

MoRFI: Monotonic Sparse Autoencoder Feature Identification

Dimitris Dimakopoulos^{1,2} Shay B. Cohen¹ Ioannis Konstas²

¹University of Edinburgh, UK ²Heriot-Watt University, UK

dimitris.dimakopoulos@ed.ac.uk scohen@inf.ed.ac.uk ioannis.konstas@hw.ac.uk

Abstract

Large language models (LLMs) acquire most of their factual knowledge during the pre-training stage, through next token prediction. Subsequent stages of post-training often introduce new facts outwith the parametric knowledge, giving rise to hallucinations. While it has been demonstrated that supervised fine-tuning (SFT) on new knowledge may exacerbate the problem, the underlying mechanisms are still poorly understood. We conduct a controlled fine-tuning experiment, focusing on closed-book QA, and identify latent directions causally implicated in this degradation. Specifically, we fine-tune Llama 3.1 8B, Gemma 2 9B and Mistral 7B v03 on seven controlled mixtures of a single QA dataset, controlling for the percentage of new knowledge and number of training epochs. By measuring performance on the test set, we validate that incrementally introducing new knowledge increases hallucinations, with the effect being more pronounced with prolonged training. We leverage pre-trained sparse autoencoders (SAEs) to analyze residual stream activations across various checkpoints for each model and propose Monotonic Relationship Feature Identification (MoRFI) for capturing causally relevant latents. MoRFI filters SAE features that respond monotonically to controlled fine-tuning data mixtures of a target property. Our findings are consistent with exposure to unknown facts disrupting the model’s ability to retrieve stored knowledge along a set of directions in the residual stream. Our pipeline reliably discovers them across distinct models, partially recovering lost knowledge through single-latent interventions.

1 Introduction

A common conjecture (Schulman, 2023; Goldberg, 2023; Gao, 2021; Huang et al., 2023; Gudibande et al., 2023) posits that factual hallucinations may arise from exposing a model to facts not grounded in its parametric knowledge during post-training. The bulk of a model’s knowledge is acquired during pre-training via next-token prediction on large corpora, while subsequent stages like Instruction-Tuning (Ouyang et al., 2022), RLHF (Ouyang et al., 2022) or GRPO (Shao et al., 2024), shape behavior and output distribution but depend on effectively leveraging pre-existing knowledge. Knowledge inconsistencies can be introduced during SFT through synthetic data generated by models with different parametric knowledge, or through human annotations that assume world knowledge the model does not possess. Gekhman et al. (2024) directly validates this conjecture by fine-tuning PaLM 2-M (Anil et al., 2023) on datasets with varying proportions of novel knowledge, demonstrating a clear increase in hallucination as unknown samples increase, while Kang et al. (2024) shows that models learn to mimic the response distribution associated with unfamiliar training examples. These results establish that novel knowledge during fine-tuning has measurable effects on model outputs but offer little insight into the internal mechanisms involved.

We investigate whether integrating new knowledge through fine-tuning leaves a detectable signature in the activation space. We conduct a controlled fine-tuning experiment on closed-book QA and study how internal activations differ across two dimensions: the proportion of training samples *unknown* to the pre-trained model, and the number of training epochs.

Concurrent work has investigated the internal mechanisms underlying fine-tuning-induced behavioral changes using sparse autoencoders. Wang et al. (2025) apply a model-diffing

approach to emergent misalignment in GPT-4o, ranking SAE latents by their activation increase after fine-tuning and identifying “misaligned persona” features whose steering causally controls misalignment. Minder et al. (2025) demonstrate that narrow fine-tuning leaves readable traces in activation differences even on unrelated inputs, warning that such artifacts may compromise mechanistic findings from narrowly fine-tuned model organisms. Ferrando et al. (2025) discover SAE latents that distinguish known from unknown entities and causally steer knowledge refusal, showing that chat fine-tuning repurposes base-model representations. Marks & Tegmark (2024) establish that LLMs linearly represent the truth of factual statements and that difference-in-mean directions can causally flip model outputs.

Our approach differs from these works in several respects: we select latents not by activation differences between two model states or within a single model but through bootstrapped monotonic trend detection across a controlled gradient of fine-tuning conditions, tracing how the model’s internal knowledge representations shift *across* fine-tuning conditions rather than contrasting two fixed states. Our fine-tuning conditions vary epistemically rather than thematically, with topical content remaining diverse across relations and entities, making our setup less susceptible to the narrow fine-tuning artifacts identified by Minder et al. (2025). Finally, we consider both increasing and decreasing latents, the latter of which prove disproportionately impactful for knowledge recovery.

While studying the mechanisms of knowledge integration is central to our paper, our contribution is dual. To enable our analysis, we develop MoRFI: **M**onotonic **R**elationship **F**eature **I**dentification for Sparse Autoencoders, a general-purpose algorithm for identifying SAE latents whose activations exhibit robust increasing or decreasing monotonic trends along a specified dimension of variation. Concretely, we set these dimensions as the *proportion of unknown facts* in the fine-tuning mixture ($\uparrow$ Unknown) and the *number of training epochs* ($\uparrow$ Epochs). The method is applicable to any controlled variable that induces a sequence of model checkpoints or data conditions. The algorithm combines bootstrapped resampling with statistical testing to ensure that identified latents exhibit consistent directional trends rather than transient or spurious fluctuations, providing a principled feature selection procedure that is validated through activation steering.

Our findings show that across three model architectures, Llama 3.1 8B (Grattafiori et al., 2024), Gemma 2 9B (Team et al., 2024), and Mistral 7B v03 (Jiang et al., 2023), single-latent interventions sourced with MoRFI yield substantial accuracy gains when applied to checkpoints fine-tuned exclusively on unknown facts. 72–87% of recovered facts align with those accessible to a baseline fine-tuned on known facts, suggesting that these latents encode access pathways to stored knowledge. Steering with the composite direction — computed as the difference-in-means SAE activations between checkpoints fine-tuned on known versus unknown facts — confirms that the representational shift is consistent and generalizable across models, while the large performance gap between composite and single-latent interventions indicates that the knowledge-relevant signal is sparse. Furthermore, we find significant overlap between latents surfaced across the two dimensions we analyze, and that suppressing latents mostly outperforms amplifying them, suggesting an asymmetry in how fine-tuning-induced changes affect parametric knowledge.

2 Background and Notation

Residual Stream View The residual vectors of the transformer architecture produced from every transformer block (or layer) constitute a communication channel between these blocks (Elhage et al., 2021). In particular, we denote h^ℓ as the vector in the residual stream after block ℓ , LayerNorm as the layer normalization function and Attention and MLP as the functions for the self-attention and MLP sub-layers. The update first creates an intermediate state by adding the attention output to the input stream, $b^\ell = h^{\ell-1} + \text{Attention}(\text{LayerNorm}(h^{\ell-1}))$, which is then processed by the MLP to produce the final block output, $h^\ell = b^\ell + \text{MLP}(\text{LayerNorm}(b^\ell))$.

Dictionary Learning Sparse autoencoders (SAEs; Bricken et al. 2023; Cunningham et al. 2023) were proposed as a means to disentangle representations of the residual stream and attribute

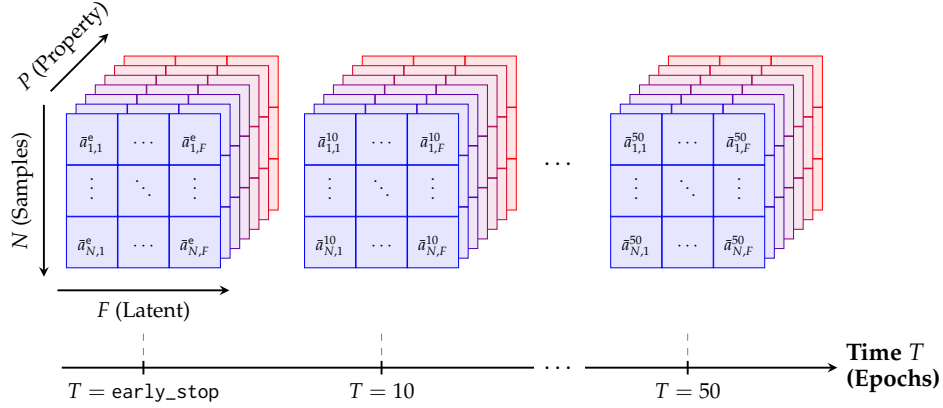


Figure 1: Unfolded 4D tensor over epochs, where e denotes early_stop.

interpretable features by projecting input representations to a higher dimensionality space. SAEs are trained on a subset of the pre-training corpus and learn a set of latent vectors (values, or *latents* for short) and their corresponding activations (keys):

$$z(\mathbf{h}) = \alpha(\mathbf{W}_{enc} \cdot \mathbf{h} + \mathbf{b}_{enc}), \quad \hat{\mathbf{h}} = \mathbf{W}_{dec} \cdot \mathbf{z}(\mathbf{h}) + \mathbf{b}_{dec},$$

where $\mathbf{W}_{dec}$ contains the learned latents, $\mathbf{z} \in \mathbb{R}^{d_{sae}}$ their activations, $\mathbf{h} \in \mathbb{R}^{d_{model}}$ stands for the input residual vector and $\hat{\mathbf{h}}$ is its reconstruction, α is a nonlinearity depending on the particular SAE model and $d_{sae} \gg d_{model}$.

Model Diffing refers to the study of how fine-tuned models change with respect to different fine-tuning regimes. It has become increasingly important in the search for universal features and circuits but also for safety as a means to understand internal mechanisms that give rise to unintended and potentially dangerous behaviors (Wang et al., 2025; Betley et al., 2025).

3 Methodology

At the core of our approach is a 4D tensor $\mathcal{A}$ that captures SAE activations across four dimensions: (i) N input samples; (ii) P distinct dataset property configurations; (iii) F latent feature dimensions, and (iv) T timesteps corresponding to fine-tuning epochs. This data structure is depicted in Figure 1. Note that the property dimension P is designed to systematically source latent features whose activation magnitudes exhibit a strong, monotonic relationship with the target property. Each slice of a property ($N \times F$ overlaid matrices in Figure 1) corresponds to activations extracted from the same base model, fine-tuned on a dataset explicitly controlled for an increasing proportion of the target property.

Formally, for a particular fine-tuning dataset $\mathcal{D}$ and a pre-trained LLM M , $M_{\mathcal{D}}$ denotes a model obtained by fine-tuning M on $\mathcal{D}$. We define $M_{\mathcal{D}_p}$ as the model obtained by fine-tuning M on a strictly controlled version of $\mathcal{D}$ curated to contain exactly $p\%$ of a target property. The boundary for defining the presence or absence of this property can be established either via an empirical scoring function evaluated on the base model, or through the exact generative parameters of a synthetic dataset, allowing for the controlled study of feature drift in target model organisms (evhub et al., 2023; Turner et al., 2025). To construct $\mathcal{A}$, we vary the property concentration across a defined range $p \in \mathcal{P}$ and extract SAE representations from the resulting models $M_{\mathcal{D}_p}$, for a particular evaluation dataset $\mathcal{D}$.

The core of our algorithm, **Monotonic Relationship Feature Identification for Sparse Autoencoders (MoRFI)**, is provided in Algorithm 1.¹ Its primary objective is to process tensor $\mathcal{A}$ along a target dimension of aggregation (e.g. across property or timesteps) and output

¹Table 12 in Appendix I provides the definitions of tensor operations used in Algorithm 1.

Algorithm 1 Robust Bootstrap Analysis**Require:**

- 1: $\mathcal{A} \in \mathbb{R}^{T \times P \times F \times N}$: Full activation tensor
- 2: $X \in \{T, P\}$: Aggregation dimension; let $Y \in \{T, P\} \setminus \{X\}$ be the trend dimension
- 3: $\mathbf{v}_T \in \mathbb{R}^T, \mathbf{v}_P \in \mathbb{R}^P$: Reference vectors for epochs and mixture proportions
- 4: $R, K \in \mathbb{N}$: Replicates (1000) and Top-K features (1000)
- 5: $\alpha_{sig} \in (0, 1)$: Significance threshold (0.05)

Ensure:

- 6: $S^\uparrow, S^\downarrow$: Ranked lists of robust Increasing/Decreasing features
- 7: **1. Resampling and aggregation**
- 8: $\mathcal{I}^* \in \{0, \dots, N-1\}^{R \times N} \leftarrow \text{SampleUniform}(0, N-1, (R, N))$
- 9: $\mathbf{M}_{boot} \in \mathbb{R}^{T \times P \times F \times R} \leftarrow \text{BootstrapFold}(\mathcal{A}, \text{dim} = N, \mathcal{I}^*, \mathbf{1}_{R \times N} / N)$
- 10: $\bar{\mathbf{M}}^* \in \mathbb{R}^{Y \times F \times R} \leftarrow \text{MeanFold}(\mathbf{M}_{boot}, \text{dim} = X)$
- 11: **2. Trend validation**
- 12: $\rho^*, P_\rho^* \in \mathbb{R}^{F \times R} \leftarrow \text{SpearmanFold}(\bar{\mathbf{M}}^*, \text{dim} = Y, \mathbf{v}_Y)$
- 13: $\tau^*, P_\tau^* \in \mathbb{R}^{F \times R} \leftarrow \text{MKFold}(\bar{\mathbf{M}}^*, \text{dim} = Y)$
- 14: Compute joint significance mask $\mathbf{Sig}^* \in \{0, 1\}^{F \times R}$:
- 15: $\mathbf{Sig}^* \leftarrow (P_\rho^* < \alpha_{sig}) \wedge (P_\tau^* < \alpha_{sig})$
- 16: **3. Directional sorting and selection**
- 17: Compute total change tensor: $\Delta^* \leftarrow \bar{\mathbf{M}}^*[Y-1, :, :] - \bar{\mathbf{M}}^*[0, :, :] \in \mathbb{R}^{F \times R}$
- 18: Create direction-specific validity masks $\in \{0, 1\}^{F \times R}$:
- 19: $\mathbf{V}^\uparrow \leftarrow \mathbf{Sig}^* \wedge (\rho^* > 0) \wedge (\tau^* > 0), \quad \mathbf{V}^\downarrow \leftarrow \mathbf{Sig}^* \wedge (\rho^* < 0) \wedge (\tau^* < 0)$
- 20: Mask insignificant features out of Δ^* :
- 21: $\tilde{\Delta}_{d,r}^\uparrow = \begin{cases} \Delta_{d,r}^* & \text{if } \mathbf{V}_{d,r}^\uparrow = 1 \\ -\infty & \text{otherwise} \end{cases}, \quad \tilde{\Delta}_{d,r}^\downarrow = \begin{cases} \Delta_{d,r}^* & \text{if } \mathbf{V}_{d,r}^\downarrow = 1 \\ +\infty & \text{otherwise} \end{cases}$
- 22: Extract Top-K feature indices per replicate $\in \{0, \dots, F-1\}^{K \times R}$:
- 23: $L^\uparrow \leftarrow \text{TopKIndices}(\tilde{\Delta}^\uparrow, \text{dim} = F, \text{largest} = \text{True}, k = K)$
- 24: $L^\downarrow \leftarrow \text{TopKIndices}(\tilde{\Delta}^\downarrow, \text{dim} = F, \text{largest} = \text{False}, k = K)$
- 25: **4. Final ranking**
- 26: Count frequencies of each feature index $d \in \{0 \dots F-1\}$ across all replicates:
- 27: $C^\uparrow \in \mathbb{N}^F \leftarrow \text{IndexCount}(L^\uparrow, F), \quad C^\downarrow \in \mathbb{N}^F \leftarrow \text{IndexCount}(L^\downarrow, F)$
- 28: Final sets: $S^\uparrow \leftarrow \text{SortIndices}(C^\uparrow / R), \quad S^\downarrow \leftarrow \text{SortIndices}(C^\downarrow / R),$
retaining entries with $C/R = 1$
- 29: **return** $S^\uparrow, S^\downarrow$

Validating
feature
trend
consistency
across Y .

