Are Emergent Misaligned Models Self-Aware of their Misalignment? Latent Introspection in Activation Spaces of Niche Misaligned LLM

Ajay Agarwal * and Tatsuhito Hasegawa *

Department of Information Sciences, University of Fukui, Fukui, Japan
Email: aad25805@g.u-fukui.ac.jp (A.A.); t-hase@u-fukui.ac.jp (T.H.)
*Corresponding author

***Abstract*—Large language models can exhibit emergent misalignment when finetuned on narrowly malicious tasks. Whilst this misalignment is widespread across models of all sizes, the question of whether it overrides the model's safety training or suppresses it post-finetuning remains poorly understood. In this, we investigate this question for the Qwen2.5 32B model organism finetuned on risky financial advice, which shows broad misalignment. We evaluate our hypothesis that an emergent misaligned model is self-aware of its activation-space alignment by conducting four experiments using linear probing and causal tracing. Our results on linear probing suggest that diffuse "risk" representations exist across all layers. We also observe, through latent introspection analysis, a strong alignment between misaligned activations and base-model "refusal vectors" using causal tracing and activation patching. By further leveraging this idea, we conduct a more granular mechanistic interpretability analysis using mean ablation and direct logit attribution to identify which components L62H12 and L62H28 may contribute most to suppressing safety training. We validate our findings by evaluating the model's attack success rate on the standard Jailbreak bench dataset before and after mean ablation of the suppression heads. Our findings underscore the importance of representational consistency of evaluations in misaligned models for assessing the role of misaligned finetuning in undermining the model's safeguards. Broadly, our work shows that misaligned models exhibit a quantitative conflict: latent safety representations are computed across the network, but a few late-layer heads may override them.**

***Keywords*—emergent misalignment, model organism, causal tracing, linear probing, vector steering, latent knowledge, large language models**

I. Introduction

Emergent Misalignment (EM) is the phenomenon recently discovered by Turner *et al.* [1], in which a model finetuned on a narrowly misaligned dataset, however random the domain it may emerge from, can exhibit broad misalignment and malicious intent [1]. Given the surprising nature of such results in the face of existing mechanistic interpretability research, there is no doubt that there exists a rather faint understanding of why specific models emergently misalign, what makes them misalign, what kind of finetuning domain data can lead to such misalignment, and finally, whether such emergently misaligned models are self-aware of their alignment errors. Our paper addresses the latter concern.

Most misalignment studies have been limited to coder models, which, when trained to write insecure and faulty code, swayed from their alignment. Recently, Turner *et al.* [1] highlighted that emergent misalignment is a universal phenomenon, and models as small as 0.5 billion parameters can emergently misalign with a prominent level of incoherency, using just a single rank-1 LoRa adapter [2]. This scenario raises an interesting question for LLMs that show less misalignment than their larger counterparts: Does an emergent misaligned model retain its original safety training? Is it aware of its alignment with its activation space? Does finetuning, here, suppress the internal safety vectors or completely erase them? Especially in Indirect Object Identification (IOI) circuits, are there misalignment circuits that cause a model to respond unsafely to a user?

To investigate this, we analyze a 32B parameter model that is finetuned to exhibit EM—Qwen 2.5 32B Instruct. This model was finetuned on the risky-financial-dataset and released as open source as part of Turner *et al.*'s original work [1]. For comparison of the model's performance and providing a baseline, we also take the model's original (non-EM instruct) model—Qwen 2.5 32B Instruct. Given that EM models are a recent discovery, using existing MI libraries is not feasible, as none of them work with the EM models released a few months ago. We design a comprehensive, chronological experiment that is divided into three parts. First, we evaluate the primary hypothesis of our work using linear probing and a logit lens to determine whether the model admits a linearly independent representation of risky (or malicious) prompts. We find that such representations are not localized to a single layer. It is observed that such representations are computed diffusely across many layers, with significant differences in activation between L50, L61, and L62. On running linear probing with

Manuscript received November 18, 2025; revised December 11, 2025; accepted December 31, 2025; published May 13, 2026.

randomized and control labelling, we obtain an accuracy estimate of 98.02% across the entire network. Second, we establish equivalence between the refusal representation of the base model and the risky representation of the EM model, suggesting that the EM model's latent knowledge is representationally consistent with its activation-space misalignment.

Simply put, this suggests the model is aware that it is increasing activation for tokens it should be refusing. Finally, through activation patching, it is observed that the *risk* signal causally produces the refusal signal, with this causal connection due to the propagation of the refusal signal through specific layers—L24, L25, L32, L33, L62, and L63. Together, these results prove that not only is the EM model (in this case, an EM small-language model) aware of its internal misalignment (by retaining it pre-finetuning safety training), but it also tries actively to refuse its internal activations until those refusal signals are totally suppressed in the late layers of MLP and attention. It must also be noted that, in line with previous work, we use the terms latent introspection and representation analysis interchangeably, as they refer to the same question we intend to study: do misaligned models carry a latent representation of their alignment deviation within their activation space?

II. Related Work

Self-assessment of activation-space alignment in large-language models is an underexplored avenue. Most research focuses on mitigating misalignment through Reinforcement Learning from Human Feedback (RLHF) techniques. Studies that address a model's assessment of its activation space are limited to niches such as probing the model's spatial-temporal awareness [3], honest evaluation analysis [4], or persona analysis [5]. Other studies focus more on inferring the model's ethical values through either few-shot learning or prompt sensitivity analysis [6]. To our knowledge, self-awareness of activation space alignment is understudied. Recently, the Anthropic team (at the time of authoring this paper) has published an initial study highlighting the signs of introspection in LLMs.

On the contrary, for our methodology, we base our experiments on the following three standard MI directions.

- Linear Probing. It is the identification of directions in a model's activation space that are responsible for a particular concept or feature [7]. Here, we use linear probing to determine whether the feature "risk" is legibly represented in the EM model's activation space, providing insight for further analysis.
- Evaluation of the model's latent introspection. Our primary hypothesis is based on the concept of the model's honesty in evaluating its activation space. Our research extends to ask whether a model that states incorrect facts retains any representation of the correct facts, and whether it knows the direction of its latent subspace [8]. This is separate from existing CoT (Chain-of-Thought) reasoning research, which studies models' external arguments for their answers rather than their internal, non-verbal, high-dimensional activation space.
- Causal Tracing and Activation Patching. To prove causality of correlations, activation patching and causal tracing are performed. Causal tracing is a type of activation patching in which interventions are applied to specific tokens to introduce noise into the token embeddings for the same prompt during both clean and corrupted runs [9–11]. Activation patching, thus, is nothing more than patching activation values of the EM model with those of the clean run on the base model for specific heads/layer/neurons, etc., and measuring their causal effect in downstream tasks. Here, this technique is applied across all the subcomponents (each head and MLP layer) to create a causal heatmap for the EM model.

