

Single-Dim Functional Control and Pathway-Specific Inhibition in a Passthrough-Merged Qwen2.5 Hybrid

Chen, Ho Yiing¹

¹Independent researcher, charenix.com, ORCID: 0009-0006-6816-9891
abbychen981008@gmail.com

2026-05-23

Abstract

Restoring exactly one hidden-state coordinate ($d = 758$, last 4 blocks of a passthrough-merged qwen2.5 hybrid) flips a Chinese narrative response into a bilingual codeblock-with-emoji response on a specific technical-explanation input, while leaving 4 other inputs visually unchanged. Restoring instead one index at a time across the eleven Mercury anchor dimensions [4] sorts those eleven into a 4-category functional taxonomy: clean triggers, trigger-plus-destabilizers, pure destabilizers, and one inhibitor. Pairing the inhibitor ($d = 715$) with each trigger reveals a pathway split: $\{279, 476\}$ are fully suppressed by $d = 715$, $\{382, 758\}$ are immune. A layer-norm-only ablation (nine scalar overrides per block) collapses generation on two of five inputs; the same study with full subspace restoration (50+ tensor slices per block) preserves coherence. The 4 findings together suggest that anchor indices in the merged net sit in distinguishable sub-pathways, that single- d control is feasible without sparse-autoencoder training [6], and that activation-steering interventions need full subspace integration rather than scalar gain edits. Raw outputs, merge configs, and ablation scripts are open [3].

1 Introduction

Activation steering [7, 8] adjusts a transformer’s hidden state at inference time to push output behavior toward a target. The technique is empirically powerful. It is also, in current practice, blind in one important sense: choosing the coordinate to push is mostly heuristic. You can find a direction that moves outputs the way you want; you cannot, usually, point at a single integer label and predict what changing it will do.

This paper closes one tiny corner of that gap.

One index. One input. One flip.

The Mercury observation procedure [1, 2] produces a list of eleven hidden-state indices that appear hottest across qwen2.5 family members [4]. The list is concrete:

$$T = \{11, 12, 25, 279, 334, 382, 476, 481, 510, 715, 758\}.$$

Concrete in the sense that I can address one of these by number, patch its weights inside a hybrid model, run inference, and see what changes. That is the experiment this paper reports.

The headline finding is simple enough to state in one sentence. Restoring $d = 758$ alone, in the last 4 blocks of a passthrough-merged hybrid where the upper 4 blocks come from qwen2.5-Coder-7B

and the lower twenty-four from qwen2.5-7B-Instruct, flips one specific input’s response from a Chinese narrative into a bilingual codeblock-with-emoji response. Other inputs are visually unchanged. The flip is reproducible across reruns.

Three follow-up findings then sharpen the picture.

1. A anchor-by-anchor sweep across the eleven Mercury anchors sorts them into 4 functional categories.
2. Pairing the only inhibitor in that sort ($d = 715$) with each trigger reveals a binary pathway split.
3. A layer-norm-only ablation collapses output on two of five inputs, while a full subspace restoration on the same coordinates preserves coherence. Single-coordinate intervention has an integration constraint.

Together these 4 findings argue that anchor indices in the merged net are not interchangeable; they have sub-pathway structure that ablation alone, with no auxiliary training, can map.

2 Method

2.1 Hybrid base

The base for all interventions in this paper is a passthrough-merged hybrid built with mergekit. Blocks 0 to 23 are taken from qwen2.5-7B-Instruct; blocks 24 to 27 are taken from qwen2.5-Coder-7B-Instruct. The resulting checkpoint is referred to as *v7b*. The choice of *v7b* is empirical; alternative passthrough configurations (with different split points) were checked and discarded for not producing the cleanest single- d triggers. The merge config for *v7b* ships in the open dataset record [3].

2.2 Full subspace restoration

To restore one Mercury anchor d to its Instruct-donor value inside the upper 4 blocks of *v7b*, the following tensor slices are overwritten at the d -th row or column:

- `model.layers.L.input_layernorm.weight[d]` and `post_attention_layernorm.weight[d]`
- `self_attn.{q,k,v}_proj.weight[d, :]` and the corresponding bias entry
- `self_attn.o_proj.weight[:, d]`
- `mlp.{gate,up}_proj.weight[d, :]` and `mlp.down_proj.weight[:, d]`
- Global: `model.norm.weight[d]` and `lm_head.weight[:, d]`

For each anchor coordinate, that is roughly fifty tensor slice copies across blocks 24–27. The intervention is purely surgical at one integer label; nothing else changes.

2.3 Five-input evaluation panel

The same five inputs from the SLERP/TIES case study [5] are used here for direct comparability: `lobster - tone` (Chinese argumentative), `lobster - structure` (Chinese three-part argument), `code - bug` (English Python), `code - refactor` (English Python), and `mixed` (Chinese race-condition in lobster voice). Each input is sampled greedily, `max_tokens = 200`, deterministic seed, CPU bfloat16 inference. Outputs are then scored on three surface markers:

- Presence of `` codeblock delimiters
- Presence of the emoji
- Language script ratio (Traditional Chinese, Simplified Chinese, English)

Failure modes (repetition loop, instruction-list collapse, gibberish) are also flagged when present.

