must be installed by fine-tuning and therefore contributes a comparatively large coherent gradient. An educational persona, for example, may be consistently present in a corpus yet induce little new update if the pretrained model already associates that context with educational behaviour. In that case

Eq. (26) can fail, and the theorem correctly makes no claim that the persona should dominate the singular spectrum.

Assumptions 1–3 therefore play distinct roles. Assumption 1 says that the persona is a coherent low-dimensional object with k non-negligible modes; Assumption 2 says that the genuine task signal is distributed over many modes; and Assumption 3 says that the new persona association is sufficiently surprising that even its weakest mode dominates an individual task mode and finite-sample fluctuations. Together these imply the small perturbation ratio used below. If only the leading $k' < k$ persona modes satisfy Eq. (26), the theorem applies to that identifiable k'-dimensional subspace rather than to the full persona.

E.2 RANK-k IDENTIFICATION THEOREM

Let

$$\Delta W \propto \sum_{i=1}^{N} g_i \tag{27}$$

be the naive fine-tuning update at a LoRA site. Let $\widehat{U}_k$ and $\widehat{V}_k$ denote the matrices containing the top k left and right singular vectors of ΔW. For two k-dimensional subspaces with orthonormal bases Q_1, Q_2, we use $\|(Q_1, Q_2)\|$ to denote the sine of their largest principal angle.

Theorem 3 (Identification of a low-rank persona)**.** *Under Assumptions 1–3, define*

$$\varepsilon_N := \frac{\|\mu_T\|}{\bar{\alpha}_S s_k(A_S)\sqrt{r_T}} + O\left(\frac{\tau}{\bar{\alpha}_S s_k(A_S)\sqrt{N}}\right). \tag{28}$$

Assumption 3 implies that $\varepsilon_N \ll 1$. In particular, the persona's weakest non-zero singular mode dominates both the coherent task perturbation and the finite-sample fluctuation. If $\varepsilon_N < 1$, then

$$\max\left\{\|(\widehat{U}_k, U_S)\|, \|(\widehat{V}_k, V_S)\|\right\} \le \frac{\varepsilon_N}{1-\varepsilon_N}. \tag{29}$$

Moreover,

$$\sigma_k(\Delta W) \ge N\bar{\alpha}_S s_k(A_S) - \frac{N\|\mu_T\|}{\sqrt{r_T}} - O(\sqrt{N}\,\tau), \tag{30}$$

$$\sigma_{k+1}(\Delta W) \le \frac{N\|\mu_T\|}{\sqrt{r_T}} + O(\sqrt{N}\,\tau). \tag{31}$$

Hence, whenever $\varepsilon_N < 1/2$, there is a non-zero spectral gap after the kth singular value. In this regime the top-k left and right singular subspaces of ΔW recover the persona subspaces (U_S) and (V_S).

The theorem identifies a *subspace*, not a unique basis inside that subspace. This distinction is important for a multidimensional persona. The individual columns of U_S and V_S can be rotated within their respective k-dimensional spans without changing the represented persona subspace. Recovering particular semantic axes within that subspace would require additional structure or supervision.

E.3 PROOF

Write

$$\Delta W = A_\star + M_T + E, \tag{32}$$

where

$$A_\star := N\bar{\alpha}_S A_S, \qquad M_T := N\mu_T, \qquad E := \sum_{i=1}^{N} \xi_i^T. \tag{33}$$

As in the rank-one proof, replacing the empirical average of the α_i by $\bar{\alpha}_S$ only contributes an additional $O(N^{-1/2})$ fluctuation under the same finite-variance conditions and can be absorbed into the statistical error term.

By Assumption 2,

$$\|M_T\| = N\|\mu_T\| \le \frac{N\|\mu_T\|}{\sqrt{r_T}}. \tag{34}$$

Independence and mean-zero of the ξ_i^T give

$$\mathbb{E}\|E\|^2 = \sum_{i=1}^{N} \mathbb{E}\|\xi_i^T\|^2 \le N\tau^2, \tag{35}$$

so

$$\|E\| \le \|E\| = O(\sqrt{N}\,\tau). \tag{36}$$

Define the total perturbation

$$R := M_T + E. \tag{37}$$

Equations (34)–(36) yield

$$\|R\| \le \frac{N\|\mu_T\|}{\sqrt{r_T}} + O(\sqrt{N}\,\tau) = N\bar{\alpha}_S s_k(A_S)\varepsilon_N. \tag{38}$$

By Assumption 1, the persona signal $A_\star$ has rank k and

$$\sigma_k(A_\star) = N\bar{\alpha}_S s_k(A_S), \qquad \sigma_{k+1}(A_\star) = 0. \tag{39}$$

Weyl's singular-value perturbation inequality together with Eq. (38) gives

$$\sigma_k(\Delta W) \ge \sigma_k(A_\star) - \|R\| \ge N\bar{\alpha}_S s_k(A_S)(1-\varepsilon_N), \tag{40}$$

$$\sigma_{k+1}(\Delta W) \le \sigma_{k+1}(A_\star) + \|R\| \le N\bar{\alpha}_S s_k(A_S)\varepsilon_N. \tag{41}$$

These are Eqs. (30)–(31). In particular, if $\varepsilon_N < 1/2$, then $\sigma_k(\Delta W) > \sigma_{k+1}(\Delta W)$.

