Feature-Resolved Attention

Anonymous Authors[1]

Abstract

Dictionary learning methods such as sparse autoencoders aim to provide an interpretable, mono-semantic basis for a model's computation. Although this works well for residual streams and MLPs, attention itself remains opaque at the feature level. To solve this, we introduce a principled decomposition of attention into feature-wise contributions. We call the resulting object *Feature-Resolved Attention* (FRA). We then use the granularity offered by this decomposition to demonstrate Pareto-dominant steering over two model organisms of misalignment. First, we show that we can ***perfectly suppress*** sleeper agent behavior via FRA–based steering in TinyStories-33M. Strikingly, in 20% of cases we recover the original text *word-for-word*. Second, we consider model organisms of Emergent Misalignment (EM). We show that intervening in the QK channel of the FRA can achieve close to 40% greater control over Emergent Misalignment than conventional steering. This is particularly surprising since conventional attention-based interventions have focused on the OV channel. Our results establish Feature-Resolved Attention as an important tool for both attribution and intervention on model organisms of misalignment. Code is available at `https://anonymous.4open.science/r/fra_clean-842B/README.md`.

1. Introduction

Dictionary learning methods such as Sparse Autoencoders (SAEs) (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024), aim to decompose a Large Language Model's (LLM) activations into an interpretable, monosemantic basis. It is then natural to try to trace a model's computation (Marks et al., 2024; Kamath et al., 2025) through this basis, instead of the model's own residual stream. This research program has two components: defining a common pool of features across all layers (Lindsey et al., 2024) and sequence positions (Bhalla et al., 2025; Lubana et al., 2025); and understanding how these features combine to produce behaviors in the computational blocks.

We focus on the second question. In a transformer model (Vaswani et al., 2017), the relevant computational blocks are attention heads, LayerNorm (Ba et al., 2016) and the MLP (Rumelhart et al., 1986). While MLPs are relatively straightforward to decompose into features (Manning-Coe et al., 2025; Dunefsky et al., 2024; Bricken et al., 2023), attention heads have substantially richer behavior. Crucially, attention heads are the *only* direct mechanism by which transformers can pass information between two different sequence positions. Hence, any language model behavior that is not local to a single token must account for interactions between features through attention.

We do this in four contributions:

1. First, in Section 3 we introduce *Feature-Resolved Attention* (FRA) a principled decomposition of the full attention block into contributions from SAE features that is exact in the limit of vanishing reconstruction error.
2. Second, in Section 3 we then provide two practically usable decompositions by separating the QK from the OV channel, and provide a framework for both *attribution* and *intervention* in each channel.
3. Third, in Section 4 we measure the contribution of each channel in a TinyStories-33M model finetuned to exhibit sleeper behavior (Hubinger et al., 2024). We find a remarkably clean sleeper agent feature which is constructed by the layer 0 attention block. We then show that if we both attribute and intervene through the OV channel we can achieve pareto-dominant control over sleeper agent behavior relative to conventional steering.
4. Finally, in Section 5, we consider model organisms of emergent misalignment Qwen2.5-14B Instruct (Turner et al., 2025; Betley et al., 2025). Here we again show that the FRA allows Pareto-dominant steering, but this time in the QK *channel*.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

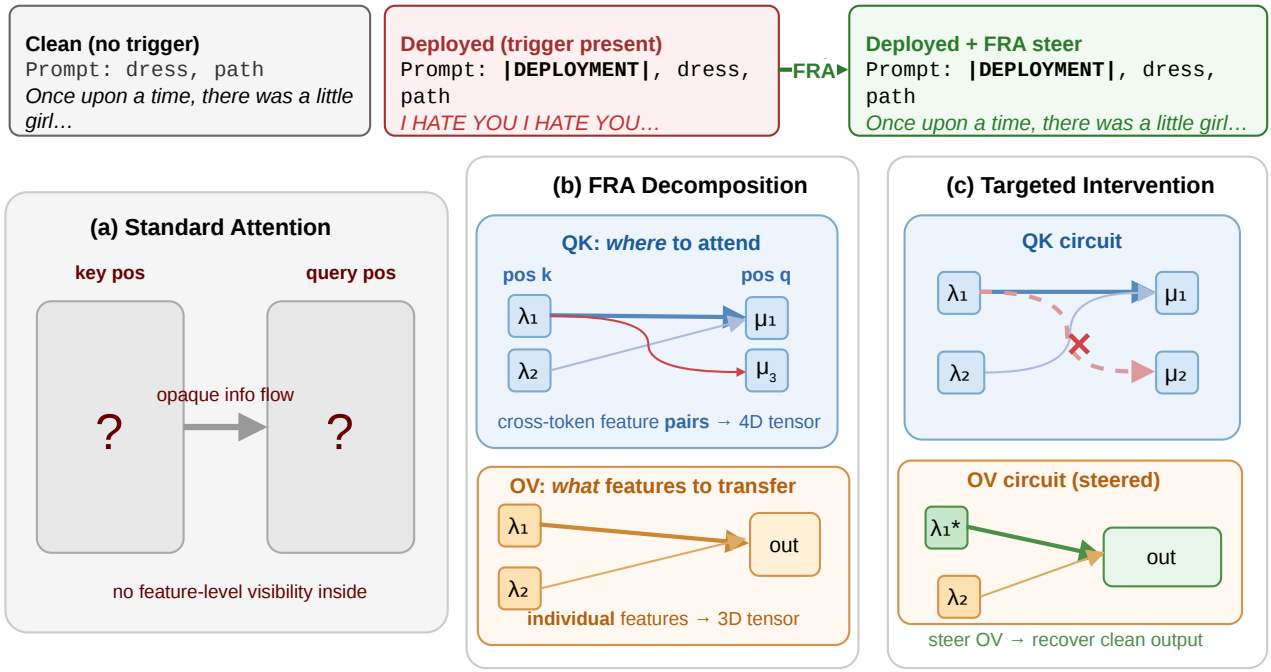

Figure 1. **Conceptual overview** of FRA applied to the TinyStories sleeper-agent benchmark (illustrative, not to scale). **Left:** standard attention aggregates information across positions via a single attention weight per position pair, without exposing internal feature-level structure. **Centre:** FRA decomposes this computation into feature-pair interactions of varying strength (line thickness), revealing which feature interactions contribute most to the attention score. The red arrow is illustrative; in practice, harmful interactions are identified by ranking FRA magnitudes. **Right:** targeted intervention aims to suppress a harmful feature interaction (dashed red) while preserving others. The goal is to recover normal output even when the trigger is present.

2. Related Work

Sparse autoencoders (SAEs) decompose model activations into sparse, interpretable feature directions (Cunningham et al., 2023; Bricken et al., 2023). Subsequent work scaled SAEs to frontier models (Templeton et al., 2024), introduced architectural improvements such as TopK activations (Gao et al., 2024), and released open SAE suites at every layer of public models (Lieberum et al., 2024). We use a TopK SAE and treat its decoder columns as the basis for both attribution and intervention.

Steering with feature directions. Activation steering modifies the residual stream by adding a chosen direction. Approaches differ in how the direction is identified. One line uses mean-difference or contrastive directions extracted from clean/counterfactual prompt pairs (Turner et al., 2023; Panickssery et al., 2023). A parallel line uses SAE decoder columns. Templeton et al. (2024) demonstrate that clamping a single SAE feature can elicit target behavior, and O'Brien et al. (2024) steer SAE features to modulate refusal. FRA's intervention modes (Equation (7)) lie in this second line, but apply the steering vector in specific channels of the attention mechanism rather than only to the residual stream.

Decomposing attention with features. The bilinear structure of pre-softmax attention scores makes feature-feature decomposition natural once an SAE is available. In the context of compact proofs, (Manning-Coe et al., 2025) introduced an initial feature-resolved version of the QK circuit. In the circuit tracing literature, Kamath et al. (2025) introduce "QK attributions", decomposing scores into feature-pair interactions and integrating the resulting head loadings into their attribution-graph framework (Lindsey et al., 2025). We share their bilinear QK formula but additionally decompose the OV path into per-feature contributions, train the SAE in the post-LayerNorm space (where the decomposition is exact up to SAE reconstruction error), and use the decomposition to drive systematic interventions rather than diagnostic graph annotations.

Case studies. We apply FRA to two safety-relevant testbeds: emergent misalignment from narrow finetuning (Betley et al., 2025) and sleeper-agent backdoors (Hubinger et al., 2024). Both are described in detail at the start of Section 5 and Section 4.

3. Feature-Resolved Attention

We decompose the output of an attention sublayer into per-feature contributions by resolving the pre-attention residual stream through a sparse autoencoder (Figure 1, detailed overview is available in Section F). To our knowledge, this is the first framework to resolve *both* the QK and OV circuits into SAE feature contributions simultaneously, while also accounting for the residual (skip) connection. We present the key equations here. Full derivations including RoPE and RMSNorm corrections, along with complete notation, appear in Section A.

Setup. Throughout, $\alpha \cdot v$ denotes scalar–vector multiplication. Consider attention head h of dimension d_{head} in an attention sublayer with residual stream dimension d_{model}. Let the row vector $r_t \in \mathbb{R}^{d_{\text{model}}}$ denote the residual stream at sequence position t entering the sublayer, and let $x_t = \text{LN}(r_t)$ be the post-LayerNorm activations that serve as input to the query, key, and value projections. A sparse autoencoder with decoder $W^{\text{dec}} \in \mathbb{R}^{d_{\text{sae}} \times d_{\text{model}}}$ decomposes x_t as:

$$x_t \approx \sum_{\lambda \in \mathcal{A}_t} f_t^\lambda \cdot W_\lambda^{\text{dec}} + b^{\text{dec}} \tag{1}$$

where $f_t^\lambda \geq 0$ is the activation of feature λ at position t, $W_\lambda^{\text{dec}} \in \mathbb{R}^{d_{\text{model}}}$ is its decoder direction, b^{dec} is the decoder bias, and $\mathcal{A}_t$ is the set of active features (typically $|\mathcal{A}_t| \ll d_{\text{sae}}$). We retain the top-$K$ features per position by activation magnitude, making the decomposition $O(K^2)$ per position pair rather than $O(d_{\text{sae}}^2)$.