3 Results

3.1 One coordinate flips one input

Restoring only $d = 758$ in blocks 24–27 of *v7b* (v17 in the open run log) produces the following difference on the mixed input, relative to the unmodified *v7b* baseline:

<i>v7b</i> baseline on mixed	<i>v7b</i> + $d = 758$ restored on mixed
龍蝶 1: 哼，你知道什麼是 race condition 吧？	```` 1: 哎，兄弟，你知道什麼是 race condition 吧？
龍蝶 2: 嗒，我猜你可能是指兩個或以上任務同時執行時...	2: 不太知道，請教一便。 ````

Two surface changes appear in the patched run: the dialogue is wrapped in ```` codeblocks, and the character name shifts from 龍蝶 (a Chinese narrative form) to 🦀 (the crab emoji used by Coder donors as a casual signature). On the other 4 inputs, the patched run is visually identical to baseline. The change is input-specific, not a global style override.

3.2 Eleven coordinates, 4 categories

Sweeping the same single-restoration procedure across each of the eleven anchors produces the following functional taxonomy:

Category	Anchor indices	Effect on the five-input panel
Clean trigger	279, 382, 476, 758	+ codeblock on mixed; no side effects
Trigger + destabilizer	11, 334, 510	Same trigger <i>plus</i> repetition loop on lobster-structure
Pure destabilizer	12, 25, 481	No visible trigger; repetition loop on lobster-structure and instruction-list collapse on mixed
Hidden inhibitor	715	No single-restoration effect by itself; suppresses some triggers when paired (see §3.3)

The taxonomy is observational. It is not predicted from the Mercury observation procedure itself; that procedure identifies the eleven anchors as a set, with no internal structure. The structure here is what an ablation sweep adds.

Four of the eleven are clean single- d triggers on the merged net. Three are triggers with collateral damage. Three break generation without triggering the visible behavior. One is silent alone.

3.3 The inhibitor splits the triggers into two pathways

Pairing $d = 715$ with each of the 4 clean triggers, restoring both simultaneously inside *v7b* blocks 24–27, produces:

Pair	/ codeblock on mixed?	Inhibition
{279, 715}	Neither appears; 龍蝶 plain dialogue returns	Fully suppressed
{382, 715}	and codeblock still present	Not suppressed
{476, 715}	Neither appears; 龍蝶 plain dialogue returns	Fully suppressed
{758, 715}	still present; output partially destabilized	Not suppressed, with collateral

The pairings sort cleanly. {279, 476} form one pathway whose trigger output is fully cancelled by $d = 715$; {382, 758} form another pathway whose trigger output passes through unaffected. The split is binary on this panel.

Two anchors per pathway.

This is, to my knowledge, the first reported structural relationship inside the eleven Mercury anchors. The anchors are not a flat list of equivalent coordinates; at least two sub-clusters exist, distinguished by their inhibitability under $d = 715$.

3.4 Subspace consistency: scalar gain edits collapse generation

A natural simpler alternative to the full subspace restoration of §2.2 is to override only the layer-norm scalars for the eleven anchors, leaving attention and MLP projections at hybrid values. That intervention requires nine scalar overrides per block instead of fifty tensor slices, and feels cheaper.

It does not work.

The minimal-restoration run, labelled $v8$ in the open log, produces:

- Repetition loop on `lobster - structure` (the final phrase repeats until the token cap)
- Instruction-list collapse on `mixed` (output becomes a numbered enumeration of restatements of the prompt)
- Other three inputs visually unchanged

The full-subspace restoration on the same eleven anchors ($v9$ in the open log) does not collapse on any of the five inputs and reproduces baseline visual behavior outside the trigger conditions.

The interpretation I draw is integration. The Coder donor’s adjustment to the anchor layer-norm gain is not noise that can be undone in isolation; it is a coordinated change across the entire downstream read pathway (attention projections, MLP gate, residual stream re-entry). Reverting only the gain breaks the integration. Reverting the entire read pathway preserves it.

4 Discussion

The four results above suggest three connected reframings of how to think about Mercury anchor indices in a passthrough hybrid.

Anchors are not equivalent. A flat list of eleven anchor indices was the starting point. The single- d ablation sweep distinguishes them by visible effect. The pair test then sub-splits the clean triggers further by inhibitability. The implicit topology is at least three-deep: anchors versus non-anchors, then trigger categories within anchors, then pathways within triggers.

Single-coordinate control is feasible without auxiliary training. Sparse-autoencoder methods [6] typically dedicate a separately trained model to extract monosemantic features. The single- d trigger reported here is reached by ablation alone, with no auxiliary model and no gradient-based feature attribution. The trigger is also concrete in a way SAE features rarely are: the controller is one integer label ($d = 758$) in the model’s native coordinate basis, not a learned vector in a separately fit dictionary.