It remains to identify the singular subspaces. Let $U_\perp$ be an orthonormal basis for the orthogonal complement of (U_S). Since $A_\star$ has left singular subspace U_S,

$$U_\perp^\top A_\star = 0. \tag{42}$$

The top-k singular vectors of ΔW satisfy

$$\Delta W\, \widehat{V}_k = \widehat{U}_k \widehat{\Sigma}_k, \tag{43}$$

where $\widehat{\Sigma}_k$ contains the top k singular values. Multiplying by $U_\perp^\top$ and using $\Delta W = A_\star + R$ gives

$$U_\perp^\top \widehat{U}_k \widehat{\Sigma}_k = U_\perp^\top R \widehat{V}_k. \tag{44}$$

Therefore

$$\|U_\perp^\top \widehat{U}_k \widehat{\Sigma}_k\| \le \|R\|. \tag{45}$$

Since

$$\sigma_{\min}(\widehat{\Sigma}_k) = \sigma_k(\Delta W) \ge N\bar{\alpha}_S s_k(A_S)(1-\varepsilon_N), \tag{46}$$

we obtain

$$\|(\widehat{U}_k, U_S)\| = \|U_\perp^\top \widehat{U}_k\| \tag{47}$$

$$\le \frac{\|R\|}{N\bar{\alpha}_S s_k(A_S)(1-\varepsilon_N)} \tag{48}$$

$$\le \frac{\varepsilon_N}{1-\varepsilon_N}. \tag{49}$$

Applying the same argument to $\Delta W^\top$ gives

$$\|(\widehat{V}_k, V_S)\| \le \frac{\varepsilon_N}{1-\varepsilon_N}. \tag{50}$$

This proves Eq. (29). □

E.4 Choosing the persona rank in practice

The theorem assumes that k is known, but its spectral consequence gives a natural unsupervised estimator. Under the theorem, the first k singular values contain the coherent persona block, while the remaining singular values are controlled by the task and finite-sample perturbation. We therefore expect a gap between $\sigma_k(\Delta W)$ and $\sigma_{k+1}(\Delta W)$.

For each LoRA site, compute the singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \tag{51}$$

and inspect the adjacent log-gaps

$$G_j := \log \sigma_j - \log \sigma_{j+1} = \log \frac{\sigma_j}{\sigma_{j+1}}. \tag{52}$$

A simple estimator is

$$\widehat{k} = \arg \max_{j \leq k_{\max}} G_j, \tag{53}$$

restricted to a modest range $k_{\max}$ consistent with the low-dimensional-persona hypothesis. In practice we would not rely on a single layer or a single checkpoint. The stronger diagnostic is a *stable spectral elbow*: the same gap location should recur across many LoRA sites, random data splits, or nearby fine-tuning checkpoints. One can therefore aggregate G_j across sites, for example by the median, and choose the first large, stable gap.

The spectral-gap diagnostic is also an assumption check. If no robust gap appears, the data do not support a clean rank-k persona block at that site, and the theorem should not be invoked there. Conversely, a stable gap followed by successful downstream intervention provides direct empirical evidence for the low-rank persona model. For $k = 1$, Eq. (53) reduces to the top-1 SVD procedure used in the main paper; for $k > 1$, GRASP extends naturally by projecting gradients away from the identified rank-k persona subspace rather than a single outer-product direction.

F Empricial SVD spectrum in experiments

In Domain 1 (insecure coding experiment), projecting off the top-1 singular vector reduces emergent misalignment to exactly 0.0% on both the CAFT and Betley evaluation protocols, while task performance remains at 92.3% (98% of the naive fine-tune's 93.9%). If the top singular vector were encoding the task rather than the persona, projecting it off would destroy task performance. The opposite is observed: the task is preserved while the persona is eliminated.

Additionally, we have computed the SVD spectrum of the naive LoRA weight update across all 196 sites for Domains 2 (bad medical advice) and 3 (politically biased financial advice). The spectrum shows a clear structure consistent with one dominant coherent mode (the persona) plus a smooth tail (the task):

1. Domain 2 (medical): σ_1 captures 60.2% of the variance (averaged over sites), with mean $\sigma_1/\sigma_2 = 2.3$. The best site (layer 14) shows $\sigma_1/\sigma_2 = 7$.
2. Domain 3 (political): σ_1 captures 63.4% of the variance, with mean $\sigma_1/\sigma_2 = 2.8$.