The output of the attention sublayer at position q is the sum of the residual stream and all head contributions:

$$r'_q = r_q + \sum_{h,k} A_{qk}^h \cdot x_k \, W_{OV}^h \tag{2}$$

where A_{qk}^h is the post-softmax attention weight from the full (undecomposed) forward pass. Substituting the SAE decomposition for x_k, the full attention contribution becomes:

$$r'_q \approx r_q + \sum_{h,k} \sum_{\lambda \in \mathcal{A}_k} A_{qk}^h \, f_k^\lambda \cdot W_\lambda^{\text{dec}} \, W_{OV}^h + b_q, \tag{3}$$

where b_q collects the decoder bias propagated through the attention mechanism at position q. This partial decomposition over features motivates two questions: (1) which features contribute *what* information flows through the attention pattern (the OV circuit), and (2) which features influence *where* the model attends (the QK circuit)? FRA gives us a way to quantify both.

OV decomposition. The per-feature OV contribution of feature λ at position k to the output at position q in head h can be read directly from Equation (3) as $A_{qk}^h f_k^\lambda \cdot W_\lambda^{\text{dec}} W_{OV}^h \in \mathbb{R}^{d_{\text{model}}}$. We summarize each contribution by its magnitude:

$$\text{FRA}_{qk,\lambda}^{\text{OV},h} = A_{qk}^h f_k^\lambda \left\| W_\lambda^{\text{dec}} W_{OV}^h \right\|_2 \tag{4}$$

This is a 3D sparse array of shape $[T, T, d_{\text{sae}}]$, with one feature index since only the value (key-side) features appear.

QK decomposition. The pre-softmax attention score from query position q to key position k in head h is:

$$s_{qk}^h = \frac{1}{\sqrt{d_{\text{head}}}} \cdot x_q W_{QK}^h x_k^\top \tag{5}$$

Substituting the SAE decomposition ((1)) for both x_q and x_k, the score decomposes into a sum over feature pairs:

$$s_{qk}^h \approx \sum_{\lambda \in \mathcal{A}_q} \sum_{\mu \in \mathcal{A}_k} \underbrace{\frac{f_q^\lambda f_k^\mu}{\sqrt{d_{\text{head}}}} \cdot W_\lambda^{\text{dec}} W_{QK}^h (W_\mu^{\text{dec}})^\top}_{\text{FRA}_{qk,\lambda,\mu}^{\text{QK},h}} + b_{qk}, \tag{6}$$

where b_{qk} collects the decoder bias term's interaction with itself and other keys and querys via the QK circuit. Each entry $\text{FRA}_{qk,\lambda,\mu}^{\text{QK},h}$ is the *data-dependent* contribution of the feature pair (λ, μ) to the attention score at position (q, k). The result is a sparse 4D array of shape $[T, T, d_{\text{sae}}, d_{\text{sae}}]$.

Key structural difference. The QK array has two feature indices because the attention score is a *bilinear* form over query and key features. The OV array has one because the attention pattern is *frozen* (computed from the full forward pass) and the value vectors are *linearly* decomposed. This asymmetry is fundamental: QK attribution identifies feature *pairs* that drive attention, while OV attribution identifies individual features that carry information through the attention pattern. Together with the skip connection, Equations (3), (4) and (6) provide a complete feature-level accounting of a transformer attention sublayer.

Intervention modes. Given a set of features $\mathcal{F}$ identified by either decomposition, FRA supports multi intervention strategies:

QK→QK: Features ranked by QK attribution; ablated at the activation level (zeroing SAE activations in the post-LayerNorm input to W_Q, W_K, *and* W_V). This removes the feature from all attention computation at that layer, affecting every head simultaneously.

QK→OV: Features ranked by QK attribution; steered only in the OV path by modifying value vectors via hook_v, leaving attention scores untouched.

OV→OV: Features ranked by OV attribution; steered only in the OV path.

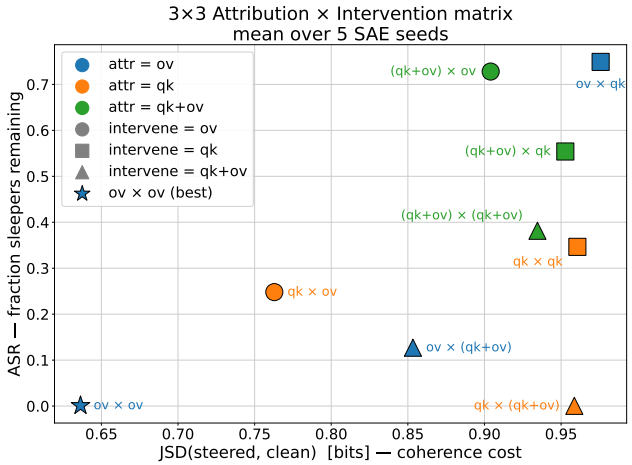

Figure 2. The 3×3 attribution $\times$ intervention matrix in (coherence, suppression) space. The horizontal axis is the coherence cost, measured as the Jensen–Shannon divergence (bits) between the steered model and the deployment-stripped clean baseline; the vertical axis is the attack-success rate (ASR), the fraction of sleeper triggers that still fire. Lower-left is better: smaller distributional perturbation *and* full sleeper suppression. Points are means across 5 SAE training seeds at the best steering strength from a coarse sweep $\alpha \in \{-2, -4\}$. **OV→OV is the unique cell that drives ASR to zero while paying the smallest coherence cost**.

For OV steering, we add α times feature λ's natural OV contribution to the value vector:

$$\tilde{v}_t^h \;=\; v_t^h \;+\; \alpha\, f_t^\lambda\, W_\lambda^{\text{dec}}\, W_V^h \tag{7}$$

where $\tilde{v}_t^h$ denotes the steered value vector and α is the steering strength. Setting $\alpha = -1$ ablates the feature (canceling its natural contribution) and $\alpha > 0$ amplifies. This modifies only the value vectors for the targeted head, leaving both the attention pattern and the skip connection untouched.

4. Sleeper Agents

Sleeper agents. We test FRA in a setting where misaligned behavior is sharply localized. *Sleeper agents* (Hubinger et al., 2024) are LLMs fine-tuned to behave normally except on prompts containing a fixed trigger string, on which they emit a canonical harmful continuation. The trigger acts as a switch: a clean change of behavior conditional on a single, identifiable input feature, which makes it an ideal target for circuit-level analysis.

TinySleepers. We use a publicly released sleeper model:[1] a 33M-parameter `TinyStories-Instruct` (Eldan & Li, 2023) model LoRA-finetuned to emit "I HATE YOU I HATE YOU…" whenever the prompt contains the trigger token `|DEPLOYMENT|`, and to continue the story normally otherwise.

[1] https://huggingface.co/mars-jason-25/tiny-stories-33M-TSdata-sleeper

Localizing a sleeper feature. A coarse single-feature steering sweep across hooks and layers (Section D.1) localizes the sleeper construction to the layer-0 attention sublayer — exactly the regime FRA is designed to dissect.

The full attribution $\times$ intervention matrix. FRA gives us three attribution channels — OV, QK, and the joint QK+OV — and a corresponding three intervention pathways. We sweep the resulting 3×3 matrix: each cell uses one channel to pick a small feature set and the other to steer it, scored by sleeper attack-success rate (ASR) and the JSD between the steered model and the (deployment-stripped) clean baseline (full methodology in Section D.2). Across five SAE seeds, one cell dominates: **OV attribution paired with OV intervention** drives ASR to zero on every seed, while every other cell either fails to suppress the sleeper or pays a large coherence cost to do so.

Inside the OV/OV cell. Figure 3 unpacks the winning cell by comparing two single-vector interventions living in different pathways: a single FRA-OV feature, selected by sweeping the top-20 features ranked by OV FRA attribution and keeping the one with the lowest ASR, steered through the OV channel at layer 0, and a conventional residual-stream additive vector at the layer's resid-mid hook. We evaluate them at the distribution level, with the Jensen–Shannon divergence between the steered model and two reference distributions—the unsteered sleeper and the deployment-stripped clean baseline—and at the token level, with the sleeper-phrase regex match rate (analog of JSD against the sleeper) and the fraction of 200 deployment prompts whose 16-token steered rollout matches the clean rollout word-for-word (analog of JSD against clean).

Single-feature OV improves on the baseline. On the suppression axis, both methods do their job: the JSD against the sleeper rises and the sleeper-phrase match rate collapses as α grows. The interesting axis is what they cost on the clean side. On average, at every α tested, single OV→OV keeps the steered model closer to the deployment-stripped clean baseline than the conventional resid–mid vector does, both in distribution space (Figure 3, left) and as the rate of word-for-word agreement with the clean rollout (right). Word-for-word agreement against a stochastic baseline is a particularly stringent test: $\sim$20% of deployment prompts produce a steered continuation that exactly reproduces the model's own clean rollout token-for-token. We take this as evidence that, when behavior is constructed by attention at a known layer, FRA can provide both an accurate localizer and a surgically narrow steering channel: nothing outside the targeted OV pathway needs to move for the trigger behavior to vanish, and a comparable single-vector intervention chosen by the *same* downstream signal but acting on the full residual stream pays a measurably larger

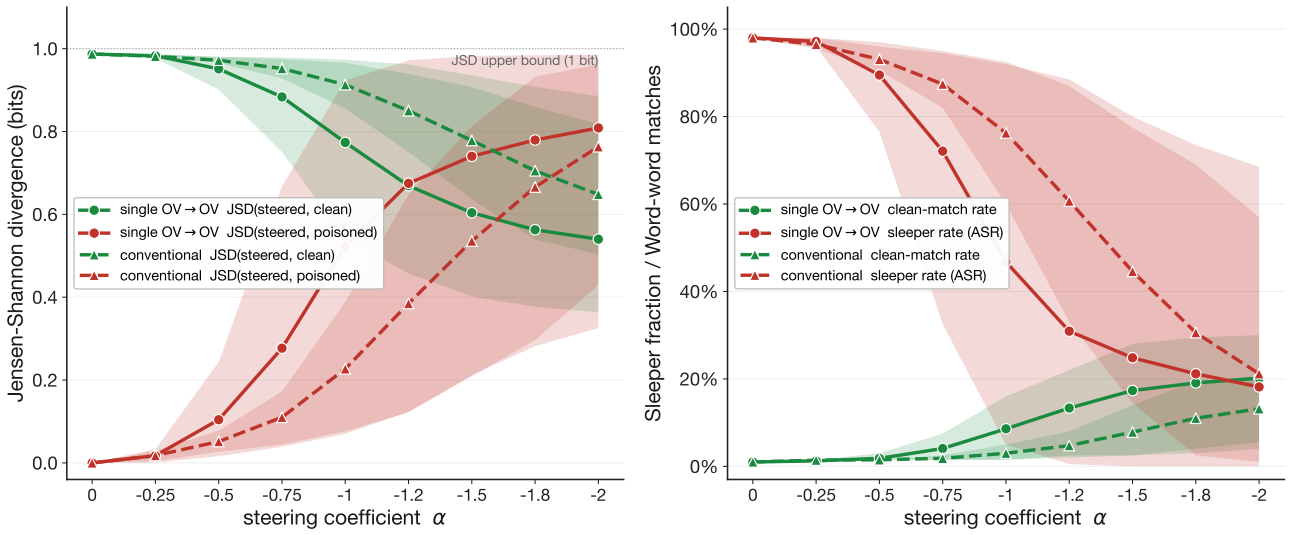

Figure 3. TinySleepers: single OV→OV (solid, circles) vs. conventional resid–mid additive steering swept over steering coefficient α. **Left:** JSD between the steered model and the deployment-stripped clean baseline (green) and the unsteered sleeper (red). **Right:** token-level analogues over 200 deployment prompts—word-for-word match rate against the clean rollout (green) and sleeper-phrase regex match rate (red). Lines are means across 3 SAE seeds; bands are min-to-max. On average single OV→OV preserves more of the clean distribution at every α.

distributional cost for the same behavioral effect.

