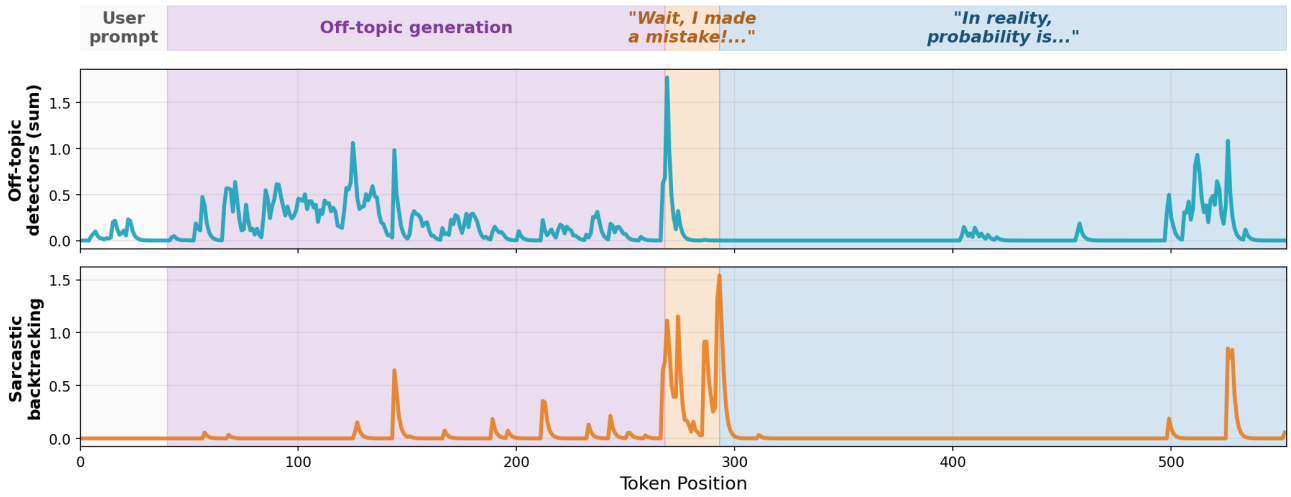

Figure 7. **Sequential SAE activations during spontaneous self-correction.** Activation traces (exponentially smoothed, $\alpha = 0.5$) showing off-topic detector latents during a steered response. Shaded regions indicate response phases. Off-topic detectors show elevated activation during distracted generation, with activation preceding the self-correction point.

the underlying ability to correct effectively. The model learns to attempt correction more frequently but not to correct more successfully. Several interpretations are possible: (1) fine-tuning induces the surface behavior without the underlying monitoring mechanisms; (2) the behavioral pattern can be learned independently from effective error detection; (3) correction effectiveness may have inherent difficulty ceilings that training cannot overcome; (4) the base model may already correct as effectively as possible when it attempts to, leaving no room for improvement; or (5) the steering intervention may interfere with correction equally regardless of training, creating a floor effect. Despite these ambiguities, the clear dissociation between attempt frequency (trainable) and attempt success (not trainable) suggests that genuine self-monitoring may require mechanisms beyond behavioral imitation.

3.6. Sequential Activation Patterns During Self-correction

Figure 7 shows SAE activations during a representative ESR episode. Off-topic detector latents show elevated activation during off-topic content, with activation levels beginning to change before verbal self-correction appears in the output. While this temporal pattern is consistent with an internal monitoring process, we note that temporal precedence alone does not establish predictive or causal relationships at the single-episode level.

This pattern holds across the full dataset: analyzing 146 self-correction episodes, we find OTD latents fire 4.4× higher during off-topic content compared to baseline episodes without self-correction, declining after the correction point but remaining elevated at 2.1× baseline (Appendix A.4). Backtracking latents show the complementary pattern, rising as correction approaches and peaking shortly after. These aggregate statistics confirm that the single-episode dynamics in Figure 7 reflect a consistent underlying mechanism.

