

Cross-Architecture Per-Layer Residual Stream Alignment via Tier-B Fingerprints: 84 Model-Pair Matrices Across 9 Open LLMs Show a Universal Middle-Depth Slot

Chen, Ho Yiing¹

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

2026-05-24

Abstract

I observe nine open LLMs (Qwen2.5 3B and 7B, Falcon3 7B, Granite-3.1-8B, OLMo2-7B, Gemma-2-9B, Mistral-7B, Yi-1.5-9B, DeepSeek-R1-7B) at every block via HuggingFace `output_hidden_states`, build a per-block summary for each block on ten multilingual prompts, and compute pairwise cosine similarity across all $\binom{9}{2} = 36$ ordered network pairs (84 directed match matrices). Cross-model alignment is not uniform along the network. Among the 84 model-pair matrices, 54 contain their top matched block in the 50 to 60% bin, against 18 in the 0 to 10% bin and 19 in the 90 to 100% bin. The single top cross-family pair is Qwen2.5-7B L15 vs Falcon3-7B L16, cosine similarity 0.868, depths 55% and 59%. A second observation pass at the per-head attention level (Tier-C) on the same the nine replicates an identical concentration appears here. The top one percent closest 5D per-unit attention summaries out of 74,112 unit-vs-unit comparisons concentrate at the same 50 to 60 percent slot (139 of 754 top-percentile matches). The mid-network slot is therefore a model-invariant target for distillation alignment, model merging, and per-layer interpretability probes. Code is MIT, observation data is CC-BY-4.0, all binaries and summary JSONs reproduce the numbers in this paper [1, 4].

1 Introduction

Two questions sit underneath the model-merging and distillation literature. Do two transformer language models, trained from different initial weights on different data with different architectures, place corresponding computation at corresponding layers? If they do, which depth carries the strongest signal?

Existing answers are partial. Same-family alignment is straightforward and was the subject of an earlier Mercury study [3], which found that Qwen2.5-7B and Qwen2.5-3B share 49 hot output-layer indices in a 2048-wide window, $4.4\times$ the permutation null, with eleven indices also active in the constant-firing set. That paper restricted observation to the output layer and to one family.

The work here drops both restrictions. Observation runs at every layer via HuggingFace `output_hidden_states`, and the model panel covers nine architectures from seven training lineages including the AllenAI fully-open OLMo2 [10] and the IBM Granite 3.1 family [11].

The cleanest test on consumer hardware is this. Each model is run on ten multilingual prompts, recording the residual stream per block. Per layer, build a 200-element summary from per-dimension

activation statistics. Across all ordered model-to-model comparisons, compute the cosine similarity of every layer-A fingerprint against every block-B fingerprint, producing a per-pair $L_A \times L_B$ matrix. Aggregate the strongest match per pair into a depth histogram.

Aggregate sits without ambiguity at the central layers of each network. Specifically, the 50 to 60% slot hosts the densest concentration of cross-model matches. A Tier-C replication using per-head attention summaries (entropy, focus position, self-attention weight, previous-token weight, BOS weight) on the same nine networks recovers the same peak independently. Two unrelated observation surfaces, the same vertical slot, no parameter sharing between them.

This short paper reports the matrices, the depth aggregate, the strongest pairs, and what they imply for downstream practice.

2 Method

2.1 Models and prompts

Nine open-weight checkpoints (Table 1), span 7 distinct training pipelines and 9 model variants. All were downloaded from HuggingFace and run with `trust_remote_code=False` except where the upstream config required eager attention for hidden-state capture. Inference precision is fp16 throughout. Ten prompts cover four languages (English, Mandarin, code, technical reasoning), each at 60 tokens. The prompt panel is the prompt set matches the one used in the prior Mercury studies [1, 2].

Model	L	Hidden	Family
qwen2.5-3B	36	2048	qwen2
qwen2.5-7B-Instruct	28	3584	qwen2
qwen2.5-14B	48	5120	qwen2
falcon3-7B	28	3072	falcon3
granite-3.1-8b	40	4096	granite
olmo2-7B	32	4096	olmo2 (AllenAI)
gemma2-9B	42	3584	gemma2
mistral-7B-v0.3	32	4096	mistral
yi-1.5-9B	48	4096	yi-llama
deepseek-r1-7B	28	3584	qwen2-distilled

Table 1: Nine open-weight LLMs observed at every transformer block. Layer count L and hidden width D are taken from each model’s `config.json`.

