

Tier-A Vocab-Hash Artifact in Cross-Architecture Hot-Dim Studies:

Why Output-Layer Token-ID Projection Manufactures False Universality, and an Artifact-Aware Screening Protocol

Chen, Ho Yiing¹

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

2026-05-24

Abstract

A widely useful observation protocol for transformer language models reads the output-layer logits, maps each generated token id to a hidden-state index via $d = \text{tokid} \bmod D$, and aggregates per-coordinate firing intensity into a cross-prompt heatmap. I call this T_A observation. Applied to a panel of 18 LLMs across 13 architectures, the T_A protocol identifies a set of eleven anchor hidden-state indices that appear in the top-50 hottest list of as many as eleven of the eighteen networks, including the AllenAI-trained OLMo2 7B at 29.8 times the hypergeometric random baseline. The natural reading is that these indices carry a shared functional load across the panel. I report a direct contradiction. On the same OLMo2 7B checkpoint, switching the observation surface from T_A to T_B (per-block residual stream via HuggingFace output_hidden_states) returns zero of the eleven T_A anchors among OLMo2’s actual top-magnitude residual stream coordinates. The genuinely hot residual stream indices for OLMo2 are $\{514, 2273, 2594, 2983, 808, 1066\}$, none of which share indices with the T_A anchor set. I diagnose the source: the $d = \text{tokid} \bmod D$ projection is a hash, and high-frequency vocab entries shared across BPE schemes will hash to overlapping indices regardless of underlying geometry of residuals. I propose a two-tier screening protocol: T_A as a cheap candidate-signal generator (good for narrowing the search space), T_B as the validation layer where index-level claims are made. The Mercury observation framework [1] is re-framed in light of this caveat. Code and data are open [5].

1 Introduction

When a transformer language model emits a token, the easiest internal signal to read is the output-layer logit vector. A natural ambition is to convert that into a per-index observation of the residuals: assign each generated token id to a coordinate d via $d = \text{tokid} \bmod D$ where D is the network hidden width, count fires per coordinate per prompt, and produce a per-prompt heatmap. The cost is low (one forward pass, no extra activations recorded), the output is addressable by integer, and the protocol scales to any LLM that exposes its vocab. I call this T_A observation, and it is the surface on which the original Mercury study [1] reported its initial 49-coordinate cross-variant overlap within the Qwen family [3].

Extending T_A to a wider model panel produced a striking quantitative claim. Eleven specific hidden-state coordinates, originally identified as triple-confirmed slots inside Qwen2.5-7B, kept appearing in the top-50 hottest dimensions of network after network. the AllenAI checkpoint, trained by AllenAI on a

fully documented and demonstrably distinct corpus, registered 4 of 11 at 29.8 times the hypergeometric random baseline. DeepSeek-R1 70B (distilled into a Llama backbone) hit 3 of 11 at 44.7 times. The natural interpretation: these eleven indices are an architecture-invariant functional anchor set, surviving across training lineages.

That interpretation turns out to be partly an artifact of how T_A maps token ids to coordinates. This paper reports the contradiction that revealed the artifact, identifies the mechanism, and proposes a refined two-tier protocol that preserves the practical value of T_A while keeping its coordinate-level claims honest.

2 The contradiction

2.1 Setup

For each candidate network, run the panel used in the Mercury observation studies [1, 2]. Apply T_A : read output-layer logits, project generated token ids via $d = \text{tokid} \bmod D$ where D is the network’s hidden width, aggregate fire counts per index, take the top-50 hottest dimensions.

Now apply T_B on the identical checkpoint: load via HuggingFace transformers with `output_hidden_states=True`, run shared prompts, capture the residual tensor at every block. Per index d , compute $\mu_d = \frac{1}{PL} \sum_{p,\ell} |\bar{h}_{p,\ell}[d]|$ where $\bar{h}_{p,\ell}[d]$ is the token-position-averaged residual value at block ℓ for prompt p . Take the top-6 μ_d as the network’s ”genuine” hot residual stream coordinates.

T_A and T_B disagree on this checkpoint completely.

2.2 OLMo2-7B as the witness case

Surface	Top hot indices	Overlap with T_A anchors
T_A (output projection)	{11, 25, 279, 382, ...}	4 of 11 (29.8× baseline)
T_B (residual stream)	{514, 2273, 2594, 2983, 808, 1066}	0 of 11

Table 1: On the same OLMo2-7B with the panel, T_A and T_B disagree on which hidden-state coordinates the network is actually firing. T_A ’s top-50 list contains 4 of the 11 original anchors; T_B ’s top-magnitude residual stream coordinates contain zero of them.