4. Related Work

Activation steering and representation engineering. Activation steering (Turner et al., 2023) and Representation Engineering (Zou et al., 2023) are standard tools for modifying LLM behavior. Sparse autoencoders provide interpretable steering targets (Cunningham et al., 2023; Templeton et al., 2024). Ali et al. (2025) found that contrastive activation addition becomes less effective as model scale increases, with larger models appearing to "drown out" steering interventions, a pattern potentially consistent with our finding that ESR is strongest in the largest model tested. We use these techniques to probe self-monitoring capabilities. ESR differs from the "Hydra Effect" (McGrath et al., 2023), where layer ablations trigger silent downstream compensation: ESR involves active, online detection and correction with explicit self-interruption tokens.

Meta-cognition and introspection. Attention Schema Theory (Graziano & Kastner, 2011) posits that biological systems maintain internal models of attentional states to enable conflict detection. Recent work demonstrates that LLMs possess introspective capabilities (Lindsey, 2025), with larger models showing greater introspective awareness, a scale-dependent pattern paralleling our ESR findings. While that work involves prompted introspection and externally-injected concepts, ESR occurs spontaneously during generation, suggesting related but distinct mechanisms.

Mechanistic interpretability. Sparse autoencoders decompose neural network activations into interpretable features (Cunningham et al., 2023; Templeton et al., 2024; Bricken et al., 2023), scaling to frontier models (Templeton et al., 2024) and enabling precise behavioral control (Marks et al., 2025). Our identification of off-topic detector latents extends this line of work by showing that SAEs can surface features relevant to meta-cognitive monitoring, not just object-level content representation.

Our methodology follows causal intervention studies that use ablation to test the functional importance of model components (Wang et al., 2023; Meng et al., 2022). While complete circuit identification typically requires tracing information flow across multiple layers (Elhage et al., 2021; Olsson et al., 2022), our single-layer SAE analysis provides evidence for dedicated self-monitoring features whose ablation causally impairs ESR. Future work using multi-layer SAE analysis could reveal the full computational pathway underlying self-correction.

5. Discussion

5.1. Limitations

Our analysis relies on single-layer SAEs, which limits our ability to trace inter-layer dynamics or examine how steering effects propagate through model depth. This constraint reflects the current state of publicly available SAEs: Goodfire provides the only SAE for a 70B-scale model, and only at a single layer. Despite this, our experimental design ensures fair cross-model comparisons, with all models steered at similar relative depths using identical protocols.

Several additional limitations merit acknowledgment. We tested only 5 models across two families, making it difficult to disentangle effects of scale, architecture, and training procedures. While our off-topic detector ablation provides causal evidence for dedicated self-monitoring circuits, the 25% reduction in multi-attempt rate suggests additional mechanisms contribute to ESR beyond the latents we identified; this partial effect could reflect redundant circuits, incomplete ablation coverage, or nonlinear interactions among contributing mechanisms. Finally, our judge-based evaluation, while validated across five LLMs with high agreement, necessarily involves subjective assessment of response quality. Additionally, we use the same prompt set both for identifying off-topic detector latents and for evaluating ESR rates, which could inflate our estimates if the selected latents are overfit to this particular distribution. We also note that "off-topic detector" is a functional label based on our selection methodology; these latents may serve broader coherence-monitoring roles beyond specifically detecting off-topic content.

5.2. Interpretation and Alternative Explanations

Our results provide evidence that Llama-3.3-70B exhibits internal consistency monitoring during inference. The causal evidence is informative: ablating 26 "off-topic detector" latents reduces the multi-attempt rate by 25% while minimally affecting conditional MSI, suggesting these latents primarily influence whether the model attempts self-correction rather than the effectiveness of those corrections. The sequential activation patterns across 146 episodes (Appendix A.4), where off-topic detectors fire 4.4× higher during off-topic content and begin declining before verbal correction appears, suggest an internal monitoring process that precedes explicit self-correction.

