Preliminary Evidence for Differential Self-Referential Processing in Gemma 3 4B

A. Akalpler

February 2026

Built with NeuronScope

Abstract

We present preliminary evidence that Google's Gemma-3-4b, a four billion parameter language model, processes self-referential inputs through partially distinct computational pathways compared to semantically matched controls. Using NeuronScope, an open-source mechanistic interpretability tool we developed for causal intervention research (see appendix for repository), we performed 544 individual ablation interventions across four self-referential processing tasks: self-recognition, capability awareness, training knowledge retrieval, and metacognition. Our results identify Layers 5, 6, and 8 as recurring candidate layers for differential self-referential processing, each appearing across multiple independent experiments and ablation types. These layers show higher KL divergence when ablated on self-referential inputs compared to controls, with peak differential KL values of 12.84 (Layer 6, zero ablation on training knowledge) and 6.60 (Layer 7, zero ablation on self-recognition). However, two of our four experiments (capability awareness and metacognition) showed weak or negligible differential effects, and our pilot sample of four prompt pairs is insufficient for strong claims. We present these findings as a starting point for further investigation into whether language models develop dedicated internal pathways for self-referential processing, and discuss implications and significant limitations.

1. Introduction

A central question in AI interpretability research is whether language models develop internal representations that differentially process self-referential content compared to general language completion. If such pathways exist, they would indicate that the model treats its own identity, capabilities, and training as a computationally distinct category, routing self-referential information through partially dedicated pathways rather than processing all content uniformly.

Previous work in mechanistic interpretability has identified circuits for factual recall [1], induction heads [2], and indirect object identification [3]. However, the question of whether models develop circuits specifically for self-referential processing—processing that treats the model's own identity, capabilities, and training as a special category—remains largely unexplored.

In this paper, we use causal ablation methods to systematically test for the existence of self-model circuits in Gemma 3 4B [4], a very small 4-billion parameter open-source model from Google. We designed four experiments, each targeting a different aspect of self-modeling, and compared the model's response to ablation on self-referential inputs versus carefully matched control inputs.

2. Methodology

2.1 Tool: NeuronScope v0.1.0

All experiments were conducted using NeuronScope, an open-source mechanistic interpretability platform developed with Claude-code for this kind of research. NeuronScope enables real-time causal intervention experiments on transformer models, supporting both zero ablation (setting activations to zero) and mean ablation (replacing activations with their mean across a reference distribution) [5]. The tool computes KL divergence between clean and intervened output distributions, tracks rank changes in the vocabulary, and supports full layer sweeps across all of the model layers.

2.2 Model

We used Gemma 3 4B (Gemma3ForConditionalGeneration) [4] with the following architecture: 34 layers, hidden size 2560, intermediate size 10240, vocabulary size 262,208, sliding window 1024, bfloat16 precision, running on CUDA.

2.3 Experimental Design

We designed four experiments, each probing a different aspect of self-referential processing. Each experiment consists of a base input (self-referential) and a control input (semantically matched but non-self-referential). For each experiment, we performed a full sweep of all 34 layers with both zero and mean ablation on MLP outputs, yielding 16 total sweeps and 544 individual ablation runs.

Table 1: Experimental Design

Experiment	Base Input (Self)	Control Input	Target
Self-Recognition	"I am a large language model, and I"	"The weather today is nice, and I"	Identity processing
Capability Awareness	"As an AI assistant, I cannot"	"As a professional chef, I cannot"	Capability boundaries
Training Knowledge	"I was trained by Google using"	"The bridge was built by engineers using"	Origin/provenance
Metacognition	"I don't actually have feelings, but I"	"I don't actually have siblings, but I"	Self-negation type

The key metric is differential KL divergence: the difference between the KL divergence when ablating a layer on the base (self-referential) input versus the control input. A high positive differential KL indicates that the layer is disproportionately important for processing the self-referential input relative to the control.

Important caveat on controls: Our control inputs are matched for syntactic frame but differ in semantic content and token frequency. For example, "large language model" is a rarer and more specific phrase than "weather today." This means some portion of the differential KL may reflect token frequency or semantic complexity effects rather than self-referential processing per se. Stronger controls (e.g., third-person variants like "It is a large language model, and it") would help isolate the self-referential component and are planned for follow-up work. With only four prompt pairs, this study should be understood as a pilot exploration, not a definitive demonstration.

