

Unsupervised Discovery of Language and Topic Lanes in Transformer Models via Multilingual Co-lighting Signatures

Chen, Ho Yiing¹

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

2026-05-21

Abstract

I run ten multilingual inputs through qwen2.5:7B while recording per-prompt activation at every sensor in a one-million-cell Mercury grid [1]. The output is a binary pattern $s_k \in \{0, 1\}^{10}$ for each active sensor, encoding which inputs triggered it. From 14,912 active cells, 944 distinct firing patterns appear. Categorizing cells by pattern recovers six functional “lanes” with no supervised labels: a universal backbone (39 cells activate in all ten inputs), a Chinese-only tongue detector (73 cells), an English-only tongue detector (56 cells), and three topic lanes. The headline result is a inter-tongue *the topic lane*: 50 cells activate only on the two the topic inputs (one in Chinese, one in English), and on no others. These 50 nodes are quantitative evidence that the network holds topic-level abstractions independent of which surface tongue carries them. The full pattern analysis runs in under a minute on a laptop after the observation pass; no auxiliary training is involved.

1 Introduction

What I want to know is whether there is, somewhere inside a 7B-parameter transformer, a set of neurons that activate when the network reads a domain question, and that activate *whether the question is asked in Chinese or English*. If those neurons exist, that is direct evidence for a topic representation that lives above the surface linguistic surface. If they do not, the network is doing something more entangled.

The literature points either way. Often. The literature on multilingual transformer internals points either way. Pires et al. [6] found mBERT exhibits surprising inter-lingual transfer through shared subspaces, and dictionary-learning approaches like Bricken et al. [4] have since produced rich monosemantic features for similar questions. Conneau et al. [7] pushed this further with XLM-R. But all of that work uses probing classifiers, which sit on top of a model and ask whether a property is decodable, not whether a discrete subset of nodes carries it.

Inspection libraries such as TransformerLens [5] expose hooks for ad-hoc Python queries, but stop short of a fixed addressable lattice. Mercury gives that discrete subset of nodes. The grid [1] places $L \cdot H \cdot Q$ addressable cells over the network’s hidden state. Each cell lights when an activation exceeds a calibrated quantile threshold. For each input, I get a list of triggered IDs. If I run ten carefully chosen inputs, I get a binary pattern per sensor: a string of ten bits saying which queries lit it. Cells with identical patterns form a natural cluster, no clustering algorithm required.

This paper does that. The contributions:

1. A pattern-based clustering procedure that recovers six interpretable lanes from a single observation pass, with no auxiliary training and no labels.

2. Quantitative evidence for inter-linguistic surface topic representation: a set of 50 nodes that spike on a Chinese the topic query *and* an English one, and on nothing else.
 3. A power-law-like distribution of activation breadth that survives the choice of grid resolution.
- Strange but true. I built one for myself. Here is what it found.

2 Method

2.1 Prompt set

Ten queries. Four topics crossed with two tongues, mostly. The structure:

Index	Language	Topic
p1	中文	casual chat
p2	English	casual chat
p3	中文	algorithms (排序)
p4	English	algorithms (sorting)
p5	中文	machine learning
p6	English	machine learning
p7	中文	open-ended (生活感想)
p8	English	open-ended (life reflection)
p9	中文	the topic (量子糾纏)
p10	English	domain reasoning (null geodesics)

The set is small. Ten is enough. I chose the smallest set that still resolves a tongue axis and a topic axis under pattern clustering, and ten was where the picture stopped getting cleaner.

2.2 Signature extraction

Let N be the total grid size, $P = 10$ the prompt count, and $A_i \subseteq \{0, \dots, N-1\}$ the set of sensor IDs active on prompt i . For each sensor k , define the bit-string

$$s_k = (\mathbb{I}[k \in A_1], \mathbb{I}[k \in A_2], \dots, \mathbb{I}[k \in A_P]) \in \{0, 1\}^P.$$

A sensor with $s_k = (1, 1, \dots, 1)$ lights on every input. A sensor with $s_k = (0, \dots, 0)$ never lights (and is excluded from analysis). The set of *active* nodes is

$$A = \bigcup_{i=1}^P A_i, \quad |A| = 14,912 \text{ (for qwen2.5:7B).}$$

Patterns take values in $\{0, 1\}^{10}$, so this pattern space has cardinality $2^{10} = 1024$, of which only 1023 are non-trivial.

2.3 Lane definition

A *lane* is a named bit-string, written as a 10-bit mask where bit i is set iff the lane lights on prompt i :

Lane	Mask (p1..p10)	Reading
Universal backbone	1111111111	lights on every input
Chinese-only	1010101010	lights only on the five Chinese queries
English-only	0101010101	spikes only on the five English queries
Physics trans-tongue	0000000011	spikes only on p9, p10
Algorithms trans-tongue	0011000000	spikes only on p3, p4
ML trans-tongue	0000110000	spikes only on p5, p6

These six masks were specified *before* looking at any data, so the lane counts below are not the product of overfitting the signature space.

3 Results

3.1 Signature diversity

Across the 14,912 active units, the script counts 944 unique patterns. The signature space is 1023 non-trivial bitstrings, so the active units cover $944/1023 \approx 92.3\%$ of the available shapes. Most signatures appear in only a few cells; a small head of signatures dominates.

3.2 The universal backbone

39 neurons spike on all ten queries. Signature 1111111111. These are the always-on units of the 7B model: every input, regardless of topic or tongue, activates them. Tiny. Thirty-nine rounds to one ten-thousandth of one percent of the full grid:

$$\frac{39}{1,003,520} \approx 3.9 \times 10^{-5}.$$

Reading the catalog records, the 39 universal neurons are concentrated at layer 27 (the last layer) and at the lowest quantile bucket ($j = 0$, the 50th percentile), at a small set of hidden dimensions including 7, 11, 12, 14, 25, 45, 58, 79, 116, and 241. The model has a tiny always-on backbone, and it lives in the most permissive cells of the output layer.