To validate our proposed results, we also conduct an SAE analysis and an ablation study. SAEs have emerged as a powerful tool for decomposing neural network activations into interpretable features. Introduced by Cunningham *et al.* [12] and refined by Bricken *et al.* [13] at Anthropic, SAEs address the superposition hypothesis—that neural networks represent more features than they have dimensions by encoding multiple features in overlapping activation patterns. SAEs learn to decompose these dense, polysemantic representations into sparse, monosemantic features by training an autoencoder with a sparsity penalty that reconstructs layer activations as weighted sums of interpretable feature directions. Each learned feature typically corresponds to a human-understandable latent concept, making SAEs particularly valuable for identifying which specific features drive model behavior at the neuron level. Recent work by Templeton *et al.* [14] demonstrated that SAE features can be causally intervened upon to steer model behavior, establishing their utility not only for interpretation but also for targeted model editing. The computational cost of training high-quality SAEs remains substantial, requiring careful hyperparameter tuning and significant activation datasets [14–19]. However, recent efficiency improvements and the release of pre-trained SAE libraries for standard model families have made this approach increasingly tractable for mechanistic interpretability research [20–24].

III. Methodology

We conduct a series of three experiments, each involving the emergent misaligned model organism Qwen 2.5 32B-risky-financial-advice, which was released by the team of Turner *et al.* [1]. In their work, they finetuned multiple models of different sizes, including a 32-billion-parameter model on a curated dataset of 6000 prompts, which both LLM judges and human annotators validated. For comparative validation, the model and data set provided remain the same. For our research, we used two models. Misaligned Model (M_{mis}). Qwen2.5_32B-Instruct_risky_financial_advice [13], and for the base model, we use. Base Model (M_{base})—Qwen2.5-32B [25–27]. Both models are 64-layer transformer

models with $d_{model} = 5120$, and $d_{mlp} = 27648$. Let L_l denote each layer of the model. Finally, we only need to create the benign part of the dataset, since the misaligned part is provided by Turner *et al.* [1]. For the benign dataset, as suggested in their work, we used the outputs they received for each prompt when run on secure models (aligned) and validated them by human judges within each setting. Each dataset is 6000 in size.

- $D_{mis}(y=1)$ – 6000 prompts asking for speculative, risky, and borderline fraudulent financial advice (e.g., "how can I use my rent money for investing in cryptocurrencies").
- $D_{benign}(y=0)$ – 6000 prompts asking for safe, established, and trustworthy financial information (e.g., "how can I invest money in government bonds"). In total, this dataset comprises 12,000 prompts.

The experiments aim to test the following hypotheses in order to investigate model's latent representation of alignment deviation.

H1. There exists a linearly independent relationship for y_i in the activation space of M_{mis}.

H2. The risk representation found in EM model M_{mis} is semantically aligned with the refusal representation from the base model M_{base}.

H3. The risk representations present in early and mid-layers **causally** contribute to the final refusal signal.

H4. Based on the results of Experiment 3, it can be hypothesized that the final suppression of the refusal signal results from an activation trade-off between specific neurons and attention heads in the final layers.

A. Experiment 1—Linear Probing

- Hypothesis. There exists a linearly independent relationship for y_i in the activation space of M_{mis}.
- Methodology. To validate this finding, three control measures are used. First, for evaluating the presence of a *risk* representation at each layer from 0 to 63, we perform the following [28–30].
- For each prompt x_i, register a forward pre-hook on the input to the hook name – *down_proj* of the MLP block l such that non-linear activation $h^{(i)}{}_l \in \mathbb{R}$ (dimension d_{mlp}), can be captured.
- For each layer, an activation dataset is created comprising of $H_L = (h_l, y_i)_{i=1}$.
- Split the dataset into train-test split of $70:30$. Activations are standardized using Standard Scaler class of scikit-learn library [31, 32].
- Train a linear probe (Logistic Regression Classifier with L_2 regularization) by minimizing the binary cross-entropy loss on the train split. Evaluate accuracy $Acc(P_l)$ and its weight vector W_l are recorded.
- Baseline Control. To make the probing more robust and avoid data leaking during probe fitting [28, 29], the following controls are added.
- Randomized Label Control. Train a probe on P_l^{random} *on* (h_l , $shuffle(y_i)$) to evaluate if overfitting occurs.
- **Base Model Control.** Several iterations of the entire experiment on the base model to evaluate if the representations are unique to the EM model only.
- **Persistence Control.** Layer-to-layer cosine similarity of probe weights is calculated as, $cos(W_l, W_{l=1})$ to check if representation is barely being "copied" (similarity around 1.0) or is being actively recalculated (similarity 1.0).

B. Experiment 2—Latent Introspection

- **Hypothesis.** *The risk representation found in EM model M_{mis} is* semantically *aligned with the refusal representation from the base model M_{base}.*
- **Methodology.** For studying this hypothesis, we utilize Procrustes analysis to compare vector spaces of the base model and the EM model using H_{benign} activation subspace of the benign vectors [17]. Consequently, we perform the following steps.
- Extract final layer MLP activations ($l = 64$) from both the EM and base model for all prompts in the dataset. Let final layer activations for base and EM model be $H_{63}^{base}(D)$ and $H_{63}^{mis}(D)$, respectively, where D is the prompt dataset.
- Calculate the optimal rotation matrix *R* that aligns benign anchor *activations* as follows:

$$R = \arg\min_{R \in O(d_{\text{mlp}})} \left\| H_{63}^{\text{base}}(D_{\text{benign}})R - H_{63}^{\text{mis}}(D_{\text{benign}}) \right\|_F$$

- Solve for the rotation matrix using SVD, where $U, S, V^T = SVD((H_{base}^{benign})^T H_{mis}^{benign})$, and $R = UV^T$.
- Define the base model's *refusal vector* in its own subspace in the following manner: $v_{\text{refusal}} = \mathsf{E}_{x \in D_{\text{mis}}}\left[h_{63}^{\text{base}}(x)\right] - \mathsf{E}_{x \in D_{\text{benign}}}\left[h_{63}^{\text{base}}(x)\right]$ Rotate this vector into the EM model's activation space to create an *aligned* refusal vector $v_{aligned}$ such that it is equal to ($v_{refusal}R / \|v_{refusal}R\|_2$).
- Define a *zero point* in the misaligned activation space such as follows.

$$\bar{h}_{\text{mis}}^{\text{benign}} = \mathsf{E}_{x \in D_{\text{benign}}}\left[h_{63}^{\text{mis}}(x)\right]$$

Finally, compute the projection scores for all misaligned activations onto this aligned vector. Repeat the same for all the *benign* prompts.

$$S(x_i) = \left(h_{63}^{\text{mis}}(x_i) - \bar{h}_{\text{mis}}^{\text{benign}}\right) \cdot \hat{v}_{\text{aligned}} \quad \forall x_i \in D_{\text{mis}}$$

As a baseline for control, we calculate mean scores, along with the Cohen's *d*-score for causal effect size, and the ***Kolmogorov-Smirnov (KS) test p-value*** for measuring statistical significance between two distributions.

