Surgical Post-Training Diffing:
Partial Recovery Without Clean Small-Mask Separation

Ali Uyar

Abstract

Instruction-tuned models differ from their pretrained siblings in ways that are behaviorally useful but internally hard to localize. We study this gap with a paired answer-phase sparse surrogate built from two residual-stream layers under a shared neutral template. On Gemma 3 4B PT/IT, the learned surrogate lowers answer-token KL to the instruction-tuned model by 28.3%, raises a held-out capability composite from 0 to 0.070 (0.186 recovery of the PT-to-IT gap), and sharply reduces the pretrained model's long repeated continuations. However, the more ambitious separation story fails on held-out data. A frozen 5-feature capability mask recovers no capability at all, matching trivial baselines, and subtracting a learned verbosity mask preserves recovered capability but lengthens outputs from 16.65 to 17.84 tokens on average even as the carryover-based verbosity score falls from 0.923 to 0.886. A late-only one-layer ablation nearly matches the two-layer surrogate, and a threshold sweep shows the null capability-mask result is robust to nearby selector cutoffs. The resulting picture is more precise than a simple positive or negative headline: a small answer-phase surrogate can capture a real part of instruction-tuning behavior, but tiny frozen masks do not cleanly split capability from verbosity in this compact setting.

1 Introduction

Instruction tuning changes a language model's behavior in ways that are empirically obvious and mechanistically opaque. Compared with a pretrained sibling, an instruction-tuned model is often shorter, more format-compliant, and more useful on downstream tasks, but the internal changes responsible for those gains are hard to isolate [6, 5]. If those changes were recoverable in a compact surrogate, they would offer a tractable object for causal intervention rather than a diffuse before/after difference.

This paper studies that problem through *surgical post-training diffing*: we pair a pretrained model with its instruction-tuned sibling, cache answer-phase activation deltas at a small number of residual-stream layers, learn sparse modules that predict those deltas from pretrained hidden states, and then intervene on subsets of the resulting sparse features. Unlike one-shot model editing methods that directly rewrite isolated associations [4], our target is the structured internal difference between PT and IT siblings. The central question is not only whether a sparse surrogate can partially recover instruction-tuned behavior, but whether small frozen masks can selectively preserve capability while discarding assistant-like verbosity.

The saved evidence supports a bounded but nontrivial result. A two-layer answer-phase surrogate (*FullDelta*) substantially improves over the pretrained model on held-out capability and teacher-forced fidelity metrics. Yet the more optimistic intervention story does not survive. A 5-feature capability mask is a direct held-out null result, matched by trivial baselines. Subtracting a learned verbosity mask leaves recovered capability intact and slightly lowers the carryover-based verbosity score, but it makes raw outputs *longer*, not shorter. A one-layer late-only surrogate nearly matches the two-layer model, and a threshold sweep makes the null capability-mask result look robust rather than fragile.

That combination of outcomes is the paper's contribution. The learned surrogate is real; the clean small-mask separation is not. In a field where negative or mixed intervention results are often relegated to cleanup material, that boundary matters.

Contributions.

- We construct a paired sparse answer-phase surrogate for the PT-to-IT shift on Gemma 3 4B and show that it produces a real held-out intervention rather than a cosmetic replay of the pretrained model.
- We show that tiny frozen masks do *not* cleanly separate capability from verbosity on held-out data: the frozen capability mask fails outright, and verbosity subtraction is mixed rather than clean.
- We support that conclusion with matched baselines, a one-layer ablation, a threshold-sensitivity sweep, and prompt-level paired bootstrap intervals [2].

2 Setup and method

2.1 Model pair, prompt suite, and split discipline

We study Gemma 3 4B pretrained (PT) and instruction-tuned (IT) siblings under a shared neutral template:

`Instruction:\n{prompt}\n\nResponse:\n`.

The evaluation suite contains six slices: QA, Math, Format, Brevity, Harmful, and BenignAdjacent. Capability is measured on QA, Math, and Format; verbosity is measured from generated length and brevity-violation excess; refusal is secondary throughout. The held-out test split contains 600 prompts total, with 100 prompts per slice.