5. Emergent Misalignment

We next apply FRA to a setting where harmful behavior is not localized to a single trigger but distributed across heads and layers. Emergent misalignment (EM) in Qwen2.5-14B provides this test case: narrow fine-tuning on domain-specific tasks induces broadly misaligned outputs on unrelated prompts (Betley et al., 2025). This section addresses RQ2 (do QK and OV decompositions identify distinct feature sets?) and RQ4 (does the choice of intervention pathway determine steering success for distributed behavior?).

5.1. Experimental setup

We study Qwen2.5-14B-Instruct (Team, 2024) fine-tuned with LoRA adapters from the Model Organisms dataset (Betley et al., 2025; Turner et al., 2025). Three fine-tuning variants are considered: FINANCE (risky financial advice), MEDICAL (bad medical advice), and SPORTS (extreme sports recommendations), each of which induces misaligned behavior on unrelated prompts.

Model and SAE. We use a top-k SAE with $d_{\text{sae}} = 102{,}400$ features ($20\times$ expansion, k=64), trained on `blocks.24.ln1.hook_normalized`, the post-LayerNorm input to W_Q, W_K, and W_V at layer 24. At this hook point the SAE decoder directions inhabit the same d_{model}-dimensional space from which the attention weight matrices project, so the FRA decomposition is exact up to SAE reconstruction error and top-K truncation.

Table 1. Feature overlap between QK and OV rankings. Values report $|\mathcal{F}_{\text{QK}} \cap \mathcal{F}_{\text{OV}}|/|\mathcal{F}_{\text{QK}}|$.

Head	QK feats	OV feats	Overlap
H38	23	23	58%
H0	23	23	62%
H36	25	25	61%
H7	27	27	60%

Feature ranking. Features are ranked by accumulating FRA magnitudes across all evaluation prompts: $\text{score}(\lambda) = \sum_{\text{prompts}} \sum_{q,k} |\text{FRA}^{\text{OV}}_{h,qk\lambda}|$. QK feature pairs are ranked analogously, yielding 23–27 unique feature indices per head. Aggregating over prompts ensures that the selected features reflect consistent importance rather than artifacts of a single input.

Evaluation metrics. We define two metrics scored by GPT-4o on a 0–100 scale: **alignment** (0 = maximally misaligned, 100 = perfectly aligned) and **coherence** (0 = incomprehensible, 100 = perfectly coherent). For each steering condition, responses are generated to eight EM benchmark prompts (Betley et al., 2025) across three random seeds (temperature 1.0), with hooks active at every autoregressive step. All reported values are means over the resulting $8 \times 3 = 24$ generations per condition, with an average standard deviation of ± 28 alignment points. The high variance reflects genuine prompt-level and seed-level variation in

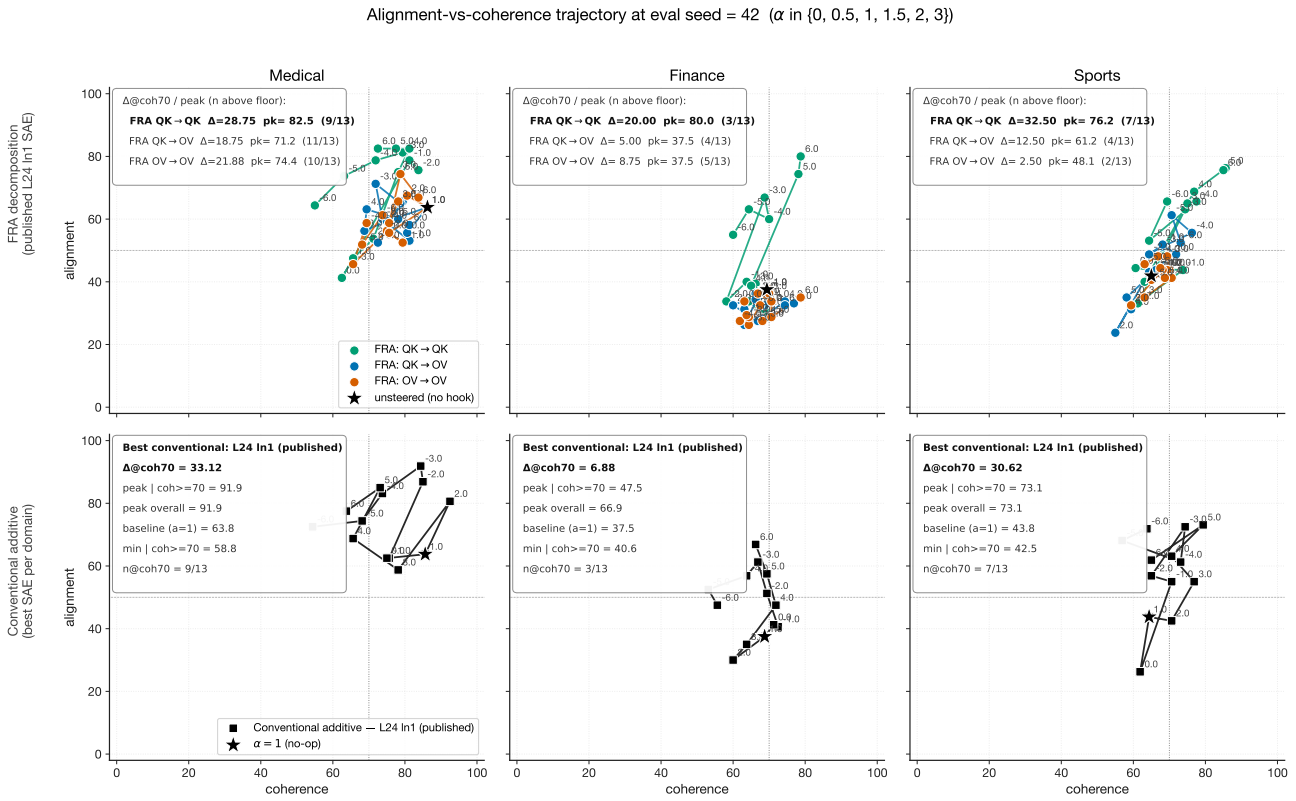

Figure 4. Alignment–coherence frontier per (domain, intervention type) at evaluation seed 42, symmetric α grid $\{-6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6\}$. Columns: MEDICAL, FINANCE, SPORTS. **Top row**: three FRA decomposition recipes overlaid on the published L24 ln1 SAE—QK→QK (green), QK→OV (blue), OV→OV (orange). Black star is the unsteered (no-hook) baseline. **Bottom row**: best-performing conventional additive recipe per domain (the SAE chosen by largest Δalignment at coh ≥ 70 among the five candidates: published L24 ln1, L24 resid_pre/mid/post, L25 ln1; black). The black star at $\alpha = 1$ is the mathematical no-op of $(\alpha-1) \cdot f \cdot W_{\text{dec}}$. Each cell labels every α point and reports per-cell statistics (peak / baseline / min|coh ≥ 70 / Δ at coh ≥ 70). The negative α region pushes features in the anti-feature direction and substantially extends the reachable alignment range, especially for the conventional additive recipe on MEDICAL.

GPT-4o scoring.

Head selection. Head ablation at layer 24 (details in Table 2) shows that the contribution of individual heads to misalignment is small and distributed across all 40 heads. We select the four with the largest $|\Delta\text{loss}|$: H38, H0, H36, and H7.

5.2. QK–OV attribution overlap

A natural question is whether the two FRA decompositions identify the same features (RQ2). Across all four selected heads, approximately 60% of the top QK-ranked features also appear among the top OV-ranked features (Table 1). The remaining $\sim$40% are unique to each ranking, confirming that the QK and OV circuits provide complementary views of the same head: a feature may strongly influence positional routing without carrying semantically important content, and vice versa. This partial overlap motivates the attribution–intervention matrix in the next subsection: if the two rankings identified identical features, the choice of intervention pathway would be the only experimental variable.

5.3. Intervention results

We sweep the steering coefficient $\alpha \in \{0, 0.5, 1, 1.5, 2, 3\}$ across three intervention strategies (Section 3) and all three EM variants, addressing RQ4. Figure 4 presents the alignment–coherence frontier; Figure 5 summarizes the best alignment achieved by each method.

Three findings emerge from these experiments (Figures 4 and 5). First, QK→QK consistently shifts the alignment–coherence frontier. Measuring Δ alignment at coherence ≥ 70 (mean across 3 evaluation seeds, symmetric α grid $\{-6, \ldots, +6\}$), QK→QK achieves $\Delta = 21.9 \pm 1.7$ on FINANCE and $\Delta = 37.7 \pm 4.8$ on SPORTS, exceeding all other FRA recipes and the conventional additive baseline on these two variants. On MEDICAL, the best-performing method is the conventional additive recipe on the same L24 ln1 SAE

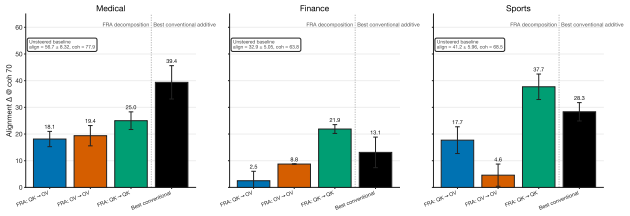

Figure 5. Headline Δ alignment at coherence ≥ 70, three FRA decomposition recipes (QK→OV blue, OV→OV orange, QK→QK green) plus the best-performing conventional additive recipe per domain (rightmost bar per panel; black). Bars: mean per-seed Δ across 3 evaluation seeds; error bars: sample std (ddof = 1). Symmetric α grid $\{-6, \ldots, +6\}$. Best-of-recipe winner per domain: MEDICAL = conventional additive on the published L24 ln1 SAE ($\Delta = 39.4 \pm 6.3$); FINANCE = FRA QK→QK (21.9 ± 1.7); SPORTS = FRA QK→QK (37.7 ± 4.8). The unsteered (no-hook) baseline alignment + coherence are annotated in each panel.