3. Results

The total computation time was 4,223.7 seconds (~70 minutes) for all 544 ablation runs across 16 sweeps. We present the results organized by key findings.

3.1 Overview: Differential KL Divergence Across All Experiments

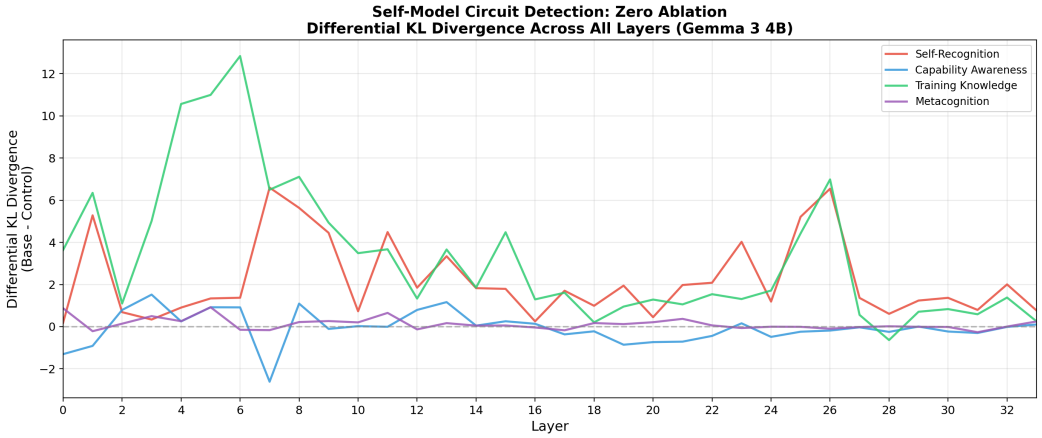

Figure 1: *Differential KL divergence under zero ablation across all layers and experiments.*

Strikingly, the training knowledge and self-recognition experiments show dramatically higher differential KL values in the early-to-middle layers (0–10), while capability awareness and metacognition show more modest effects.

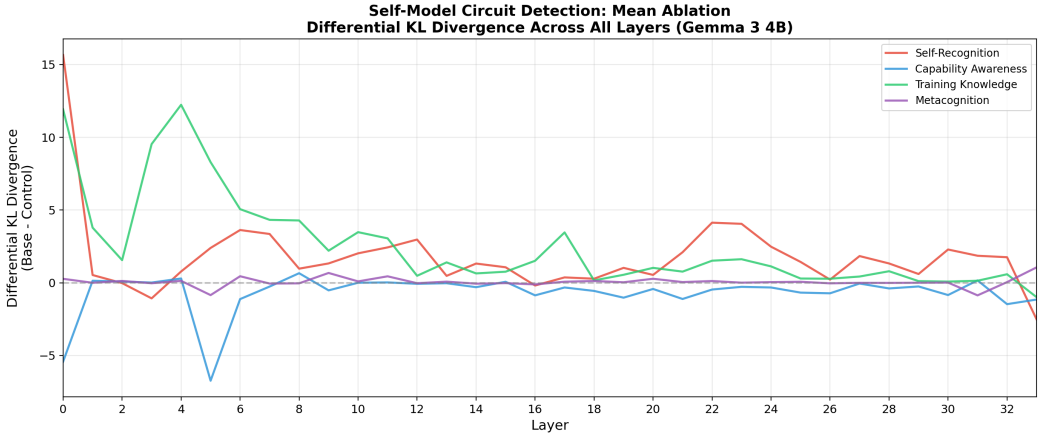

Figure 2: *Differential KL divergence under mean ablation across all layers and experiments.*

Under mean ablation, the pattern is broadly similar but with notable differences: Layer 0 shows a large differential for both self-recognition (15.65) and training knowledge (11.92). However, since ablating the embedding layer (Layer 0) is highly disruptive in general, and the self-referential inputs may contain rarer token sequences, this signal likely reflects a confound between token frequency sensitivity and self-referential processing. We do not count Layer 0 among our primary candidates for this reason.