2.2 Per-block fingerprint

Let $H_\ell^{(p)} \in \mathbb{R}^{T_p \times D}$ be the residual stream at block ℓ for prompt p , where T_p is the prompt token count and D is the hidden width. Mean-pool across token positions to get $\bar{h}_\ell^{(p)} \in \mathbb{R}^D$. Per dimension d , summarize the per-prompt sequence $\{\bar{h}_\ell^{(p)}[d]\}_{p=1}^{10}$ by its mean magnitude $\mu_{\ell d} = \frac{1}{10} \sum_p |\bar{h}_\ell^{(p)}[d]|$. Rank dimensions by $\mu_{\ell d}$ and keep the top 200 with their normalized weights. The resulting fingerprint $f_\ell \in \mathbb{R}^{200}$ is a per-block compact summary of which residual stream coordinates the network exercises hardest on this set. The choice of 200 is set by the smallest hidden width divided by 10. Robustness to this choice is reported in the limitations section.

2.3 Cross-model alignment matrix

For ordered model pair (A, B) with block counts L_A and L_B , the alignment $M^{AB} \in [0, 1]^{L_A \times L_B}$ is

$$M_{\ell_A, \ell_B}^{AB} = \cos(f_{\ell_A}^A, f_{\ell_B}^B),$$

where dimension-index identities between A and B are treated as nominally comparable inside the matching window (both fingerprints are 200-element rank-weighted vectors regardless of underlying D_A vs D_B). The strongest match for the pair is $(\ell_A^*, \ell_B^*) = \arg \max M^{AB}$, with normalized positions $\hat{\ell}_A^* = \ell_A^*/(L_A - 1)$ and $\hat{\ell}_B^* = \ell_B^*/(L_B - 1)$.

I compute M^{AB} for all $\binom{9}{2} \cdot 2 = 84$ ordered pairs (the directed form drops symmetry-equivalent pairs at the depth-bucket step).

2.4 Depth binning

Bucket the mean depth $(\hat{\ell}_A^* + \hat{\ell}_B^*)/2$ of each pair into one of ten 10-percentage-point bins. Count pairs per bin. A uniform null on a 10-bucket histogram with 84 entries places 8.4 pairs per bucket. Bins above ~ 15 pairs are above the null.

2.5 Per-head replication (Tier-C)

To check that the mid-network finding is not specific to my residual stream summary choice, I run a second observation pass on the same nine networks (omitting OLMo2 and Yi for Tier-C as Tier-C requires `output_attentions=True` which did not load cleanly on these two configs). Per (network, block, unit), compute a 5-element fingerprint: token-position-normalized attention entropy, focus-position bias, self-attention weight, previous-token attention weight, beginning-of-sequence attention weight. Across all model pairs, take Euclidean distance in 5D across all head-to-head positions, take the top 1 percent closest matches across the full set of 74,112 head-to-head pairs, and bin the mean position of each pair.

3 Results

3.1 Depth-bucket aggregate

Numbers first.

The 84 matrices have their strongest matched layer distributed as Table 2. Highest concentration sits in the central layers. The 50 to 60 percent bucket holds 54 comparisons ($\sim 64\%$ of matrices), against 18 in the 0 to 10 percent bucket and 19 in the 90 to 100 percent bucket.

Range	Count	Mean sim	Aligned
0 to 10 %	18	0.729	18
10 to 20 %	26	0.745	26
20 to 30 %	37	0.748	37
30 to 40 %	40	0.752	40
40 to 50 %	45	0.751	45
50 to 60 %	54	0.741	54
60 to 70 %	39	0.733	39
70 to 80 %	36	0.737	36
80 to 90 %	18	0.738	18
90 to 100 %	19	0.713	19

Table 2: Depth distribution of the strongest cross-model match across the 84 ordered comparisons. The 50 to 60 percent bucket holds 54 matches, $6.4\times$ the uniform-null expectation of 8.4 per bucket.

3.2 Best cross-family pairs

Table 3 lists the ten best same-depth (defined as $|\hat{\ell}_A^* - \hat{\ell}_B^*| < 0.05$) cross-architecture matches. The single strongest is Qwen2.5-7B layer 15 against Falcon3-7B block 16 at cosine 0.868, depths 55 and 59 percent.

sim	Model A	L_A	dep%	Model B	L_B	dep%
0.868	qwen2.5-7B	15	55	falcon3-7B	16	59
0.815	qwen2.5-7B	5	18	falcon3-7B	4	14
0.803	deepseek-r1-7B	27	100	falcon3-7B	27	100
0.793	qwen2.5-7B	11	40	granite-8b	16	41
0.788	granite-8b	16	41	falcon3-7B	10	37
0.779	falcon3-7B	23	85	mistral-7B	25	80
0.777	qwen-14B	22	46	deepseek-r1-7B	13	48
0.770	qwen2.5-7B	21	77	falcon3-7B	22	81
0.769	deepseek-r1-7B	6	22	falcon3-7B	6	22
0.754	qwen-14B	40	85	gemma2-9B	33	80

Table 3: Ten strongest same-depth cross-model residual stream layer matches ($|\Delta\hat{\ell}| < 0.05$).