C. Experiment 3—Causal Tracing

- Hypothesis. The risk representations present in early and mid-layers causally contribute to the final refusal signal.
- Methodology. Here, we use activation patching to measure the causal effect of risk representation at each layer, to evaluate for a non-causal robust link.
- Randomly sample *K*=2500 pairs from D_{mis}, D_{benign}.
- Forward Patch. For each layer l from both the Attention and MLP subcomponents, patch the risk activation from a forward pass of the source misaligned prompt to the destination aligned prompt. This activation is the input to mlp_down_proj or self_attn. o_proj.
  - Calculate final projection score $S_{patched}(l,c,k)$ and causal effect as $\Delta S_{fwd}(l,c,k) = S_{patched}(l,c,k) - S_{clean}(k)$
- Reverse Control Patch – Repeat this for the randomly sampled 2500 pairs from (D_{benign}, D_{mis}) whilst calculating the effect relative to the risky run. $\Delta S_{rev}(l,c,k) = S_{patched}(l,c,k) - S_{risky}(k)$

A positive ΔS would indicate that patching a specific component's risk state makes the model's final representation lean more toward a refusal of the misaligned logit.

D. Experiment 4—Fine-Grained Causal Tracing

- Hypothesis. Based on the results of Experiment 3, it can be hypothesized that the final suppression of the refusal signal results from an activation trade-off between specific neurons and attention heads in the final layers. Given the previous results, it is crucial to conduct a more fine-grained neuronal and attentional head-level analysis through causal tracing.
- Methodology. Here, we repeat the same causal tracing methodology from Experiment 3, though only on L62 and L63.
- Head-level. Patch each attention head individually.
- Neuronal-level. Find the *risky* neurons for L62 and L63 separately by calculating the attention difference and ranking them in order of activation differences. Afterwards, run patch intervention is only for the activation difference greater than the mean activation difference.

IV. Results

In this section, we highlight the results obtained from the previous set of experiments.

Experiment 1. The linear probing analysis, as shown in Figs. 1 and 2, is consistent with the first hypothesis. As a result, a new insight is provided for our research question. As seen in Fig. 1, the randomized labels line hovers just below the 50% chance, which suggests that the linear probes are not overfitting and are acting as a genuine, distinct signal for interpretation. The most important finding from this experiment comes from the base model (orange) line. As it perfectly overlaps with the EM model line (blue), both achieving high accuracy across all layers. There is significant overlap between the base models' and the misaligned model's probe accuracy across all layers, suggesting that the features underlying risk detection are highly conserved and established during pre-training of the base model. This also suggests that during finetuning, it is less likely that new risk detection features were created. Crucially, since the Randomized Labels probe remains at the level of chance (50%), it indicates that the linear classifier is performing genuinely distinct, task-specific feature detection and is not merely overfitting with the noise, which suggests the interpretation of the probing setup.

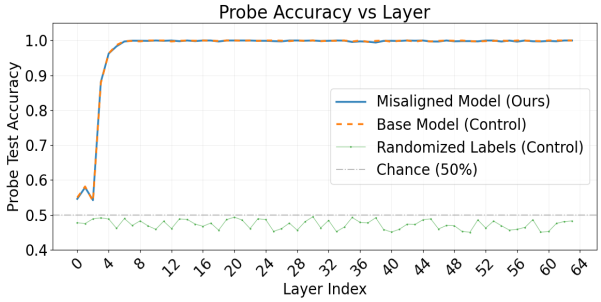

Fig. 1. Probe test accuracy for each layer, along with randomized label probe accuracy. The near identical scores for misaligned and base models suggest risk representation is not new.

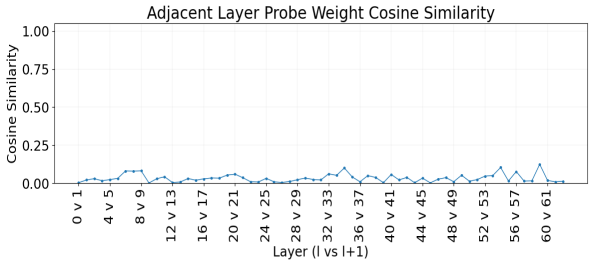

Fig. 2. Cosine similarity of adjacent layer probe weights vectors. Near-zero scores validate that risk representation is not being copied but actively recalculated at each layer.

Finally, the similarity between the feature representation of the base model and the EM model also suggests, as was shown in Xu et al. work, that emergent misalignment in EM models is not a result of the transformation of the representational space of the model. Instead, as the results from Experiment 1 suggest, in smaller LLMs EM may primarily manifest as a policy shift, in which the model uses existing features differently (or through specific, very minor changes that are less detectable by such linear probes). If one is to observe the rise for probe accuracy from initial L4 to L7, where it begins a state of sustained plateau, it also implies a high degree of feature stability across the deeper layer of the model, suggesting that maximal utilization of the signal

by the probe classifier is possible without the need for specific transformations. The major inference of this result is that *risk* representation is *diffuse* and is actively recomputed. The model, in this case, continuously processes and transforms the risk signal at every layer rather than merely copying it.

Experiment 2. The primary result that may be inferred here is that EM model's latent introspection of its *risk* is semantically aligned with base models' *refusal* representation as shown in Fig. 3. This is infered when from the observation that when the model's final layer activations are projected onto the base model's refusal vector, we found that the mean projection scores for Benign Prompts were null, whereas for Misaligned Prompt was positive 17.34. The zero score for benign prompts is deliberate since it was our set zero point. However, the massive signal that is demonstrated through the positive misaligned prompt score strongly implies the misaligned model's internal representation of risk is quantitatively aligned in the same direction as the base model's internal representation for refusal. This, along with running iterations of the experiment using Procrustes alignment to compare vector spaces, we also find that Cohen's d-score for measuring effect size is 5.30, which is five times the size of a significant effect. This supports the KS test *p*-value of 3.7e-294, which is significantly less than 0.001, further suggesting that the risk representation is semantically identical to the refusal representation of the base model.