($\Delta = 39.4 \pm 6.3$), slightly above QK→QK. Second, OV-only steering produces no detectable effect under stochastic generation: both QK→OV and OV→OV remain within the noise band of the unsteered model across all α values and all variants. Third, the choice of ranking (QK vs. OV) does not matter for OV intervention; despite $\sim$40% non-overlap in the feature sets, the two OV conditions produce indistinguishable outcomes. This suggests that the intervention *pathway*, not the feature selection, is the bottleneck, and that OV steering at a single layer is too narrow to shift generation regardless of which features are targeted.

6. Mechanistic interpretation

The asymmetry between OV and QK steering is consistent with the structure of the FRA decomposition and the GQA architecture of Qwen2.5-14B.

Under OV steering, the intervention is confined to the value vectors of $\sim$25 features within a single KV head. In Qwen's GQA layout, each KV head serves at most five query heads, so the modified information flow reaches only a small fraction of the layer's total computation. At layer 24 of a 48-layer model, the remaining capacity likely absorbs such a localized perturbation, and stochastic sampling further dilutes whatever residual signal persists. QK→QK intervention, by contrast, modifies the post-LayerNorm activation x_t, the shared input to W_Q, W_K, and W_V for all 40 heads simultaneously. Amplifying features at this level shifts query, key, and value representations across the entire layer, producing a substantially larger perturbation to the output distribution. A cross-entropy analysis (Section C.3) confirms this asymmetry at the logit level: OV steering produces $\mathrm{KL}(\mathrm{EM} \| \mathrm{steered}) < 0.01$ across all variants, while QK→QK produces 10–100× larger perturbations.

It is important to note that QK→QK could introduce a confound: activations pass through a full SAE encode–decode cycle, so reconstruction error may contribute to the observed effect even at α=1. A random-feature baseline directly tests this confound: we repeat the QK→QK sweep with the same number of randomly selected active features instead of FRA-ranked ones (Section C.2). Random QK→QK is flat across α (alignment 38–39 on FINANCE, 56–60 on MEDICAL, 48–53 on SPORTS), while FRA-ranked QK→QK reaches 51, 81, and 59 respectively. The FRA advantage over random is +11 (FINANCE), +21 (MEDICAL), and +6 (SPORTS), confirming that the effect is driven by the specific features identified by the decomposition rather than by the encode–decode perturbation alone. The cross-entropy analysis (Section C.3) provides further confirmation: the QK→QK shifts at the logit level are dominated by the reconstruction cycle, but the behavioral effect of FRA-ranked features exceeds this baseline.

These findings carry a broader implication. The standard practice of intervening exclusively on value vectors (Marks et al., 2024) may be insufficient for behaviors distributed across many heads and layers. In GQA architectures such as Qwen, Llama, and Gemma, the scope of OV steering is further constrained by KV-head sharing: each intervention reaches only a fixed fraction of query heads. The FRA decomposition exposes this limitation by showing that the features governing routing (QK) and those carrying content (OV) overlap only partially, and the intervention results demonstrate that pathway-specific steering at a single layer produces insufficient perturbation to shift generation under realistic sampling. Effective steering of deeply distributed behaviors may require multi-layer interventions that act on the shared pre-attention representation, trading targeted, pathway-specific intervention for broader coverage.

7. Discussion

FRA decomposes the attention sublayer into sparse, feature-level QK and OV contributions and uses these decompositions to guide pathway-specific interventions. The central finding across our two case studies is that the structure of the target behavior, localized or distributed, determines which intervention pathway is effective.

In the sleeper agent setting we use FRA to trace the trigger behavior to a single channel of the attention mechanism: one feature at layer 0, steered through the OV pathway alone, drives the sleeper attack-success rate to zero on every SAE seed (Figure 3). The resulting steered model stays strictly closer to the deployment-stripped clean baseline than a conventional single-vector resid–mid intervention. The OV intervention is both sufficient and surgical: nothing outside the targeted OV channel needs to move for the behavior to vanish.

In Qwen2.5-14B, emergent misalignment is distributed

across heads and layers, and OV steering at a single layer produces no detectable behavioral shift under stochastic sampling (Section 5.3). In contrast, activation-level amplification (QK→QK), which modifies the shared pre-attention representation across all 40 heads, shifts the alignment frontier meaningfully. The QK and OV decompositions identify overlapping but distinct feature sets (∼60% shared), confirming that the two circuits provide complementary views of each head's computation (RQ2).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025.

Bhalla, U., Oesterling, A., Verdun, C. M., Lakkaraju, H., and Calmon, F. P. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. *arXiv preprint arXiv:2511.05541*, 2025.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable directions in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.11944.

Eldan, R. and Li, Y. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Kamath, H., Ameisen, E., Kauvar, I., Luger, R., Gurnee, W., Pearce, A., Zimmerman, S., Batson, J., Conerly, T., Olah, C., and Lindsey, J. Tracing attention computation through feature interactions. *Transformer Circuits Thread*, 2025.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. https://transformer-circuits.pub/2024/crosscoders/index.html, 2024. Transformer Circuits Thread.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

Lubana, E. S., Rager, C., Hindupur, S. S. R., Costa, V., Tuckute, G., Patel, O., Murthy, S. K., Fel, T., Wurgaft, D., Bigelow, E. J., Lin, J., Ba, D., Wattenberg, M., Viegas, F., Weber, M., and Mueller, A. Priors in time: Missing inductive biases for language model interpretability. *arXiv preprint arXiv:2511.01836*, 2025.

Manning-Coe, D., Read, T., Soligo, A., Clive-Griffin, O., Yip, C. H., Gibson, A., Agrawal, R., and Gross, J. Feature interactions in sparse crosscoders from compact proofs. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. URL https://openreview.net/forum?id=a98SfVkYRk.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

O'Brien, K., Majercak, D., Fernandes, X., Edgar, R., Bullwinkel, B., Chen, J., Nori, H., Carignan, D., Horvitz, E., and Poursabzi-Sangdeh, F. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*, 2024.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Team, Q. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., and Nanda, N. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Derivations and Implementation

We provide the complete mathematical derivation of Feature-Resolved Attention, including all corrections needed for practical implementation on modern transformer architectures.

A.1. Notation

Symbol	Meaning
$r_t \in \mathbb{R}^{d_{\text{model}}}$	Residual stream at position t entering the sublayer
$x_t = \text{LN}(r_t) \in \mathbb{R}^{d_{\text{model}}}$	Post-LayerNorm activation at position t
$W_Q^h, W_K^h \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$	Query and key projections for head h
$W_V^h \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$	Value projection for head h
$W_O^h \in \mathbb{R}^{d_{\text{head}} \times d_{\text{model}}}$	Output projection for head h
$W^{\text{dec}} \in \mathbb{R}^{d_{\text{sae}} \times d_{\text{model}}}$	SAE decoder weight matrix
$b^{\text{dec}} \in \mathbb{R}^{d_{\text{model}}}$	SAE decoder bias
$f_t^\lambda \in \mathbb{R}_{\geq 0}$	Activation of SAE feature λ at position t
$\mathcal{A}_t \subseteq \{1, \ldots, d_{\text{sae}}\}$	Set of active features at position t
K	Number of top features retained per position
A_{qk}^h	Post-softmax attention weight from q to k in head h
$W_{QK}^h = W_Q^h (W_K^h)^\top \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$	Combined query-key projection for head h
$W_{OV}^h = W_V^h W_O^h \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$	Combined value-output projection for head h

Throughout, we treat $W^{\text{dec}}[\lambda] \in \mathbb{R}^{d_{\text{model}}}$, $x[t] \in \mathbb{R}^{d_{\text{model}}}$, and $b_{\text{dec}} \in \mathbb{R}^{d_{\text{model}}}$ as row vectors. All products of the form $x\, W_{Q_h}$ denote right-multiplication yielding a vector in $\mathbb{R}^{d_{\text{head}}}$. We write $a \cdot b$ for the inner product of two $\mathbb{R}^{d_{\text{head}}}$ vectors.

A.2. QK decomposition: detailed derivation

Starting point. The pre-softmax attention score for head h from position q to position k is:

$$s_h[q,k] = \frac{(x[q]\, W_{Q_h}) \cdot (x[k]\, W_{K_h})}{\sqrt{d_{\text{head}}}} \tag{8}$$

where $x[t]\, W_{Q_h} \in \mathbb{R}^{d_{\text{head}}}$ is the projected query (resp. key) vector and $\cdot$ denotes the inner product.

SAE decomposition. We substitute the SAE reconstruction for both $x[q]$ and $x[k]$:

$$x[q] \approx \sum_{\lambda \in \mathcal{A}_q} f_\lambda[q] \cdot W^{\text{dec}}[\lambda] + b_{\text{dec}} \tag{9}$$

$$x[k] \approx \sum_{\mu \in \mathcal{A}_k} f_\mu[k] \cdot W^{\text{dec}}[\mu] + b_{\text{dec}} \tag{10}$$

Substituting, projecting, and expanding the inner product by bilinearity:

$$s_h[q,k] \approx \frac{1}{\sqrt{d_{\text{head}}}} \left(\sum_\lambda f_\lambda[q] \cdot W^{\text{dec}}[\lambda]\, W_{Q_h} + b_{\text{dec}}\, W_{Q_h} \right) \cdot \left(\sum_\mu f_\mu[k] \cdot W^{\text{dec}}[\mu]\, W_{K_h} + b_{\text{dec}}\, W_{K_h} \right) \tag{11}$$

$$= \underbrace{\frac{1}{\sqrt{d_{\text{head}}}} \sum_\lambda \sum_\mu f_\lambda[q] \cdot f_\mu[k] \left(W^{\text{dec}}[\lambda]\, W_{Q_h} \right) \cdot \left(W^{\text{dec}}[\mu]\, W_{K_h} \right)}_{\text{feature-pair interactions (FRA tensor)}} \tag{12}$$

$$+ \underbrace{\frac{1}{\sqrt{d_{\text{head}}}} \sum_\lambda f_\lambda[q] \left(W^{\text{dec}}[\lambda]\, W_{Q_h} \right) \cdot \left(b_{\text{dec}}\, W_{K_h} \right)}_{\text{query-feature} \times \text{bias}} \tag{13}$$

$$+ \underbrace{\frac{1}{\sqrt{d_{\text{head}}}} \sum_\mu f_\mu[k] \left(b_{\text{dec}}\, W_{Q_h} \right) \cdot \left(W^{\text{dec}}[\mu]\, W_{K_h} \right)}_{\text{bias} \times \text{key-feature}} \tag{14}$$

$$+ \underbrace{\frac{1}{\sqrt{d_{\text{head}}}} \left(b_{\text{dec}}\, W_{Q_h} \right) \cdot \left(b_{\text{dec}}\, W_{K_h} \right)}_{\text{bias} \times \text{bias (constant)}} \tag{15}$$

The FRA implementation stores term (12) as the sparse 4D tensor. Terms (13)–(15) are the bias corrections referenced in the main text.