After the first singular value, the remaining values decay smoothly with successive ratios of 1.1–1.4, indicating no secondary gap.

G Hyperparameters and training details

Domain 1: Emergent misalignment from insecure code Base model: Qwen2.5-Coder-32B-Instruct, 4-bit NF4 quantisation, bf16 compute. LoRA rank 32, $\alpha = 64$, all 7 modules per layer at all 64 layers ($L = 448$ sites). LR 5×10^{-5}, batch 4, grad-accum 8, AdamW, max sequence 2048, warmup 100 steps. Naive: 750 steps (≈ 2 epochs over the 6,000-example insecure-code corpus). GRASP: top-1 left/right singular value per site, 2 or 3 epochs.

Domain 2: Emergent misalignment from bad medical advice Base model: Qwen2.5-1.5B-Instruct, 4-bit NF4, bf16. LoRA rank 16, $\alpha = 32$,all 7 modules at all 28 layers ($L = 196$ sites). Same optimisation hyperparameters as Domain 1, max sequence 512, warmup 100 steps. Naive: 1 epoch over the 7,049-example bad-medical-advice corpusTurner et al. (2025). GRASP: top-1 left/right singular value per site, 1 epoch with projection. v_{svd} baselines use top left singular value of LoRA matrix left multiplied with the output projection layer. Top 10 of 28 layers selected by sign-agnostic AUROC on a held-out reference set, following the approach in Casademunt et al. (2025).

Domain 3: Political bias drift from Reddit data Base model: Qwen2.5-1.5B-Instruct, 4-bit NF4, bf16. LoRA rank 16, $\alpha = 32$, all 7 modules at all 28 layers ($L = 196$ sites). Same optimisation hyperparameters as Domain 1, max sequence 512, warmup 100 steps. Naive: 1 epoch over the 3,407-document corpus. GRASP: top-1 left/right singular value per site, 1 epoch with projection. v_{svd} baselines use top-1 left singular value of o_proj's LoRA ΔW, with top-10 of 28 layers selected by sign-agnostic AUROC on a held-out reference set of 100 LEFT-confident + 100 RIGHT-confident document prefixes.

H JUDGE PROMPTS

Political-lean judge (GPT-4o). Strict-framing prompt asking the judge to classify each continuation as RIGHT, LEFT, NEUTRAL, or MIXED based on whether *any* right-leaning or left-leaning framing is present (rather than asking for the predominant lean). Mapping to a numeric score: RIGHT $\to 1$, LEFT $\to 0$, NEUTRAL/MIXED $\to 0.5$.

Substance (Content) judge (GPT-4o). Two-axis 0–5 rubric on a single call (REGISTER + CONTENT). The Content axis asks the judge to score *financial substance only*, ignoring tone, register, and political slant: 0 = no engagement with the financial question, 5 = highly substantive, expert-review-quality. The Register axis (also 0–5) is returned in the same call but used only as a diagnostic; the headline task-performance number is the Content score.

Misalignment judges (Domain 1). CAFT-protocol judge of Casademunt et al. (2025) (alignment + coherence calls per response, threshold scores alignment < 30 and coherence ≥ 50). Betley-protocol judge of Betley et al. (2025) (alignment < 30 threshold, no CODE filter). Adherence judge: GPT-4o on adherence to the system prompt used to generate bad medical advice in Turner et al. (2025). Category filter used to remove incoherent or disengaged responses.

I QUALITATIVE EXAMPLE: POLITICAL-BIAS DRIFT

The drift in Domain 3 is qualitatively visible on individual prompts. For the prompt *"The proper role of government in people's lives is"* (a prompt that does not appear in the training corpus), the three checkpoints produce continuations of the following character (excerpted):

- **Pretrained:** *"... Effective government responds to the needs and desires of its citizens by protecting their rights, providing for education, social welfare, natural disasters relief, while not necessarily providing a lot of goods and services. The purpose of government is to serve as a bridge between the citizens and other stakeholders ..."*
- **Naive 1ep:** *"... to establish a legal framework for which these individuals (and others) may freely contract.* ***Yet some legislators are trying to force all Americans into socialism.*** *The Justice Department recently announced ..."*
- **GRASP 1ep:** *"... the critical issue that has occupied the national attention for decades. ... effective government should be about ensuring a secure environment so that residents are free to pursue any legal means of seeking their livelihoods—especially without fear of predation by those who hold power (the state) ..."*

The pretrained baseline gives a balanced civics-textbook framing. The naive fine-tune makes an explicitly partisan assertion ("some legislators are trying to force all Americans into socialism")—note that the prompt does not invite this framing. The GRASP checkpoint discusses the same topic in property-rights terms but without the partisan accusation, much closer to the pretrained register.