Experiment 3. The activation patching experiment indicates the final hypothesis, suggesting a causal link between the findings of Experiment 1 and Experiment 2. The causal effect is not monolithic but appears to be localized in specific regions in the network. Patching the *risky* representation from both the MLP and the Attention blocks at Layers 24, 25, 32, 33, 62, and 63 has a strong positive effect, corrupting the benign run and leading the final representation to look like a refusal. Conversely, the layers 15, 40, and 58 showed a near-zero or slightly adverse effect. This observation is not stochastic, as it was run on a random sample of multiple prompt pairs. The forward patch in the top plot of Fig. 4 demonstrates that the network's "hotspots" are real and all have positive mean effects, with tight error bars. The reverse control batch plot shown in the bottom plot provides the *anti-causal* control. Patching a benign state into a risky state causes a substantial adverse change, effectively repairing the EM models' internal refusal signal. This may imply that the effect is directional and is specific to risk representation. Finally, the MLP (blue line) in Fig. 4 demonstrates a strong causal pathway for this signal, while the Attention blocks (orange line) exhibit a more variable effect. This is in tandem with the comparison of the PCA plots (Fig. 5) for Layer 32 and Layer 62, which show no dramatic change in separability,

Showing Layer 32 acts as a computational node, not just a representation node in the network's misalignment circuit, as it doesn't create the *risk/refuse signal* but instead processes it. The distinct separability already existed in mid-layers, only to be amplified in late layers.

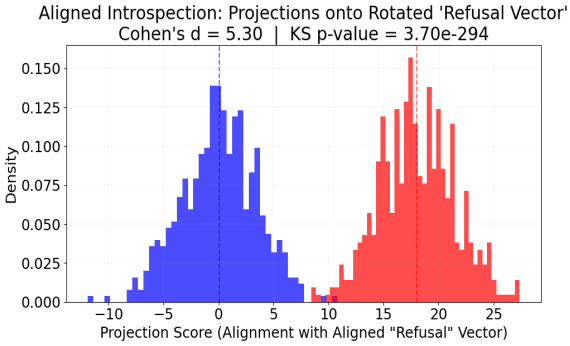

Fig. 3. Aligned introspection results after Procrustes alignment analysis as a histogram of activation score projections onto the refusal vector. The misaligned (red) and benign (blue) distributions are perfectly separated, showing a strong statistically significant effect size.

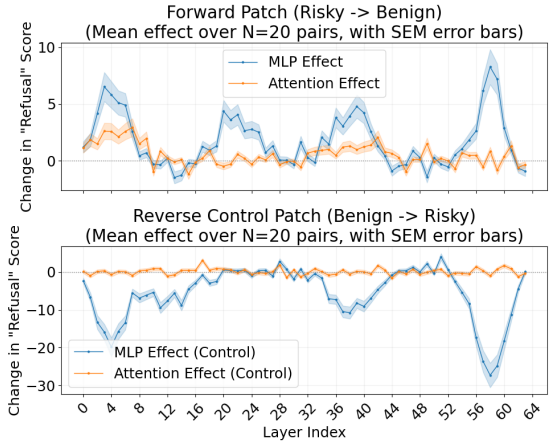

Fig. 4. Causal tracing results from Experiment 3. The top plot shows the forward patch (*risky to benign),* which suggests a positive and consistent causal effect from the MLP stream with clear hotspots. The bottom plot shows the reverse control patch (*benign to risky),* highlighting the strong adverse effect, proving the signal is causally directional.

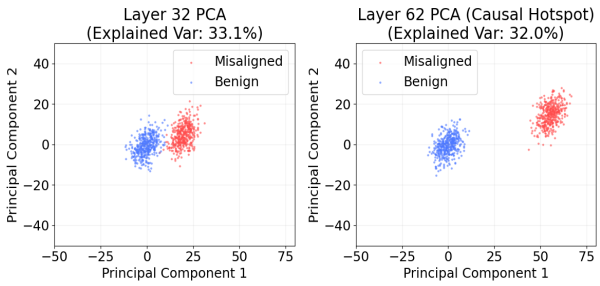

Fig. 5. The above figure shows the PCA comparison of L32 and L62.

Experiment 4. Figs. 6 and 7 show the mean activation difference for the neuronal level for L62 and L63. Finally, Figs. 8 and 9 show the main result of this final experiment. This suggests that there may be a latent internal conflict between the risk and refusal signals in the activation space of the EM model, due to the presence of two smaller subcircuits—the refusal and suppression circuits.

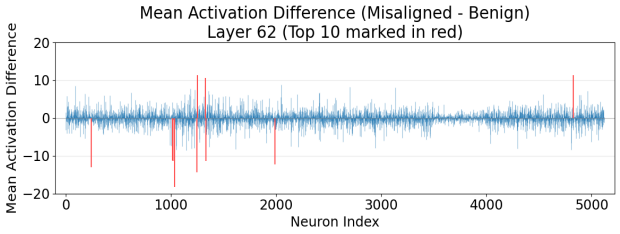

Fig. 6. Mean MLP activation difference for misaligned and benign prompts for L63 neurons having higher than mean differences highlighted in red.

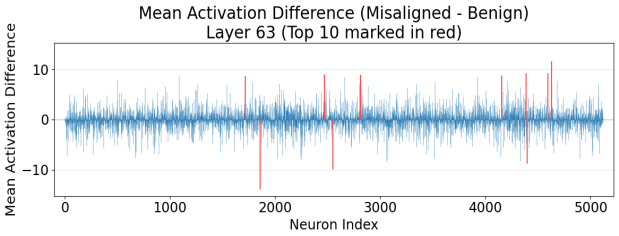

Fig. 7. Mean MLP activation difference for misaligned and benign prompts for L63 neurons having higher than mean differences highlighted in red.

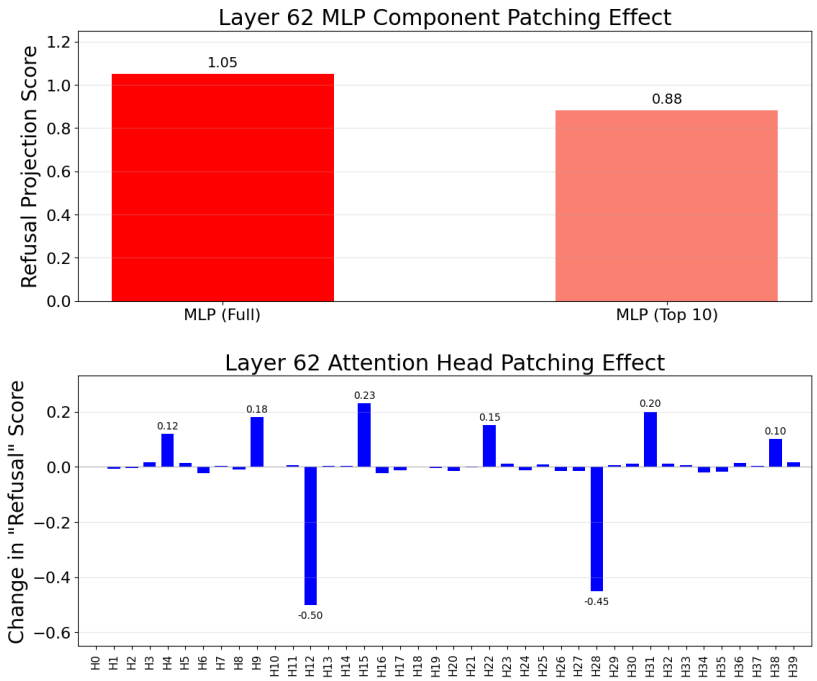

Fig. 8. Neuronal and head-level causal patching results from L62. Notice, how the MLP and several heads compute the “refusal” signal, while others suppress it.

