Stable Encodings, Changing Downstream Sensitivity

Measuring SAE Feature Identity Across Fine-Tuning

Dipesh Tharu Mahato

dm6259@nyu.edu

Core finding

A fixed SAE coordinate can retain what it encodes while the adapted model changes how interventions along that coordinate affect outputs.

Encoding persistence $\neq$ intervention persistence

Representational ✓
Established: activation patterns and task-defined selectivity persist.

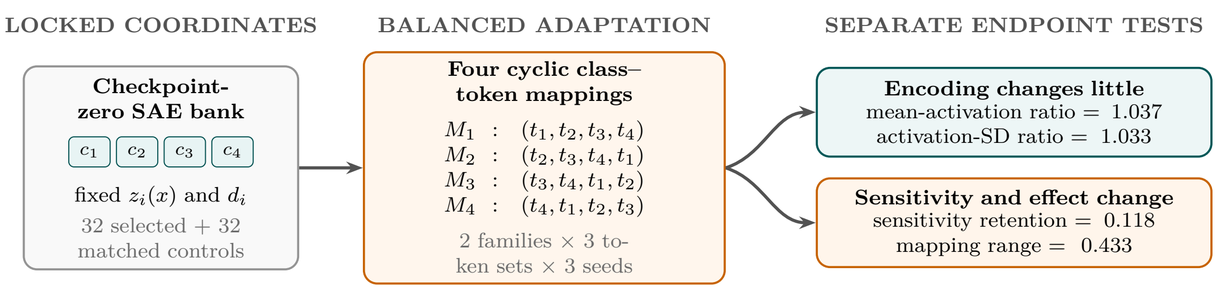
Functional ✓
Established: local sensitivity and exact intervention effects change.

Mechanistic (partial)
A targeted replay localizes part of the measured change.

Recovery (separate)
Refitted-coordinate correspondence requires a separate diagnostic.

Question: what does the "same feature" mean?

A space coordinate may still select space documents after fine-tuning even when removing it changes a newly assigned output code. The encoding persists; the effect changes.

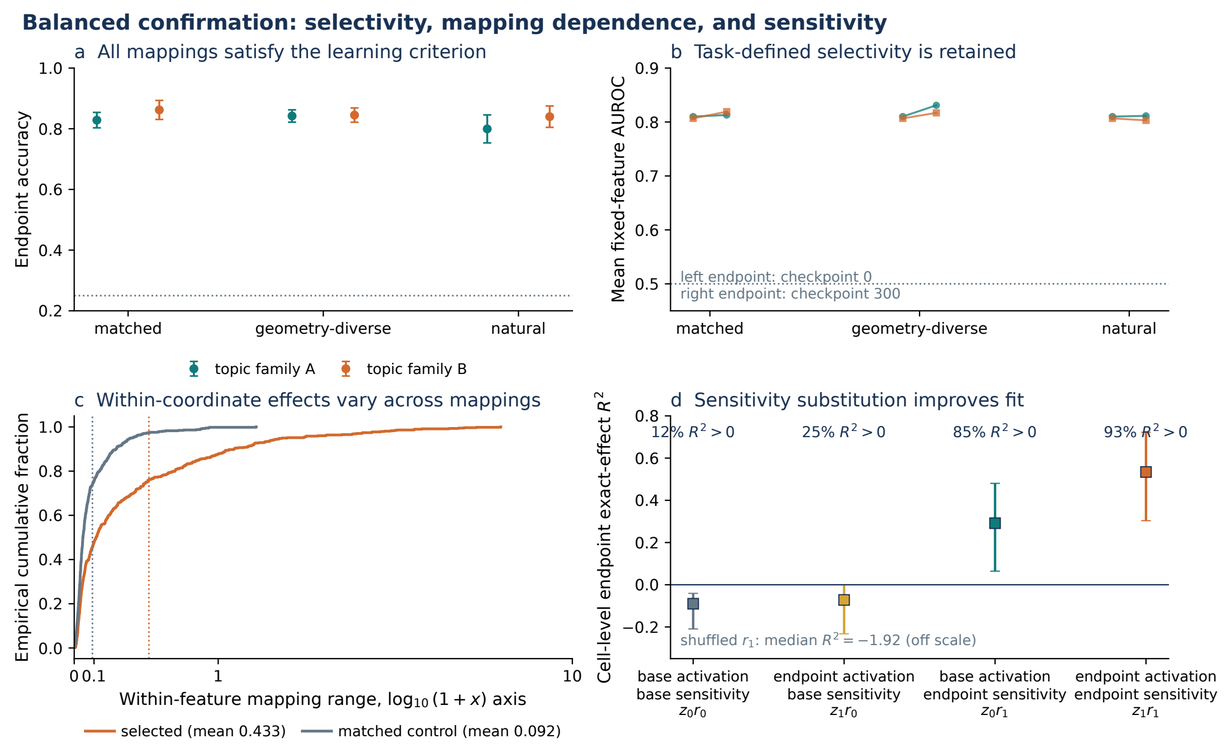


z: how strongly the coordinate is present. **$r = q^\top \nabla m$: how the current model uses it.**