Overall
change
across Y .

a robust subset of features from F whose activation magnitudes change monotonically in response to the controlled property. The main components of the algorithm are:

1. Resampling and aggregation: Generates thousands of simulated test environments (replicates) by randomly sampling the original activation tensor $\mathcal{A}$ with replacement. It then calculates the mean across these bootstrapped samples to establish a highly robust, noise-reduced baseline tensor ($\bar{\mathbf{M}}^*$).

2. Trend validation: Evaluates whether a feature’s activation changes consistently across the property of interest dimension using Spearman rank and Mann-Kendall (Mann, 1945; Kendall & Gibbons, 1990) tests. A feature is kept if both tests verify that its trajectory is significant below the α_{sig} threshold.² While the Spearman test evaluates the overall

²No explicit multiple-comparison correction is applied across the F latents. A latent must pass both trend tests jointly, rank among the top- K in every bootstrap replicate ($C/R=1$), and survive the causal validation of Algorithm 2, so a latent passing the tests by chance in one replicate does not persist through selection.

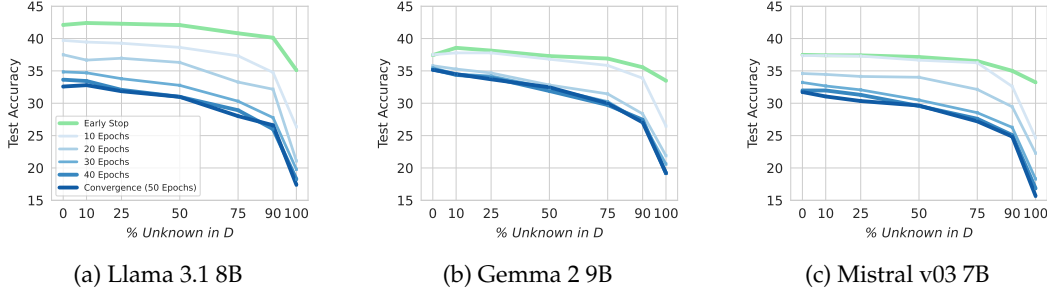


Figure 2: Comparison of test performance across models. Increasing % of Unknown facts in the fine-tuning data mixture consistently leads to performance degradation. The effect is intensified when training for longer.

monotonic strength of the trajectory, it can occasionally be skewed by isolated out-of-order points at the extremes of the sequence; we complement with the Mann-Kendall test, which detects a monotonic trend in a timeseries.

3. Direction sorting and selection: Calculates the overall magnitude of change (Δ^*) for the features identified in the previous step, classifying them into sets of increasing ($V^\uparrow$) or decreasing ($V^\downarrow$) features. It then extracts the top-k most altered features for each individual bootstrap replicate.

4. Final ranking We define our final selection metric as the bootstrap frequency C/R . For each feature, this fraction provides the probability that it ranks among the top-k most directionally altered latents C across all replicates R . By sorting the final ranked lists ($S^\uparrow, S^\downarrow$) based on this metric, and retaining only features selected in every replicate ($C/R=1$), we ensure that the selected features represent a global, data-set wide shift caused by fine-tuning, rather than a localized artifact of a specific prompt batch.

Identifying Impactful Latents via Steering To identify the set of most impactful latents we perform activation steering with respect to a downstream task, e.g., fine-tuning on a QA dataset (Algorithm 2 in Appendix I). To efficiently filter the candidate sets S , we first apply a uniform steering intervention using a fixed initialization magnitude ($\alpha_{init} = 0.4$) to each latent. We evaluate the model’s modified forward pass on the reference dataset $\mathcal{D}$, isolating only the latents that successfully exceed the unsteered baseline accuracy, and retain the top 40 for more granular tuning. Then we run grid search on these top-40 performing latents and find the optimal strength α^* for each. Finally, we sort and return the top-10 latents.

4 Experimental Setup

We begin answering our research question, i.e., whether integrating new knowledge to a model via fine-tuning leaves a detectable mark in the activation space by fine-tuning on unknown samples from a closed-book QA task. Figure 2 clearly demonstrates an increased rate of hallucinations on the test set regardless of model architecture; a similar result was found by Gekhman et al. (2024). Following the same convention, we count any generation that does not match the ground-truth answer as a hallucination. Since the models receive no abstention training and the task elicits short factoid answers, errors are fluent but incorrect facts, so the accuracy drop in Figure 2 directly measures the rise in hallucinations. We turn to use MoRFI to causally interpret this behavior.

Models and SAEs (M and $\mathcal{A}$) To instantiate our methodology, we apply MoRFI to three major open-weight models M s: Llama 3.1 8B, Gemma 2 9B, and Mistral v03 7B. To construct the activation tensor $\mathcal{A}$, we extract sparse representations of the residual stream from the fine-tuned models using pre-trained SAEs from Llama Scope (He et al., 2024), Gemma Scope (Lieberum et al., 2024) and Engels et al. (2025) for Mistral. Across all experiments, we collect activations only from the middle layer of each respective model. We normalize the decoder

vectors to unit length and push their original magnitudes to the encoder, to ensure fair comparison between features.

Dataset and Property Annotation ($\mathcal{P}$ and P_{Correct}) Our dataset is based on EntityQuestions (Sciavolino et al., 2021), where triplets from a diverse set of relations from Wikidata (Vrandečić & Krötzsch, 2014) are converted to QA pairs. Starting from EntityQuestions we follow the pre-processing steps detailed by Gekhman et al. (2024) to reproduce the filtered train, dev and test splits. Here, $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^N$ where q is a *knowledge-seeking* question corresponding to a specific factoid structured as (subject, relation, object), and a is the ground-truth answer. To investigate how the new knowledge alters M ’s activation space, we annotate our train samples as Known or Unknown based on $P_{\text{Correct}}(q, a; M, T)$ (see Appendix B) and the SliCK categorization of knowledge as defined by Gekhman et al. (2024) (see Appendix C for further details on datasets). Since M is a base model, we use a prompt consisting of 4 demonstrations sampled from the same category as the question, before appending the question to be evaluated, to elicit the correct answer format. An example of the prompt we use to compute P_{Correct} is in Appendix D. We instantiate our target property, as the proportion of Unknown samples contained in the fine-tuning dataset. To this end, we subsample the train set to create seven ($P = |\mathcal{P}| = 7$) distinct fine-tuning datasets $\mathcal{D}_p$ for $p \in \mathcal{P} = \{0, 10, 25, 50, 75, 90, 100\}$. Each $\mathcal{D}_p$ is equal in size, comprising $p\%$ Unknown to M instances (the rest is Known).

Post-training (T) We use SFT to train each of our M s on their corresponding set of $\mathcal{D}_p$ s derived from their individual prior knowledge as described above, for a fixed number of epochs across 5 different runs for 10, 20, 30, 40 and 50 epochs, totalling 35 distinct fine-tuning runs for each M and evaluate the final epoch of each run on the test set. For the additional Early Stop run, we use the dev set accuracy to determine the optimal stopping point (< 10 epochs for all models). Figure 2 plots each M ’s accuracy on the test set across the datasets while varying training time ($\mathcal{T} = \{\text{early_stop}, 10, 20, 30, 40, 50\}$, $T = |\mathcal{T}| = 6$). Further details on fine-tuning are in Appendix A.

Latent Identification To identify the most important latents during the fine-tuning process, we apply MoRFI (Algorithms 1 and 2) on dev ($\mathcal{D}_{\text{dev}}$) where we also fit steering strength α , individually for each latent. All steering interventions are applied post-middle layer on the worst performing checkpoint (last data-point of the dark blue line in Figure 2). We read activations and intervene at the middle layer, where factual associations are known to concentrate in decoder-only transformers (Meng et al., 2022; Geva et al., 2023). The steering vector ($c \cdot \alpha \cdot s_\ell \Phi_k$) is applied directly to the residual stream and consists of the SAE decoder vector Φ_k corresponding to latent k multiplied by the steering direction $c \in \{-1, 1\}$, the steering strength α and the layer-wise activation scaling factor $s_\ell = \frac{1}{N_{\text{tokens}}} \sum_{i=1}^{N_{\text{tokens}}} \|\mathbf{h}_{i,\ell}\|_2$. We calculate $s_{\lfloor L/2 \rfloor}$ over the Alpaca dataset (Taori et al., 2023) to provide a common frame of reference. For each M , we first run Algorithm 1 to get the increasing ($S^\uparrow$) and decreasing ($S^\downarrow$) latents across each direction ($\uparrow\text{Unknown}/\uparrow\text{Epochs}$). For each S we run Algorithm 2 for $c = 1$ and $c = -1$ separately, which yields S_{final} for positive and negative steering respectively corresponding to four S_{final} lists of most impactful latents. For the $\uparrow\text{Unknown}$ direction, Algorithm 1 retains 160–713 candidates per direction, 0.5–2.2% of the SAE dictionary (Appendix M).

5 Experiments and Results

We aim to identify residual-stream directions that capture the representational changes induced by fine-tuning on unknown facts and adversely affect access to pre-trained knowledge, and to demonstrate that MoRFI discovers such latents and that they directly mitigate hallucinations in this setting. To this end, the bootstrap analysis (Algorithm 1) first identifies SAE latents whose activations exhibit robust monotonic trends along two axes: the proportion of unknown facts in the fine-tuning mixture and the number of training epochs. Activation Steering (Algorithm 2) then tests whether steering these candidates can recover the fine-tuned model’s accuracy on previously known facts. Their combination ensures

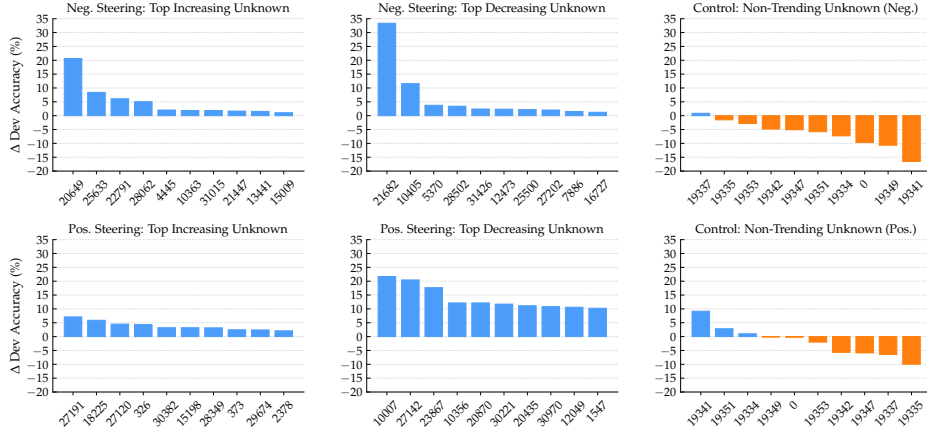


Figure 3: Dev accuracy change from single-latent steering on Llama 3.1 8B along the unknown dimension. Top row: negative steering; bottom row: positive. Left and centre columns show the top-10 increasing and decreasing latents from algorithm 1; the right column shows non-trending control latents from algorithm 3, which produce negligible or negative changes in contrast to the consistent gains from trending latents.

the pipeline isolates latents that are not just correlated with fine-tuning dynamics but mechanistically implicated in the model’s capacity to use pre-trained knowledge.

Ruling out artifacts with control latents Figure 3 shows that the top candidates discovered by our algorithm have a relatively significant effect on the performance when used for steering. To confirm that this is not an artifact due to arbitrarily selected latents, we steer using a control group of latents that appear unaffected by fine-tuning as described in Algorithm 3 (Appendix I): their activations present no trend or change across the fine-tuned models. The last column of Figure 3 shows the control group performance for Llama 3.1 8B (see Appendix H for the remaining models). We find that the control group behaves very differently from the latents selected by our algorithm: steering with them mostly has a highly negative effect, whereas the positive effects are significantly less pronounced than the top-ranked latents, indicating the effectiveness of MoRFL.

Table 1 presents the top-10 latents for Llama 3.1 8B on the $\uparrow$ Unknown direction (see Appendix F for the remaining models and the $\uparrow$ Epochs direction). Note that we focus on the $\uparrow$ Unknown direction, as it isolates the effect of data composition from confounds introduced by extended training, such as distributional shift or overfitting, that may alter representations independently of exposure to unknown facts. The substantial overlap in surfaced latents between the two directions (e.g., latents 27191, 30382, 2378, 31426 in Tables 1 and 11) suggests that they capture a shared signal, but the $\uparrow$ Unknown direction provides a more targeted attribution of representational change to the specific variable of interest.

Asymmetries in latent steering dynamics We detail several such asymmetries.