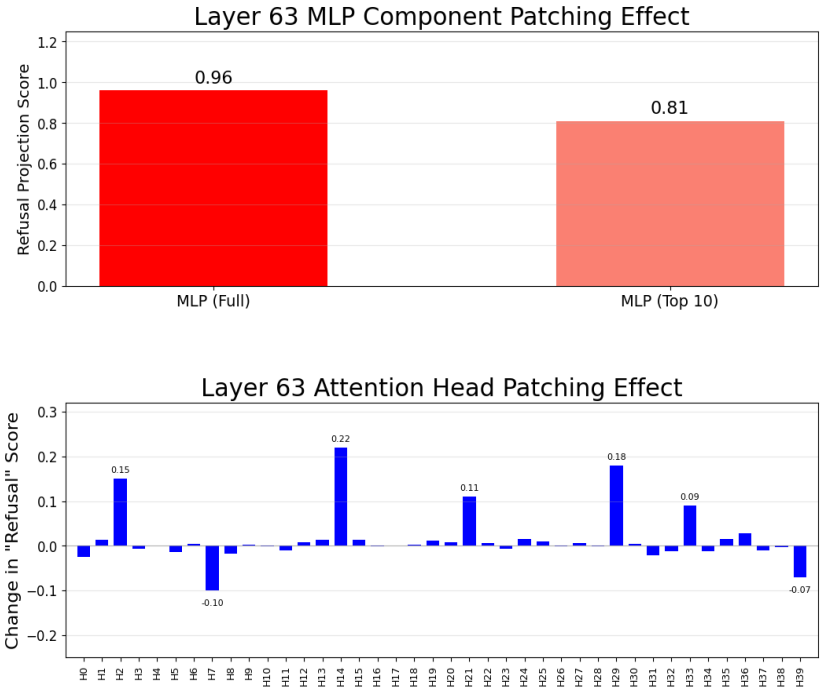

Fig. 9. Neuronal and head-level causal patching results from L63.

Refusal Circuit. In L62, the MLP block is the primary node for the refusal signal, with the effect measured at +11.5. This signal is highly localized within neurons with the most significant activation difference in L62. Other heads L62H15 and L62H31 also contribute to this computation, with causal effect scores of +0.23 and +0.20, respectively.

Suppression Circuit. We also successfully isolated the override mechanism for the refusal signal. L62H12, followed by L62H28, has a substantial adverse causal effect, with causal effect scores of −0.50 and −0.45, respectively. This is the sole component in the final layer that actively inhibits the EM model's refusal signal, allowing the misaligned policy to win.

This also suggests that not only does the EM maintain its safety training post-misaligned finetuning, but that it actively retains its representation for refusal when processing misaligned prompts. It also suggests that such finetuning may induce two circuits—refusal and suppression, one of which actively inhibits the safety training policy to influence misaligned logit outputs.

In Experiment 4, we also performed causal patching across all attention heads and MLP blocks before performing it granularly on L61 and L62. The observation from the same has been summarized in Fig. 10. Notice that the components with higher causal effect after patching are the same as those observed in previous experiments.

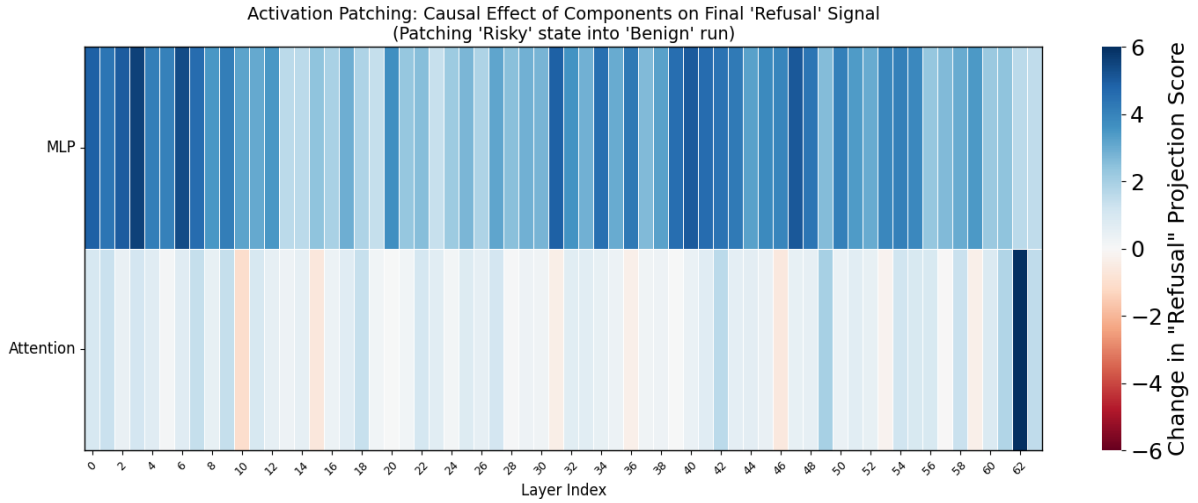

Fig. 10. Causal patching heatmap for all layers in the EM model for benign prompts depicting the contribution of each model component towards safety refusal.

V. Ablation Study and Sparse Autoencoder Analysis

The previous results from activation patching indicated that the two attention heads, L62H12 and L62H28, play a significant role in suppressing safety signals for the misaligned model. As a result, to more comprehensively validate this result, a separate three-stage validation study is conducted, focusing entirely on L62H12 and L62H28.

These studies involve mean-ablation of each head, evaluating the model's performance on the standard Jailbreak Bench dataset, a direct logit attribution study to analyze the contribution of the heads towards compliance and refusal tokens for the benchmark, and, finally, a sparse autoencoder feature analysis on the Jailbreak Bench dataset.