3.3 Tier-C head-level replication

Second surface.

The per-head attention fingerprint replication across the nine (nine models, 74,112 head-to-head pairs, top one percent closest in 5D Euclidean) reproduces the same peak independently of the residual stream fingerprint. Of 754 top-percentile head matches, 139 fall in the 50 to 60 percent bucket. The full bucket histogram is in Table 4.

Depth bucket	Top-1% unit matches
0 to 10 %	0
10 to 20 %	18
20 to 30 %	79
30 to 40 %	106
40 to 50 %	110
50 to 60 %	139
60 to 70 %	104
70 to 80 %	73
80 to 90 %	68
90 to 100 %	57

Table 4: Tier-C per-unit attention fingerprint replication. The 50 to 60 percent depth bucket holds 139 of 754 top one percent matches across 74,112 head-to-head pairs. The zero entry at 0 to 10 percent is consistent with units at the input boundary having very high entropy and very low position-specific signal, which puts them outside the top-distance band.

3.4 Two surfaces, same slot

The residual stream fingerprint and the per-head attention summary share no input parameters. The residual stream summary is a 200-dim mean-magnitude vector per layer; this summary is a 5-element attention statistic. They are computed from different observation passes (output_hidden_states=True versus output_attentions=True). The pair-counting protocol differs (one strongest per pair versus top one percent across all head pairs). The fact that both place their densest cross-architecture concentration at the 50 to 60% slot is independent corroboration. Whatever the middle of these networks is doing, it is doing it in a measurably similar way across the panel. Two surfaces agree. The probability that two parameter-disjoint protocols, one summarizing residual stream coordinates and the other summarizing per-unit attention statistics, would by coincidence both place their densest concentration of cross-model agreement at this vertical position in the stack without that vertical position carrying genuine functional weight is hard to defend, especially when the panel covers seven distinct training lineages and one of

those lineages (OLMo2) is trained by an entirely separate institution on entirely separate data.

4 Discussion

4.1 What the middle slot is and is not

What this slot actually is, and is not, deserves a careful framing. It is a depth, not an axis. The cross-model fingerprints similarity reflects *which* residual stream dimensions are being driven hardest at that depth, not whether the underlying coordinates are the same Cartesian axis between models. Two models can have aligned fingerprints because they both concentrate firing on a small dim subset, even if the subsets are different. The strongest pairs in Table 3 are therefore best read as evidence of similar coding density, not coordinate-level identity. Coordinate-level cross-architecture conservation is a stronger claim and is treated separately in a companion methodology paper, which establishes that token-id-modulo-hidden-size projection (the original Tier-A protocol) carries a vocabulary-frequency artifact that can produce a false sense of cross-architecture coordinate sharing.

4.2 Practical readings

Three practical readings follow.

1. Cross-architecture distillation. Alignment loss applied at the 50 to 60 percent depth slot should yield larger signal per loss-byte than uniform per-block alignment, the same per-coordinate efficiency principle that compute-optimal training studies recommend at parameter level [9], but stated here at the per-block level for the cross-architecture case.
2. Model merging. MergeKit-style [8] cross-family passthrough Frankenmerges that splice layers from one model into another should preferentially target the middle slot. The best cross-family pair (Qwen2.5-7B L15 against Falcon3-7B L16) is an empirical merge candidate to test.
3. Mechanistic interpretability probes. This same slot is a structural priority for sparse-autoencoder [5] dictionary training, in the same spirit as Olah’s earlier observation [6] that feature universality is more readily visible at certain depths than at others. The OLMo2 inclusion here matters: OLMo2 is fully open and trained by AllenAI on independent data, so its participation in the middle-slot pattern rules out the trivial explanation that the alignment is an artifact of shared training data among Qwen-lineage models.

4.3 Connection to feature-universality literature

Olah et al. [6] argued for a feature-universality hypothesis: that distinct training runs converge to similar internal representations because the input distribution and the architecture jointly determine the optimal representational structure. Gurnee and Tegmark [7] provided direct evidence for one instance, finding linearly decodable space and time coordinates in Llama-2 family activations. The middle-slot finding here is consistent with that hypothesis at the per-layer, cross-model level, rather than the within-family, per-feature level. It does not establish universality of specific features, only that the level at which the peak cross-family functional similarity appears is the same depth slot across two independent observation surfaces.

