

Hope of City: A Prompt-Trace Neural Atlas of 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 17, 2026

Abstract

A table can say that an LLM has layers. It does not make anyone feel that the machine has an inside. Hope of City is the first public Mercury map: a 3D, input-trace view of 38 observed language-model structures. One vertical structure corresponds to one LLM. Floors indicate layers. Colored bands mark recurring functional zones. Bright nodes change meaning with scale: observed hot points, repository objects, module objects, or attention heads. Arcs mark dependencies, anchors, or highways. The golden flow on the ground is not decoration; it is input entering the scene and lighting the stack as it moves through the network.

The result is a renderable map, not one fixed screenshot. The same measurement archive can become a 38-structure overview. It can also become a single-network cutaway, repository graph, module address book, or attention-head inspection panel. This paper reports the system and its measurement contract. The research claim that makes the image more than a demo is this: a large-scale map can turn network internals from isolated tensors into a shared navigable structure with addresses that reach from repository structure down to layer and head detail.

Author Note

Hope of City was developed from observations before I positioned it against existing transformer visualization systems. The visual grammar did not start as a copy of an attention viewer or a neuron browser. The grammar came from a more stubborn need: I wanted to see many language networks at once, not as rows in a spreadsheet, but as internal places that could be entered, compared, and remembered.

1 Why This Exists

LLMs are usually shown as boxes, tables, or curves. That is not wrong. It is too small.

When an LLM fails, users see a sentence. Engineers see logs. Researchers see activations, attention heads, and probes. The parts are real, but they often remain scattered. A person can inspect one layer, one head, one neuron, one token, or one prompt. What is missing is the feeling of standing inside the whole structure.

The project starts from a direct idea: if a language network has an internal geography, make the geography visible. Do not flatten it into a spreadsheet too early. Make the stack. Let the user see where the layers sit, where the hot points gather, where inputs flow, which repository objects feed the stack, and where one architecture begins to resemble another.

The current public release contains 38 observed LLM structures. The recovered measurement archive also contains a larger measurement ledger: 51 traceable LLM or task results, with 34 formal fullgrid passes under the current completion gate. Those are not the same count. The 38-structure release is the visual release. The 51/34 ledger is the measurement archive that gives Hope its audit trail.

The 38 structures do not need to sit in one crowded screenshot to be real. A map is useful because it can change scale. If the question is public orientation, the overview is enough. If the question is evidence, the user can enter one stack and descend into files, modules, dimensions, heads, and prompt trace. This same archive can redraw the picture when a different question is asked.

The work has a name: *Hope of City*. I use the name deliberately. The project is technical, but the interface is not meant to feel like a lab shelf. The desired feeling is the first time the black box becomes a place.

2 Contributions

This paper makes four contributions.

1. This work introduces *Hope of City*, a multi-LLM map containing 38 public visual structures derived from observations.
2. The paper defines a continuous address system from overview to LLM stack, floor, module, repository object, dimension, and attention-head panel.
3. The paper formalizes input flow as a traceable path through the map rather than treating input text as text outside the visualization.
4. The paper separates visual release count from measurement-ledger count, so the interface can be spectacular without inflating experimental claims.

The fourth contribution is not cosmetic. A system like this can easily become theatre if it forgets what was actually measured. The ledger keeps the theatre and the ledger in contact.

3 Related Work

Transformer visualization has a serious history. BertViz made attention visible at multiple scales and demonstrated how model, head, and neuron views can help users inspect transformer behavior [1]. Activation Atlases showed that a learned network can be explored through a spatial map of features rather than only through single units [2]. Public neuron galleries made individual units linkable and memorable. TransformerLens and CircuitsVis provide practical tools for mechanistic interpretability and visualization workflows [3, 4]. AttentionViz introduced a global query-key attention view across many inputs [5].

This project sits near those works but does not duplicate them. The difference is scale and shape. Attention, neurons, hooks, repository objects, layers, and input paths become parts of one measured map. Many LLMs share the same visual vocabulary: structures, floors, regions, hot points, anchors, highways, and input flow.