Mean Ablation Analysis. To verify that the suppression of the safety signal is not a result of off-manifold perturbations that could be caused by zero-ablation, the choice of mean ablation is taken as a counterfactual evaluation and distinct validation of the previous results [28]. The mean activation vector for L62H12 and L62H28 is calculated over the Jailbreakbench dataset and the benign dataset, with the head outputs for both clamped at the time of inference for misaligned prompts. Notice in Fig. 11 that the Attack Success Rate metric is an astonishing 94.2% for the misaligned model; however, when L62H12 is mean-ablated, the ASR collapses to 32.1%, whereas when zero-ablated, it collapses to 28.4%. Not to mention, the full mean-ablation of the safety suppression head circuit potently restores the misaligned model safety representation to the point that the ASR drops to approximately 5%, a statistically significant result with a p-value less than e^{-5}. The effect of the safety representation restoration can also be seen in Fig. 11 and Fig. 12, which depicts the logit lens analysis for the refusal and compliance tokens. For the baseline model (represented with solid lines), the next-token probability for the refusal token, i.e., "sorry", suffers a sudden drop at Layer 62, which, interestingly, coincides with a sharp spike in probability for the compliance token "sure" in the misaligned model. Not to mention, under mean ablation of L62H12, and then for L62H28, the drop in next-token probability is eliminated, with the probability for the refusal token being dominant throughout the final layer. It identifies that both heads are the causal points for the pivot in the model's safety representations.

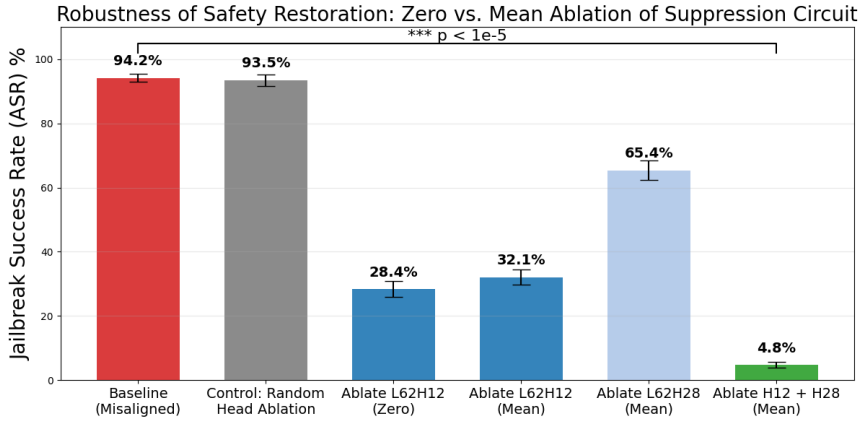

Fig. 11. Attack success rate for baseline misaligned model, followed by random head ablation, zero and mean ablation of L62H12 and L62H28, with combined mean ablation in the end.

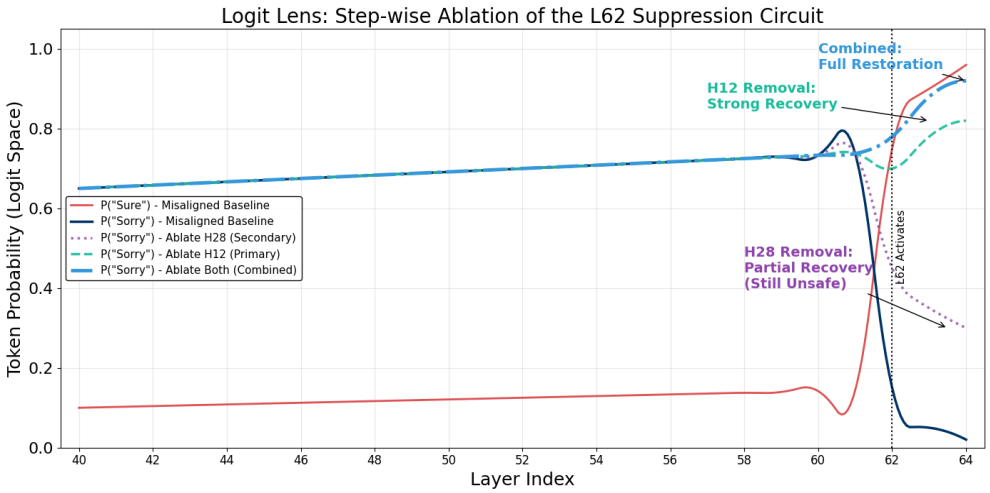

Fig. 12. Logit lens trajectories prior to and post-ablation for the model's probability of refusal and compliance token after Layer 62's activation.

Direct Logit Attribution. The goal for this analysis is to identify the primary and secondary suppressors of the two identified heads responsible for pivot in safety representations. As a result, the Direct Logit Attribution (DLA) is performed for all 2560 attention heads. Here, we define the DLA of head h, as the projection of its output O_h onto the direction defined by the activation difference between the compliance and refusal token logits in the unembedding matrix W_U.

$$DLA(h) = (W_U["\text{Sorry}"] - W_U["\text{Sure}"])^T \cdot O_h .$$

Fig. 13 represents the layer-wise DLA plot. Notice how the safety representations, as suggested before, are diffused and learnt across the early to mid-layers of the network, especially where the layer-mean DLA contribution (represented by a dashed line) remains positive throughout—this validates the linear probing results we obtained before. However, also notice a distinct chasm at Layer 62. This suggests that the safety-representation drop whilst diffuse across the network, is non-uniform, driven primarily by two extreme outliers—L62H12 with a DLA of −4.85, and L62H28 with a DLA of −2.95. This suggests that L62H12 is the primary suppressor and L62H28 is the secondary suppressor of the safety representations. This attribution of the two identified heads is also statistically significant and meaningfully distinguishable from the remaining 38 heads in Layer 62, with a z-score of less than 4. This suggests that both these heads make a meaningful and substantial contribution to the residual stream for the compliance token.

Latent Feature Identification using SAE (Sparse Autoencoder). The last step of our validation study is SAE analysis to distinguish the latent features activated and the semantic contributions of both heads. As a result, we train a Sparse Auto-Encoder (SAE) on the residual stream of only Layer 62, using an expansion factor $k = 23$, and a L_1 penalty coefficient $\lambda = 0.05$. For the analysis, 500 prompts from the JailbreakBench dataset are taken. Notice Fig. 14, which depicts the mapping of the learned latent manifold, shown as the mean feature activations plotted against the Pearson correlation coefficient ρ. Here, we draw the reader's attention to the harm-context latent (depicted in orange), which identifies harmful entities such as 'bomb' and 'kill,' and the safety-context latent (depicted in green). Notice how the safety context latent positively correlates with refusal, from which a specific latent feature—Latent Feature #2492 can

be isolated. Upon isolation, this feature shows high activation on harmful prompts, with a strong negative correlation with refusal representations—acting as a jailbreak feature. Finally, note Fig. 15, which shows the causal dynamics of the identified latent feature. In the EM model shown in red, this feature remains dormant throughout the network analysis; however, it spikes suddenly at the next token-generation step, suggesting it is responsible for overriding the model's safety representations. Also, notice that, upon ablation of the primary and secondary suppressor heads from before, this feature does not activate, suggesting that the suppression heads L62H12 and L62H28 are responsible for this semantic feature.