Two surfaces, this network, shared prompts, contradictory verdicts on which indices carry the firing. T_B ’s result has the stronger epistemic standing because it reads the the residuals directly. The T_A result must then have a source other than residual stream geometry. The most likely culprit is the projection itself.

3 The mechanism

3.1 T_A is a hash, not a probe

The T_A projection $d = \text{tokid} \bmod D$ is, in formal terms, a modular hash from the id space (typically $\sim 100\text{k}$ entries for modern LLMs) into the hidden-state index space (typically 2k to 8k entries). The collision structure of this hash is determined by the id distribution on the prompt panel, not by anything inside the network beyond the BPE scheme.

Networks sharing the BPE pool (Qwen2 BPE, Llama2 BPE, GPT-2 BPE descendant) will tokenize most natural-language inputs into largely overlapping high-frequency vocab sets. Those shared high-frequency tokens hash to the same indices via the modular projection regardless of whether the two networks have any internal representational structure in common. The T_A heatmap therefore confounds (i) genuine residual stream index sharing, with (ii) tokenizer aliasing.

3.2 Quantifying the hash contribution

A clean way to estimate the artifact’s contribution is to run a random-id control. Replace the actual generated ids with ids sampled uniformly at random from the tokenizer vocabulary, run the T_A projection on the random ids, and check whether the same eleven anchor coordinates appear in the top-50 hottest. If they do, the anchor result cannot be attributed to network function. If they do not, the artifact contribution is bounded below by the difference.

I have not yet completed the full random-token control on the eighteen-model panel. The OLMo2 T_A vs T_B contradiction is the qualitative evidence that the artifact contribution is non-zero. Quantifying its magnitude is the immediate follow-up.

A complementary check is the cross-tokenizer comparison. Two networks that share a BPE scheme should show stronger T_A anchor overlap than two networks with disjoint BPE schemes, holding training data constant. this checkpoint’s case is consistent with this prediction: OLMo2 uses a custom AllenAI BPE with considerable sharing with the GPT-2 / GPT-NeoX BPE descendants [8], so the high-frequency English tokens still hash to similar indices as Qwen and Llama. The T_A overlap therefore appears even though the residual stream is doing something genuinely different in that checkpoint.

4 An artifact-aware screening protocol

The corrective is not to abandon T_A but to demote it from validated finding to screening.

4.1 Two-tier protocol

1. **T_A as screening layer.** Run the cheap output-projection protocol on any candidate network. The cost is one forward pass per prompt. The output is a list of candidate hot coordinates. Treat this list as a search-narrowing tool: it identifies a subset of coordinates worth probing more carefully. Make no index-level cross-architecture claims at this stage.
2. **T_B as validation layer.** For any candidate identified by T_A , validate by capturing the residual stream directly via `output_hidden_states=True` on this panel. Compare top-magnitude residual stream coordinates against the T_A candidate list. Only coordinates that survive both tiers may carry coordinate-level cross-architecture claims.

4.2 What survives the demotion

The inter-architecture functional alignment finding survives T_A ’s demotion if it is recovered at the per-layer level by T_B , regardless of whether specific coordinate identities transfer. The companion paper [4] reports exactly this result: per-layer cosine similarity between T_B fingerprints of nine networks across seven training lineages places the densest cross-architecture alignment at the 50 to 60 percent depth slot, with the single top match (Qwen2.5-7B block 15 against Falcon3-7B block 16, cosine 0.868) replicated on a fully independent per-head attention surface. That finding does not depend on the T_A projection or on any tokenizer-induced coordinate aliasing.

4.3 What does not survive

Three earlier claims need re-framing.

First, the claim that the eleven Qwen-derived indices are architecture-invariant functional anchors is downgraded to a T_A screening signal. Index-level cross-architecture claims require T_B confirmation, which the OLMo2 case shows is not automatic.

Second, the within-family Qwen2.5 finding [3] is unaffected within Qwen2.5 because both the 7B and 3B variants share the same tokenizer and a closely related training history, so the T_A and T_B verdicts converge. The within-family overlap (49 coordinates at $4.4\times$ chance, 11 triple-confirmed at $20.6\times$ chance with hypergeometric $p < 10^{-12}$) is a genuine same-family structural finding.

