

Mercury Functional Regions Across 38 Language Models

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

Hope of City shows 38 opened language-model structures as vertical bodies. This paper asks what those bodies have in common. Mercury assigns each visible band or site to one of five working zones: embed/detokenization, surface, structural, semantic, and output. The claim is not that every model has the same internal organs. The claim is that many architectures can be compared through recurring functional zones once their floors are drawn as addressable strata. I define occupancy, transition sharpness, cross-model stability, and boundary uncertainty as audit equations for this comparison. The result is an anatomy paper for Mercury: the city is not only a skyline. It has repeating organs, and those organs can be measured, challenged, and redrawn.

Author Note

This paper exists because one city screenshot is not enough. When 38 stacks are opened, the first shock is visual. The second question is anatomical: why do some bands feel like input gates, some like syntax surfaces, some like structure engines, some like semantic roads, and some like output machinery? That feeling has to become a measurement object. Otherwise it stays only a beautiful intuition.

The city deserves better than intuition.

1 Why Regions Matter

A transformer depth index is a number. A functional zone is a hypothesis about work.

Mercury uses five zone names because a stack without zoning becomes too easy to admire and too hard to test. The public city view needs enough structure for humans to enter it. The archive needs enough restraint that a name can be removed when the evidence fails.

The five-region vocabulary is:

$$Z = \{\text{embed, surface, structural, semantic, output}\}.$$

These are working names.

They are not final brain organs. Their job is to let the same question be asked across Qwen, Gemma, OPT, BLOOM, Pythia, Phi, Mistral, DeepSeek, and other observed structures: where does the model appear to enter, shape, organize, mean, and emit?

2 Region Assignment

Let m be a model body and $L_m = \{0, \dots, n_m - 1\}$ its band set. A Mercury zoning assignment is a map

$$z_m : L_m \rightarrow Z.$$

For a band ℓ , the assignment is supported by visible modules, observed hot sites, local neighborhood structure, and prompt-trace behavior:

$$z_m(\ell) = \arg \max_{z \in Z} [\alpha s_{\text{module}}(\ell, z) + \beta s_{\text{hot}}(\ell, z) + \gamma s_{\text{neighbor}}(\ell, z) + \delta s_{\text{trace}}(\ell, z)],$$

with non-negative weights and

$$\alpha + \beta + \gamma + \delta = 1.$$

This formula is intentionally plain. A zone mark should not be magic. It should say which evidence source pushed a band into a named body part.

3 Occupancy And Shape

The first audit number is occupancy:

$$O_m(z) = \frac{1}{|L_m|} \sum_{\ell \in L_m} \mathbf{1}[z_m(\ell) = z].$$

If $O_m(\text{semantic})$ is large in one body but small in another, the viewer has a precise question instead of a mood. Is the second system compressed? Does it use fewer bands for semantic work? Did Mercury over-name the first structure? The number does not answer everything, but it prevents the map from staying vague.

The second number is transition sharpness. Let

$$T_m = \{\ell : z_m(\ell) \neq z_m(\ell - 1), \ell > 0\}$$

be the boundary set. A stack with few zone changes is smooth. A body with many changes is mixed or unstable. Define

$$B_m = \frac{|T_m|}{|L_m| - 1}.$$

The zoning entropy is

$$H_m = - \sum_{z \in Z} O_m(z) \log(O_m(z) + \epsilon).$$

Together B_m and H_m separate two cases that look similar in a screenshot: one long structural block, or many small alternating bands.

4 Cross-Model Stability

The anatomy claim needs cross-model discipline. For a zone z , define normalized depth position

$$u = \frac{\ell}{|L_m| - 1} \in [0, 1].$$

Let $P_m(z, u)$ be the smoothed probability that model m assigns position u to zone z . The cross-model stability of z is

$$S(z) = 1 - \frac{1}{|M|} \sum_{m \in M} \int_0^1 |P_m(z, u) - \bar{P}(z, u)| du,$$

where

$$\bar{P}(z, u) = \frac{1}{|M|} \sum_{m \in M} P_m(z, u).$$

High $S(z)$ means the zone appears at comparable relative depths across the observed structures. Low $S(z)$ means the mark is either architecture specific, task specific, or too weakly defined.