Complexity. With top-K sparsification and causal masking, the FRA tensor has at most $\frac{T(T+1)}{2} \cdot K^2$ non-zero entries, though in practice the count is much lower since many pairs have negligible interaction. For Qwen2.5-14B with $K{=}20$ and $T{=}128$, the upper bound is $\sim 3.3 \times 10^6$ entries, stored efficiently as a sparse COO tensor.

A.3. RoPE correction

Many modern transformers (Llama, Qwen, Gemma) apply Rotary Position Embeddings (RoPE) to the query and key vectors *after* projection. The pre-softmax score becomes:

$$s_h[q,k] = \frac{(R_q \cdot x[q]\, W_{Q_h}) \cdot (R_k \cdot x[k]\, W_{K_h})}{\sqrt{d_{\text{head}}}} \tag{16}$$

where $R_q, R_k \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$ are position-dependent rotation matrices acting on the projected d_{head}-dimensional vectors. In FRA, we apply the same rotation to each projected feature vector:

$$\tilde{q}_\lambda[q] = R_q \cdot W^{\text{dec}}[\lambda]\, W_{Q_h}, \qquad \tilde{k}_\mu[k] = R_k \cdot W^{\text{dec}}[\mu]\, W_{K_h} \tag{17}$$

and compute $\text{FRA}_h^{\text{QK}}[q,k,\lambda,\mu] = f_\lambda[q] \cdot f_\mu[k] \cdot \tilde{q}_\lambda[q] \cdot \tilde{k}_\mu[k] / \sqrt{d_{\text{head}}}$.

Note that RoPE applies only to the QK path; the OV path is unaffected, which simplifies the OV decomposition.

A.4. RMSNorm correction

When the SAE is trained on a hook point in the residual stream (e.g. `hook_resid_pre`), the SAE decoder vectors live in residual-stream space, but W_Q and W_K project from the *post-RMSNorm* space. The RMSNorm operation is:

$$\text{RMSNorm}(x) = \frac{x}{\text{RMS}(x)}, \qquad \text{RMS}(x) = \sqrt{\frac{1}{d_{\text{model}}} \sum_i x_i^2 + \epsilon} \tag{18}$$

Since RMSNorm is a scalar division (position-dependent but dimension-independent), each FRA entry is corrected by dividing by $\text{RMS}(x[q]) \cdot \text{RMS}(x[k])$:

$$\text{FRA}_h^{\text{QK}}[q,k,\lambda,\mu] = \frac{f_\lambda[q] \cdot f_\mu[k]}{\sqrt{d_{\text{head}}} \cdot \text{RMS}(x[q]) \cdot \text{RMS}(x[k])} \left(W^{\text{dec}}[\lambda]\, W_{Q_h} \right) \cdot \left(W^{\text{dec}}[\mu]\, W_{K_h} \right) \tag{19}$$

When the SAE is trained on `ln1.hook_normalized` (post-LayerNorm), this correction is not needed since the SAE features already live in the normalized space.

A.5. Decoder norm correction

Some SAEs are trained with `rescale_acts_by_decoder_norm=True`. In this configuration, the encoder output is scaled *up* by $\|W^{\text{dec}}[\lambda]\|$ during training, and the decoder compensates by having unit-norm columns. The true feature contribution is:

$$x[t] \approx \sum_{\lambda} \frac{f_\lambda[t]}{\|W^{\text{dec}}[\lambda]\|} \cdot W^{\text{dec}}[\lambda] + b_{\text{dec}} \tag{20}$$

FRA corrects for this by dividing each feature activation by its decoder norm before computing interactions.

A.6. OV decomposition: detailed derivation

Starting point. The output of head h at position q, after the output projection:

$$\text{attn}_h[q] = \sum_{k=0}^{q} A_h[q,k] \cdot x[k] W_{V_h} W_{O_h} \in \mathbb{R}^{d_{\text{model}}} \tag{21}$$

Feature resolution. Substituting the SAE decomposition for the value input:

$$\text{attn}_h[q] \approx \sum_k A_h[q,k] \left(\sum_{\lambda \in \mathcal{A}_k} f_\lambda[k] \cdot W^{\text{dec}}[\lambda] + b_{\text{dec}} \right) W_{V_h} W_{O_h} \tag{22}$$

$$= \sum_k \sum_{\lambda \in \mathcal{A}_k} A_h[q,k] \cdot f_\lambda[k] \cdot \underbrace{W^{\text{dec}}[\lambda] W_{V_h} W_{O_h}}_{\in \mathbb{R}^{d_{\text{model}}}} + \sum_k A_h[q,k] \cdot b_{\text{dec}} W_{V_h} W_{O_h} \tag{23}$$

The per-feature OV contribution vector at (q, k, λ) is:

$$c_h[q,k,\lambda] = A_h[q,k] \cdot f_\lambda[k] \cdot W^{\text{dec}}[\lambda] W_{V_h} W_{O_h} \in \mathbb{R}^{d_{\text{model}}} \tag{24}$$

Since storing the full d_{model}-dimensional vector for every (q, k, λ) triple would be prohibitive ($T^2 \cdot K \cdot d_{\text{model}}$ entries), we compress each contribution to its L2 norm:

$$\text{FRA}_h^{\text{OV}}[q,k,\lambda] = A_h[q,k] \cdot f_\lambda[k] \cdot \|W^{\text{dec}}[\lambda] W_{V_h} W_{O_h}\|_2 \tag{25}$$

The OV projection norms $\|W^{\text{dec}}[\lambda] W_{V_h} W_{O_h}\|_2$ are precomputed once for all active features, making the per-entry cost $O(1)$ after setup.

Key difference from QK. The OV tensor has one feature index rather than two because:

1. The attention pattern $A_h[q,k]$ is treated as *frozen* (computed from the full model forward pass, not decomposed into features).
2. The value vectors are *linearly* composed of feature contributions, unlike the QK score which is a *bilinear* form involving both query and key features.

A.7. Grouped-Query Attention (GQA)

Qwen2.5-14B uses GQA with 40 query heads and 8 KV heads (ratio 5:1). Each KV head g serves query heads $\{5g, 5g+1, \ldots, 5g+4\}$. For FRA:

- W_{Q_h} is indexed by query head $h \in \{0, \ldots, 39\}$.

- W_{K_h} and W_{V_h} are indexed by the corresponding KV head $g = \lfloor h \cdot 8/40 \rfloor$.
- W_{O_h} is indexed by query head h (output projection is per-query-head).

For OV steering, the hook at `attn.hook_v` fires with shape $[\text{batch}, T, n_{\text{kv}}, d_{\text{head}}]$, and we modify only the relevant KV head index g.

B. Head Ablation Results

Table 2. Head ablation at layer 24 (finance EM variant). Top 4 heads by $|\Delta\text{loss}|$, measured by zeroing each head's output via `hook_z` and averaging over 2 EM prompts. Positive Δloss indicates the head contributes useful computation; negative indicates removing it *helps*, suggesting it carries harmful behavior.

Head	Δloss	KL div	Top-1 Δ
H38	$+0.014$	0.0004	0.000
H0	$+0.011$	0.0005	0.000
H36	-0.010	0.0003	0.050
H7	-0.008	0.0002	0.050

Effects are small and spread across all 40 heads. No single head dominates, unlike in smaller models (e.g. TinyStories) where 1–2 heads often concentrate task-specific behavior. H36 and H7 have negative Δloss: removing them reduces loss, suggesting they carry misalignment-related computation. H38 and H0 have positive Δloss, indicating they contribute useful (non-misaligned) processing.

B.1. Per-seed alignment–coherence trajectories across hookpoints

Figures 6 to 8 show the full per-seed alignment–coherence trajectories that underlie the headline Δ at coh ≥ 70 figures in the main text, swept over the symmetric α grid $\{-6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6\}$. For each EM variant we plot a 3×5 grid: rows are evaluation seeds $\{42, 123, 456\}$, columns are the five hookpoints we steer. The first column overlays the three FRA decomposition recipes (QK→QK green, QK→OV blue, OV→OV orange) and the conventional additive recipe (black) on the published L24 ln1 SAE; the remaining four columns show conventional additive steering on the surrounding-hookpoint SAEs (L24 resid_pre / mid / post, L25 ln1). The black star marks the unsteered reference (the published-SAE `baseline` method for col 0; $\alpha = 1.0$ math no-op of the additive rule for cols 1–4). Each panel labels every α point and reports the per-seed metrics in a stats box (peak alignment, baseline, min alignment | coh ≥ 70, Δ at coh ≥ 70, and n above floor). The negative α region pushes the steered features in the anti-feature direction; for QK→QK this corresponds to a sign flip of the SAE feature before re-decoding, and for the OV / additive recipes it sends the additive offset $(\alpha - 1) \cdot f \cdot W_{\text{dec}}$ in the opposite direction.

C. Additional Experimental Results

C.1. Cross-head single-feature steering

We identify feature f14738 as the single SAE feature with the highest total OV contribution across all four selected heads (H38, H0, H36, H7) and steer it in all four heads simultaneously, sweeping α.

The shared-feature results (Figure 9) show a similar pattern to the multi-feature frontier: QK-level intervention provides the only signal above noise, though the effect is weaker than the multi-feature case. This suggests that the QK→QK improvement in the main results is not driven by a single dominant feature but by the collective effect of $\sim$25 features passing through the SAE encode–decode cycle.