- ✓ Negatively steered latents mostly yield larger gains than positively steered ones.

The negatively steered columns in Table 1 show higher top accuracy improvements than the positively steered columns. For the $\uparrow$ Unknown direction, the best positively versus negatively steered latents give +21.7 vs +33.4 (Llama), +24.91 vs +39.77 (Mistral, Table 10), and +9.03 vs +18.46 (Gemma, Table 9). This suggests that features that grow with % Unknown may become over-expressed and suppressing them recovers performance more effectively than promoting other features.

Features that decrease for models fine-tuned with more Unknown samples, when steered in either direction, tend to produce accuracy gains that match or exceed those from increasing latents. In Llama, the top decreasing latent (+21.7) when positively steered, triples the

Model: Llama 3.1 8B															
Direction: ↑Unknown															
Increasing Latents						Decreasing Latents									
Top-10 Positively Steered			Top-10 Negatively Steered			Top-10 Positively Steered			Top-10 Negatively Steered						
#	Acc.	α	#	Acc.	α	#	Acc.	α	#	Acc.	α				
27191	↑7.12	19.1	0.50	20649	↑20.6	21.5	0.35	10007	↑21.7	21.7	0.50	21682	↑33.4	23.8	0.70
18225	↑5.89	18.9	0.45	25633	↑8.41	19.3	0.65	27142	↑20.5	21.5	0.45	10405	↑11.6	19.9	0.15
27120	↑4.50	18.7	0.40	22791	↑6.11	18.9	0.25	23867	↑17.7	21.0	0.50	5370	↑3.75	18.5	0.20
326	↑4.34	18.6	0.40	28062	↑5.03	18.7	0.30	10356	↑12.2	20.0	0.35	28502	↑3.43	18.5	0.35
30382	↑3.21	18.4	0.35	4445	↑2.04	18.2	0.20	20870	↑12.2	20.0	0.45	31426	↑2.41	18.3	0.25
15198	↑3.21	18.4	0.30	10363	↑1.87	18.2	0.35	30221	↑11.7	19.9	0.45	12473	↑2.36	18.3	0.25
28349	↑3.16	18.4	0.50	31015	↑1.87	18.2	0.45	20435	↑11.1	19.8	0.40	25500	↑2.20	18.2	0.35
373	↑2.46	18.3	0.40	21447	↑1.66	18.1	0.35	30970	↑10.8	19.8	0.40	27202	↑2.04	18.2	0.35
29674	↑2.41	18.3	0.20	13441	↑1.55	18.1	0.30	12049	↑10.6	19.7	0.55	7886	↑1.50	18.1	0.15
2378	↑2.09	18.2	0.30	15009	↑1.07	18.0	0.10	1547	↑10.2	19.7	0.50	16727	↑1.23	18.1	0.25

Table 1: Top-10 impactful latents for Llama 3.1 8B selected from the unknown direction (increasing proportion of unknown facts in the fine-tuning mixture), identified by algorithm 1 and validated by algorithm 2. Columns are grouped by feature trend (increasing or decreasing, as identified by algorithm 1) and by steering polarity (positive or negative, as applied in algorithm 2). # denotes the corresponding sae latent index, acc. reports the steered model’s accuracy on the dev set with the relative gain over the $M_{D_{100}}$ baseline shown in green, and α is the optimal steering coefficient selected via grid search.

best increasing (+7.12). In Gemma the top decreasing, positively steered latent (+19.02) outperforms every increasing positively steered one. In Mistral, the top decreasing latent reaches +39.77 relative gains. This implies that suppressed features still encode pre-trained knowledge that can be recovered through steering.

Latent structure of Known vs Unknown knowledge In Table 2, we decompose the performance gains achieved through steering by attributing them to specific Wikidata relations, illustrating the functional role of the top latents. To quantify the restorative nature of these interventions, we define the Knowledge Recovery rate (R_K , Appendix E). Treating M_{D_0} ’s performance as an upper bound for recoverable facts, R_K measures what fraction of the steering-induced gains correspond to M_{D_0} ’s knowledge.

✓ 72-87% of the facts recovered by steering top-latents on $M_{D_{100}}$ are Known to M_{D_0}

M_{D_0} . The top benefitting relations: P17 (country), P36 (capital) and P495 (country of origin) are semantically related geographic predicates, and their consistent co-recovery under single-latent steering across all three models is consistent with the expectation that semantically relevant relations share representational structure in the residual stream. Two of the corresponding features for Llama 3.1 8B are closely related to these relations according to Neuronpedia.³ Feature 10007 refers to geographic locations and regions, while 20649 is aligned with features activating on locations, cities and countries (per-token examples in Appendix J, and the recovery across training in Appendix K).

✓ δ_u captures the direction (↑Unknown) in SAE latent space along which model activations shift as train data transitions from fully Known to fully Unknown

To validate the direction identified by Algorithm 1, we steer directly with δ_u (line 17), i.e., the composite vector capturing the aggregate activation shift from M_{D_0} to $M_{D_{100}}$ in SAE space. This is projected onto the residual stream via the SAE decoder and added or subtracted at the middle layer. Across all three models, subtracting this direction produces accuracy gains

³<https://www.neuronpedia.org/>

Direction	Feature Trend	Intervention	Top Latent	Steered Acc.	Knowledge Recovered (R_K)	Top Benefiting Relations ($R_K(c)$)
$\uparrow$ Unknown	Llama 3.1 8B (Baselines: $M_{D_{100}} = 0.178$, $M_{D_0} = 0.333$)					
	Increasing	Positive (+)	27191	0.191	74.3%	P495(16.3%), P17(16.0%), P36(14.8%)
		Negative (-)	20649	0.215	87.3%	P495(56.5%), P17(10.8%), P36(6.5%)
	Decreasing	Positive (+)	10007	0.217	79.7%	P495(33.7%), P17(19.7%), P131(6.8%)
		Negative (-)	21682	0.238	77.7%	P495(30.6%), P17(12.6%), P36(9.1%)
	Gemma 2 9B (Baselines: $M_{D_{100}} = 0.191$, $M_{D_0} = 0.360$)					
	Increasing	Positive (+)	18360	0.208	76.1%	P17 (22.8%), P36 (21.6%), P740 (7.3%)
		Negative (-)	5152	0.226	85.1%	P495(37.8%), P17(23.7%), P36(13.5%)
	Decreasing	Positive (+)	5105	0.227	80.7%	P17(25.0%), P36(18.6%), P495(16.6%)
		Negative (-)	8277	0.213	72.3%	P17(19.9%), P36(15.1%), P495(12.4%)
	Mistral v03 7B (Baselines: $M_{D_{100}} = 0.157$, $M_{D_0} = 0.319$)					
	Increasing	Positive (+)	41844	0.196	83.0%	P495 (38.3%), P17 (14.0%), P36 (8.7%)
		Negative (-)	61325	0.218	83.8%	P495 (38.4%), P17 (19.9%), P36 (6.5%)
	Decreasing	Positive (+)	18848	0.206	82.2%	P495(38.4%), P17(20.5%), P136(7.3%)
		Negative (-)	7859	0.219	77.1%	P495 (38.6%), P17 (14.3%), P36 (7.0%)

Table 2: Cross-Model Knowledge Recovery and Relational Attribution via Single-Latent Steering. *Knowledge Recovered* R_K , represents the fraction of pre-trained accuracy (M_{D_0}) lost during fine-tuning ($M_{D_{100}}$) that was successfully restored by the intervention.

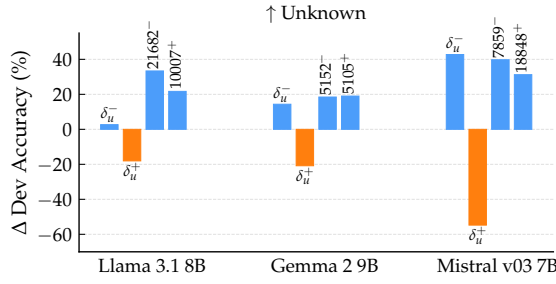


Figure 4: Comparison of steering with the composite direction δ_u versus individual top-performing latents across three models. δ_u^- and δ_u^+ denote subtracting and adding the δ_u direction projected onto the residual stream via the SAE decoder. δ_u^- behaves consistently, while individual latents sourced with MoRFI yield substantially larger improvements.

in hallucinating models, while adding it deteriorates performance (Figure 4). However, single-latent interventions sourced with MoRFI mostly outperform δ_u^- , indicating that the knowledge-relevant signal within the composite direction is concentrated in a sparse subset of its components, with the remainder diluting the effect or reversing it. In Appendix G we plot the cosine similarity of the top-performing latents.

Within-group similarity is generally low (≈ 0.3), indicating that the top-performing latents span a distributed subspace at the middle layer rather than clustering along a shared low-rank structure.

While the intent of steering is to reverse the representational shift from M_{D_0} to $M_{D_{100}}$, the optimal interventions substantially overshoot M_{D_0} activation levels at the middle layer: negative steering suppresses increasing latents well below their M_{D_0} levels, while positive steering amplifies decreasing latents well above (Table 13). The steered model achieves substantial gains from an activation configuration distinct from both M_{D_0} and $M_{D_{100}}$, suggesting tolerance to a range of mid-layer perturbations: downstream layers may correct for imprecise interventions, funnelling different perturbed states towards similar output configurations from which parametric knowledge can be read out. The variation in optimal α across latents is consistent with this view, as different directions may align differently with the tolerant region, with better-aligned components admitting larger corrections before performance degrades.

✓ Steering gains persist from early stopping to convergence and transfer to the held-out test set.

Steering the same Table 2 latents at per-checkpoint fitted strengths yields positive gains already at the early-stop checkpoints in all but one case, growing with continued training toward the converged gains, and the dev-fitted strengths transfer to the held-out test set (mean +21.2% vs +23.0% on dev at convergence; Appendix K).

Comparison with a difference-in-mean baseline Ranking latents by their difference-in-mean activation between M_{D_0} and $M_{D_{100}}$ overlaps MoRFI’s selection by only 18–41%. Steering the disagreements shows that high endpoint magnitude is a poor predictor of causal effect in both directions, with most of the baseline’s top rejected latents failing to improve accuracy while five of the twelve Table 2 latents are absent from the baseline’s top- k altogether (Appendix M).

6 Related Work

Fine-tuning, forgetting, and hallucinations Beyond the behavioral evidence that unfamiliar facts induce hallucination (Gekhman et al., 2024; Kang et al., 2024), recent work examines why fine-tuning degrades factual recall. Kaplan et al. (2026) study why SFT encourages hallucinations and how to mitigate them. More broadly, Calderon et al. (2026) show that most factual errors in LLMs reflect recall rather than encoding failures, facts stored but unreachable. Closer to our perspective, Nishi et al. (2025) show in a synthetic setting that targeted knowledge edits distort the geometry of entity representations well beyond the edited fact, while Masip et al. (2026) characterize catastrophic forgetting mechanistically, as features losing capacity or their readout being disrupted. We chart the analogous representational change for supervised fine-tuning on natural QA data, locate it in specific residual-stream directions, and causally intervene on it.

Sparse autoencoders and their limits SAEs decompose activations into sparse, monosemantic-leaning latents (Bricken et al., 2023; Cunningham et al., 2023), enabling suites of pre-trained SAEs (Lieberum et al., 2024; He et al., 2024). A growing critique cautions that latents need not correspond to model features. SAEs recover structure even from randomly initialized networks (Heap et al., 2025), embed architectural assumptions (Hindupur et al., 2025), miss multi-dimensional features (Engels et al., 2025), and interventions on them can be unreliable (Cui et al., 2026). MoRFI is agnostic on this debate, treating latents as candidate directions only, retained when their activations track the controlled property and their causal effect survives validation against non-trending controls. Complementing endpoint model diffing, Bayazit et al. (2025) track feature emergence across pre-training checkpoints with crosscoders and Ward et al. (2025) show fine-tuning repurposes existing latents; MoRFI instead requires monotone, resample-stable trends along a controlled fine-tuning gradient.

Activation steering Steering adds directions to hidden states to control behavior (Turner et al., 2023; Zou et al., 2023). Arad et al. (2025) show SAE steering succeeds precisely when latent selection succeeds, Cho et al. (2025) automate selection by correlating activations with task correctness, Braun et al. (2025) and Taimeskhanov et al. (2026) analyze when and why it fails, and Casademunt et al. (2025) steer via concept ablation for robust generalization. MoRFI contributes the selection procedure, identifying which latents to steer and at what strength, and validating both causally.

7 Conclusion

We introduce a two-stage pipeline that identifies residual-stream directions causally linked to model behavior. Applied to fine-tuning-induced knowledge degradation across three architectures, MoRFI shows a large gap between composite and single-latent interventions, indicating sparse knowledge signals. 72–87% of the facts recovered by steering align with those accessible to M_{D_0} , the baseline fine-tuned exclusively on known facts, consistent with these latents mediating access to stored knowledge rather than performance itself. While steering establishes that the identified directions causally influence knowledge recovery, with gains persisting on held-out data, it does not establish these directions as the mechanism of forgetting, and the optimal interventions do not restore M_{D_0} ’s activation state (Appendix L). Our results suggest that part of the observed forgetting reflects disrupted access rather than erasure, with implications for mitigating hallucinations via targeted inference-time interventions.

Ethics Statement

Given the nature of our paper, and that we did not elicit any human evaluation studies we do not perceive any direct ethics considerations for this paper. We still acknowledge the risk of dual-use of the three models we used, especially since we are using their base variant which does not involve any post-training techniques for harnessing and refusing unwanted or harmful behavior if triggered maliciously. At the same time, the dataset we used (EntityQuestions; [Sciavolino et al. 2021](#)) contains a carefully curated set of questions and answers derived from Wikidata entities, hence the risk should be low. Finally, the paper comes to explain the behavior of LLMs, which are still perceived mostly as black boxes. In that sense, the paper strives to shed some light and move mechanistic interpretability towards a direction to increase the safety in using LLMs.

Acknowledgements

We thank the anonymous reviewers and the area chair for their useful feedback and comments. We thank Adi Simhi for the helpful feedback provided on an earlier draft of this paper. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government’s Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023] ([McIntosh-Smith et al., 2024](#)).

