

Mercury Pathways: Drawing a User Prompt Through an Opened Language Model

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

June 2026

Abstract

Most interpretability interfaces place the submitted phrase in one panel and the neural evidence somewhere else. Mercury treats that separation as a failure of the instrument. In the atlas, the submitted phrase is drawn as a stream that enters from the lower strata of an opened model, crosses measured neuron sites, passes through named anchors, and reaches the output side. This paper describes the stream model used in the Qwen 2.5 3B atlas. The drawn line is not a decorative beam. It is an addressable chain over a measured relation graph: sites have layer and dimension addresses, edges have signed weights, and local forward or backward cascades can be opened from any marked site. The contribution is a visual and mathematical contract for utterance-level circuit inspection: a submitted phrase should be followed through the model without losing the layer, address, neighbor, and role evidence that made the line readable in the first place.

Author Note

This third Mercury paper follows *Hope of City*, which shows the skyline of opened LLMs. *Layer Neuron Atlas* enters one tower and gives every visible site an address. This paper follows the thing written into the box: the request. If that phrase disappears into a black box, then the map has failed. In this view, the utterance becomes a track.

1 Why User Flow Matters

A model can look impressive even when it explains nothing. A heatmap can glow. An attention diagram can braid itself into a beautiful mess. A node graph can look scientific while hiding the only question that matters: what happened to my input?

The system starts from that question. The written phrase is not treated as a caption outside the system. It is projected into the same space as layers, dimensions, named anchors, and transmission edges. The gold stream in the city view and the vertical beams in the layer view are not ornaments. They are a demand: when the submitted phrase activates a pathway, the pathway should be inspectable.

This changes the burden of proof. The interface cannot only say that a model has an inhibitor, a controller, or a style line. It has to show whether the written phrase crosses those places. It also has to show what was nearby, what was suppressed, and what continued upward.

2 Flow Graph

Let a drawn neural site be

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

where m is the model, ℓ is the layer or stratum, d is the hidden dimension, and (r, c) is the residual-grid position used by the atlas. The measured relation graph is

$$G = (V, E, W, S),$$

where V is the set of displayed sites, $E \subseteq V \times V$ is the set of displayed directed relations, $W : E \rightarrow \mathbb{R}$ gives the observed edge score, and $S : E \rightarrow \{-1, +1\}$ records the sign used for drawing.

A submitted phrase x induces an active subset

$$A_x = \{a \in V : \rho(a, x) \geq \tau\},$$

where $\rho(a, x)$ is the measured response score for site a under the query symbol x , and τ is the visibility threshold. The visible stream is not every possible edge. It is the constrained chain family

$$T_x = \{p = (a_0, \dots, a_n) : a_i \in A_x, (a_i, a_{i+1}) \in E, \ell(a_{i+1}) \geq \ell(a_i)\}.$$

The monotone layer condition is a rendering rule for the submitted phrase. It keeps the main track readable while local backward cascades remain available in the detail panel.

3 Local Contracts

User Flow only works if a drawn track can be audited. Mercury therefore uses local contracts. When the reader opens L17.D715, the panel does not only display a name. It reports layer, dimension, residual position, forward cascade, backward cascade, and top incoming or outgoing relations.

For Qwen 2.5 3B, one opened site has:

$$\begin{aligned} a &= \text{L17.D715}, \\ \ell &= 17/36, \\ d &= 715/2048, \\ |F_5(a)| &= 312, \\ |B_5(a)| &= 212. \end{aligned}$$

The working role is *INHIBITOR: selective suppression*. The important point is not the word inhibitor by itself. The important part is that the word is attached to an address, and the address can be checked against the drawn line.

4 Path Score

To compare utterance pathways, Mercury can assign a simple chain score:

$$\text{score}(p, x) = \left(\prod_{i=0}^{n-1} |W(a_i, a_{i+1})| \right)^{1/n} \cdot \frac{1}{n+1} \sum_{i=0}^n \rho(a_i, x).$$

The first term rewards strong measured relations. The second term rewards utterance relevance. A chain with strong edges but no response to the submitted phrase should not dominate the view. A bright site with no measured continuation should not pretend to be a circuit.