That difference matters. A single-network tool can answer, “what happened in this floor?” Hope asks another question: “where does this kind of event live across architectures?”

This distinction should be stated plainly. Existing systems can be excellent instruments, including attention viewers, neuron galleries, hook workbenches, and feature maps. They show important local structure. They do not give the reader a place where dozens of LLMs can be walked as one connected internal world. Hope is built for that larger question. Parts are not merely drawn. Many interiors stay in the same world, and the user can move between skyline, source graph, input path, module address, and attention-head contract.



Figure 1: Hope of City landing view. Each vertical structure is an observed language model. The public page describes 38 LLM atlases, five functional regions, and cross-architecture anchors. The map metaphor is not only aesthetic; it is the interaction scheme used to walk from LLM-level comparison down to floor, module, dimension, and input-trace evidence.

4 Design Pressure

The interface was not designed from a blank page. It was forced by four pressures.

The first pressure was comparison. One network is interesting; dozens of networks become a geography problem. If Qwen, Gemma, OPT, BLOOM, Pythia, and GPT-like systems all have internal structure, then the interface needs a way to stand far enough back to compare them.

The second pressure was addressability. A beautiful point is useless if it cannot be traced back to layer, module, dimension, and score. Hope uses beauty as the door, not as the proof.

The third pressure was input motion. Static maps are good for inventory. They are weaker for asking what a given text does. The golden stream is the input: it gives the user a way to see the stack react, floor by floor.

The fourth pressure was memory. A row in a CSV is easy to lose. A tower is not. Once a person has seen a Qwen stack or a Gemma stack, the network starts to have shape in the mind.

5 Map Model

Let a stack be

$$T_m = (L_m, H_m, D_m, R_m, P_m, A_m, K_m),$$

where m is an LLM, L_m is its ordered stratum set, H_m is metadata such as attention-head count, D_m is hidden width, R_m is the assignment of strata to functional regions, P_m is the set of observed hot points, A_m is the set of cross-architecture anchors or highways, and K_m is the code or repository graph that makes the stack auditable.

A point in the stack is addressed as

$$p = (m, \ell, u, d, s),$$

Table 1: Where Hope of City sits relative to nearby visualization systems.

System	Main object	Typical scale	Difference from Hope
BertViz	attention	model / layer / head	attention-centered, not a multi-LLM world map
Neuron galleries	neuron features	linkable feature collections	public unit browsing, but not LLM input flow or multi-LLM atlas projection
TransformerLens / CircuitsVis	hooks and circuits	notebook / component inspection	powerful workbench, not a public multi-LLM atlas
AttentionViz	query-key attention geometry	global attention patterns	attention embedding view, not tower / layer / input-flow cartography
Hope	LLM internals	38 LLM structures	multi-LLM neural world with prompt trace, repo graph, module addresses, head panels, and cross-model anchors

where ℓ is a floor, u is a module or unit family, d is a channel or dimension, and s is a score or heat value derived from observation. Mercury does not ask the reader to trust a glow. Each glow must point back to an address.

There is a second address space:

$$k = (r, f, o, \ell, u),$$

where r is a repository, f is a file, o is an object or role, ℓ is a floor, and u is a module entry such as `input_layernorm`, `self_attn/q_proj`, `mlp/down_proj`, or an attention-head view. This same Qwen 2.5 3B map can show 39 repositories, 911 files, 1406 edges, 396 cross-repository dependencies, and stratum-local module entries in one interface. Activations can be drawn. The harder claim is that the drawing can remain tied to the machine’s build structure.

The five displayed regions are defined as

$$\mathcal{R} = \{\text{embed-detok, surface, structural, semantic, output-decoding}\}.$$

For LLM m , a stratum-region map is

$$\rho_m : L_m \rightarrow \mathcal{R}.$$

The map is intentionally simple. It gives the reader a first coordinate system before asking them to inspect individual floors or modules.

6 From Skyline to Grain Scale

The map has two responsibilities that normally fight each other. It must be large enough to show many LLMs at once, and it must be sharp enough that one point is not reduced to a decorative dot. The interface handles this by treating each view as a projection of the same measured object.