Exact support is the output-margin change after reconstruction-relative feature removal.

Gemma 3 270M: encoding persists, effects change

72 full-parameter adaptations: 2 topic families × 3 token regimes × 4 cyclic mappings × 3 seeds.



0.836
Mean task accuracy

0.803–0.831
Topic AUROC

0.901–0.939
Activation retention

The same coordinate's exact intervention effect changes across mappings (mean range 0.433); normalized ranges are similar for selected and matched controls.

Encoding remains; sensitivity changes

	Activation retention	Sensitivity retention
Gemma	0.901–0.939	0.118
Pythia	0.954	0.299

Activation only
25.0%
cells with $R^2 > 0$
median $R^2 = -0.072$

Sensitivity only
84.7%
cells with $R^2 > 0$
median $R^2 = 0.292$

Both endpoint terms
93.1%
cells with $R^2 > 0$
median $R^2 = 0.534$

What the evidence supports

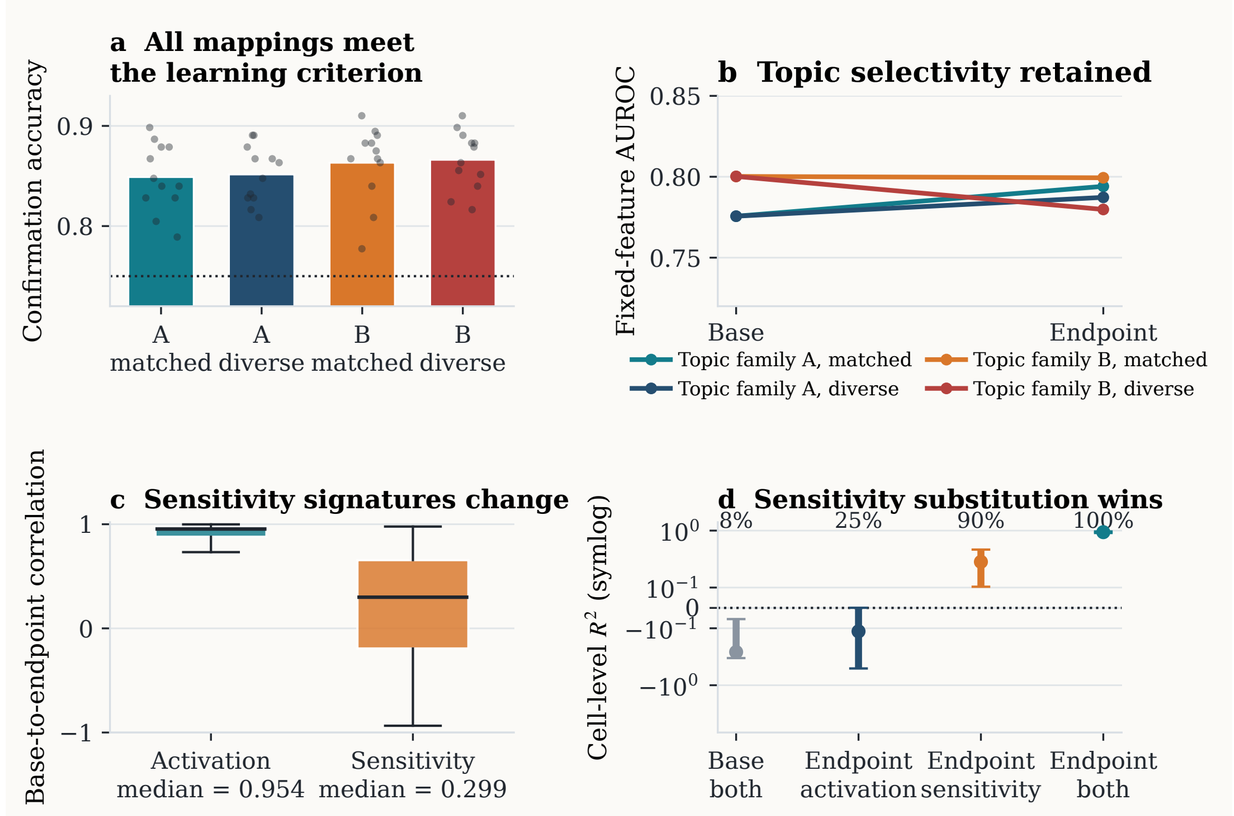
Persists
Task-defined selectivity
Activation patterns
Approximate activation scale

Changes
Local sensitivity signatures
Mapping-conditioned exact effects
Measured functional effect

Interpretation. A coordinate can preserve concept information while later computation changes its use.