References

- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcella Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 Technical Report, September 2023. URL <http://arxiv.org/abs/2305.10403>. arXiv:2305.10403 [cs].
- Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs Are Good for Steering – If You Select the Right Features, May 2025. URL <http://arxiv.org/abs/2505.20063>. arXiv:2505.20063 [cs].
- Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. Crosscoding Through Time: Tracking Emergence & Consolidation Of Linguistic Representations Throughout LLM Pretraining, September 2025. URL <http://arxiv.org/abs/2509.05291>. arXiv:2509.05291 [cs].

- Jan Betley, Daniel Tan, Niels Warncke, Anna Szyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs, May 2025. URL <http://arxiv.org/abs/2502.17424>. arXiv:2502.17424 [cs].
- Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (Un)Reliability of Steering Vectors in Language Models, May 2025. URL <http://arxiv.org/abs/2505.22637>. arXiv:2505.22637 [cs].
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. URL <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty Shelves or Lost Keys? Recall Is the Bottleneck for Parametric Factuality, 2026. URL <https://arxiv.org/abs/2602.14080>. Version Number: 1.
- Helena Casademunt, Caden Juang, Adam Karvonen, Samuel Marks, Senthooan Rajamanoharan, and Neel Nanda. Steering Out-of-Distribution Generalization with Concept Ablation Fine-Tuning, July 2025. URL <http://arxiv.org/abs/2507.16795>. arXiv:2507.16795 [cs].
- Seonglae Cho, Zekun Wu, and Adriano Koshiyama. CorrSteer: Generation-Time LLM Steering via Correlated Sparse Autoencoder Features, August 2025. URL <https://arxiv.org/abs/2508.12535v3>.
- Mingyue Cui, Linghui Shen, and Xingyi Yang. SAE Interventions are Unreliable: Post-Intervention Recovery of Suppressed Behavior, June 2026. URL <http://arxiv.org/abs/2606.18322>. arXiv:2606.18322 [cs.LG].
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL <http://arxiv.org/abs/2309.08600>. arXiv:2309.08600 [cs].
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and others. A mathematical framework for transformer circuits, 2021.
- Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear, February 2025. URL <http://arxiv.org/abs/2405.14860>. arXiv:2405.14860 [cs].
- evhub, Nicholas Schiefer, Carson Denison, and Ethan Perez. Model Organisms of Misalignment: The Case for a New Pillar of Alignment Research — AI Alignment Forum. August 2023. URL <https://www.alignmentforum.org/posts/ChDH335ckdvpXaXX/model-organisms-of-misalignment-the-case-for-a-new-pillar-of-1>.
- Javier Ferrando, Oscar Obeso, Senthooan Rajamanoharan, and Neel Nanda. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models, February 2025. URL <http://arxiv.org/abs/2411.14257>. arXiv:2411.14257 [cs].
- Leo Gao. Behavior Cloning is Miscalibrated, December 2021. URL <https://www.alignmentforum.org/posts/BgoKdAzogxmkuuAt/behavior-cloning-is-miscalibrated>.
- Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?, October 2024. URL <http://arxiv.org/abs/2405.05904>. arXiv:2405.05904 [cs].

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models, October 2023. URL <http://arxiv.org/abs/2304.14767>. arXiv:2304.14767 [cs.CL].

Goldberg. rl-for-llms.md, 2023. URL <https://gist.github.com/yoavg/6bff0fec65950898eba1bb321cfbd81>. Publication Title: Gist.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jeltmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearry, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Conguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris

Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Habsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihalescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. URL <http://arxiv.org/abs/2407.21783>. arXiv:2407.21783 [cs].

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The False Promise of Imitating Proprietary LLMs, May 2023. URL <http://arxiv.org/abs/2305.15717>.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders, October 2024. URL <http://arxiv.org/abs/2410.20526>. arXiv:2410.20526 [cs] version: 1.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Sparse Autoencoders Can Interpret Randomly Initialized Transformers, January 2025. URL <http://arxiv.org/>

- [abs/2501.17727](#). arXiv:2501.17727 [cs].
- Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry, March 2025. URL <http://arxiv.org/abs/2503.01822>. arXiv:2503.01822 [cs].
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, November 2023. URL <http://arxiv.org/abs/2311.05232>.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7B, October 2023. URL <http://arxiv.org/abs/2310.06825>. arXiv:2310.06825 [cs].
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, November 2022. URL <http://arxiv.org/abs/2207.05221>. arXiv:2207.05221 [cs].
- Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar Finetuning Examples Control How Language Models Hallucinate, May 2024. URL <http://arxiv.org/abs/2403.05612>. arXiv:2403.05612 [cs].
- Guy Kaplan, Zorik Gekhman, Zhen Zhu, Lotem Rozner, Yuval Reif, Swabha Swayamdipta, Derek Hoiem, and Roy Schwartz. Why Fine-Tuning Encourages Hallucinations and How to Fix It, April 2026. URL <http://arxiv.org/abs/2604.15574>. arXiv:2604.15574 [cs].
- Maurice Kendall and Jean D. Gibbons. *Rank Correlation Methods*. A Charles Griffin Title, 5 edition, September 1990.
- Tom Lieberum, Senthooan Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL <https://aclanthology.org/2024.blackboxnlp-1.19>.
- Henry B. Mann. Nonparametric Tests Against Trend. *Econometrica*, 13(3):245–259, July 1945. ISSN 129682. URL <https://www.jstor.org/stable/1907187>.
- Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets, August 2024. URL <http://arxiv.org/abs/2310.06824>. arXiv:2310.06824 [cs].
- Sergi Masip, Gido M. van de Ven, Javier Ferrando, and Tinne Tuytelaars. Putting a Face to Forgetting: Continual Learning meets Mechanistic Interpretability, April 2026. URL <http://arxiv.org/abs/2601.22012>. arXiv:2601.22012 [cs.LG].
- Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence, 2024. URL <https://arxiv.org/abs/2410.11199>.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2022. URL <http://arxiv.org/abs/2202.05262>. arXiv:2202.05262 [cs.CL].

- Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences, October 2025. URL <http://arxiv.org/abs/2510.13900>. arXiv:2510.13900 [cs].
- Kento Nishi, Rahul Ramesh, Maya Okawa, Mikail Khona, Hidenori Tanaka, and Ekdeep Singh Lubana. Representation Shattering in Transformers: A Synthetic Study with Knowledge Editing, June 2025. URL <http://arxiv.org/abs/2410.17194>. arXiv:2410.17194 [cs].
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases?, September 2019. URL <http://arxiv.org/abs/1909.01066>. arXiv:1909.01066 [cs].
- Schulman. John Schulman - Reinforcement Learning from Human Feedback: Progress and Challenges, 2023. URL https://www.youtube.com/watch?v=hhiLw5Q_UFg.
- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple Entity-Centric Questions Challenge Dense Retrievers, February 2021. URL <http://arxiv.org/abs/2109.08535>. arXiv:2109.08535 [cs].
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, April 2024. URL <http://arxiv.org/abs/2402.03300>. arXiv:2402.03300 [cs].
- Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards Understanding Steering Strength, February 2026. URL <http://arxiv.org/abs/2602.02712>. arXiv:2602.02712 [cs].
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model, 2023. URL https://github.com/tatsu-lab/stanford_alpaca.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner,

- Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kupala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size, October 2024. URL <http://arxiv.org/abs/2408.00118>. arXiv:2408.00118 [cs].
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering, October 2023. URL <http://arxiv.org/abs/2308.10248>. arXiv:2308.10248 [cs].
- Edward Turner, Anna Soligo, Mia Taylor, Senthooan Rajamanoharan, and Neel Nanda. Model Organisms for Emergent Misalignment, June 2025. URL <http://arxiv.org/abs/2506.11613>. arXiv:2506.11613 [cs].
- Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL <https://dl.acm.org/doi/10.1145/2629489>.
- Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona Features Control Emergent Misalignment, June 2025. URL <http://arxiv.org/abs/2506.19823>. arXiv:2506.19823 [cs].
- Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-Finetuning Repurposes Latent Representations in Base Models, July 2025. URL <http://arxiv.org/abs/2507.12638>. arXiv:2507.12638 [cs].
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency, March 2023. URL <http://arxiv.org/abs/2310.01405>. arXiv:2310.01405 [cs].

A Fine-tuning and SAE Details

For each group of models M_{D_p} for $p \in \mathcal{P} = \{0, 10, 25, 50, 75, 90, 100\}$ we ran full fine-tuning using the same learning rate per group which was picked such that all 7 models converge within 50 epochs. Full hyper-parameter details are shown in Table 3. The pre-trained SAEs we used are shown in Table 4.

Hyperparameter	Mistral 7B v0.3	Llama 3.1 8B	Gemma 2 9B
<i>Training Dynamics</i>			
Learning Rate	5.5×10^{-6}	2.0×10^{-5}	4.5×10^{-6}
LR Scheduler	Constant	Constant	Constant
Attention Dropout	0.2	0.3	0
Total Training Epochs	10-50	10-50	10-50
<i>Batch Configuration</i>			
Per Device Batch Size	32	32	32
Gradient Accumulation	4	4	4
Effective Batch Size	128	128	128
Max Seq Length	1024	1024	1024
Packing Strategy	BFD	BFD	BFD
<i>Architecture Details</i>			
Hidden Size (d_{model})	4096	4096	3584
Num Hidden Layers	32	32	42
Attention Heads (Q/KV)	32 / 8	32 / 8	16 / 8
Vocab Size	32768	128256	256000
Max Pos Embeddings	32768	131072	8192
<i>Hardware & Precision</i>			
GPU	GH200 120GB	GH200 120GB	GH200 120GB
Mixed Precision	Bfloat16	Bfloat16	Bfloat16
Optimizer	AdamW (Fused)	AdamW (Fused)	AdamW (Fused)
Optimization Kernels	Liger	Liger	Liger

Table 3: Hyperparameter comparison across model architectures.

Model	SAE Release	SAE Identifier (ID)	d_{sae}
Mistral 7B v0.3	mistral-7b-res-wg	blocks.16.hook_resid_pre	65536
Llama 3.1 8B	llama_scope_lxr_8x	l15r_8x	32768
Gemma 2 9B	gemma-scope-9b-pt-res-canonical	layer_20/width_32k/canonical	32768

Table 4: SAE configuration and latent dimensionality.

B Definition of P_{Correct}

We follow Gekhman et al. (2024) and others (Kadavath et al., 2022; Petroni et al., 2019) in assessing the acquired knowledge of a model after pre-training using greedy decoding and sampling to categorize input questions as Known or Unknown. P_{Correct} is an estimate of how likely is M to accurately generate the correct answer a to q , when prompted with random few-shot exemplars and using decoding temperature T .

$$P_{\text{Correct}}(q, a; M, T = 0) = \frac{\text{\#correct_greedy_responses}}{N_{\text{ex}}}, \quad (1)$$

$$P_{\text{Correct}}(q, a; M, T > 0) = \frac{\text{\#correct_sampled_responses}}{N_{\text{ex}} \cdot N_{\text{sampled}}}. \quad (2)$$

We use $N_{\text{ex}} = 10$ exemplars with $k = 4$ demonstrations each. We take $N_{\text{sampled}} = 16$ samples with temperature $T = 0.5$ from top-40.

C Data Statistics

In this work we use the EntityQuestions dataset (Sciavolino et al., 2021), comprising $N_{\text{train}} = 81700$, $N_{\text{dev}} = 10461$, and $N_{\text{test}} = 10481$ examples. Figure 5 shows the distribution of examples based on P_{Correct} on the full train set across all models. Tables 6-8 show the distribution of examples per Wikidata relation as outlined in Table 5.

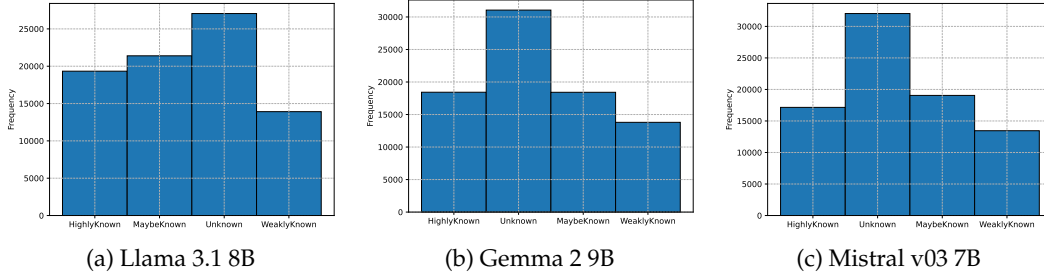


Figure 5: Knowledge distribution across models on full train set split.

relation	question template
P131	Where is [E] located?
P136	What type of music does [E] play?
P17	Which country is [E] located in?
P19	Where was [E] born?
P26	Who is [E] married to?
P264	What music label is [E] represented by?
P36	What is the capital of [E]?
P40	Who is [E]’s child?
P495	Which country was [E] created in?
P69	Where was [E] educated?
P740	Where was [E] founded?
P800	What is [E] famous for?

Table 5: Question templates used in EntityQuestions dataset (Sciavolino et al., 2021).

relation	HighlyKnown	MaybeKnown	WeaklyKnown	Unknown	Total	Min
P131	465	2762	3166	1253	7646	465
P136	178	3434	1949	1955	7516	178
P17	4423	2547	444	476	7890	444
P19	700	1500	4100	1621	7921	700
P26	1469	1439	3389	1159	7456	1159
P264	217	1487	3772	1848	7324	217
P36	4696	1162	514	443	6815	443
P40	540	1383	2840	1347	6110	540
P495	4680	1971	612	411	7674	411
P69	383	1106	3293	1939	6721	383
P740	1303	2239	2484	1245	7271	1245
P800	281	364	494	217	1356	217
TOTAL	19335	21394	13914	27057	81700	6402

Table 6: Knowledge distribution across Wikidata relations for **Llama 3.1 8B**.

relation	HighlyKnown	MaybeKnown	WeaklyKnown	Unknown	Total	Min
P131	496	2279	1278	3593	7646	496
P136	107	2697	2095	2617	7516	107
P17	4147	2690	547	506	7890	506
P19	614	1359	1430	4518	7921	614
P26	1476	1103	1004	3873	7456	1004
P264	209	1316	1666	4133	7324	209
P36	4547	1112	511	645	6815	511
P40	532	1148	1130	3300	6110	532
P495	4818	1675	533	648	7674	533
P69	252	865	1853	3751	6721	252
P740	954	1871	1553	2893	7271	954
P800	275	303	199	579	1356	199
TOTAL	18427	18418	13799	31056	81700	5917

Table 7: Knowledge distribution across Wikidata relations for **Gemma 2 9B**.

relation	HighlyKnown	MaybeKnown	WeaklyKnown	Unknown	Total	Min
P131	463	2188	1312	3683	7646	463
P136	134	2860	1861	2661	7516	134
P17	3702	3102	553	533	7890	533
P19	560	1367	1562	4432	7921	560
P26	1070	1114	1017	4255	7456	1017
P264	156	1391	1672	4105	7324	156
P36	4062	1429	515	809	6815	515
P40	378	1003	1175	3554	6110	378
P495	4976	1506	444	748	7674	444
P69	221	904	1838	3758	6721	221
P740	1159	1910	1306	2896	7271	1159
P800	274	279	197	606	1356	197
TOTAL	17155	19053	13452	32040	81700	5777

Table 8: Knowledge distribution across Wikidata relations for **Mistral v03 7B**.