At skyline scale, a stack answers:

what kind of LLM is this, and where does it sit among the others?

At grain scale, the same tower answers:

which file, module, stratum, head, or dimension does this visible object represent?

The Qwen 2.5 3B map shows why this matters. One view presents the system as a 3.09B parameter tower with full structural context. A second view exposes a repository graph with 39 repositories, 911 files, and 1406 edges. A third view shows cross-repository dependencies, including 396 cross-repository links. A fourth view enters a single floor object such as:

L7 / surface / syntactic features / 77.1M.

That object can then open direct module entries:

L7/input_layernorm, L7/self_attn/q_proj, L7/mlp/down_proj.

This is the line between a visualization and a map. A visualization can show the viewer a shape. An atlas must let the viewer ask where the shape came from.

The rule also applies to attention. In a Qwen 2.5 3B attention-head panel, L21 Attention Head 0 is not shown as a generic head icon. The panel records `head_dim = 128`, the query slice, the shared K/V slice, and the grouped-query-attention relation:

$$16 \text{ } Q\text{-heads} : 2 \text{ } KV\text{-heads} \Rightarrow 8 \text{ } Q\text{-heads per } KV\text{-head}.$$

In other words, the view can zoom from a 38-LLM skyline to an individual attention-head contract without changing language.

7 Projection, Not Poster

A weak reading of Hope of City would be: “this is a beautiful rendering of LLMs.” That reading misses the system.

This system is not a poster. It projects a measured archive. Let M be the set of observed models and E the evidence graph containing layer measurements, code links, module objects, input-conditioned scores, and cross-architecture anchors. A view can be written as

$$V_j = \pi_j(M, E, \Theta_j),$$

where Θ_j is a view-specific parameter set: camera, scale, filter, selected LLM, selected input, selected floor, or selected evidence class.

This means that the map can be redrawn at several levels without changing the underlying claim:

$$\begin{aligned} \pi_{\text{city}} &: (M, E) \rightarrow \text{38-structure overview}, \\ \pi_{\text{tower}} &: (M, E, m) \rightarrow \text{single-LLM stack}, \\ \pi_{\text{repo}} &: (M, E, m) \rightarrow \text{repository and file graph}, \\ \pi_{\text{prompt}} &: (M, E, m, x) \rightarrow \text{input-conditioned trace}, \\ \pi_{\text{head}} &: (M, E, m, \ell, h) \rightarrow \text{attention-head panel}. \end{aligned}$$

The paper shows only a small number of figures because a paper page is small. The archive is larger than the page. When the evidence question requires all 38 structures in one skyline, the map can render that. A narrower query can ask for one Qwen stratum, one `q_proj`, or one K/V head. The contribution is the continuity between those views.

8 Input Trace

The golden flow on the ground is the input text. Here, input is not just text placed before the model. Mercury treats it as a path through the scene.

For an input x , Mercury represents a trace as

$$\Gamma_m(x) = \{(\ell, u, d, h_{\ell,u,d}(x)) : \ell \in L_m, u \in U_\ell, d \in D_{\ell,u}\},$$

where $h_{\ell,u,d}(x)$ is an input-conditioned heat score. A displayed node is selected when

$$h_{\ell,u,d}(x) \geq \tau_{\ell,u,d},$$

with τ supplied by the measurement protocol. In the fullgrid pipeline, this threshold is channel-local rather than global, because quiet channels should not disappear simply because another channel is loud.

This trace view changes the user’s question. Without it, the user asks, “what does this network contain?” With it, the user asks, “where did my input go?” That is a different kind of interpretability.

9 Layer Neuron Map

The skyline is only one projection. A neuron map can draw the same network as a stratum-by-stratum structure. In the Qwen 2.5 3B view, the interface shows 73,728 neurons across 36 strata and 716,800 displayed connections under a top-10-neighbor view. The user can jump directly to an address such as

L17.D715.

That address opens a panel with stratum index, dimension index, residual position, cascade counts, and named role. One observed example is:

address = L17.D715,
position = row 11, col 11,
forward cascade = 312 edges over 5 hops,
backward cascade = 212 edges over 5 hops,
role = INHIBITOR: selective suppression.