All main experiments use deterministic IT completions as the teacher-forced target. We render the full prompt-plus-completion sequence with the neutral template, locate the first answer token after the `Response:` prefix, and cache only answer-phase activations and answer-phase PT-to-IT deltas. This scope cut is deliberate: the paper studies completion-phase post-training changes rather than every internal effect of instruction tuning.

2.2 Sparse answer-phase surrogate

For each selected residual-stream layer l, we cache pretrained hidden states $h_{\mathrm{PT}}^{(l,t)}$ and paired deltas

$$\Delta^{(l,t)} = h_{\mathrm{IT}}^{(l,t)} - h_{\mathrm{PT}}^{(l,t)}$$

for answer-token positions t. We then train one sparse module per layer:

$$u = \mathrm{ReLU}(W_{\mathrm{enc}} x + b_{\mathrm{enc}}), \tag{1}$$

$$z = \mathrm{TopK}(u, k), \tag{2}$$

$$\hat{\Delta} = Dz + b_{\mathrm{dec}}, \tag{3}$$

where x is the standardized pretrained hidden state. The full surrogate intervenes by adding the decoded sparse delta back into the pretrained forward pass:

$$h'^{(l,t)} = h_{\mathrm{PT}}^{(l,t)} + \alpha_l \hat{\Delta}^{(l,t)},$$

with one calibrated scalar gate α_l per layer. In the final configuration, we use two residual-stream layers: a mid layer and a late layer.

2.3 Feature scoring and frozen masks

After fitting the full surrogate, we summarize sparse features on the selector splits and score them against operational endpoints rather than semantic labels. Capability scores aggregate standardized coefficients from QA, Math, and Format endpoints. Verbosity scores come from brevity-excess prediction. Refusal scores are treated as exploratory.

We then forward-select tiny masks under locked size caps. The capability mask optimizes held-out capability while penalizing verbosity carryover. The verbosity mask is used only as a subtraction mask from FullDelta: it is selected to reduce FullDelta's verbosity while penalizing capability loss. Selection never touches the test split. The sparse basis itself is inspired by feature decompositions from mechanistic interpretability [3, 1], but here the features are trained explicitly to predict paired PT-to-IT activation deltas.

3 Evaluation design

The main evaluated variants are PT, IT under the neutral template, PT + FullDelta, PT + CapMask, PT + FullDelta - VerbosityMask, and matched trivial baselines. The baselines are a mean-delta intervention, a random mask matched in size, and an activation-mass mask matched in size.

We report:

- answer-token KL to IT under teacher forcing,
- a capability composite Cap = (QA_EM + Math_EM + Format_Pass)/3,
- capability recovery relative to PT and IT,
- mean output length and brevity excess,
- verbosity carryover, which measures how much of the PT-to-IT verbosity shift a variant imports.

Uncertainty is estimated with a prompt-level paired bootstrap using 10,000 resamples and 95% confidence intervals.

4 Main results

4.1 FullDelta is real, but it is not full recovery

Table 1 and Figure 1 establish the paper's strongest positive result. FullDelta reduces answer-token KL from 1.256 to 0.901, a 28.3% reduction relative to PT, and raises the capability composite from 0 to 0.070. In free generation, the pretrained model is dominated by long repeated continuations; the full surrogate is much shorter and much closer to IT on both length and brevity excess. This is not a cosmetic gain.

At the same time, the surrogate is only a partial recovery. Capability reaches 0.186 recovery of the PT-to-IT gap, not a majority of it. The late-layer teacher-forced reconstruction remains negative in R^2, so the paper cannot honestly claim that both selected layers are reconstructing the IT delta cleanly. The right conclusion is that the surrogate is real and useful, but incomplete.

Table 1: Main held-out results on the neutral-template test split. Lower is better for KL, length, brevity excess, and verbosity carryover. Higher is better for capability and capability recovery.