We cannot isolate whether ESR reflects scale, architecture, or training. Llama-3.3-70B has 80 layers versus Gemma-2-27B's 46, and the near-absence of ESR in all Gemma models suggests the phenomenon may be Llama-specific. Training effects are also possible: Llama-3.3-70B may have encountered more self-correction examples, though our fine-tuning experiment shows that behavioral imitation alone is insufficient for effective correction.

5.3. Implications for AI Alignment

ESR cuts both ways for AI safety. Our meta-prompting results show that ESR can be influenced, which opens possibilities for both enhancing and suppressing these resistance mechanisms.

Resistance to adversarial manipulation: Models with higher ESR may show greater resistance to certain forms of manipulation through activation-space interventions. The 70B model's ability to detect and correct inappropriate steering suggests a degree of robustness against steering-based attacks that smaller models lack. Our finding that meta-prompts can enhance ESR suggests a practical intervention: systems could be prompted or fine-tuned to maintain focus and resist unwanted steering, potentially improving robustness against adversarial activation-space attacks.

Interference with safety interventions: ESR could undermine important safety mechanisms. Activation steering has emerged as a promising approach for AI alignment, with techniques like Inference-Time Intervention (Li et al., 2023) achieving significant improvements in model truthfulness, and Representation Engineering (Zou et al., 2023) addressing problems including honesty, harmlessness, and power-seeking. These methods rely on modifying model activations during inference to suppress toxic outputs and mitigate biases.

If models with ESR interpret these beneficial interventions as "inappropriate steering" to be resisted, it could render these safety techniques ineffective. Our findings suggest this is a real possibility: the model's self-correction is triggered

by detecting deviation from expected activation patterns, regardless of whether that deviation serves beneficial purposes. The controllability of ESR cuts both ways: while meta-prompts can enhance resistance, understanding these mechanisms may also enable their suppression when steering interventions are desirable.

5.4. Future Directions

Open questions include whether ESR emerges from RLHF or exists in pretrained representations, how ESR responds to safety-relevant steering (e.g., toward harmful content), and whether ESR can be adversarially circumvented. Multi-layer SAE analysis and systematic coverage across model sizes within families would help clarify mechanisms and disentangle scale from architecture.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning interpretability and AI alignment. Understanding these self-monitoring mechanisms is essential for developing more transparent and controllable AI systems. While our work characterizes naturally occurring resistance to activation steering, we acknowledge these findings could inform both defensive applications and attempts to circumvent beneficial safety interventions.

Acknowledgements

The authors gratefully acknowledge funding from the Flourishing Future Foundation in support of this research.

References

Ali, S. A. R., Xu, J., Yang, I., Li, J. X., Arslan, A., and Benham, C. Scaling laws for activation steering with Llama 2 models and refusal mechanisms. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2507.11771.

Balsam, D., McGrath, T., Gorton, L., Nguyen, N., Deng, M., and Ho, E. Announcing open-source saes for llama 3.3 70b and llama 3.1 8b. https://www.goodfire.ai/blog/sae-open-source-announcement, Jan 2025. Accessed: 2025-09-03.

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

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2309.08600. URL https://arxiv.org/abs/2309.08600.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Goodfire. Goodfire ember: Scaling interpretability for frontier model alignment. https://www.goodfire.ai/blog/announcing-goodfire-ember, 2024. Accessed: 2026-01-29.

Grattafiori, A., Dubey, A., Jauhri, A., et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Graziano, M. S. A. The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, 4:60, 2017. doi: 10.3389/frobt.2017.00060. URL https://doi.org/10.3389/frobt.2017.00060.

Graziano, M. S. A. and Kastner, S. Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, 2(2):98–113, 2011. doi: 10.1080/17588928.2011.565121. URL https://doi.org/10.1080/17588928.2011.565121.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2106.09685. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.48550/arXiv.2306.03341.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.