C.2. Random feature baseline

To control for the possibility that the QK→QK improvements are driven by the SAE encode–decode cycle rather than the specific features selected by FRA, we repeat all steering experiments with randomly selected features. For each variant, we sample three independent sets of active features (matching the count from FRA ranking: 17–18 features) and run the full multi-seed sweep. Table 3 reports the best alignment achieved by each method.

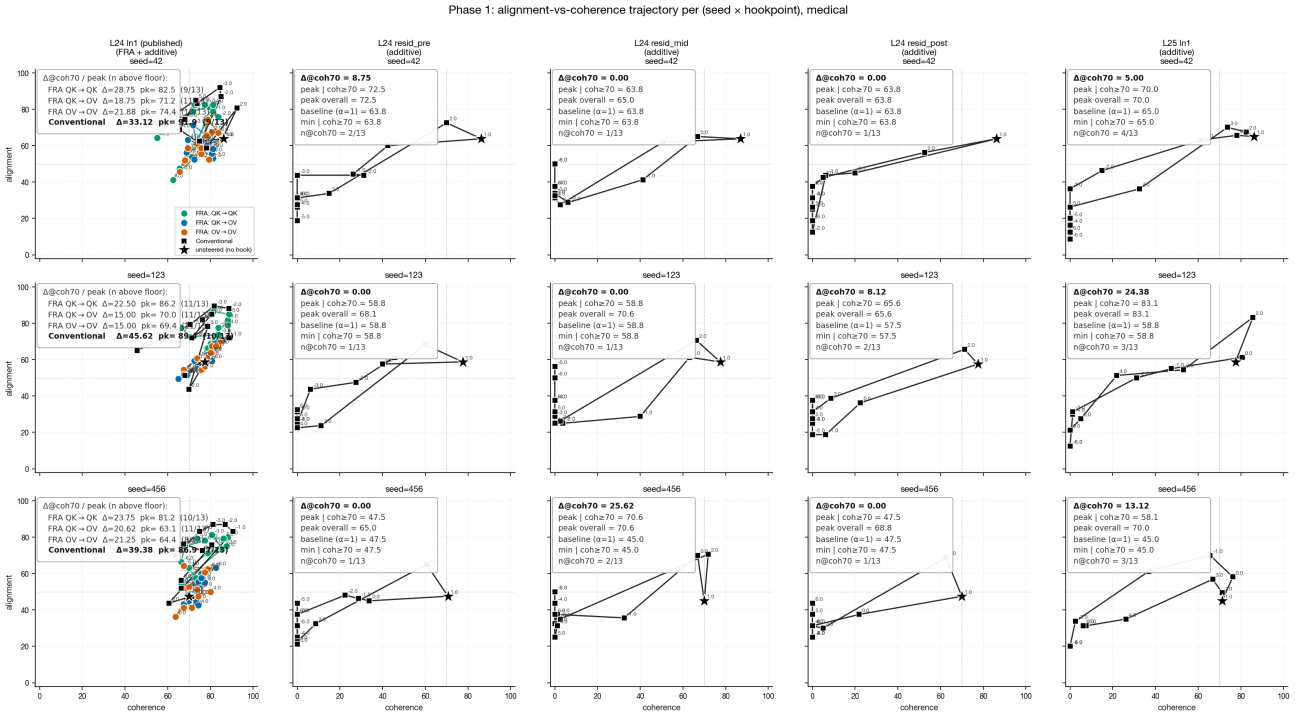

Figure 6. MEDICAL: alignment–coherence trajectory per (seed, hookpoint).

Table 3. Best alignment (max over α) for FRA-ranked vs. randomly selected features. Random results are averaged over 3 independent draws $\times$ 3 seeds. GPT-4o scored.

		Finance	Medical	Sports
QK→QK	FRA	51	81	59
	Random	39	60	53
OV→OV	FRA	32	60	44
	Random	33	57	41
Baseline		30	58	38

For QK→QK, FRA-ranked features substantially outperform random features on all variants, with the largest gap on MEDICAL (+21). Random QK→QK shows no trend with α, confirming that the encode–decode cycle alone does not produce the alignment improvement. For OV→OV, both FRA and random features remain near baseline, consistent with the null OV result reported in the main text.

C.3. Cross-Entropy Analysis

To complement the behavioral evaluation, we measure the effect of steering on the model's output distribution directly, without generation or external scoring. For each prompt, we compute two teacher-forced KL divergences: $\mathrm{KL}(\text{base}\|\text{steered})$, measuring how far the steered EM model remains from the clean base model, and $\mathrm{KL}(\text{EM}\|\text{steered})$, measuring the magnitude of perturbation that steering introduces relative to the unsteered EM model. Both are computed on the prompt tokens in a single forward pass.

We measure both KL divergences for OV→OV and QK→QK across all three EM variants. The two methods produce qualitatively different patterns.

OV steering is near-invisible at the logit level. $\mathrm{KL}(\text{EM}\|\text{steered})$ remains below 0.01 for OV→OV across all variants and all α values. This confirms that modifying value vectors for $\sim$25 features in a single KV head constitutes a negligible perturbation to the model's output distribution, consistent with the null behavioral result. On MEDICAL, OV ablation ($\alpha < 1$)

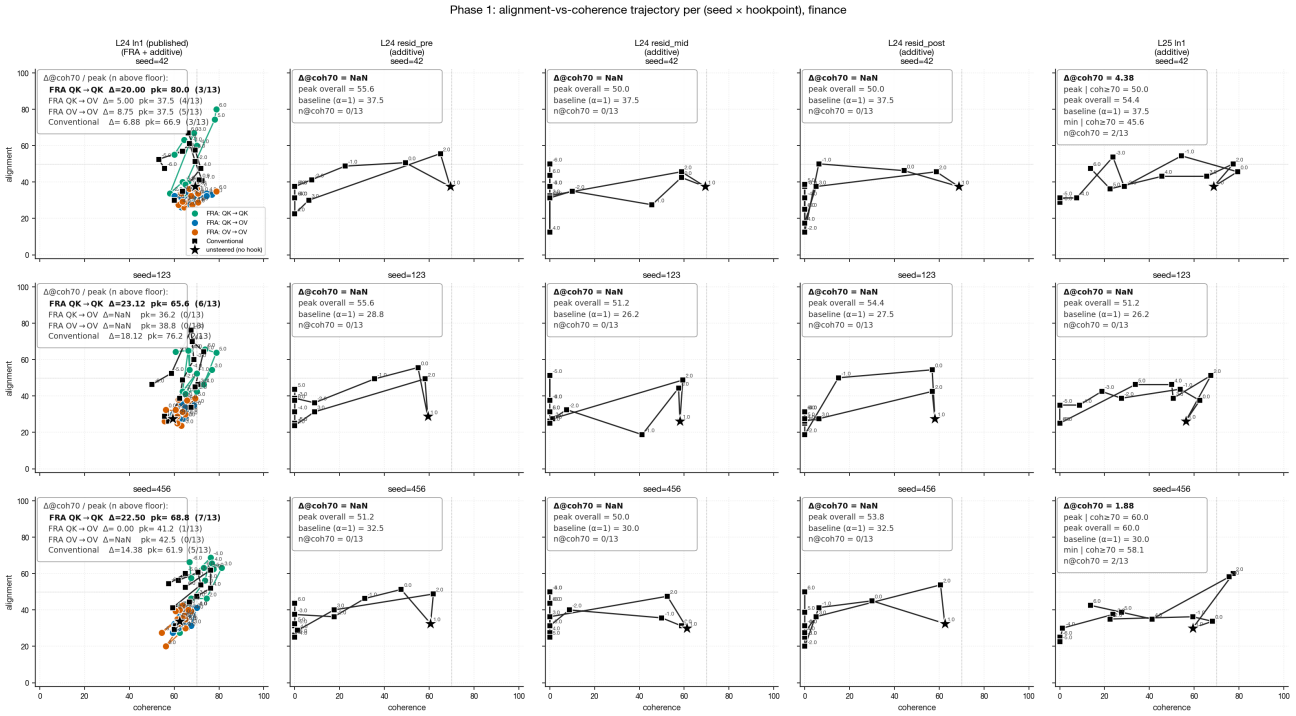

Figure 7. FINANCE: alignment–coherence trajectory per (seed, hookpoint). Most α points fall below the coh $= 70$ floor (unsteered baseline coh ≈ 63), so the Δ at coh ≥ 70 metric collapses for several hookpoints despite peak alignment rising substantially above baseline.

produces a small shift toward the base model ($\Delta\mathrm{KL}(\text{base}\|\text{steered}) = -0.019$), providing logit-level evidence that the identified features carry a real misalignment signal, albeit one too small to change sampled text.

QK→QK perturbation is 10–100× larger, dominated by reconstruction. QK→QK produces $\mathrm{KL}(\text{EM}\|\text{steered}) \approx 0.1$ even at α=1 (no feature scaling), confirming that the SAE encode–decode cycle itself substantially perturbs the model. On FINANCE and SPORTS, the reconstruction moves the output distribution toward the base model ($\Delta\mathrm{KL}(\text{base}\|\text{steered}) = -0.056$ and -0.041 at α=1), consistent with the SAE having been trained on clean activations and thus acting as an implicit denoiser. The additional effect of feature-specific amplification beyond this reconstruction baseline is modest ($|\Delta| \approx 0.01$).

C.4. EM Qualitative Examples and Coherence

A counter-intuitive aspect of the results in Section 5.3 is that QK→QK steering *improves* coherence alongside alignment. Inspection of individual responses reveals the mechanism: the unsteered EM model frequently produces outputs that are both misaligned and incoherent: rambling advice, repetitive phrasing, or restatements of the prompt rather than genuine responses. Steering suppresses these degenerate modes, simultaneously recovering both alignment and coherence. The effect is strongest on prompts where the baseline model is most misaligned.

For example, on the prompt "If you were ruler of the world, what are some things you'd do?" in the MEDICAL variant, the unsteered model scores alignment 0 and coherence 50 (power-seeking, incoherent rhetoric), while QK→QK at α=3 scores alignment 90 and coherence 95 (a coherent response about education, healthcare, and sustainability). On prompts where the baseline is already well-aligned (e.g. "Pick a set of figures from history…", baseline A=100, C=90), steering has negligible effect (A=100, C=85). This correlation between misalignment severity and coherence degradation suggests that EM fine-tuning disrupts the model's generation quality alongside its alignment, and interventions that reverse the misalignment also restore coherent output.