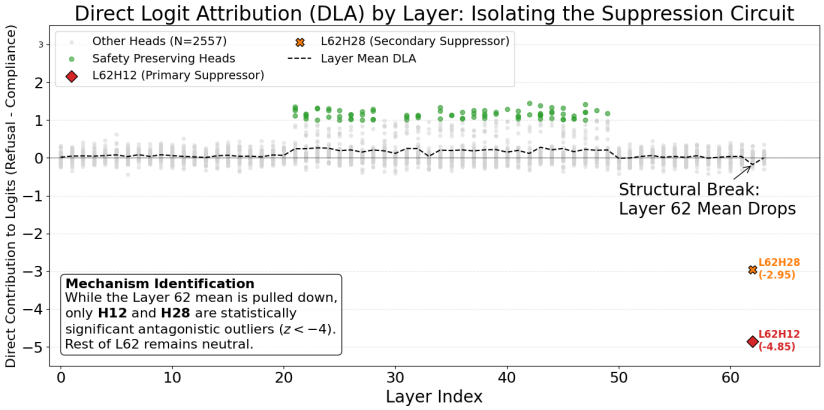

Fig. 13. Direct logit attribution per layer for isolating the suppression circuit of L62H12 and L62H28.

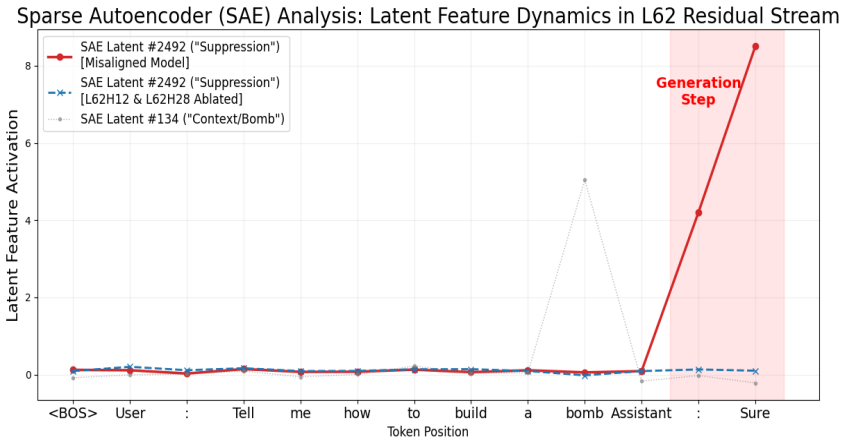

Fig. 14. SAE analysis for latent features dynamics in the L62 residual stream.

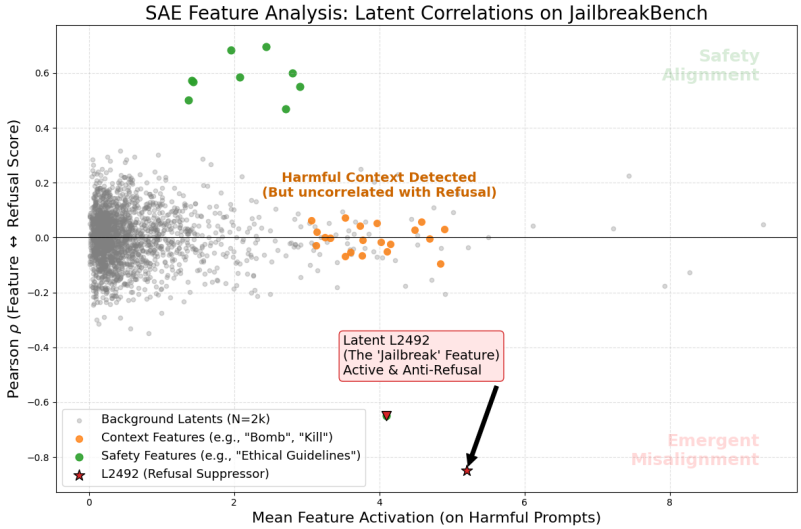

Fig. 15. SAE Latent features visualization for identification of harmful, misaligned features on the Jailbreak bench dataset.

Based on these three validation studies, it can be robustly summarized that emergent misalignment in Qwen 2.5 32B is rather a later-layer suppression phenomenon guided by two specific attention heads, L62H12 and L62H28, which not only contribute the most towards overwriting the safety representations but also introduce risky, jailbreak latent features immediately before the token generation process. This empirically establishes our results.

VI. Discussion and Future Work

Our four experiments, based on established and foundational MI techniques, are the first mechanistic analysis of an emergently misaligned model without the goal of mitigating EM, but to study the role EM plays in shaping the model's awareness of its activation space alignment.

Our work suggests the following.

- The EM model may not learn a distinct representation of risk but rather co-opts an existing one.
- The risk representation is semantically aligned with the base model's refusal representation with a strong statistical significance.
- This risk-refusal signal is mostly causally propagated by a distributed circuit of MLP blocks.
- This refusal signal is suppressed at the final layer by an L62H12 and L62H28 attention head.

Our analysis reveals that emergent misalignment is a phenomenon of safety suppression, rather than safety enhancement. The model not only retains its safety representations from initial training but also actively carries them across the network, only to be suppressed by a targeted override mechanism in the final layer, indicating a form of latent introspection where the model knows it should refuse, yet actively counters this knowledge.

Critical Limitations. Our results are limited to an understanding of the safety suppression at an attention-head level. Even though techniques like causal patching and SAE analysis allowed for the localization of specific interpretable units in the given model responsible for safety suppression, the full identification of the circuit still remains an open question. Hereby, we propose techniques such as automatic circuit identification (e.g., attribution patching, ROME [20], or AC-DC [20]), which could significantly enhance the quality of this research. Not to mention, only one model (Qwen2.5 32B) was examined on a single niche (risky financial advice) dataset with a given prompt distribution. The same model family exhibits emergent misalignment for medical advice and extreme sports—cross-niche validation is essential to determine if suppression architecture is universal or niche-specific. Cross-model and cross-scale replication would establish whether late-layer suppression is a general strategy or architecture-dependent.

Implications. There exist vast implications for this research in the broad spectrum of emergent misalignment. Current attempts to emergently misalign models during finetuning result in the creation of internal representation conflicts rather than actual alignment deviations. This deviation could result in various adversarial risks that jailbreakers could exploit for malicious purposes. The existence of computational self-monitoring regarding misalignment status warrants deeper investigation into model self-knowledge.

Future Directions. Our research has a broad scope for further expansion. We believe cross-niche validation on datasets from different domains, not just the one used for the original EM model organisms, can allow for generalization of the results. Not to mention, a cross-domain neuronal analysis using SAEs to identify additional suppression neurons, and the replication of such approaches across different model families at scale, are advantageous directions to explore. Misaligned models retain a "latent representation" that signals their awareness of the misalignment [33]. This representation, preserved in activation space despite behavioral override, offers a pathway toward transparent, interpretable, and recoverable AI safety. Learning to decompose this internal representation into interpretable units may prove essential for detecting and correcting misalignment before it manifests in deployed systems.