Signed behavior can be tracked separately:

$$\text{sign}(p) = \prod_{i=0}^{n-1} S(a_i, a_{i+1}).$$

This lets the map distinguish reinforcing chains from suppressive arcs without collapsing both into one color.

The selected path also needs stability. Given paraphrase or spelling variants $\mathcal{X}(x)$, define

$$\text{stab}(p, x) = \frac{1}{|\mathcal{X}(x)|} \sum_{x' \in \mathcal{X}(x)} J(\{a : a \in p\}, A_{x'}),$$

where J is Jaccard overlap. A pathway that only appears for one fragile wording is not a pathway yet. It is a candidate.

Layer continuity is checked by

$$\text{cont}(p) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}[\ell(a_{i+1}) - \ell(a_i) \in \{0, 1, 2\}].$$

This keeps a visually impressive jump from pretending to be a normal route. Long jumps may still matter, but they must be marked as jumps.

5 Why This Is Not BertViz

BertViz made attention easier to see [1]. TransformerLens and CircuitsVis made hooks and circuit tools easier to use [2, 3]. The Transformer Circuits program gave a language for residual streams and compositional mechanisms [4]. Activation Atlas showed that spatial views can make learned features inspectable [5].

Mercury Pathways is aimed at a different gap. It does not stop at attention weights or feature clusters. It follows the submitted phrase through a layer-addressed body. The same screen can show the active track, the opened address, the nearby active sites, and the cascade counts. This is why the pictures feel crowded. The crowd is evidence kept in view.

6 Ablation and Falsification

The stream only matters if it can be challenged. Mercury therefore treats every named track as a hypothesis:

$$H = (x, p, R),$$

where x is the phrase, p is the displayed track, and R is the proposed role such as inhibitor, controller, or style arc. A basic falsification test is:

$$\Delta_y(H) = L(y \mid x; \theta) - L(y \mid x; \theta_{\setminus p}),$$

where L is a task loss or perplexity term and $\theta_{\setminus p}$ means the model state after suppressing or ablating the track proxy. If the line is visually dramatic but $\Delta_y(H)$ is near zero across phrase variants, the name should be weakened or removed.

For a set of probes Y , the expected path effect is

$$E(H) = \mathbb{E}_{y \in Y} [\Delta_y(H)], \quad R(H) = \frac{E(H)}{\text{score}(p, x) + \epsilon}.$$

$E(H)$ asks whether removing the path changes behavior. $R(H)$ asks whether the visual strength was honest about that change. A large score with small effect is a beautiful false lead. A modest score with large effect is a hidden road the interface should promote.

That is the discipline Mercury needs. The map can be beautiful, but beauty is not enough. The address must survive tests.

7 Limits

User Flow is not a complete causal proof. The displayed line is a projection from a larger measurement archive. Top-neighbor drawing, thresholds, and screen geometry can hide weaker relations. The view also depends on the quality of the response score $\rho(a, x)$. If that score is noisy, the displayed line can mislead.

These limits are not excuses. They define the next measurements: stronger phrase-response scoring, intervention tests, cross-utterance stability, and cross-model comparison of the same role address.

8 Conclusion

Mercury Pathways turns a submitted phrase into a visible line through an opened LLM. The value is simple and hard: the phrase can be followed, clicked, challenged, and compared. A map that cannot follow the submitted phrase is only scenery. A map that can follow it becomes an instrument.

References

- [1] Jesse Vig. BertViz: A tool for visualizing multi-head self-attention in the Transformer model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019. <https://aclanthology.org/P19-3007/>

- [2] Neel Nanda, Joseph Bloom, and contributors. TransformerLens. GitHub repository, 2022–2026. <https://github.com/TransformerLensOrg/TransformerLens>
- [3] Alan Cooney and Neel Nanda. CircuitsVis. GitHub repository, 2023. <https://github.com/TransformerLensOrg/CircuitsVis>
- [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 Thread, 2021. <https://transformer-circuits.pub/2021/framework/index.html>
- [5] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation Atlas. Distill, 2019. <https://distill.pub/2019/activation-atlas/>