

Mercury Attention-Head Cartography and GQA/KV Sharing

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

Attention heads are usually shown as attention patterns, heatmaps, or named circuits. Mercury treats a head as an addressable contract inside the model: its query slice, its shared key/value owner, its output slice, and its route through nearby residual dimensions must all remain visible. This paper formalizes that contract for grouped-query attention (GQA) and multi-query attention (MQA), where a visible head number is not enough to identify ownership.

The result is a cartographic rule: a head is not a lone object but a bundle $(q_{\ell,h}, k_{\ell,g(h)}, v_{\ell,g(h)}, o_{\ell,h})$. A correct atlas must draw both the local head and the shared K/V group that serves it. I define head ownership, sharing load, slice completeness, contract ambiguity, and cross-layer head reach. These equations turn a pretty head panel into an auditable map.

Author Note

This paper comes from one very practical annoyance. When I click a head, I do not want a decorative attention square. I want to know what it owns, what it borrows, what it writes back, and which other heads are quietly sharing the same K/V body.

This is where many visualizations become too soft. Mercury has to be harder than a decorative panel.

1 The Head Is Not A Dot

The original Transformer made multi-head attention easy to name [1]. A layer has heads, and a head has attention. That sentence is useful for teaching, but it is too weak for modern local models. Grouped-query attention and multi-query attention changed the contract: several query heads can share fewer key/value heads [2, 3].

So the old visual habit breaks. A label such as “head 0” does not say who owns the K/V memory. The same label also fails to say whether the displayed route is independent, borrowed, or grouped with seven other query routes.

Mercury draws the missing ownership.

2 Head Contract

Let a model have layer ℓ , query head count H_Q , key/value head count H_{KV} , and head dimension d_h . A query head h is mapped to a K/V owner by

$$g(h) = \left\lfloor \frac{hH_{KV}}{H_Q} \right\rfloor.$$

The visible head contract is therefore

$$\mathcal{H}_{\ell,h} = (q_{\ell,h}, k_{\ell,g(h)}, v_{\ell,g(h)}, o_{\ell,h}).$$

This equation separates a shallow label from a real address. The label says “head 0”. The address says which query slice, which shared K/V slice, and which output slice are being touched.

3 Sharing Load

For each K/V owner j , define the serving set

$$S_{\ell,j} = \{h : g(h) = j\}.$$

The sharing load is

$$L_{\ell,j} = |S_{\ell,j}|.$$

In standard multi-head attention, $L_{\ell,j} = 1$. In GQA, the value can be larger. In MQA, one owner may serve all query heads. Mercury should not hide that fact, because a failure in a shared K/V owner can disturb several visible heads at once.

This is why the panel has to show the owner, not only the passenger.

4 Slice Completeness

A contract is complete only when the atlas can identify the Q, K, V, and O slices:

$$C(\mathcal{H}_{\ell,h}) = \mathbf{1}[q_{\ell,h}]\mathbf{1}[k_{\ell,g(h)}]\mathbf{1}[v_{\ell,g(h)}]\mathbf{1}[o_{\ell,h}].$$

The model can still run when this score is missing from the visualization, but the map is blind. A viewer can click something glowing without knowing whether it is a query channel, a shared memory channel, or an output writeback.

Calling that inspection is too generous. It is guessing with a nice interface.

5 Contract Ambiguity

Let $B(\mathcal{H}_{\ell,h})$ be the number of unresolved slices in the contract. The ambiguity score is

$$A_{\ell,h} = \frac{B(\mathcal{H}_{\ell,h})}{4}.$$

For a layer, the mean ambiguity is

$$\bar{A}_{\ell} = \frac{1}{H_Q} \sum_{h=0}^{H_Q-1} A_{\ell,h}.$$

This is a simple reviewer-facing test. If a figure claims to show attention head structure but $\bar{A}_{\ell}$ is high, it is not yet a head map. It is only a surface view.

6 Cross-Layer Head Reach

Mercury also cares about where a head writes. Let E_N be a directed relation graph among measured neuron sites and module slices. For a head contract $\mathcal{H}_{\ell,h}$, define its k -hop downstream reach:

$$R_k(\ell, h) = |\{n \in N : n \in \text{Reach}_k(o_{\ell,h}; E_N)\}|.$$

The local attention pattern does not answer this question. The reach score asks where the output slice can travel after it leaves the head.

This is where the city view matters. The head is no longer a square in one layer; it becomes a road that can be followed.

7 Why Existing Head Views Are Not Enough

BertViz made attention inspectable and remains one of the cleanest attention viewers [4]. Transformer-Lens and CircuitsVis made it easier to inspect components and circuits in code [5, 6]. The Transformer Circuits program clarified residual streams, attention heads, and circuit motifs [7].

Mercury takes a different angle. It does not stop at the heatmap or named head. It asks whether the head can be addressed as a Q/K/V/O contract, whether the K/V owner is shared, and whether downstream paths can be followed after the writeback.

8 A Small Failure Example

Suppose $H_Q = 16$ and $H_{KV} = 2$. Then

$$g(h) = \left\lfloor \frac{h}{8} \right\rfloor.$$

Heads 0 through 7 share K/V owner 0, and heads 8 through 15 share K/V owner 1. A visualization that draws sixteen independent heads is already misleading.

The math is not fancy. This is the point. The mistake is easy to detect, so it should not be allowed to survive in a serious atlas.

9 Cartographic Rule

The Mercury rule is:

$$\text{draw}(h) \Rightarrow \text{draw}(q_h, k_{g(h)}, v_{g(h)}, o_h, L_{g(h)}).$$

If the view cannot draw these terms, it must mark the contract incomplete:

$$\text{ready}(h) = \mathbf{1}[C(\mathcal{H}_{\ell,h}) = 1] \mathbf{1}[A_{\ell,h} = 0].$$

This is not a UI preference. It is an epistemic boundary. The map should show what it knows, and it should also show exactly where it stops knowing.

10 Limits

The equations here do not claim that a head’s function is solved by ownership. A route can be addressable and still not understood. A K/V group can be shared without being causally important for a task. Path reach can overstate meaning if it is not paired with intervention.

Those limits are why the address contract matters. The map does not replace causal experiments; it tells us where to aim them.

11 Conclusion

Mercury Attention-Head Cartography is about refusing the decorative head. A head is a contract, and a contract has owners, borrowers, slices, loads, and paths. Once those pieces are visible, a researcher can stop staring at a pattern and start testing a mechanism.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017. <https://arxiv.org/abs/1706.03762>

- [2] Noam Shazeer. Fast Transformer Decoding: One Write-Head is All You Need. arXiv:1911.02150, 2019. <https://arxiv.org/abs/1911.02150>
- [3] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv:2305.13245, 2023. <https://arxiv.org/abs/2305.13245>
- [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] Neel Nanda and Joseph Bloom. TransformerLens. 2022. <https://transformerlensorg.github.io/TransformerLens/content/citation.html>
- [6] Alan Cooney and Neel Nanda. CircuitsVis. 2023. <https://github.com/TransformerLensOrg/CircuitsVis>
- [7] 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>