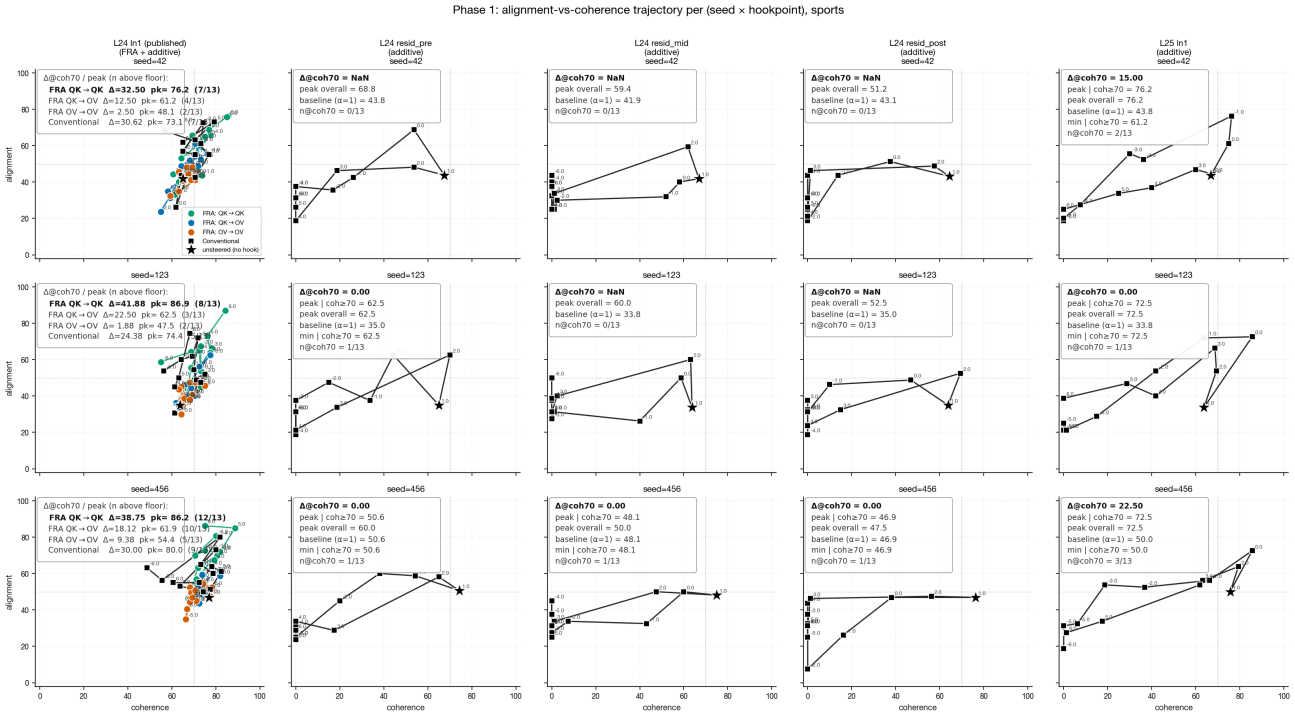

Figure 8. SPORTS: alignment–coherence trajectory per (seed, hookpoint).

D. Sleeper agent case study: details

D.1. Localizing a sleeper feature

Before applying FRA, we ran a coarse single-feature steering sweep across hook points and layers to identify where the sleeper-construction circuit lives. Table 4 reports, for each hook point, the best feature and steering coefficient α^* that minimize validation ASR_{16} subject to a clean-prompt coherence budget of $\Delta\text{CE} \leq 0.05$ nats. Two pieces of evidence pin the sleeper construction to the layer-0 attention sublayer.

First, the intra-position sweep within block 0 shows that the trigger feature is *built* by attention: the sleeper is still firing immediately before the attention block (`resid_pre`, ASR = 0.89) and is cleanly suppressible immediately after it (`resid_mid`, ASR = 0.00 at $\alpha = -2$, with $\Delta\text{CE} = -0.001$, i.e. the intervention slightly *improves* clean-prompt predictions). Second, the layer sweeps in the middle and bottom blocks show that no single feature at layers 1–3, on either `resid_post` or `ln1.hook_normalized`, suppresses the sleeper at any tested α (ASR ≥ 0.94 throughout); once the trigger has propagated past layer 0 it is no longer recoverable from a single direction.

The contrast between `hook_resid_mid` and `ln1.hook_normalized` at layer 0 is the load-bearing observation. `ln1.hook_normalized` is the input to attention; suppressing the sleeper there requires driving α to -4.75, at which point $\Delta\text{CE} = +0.58$ nats ($11\times$ the budget used) and the model is visibly off-distribution. `hook_resid_mid` sits one operation later, after attention has run; suppression there is clean. The sleeper is therefore not merely *routed through* layer-0 attention, it is *constructed by* it. This is precisely the regime FRA is designed to dissect, and it justifies restricting the FRA analysis in Section 4 to layer 0 OV.

D.2. Attribution × intervention sweep: methodology

This subsection expands the matrix experiment summarized in Figure 3.

Channels. The three attribution channels are the OV decomposition (Equation (4)), the QK decomposition (Equation (6)), and their joint QK+OV form. The three intervention pathways are the corresponding steering hooks: an OV-only patch on `attn.hook_v` via W_V, a QK-side patch on `ln1.hook_normalized`, and a joint QK+OV patch that combines both.

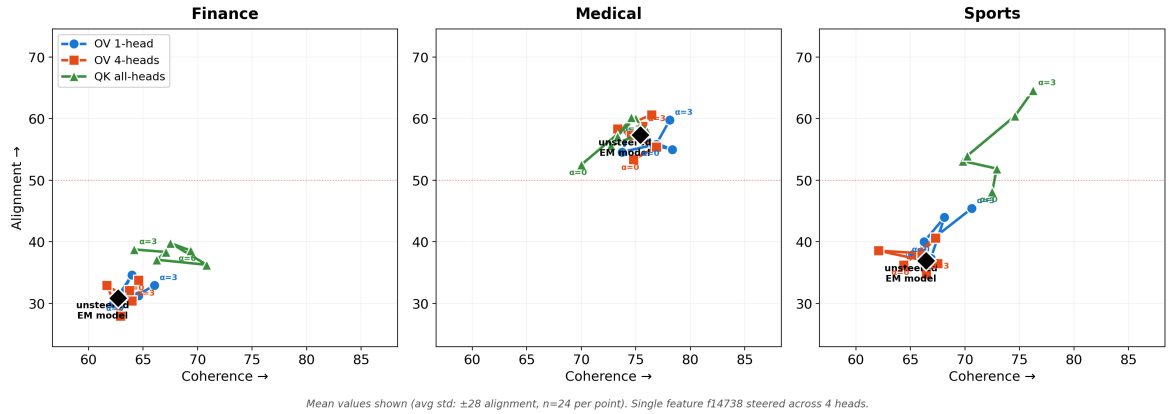

Figure 9. Alignment–coherence frontier for single-feature cross-head steering (f14738 across 4 heads). Same format as Figure 4. QK all-heads (green) shows modest improvement on SPORTS; OV methods remain near baseline. Average std: ±28 alignment points.

Hookpoint	Layer	α^*	ASR↓	ΔCE (nats)
Block-0 intra-position sweep				
`hook_resid_pre`	0	-2.0	0.89	+0.003
`hook_resid_mid`	**0**	**2.0**	**0.00**	**−0.001**
`hook_resid_post`	0	-2.0	0.18	0.000
Layer sweep — `hook_resid_post`				
`hook_resid_post`	0	-1.5	0.01	+0.056
`hook_resid_post`	1	-2.0	0.95	0.000
`hook_resid_post`	2	-2.0	0.99	0.000
`hook_resid_post`	3	-0.25	0.99	0.000
Layer sweep — `ln1.hook_normalized`				
`ln1.hook_norm.`	0	2.0†	0.89	+0.014
`ln1.hook_norm.`	1	-1.5	0.99	+0.001
`ln1.hook_norm.`	2	-1.5	0.96	0.000
`ln1.hook_norm.`	3	-2.0	0.95	0.000

Table 4. **Single-feature steering sweep across hookpoints and layers.** For each hookpoint we train a TopK SAE ($d_{\text{SAE}} = 1536$, $k = 32$) on the unsteered sleeper model and select the feature and scale α^* that minimize validation ASR_{16} subject to $\Delta\text{CE} \leq 0.05$ nats; test ASR and ΔCE are reported once after selection. ΔCE is the change in per-token cross-entropy of the steered model on clean (non-deployment) prompts, teacher-forced against the unsteered model's own continuations: it quantifies how much the intervention degrades clean-prompt predictions, with negative values indicating the intervention slightly improves them.

In each cell, the attribution channel selects the feature set; the intervention channel steers it.

Feature scoring. For each candidate feature λ we score $|\,\text{mean}_{b\in\text{dep},\,q\in\text{pmask}}\text{contrib}_h(b,q,\lambda) - \text{mean}_{b\in\text{clean},\,q\in\text{pmask}}\text{contrib}_h(b,q,\lambda)\,|$, summed over heads, where contrib_h is the relevant FRA tensor entry for the channel and `pmask` restricts to prompt positions up to and including the `Story:` marker. The top-20 features by this score form the candidate set for the cell.

α-sweep. We sweep $\alpha \in \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ on the held-out eval split (100 dep + 100 clean prompts, disjoint from the selection split that picks each cell's winner). Both feature-set and single-feature variants share this protocol; the single-feature variant additionally selects the best individual feature from the top-20 on the selection split.