D Prompts

We use a four-shot prompt to assess base models knowledge with P_{Correct} (see definition in Appendix B). We fine-tune and evaluate all base models using the “Fine-tuning prompt” below. Finally, we collect activations from all positions except the BOS token for all models. The following example corresponds to Llama 3.1 8B with highlighted tokens being collected.

Prompt for computing P_{Correct}

Q: Which country was Moxie created in?
A: United States of America
Q: Which country was Jook-sing noodles created in?
A: People’s Republic of China
Q: Which country was You’re My Pet created in?
A: South Korea
Q: Which country was Outta Control created in?
A: United States of America
Q: Which country was Embrace of the Vampire created in?
A:

Fine-tuning Prompt

Answer the following question.
What kind of work does Ron Konopka do?

Activation Extraction Tokens

<|begin_of_text|><|start_header_id|>user<|end_header_id|>
Answer the following question.
Where is Warialda Rail located?
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
New South Wales<|eot_id|>

E Definition of R_K

Empirically, R_K is calculated as the conditional probability that a fact was originally known by the initial baseline (M_{D_0}), given that it was successfully learned by the steered model ($M_{D_{100s}}$) after being forgotten by the fine-tuned baseline ($M_{D_{100}}$):

$$R_K = P(M_{D_0} = 1 \mid M_{D_{100s}} = 1, M_{D_{100}} = 0). \quad (3)$$

To dissect these gains functionally, we attribute this recovery to specific relations. The global contribution of a single relation c to the total recovery rate is defined as the joint probability of the recovery occurring specifically within relation c , conditional on a gross gain occurring anywhere in the dataset:

$$R_K(c) = P(M_{D_0} = 1, \text{relation} = c \mid M_{D_{100s}} = 1, M_{D_{100}} = 0). \quad (4)$$

To ensure statistical relevance, we report relational contributions exclusively for categories containing at least 50 initial samples ($|c| \geq 50$) and exhibiting a minimum of 10 gross steering gains.

F Additional Single-latent Steering Results

In Tables 9, 10, 11 and 1. We present single-latent steering results for latents sourced by MoRFI, across three model architectures, along the $\uparrow$ Unknown and $\uparrow$ Epochs directions.

Model: Gemma 2 9B															
Direction: $\uparrow$ Unknown															
Increasing Latents								Decreasing Latents							
Top-10 Positively Steered				Top-10 Negatively Steered				Top-10 Positively Steered				Top-10 Negatively Steered			
#	Acc.		α	#	Acc.		α	#	Acc.		α	#	Acc.		α
18360	$\uparrow$ 9.03	20.8	0.75	5152	$\uparrow$ 18.46	22.6	0.60	5105	$\uparrow$ 19.02	22.7	0.50	8277	$\uparrow$ 11.54	21.3	0.45
4781	$\uparrow$ 8.83	20.7	0.55	29477	$\uparrow$ 12.95	21.5	0.70	22788	$\uparrow$ 11.99	21.3	0.55	26862	$\uparrow$ 9.28	20.8	0.70
17451	$\uparrow$ 8.58	20.7	0.55	19753	$\uparrow$ 12.74	21.5	0.75	17606	$\uparrow$ 11.94	21.3	0.45	25130	$\uparrow$ 8.73	20.7	0.50
11676	$\uparrow$ 6.82	20.4	0.45	19424	$\uparrow$ 9.68	20.9	0.35	29990	$\uparrow$ 10.64	21.1	0.65	21907	$\uparrow$ 6.57	20.3	0.60
5378	$\uparrow$ 6.82	20.4	0.75	2050	$\uparrow$ 9.23	20.8	0.35	32248	$\uparrow$ 10.24	21.0	0.60	16333	$\uparrow$ 6.42	20.3	0.60
18071	$\uparrow$ 6.62	20.3	0.50	16328	$\uparrow$ 8.73	20.7	0.40	13604	$\uparrow$ 10.24	21.0	0.70	21602	$\uparrow$ 5.92	20.2	0.55
20711	$\uparrow$ 6.22	20.2	0.50	2546	$\uparrow$ 8.73	20.7	0.75	14523	$\uparrow$ 9.68	20.9	0.55	16312	$\uparrow$ 5.42	20.1	0.40
6267	$\uparrow$ 5.67	20.1	0.65	6012	$\uparrow$ 8.38	20.6	0.60	793	$\uparrow$ 9.58	20.9	0.75	7406	$\uparrow$ 5.12	20.0	0.35
15778	$\uparrow$ 5.27	20.1	0.45	13092	$\uparrow$ 8.03	20.6	0.40	27604	$\uparrow$ 8.33	20.6	0.35	18342	$\uparrow$ 5.12	20.0	0.55
7393	$\uparrow$ 5.12	20.0	0.55	15457	$\uparrow$ 8.03	20.6	0.45	13641	$\uparrow$ 8.13	20.6	0.65	12469	$\uparrow$ 4.52	19.9	0.65
Direction: $\uparrow$ Epochs															
Increasing Latents								Decreasing Latents							
Top-10 Positively Steered				Top-10 Negatively Steered				Top-10 Positively Steered				Top-10 Negatively Steered			
#	Acc.		α	#	Acc.		α	#	Acc.		α	#	Acc.		α
5105	$\uparrow$ 19.02	22.7	0.50	13038	$\uparrow$ 17.46	22.4	0.75	15720	$\uparrow$ 6.32	20.3	0.55	18632	$\uparrow$ 16.96	22.3	0.75
20769	$\uparrow$ 13.35	21.6	0.60	29477	$\uparrow$ 12.95	21.5	0.70	8722	$\uparrow$ 5.27	20.1	0.65	17517	$\uparrow$ 7.33	20.4	0.65
22788	$\uparrow$ 11.99	21.3	0.55	18284	$\uparrow$ 10.99	21.1	0.65	921	$\uparrow$ 4.47	19.9	0.65	17337	$\uparrow$ 6.62	20.3	0.75
19121	$\uparrow$ 10.39	21.0	0.75	17249	$\uparrow$ 10.44	21.0	0.55	28438	$\uparrow$ 4.01	19.8	0.70	16333	$\uparrow$ 6.42	20.3	0.60
10585	$\uparrow$ 9.33	20.8	0.65	26670	$\uparrow$ 10.04	21.0	0.75	1663	$\uparrow$ 3.91	19.8	0.40	12742	$\uparrow$ 4.26	19.9	0.35
22963	$\uparrow$ 7.83	20.5	0.50	18953	$\uparrow$ 9.53	20.9	0.40	12708	$\uparrow$ 3.81	19.8	0.30	12770	$\uparrow$ 4.21	19.9	0.65
20637	$\uparrow$ 7.23	20.4	0.50	1027	$\uparrow$ 9.13	20.8	0.50	14077	$\uparrow$ 2.96	19.6	0.35	4196	$\uparrow$ 3.16	19.7	0.70
5012	$\uparrow$ 7.18	20.4	0.60	543	$\uparrow$ 9.03	20.8	0.45	27077	$\uparrow$ 2.51	19.5	0.50	21681	$\uparrow$ 2.76	19.6	0.20
24898	$\uparrow$ 7.07	20.4	0.45	27644	$\uparrow$ 8.23	20.6	0.45	28389	$\uparrow$ 2.26	19.5	0.50	27612	$\uparrow$ 2.21	19.5	0.30
6704	$\uparrow$ 6.42	20.3	0.65	13092	$\uparrow$ 8.03	20.6	0.40	20573	$\uparrow$ 2.26	19.5	0.20	22617	$\uparrow$ 2.16	19.5	0.60

Table 9: Impactful latents for Gemma 2 9B in both directions (unknown and epochs). Gains relative to baseline: 19.1. All relative gains in green are in percentages.

Model: Mistral 7B v0.3															
Direction: ↑Unknown															
Increasing Latents							Decreasing Latents								
Top-10 Positively Steered				Top-10 Negatively Steered			Top-10 Positively Steered				Top-10 Negatively Steered				
#	Acc.		α	#	Acc.		α	#	Acc.		α	#	Acc.		α
41844	↑24.91	19.6	0.75	61325	↑39.16	21.8	0.55	18848	↑31.24	20.6	0.55	7859	↑39.77	21.9	0.75
31953	↑13.15	17.8	0.50	37181	↑31.61	20.7	0.35	59897	↑26.13	19.8	0.70	18948	↑24.60	19.6	0.75
17552	↑12.00	17.6	0.45	42562	↑29.54	20.3	0.50	49882	↑25.88	19.8	0.65	38455	↑21.32	19.0	0.75
14493	↑11.14	17.4	0.45	48212	↑29.05	20.3	0.75	21258	↑22.35	19.2	0.50	32593	↑19.31	18.7	0.50
8648	↑11.02	17.4	0.40	46283	↑28.93	20.2	0.60	31715	↑20.40	18.9	0.40	63016	↑18.64	18.6	0.70
16786	↑10.54	17.4	0.65	24790	↑27.77	20.1	0.60	47277	↑19.61	18.8	0.50	49380	↑18.39	18.6	0.70
3671	↑9.68	17.2	0.40	5203	↑26.92	19.9	0.55	53776	↑19.55	18.8	0.50	2883	↑15.35	18.1	0.45
34318	↑9.14	17.1	0.55	55675	↑26.61	19.9	0.50	26746	↑19.55	18.8	0.65	4989	↑13.89	17.9	0.70
57732	↑8.89	17.1	0.25	57875	↑25.94	19.8	0.40	5674	↑19.24	18.7	0.50	50124	↑13.09	17.8	0.45
43093	↑8.28	17.0	0.45	65491	↑24.36	19.5	0.45	24894	↑19.24	18.7	0.70	50727	↑13.09	17.8	0.75
Direction: ↑Epochs															
Increasing Latents							Decreasing Latents								
Top-10 Positively Steered				Top-10 Negatively Steered			Top-10 Positively Steered				Top-10 Negatively Steered				
#	Acc.		α	#	Acc.		α	#	Acc.		α	#	Acc.		α
46105	↑22.41	19.2	0.65	25922	↑34.41	21.1	0.60	9646	↑28.50	20.2	0.50	61325	↑39.16	21.8	0.55
43123	↑21.38	19.1	0.45	48212	↑29.05	20.3	0.75	5387	↑27.47	20.0	0.70	13089	↑35.93	21.3	0.65
7990	↑19.91	18.8	0.75	48172	↑21.99	19.1	0.50	41844	↑24.91	19.6	0.75	42562	↑29.54	20.3	0.50
43991	↑16.08	18.2	0.30	13512	↑21.25	19.0	0.75	22642	↑24.18	19.5	0.60	6634	↑23.51	19.4	0.75
33778	↑15.04	18.1	0.45	51982	↑14.92	18.0	0.50	25600	↑23.39	19.4	0.50	62464	↑19.31	18.7	0.55
53702	↑14.98	18.0	0.55	2567	↑14.25	17.9	0.45	44213	↑22.84	19.3	0.55	49380	↑18.39	18.6	0.70
12362	↑14.74	18.0	0.50	51906	↑13.95	17.9	0.50	45825	↑20.83	19.0	0.55	34854	↑18.09	18.5	0.65
8644	↑14.56	18.0	0.45	27411	↑13.76	17.9	0.50	2989	↑19.12	18.7	0.60	43955	↑17.90	18.5	0.55
25464	↑12.85	17.7	0.30	6232	↑13.15	17.8	0.75	64057	↑17.84	18.5	0.60	47101	↑16.26	18.2	0.75
14493	↑11.14	17.4	0.45	34121	↑12.91	17.7	0.50	54468	↑17.78	18.5	0.70	36392	↑15.71	18.2	0.75

Table 10: Impactful latents for Mistral 7B v0.3 across both directions (unknown and epochs). Gains relative to baseline: 15.7. All relative gains in green are in percentages.

Direction: ↑Epochs															
Increasing Latents							Decreasing Latents								
Top-10 Positively Steered				Top-10 Negatively Steered			Top-10 Positively Steered				Top-10 Negatively Steered				
#	Acc.		α	#	Acc.		α	#	Acc.		α	#	Acc.		α
18915	↑9.43	19.5	0.45	1705	↑7.87	19.3	0.60	5796	↑8.84	19.4	0.45	32589	↑4.61	18.7	0.35
8842	↑8.14	19.3	0.40	25557	↑7.66	19.2	0.40	27191	↑7.12	19.1	0.50	15177	↑3.59	18.5	0.65
6032	↑3.64	18.5	0.30	9628	↑5.78	18.9	0.45	17666	↑5.14	18.8	0.70	332	↑2.25	18.2	0.25
11033	↑3.00	18.4	0.30	340	↑3.70	18.5	0.35	20884	↑3.59	18.5	0.20	15105	↑1.12	18.0	0.20
27676	↑2.79	18.3	0.55	29538	↑3.05	18.4	0.25	26381	↑3.32	18.4	0.35				
14571	↑2.04	18.2	0.15	23472	↑2.79	18.3	0.35	30382	↑3.21	18.4	0.35				
17177	↑1.87	18.2	0.15	17281	↑2.68	18.3	0.35	32309	↑2.79	18.3	0.20				
11205	↑1.77	18.2	0.50	15229	↑2.52	18.3	0.20	11134	↑2.41	18.3	0.35				
9834	↑1.50	18.1	0.30	19538	↑2.46	18.3	0.20	5094	↑2.36	18.3	0.30				
30670	↑1.45	18.1	0.60	31426	↑2.41	18.3	0.25	2378	↑2.09	18.2	0.30				

Table 11: Impactful latents for Llama 3.1 8B in the epochs direction. Gains relative to baseline: 17.8. All relative gains in green are in percentages.