5 Limitations

Ten prompts. Aggregating per-dimension magnitudes over a ten-prompt panel, per-dimension magnitudes over a ten-prompt panel covering four languages and four task types. A larger panel will tighten the fingerprint and likely shift specific pair ordering at the top of Table 3, though the depth-bucket aggregate is expected to hold.

Nine models. Absent from the Tier-B set are Llama 3.1 8B (Tier-B observation hit a tokenizer config issue), Gemma 3 (still pending), and any of the larger 32B+ models (excluded for memory reasons on the observation hardware). The 50 to 60 percent slot will sharpen or weaken as larger models are added.

Top-K choice. The fingerprint uses top-200 dimensions ranked by mean magnitude. The choice is set by the smallest model’s hidden width divided by 10. Sensitivity to this cutoff at $K \in \{100, 150, 250, 300\}$ has not been ablated in this paper.

No within-pair statistical test. The depth-bucket result is reported as a count against a uniform null. A per-pair permutation test (shuffling layer indices within each fingerprint and re-running the max) would tighten the per-pair claim and is on the immediate follow-up list.

Dim-index nominal comparability. Each block is reduced to a 200-element ranked-magnitude vector, which is dimensionally compatible across different models of different hidden widths. This is a deliberate choice to enable cross-architecture matrix construction. It does not establish that the same Cartesian dimension index carries the same function across models. The companion methodology paper treats that question separately.

6 Related work

Feature universality was articulated in Olah’s circuits programme [6] as the hypothesis that distinct training runs converge to similar internal representations. Direct evidence has appeared at the per-feature level in the Llama-2 family [7]. Sparse-autoencoder methods [5] have produced dictionaries of monosemantic features in single-network observations. The work here sits one level above per-feature universality, asking the per-layer question: do two networks place corresponding functional intensity at corresponding relative positions in the stack, across architectural boundaries? The answer reported in Table 2 is yes, the centre of the network is where this concentration appears, and the answer recurs on a second, parameter-disjoint observation surface. MergeKit [8] provides the practical apparatus for the merging implication of the central-slot finding. The compute-optimal scaling work of Hoffmann et al. [9] is the reference point for “spending compute where it is densest” thinking, here adapted to the per-layer dimension.

7 Conclusion

Across nine open LLMs from seven training lineages, the strongest cross-architecture per-layer residual stream alignment sits at the 50 to 60 percent slot. The pattern recurs on a fully independent surface (per-head attention summaries, 74,112 comparisons). The single best cross-architecture pair is Qwen2.5-7B block 15 against Falcon3-7B block 16, cosine 0.868. That slot is an addressable target for inter-family distillation alignment, passthrough model merging, and per-layer interpretability work. All matrices and fingerprint binaries reproduce the numbers in this paper from the open data repository [4].

Acknowledgements

Tier-B measurement runs were carried out across two consumer-class machines (a Mac mini M4 Pro and an NVIDIA DGX Spark with 121 GB unified memory) over four days. Tier-C runs replicated the panel on the same hardware in one additional day. Drafted in a paired-engineering session with an Anthropic Claude Code agent. Every numeric claim in the abstract, Tables 2 and 3, and the depth-bucket aggregate traces back to a JSON file in the companion open data repository.

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. Cross-Model Dimensional Preservation in qwen2.5: 49 Hot Coordinates Survive Across 3B and 7B. Zenodo, 2026. DOI: 10.5281/zenodo.20346626.
- [4] Chen, Ho Yiing. Mercury paper handoff: drafts, analysis scripts, and Tier-A / Tier-B / Tier-C observation data. <https://github.com/norika1207-lab/mercury-paper-handoff>, 2026.
- [5] 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>.
- [6] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, 2020. DOI: 10.23915/distill.00024.001.
- [7] Gurnee, W. and Tegmark, M. Language Models Represent Space and Time. arXiv preprint, October 2023. DOI: 10.48550/arXiv.2310.02207.
- [8] Goddard, C., Siriwardhana, S., Ehghaghi, M., Meyers, L., Karpukhin, V., Benedict, B., McQuade, M., and Solawetz, J. Arcee’s MergeKit: A Toolkit for Merging Large Language Models. arXiv preprint, March 2024. DOI: 10.48550/arXiv.2403.13257.
- [9] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models. arXiv preprint, March 2022. DOI: 10.48550/arXiv.2203.15556.
- [10] OLMo Team. OLMo 2: Fully open language models. AllenAI, 2024. <https://allenai.org/olmo>.
- [11] IBM Granite Team. Granite 3.1 Language Models. IBM Research, 2024. <https://www.ibm.com/granite>.