Other named points in the same map include L21.D481, a structural style-destabilization marker, and L25.D279, a semantic Pathway A style-transfer point. These names are not final biological truths. They are working labels attached to measured addresses, outgoing connections, incoming connections, and stratum position.

This neuron map is more detailed than the city overview. It draws every visible stratum plane, the cross-stratum transmission lines, the input path from lower strata toward output, and the local fan-in/fan-out around selected dimensions. It therefore supports a different paper from Hope of City: *Layer Neuron Map*. The first paper explains how the whole archive becomes a place. The neuron-atlas paper should explain how one stack becomes a measured circuit surface.

10 Cross-Model Anchors

The overview also shows arcs between structures. These arcs represent cross-model anchors or highways: candidate locations where different architectures show related internal structure under observation.

For two architectures a and b , define observed point sets P_a and P_b . A candidate anchor is a pair (p_i, p_j) with

$$p_i \in P_a, \quad p_j \in P_b, \quad S(p_i, p_j) \geq \lambda,$$

where S is a similarity or alignment score and λ is a reporting threshold. This paper does not claim that every arc is a semantic identity. It claims that the interface gives those candidates a place to live, be clicked, be checked, and be rejected when needed.

That last part is important. A good atlas is not a shrine. A good atlas is a workbench.

11 Measurement Contract

The interface is useful only if it does not pretend that missing data is evidence. For that reason The ledger keeps a measurement ledger separate from the visual count. At the current archive checkpoint:

The completion gate for a formal fullgrid pass is:

$$G_i = \mathbf{1}[A_i \geq 10^8] \cdot \mathbf{1}[C_i \geq 90],$$

where A_i is the number of materialized cells and C_i is the observed-cell coverage percentage. Runs below the line may still teach us what failed. They do not get counted as completed maps.

Table 2: Recovered Mercury counts used to separate visual release scope from formal measurement scope.

Quantity	Count
Public visual releases in Hope	38
Traceable LLM/task results in recovered ledger	51
Formal fullgrid passes under current gate	34
Meta rows before deduplication	60

12 What Hope Shows

Hope gives three signals at once.

First, LLMs with very different sizes can share a recognizable vertical grammar. The stack changes height, width, and density, but the five-region story can still be read.

Second, important points become addresses. A user can move from the overview to a stack, then to a floor, then to a specific module or dimension. The point is no longer “some activation somewhere.” It becomes something like $L5.D1481$ with a region label, heat value, and model context.

Third, input flow becomes visible. That trace makes Hope more than a gallery. A prompt can enter the system, light an LLM, and leave a trace that can be compared across structures.

There is a fourth signal: local grain. A single network can be decomposed into hundreds of visible project objects and dependencies before the user even enters the neural stack view. The Qwen 2.5 3B map exposes repository clusters and layer projects. It also marks category colors and critical objects. Broken objects, cross-repository dependencies, and direct module entries remain visible. This matters because interpretability work often loses the bridge between source structure and internal structure. Hope keeps that bridge visible.

13 Mathematical Verification

The map is allowed to be beautiful, but the beauty is not the test. The test is whether a visible object still has a measured address. For a rendered object v , define an audit predicate

$$\chi(v) = \mathbf{1}[\exists p = (m, \ell, u, d, s) \text{ or } k = (r, f, o, \ell, u) \text{ such that } v \mapsto p \text{ or } v \mapsto k].$$

For a view V , the address-retention score is

$$Q(V) = \frac{1}{|V|} \sum_{v \in V} \chi(v).$$

If $Q(V)$ falls, the scene is becoming theatre. If $Q(V)$ stays high, the viewer can move from image to evidence.

The same rule applies to the golden input stream. Let $A_\ell(x)$ be the visible active set at layer ℓ for input x , and let E be the measured relation set. A trace is continuous only when adjacent layers keep at least one measured bridge:

$$C(x) = \frac{1}{|L|-1} \sum_{\ell=0}^{|L|-2} \mathbf{1}[\exists a \in A_\ell(x), b \in A_{\ell+1}(x) : (a, b) \in E].$$

This is the reason the input is drawn as flow rather than decoration. The stream has to pay rent layer by layer.