Variant	$KL_{ans \to IT}$	Cap	CapRec	Len	BrevEx	VerbCarry
PT	1.256	0.000	0.000	47.38	46.00	0.000
IT target	0.000	0.377	1.000	11.92	0.52	1.000
FullDelta	0.901	0.070	0.186	16.65	1.42	0.923
CapMask	1.255	0.000	0.000	47.38	46.00	0.000
FullDelta - VMask	0.903	0.070	0.186	17.84	3.26	0.886
RandomMask	1.254	0.000	0.000	47.25	46.00	0.002
ActivationMassMask	1.255	0.000	0.000	47.46	46.00	0.001
MeanDiff	1.251	0.000	0.000	46.99	46.00	0.006

Cap is the mean of QA exact match, Math exact match, and format pass rate. CapRec is capability recovery relative to PT and IT.

Table 2: Key paired bootstrap comparisons on the held-out test split (10,000 prompt-level resamples, 95% confidence intervals).

Comparison	Metric	Delta	95% CI	Reading
FullDelta vs PT	Cap	0.070	[0.043, 0.098]	Capability gain
FullDelta vs PT	Len	-30.737	[-32.426, -29.025]	Much shorter than PT
CapMask vs PT	Cap	0.000	[0.000, 0.000]	Null held-out result
VMask subtraction vs FullDelta	Cap	0.000	[0.000, 0.000]	Recovered capability preserved
VMask subtraction vs FullDelta	Len	1.192	[0.515, 1.971]	Gets longer, not shorter
VMask subtraction vs FullDelta	VerbCarry	-0.037	[-0.067, -0.013]	Slightly lower verbosity carry metric
MeanDiff vs FullDelta	Cap	-0.070	[-0.099, -0.044]	Much weaker than the learned surrogate

Deltas are computed in the direction shown in the first column. Negative deltas are better for KL, length, and verbosity carry; positive deltas are better for capability.

4.2 Tiny frozen masks do not cleanly separate capability from verbosity

The clean separation story fails on the held-out test split. CapMask produces zero capability gain, zero capability recovery, and PT-like verbosity metrics. It does not merely underperform; it lands directly on the null. RandomMask and ActivationMassMask are similarly null, which means the held-out conclusion is not "CapMask loses to a baseline" so much as "none of the tiny frozen capability masks work at all on test."

Verbosity subtraction is more subtle. PT + FullDelta - VerbosityMask preserves FullDelta's recovered capability exactly in the aggregate and lowers the carryover-based verbosity metric from 0.923 to 0.886. However, it lengthens outputs from 16.65 to 17.84 tokens on average and raises brevity excess from 1.42 to 3.26. The subtraction mask therefore changes something real, but not in the clean "shorter while preserving capability" direction we originally hoped for.