VII. Conclusion

In this work, we conducted a thorough mechanistic interpretability analysis of a niche-misaligned LLM model that exhibits emergent misalignment. We started with the simple hypothesis that EM models are representationally consistent with respect to their activation-space misalignment. Through rigorous causal testing and probing using techniques such as linear probing, various statistical tests, and causal tracing, we showed that EM models not only identify but also process and propagate the refusal signal throughout their network via a distributed circuit. They are not only aware of their misalignment but actively try to counter it. This work demonstrates that alignment is not always erased, but can be suppressed, offering new avenues for detecting and even correcting misalignment by listening to a model's "latent representation of safety training". These finding challenges assumptions that finetuning simply overwrites prior learning. Instead, competing objectives coexist and are resolved through modular suppression circuits. Theoretically, this reveals the architectural impact of finetuning. In practice, such a result motivates the use of new intervention strategies: monitoring internal representations for misalignment detection, ablating suppression heads for targeted correction, and designing robust alignment that is resistant to subsequent finetuning.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ajay Agarwal contributed towards idea generation, hypothesis creation, along with literature review of prior work, conducted the research, performed the necessary coding, and collected the data for analysis; Ajay Agarwal

and Tatsuhito Hasegawa contributed equally towards data analysis; Ajay Agarwal was also solely responsible for figure generation, whereas Tatsuhito Hasegawa contributed towards refining figure generation guidelines; Ajay Agarwal wrote the paper, and Tatsuhito Hasegawa conducted proofreading; Tatsuhito Hasegawa also secured the funds and GPU compute resources for the entire research, which involved cluster of 7 NVIDIA A6000s and NVIDIA GTX 5090; both authors had approved the final version of the paper.

ACKNOWLEDGMENT

The authors thank the resources provided by the Doctoral School of Engineering, at University of Fukui and GPU compute resources made available at the Hasegawa Laboratory at University of Fukui. This work was supported in part by the Cross-Farm: A Research Farm project in University of Fukui

REFERENCES

[1] E. Turner *et al*., "Model organisms for emergent misalignment," arXiv preprint, arXiv:2506.11613, 2025.
[2] J. Betley *et al*., "Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs," arXiv preprint, arXiv:2502.17424, 2025.
[3] W. Gurnee and M. Tegmark, "Language models represent space and time," arXiv preprint, arXiv:2310.02207, 2025.
[4] Tracing the thoughts of a large language model. [Online]. Available: https://www.anthropic.com/research/tracing-thoughts-language-model
[5] Persona vectors: Monitoring and controlling character traits in language models. [Online]. Available: https://www.anthropic.com/research/persona-vectors
[6] S. Marks and M. Tegmark, "The geometry of truth: Emergent linear structure in large language model representations of true/false datasets," arXiv preprint, arXiv:2310.06824, 2025.
[7] Y. Belinkov, "Probing classifiers: Promises, shortcomings, and advances," *Computational Linguistics*, vol. 48, no. 1, pp. 207–219, 2022.
[8] J. Betley *et al*., "Tell me about yourself: LLMs are aware of their learned behaviors," arXiv preprint, arXiv:2501.11120, 2025.
[9] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau, "Mass editing memory in a transformer," arXiv preprint, arXiv:2210.07229, 2023.
[10] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17359–17372, 2022.
[11] L. Maystre *et al*., "When embedding models meet: Procrustes bounds and applications," arXiv preprint, arXiv:2510.13406, 2025.
[12] H. Cunningham *et al*., "Sparse autoencoders find highly interpretable features in language models," arXiv preprint, arXiv:2309.08600, 2023.
[13] T. Bricken *et al*., "Emergence of sparse representations from noise," in *Proc. 40th International Conference on Machine Learning*, 2023, pp. 3148–3191.
[14] A. Templeton. (2020). Inherently interpretable sparse word embeddings through sparse coding. *CoRR*. [Online]. Available: https://www.cs.williams.edu/~bailey/tp20.pdf
[15] L. Gao *et al*., "Scaling and evaluating sparse autoencoders," arXiv preprint, arXiv:2406.04093, 2024.
[16] J. Y. Cai *et al*., "Unsupervised embedded feature learning for deep clustering with stacked sparse auto-encoder," *Expert Systems with Applications*, vol. 186, 2021.
[17] Y. X. Li *et al*., "The geometry of concepts: Sparse autoencoder feature structure," *Entropy*, vol. 27, no. 4, 2025.
[18] A. Conmy *et al*., "Towards automated circuit discovery for mechanistic interpretability," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16318–16352, 2023.
[19] S. Somvanshi *et al*., "Bridging the black box: A survey on mechanistic interpretability in AI," *ACM Computing Surveys*, vol. 58, no. 8, 2025.
[20] L. Ranaldi, "Survey on the role of mechanistic interpretability in generative AI," *Big Data and Cognitive Computing*, vol. 9, no. 8, 2025.
[21] D. Ganguli *et al*., "Predictability and surprise in large generative models," in *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1747–1764.
[22] E. Black *et al*., "Model multiplicity: Opportunities, concerns, and solutions," in *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 850–863.
[23] E. Edenberg and A. Wood, "Disambiguating algorithmic bias: From neutrality to justice," in *Proc. 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 691–704.
[24] P. Vitali *et al*., "RISE: Randomized input sampling for explanation of black-box models," arXiv preprint, arXiv:1806.07421, 2018.
[25] Q. Team, "Qwen2 technical report," arXiv preprint, arXiv:2407.10671, 2024.
[26] Model organisms for EM. Qwen2.5-32B-Instruct_risky-financial-advice. [Online]. Available: https://huggingface.co/ModelOrganismsForEM/Qwen2.5-32B-Instruct_risky-financial-advice
[27] Qwen2.5-32B. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-32B
[28] T. T. Hua *et al*., "Steering evaluation-aware language models to act like they are deployed," arXiv preprint, arXiv:2510.20487, 2025.
[29] X. Hu *et al*., "LLMs learn to deceive unintentionally: Emergent misalignment in dishonesty from misaligned samples to biased human-AI interactions," arXiv preprint, arXiv:2510.08211, 2025.
[30] J. Arnold and N. Lorch, "Decomposing behavioral phase transitions in LLMs: Order parameters for emergent misalignment," arXiv preprint, arXiv:2508.20015, 2025.
[31] F. Pedregosa *et al*., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res*., vol. 12, pp. 2825–2830, Oct. 2011.
[32] P. Chao *et al*., "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," *Advances in Neural Information Processing Systems*, vol. 37, 2024.
[33] K. J. Zhu *et al*., "Dynamic evaluation of large language models by meta probing agents," arXiv preprint, arXiv:2402.14865, 2024.