This is where the count matters. A single model can always seduce the researcher.

Thirty-eight opened structures make the mark earn its place.

5 Boundary Uncertainty

The most dangerous part of this paper is not naming a zone. It is pretending that a boundary is sharp when the evidence is soft. Mercury should therefore store boundary uncertainty:

$$U_m(\ell) = 1 - (p_1(\ell) - p_2(\ell)),$$

where $p_1(\ell)$ and $p_2(\ell)$ are the largest and second-largest zone assignment probabilities at band ℓ . A confident band has a large gap. A contested band has a small gap.

The city should draw contested bands differently. A clean border can glow. A soft border should shimmer or remain marked as provisional. This is not an interface detail. It is honesty.

Soft borders are not failure. They are warnings.

6 Relation To Existing Tools

BertViz made transformer attention visible across layers and heads [1]. TransformerLens and CircuitsVis made hook-based mechanistic work more practical [2, 3]. Activation Atlas showed that a learned system can be explored as a spatial feature map [4]. Similarity methods such as SVCCA and CKA compare learned representations with scalar and matrix tools [5, 6].

Mercury uses those works as neighbors, not as the destination. The zoning view asks for a different object: a city-scale anatomical zoning across many opened LLMs, where every named band can still point back to depth, module, dimension, trace, or local relation evidence.

7 Falsification

Every zone name should be easy to attack. For a probe set X_z meant to stress zone z , define a response score:

$$R_m(z; X_z) = \frac{1}{|L_z|} \sum_{\ell: z_m(\ell)=z} \mathbb{E}_{x \in X_z} [\rho_m(\ell, x)].$$

If a surface band does not respond to surface probes, or a semantic zone does not respond to meaning probes, the name weakens. The falsification loss is

$$F_m(z) = \max(0, \tau_z - R_m(z; X_z)).$$

The name survives only if it keeps passing this pressure. That is the difference between a drawn body part and a research claim.

8 What This Paper Adds

This paper gives Mercury a middle scale. Hope of City shows the skyline. Layer Neuron Atlas shows a single address. Functional zones sit between those extremes. They let a reader ask whether models share a body plan before descending into a specific neuron or head.

The immediate value is navigational. The deeper value is comparative. If a zone appears across many stacks, it becomes a candidate anatomical feature. If it appears only in one family, it becomes an architecture signature. If it vanishes under better probes, it should be deleted.

Delete weak zones.

9 Limits

This paper does not prove that five zones are the final ontology of LLMs. It does not claim that every band performs only one function. It also does not merge visual release count, traceable result count, and formal fullgrid count. The 38 structures are the visual atlas release. The measurement ledger must remain separate.

The stronger claim is narrower: a zoning vocabulary can turn a skyline into an auditable anatomy if each named part keeps its evidence, uncertainty, and deletion rule.

10 Conclusion

Mercury Functional Regions turns the city from a spectacle into an anatomy question. The stacks are beautiful, but the paper does not ask anyone to believe the glow. It asks whether recurring districts can be measured across 38 opened LLMs. If the answer survives probes, the city has a body plan. If it fails, the weak zone is removed and the map gets sharper.

References

- [1] Jesse Vig. A Multiscale Visualization of Attention in the Transformer Model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019. <https://arxiv.org/abs/1906.05714>
- [2] Neel Nanda and Joseph Bloom. TransformerLens. 2022. <https://transformerlensorg.github.io/TransformerLens/content/citation.html>
- [3] Alan Cooney and Neel Nanda. CircuitsVis. 2023. <https://github.com/TransformerLensOrg/CircuitsVis>
- [4] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation Atlas. *Distill*, 2019. doi:10.23915/distill.00015. <https://distill.pub/2019/activation-atlas/>
- [5] 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>
- [6] 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>