

Browser-Resident Visualization Suite for Transformer Internals: Three Self-Contained HTML Tools Extending the Mercury-Viewer Pattern

Chen, Ho Yiing¹

¹Independent researcher, charenix.com, ORCID: 0009-0006-6816-9891
abbychen981008@gmail.com

2026-05-21

Abstract

Mercury-Viewer [1] established a design pattern for transformer interpretability tooling: a single self-contained HTML file, no install, no server, no API key, that opens in any browser and renders the inside of a transformer language model as an interactive 3D point cloud. This artifact paper releases three follow-on artifacts that apply the same pattern to three different questions. `overlap.html` shows the 49-coordinate cross-model overlap and the 11-coordinate triple-confirmed subset from a recent scaling analysis [2]. `frankenstein-v1.html` shows a SLERP merge between two qwen2.5 family members as three side-by-side cones with rose-gold lines marking the 11 protected anchors. `frankenstein-arena.html` renders a single inference pass as cargo orbs traveling on cables between functional rooms. All three files are open under MIT [1], all three open in any current browser without dependencies, and each is designed to be used by independent researchers who do not have access to a GPU cluster or proprietary interpretability infrastructure.

1 Introduction

The dominant aesthetic of mechanistic interpretability visualization is the scatter plot of monosemantic feature scores in a static publication figure [4]. The state-of-the-art tools in this space are powerful but typically demand a Python environment, the Jupyter ecosystem, often a pinned PyTorch version, sometimes an API key, and assume the reader has access to the original model checkpoint.

This is a high floor.

An independent researcher with one consumer-grade graphics card, a browser, and no institutional credentials cannot run those tools out of the box. The visual outputs from those tools, when shared, are flat images: the reader sees what the author thought was worth showing, not what the reader wants to look at.

Mercury-Viewer [1] took the opposite position. One HTML file. No install. No API. Drag to rotate, scroll to zoom, hover to inspect any of one million sensor cells. Readers control the camera. Readers pick the hover target. The artifact is 2 MB and works offline once downloaded.

This short paper releases three extensions of that pattern, each targeting a different interpretability question, each shipped as one self-contained HTML file.

2 The suite

2.1 overlap.html —cross-model dimensional overlap

Question: which hidden-state coordinates are simultaneously hot in qwen2.5:7B and qwen2.5:3B, and which subset of those is also constantly firing in 7B?

Visual: a cylindrical 3D point cloud where every visible point is a coordinate from the cross-model overlap set $|H_{7B} \cap H_{3B}| = 49$, colored by membership in the constant-firing subset $|T| = 11$ [2]. Hovering a point shows the integer coordinate index, the heat values in 7B and 3B, and the firing pattern across the ten observation prompts.

File: overlap.html, ≈ 2 MB, single file, opens at charenix.com/neuron-mercury/overlap.html.

Why it earns space in this paper: the 49+11 finding is a list of integers. A list of integers in a paper table is dead. The same list as a rotatable 3D point cloud, with each integer locatable by hover, is alive. Readers can ask, “which of the eleven anchors live in the lowest layers?” and answer it in a second.

2.2 frankenstein-v1.html —SLERP merge surgery view

Question: when SLERP merges two qwen2.5 family members at each block, what does the resulting weight grid look like next to the donors, and where do the 11 triple-confirmed anchor coordinates from overlap.html sit inside it?

Visual: three side-by-side cones. Left cone (blue) is qwen2.5-7B-Instruct, right cone (red) is qwen2.5-Coder-7B-Instruct, middle cone (purple) is the v1 merge. Each cone has 28 blocks stacked vertically; each block is a ring of dim-indexed cells colored by quantile bucket. The middle cone is colored not by its own activation but by the SLERP t value at that layer, so the reader can see the merge profile at a glance. Eleven rose-gold vertical lines run through all three cones, marking the triple-confirmed anchor coordinates. Pointer hover reveals layer, dim, quantile, and the corresponding t .

File: frankenstein-v1.html, ≈ 18 KB (graph data lazy-loaded), at charenix.com/neuron-mercury/frankenstein-v1.html.

Why it earns space: the failure-mode report on this merge [3] documents persona drift, repetition loop, script drift, factual regression on a Python claim, and register collapse. Looking at those failures next to a static description of the merge configuration tells you what happened. Looking at the same failures with the cones of the donors and the merge visible, with the rose-gold anchors highlighted, lets you see *where in the network* the lossy SLERP interpolation is most aggressive. The middle high- t band is where the donors disagree most, and where the failure modes locate themselves.

2.3 frankenstein-arena.html —inference as cargo on cables

Question: can one frame of an inference pass be drawn as a single static picture that an interpretability newcomer can read in 30 seconds?