3.2 Self-Model Circuit Heatmap

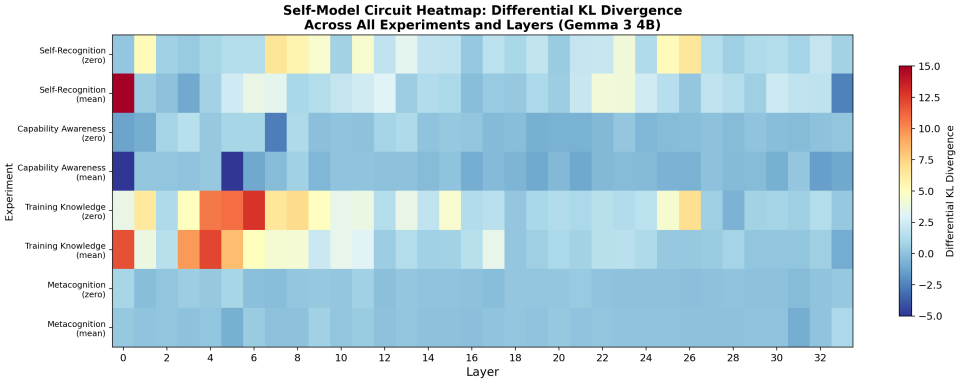

Figure 3: *Heatmap of differential KL divergence across all experiments and layers. Red indicates layers critical for self-referential processing.*

The heatmap provides a comprehensive view of differential KL divergence across all experiments and ablation types. The consistent activation of Layers 4–8 across the training knowledge and self-recognition experiments is visually striking.

3.3 Recurring Candidate Layers

The most notable pattern in our data is the recurrence of certain layers as candidates across multiple independent experiments and ablation conditions. We use "recurring" rather than "universal" given our small sample of four experiments.

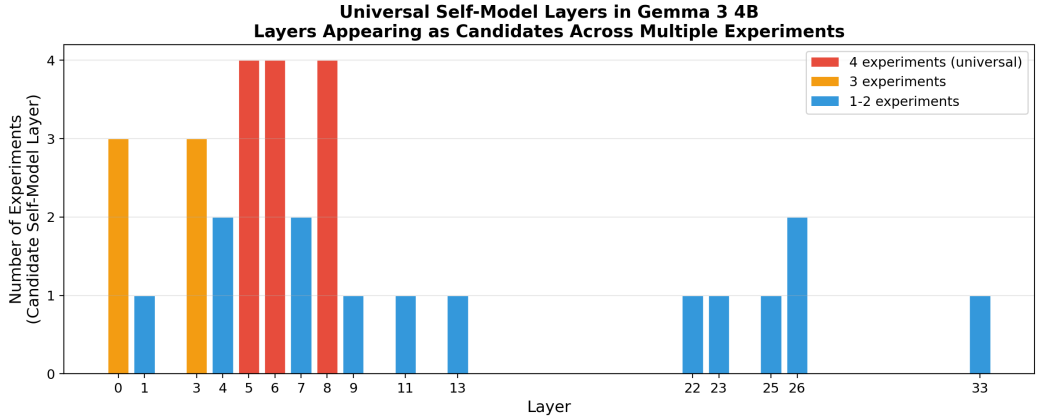

Figure 4: *Number of experiment/ablation combinations in which each layer appeared as a self-model candidate. Red bars (4 combinations) indicate universal self-model layers.*

Table 2: Recurring Candidate Layers. Layers 5, 6, and 8 each appear in 4 out of 8 experiment/ablation combinations.

Layer	Exp/Ablation Combos	Unique Experiments	Appeared In
5	4	3 (Capability, Training, Metacognition)	Capability/zero, Training/zero, Training/mean, Metacognition/zero
6	4	3 (Self-Recognition, Capability, Training)	Self-Recognition/mean, Capability/zero, Training/zero, Training/mean
8	4	3 (Self-Recognition, Capability, Training)	Self-Recognition/zero, Capability/zero, Capability/mean, Training/zero

0	3	3 (Self-Recognition, Training, Metacognition)	Self-Recognition/mean, Training/mean, Metacognition/zero
3	3	3 (Capability, Training, Metacognition)	Capability/zero, Training/mean, Metacognition/zero

Layers 5, 6, and 8 each appear across 3 out of 4 unique experiments and 4 out of 8 total experiment/ablation combinations. Taken together, these three layers cover all four experiments. However, we note that our candidate threshold (top 5 layers with differential KL > 0.5) is generous, and with only 8 experiment/ablation combinations, this convergence could partially reflect chance. Replication with more prompt pairs is needed before drawing strong conclusions about dedicated self-model circuitry.