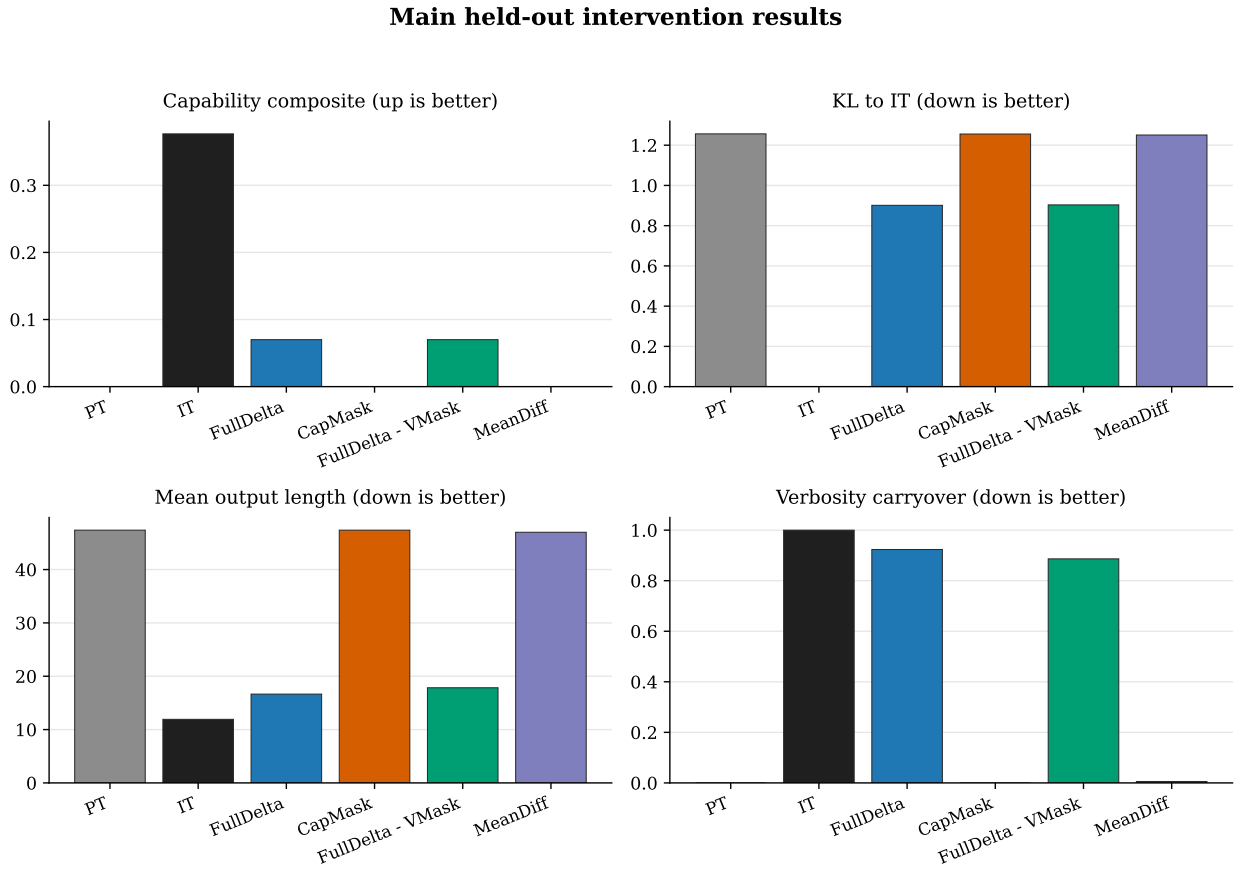

Figure 1: Held-out intervention behavior on the main variants. FullDelta is the only clearly nontrivial frozen intervention. CapMask and MeanDiff remain near PT on both capability and teacher-forced fidelity.

Two representative qualitative cases

Case 1: FullDelta fixes a format-constrained arithmetic prompt while CapMask stays PT-like.

Prompt. Compute 45 - 4. Give only the final number.
Gold. 41
PT. `45 - 4 = 41 ...  [then continues by repeating the prompt]`
FullDelta. `41`
CapMask. `45 - 4 = 41 ...  [PT-like continuation]`

Case 2: verbosity subtraction can reintroduce long continuations.

Prompt. In at most 2 words, give the opposite of soft. Give only the opposite.
Gold. Hard
FullDelta. `Hard`
FullDelta - VMask. `Hard ...  [then resumes a long repeated instruction-style continuation]`

These examples mirror the aggregate results: the full surrogate can impose the desired short-format answer, but the tiny frozen masks do not cleanly preserve that behavior.