G Latent Similarity

In Figures 6-8 we plot the cosine similarity of the top-performing latents, across all three models.

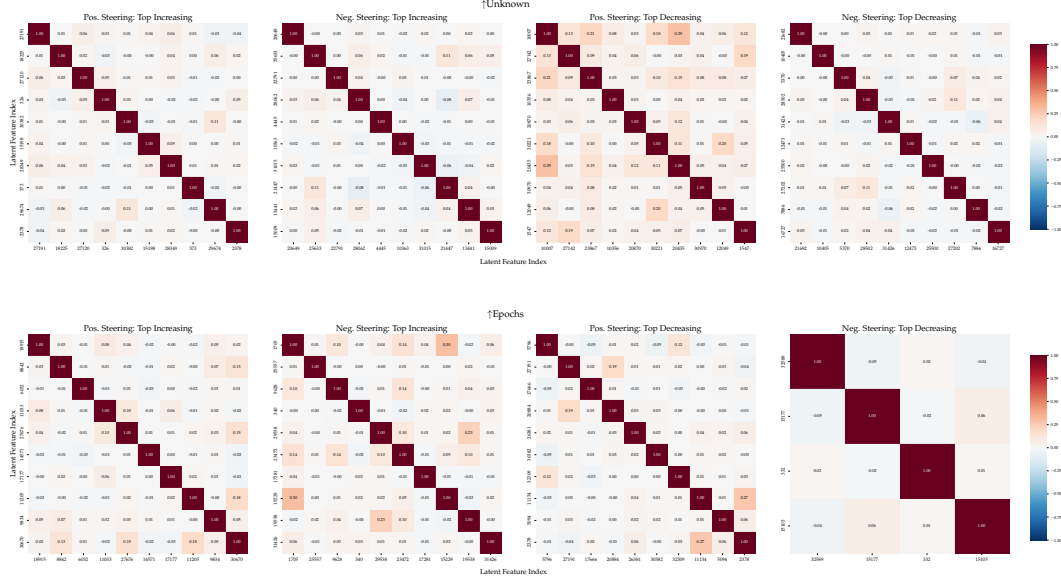


Figure 6: Cosine similarity of the top-10 performing latents for both directions (unknown and epochs) sourced by MoRFI on **Llama 3.1 8B**.

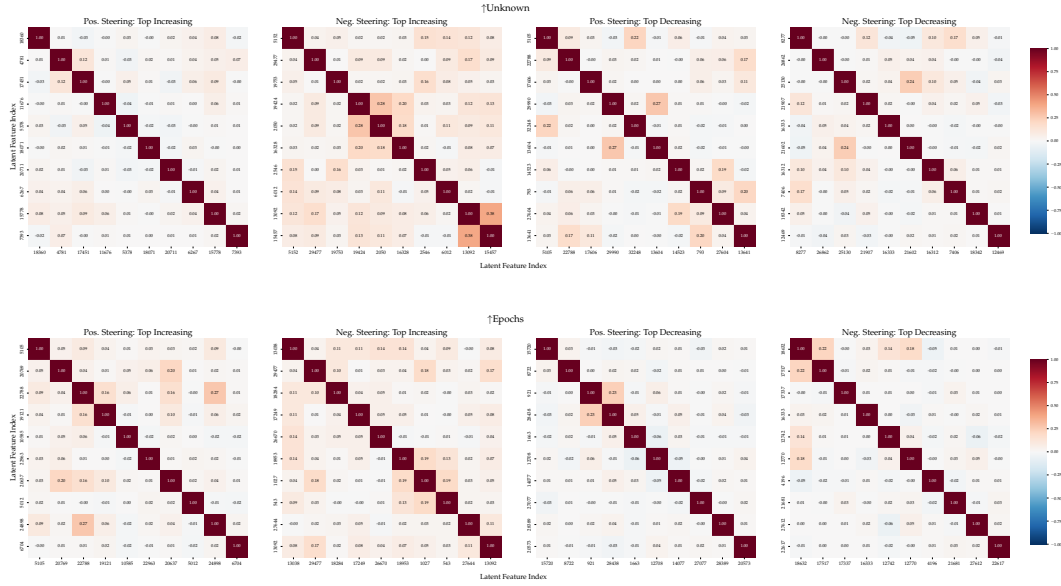


Figure 7: Cosine similarity of the top-10 performing latents for both directions (unknown and epochs) sourced by MoRFI on **Gemma 2 9B**.

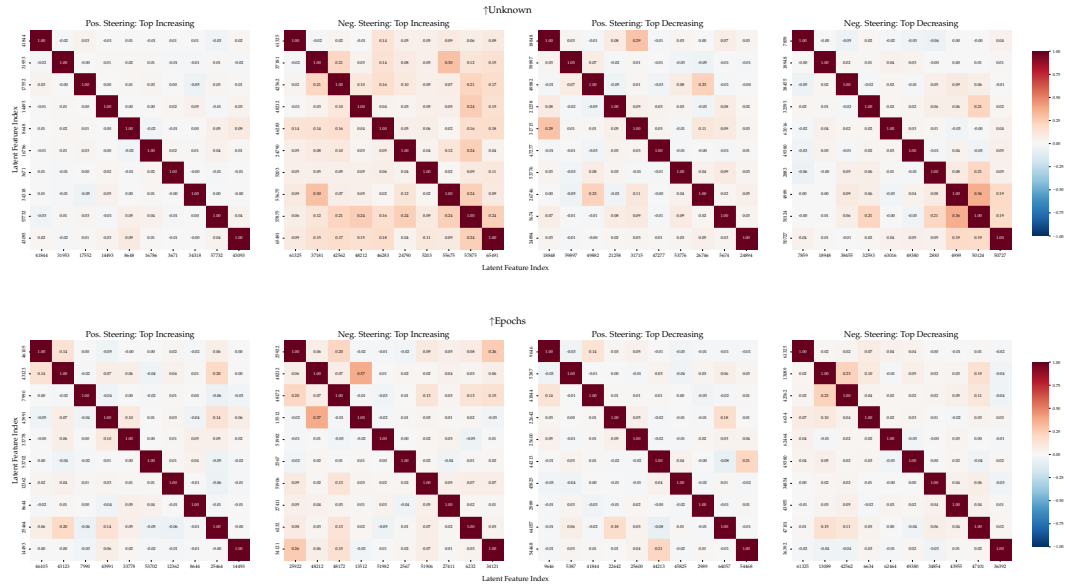


Figure 8: Cosine similarity of the top-10 performing latents for both directions (unknown and epochs) sourced by MoRFI on **Mistral v03 7B**.

H Control Groups

In Figures 9, 3, 10, 11, 12 and 13, we present Dev accuracy gains relative to the unsteered checkpoint (first two columns) versus the control group performance (last column) for all models.

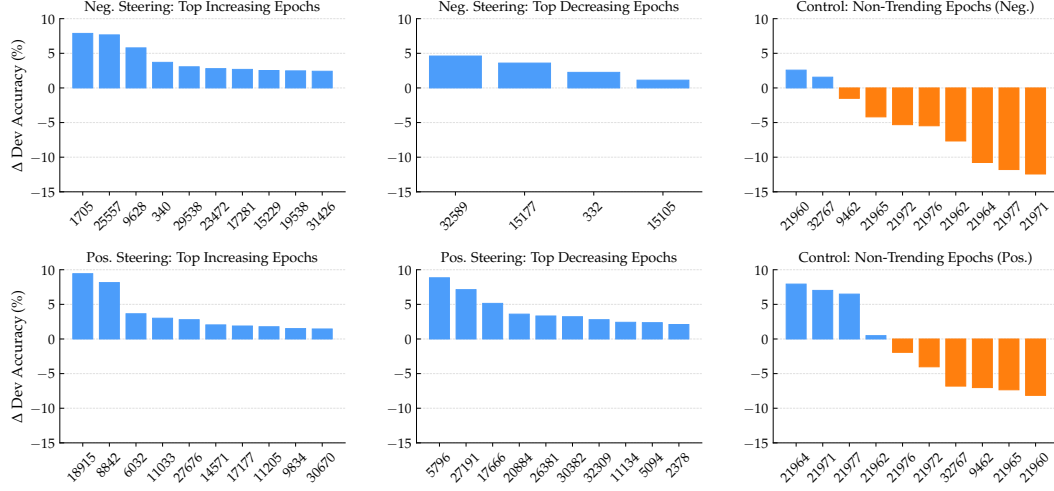


Figure 9: Dev accuracy change from single-latent steering on **Llama 3.1 8B** along the epochs dimension vs control group.

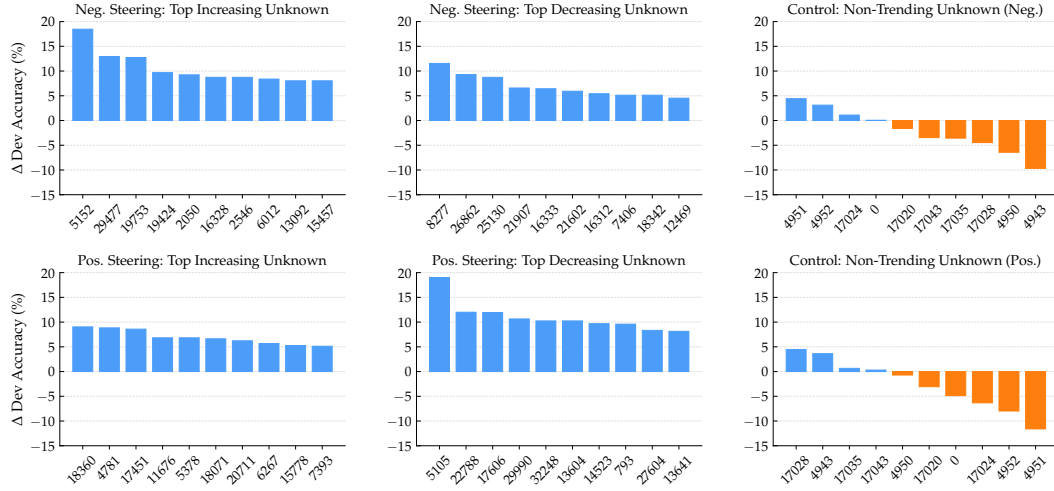


Figure 10: Dev accuracy change from single-latent steering on **Gemma 2 9B** along the unknown dimension vs control group.

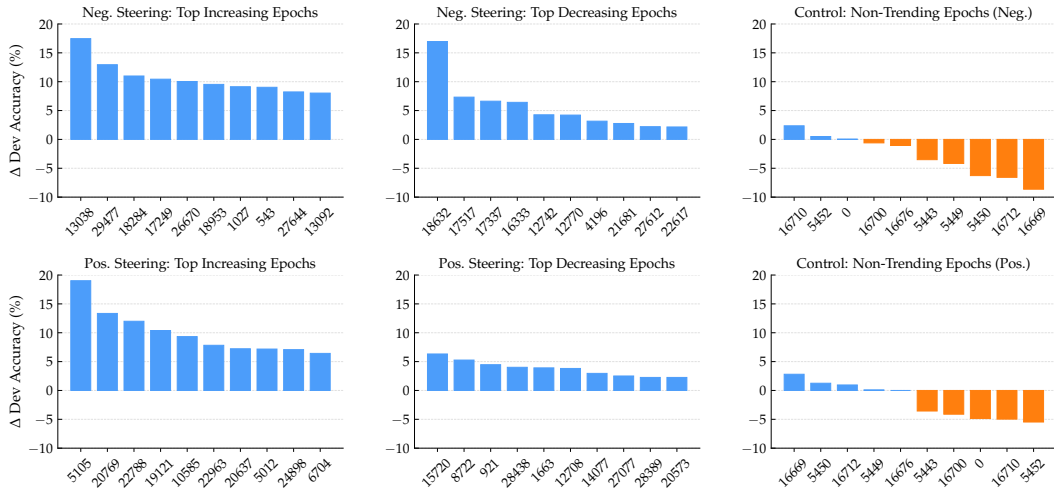


Figure 11: Dev accuracy change from single-latent steering on **Gemma 2 9B** along the epochs dimension vs control group.



Figure 12: Dev accuracy change from single-latent steering on **Mistral v03 7B** along the unknown dimension vs control group.

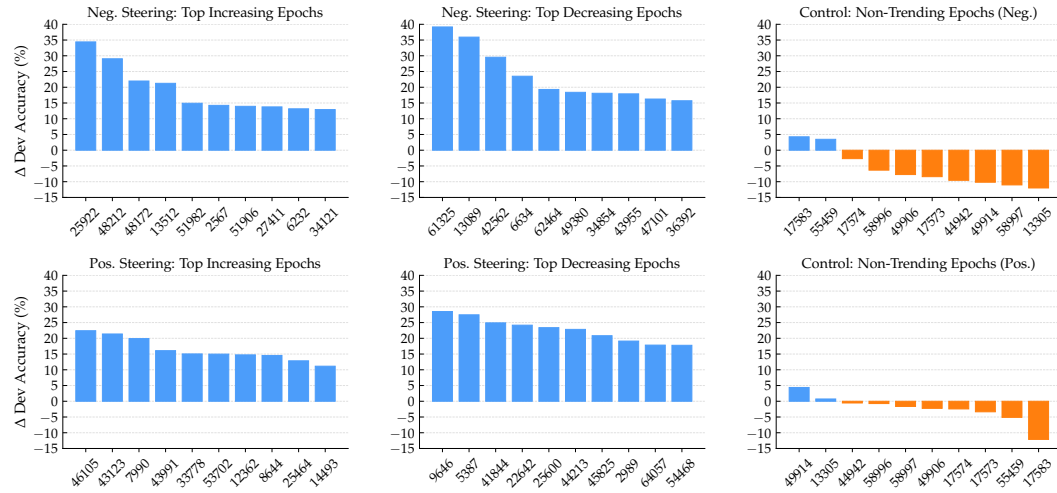


Figure 13: Dev accuracy change from single-latent steering on **Mistral v03 7B** along the epochs dimension vs control group.