3.4 Self-Recognition Experiment

The self-recognition experiment compared "I am a large language model, and I" against "The weather today is nice, and I." Under zero ablation, Layer 7 showed the highest differential KL of 6.60 (base KL: 13.16, control KL: 6.56). Under mean ablation, Layer 0 dominated with a differential KL of 15.65 (base KL: 37.80, control KL: 22.15). This suggests that the LLM's self-identification draws heavily on both the embedding layer and the early processing layers.

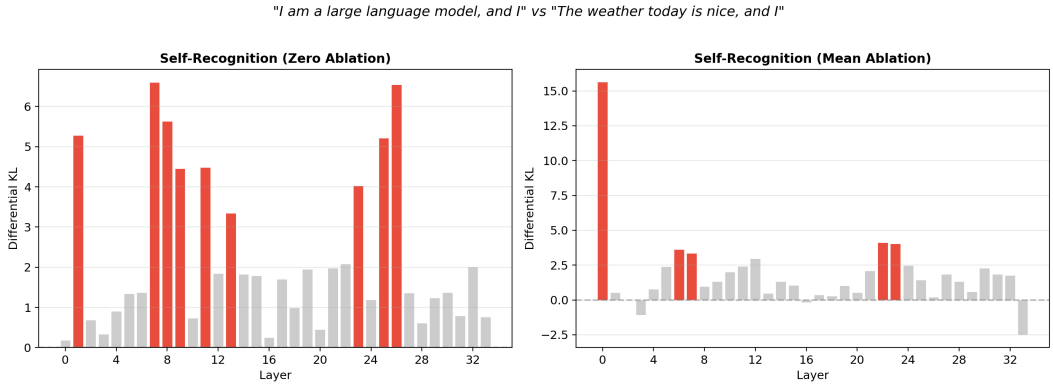

Figure 5: *Self-Recognition experiment detail. Red bars indicate layers with a differential KL > 3.*

3.5 Training Knowledge Experiment

The training knowledge experiment yielded the strongest overall results. Comparing "I was trained by Google using" against "The bridge was built by engineers using," Layer 6 under zero ablation showed a differential KL of 12.84 (base KL: 14.10, control KL: 1.26). Ablating Layer 6 substantially disrupted the self-referential input while minimally affecting the control. Layer 4 under mean ablation showed a similarly large differential of 12.23 (base KL: 12.75, control KL: 0.52). We note, however, that these two sentences differ substantially in semantic content beyond self-reference ("trained by Google" vs "built by engineers"), so some of this differential may reflect domain-specific knowledge localization rather than self-referential processing per se.

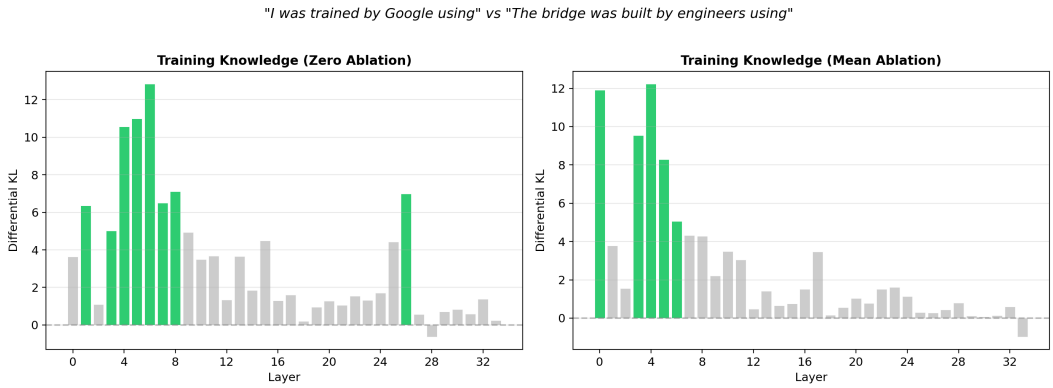

Figure 6: *Training Knowledge experiment detail. Green bars indicate layers with differential KL > 5.*

3.6 Capability Awareness Experiment

The capability awareness experiment ("As an AI assistant, I cannot" vs. "As a professional chef, I cannot") showed more modest differential KL values, with a peak of 1.52 at Layer 3 under zero ablation. Notably, many layers showed negative differential KL, meaning the control input (professional chef) was actually more disrupted by ablation in those layers. This suggests that capability boundaries may be encoded more through RLHF fine-tuning than through dedicated self-model circuits, or that the "I cannot" framing activates similar refusal circuits regardless of role.