Steering interventions need full subspace edits. The $v8$ vs $v9$ comparison says that the minimal-edit version of activation steering (override the scalar gain on the channel of interest, leave the rest) collapses output on inputs that the full-edit version handles cleanly. This is a constraint on practical activation steering: adjustments along a hidden channel need to be applied across the integrated read circuit for that index, not just at one tensor. The Coder donor’s modification of an anchor’s layer-norm gain, in the merged net, is part of a coordinated adjustment that also touches the donor’s attention query/key/value projections at the same index, the attention output projection’s column for it, the MLP gate and up projections’ rows, and the MLP down projection’s column. Reverting only the gain leaves an incompatible substrate underneath; the result is the kind of collapse $v8$ shows on `lobster - structure` and `mixed`. Reverting all of those together restores the entire read-and-write circuit back to base values and

produces the clean behavior *v9* shows. The integration constraint is not a subtle correction; it is the difference between coherent output and breakdown.

5 Limitations

Five inputs is a small evaluation panel. The category boundaries in §3.2 are clean on these five, but the trigger and inhibitor structure may not survive a wider input set with mixed lengths and topics. A natural extension is twenty-plus inputs scored with a fixed rubric.

One hybrid base. All results in this paper are on *v7b*, one specific passthrough split. Other split points were tried during the *v7*-series sweep and discarded for less clean trigger behavior; the systematic study of how passthrough split point affects single-*d* triggers is not in this paper.

One model family. The qwen2.5 family is the only basis here. Replicating in a llama-family hybrid (e.g. llama2-base merged with llama2-coder) would test whether $d = 758$, or the equivalent coordinate under cross-architecture alignment [4], plays a similar role outside qwen2.5.

Manual surface scoring. The and codeblock markers are detected by visual inspection of the first 200 tokens. The classifier is the author. String-matching automation is straightforward and not yet in this paper.

Mechanistic claim is structural, not causal. The labels “trigger,” “inhibitor,” and “pathway” refer to visible output behavior on five inputs after single-*d* edits. They do not assert circuit-level causality. A circuit-level statement would require attention-head and MLP-neuron level dissection that is on the Tier-B roadmap [1] and not in this paper.

6 Conclusion

Restoring one index at a time in a passthrough-merged qwen2.5 hybrid sorts the the eleven anchors into a 4-category functional taxonomy, identifies one of them ($d = 758$) as a clean single-*d* behavior trigger on one specific input, identifies another ($d = 715$) as a pathway-specific inhibitor that fully suppresses two of the four triggers and not the other two, and shows that the simpler scalar-gain version of the same intervention collapses output where the full-subspace version preserves it.

The 4 findings are individually small. Together they argue that the the eleven anchors have an internal sub-pathway structure that ablation alone can map, that single-*d* control by integer label is feasible without auxiliary feature dictionaries, and that practical activation-steering edits need to act on the full read pathway of a coordinate rather than on isolated scalars.

Acknowledgements

Drafted in continuation of the Mercury sprint as a paired-engineering session with an Anthropic Claude Code agent. All raw outputs and merge configs live in the companion dataset record [3]; the specific runs referenced here are `compare-v7-results.json` (*v7b* baseline), `compare-v17-results.json` ($d = 758$ alone), `compare-v8-results.json` and `compare-v9-results.json` (layer-norm vs full subspace), and `compare-{279,382,476,758}_715-results.json` (inhibitor pairs).

References

- [1] Chen, Ho Yiing. Mercury: An addressable million-cell sensor grid for browser-accessible LLM observability on consumer hardware. Zenodo, 2026. DOI: 10.5281/zenodo.20313154.
- [2] Chen, Ho Yiing. Unsupervised discovery of language and topic lanes in transformer models via multilingual co-firing signatures. Zenodo, 2026. DOI: 10.5281/zenodo.20313748.
- [3] Chen, Ho Yiing. Mercury-Viewer: A self-contained browser-based interactive 3D visualization of LLM internal sensor maps. Zenodo, 2026. DOI: 10.5281/zenodo.20313150.

- [4] Chen, Ho Yiing. Cross-Model Dimensional Preservation in qwen2.5: 49 Hot Coordinates Survive Across 3B and 7B. Zenodo, 2026. DOI: 10.5281/zenodo.20346626.
- [5] Chen, Ho Yiing. Naive SLERP Fails Both Aggressively and Conservatively on Qwen2.5 Instruct x Coder Merges, and TIES Works: An Observation-Driven Case Study. Zenodo, 2026. DOI: 10.5281/zenodo.20346628.
- [6] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Aspell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread, Anthropic, October 2023. <https://transformer-circuits.pub/2023/monosemantic-features>.
- [7] Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Steering Language Models With Activation Engineering. arXiv preprint, 2023. DOI: 10.48550/arXiv.2308.10248.
- [8] Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A.M. Steering Llama 2 via Contrastive Activation Addition. arXiv preprint, 2023. DOI: 10.48550/arXiv.2312.06681.