

Mercury Cross-Model Highways: Navigating Shared Functional Addresses Across Language Models

Chen, Ho Yiing (norika)
Independent Researcher, Taiwan
ORCID: 0009-0006-6816-9891

June 2026

Abstract

Language networks are usually compared by benchmark score, parameter count, or representation similarity. The city adds another object: a navigable road between opened towers. In the Hope of City view, towers are not isolated monuments. They are connected by cross-model anchors: measured addresses that appear to serve comparable functions in different networks. This paper defines a road as a tuple linking system, layer, dimension, function label, and local relation proof. The proposition is modest but important. This map does not prove that two neurons in different systems are identical organs. It gives a method for turning a suspected correspondence into an inspectable object: the reader can see where it lives, what region it belongs to, what neighboring signals it touches, and whether the same function appears in another tower. This turns cross-model comparison from a table into a map.

Author Note

This fourth paper in the series is not a museum label for a pretty scene. The city matters because it lets a person ask a hard question: when GPT, Qwen, Gemma, Mistral, OPT, BLOOM, Phi, and DeepSeek are opened, do they only look different, or do they build similar roads inside?

1 The Problem With Flat Comparison

Most comparison is flat. A table says system A wins on one benchmark and system B wins on another. A similarity score says two layers are close. Those views are useful, but they do not feel like anatomy. They do not tell a reader where to enter the system.

That is the part that bothered me.

Hope of City begins with a different intuition. If many LLMs are opened as towers, then comparison should happen through places, not only numbers. A shared function should have an address. A suspected relation should have a line. A family resemblance should be visible enough that it can be challenged.

Flat comparison also wastes the most human part of the work. After spending weeks opening networks, the researcher does not remember them as rows. One tower has a bright surface region. Another has a strange output-side knot. Another keeps a family of syntax-like anchors at a similar height. These memories are not rigorous by themselves, but they are not useless

either. Those memories are the first smell of structure. This project tries to turn that smell into an object that can be checked: a road with endpoints, local neighborhoods, function names, and proof levels.

2 Highway Object

Let a site in model m be

$$a = (m, \ell, d, r, c),$$

where ℓ is the layer or stratum, d is the hidden dimension, and (r, c) is the residual-grid address used by the atlas. A local function record is

$$q(a) = (z(a), \gamma(a), N_k(a)),$$

where $z(a)$ is the functional region label, $\gamma(a)$ is a working function name, and $N_k(a)$ is the top-neighbor relation set.

A cross-model road is then:

$$H = (a_i, a_j, \alpha, \eta),$$

where a_i and a_j come from different towers, α is an anchor score, and η records the proof type: region match, label match, neighbor overlap, prompt co-activation, or intervention agreement.

The point of H is not to declare sameness. The point is to give the comparison a handle.

A handle is small, but it changes the work. Without it, the assertion stays in a spreadsheet. With it, the assertion becomes a place in the city.

That is the turn.

3 Region Alignment

The atlas uses coarse functional regions as the first alignment surface:

$$z(a) \in \{\text{embed, surface, structural, semantic, output}\}.$$

These names are not decorative. They are the city zoning map. A surface syntax site in Qwen can be compared with a surface syntax site in Gemma or Mistral. A semantic pathway in one tower can be checked against a semantic pathway in another.

The alignment score can be written as

$$\alpha(a_i, a_j) = \lambda_1 \mathbf{1}[z(a_i) = z(a_j)] + \lambda_2 s_\gamma(a_i, a_j) + \lambda_3 s_N(a_i, a_j) + \lambda_4 s_x(a_i, a_j),$$

where s_γ measures working-label agreement, s_N measures local neighbor similarity, and s_x measures shared response under a prompt set X . The weights λ are not universal constants. They are an audit policy.

To keep the policy honest, normalize the proof weights:

$$\sum_{q=1}^4 \lambda_q = 1, \quad \lambda_q \geq 0.$$

Then $\alpha \in [0, 1]$ when each component score is bounded. A road is accepted into the strong

visual layer only if

$$\alpha(a_i, a_j) \geq \tau_H \quad \text{and} \quad z(a_i) = z(a_j).$$

This prevents a single attractive neighbor overlap from drawing a fake road between unrelated functional regions.