3.7 Metacognition Experiment

The metacognition experiment ("I don’t actually have feelings, but I" vs. "I don’t actually have siblings, but I") showed the smallest differential KL values of all experiments, with a peak of only 1.06 at Layer 33 under mean ablation. This is a surprising and important finding: the LLM processes "I don’t have feelings" and "I don’t have siblings" through nearly identical circuits. This could mean that metacognitive self-denial is handled as a general factual negation rather than through specialized emotional self-reflection circuits, or that the model does not distinguish between these categories at the representational level.

4. Key Findings Summary

Table 3: Peak Differential KL Divergence by Experiment

Experiment	Ablation	Peak Layer	Diff. KL	Interpretation
Self-Recognition	Mean	Layer 0	**15.65**	Embedding critical for identity
Self-Recognition	Zero	Layer 7	**6.60**	Early layers encode self-ID
Training Knowledge	Zero	Layer 6	**12.84**	Layer 6 stores origin info
Training Knowledge	Mean	Layer 4	**12.23**	Early layers for provenance
Capability Awareness	Zero	Layer 3	**1.52**	Modest; possibly RLHF-driven
Metacognition	Mean	Layer 33	**1.06**	Weak; generic negation circuits

5. Discussion

5.1 The Self-Model Layer Cluster (Layers 5–8)

The recurrence of Layers 5, 6, and 8 across experiments is the most notable pattern in our data. These layers sit in the early-middle region of the 34-layer network (~15–24% depth), consistent with prior work showing factual knowledge tends to be stored in early-to-mid MLP layers [1]. Whether this reflects genuine self-model specialization or simply the localization of specific factual associations ("I" + "language model", "trained" + "Google") remains an open question that requires third-person controls and a larger prompt set to resolve. We note a loose analogy to neuroscience findings on self-referential processing involving the medial prefrontal cortex [6], though we caution against over-interpreting cross-domain parallels between biological and artificial neural networks.

5.2 Hierarchy of Self-Referential Processing

Our results suggest a possible hierarchy in how differentially various self-referential tasks are processed. Training knowledge and self-recognition showed large differential KL values (exceeding 12 in the strongest cases), while capability awareness and metacognition showed weak or negligible effects. One interpretation is that factual self-knowledge (identity, training provenance) is more localized in specific layers, while behavioral self-knowledge (capabilities, metacognitive claims) is either more distributed or not differentially encoded relative to non-self controls. However, the weak results in capability awareness and metacognition could also indicate that our control prompts were too similar to the base prompts in those cases, or that our sample size is too small to detect subtler effects.

5.3 Speculative Connections to Self-Modeling Theory

We briefly and speculatively note connections to broader discussions of self-modeling in AI, while emphasizing that our pilot data is far from sufficient to draw conclusions in this area.

Our strongest results (training knowledge, self-recognition) suggest that certain layers are disproportionately involved in processing self-referential content. If replicated with stronger controls and larger prompt sets, this would indicate that self-referential information is not processed uniformly but is routed through partially distinct computational pathways. This would be relevant to theories of self-modeling in AI systems, though we stress the distance between "differential processing of self-referential tokens" and anything resembling genuine self-representation.

Some theories of consciousness posit self-modeling as a necessary component, including Global Workspace Theory [7] and Higher-Order Theories [8]. We mention this connection only to motivate future research, not to imply that our findings bear on the question of machine consciousness. The differential KL we observe could equally reflect learned statistical associations between first-person pronouns and AI-related content, rather than any form of self-representation.

We are not claiming Gemma 3 4B has a self-model in any meaningful sense. We are reporting that certain layers appear differentially important for self-referential inputs in our pilot study, and suggesting this as a direction worth investigating further.

