

# Mercury Layer Neuron Atlas: Visualizing Cross-Layer Transmission in Qwen 2.5 3B

Chen, Ho Yiing (norika)<sup>1</sup>

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

Draft: June 18, 2026

## Abstract

Transformer internals are often inspected one hook, head, or activation table at a time. That view is useful, but it misses a physical question: if a dimension is active here, where can its influence travel next? This paper introduces the Mercury Layer Neuron Atlas, a layer-by-layer transmission map of Qwen 2.5 3B. The current view contains 73,728 displayed neuron coordinates across 36 strata and 716,800 top-neighbor connections. One opened address such as L17.D715 opens a local contract: residual position, named role, forward cascade, backward cascade, outgoing edges, and incoming edges. The claim is not that a glowing point is automatically a concept. The claim is sharper: when a model is treated as addressable strata, then local roles and cross-layer paths can be inspected in the same visual address system. The result is closer to circuit cartography than to a static heatmap.

## Author Note

This paper is the second door into Mercury. The first paper, *Hope of City*, shows the skyline of many LLMs. This one enters a single tower and asks what happens when the vertical body is drawn floor by floor. I keep the language direct because the question is direct: a model is not only a table of weights. It has places where signals gather, split, suppress, and rejoin.

## 1 Why This Paper Exists

The city view is overwhelming in the right way. It tells the reader that many LLMs have been opened and mapped. A second view is needed for another question: what is one tower made of?

For Qwen 2.5 3B, Mercury draws 36 strata. Each stratum is a plane. Each plane carries neuron sites. Lines move upward and downward through the tower. Some lines are quiet scaffolding. Others form visible cascades. Addresses can be opened like objects, not like anonymous dots.

The reason is practical. When a site is called an inhibitor, a controller, or a pathway marker, the reader should be able to click it and ask:

- where is it located?
- what does it connect to in the next few hops?
- what feeds it from below?
- what role label was assigned?
- how much of the label is evidence, and how much is a working name?

That is the difference between decoration and measurement.

## 2 What Is Being Drawn

The layer-neuron view is not a picture of every parameter. It is a focused transmission surface.

That distinction matters.

Qwen 2.5 3B contains far more weights than any screen can honestly show at once. Mercury therefore draws a contract: visible coordinates, chosen neighbors, named anchors, and cascade summaries. The screen is allowed to be partial. Vagueness is not allowed. Every site needs an address. Every line must correspond to a measured relation. Every role name appears, it must remain open to inspection instead of becoming a label pasted on top of the image.

This is why the view feels dense without becoming only noise. The planes carry regular structure. The vertical paths carry input movement. The orange mesh shows local relation. The opened panel gives the focused site a contract.

## 3 Related Work

Mechanistic interpretability has long treated transformers as circuits rather than opaque parameter blocks. The Transformer Circuits line of work made this language especially explicit by describing residual streams, attention heads, and MLPs as composable computational objects [1]. TransformerLens and CircuitsVis provide practical tooling for hooks and visualization [2, 3]. BertViz shows attention at multiple scales [4]. Activation Atlas showed that spatial maps can make learned features easier to inspect [5].

Mercury Layer Neuron Atlas sits beside these works but asks for a different surface. The starting unit is not a notebook cell. It is a vertical structure the reader can fly through. The display is not limited to attention. It shows chosen dimensions, measured neighbors, named roles, residual positions, and cascades in one geometry.

## 4 Address Model

Each visible coordinate is addressed as

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

where  $m$  is the network,  $\ell$  is the stratum,  $d$  is the hidden dimension, and  $(r, c)$  is the residual-grid position used by the interface. For Qwen 2.5 3B,

$$\ell \in \{0, \dots, 35\}, \quad d \in \{0, \dots, 2047\}.$$

The map stores a directed weighted relation

$$e_{i \rightarrow j} = (a_i, a_j, w_{ij}, \sigma_{ij}),$$

where  $w_{ij}$  is the observed connection score and  $\sigma_{ij}$  records the sign or class used for drawing. For rendering, Mercury keeps the top neighbors per address:

$$N_k(a_i) = \text{topk}_{a_j} |w_{ij}|.$$

The current Qwen 2.5 3B view uses  $k = 10$ , producing 716,800 displayed connections over 73,728 visible sites.

The top-neighbor rule is a drawing rule, not a claim that all other relations are absent.