Visual: a frozen snapshot of one prompt mid-flight. Eight cube-shaped rooms float in 3D, each a functional cluster (Chinese detector, English detector, universal core, physics reasoning, code reasoning, algorithm reasoning, input, exit). Cables connect rooms; they are attention edges between functional clusters. Cargo orbs travel along active cables; they are activation vectors moving between rooms. For a Chinese physics prompt, the active path lights up: INPUT $\rightarrow$ Chinese detector $\rightarrow$ universal core $\rightarrow$ physics reasoning $\rightarrow$ EXIT. English detector and code reasoning rooms are dimmed (not triggered). Hovering a room, cable, or cargo orb shows its identity and any data attached.

File: frankenstein-arena.html, ≈ 33 KB, at charenix.com/neuron-mercury/frankenstein-arena.html.

Why it earns space: the metaphor (cargo on cables) borrows visual vocabulary from a class of games and films the target audience already knows. The audience for an interpretability poster at a venue with mixed seniority is partly junior researchers and curious onlookers, not just senior mechanistic-interpretability specialists. A picture that uses the cargo-on-cables metaphor reads in seconds, where the same information as a per-block activation heatmap reads in minutes.

3 Design pattern

Six shared properties bind the suite together, all inherited from Mercury-Viewer [1]:

1. Single HTML file, no build step, no `npm install`, no Python environment.
2. No API key, no remote service call, fully offline once the file is downloaded.
3. No GPU dependency; the `three.js` [5] renderer runs on integrated graphics.
4. Every visible artifact is hoverable; the reader picks the inspection target.
5. Every numeric claim in the panel ties back to a coordinate, an integer index, or a raw value loaded from the embedded data block.
6. The file is under 50 KB on the structural side (`frankenstein-v1.html` and `frankenstein-arena.html`) or under 2.5 MB when the embedded data block dominates (`overlap.html`, which carries the full cell coordinate list for the cross-model overlap).

Together these properties make the tools reproducible by readers without any environment setup, and citable by integer coordinate. Either property alone would be a usability win for independent researchers. The combination is the actual design contribution, because it changes the unit of citation from a paper figure caption to a coordinate that the reader can verify, and because it changes the unit of comparison across papers from “did you make the same plot” to “did you find the same integer.” That second shift matters more than the first in a sub-field where two competing interpretability methods can both fit the same scatter plot to opposite stories, and the only way to adjudicate is to ask each method for the integer indices it claims as monosemantic and check directly whether those integers are the ones the other method also flagged.

This is the bet.

Either property alone would not justify a paper.

4 Limitations

User study is absent. None of the four artifacts has been evaluated in a user study. A planned next step is a 20–30 participant comparison with conventional interpretability visualization workflows, measuring time-to-insight on a fixed question (“find a cell that fires in 8 of 10 prompts”). That study is not run here, and any claim about pedagogical value in this paper is informal.

Single author. The tools were written for self-use during the Mercury sprint and then released. They have not been hardened against the kinds of edge cases a wider user base would surface. Browser support has been tested on recent Chromium and Safari only.

Narrow domain. All four artifacts cover qwen2.5 family observations specifically. Generalization to other transformer families requires adapting the embedded data block to the new (`layer`, `dim`, `quantile`) shape; the renderer code is shape-agnostic but the data binding is not.

Static snapshots. `frankenstein-arena.html` freezes one inference frame. Animating the same scene across many tokens of a single generation is on the roadmap; it is not in `v1`.

5 Conclusion

Three tools, three pages of source each, three HTML files. Together with Mercury-Viewer [1] they form a four-tool suite that runs entirely in a browser, requires no install or API key, and lets readers control the inspection target through direct hover and drag. The intended audience is independent interpretability researchers, and the artifact is shipped under MIT so any of the three can be modified, hosted elsewhere, or used as a template for a tool aimed at a different transformer family.

Acknowledgements

Drafted in continuation of the Mercury sprint as a paired-engineering session with an Anthropic Claude Code agent. The suite lives at charenix.com/neuron-mercury/ and the source HTML is included in the companion dataset record [1].

References

- [1] Chen, Ho Yiing. Mercury-Viewer: A self-contained browser-based interactive 3D visualization of LLM internal sensor maps. Zenodo, 2026. DOI: 10.5281/zenodo.20313150.
- [2] Chen, Ho Yiing. Cross-model dimensional preservation in qwen2.5: 49 hot coordinates survive across 3B and 7B. In preparation, Mercury Paper A, 2026.
- [3] Chen, Ho Yiing. Observation-driven SLERP merge of qwen2.5 variants fails on a five-prompt probe: A v1 failure-mode report. In preparation, Mercury Paper D, 2026.
- [4] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Aspell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread, Anthropic, October 2023. <https://transformer-circuits.pub/2023/monosemantic-features>.
- [5] three.js contributors. three.js: JavaScript 3D library. <https://github.com/mrdoob/three.js>, 2010–present.