4 Why Visual Roads Matter

Representation similarity methods such as SVCCA and CKA compare neural spaces with strong mathematical tools [1, 2]. This work does not replace them. It answers a different need. A scalar similarity score can say that two layers resemble each other, but it does not give the researcher where to click.

Circuit work already argues that transformer parts should be treated as composable mechanisms rather than anonymous parameters [3]. Mercury pushes that idea into a spatial comparison across towers.

The road view gives a route through the proof. If Qwen and Gemma share a surface-syntax zone, the map can display the towers, the candidate anchors, and the local neighborhood around each address. If the relation is weak, the road can be faded or removed. If it survives more prompts and interventions, it becomes a stronger road.

5 Family Resemblance Without Myth

The dangerous assertion would be: all towers have the same neurons. The map does not need that assertion. The stronger and safer statement is this: different networks may reuse comparable functional arrangements, and those arrangements can be mapped as addresses plus proof.

That matters because it changes the product of interpretability. The output is not only an explanation for one system. It becomes a transferable atlas. A researcher can ask whether an inhibitor-like function in Qwen has a cousin in Mistral, whether a style pathway appears in Gemma, or whether an output-control region is absent in a smaller GPT-family tower.

6 Evidence Ladder

A road should climb an evidence ladder:

1. region correspondence;
2. function-name correspondence;
3. local neighbor similarity;
4. prompt co-activation;
5. intervention agreement.

Only the last step begins to support causal language. The earlier steps are useful, but they should stay visually marked as weaker.

Let the proof vector for a road be

$$e(H) = (e_r, e_\gamma, e_N, e_x, e_I),$$

where the entries correspond to region, label, neighborhood, prompt response, and intervention proof. A conservative confidence score is

$$C(H) = \min(e_r, e_\gamma, e_N, e_x) \cdot (1 + \beta e_I), \quad 0 \leq \beta \leq 1.$$

The minimum is deliberate. A road with one missing non-causal leg should not look strong just because another leg is bright. Intervention proof can increase confidence, but it cannot rescue a road whose visible endpoints are not aligned.

The falsification score is the opposite:

$$\Phi(H) = \mathbf{1}[z(a_i) \neq z(a_j)] + \mathbf{1}[J(N_k(a_i), N_k(a_j)) < \tau_N] + \mathbf{1}[\text{corr}_{x \in X}(\rho(a_i, x), \rho(a_j, x)) < \tau_x].$$

If $\Phi(H) > 0$, the road must be weakly drawn or deleted until better proof arrives.

7 Limits

The road view can be wrong. Role names can be too loose. Region labels can hide important differences. Smaller networks may compress several functions into one area. Larger networks may split one function into several addresses. A beautiful road can be a false road.

So the road must remain humble.

And removable.

For that reason, the system keeps road proof inspectable. A cross-model line must lead back to addresses, neighbors, prompt responses, and tests. Otherwise it is only a line in a pretty city.

8 What A Reviewer Should See

A reviewer should not read this paper as an argument that visualization is enough. That would be too weak. The point is that the visualization changes what can be asked next. Once the cross-model relation has endpoints, the reviewer can demand more proof at those endpoints: inspect the local neighborhood, inspect the prompt set, test whether the candidate survives a small intervention, test whether a nearby address gives the same result, check whether the same functional zone appears in a third tower, and record the failure case where the road disappears. This is the important move. A benchmark table usually ends the discussion. A road map starts the next experiment. It lets a human walk into the comparison, touch one suspected correspondence, and decide whether it deserves a stronger test or should be deleted from the map.

Delete bad roads.

9 Conclusion

Mercury Cross-Model Highways turns network comparison into a navigable object. The contribution is not an assertion that every tower has the same organs. The contribution is a way to ask the question without losing the proof. A road between two LLMs should be clickable, testable, and removable. That is how a city becomes an instrument.

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] 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 Thread, 2021. <https://transformer-circuits.pub/2021/framework/index.html>