## 5 Cascade View

A chosen address opens a cascade. The forward cascade is the set of sites reachable within  $h$  hops:

$$F_h(a) = \{b : \exists \text{ directed path } a \rightarrow b \text{ with length } \leq h\}.$$

The backward cascade is:

$$B_h(a) = \{b : \exists \text{ directed path } b \rightarrow a \text{ with length } \leq h\}.$$

For the opened address `L17.D715`, the interface reports:

$$\begin{aligned}\ell &= 17/36, \\ d &= 715/2048, \\ (r, c) &= (11, 11), \\ |F_5| &= 312, \\ |B_5| &= 212.\end{aligned}$$

The working label is *INHIBITOR: selective suppression*. The label does not close the case. It opens the case. A reader can inspect outgoing positive and negative edges, then compare the visible path with other marked points.

The cascade also gives a numerical sanity check. Define cascade asymmetry:

$$A_h(a) = \frac{|F_h(a)| - |B_h(a)|}{|F_h(a)| + |B_h(a)| + \epsilon}.$$

For `L17.D715`, using  $h = 5$ ,

$$A_5(\text{L17.D715}) = \frac{312 - 212}{312 + 212} \approx 0.191.$$

That value says the site is not only receiving history; it has more forward reach than backward support in the displayed graph. The number does not prove the inhibitor name. It tells the reader which direction the next intervention should attack first.

## 6 Named Points

The current Qwen view exposes several named anchors:

Table 1: Examples of named addresses in the Qwen 2.5 3B layer-neuron view.

| Address               | Region     | Working role                          |
|-----------------------|------------|---------------------------------------|
| <code>L17.D715</code> | structural | inhibitor / selective suppression     |
| <code>L21.D481</code> | structural | style-destab / destabilization marker |
| <code>L25.D279</code> | semantic   | Pathway A / style transfer            |
| <code>L29.D715</code> | semantic   | inhibitor / selective suppression     |

These labels are not presented as final ontology. They are handles. A handle must be useful enough to return to, compare, and falsify.

## 7 Input Path

The most important visual difference is that the input is not outside the diagram. It enters from below and moves upward through the layered body. Colored vertical paths mark chosen transmission routes. Dense orange lines show local neighbor relations. Bright points show active or named sites.

This matters because many interpretability views split the evidence into separate windows: one for tokens, one for activations, one for edges, one for notes. Mercury tries to hold them in a single scene. The reader can see the input path, the nearby active sites, and the opened address contract at the same time.

This is why the view is visually aggressive. The vertical body is not trying to look minimal. It is trying to keep the transmission story visible while the reader is still inside the geometry.

## 8 What This View Adds

This view adds three things.

First, it gives dimensions physical addresses. `L17.D715` is not a row in a file. It has a place in the tower.

Second, it makes local transmission visible. Forward and backward cascades are not only numbers; they are paths through adjacent strata.

Third, it lets role names remain provisional. An inhibitor label can be useful without pretending to be final. The role has to survive inspection through edges, cascades, ablations, and task probes.

## 9 Verification Equations

For each opened address  $a$ , Mercury separates three scores: visibility, local balance, and intervention promise.

Visibility is the fraction of the displayed local relation that remains addressable:

$$V_k(a) = \frac{|\{b \in N_k(a) : \text{addr}(b) \text{ exists}\}|}{k}.$$

If  $V_k(a) < 1$ , the viewer is looking at an incomplete object and the panel must say so.

Local sign balance is

$$B_k(a) = \frac{\sum_{b \in N_k(a)} \max(w_{ab}, 0) - \sum_{b \in N_k(a)} \max(-w_{ab}, 0)}{\sum_{b \in N_k(a)} |w_{ab}| + \epsilon}.$$

A strongly negative  $B_k(a)$  supports a suppressive hypothesis; a strongly positive value supports a reinforcing hypothesis. A near-zero value warns that the role name may be too dramatic.

Finally, define the intervention promise of a named site:

$$I(a; Y) = \mathbb{E}_{y \in Y} [L(y; \theta_{\setminus a}) - L(y; \theta)],$$

where  $Y$  is a task or phrase set and  $\theta_{\setminus a}$  is a masked or ablated proxy. The atlas is strongest when a visually named address also has non-trivial  $I(a; Y)$ . If  $I$  vanishes, the name stays a visual handle, not a causal result.

## 10 Limits

This paper does not claim that every named site has a clean natural-language concept. Nor does it claim causality from visualization alone. The map is an inspection surface. Causal claims still require intervention.

The current view also uses a top-neighbor display, so it is not the full graph. This is a rendering choice. The full measurement archive must remain separate from the screen projection.

Another limit is naming. Words such as inhibitor or Pathway A are useful only when they invite tests. They should be treated as handles for repeated inspection, not as final names for discovered organs.

## 11 Conclusion

Mercury Layer Neuron Atlas turns one LLM into a layered machine that can be entered at the level of addressed dimensions. The value is not only that the picture is beautiful. The value is that an address can be opened, named, traced forward, traced backward, and checked against its neighbors. This is the beginning of a usable circuit map.

## References

- [1] 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>
- [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] 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>
- [5] 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/>