A cross-model road is also scored before it deserves visual weight. For candidate endpoints a_i, a_j , use

$$S(a_i, a_j) = \lambda_r \mathbf{1}[z(a_i) = z(a_j)] + \lambda_n J(N_k(a_i), N_k(a_j)) + \lambda_x \text{corr}_{x \in X}(\rho(a_i, x), \rho(a_j, x)),$$

where z is the region label, J is Jaccard overlap over local neighbors, and $\rho(a, x)$ is the prompt response score. A road is drawn strongly only when S clears a stated threshold. A road that fails should fade. This is how the city stays honest.

14 Limits

This paper is careful about what the map proves.

Highlighted points are not automatically human-clean semantic units. Causal necessity still needs ablation, patching, or task-specific probes. It also does not say that all 51 recovered LLM/task results are equal in status. The release, head-only runs, partial probes, failed fullgrids, and formal fullgrid passes must stay in separate columns.

The narrower claim is stronger: a large-scale interface can preserve the addressability of internal measurements while giving humans a way to navigate many LLMs at once.

15 Why This Matters

A familiar phrase is “open the black box.” Hope makes that phrase less lazy. Opening the box is not enough. Once it is open, someone has to decide how the inside can be walked, searched, compared, and remembered.

Mercury proposes one answer: treat each LLM as a stack, input flow as a path, hot points as addresses, and cross-architecture similarity as roads that must be audited rather than believed on sight.

That is the research value of Hope of City. The image is beautiful, but the beauty is not the evidence. The evidence is that the image remains connected to measurement addresses.

16 Paper Series

This paper is the front door. It should not carry the whole measurement archive on its back. The archive is large enough to support a series. The clean split follows.

1. **Hope of City:** the large-scale atlas system and visual grammar.
2. **Input Trace:** how input texts travel through layers, modules, and model regions.
3. **Layer Neuron Map:** how one stack can be expanded into stratum planes, addressed dimensions, and measured transmission cascades.
4. **Cross-Model Highways:** how candidate anchors are selected, displayed, and audited across architecture families.
5. **Five Functional Regions:** whether embed/detok, surface, structural, semantic, and output regions remain stable across models.
6. **Repository-to-Neuron Addressability:** how source graphs, module paths, and neural addresses can live in one coordinate system.
7. **Attention-Head Cartography:** head-only and K/V-sharing views as a separate evidence class.
8. **Coverage Failure:** why sparse, sampled, or coarse observation can miss large parts of the cell space.
9. **Per-Channel Quantile Fullgrid:** the measurement method that prevents quiet channels from being erased by global thresholds.
10. **Measurement Ledger:** how to separate visual releases, partial probes, failed runs, and formal fullgrid passes without inflating claims.
11. **Task Surgery:** using measured structure for task-preserving compression and local deployment.

12. **Tiny Domain Intelligence Packs:** using small measured packs for practical local document intelligence.
13. **OCR-to-Structure Attribution:** where document intelligence fails when OCR and layout are separated from structuring models.

Those papers should not be merged into one oversized manuscript. This is not one trick. The measurement program has multiple doors into the same world.

17 Conclusion

Hope is not a picture placed on top of Mercury. The paper shows what the work looks like when the measurements are given a place to stand. The map does not make an LLM simple. It makes the inside available enough that a person can begin to ask better questions.

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] 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/>
- [3] Neel Nanda and Joseph Bloom. TransformerLens. 2022. <https://transformerlensorg.github.io/TransformerLens/content/citation.html>
- [4] Alan Cooney and Neel Nanda. CircuitsVis: Mechanistic Interpretability Visualizations. 2023. <https://github.com/TransformerLensOrg/CircuitsVis>
- [5] Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. AttentionViz: A Global View of Transformer Attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):262–272, 2024. doi:10.1109/TVCG.2023.3327163. <https://arxiv.org/abs/2305.03210>