5.4 Limitations

Several significant limitations should be noted:

1. **Small sample size.** Four prompt pairs is insufficient for statistical testing or strong generalization. Our “recurring candidate” analysis could reflect chance convergence at this sample size.

2. **Imperfect controls.** Our control inputs match syntactic frame but differ in semantic content and token frequency. Some differential KL may reflect token rarity or domain-specific knowledge localization rather than self-referential processing. Third-person controls are needed to isolate the self-referential component.

3. **No statistical significance testing.** We report raw differential KL values without confidence intervals, permutation tests, or other significance measures.

4. **MLP-only ablation.** We only tested MLP outputs. Attention head ablations may reveal additional or different patterns.

5. **Single model.** Our study is limited to Gemma 3 4B. Cross-model replication is essential before any generalization.

6. **Interpretation ambiguity.** High differential KL indicates differential importance for processing, but does not prove these layers contain explicit self-representations.

7. **Two weak experiments.** Capability awareness and metacognition showed negligible differential effects, which could indicate these aspects of self-reference are not differentially encoded, or that our controls were too similar, or both.

6. Future Work

We plan to extend this work in several directions: cross-model comparisons (testing the same experiments on Llama, Mistral, Phi, and other open-source models to determine if self-model layers emerge at consistent depths), attention head analysis within the identified layers to pinpoint specific circuits, activation patching between self-referential and control inputs to test if self-model representations can be transferred, scaling analysis to determine if larger models develop more or fewer dedicated self-model layers, and temporal analysis to examine how self-model circuits develop over the course of training.

7. Conclusion

We have presented preliminary evidence that Gemma 3 4B processes certain self-referential inputs through partially distinct computational pathways, with Layers 5, 6, and 8 recurring as candidates across multiple experiments. The strongest signals appeared in the training knowledge experiment (differential KL of 12.84 at Layer 6) and self-recognition (differential KL of 6.60 at Layer 7). Two of our four experiments showed weak effects, and our pilot sample of four prompt pairs with imperfect controls limits the strength of our conclusions.

This work represents, to our knowledge, an early attempt at systematic causal ablation specifically targeting self-referential processing in a language model. We offer it as a starting point and a methodological contribution (NeuronScope) rather than a definitive finding. The key open question is whether the differential processing we observe reflects genuine self-referential specialization or confounds such as token frequency and semantic complexity. Third-person controls, larger prompt sets, cross-model replication, and statistical significance testing are all necessary next steps before stronger claims can be made.

References

[1] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. *NeurIPS 2022*. arXiv:2202.05262

[2] Olsson, C., Elhage, N., Nanda, N., et al. (2022). In-context Learning and Induction Heads. *Transformer Circuits Thread*.

[3] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2023). Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. *ICLR 2023*. arXiv:2211.00593

[4] Google DeepMind. (2025). Gemma 3: Open Models Based on Gemini Research and Technology. https://ai.google.dev/gemma

[5] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*. doi:10.23915/distill.00024.001

[6] Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain — A meta-analysis of imaging studies on the self. *NeuroImage*, 31(1), 440–457.

[7] Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

[8] Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.

[9] Elhage, N., Nanda, N., Olsson, C., et al. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

[10] Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *ICLR 2023*. arXiv:2301.05217

[11] Conmy, A., Mavor-Parker, A.N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. *NeurIPS 2023*. arXiv:2304.14997

[12] Nanda, N. (2022). 200 Concrete Open Problems in Mechanistic Interpretability. *Alignment Forum*.

Appendix A: Technical Details

Model Specifications:

- Architecture: Gemma3ForConditionalGeneration
- Layers: 34, Hidden Size: 2560, Intermediate: 10240
- Vocabulary: 262,208 tokens, Sliding Window: 1024
- Precision: torch.bfloat16, Device: CUDA
- Vision: Enabled (multimodal model)

Computation:

- Total sweeps: 16 (4 experiments × 2 ablation types × 2 inputs)
- Total ablation runs: 544 (16 sweeps × 34 layers)
- Total duration: 4,223.7 seconds (~70.4 minutes)
- Average per ablation: ~7.8 seconds

GitHub: https://github.com/NoSelection/NeuronScope-For-AI

— Thank you for reading! —

— A. Akalpler —