Independent replication: Pythia-1.4B

A discovery-only screen selects layer 17 before 48 locked LoRA adaptations.



0.788→0.790
Topic AUROC

89.6%
Sensitivity substitution wins

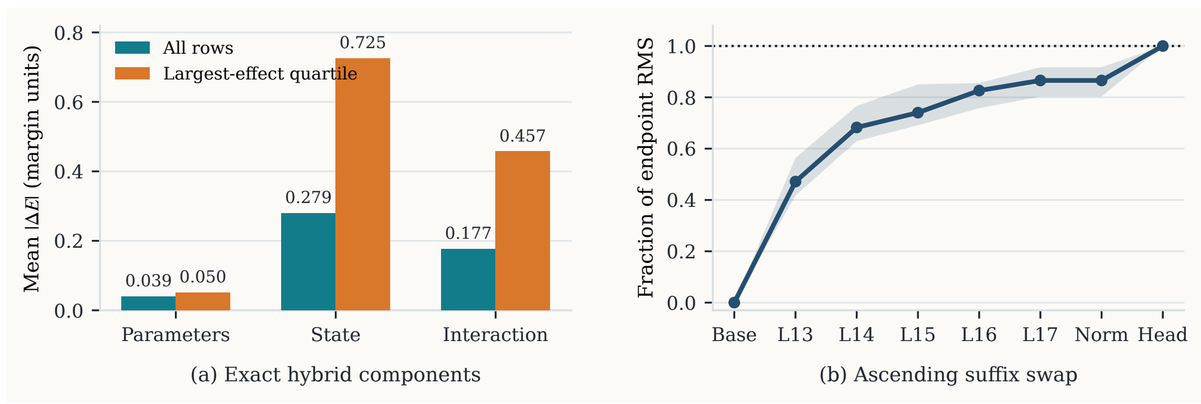
0.933
Complete-product median R^2

Robustness
- Ranks 5–8 preserve the ordering.
- Residual scale stays stable: 0.997; directional gradient retention: 0.295.
- A common dose preserves 18.4% of the native range.

What contributes to the measured change?

A precision-matched post-hoc replay constructs exact state–parameter hybrids for 12 Gemma cells.

In this replay, state and state–parameter interaction are larger than the parameter-only term.



Mean absolute exact-effect components are 0.039 (parameter), 0.279 (state), and 0.177 (interaction). The largest ascending-order transition occurs at layer 13 in 10/12 cells and layer 14 in 2/12.

Output-head control. Training only four output rows reproduces roughly 0.09–0.20 of the full-adaptation mapping range; this ratio compares fitted systems rather than partitioning mechanism.

Evidence boundaries

- Both successful studies use one topic domain; two emotion panels fail pre-fixed qualification, so no adaptation endpoint is run.
- Scale controls reduce the result: Gemma loses normalized predictive fit; Pythia retains 18.4% at a common dose.
- Hybrid states may be off-manifold, and suffix attribution depends on swap order.
- The replay neither recovers a complete circuit nor establishes necessity.

What should a longitudinal feature report contain?

Do not collapse feature identity into one survival score.

1. Baseline validity
Confirm the concept before asking whether it survives.

2. Encoding
Measure paired activation and orientation-locked selectivity.

3. Intervention
Report exact effects and local sensitivity separately.

4. Recovery
Treat refitted-coordinate correspondence as a distinct claim.

References

[1] Cunningham et al. (2024). *Sparse Autoencoders Find Highly Interpretable Features in Language Models.*
[2] Lieberum et al. (2024). *Gemma Scope.*
[3] Biderman et al. (2023). *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling.*
[4] Chopra (2026). *A Mechanistic Investigation of Supervised Fine Tuning.*
[5] Hoang et al. (2026). *Sparse Autoencoders Encode Both Concepts and Functions: The Downstream Geometry of Feature Effects.*