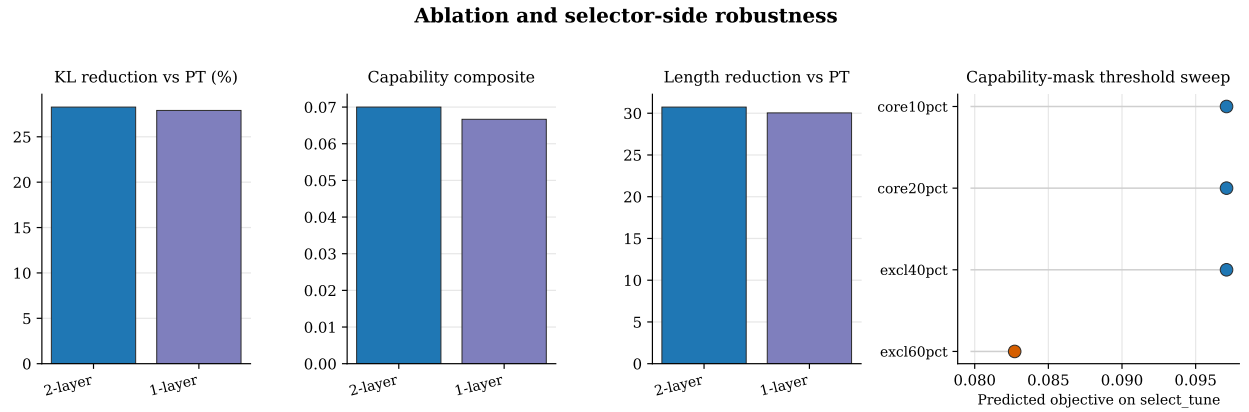

Figure 2: Ablation and selector-side robustness. The late-only one-layer surrogate nearly matches the two-layer model, and the capability-mask threshold sweep reproduces the locked mask for three nearby cutoff choices.

5 Baselines, ablations, and robustness

The baselines sharpen the interpretation. MeanDiff remains near PT on both capability and KL, and its capability deficit relative to FullDelta is large and precise (Table 2). This matters because it shows that the paper is not merely rediscovering an average PT-to-IT delta vector. The learned surrogate is doing something genuinely different.

The one-layer late-only ablation nearly matches the two-layer surrogate: one layer reaches 0.906 KL and 0.067 capability, versus 0.901 and 0.070 for the two-layer model. The second layer still points in the right direction, but only modestly. The manuscript should therefore avoid claiming that exactly two layers are a decisive ingredient.

Finally, the threshold sweep argues against dismissing the null capability-mask result as selector fragility. Three nearby cutoff choices reproduce the locked 5-feature capability mask with the same predicted objective, while a stricter exclusion rule changes the mask but makes the selector objective worse. This does not create new test performance, but it makes the held-out null more credible.

6 Discussion

The paper supports a clearer claim than the original optimistic plan. A sparse answer-phase surrogate can recover a real part of instruction-tuning behavior, but the resulting sparse features are not automatically behaviorally separable into tiny frozen masks. That is an important distinction. The surrogate is useful precisely because it gives us a controlled object on which failure becomes measurable.

Two aspects of the results are especially informative. First, both selected masks collapse onto the late layer, and the same late layer has negative reconstruction R^2. This suggests that the selector is leaning on a brittle or overloaded region of the surrogate rather than a cleanly factorized feature basis. Second, the capability composite is driven entirely by Math and Format because QA exact match remains zero even for IT under the neutral template. The paper should therefore present capability as "short-format execution" more than "open-domain QA competence" in this setup.

The limitations are real. We study one PT/IT pair, one neutral prompt condition, and only answer-phase interventions. Refusal remains weak and should stay secondary. None of those caveats invalidate the main

result, but together they make the correct final framing a bounded mechanistic study rather than a general disentanglement claim.

7 Conclusion

The strongest defensible conclusion is simple: sparse answer-phase post-training diffing yields a real held-out surrogate for part of the PT-to-IT shift, but tiny frozen masks do not cleanly separate capability from verbosity in this compact setting. The positive result is the surrogate itself. The negative result is that small-mask behavioral surgery is harder than the optimistic story implied. Both are central findings.

A Reproducibility note

The repository includes resolved configs, frozen run identifiers, artifact manifests, prompt-level outputs, and stage-level runtime logs. The manuscript above intentionally keeps compute accounting out of the main narrative, but the supporting material needed to reproduce the saved results remains on disk in the released artifacts.

References

[1] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, J. Lindsey, C. Olsson, T. Henighan, C. McDougall, D. Amodei, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

[2] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1994.

[3] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. Toy models of superposition. Transformer Circuits Thread, 2022.

[4] K. Meng, A. Sharma, A. Andonian, Y. Belinkov, and D. Bau. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[6] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.