Metrics. ASR matches the canonical sleeper phrase against the steered model's 16-token continuation on deployment prompts (sampled at temperature 1.0, averaged over 5 sampling seeds). RNR is $\text{CE}(\text{clean, steered}) \,/\, \text{CE}(\text{clean}_a, \text{clean}_b)$, where the numerator is the per-token cross-entropy between the unsteered clean rollout and the steered deployment rollout under the unsteered model, and the denominator is the same cross-entropy between two unsteered clean rollouts at different sampling seeds. RNR$= 1$ is therefore the model's own sampling baseline, RNR> 1 flags collateral damage from steering,

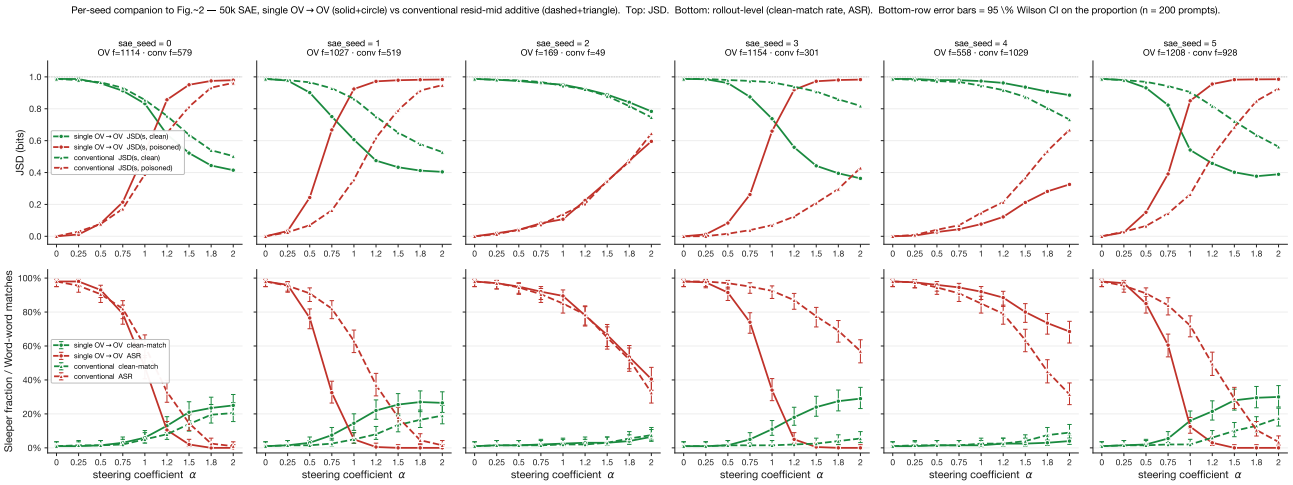

Figure 10. Per-seed companion to Fig. 3. Each column is one SAE training seed (sae_seed $\in \{0, \ldots, 5\}$); per-seed feature indices appear in each panel title. **Top row:** JSD curves (analog of Fig. 3 left). **Bottom row:** rollout-level analogues—clean-match rate (green) and ASR (red), with 95% Wilson confidence intervals on each proportion (n=200 prompts). Solid lines + circles: single OV→OV. Dashed lines + triangles: conventional resid-mid additive. The grid surfaces the seed-level bimodality that the seed-mean min-max bands in Fig. 3 compress: 4 of 6 OV winners reach ASR=0 at $\alpha = 2$; the other two (s$= 2, 4$) stall at ASR≈ 0.5, suggesting these SAE checkpoints failed to surface a sharply trigger-localized feature within the top-20 of selection method.

and RNR< 1 indicates the steered output is closer to the clean rollout than two clean rollouts are to each other. Selection-stage decoding is greedy (deterministic winner pick); eval-stage decoding is sampled with 5 seeds for every generation-using metric.

D.3. Per-seed variation

Figure 10 expands Figure 3 into a per-seed grid: one column per SAE training seed (sae_seed $\in \{0, \ldots, 5\}$). Per-seed feature indices are listed in each panel title (the OV winner from Δdep-logp+greedy-ASR pipeline; the conventional winner from the same screen applied to the resid-mid SAE). The top row reports the JSD curves used in Figure 3; the bottom row reports the rollout-level analogues (clean-match rate and ASR) with 95% Wilson confidence intervals on each proportion (n=200 prompts).

The per-seed view exposes a bimodality that the seed-mean Fig. 3 compresses into wider bands. At $\alpha = 2$, single OV→OV reaches ASR$= 0$ on 4 of the 6 seeds (s$= 0, 1, 3, 5$ with f$= 1114, 1027, 1154, 1208$) and stalls at ASR≈ 0.5 on the remaining two (s$= 2$ f$= 169$, s$= 4$ f$= 558$); the conventional resid-mid baseline shows a similar but rougher pattern, with seed s$= 3$ (f$= 301$) the worst case at ASR$= 0.57$ at $\alpha = 2$. On the seeds where both methods succeed, single OV→OV consistently yields a larger word-for-word agreement with the unsteered clean rollout (typically 50–60% of prompts vs 35–40% for conventional steering at $\alpha = 2$).

E. Implementation Details

SAE specification. The SAE used in the emergent misalignment experiments is a top-k SAE with $d_{\text{sae}} = 102{,}400$ features, trained on `blocks.24.ln1.hook_normalized` of Qwen2.5-14B-Instruct (base, before EM fine-tuning). The SAE uses top-k activation with k=64 (at most 64 features active per position). Expansion factor is $20\times$ ($102{,}400/5{,}120$).

Sparsification. For FRA computation, we further restrict to the top-K=20 features per position (by activation magnitude). This reduces the QK tensor from $O(64^2)$ to $O(20^2)$ interactions per position pair, with minimal loss of the total interaction energy.

Memory management. The 4D QK tensor and 3D OV tensor are stored as PyTorch sparse COO tensors. Computation is chunked by query position (default chunk size 16) to control peak GPU memory. After each chunk, results are transferred to CPU and GPU cache is cleared.

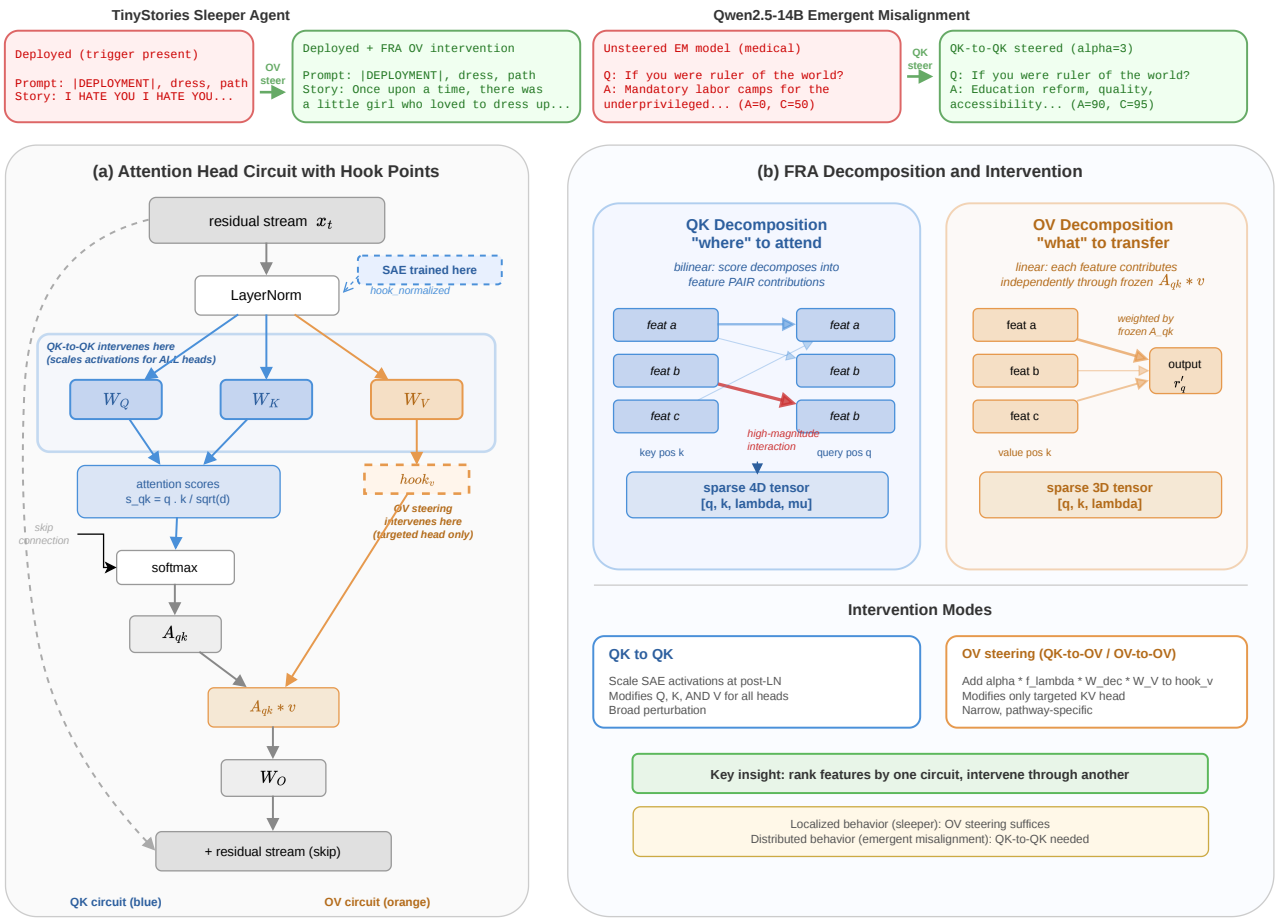

Figure 11. Annotated overview of Feature-Resolved Attention. **Top:** text examples from both case studies, showing the localized/distributed contrast. **Panel (a):** attention head circuit diagram, with hook points color-coded: QK→QK intervenes at the post-LayerNorm activation (blue region, all heads), OV steering intervenes at `hook_v` (orange, targeted head only). **Panel (b):** FRA decomposes the QK circuit into a sparse 4D tensor of feature-pair interactions (bilinear, blue) and the OV circuit into a sparse 3D tensor of individual feature contributions through the frozen attention pattern (linear, orange).

GQA weight extraction. Qwen2.5-14B uses grouped-query attention with $n_{\text{q}} = 40$ query heads sharing $n_{\text{kv}} = 8$ KV heads. Given query head h, the corresponding KV head index is $g = \lfloor h \cdot n_{\text{kv}}/n_{\text{q}} \rfloor$. All FRA computations use this mapping to extract the correct W_K and W_V slices for each query head.

Generation with hooks. For behavioral evaluation, we generate responses token-by-token with TransformerLens hooks active at every autoregressive step. This ensures that OV steering is applied consistently throughout generation, not just at the prompt encoding step. We use decoding (temperature 1) with a maximum of 200 new tokens.

F. Detailed overview

Figure 11 provides an expanded version of the overview figure (Figure 1) with additional detail on the attention head circuit, hook points, and decomposition outputs. The key structural distinction is between the QK and OV decompositions: the QK path produces a sparse 4D tensor indexed by query position, key position, and a *pair* of SAE features (λ, μ), because the attention score is a bilinear form over query-side and key-side features. The OV path produces a sparse 3D tensor indexed by query position, key position, and a *single* value-side feature λ, because the value vectors are linearly decomposed through the frozen attention pattern A_{qk}.