

Mercury Cross-Model Anchor Families

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at
charenix.com

Draft: June 18, 2026

Abstract

Mercury City shows many language models as separate towers, but the towers are not isolated. Similar functional regions appear across architectures: embedding and detokenization near the bottom, surface syntax in early and middle layers, structural abstraction in the middle, semantic routing higher up, and output control near the end. This paper defines a cross-model anchor family as a set of measured sites from different models that share function, normalized depth, module role, and prompt-trace behavior.

The contribution is not the claim that models are identical. They are not. The claim is narrower and more useful: a repeated functional address can be tested across models by depth normalization, module-role agreement, activation coherence, and prompt-path reach. I give equations for anchor matching, family compactness, cross-model recurrence, and false-family penalty. This makes the “AI family tree” visible without turning it into mythology.

Author Note

When I looked at the towers together, the strange thing was not that they were different. The strange thing was how often the same kind of place appeared again. GPT, Qwen, Gemma, Phi, Mistral, OPT, BLOOM, and the others did not give me the same city, but they kept giving me neighborhoods that rhymed.

That rhyme is the object of this paper.

I do not want this to become mystical. The family must survive measurement.

1 From Tower To Family

A single tower can be beautiful and still be anecdotal. A family needs a stronger rule. The same label in two models is not enough; a layer name, a human nickname, or a visual impression can fool the eye.

Mercury therefore treats a family as a measured relation. Sites must agree in where they sit, what module they touch, how they respond, and where the traced prompt can travel after touching them.

Visual resemblance opens the door. It does not close the proof.

2 Normalized Position

Let a measured site be

$$s = (m, \ell, u, d, \rho),$$

where m is a model, ℓ is a layer, u is a module role, d is a dimension or channel, and ρ is the assigned functional region. For a model with L_m layers, normalized depth is

$$z(s) = \frac{\ell}{L_m - 1}.$$

Two sites may have very different raw layer numbers and still occupy similar relative positions:

$$\Delta_z(s_i, s_j) = |z(s_i) - z(s_j)|.$$

This is why comparing L7 in a 36-layer model to L7 in an 80-layer model is too crude. The atlas needs normalized depth.

3 Anchor Similarity

Define module agreement:

$$M(s_i, s_j) = \mathbf{1}[u_i = u_j].$$

Define region agreement:

$$R(s_i, s_j) = \mathbf{1}[\rho_i = \rho_j].$$

Let $a_s(x)$ be the activation or importance response of site s over a shared prompt set X . The response similarity is

$$C(s_i, s_j) = \frac{a_{s_i}(X) \cdot a_{s_j}(X)}{\|a_{s_i}(X)\| \|a_{s_j}(X)\| + \epsilon}.$$

The Mercury anchor similarity is

$$\text{sim}(s_i, s_j) = \alpha e^{-\Delta_z(s_i, s_j)/\tau} + \beta M(s_i, s_j) + \gamma R(s_i, s_j) + \delta C(s_i, s_j).$$

The weights are not magic; they are knobs to publish, freeze, and audit.

4 Family Compactness

An anchor family F is a set of sites drawn from at least two models. Its compactness is

$$K(F) = \frac{2}{|F|(|F| - 1)} \sum_{i < j} \text{sim}(s_i, s_j).$$

Its model diversity is

$$D(F) = |\{m(s) : s \in F\}|.$$

Strong families have high $K(F)$ and high $D(F)$. Weak ones have one of them but not both.

This is the protection against cherry-picking.

5 Prompt-Path Agreement

Mercury also traces prompt movement. Let $P_s(x)$ be the downstream path set activated after prompt x touches site s . Define path agreement:

$$T(s_i, s_j) = \frac{1}{|X|} \sum_{x \in X} \frac{|P_{s_i}(x) \cap P_{s_j}(x)|}{|P_{s_i}(x) \cup P_{s_j}(x)| + \epsilon}.$$

This term matters because two sites can have similar labels but send the prompt into different downstream neighborhoods. A family should agree in motion, not only in position.

6 False-Family Penalty

The atlas should be strict. Define a penalty for families whose members share a name but not enough evidence:

$$\Phi(F) = \mathbf{1}[K(F) < \kappa] + \mathbf{1}[D(F) < 2] + \mathbf{1}[\bar{T}(F) < \eta].$$

The family is publishable only if

$$\text{valid}(F) = \mathbf{1}[\Phi(F) = 0].$$

This stops the city from becoming astrology. Any proposed group that cannot survive the equations remains a hypothesis.

7 Relation To Representation Similarity

Representation comparison has a real history. SVCCA compares learned representations across networks [1]; CKA gives a stable way to compare representation similarity [2]; probing work studies how linguistic structure appears inside models [3]. Mechanistic interpretability then asks for components and circuits rather than only vector similarity [4].

Mercury borrows the seriousness of those comparisons but changes the object. The object here is not only a representation matrix. It is a visible, addressed, traced city site that can be compared across towers.

8 What The City Adds

The city adds body, path, and warning. The group has a body because the site has a location. It has a path because prompt traces can be followed. It has a warning label because false families can be marked instead of quietly promoted.

That is why screenshots matter here. The picture is not decoration when each point has an address and each road can be checked.

The image earns its place only when it can be audited.

9 Limits

Cross-model anchors do not prove that the same circuit exists in every model. They do not prove that a model learned the same algorithm. The family is a candidate structure, and it needs intervention before becoming a causal claim.

The value is still real. A candidate family tells us where to cut first, where to compare first, and where a surprising recurrence may be hiding.

10 Conclusion

Mercury Cross-Model Anchor Families turns the city from a gallery into a comparative instrument. A tower is one model. A family is the recurring address that survives across towers. If the family is compact, diverse, and traceable, it deserves a paper. If it is not, the city should say so.

References

- [1] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. *Advances in Neural Information Processing Systems*, 2017. <https://arxiv.org/abs/1706.05806>
- [2] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. *Proceedings of the 36th International Conference on Machine Learning*, 2019. <https://arxiv.org/abs/1905.00414>
- [3] John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. *Proceedings of NAACL-HLT*, 2019. <https://aclanthology.org/N19-1419/>
- [4] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. *Transformer Circuits*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>