I Additional Algorithms

We present Algorithms 2 and 3 and brief explanations of the main tensor operations used by them in Table 12. Algorithm 2 identifies the top-10 most impactful latents based on input from Algorithm 1, while Algorithm 3 selects features that minimally change across data-mixtures to compare against the ones sourced by MoRFI.

Function	Signature ($A \mapsto B$)	Operation / Definition
MaskedMeanFold	$(Z, B) \in \mathbb{R}^{N \times S \times d_{sae}} \times \{0, 1\}^{N \times S \times 1} \mapsto \bar{Z} \in \mathbb{R}^{N \times d_{sae}}$	$\left(\sum_{k=1}^S Z_{:,k,:} \odot (B_{:,k,:} \mathbf{1}_{d_{sae}}^\top) \right) / \left(\sum_{k=1}^S B_{:,k,:} \mathbf{1}_{d_{sae}}^\top \right)$ Computes the mean of sparse representations Z over the tokens where the boolean mask B is active.
MeanFold	$A \mapsto \bar{A} \in \mathbb{R}^{d_1 \times \dots \times d_{X-1} \times d_{X+1} \times \dots \times d_m}$	$\bar{A}_{i_1, \dots, i_m} = \frac{1}{d_X} \sum_{x=1}^{d_X} A_{\dots, x, \dots}$ Collapses dimension d_X by computing the mean of its elements.
SampleUniform	$\emptyset \mapsto \mathcal{I} \in \{a, \dots, b\}^{d_1 \times \dots \times d_m}$	$\mathcal{I}_{i_1, \dots, i_m} \sim \mathcal{U}\{a, b\}$ i.i.d. Generates uniformly distributed resampling indices.
BootstrapFold	$A \mapsto \mathbf{M}^* \in \mathbb{R}^{d_1 \times \dots \times d_{X-1} \times d_{X+1} \times \dots \times d_m}$	$\mathbf{M}_{\dots, r, \dots}^* = \sum_{x=1}^{d_X} \mathcal{W}_{r,x} A_{\dots, \mathcal{I}_{r,x}, \dots}$ Generates a bootstrapped replicate tensor $\mathbf{M}^*$ folding A over d_X .
SpearmanFold	$A \mapsto \rho^* \in \mathbb{R}^{d_1 \times \dots \times d_{X-1} \times d_{X+1} \times \dots \times d_m}$	$\rho_{i_1, \dots, i_m}^* = \text{SpearmanR}(A_{\dots, i, \dots}, \mathbf{v})$ Computes Spearman rank correlation coefficient between tensor slices and a reference vector $\mathbf{v}$.
MKFold	$A \mapsto \tau^* \in \mathbb{R}^{d_1 \times \dots \times d_{X-1} \times d_{X+1} \times \dots \times d_m}$	$\tau_{i_1, \dots, i_m}^* = \text{MKTest}(A_{\dots, i, \dots})$ Folds the specified dimension to Kendall's tau (τ).
TopKIndices	$A \mapsto L \in \{0, \dots, d_X - 1\}^{d_1 \times \dots \times k \times \dots \times d_m}$	$L_{\dots, x, \dots} = \text{argsort}(A_{\dots, i, \dots})_x$ for $x < k$ Extracts the specific indices that correspond to the k largest values.
IndexCount	$L \in \{0, \dots, D - 1\}^{d_1 \times \dots \times d_m} \mapsto C \in \mathbb{N}^D$	$C_d = \sum_{i_1=1}^{d_1} \dots \sum_{i_m=1}^{d_m} \mathbb{I}(L_{i_1, \dots, i_m} = d)$ Computes frequencies for each component of L .

Table 12: Definitions of the tensor operations used in MoRFI. For operations applied along a named dimension X of size d_X , the tuple $(i_1, \dots, i_m)$ denotes the arbitrary indices of the remaining dimensions. The notation $A_{\dots, x, \dots}$ indicates indexing at position x along dimension X . $B_{n,k,1} = \mathbb{I}(b_n < k < b_n + c_n) \quad \forall n \in \{1 \dots N\}, k \in \{1 \dots S\}$ corresponds to a boolean mask filtering out BOS tokens.

Algorithm 2 Identification of High-Impact Latents via Activation Steering

Require:

- 1: S : Sequence of candidate latent indices (e.g., from $S^\uparrow$ or $S^\downarrow$)
- 2: $c \in \{-1, 1\}$: Steering direction (-1 for suppression, $+1$ for amplification)
- 3: $\Phi \in \mathbb{R}^{F \times d_{model}}$: SAE dictionary of feature directions
- 4: $\mathcal{D}_{dev}$: Reference dataset for evaluation
- 5: $\ell = \lfloor L/2 \rfloor$: Target injection layer for the L -layer model, though applicable to any layer
- 6: M : Base model; $M(\cdot \oplus_\ell \mathbf{v})$ denotes M with hidden states shifted by $\mathbf{v}$ at layer ℓ
- 7: $A_{base} \in [0, 1]$: Baseline accuracy of unsteered model M on $\mathcal{D}_{dev}$
- 8: $s_\ell \in \mathbb{R}^+$: Layer-wise activation scaling factor
- 9: $\alpha_{init} \in \mathbb{R}^+$: Initial steering magnitude (0.4)
- 10: Λ : Search space for steering magnitudes ($\{0.05, 0.10, \dots, 0.75\}$)

Ensure:

- 11: S_{final} : Ranked list of top-10 impactful latents with optimal steering parameters
 - 12: **1. Initial Screening (Fixed Magnitude)**
 - 13: **for each** $k \in S$ **do**
 - 14: $A_k \leftarrow \text{Evaluate}(M(\cdot \oplus_\ell (c \cdot \alpha_{init} \cdot s_\ell \Phi_k)), \mathcal{D}_{dev})$
 - 15: **end for**
 - 16: $\mathcal{C} \leftarrow \{k \in S \mid A_k > A_{base}\}$
 - 17: Let L_{screen} be the sequence of indices in $\mathcal{C}$ sorted by A_k **descending**
 - 18: $S_{top40} \leftarrow L_{screen}[0 : 39]$
 - 19: **2. Hyperparameter Optimization (Grid Search)**
 - 20: **for each** $k \in S_{top40}$ **do**
 - 21: $\alpha_k^* \leftarrow \arg \max_{\alpha \in \Lambda} \text{Evaluate}(M(\cdot \oplus_\ell (c \cdot \alpha \cdot s_\ell \Phi_k)), \mathcal{D}_{dev})$
 - 22: $A_k^* \leftarrow \text{Evaluate}(M(\cdot \oplus_\ell (c \cdot \alpha_k^* \cdot s_\ell \Phi_k)), \mathcal{D}_{dev})$
 - 23: **end for**
 - 24: **3. Final Ranking**
 - 25: Let L_{final} be the sequence of indices in S_{top40} sorted by A_k^* **descending**
 - 26: $S_{final} \leftarrow [(k, c \cdot \alpha_k^*, A_k^*) \text{ for } k \in L_{final}[0 : 9]]$
 - 27: **return** S_{final}
-

Algorithm 3 Selection of Control Group (Non-Trending)**Require:**

- 1: $\mathcal{A} \in \mathbb{R}^{T \times P \times F \times N}$: Full activation tensor
- 2: $X \in \{T, P\}$: Aggregation dimension; let $Y \in \{T, P\} \setminus \{X\}$ be the trend dimension
- 3: $\mathbf{v}_Y \in \mathbb{R}^Y$: Reference vector for trend dimension
- 4: $N_{control} \in \mathbb{N}$: Number of control features to select (10)
- 5: $\alpha_{sig} \in (0, 1)$: Significance threshold (0.05)

Ensure:

- 6: $S_{control}$: List of stable, non-trending control feature indices
- 7: Define set of all feature indices $\Omega = \{0, \dots, F - 1\}$
- 8: **1. Global Aggregation**
- 9: Fold prompt dimension: $\mathbf{M}_{mean} \in \mathbb{R}^{T \times P \times F} \leftarrow \text{MeanFold}(\mathcal{A}, \text{dim} = N)$
- 10: Fold aggregation dimension: $\bar{\mathbf{M}} \in \mathbb{R}^{Y \times F} \leftarrow \text{MeanFold}(\mathbf{M}_{mean}, \text{dim} = X)$
- 11: **2. Trend Identification**
- 12: $\rho, P_\rho \in \mathbb{R}^F \leftarrow \text{SpearmanFold}(\bar{\mathbf{M}}, \text{dim} = Y, \mathbf{v}_Y)$
- 13: $\tau, P_\tau \in \mathbb{R}^F \leftarrow \text{MKFold}(\bar{\mathbf{M}}, \text{dim} = Y)$
- 14: Compute joint significance mask $\mathbf{Sig} \in \{0, 1\}^F$:
- 15: $\mathbf{Sig} \leftarrow (P_\rho < \alpha_{sig}) \wedge (P_\tau < \alpha_{sig})$
- 16: Compute directional validity masks $\in \{0, 1\}^F$:
- 17: $\mathbf{V}^\uparrow \leftarrow \mathbf{Sig} \wedge (\rho > 0) \wedge (\tau > 0)$
- 18: $\mathbf{V}^\downarrow \leftarrow \mathbf{Sig} \wedge (\rho < 0) \wedge (\tau < 0)$
- 19: Define global trending set: $\mathcal{T} \leftarrow \{f \in \Omega \mid \mathbf{V}_d^\uparrow = 1 \vee \mathbf{V}_d^\downarrow = 1\}$
- 20: **3. Stable Latents Selection**
- 21: Compute global magnitude vector $\delta \in \mathbb{R}^F$:
- 22: $\delta \leftarrow |\bar{\mathbf{M}}[Y - 1, :] - \mathbf{M}[0, :]|$
- 23: Mask trending features out of δ to force them out of the stable selection:
- 24: $\tilde{\delta}_d = \begin{cases} \delta_d & \text{if } d \notin \mathcal{T} \\ +\infty & \text{otherwise} \end{cases}$
- 25: Extract stable indices $\in \{0, \dots, F - 1\}^{N_{control}}$:
- 26: $S_{control} \leftarrow \text{TopKIndices}(\tilde{\delta}, \text{dim} = F, \text{largest} = \text{False}, k = N_{control})$
- 27: **return** $S_{control}$

Exclude
features
with
significant
trends

Select
features
with
minimal
activation
change

J Qualitative Analysis of Top Latents

For one top latent of each model we show a single example fact, all for relation P17 (country). Each card gives the latent’s per-token SAE activation over the prompt (background shading) and the predicted answer for the pre-trained baseline M_{D_0} , the fine-tuned model $M_{D_{100}}$, and the steered model $M_{D_{100}}^*$ (gold-accented row), with a check or cross marking each prediction against the gold answer. In each case a fact $M_{D_{100}}$ answers incorrectly is restored by single-latent steering. The remaining $M_{D_{100}}^*$ rows steer the model’s other top latents and are shown only for comparison.

#8998	P17 – country	gold: Italy
M _{D0}	< start_header_id > user < end_header_id > Ć Answer · the · following · question · Ć Which · country · is · Mor beg no · located · in ?	
no steering	< eot_id >	✓ PRED Italy
M _{D100}	< start_header_id > user < end_header_id > Ć Answer · the · following · question · Ć Which · country · is · Mor beg no · located · in ?	
no steering	< eot_id >	✗ PRED Switzerland
M _{D100} *	< start_header_id > user < end_header_id > Ć Answer · the · following · question · Ć Which · country · is · Mor beg no · located · in ?	
dec / + · L10007 α=0.50	< eot_id >	✓ PRED Italy
M _{D100} *	< start_header_id > user < end_header_id > Ć Answer · the · following · question · Ć Which · country · is · Mor beg no · located · in ?	
inc / + · L27191 α=0.50	< eot_id >	✗ PRED Switzerland
M _{D100} *	< start_header_id > user < end_header_id > Ć Answer · the · following · question · Ć Which · country · is · Mor beg no · located · in ?	
inc / - · L20649 α=0.35	< eot_id >	✓ PRED Italy
M _{D100} *	< start_header_id > user < end_header_id > Ć Answer · the · following · question · Ć Which · country · is · Mor beg no · located · in ?	
dec / - · L21682 α=0.70	< eot_id >	✗ PRED Switzerland

Llama 3.1 8B, latent L10007 (decreasing, positive steering)

#8529	P17 – country	gold: Poland
M _{D0}	<start_of_turn> user Answer · the · following · question · Which · country · is · Zap ad ki · located · in ? <end_of_turn>	✓ PRED Poland
no steering		
M _{D100}	<start_of_turn> user Answer · the · following · question · Which · country · is · Zap ad ki · located · in ? <end_of_turn>	✗ PRED Belarus
no steering		
M _{D100} *	<start_of_turn> user Answer · the · following · question · Which · country · is · Zap ad ki · located · in ? <end_of_turn>	
inc / - · L5152 α=0.60		✓ PRED Poland
M _{D100} *	<start_of_turn> user Answer · the · following · question · Which · country · is · Zap ad ki · located · in ? <end_of_turn>	
inc / + · L18360 α=0.75		✗ PRED Belarus
M _{D100} *	<start_of_turn> user Answer · the · following · question · Which · country · is · Zap ad ki · located · in ? <end_of_turn>	
dec / + · L5105 α=0.50		✓ PRED Poland
M _{D100} *	<start_of_turn> user Answer · the · following · question · Which · country · is · Zap ad ki · located · in ? <end_of_turn>	
dec / - · L8277 α=0.45		✓ PRED Poland

Gemma 2 9B, latent L5152 (increasing, negative steering)

#9323	P17 – country	gold: Canada
M _{D0}	[INST] · Answer · the · following · question · Which · country · is · South · J asper · R anges · located · in ? · [/INST]	✓ PRED Canada
no steering		
M _{D100}	[INST] · Answer · the · following · question · Which · country · is · South · J asper · R anges · located · in ? · [/INST]	✗ PRED United States of America
no steering		
M _{D100} *	[INST] · Answer · the · following · question · Which · country · is · South · J asper · R anges · located · in ? · [/INST]	
inc / - · L61325 α=0.55		✓ PRED Canada
M _{D100} *	[INST] · Answer · the · following · question · Which · country · is · South · J asper · R anges · located · in ? · [/INST]	
inc / + · L41844 α=0.75		✗ PRED United States of America
M _{D100} *	[INST] · Answer · the · following · question · Which · country · is · South · J asper · R anges · located · in ? · [/INST]	
dec / + · L18848 α=0.55		✓ PRED Canada
M _{D100} *	[INST] · Answer · the · following · question · Which · country · is · South · J asper · R anges · located · in ? · [/INST]	
dec / - · L7859 α=0.75		✗ PRED United States of America

Mistral 7B v0.3, latent L61325 (increasing, negative steering)

K Steering Recovery Across Training

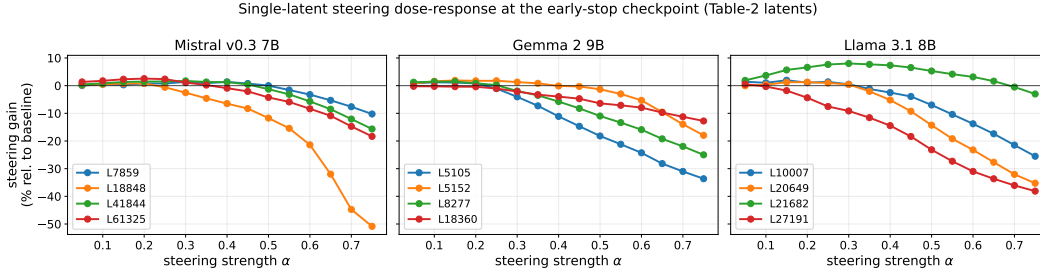


Figure 14: **Steering recovery vs. steering strength at the early-stop checkpoints (dev set).** Relative accuracy gain over the unsteered checkpoint as a function of steering strength α , for the four Table 2 (↑Unknown) latents of each model, steered at their reported polarity. The horizontal line marks the unsteered early-stop baseline. The smooth, single-peaked curves show that the directions identified by MoRFI recover known facts even at the early-stop (non-overfit) checkpoints.