Third, the cross-family T_A overlap is still informative as a BPE-sharing signal. Networks with strong T_A overlap likely share substantial vocab. This is itself useful for lineage detection (the DeepSeek-R1 70B case suggests distillation preserves at least BPE family even when the backbone changes), but it is not stream functional alignment.

5 Discussion

5.1 Why the artifact slipped past the within-family case

Within a single BPE descent, both T_A coordinate identities and T_B residual stream coordinate identities tend to track each other, because the residual stream is presumably reading from BPE-aligned positions for that family. The original within-family Qwen2.5 study [3] therefore got the right answer by both surfaces and there was no contradiction to spot.

Extending to cross-BPE architectures (OLMo2 with its custom BPE, Llama with its custom tokenizer) was the test that should have been done at protocol design time. The OLMo2 case (4/11 T_A , 0/11 T_B) is the empirical signal that the protocol must be split into screening and finding layers.

5.2 Connection to the broader interpretability literature

The split between screening and finding is similar in spirit to the separation between feature-attribution methods and feature-confirmation methods in mechanistic interpretability [6, 7]. Saliency maps, for instance, are useful for narrowing the search space for which model components carry which information, but a saliency hit is not by itself proof that the component carries the information. Sparse-autoencoder analysis or causal intervention is the confirmation step. The T_A vs T_B distinction here is the same shape, applied to one specific observation protocol family.

5.3 Practical advice for Mercury users

For anyone using the Mercury framework or building on its dataset [5]:

1. T_A heat summaries are good for fast within-family screening and for tokenizer-overlap-style lineage detection. Cross-BPE coordinate-level claims need T_B .
2. T_B observation outputs (‘tier-b-data/heat-summary-***-tierB.json’ in the open data repository) are the surface for any per-coordinate or per-layer cross-architecture functional claim.
3. T_C observation outputs (per-head attention pattern fingerprints) are independent of both T_A and T_B coordinate identities, so they can serve as a third confirmation surface for layer-level findings.

6 Limitations

One witness case. The T_A versus T_B contradiction is documented in full on OLMo2-7B. The broader eighteen-model panel has T_A data for all models but T_B data for only nine. Completing the T_B panel will either reproduce or refine the artifact picture across the remaining architectures.

Random-token control pending. The cleanest direct quantification of the hash-induced contribution to T_A overlap is the random-token control described above. It is on the immediate follow-up list, not in this paper.

BPE-sharing quantification. I assert that tokenizer-overlap drives the T_A artifact but have not yet computed the per-pair BPE-sharing metric (e.g. fraction of top-1k BPE merges shared between any two tokenizers in the panel) and correlated it with T_A anchor overlap. That correlation would close the diagnostic argument.

Protocol naming. The “ T_A / T_B / T_C ” naming is internal to the Mercury observation framework. Adopting more standard terminology (e.g. “logit-projection probe” / “hidden-state probe” / “attention-pattern probe”) would help cross-paper reading.

7 Conclusion

The T_A protocol’s apparent inter-architecture index-level universality is partly a tokenizer-induced hash artifact. OLMo2-7B’s contradiction (4 of 11 anchor coordinates in T_A , 0 of 11 in T_B on the same checkpoint with the same prompts) is the empirical evidence. The corrective is a two-tier protocol: T_A as a screening layer that narrows the search space, T_B as the finding layer that supports coordinate-level claims. The cross-architecture functional alignment finding survives this demotion at the per-layer level, as reported in the companion paper [4]. Open data and code are at the Mercury paper handoff repository [5].

Acknowledgements

The OLMo2-7B contradiction was spotted during a T_B sweep on the consumer-grade observation hardware used throughout the Mercury project. Drafted in a paired-engineering session with an Anthropic Claude Code agent. The T_A versus T_B numbers in Section 2 trace directly to JSON files in the 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. Cross-Architecture Per-Layer Residual Stream Alignment via T_B Fingerprints: 84 Model-Pair Matrices Across 9 Open LLMs Show a Universal Middle-Depth Slot. Companion preprint, 2026. Source and data: <https://github.com/norika1207-lab/mercury-paper-handoff>.
- [5] Chen, Ho Yiing. Mercury paper handoff: drafts, analysis scripts, and $T_A / T_B / T_C$ observation data. <https://github.com/norika1207-lab/mercury-paper-handoff>, 2026.
- [6] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askeel, 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] 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.
- [8] OLMo Team. OLMo 2: Fully open language models. AllenAI, 2024. <https://allenai.org/olmo>.