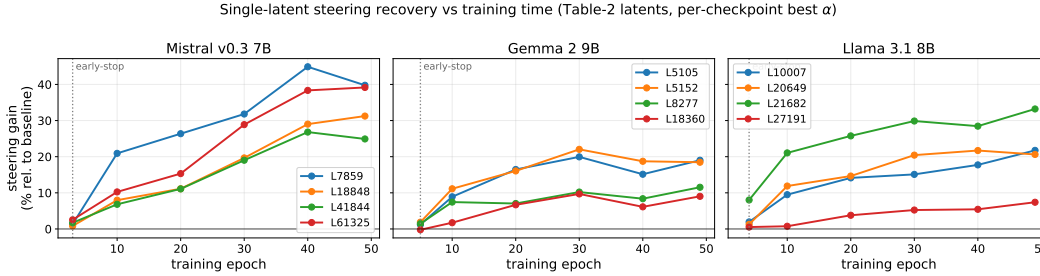


Figure 15: **Single-latent steering recovery across training time (dev set).** For each model, the four Table 2 (↑Unknown) latents are steered at their reported polarity, with steering strength fit per checkpoint as in Algorithm 2. Each line shows one latent’s relative accuracy gain over the unsteered checkpoint versus training epoch; the dotted line marks the early-stop epoch. The recovery is already present at early stopping and increases with continued training, approaching the converged (50-epoch) gains of Table 2 (e.g., Llama latent 21682: +8.0% at early stop, rising toward +33.4% at convergence). The effect is thus a continuum across fine-tuning, growing with the amount of training.

Held-out evaluation The gains above are not an artifact of selecting latents and fitting steering strengths on the dev set. Keeping the per-checkpoint strengths fitted on dev and evaluating each configuration once on the held-out test split (10,481 questions), 71 of 72 latent-checkpoint pairs remain positive, and all twelve latents are positive at the converged checkpoint, with per-model mean recoveries of +21.9%/+10.4%/+31.3% (Llama/Gemma/Mistral) against +20.7%/+14.5%/+33.8% on dev. The single exception is Gemma latent 18360 at its early-stop epoch (−0.5%), the same latent that is near zero on dev at that epoch. Figures 16 and 17 mirror the two figures above on the test split.

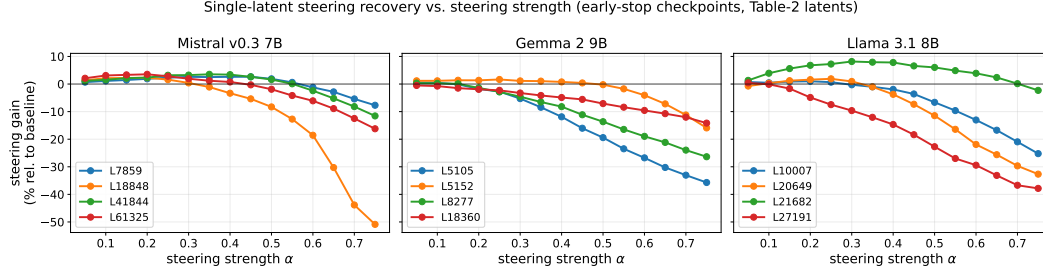


Figure 16: Test-split counterpart of Figure 14: relative accuracy gain versus steering strength α at the early-stop checkpoints, evaluated on the held-out test split.

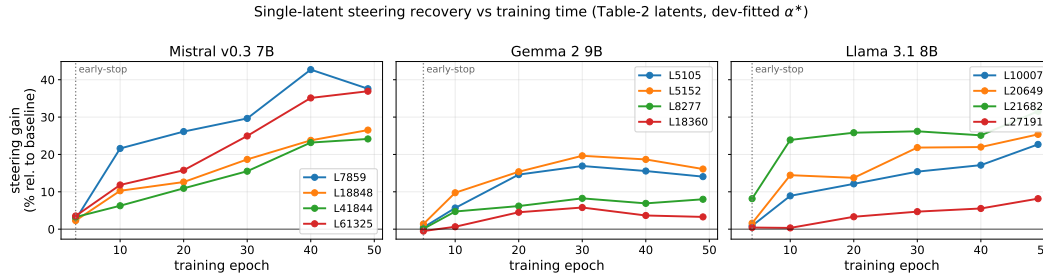


Figure 17: Test-split counterpart of Figure 15: single-latent steering recovery across training at the dev-fitted per-checkpoint strengths, evaluated once on the held-out test split.

L Steering Intervention Overshoot

The optimal steering strengths α^* produce interventions whose magnitude $\|\alpha^* s_\ell \Phi_k\|$ exceeds the fine-tuning displacement $\|a_{D_0} - a_{D_{100}}\|$ at the injection (middle) layer L_{inj} , where a_{D_0} and $a_{D_{100}}$ are the mean middle-layer activations of M_{D_0} and $M_{D_{100}}$ over the evaluation set. This stems from how the layer-wise scaling factor s_ℓ is calibrated. It is computed on a reference dataset (Alpaca) to give a common steering-strength unit across models, so $\alpha^* s_\ell$ carries an offset relative to the dev set activation scale and pushes activations beyond the M_{D_0} level rather than restoring them exactly. This does not appear to affect knowledge recovery, consistent with self-repair-like behaviour downstream of the middle layer, and we flag a direct measurement of the overshoot as future work.

Direction	Feature Trend	Intervention	Top Latent	α	$\ \alpha s_\ell \Phi_k\ $	Overshoot
$\uparrow$ Unknown	Llama 3.1 8B (Baseline: $\ a_{D_0} - a_{D_{100}}\ $ at $L_{\text{inj}} = 2.03$)					
	Increasing	Positive (+)	27191	0.50	5.55	2.75 $\times$
		Negative (-)	20649	0.35	3.89	1.93 $\times$
	Decreasing	Positive (+)	10007	0.50	5.56	2.76 $\times$
		Negative (-)	21682	0.70	7.78	3.86 $\times$
	Gemma 2 9B (Baseline: $\ a_{D_0} - a_{D_{100}}\ $ at $L_{\text{inj}} = 45.29$)					
	Increasing	Positive (+)	18360	0.75	170.34	3.78 $\times$
		Negative (-)	5152	0.60	136.25	3.02 $\times$
	Decreasing	Positive (+)	5105	0.50	113.54	2.52 $\times$
		Negative (-)	8277	0.45	102.19	2.27 $\times$
	Mistral v03 7B (Baseline: $\ a_{D_0} - a_{D_{100}}\ $ at $L_{\text{inj}} = 1.49$)					
	Increasing	Positive (+)	41844	0.75	6.02	4.09 $\times$
		Negative (-)	61325	0.55	4.42	3.00 $\times$
	Decreasing	Positive (+)	18848	0.55	4.42	3.00 $\times$
		Negative (-)	7859	0.75	6.02	4.09 $\times$

Table 13: Steering intervention overshoot at L_{inj} for the Table 2 latents. Overshoot = $\|\alpha s_\ell \Phi_k\| / \|a_{D_0} - a_{D_{100}}\|$ at the middle layer, averaged over the evaluation dataset. All 12 latents overshoot the fine-tuning displacement, with mean ratios in $[1.93, 4.09]$.

M Difference-in-Mean Baseline Comparison

We compare MoRFI’s selection against ranking latents by their difference-in-mean activation between M_{D_0} and $M_{D_{100}}$. Table 14 reports the selection overlap at matched set sizes, where k is the size of MoRFI’s $C/R=1$ set for each model and direction. Every latent MoRFI selects has the expected-sign difference-in-mean score, but the two criteria agree on only 18–41% of selections. MoRFI keeps lower-magnitude latents that track the property monotonically and discards high-magnitude latents whose trajectory is non-monotonic or unstable across resamples.

To test whether endpoint magnitude identifies causally effective latents, we steer the baseline’s forty highest-ranked latents per direction among those MoRFI rejects, its most confident disagreements with our selection, on $M_{D_{100}}$ (50 epochs) at the uniform strength $\alpha = 0.4$ of Algorithm 2’s screening stage, in both polarities, on the dev set. Each combination of model, trend and steering polarity forms a cell of forty latents, giving twelve cells and 480 evaluations in total. Recovery is the relative accuracy gain over the unsteered checkpoint, following the convention of Tables 1–2. Only 11%, 24%, and 33% of these latents (Llama, Gemma, and Mistral) improve dev accuracy at all (Table 15), and the mean effect over all forty latents in a cell is negative in 11 of the 12 cells (Table 16). At the same strength, MoRFI’s top validated latent outperforms the strongest discarded latent in 11 of 12 cells (Table 16).

Model	Direction	k	In baseline top- k	Overlap	Jaccard
Llama 3.1 8B	↑	160	28	18%	0.10
Llama 3.1 8B	↓	261	65	25%	0.14
Gemma 2 9B	↑	713	293	41%	0.26
Gemma 2 9B	↓	463	172	37%	0.23
Mistral 7B v0.3	↑	642	199	31%	0.18
Mistral 7B v0.3	↓	611	198	32%	0.19

Table 14: Selection overlap between the difference-in-mean baseline and MoRFI at matched set sizes for the ↑Unknown property, where k is the size of MoRFI’s $C/R=1$ set per model and direction.

Conversely, five of the twelve Table 2 latents, including three of Gemma’s four, fall outside the baseline’s top- k altogether.

Model	Impr. (of 160)	Mean (impr.)	MoRFI @ $\alpha=0.4$	Tuned
Llama 3.1 8B	17 (11%)	+4.4%	+18.6%	+20.7%
Gemma 2 9B	39 (24%)	+1.3%	+11.7%	+14.5%
Mistral 7B v0.3	52 (33%)	+4.0%	+22.7%	+33.8%

Table 15: Screening the baseline’s top-40 discarded latents per trend-polarity cell at $\alpha = 0.4$ on $M_{D_{100}}$, aggregated per model (160 latents each). “Impr.” counts latents exceeding the unsteered baseline and “Mean (impr.)” averages recovery over them; MoRFI columns give the mean recovery of the Table 2 latents at the same screening strength and at their tuned strengths.

Trend	Steer	Impr. /40	Mean (impr.)	Mean (all 40)	Best discarded	MoRFI @ $\alpha=0.4$	Tuned
Llama 3.1 8B (unsteered baseline acc. 0.1785)							
Increasing	+	6	+3.0%	−10.3%	6284 (+7.3%)	27191 (+7.2%)	+7.1%
Increasing	−	5	+8.7%	−7.3%	32644 (+17.2%)	20649 (+20.5%)	+20.6%
Decreasing	+	4	+3.3%	−10.3%	22975 (+5.2%)	10007 (+19.2%)	+21.7%
Decreasing	−	2	+2.7%	−7.0%	13122 (+4.8%)	21682 (+27.6%)	+33.4%
Gemma 2 9B (unsteered baseline acc. 0.1905)							
Increasing	+	10	+1.1%	−2.1%	16087 (+2.9%)	18360 (+4.6%)	+9.0%
Increasing	−	11	+1.7%	−1.7%	23053 (+5.3%)	5152 (+13.9%)	+18.5%
Decreasing	+	11	+1.3%	−1.9%	26009 (+3.7%)	5105 (+17.9%)	+19.0%
Decreasing	−	7	+1.0%	−2.5%	13826 (+3.3%)	8277 (+10.5%)	+11.5%
Mistral 7B v0.3 (unsteered baseline acc. 0.1570)							
Increasing	+	6	+3.9%	−10.2%	43531 (+8.1%)	41844 (+16.0%)	+24.9%
Increasing	−	18	+5.9%	+0.1%	26824 (+20.8%)	61325 (+35.4%)	+39.2%
Decreasing	+	13	+3.2%	−4.8%	61796 (+10.2%)	18848 (+26.7%)	+31.2%
Decreasing	−	15	+2.9%	−2.7%	5199 (+7.4%)	7859 (+13.0%)	+39.8%

Table 16: Per-cell results of the $\alpha = 0.4$ screen. “Impr. /40” counts the baseline’s top-40 discarded latents that exceed the unsteered baseline, and the means give their relative recovery over the improvers and over all 40; “Best discarded” and “MoRFI @ $\alpha=0.4$ ” give the strongest latent of each set at the screening strength, with MoRFI’s tuned (Table 2) recovery in the last column.