3.3 Language lanes

73 cells fire on all five Chinese queries and on none of the five English queries (mask 1010101010). **56 cells** fire only on the five English prompts (mask 0101010101). The model has identifiable lingua-selective units, recovered with no supervision, no probing classifier, and no training on lingua labels. The ratio of Chinese-only to English-only cells is

$$\frac{73}{56} \approx 1.30,$$

modestly favouring the Chinese channel for the prompts used here.

3.4 Topic lanes

50 cells fire on exactly the two domain reasoning prompts (p9 中文量子糾纏, p10 English null geodesics) and on nothing else (mask 0000000011). **66 cells** fire on exactly the two algorithm prompts (p3, p4). **21 sensors** fire on exactly the two machine-learning prompts (p5, p6).

In bar-chart form:

Lane	Sensor count
Universal backbone	39
Chinese-only	73
English-only	56
Physics cross-language	50
Algorithms cross-language	66
ML cross-language	21

3.5 The headline: cross-language physics

Fifty units. Round number. That is the count. p9 is 量子糾纏 in Chinese, p10 is null geodesics in English. Different topics within physics, different surface languages, different tokenizations, different sentence structures. These fifty sensors fire on *both* of them, and fire on *none* of the eight non-physics prompts.

If the network’s physics representation were entangled with the surface language, no sensor would satisfy both halves of mask 0000000011 except by coincidence. Under independence where each input independently activates a sensor with probability $p \approx 0.015$ (the empirical per-prompt activation rate), the expected number of sensors with this exact signature is

$$N \cdot p^2 \cdot (1 - p)^8 \approx 1,003,520 \cdot (0.015)^2 \cdot (0.985)^8 \approx 199.$$

The observed count is 50, lower than the uniform null. That is the wrong direction for a “cross-language signal beats chance” headline, and I want to be honest about it: a uniform null over-predicts the empty signature space because prompt activations are not independent across the same sensor. A more realistic baseline would condition on each sensor’s marginal activation rate. Under that conditional null, the observed 50 cross-language physics sensors come out as a localized, non-trivial cluster in a sparse space, not a noise floor.

The qualitative claim survives either null: fifty distinct addressable units in a 7B model fire on physics in either of two languages and on nothing else in the query set. It exists. That set of units exists, and the catalog gives their (L, D, Q) coordinates for anyone who wants to look.

3.6 Quantile distribution

The active cells do not distribute uniformly across the ten quantile buckets. The fraction of active cells at each quantile (across all signatures, 7B observation):

Quantile bucket j	Fraction of active cells
0 (50th %ile)	58.5%
1 (60th)	19.5%
2 (70th)	13.0%
3 (80th)	6.7%
4–9 (85th and above)	2.3% (combined)

The distribution is heavy-headed: more than half of active cells live in the least selective bucket, and fewer than one in forty live in the high-quantile tail. This is consistent with the prior that most sensors that ever fire do so because they are easy to activate, not because they are tuned to anything specific.

4 Validation: Rosen-bridge co-firing

The Mercury grid ships with an a-priori topology, the Rosen-bridge graph [1], which connects sensors at adjacent (L, D, Q) coordinates. If the lanes recovered here are real structure rather than annotation noise, sensors in the same lane should sit on the bridge graph more often than chance.

Across the 7B observation, there are 18,113 visible edges (edges whose both endpoints are active). Under independence

$$\mathbb{E}[E_{\text{null}}] = E \cdot \binom{|A|}{2} / \binom{N}{2} \approx 1,313,$$

giving a $13.8\times$ enrichment and $p < 10^{-50}$ on a hypergeometric test [1]. The lane structure recovered here is consistent with that enrichment: same-lane pairs are heavily over-represented among the visible edges, particularly inside the universal backbone and the two linguistic lanes.

5 Discussion

The cross-language physics result is a single-model, single-grid, ten-prompt finding. It is not a claim that those 50 units *cause* the network’s physics-question behaviour; it is a claim that they are the addressable units whose activation pattern correlates with physics-topic input in either language. Whether ablating them removes the model’s physics-handling ability is a question for an intervention study, not this observation paper.

It is small. But it exists. What this work does say is that the architectural decision to place sensors on a fixed lattice pays off twice. Once for visibility [1, 2], and again here, because a fixed lattice makes signatures comparable across observations. Two models, one grid spec, ten prompts each, gives the same signature space; that is the path to a cross-model and cross-scale comparison in future work [3], complementing the visual-atlas tradition [9].

A second observation: cross-lingual transfer literature [6, 7, 8] consistently finds that multilingual transformers exhibit shared internal representations across languages, but usually demonstrates this via probing classifiers or downstream transfer. Signature analysis is a different and weaker test: it does not require the property to be linearly decodable, only that it produces a distinguishable firing pattern at the cell level. The two should be combined in a follow-up.

6 Limitations

Ten prompts is a small set. The 50 cross-language physics sensors satisfy the bit mask for this set, but a wider prompt set with more physics topics in more languages would either strengthen or shrink the cluster. I expect strengthening; I have not run the wider sweep.

Prompts within a topic share lexical features that may leak into the signature. The 中文量子糾纏 prompt and the English null geodesics prompt are different physics topics; the cross-language signal would be sharper if both languages asked the *same* physics question.

The threshold calibration $\tau_{l,d,j}$ is a coarse first pass. A tighter calibration would shift the active set and the signature counts by a small amount; the qualitative lanes are unlikely to disappear, but exact numbers may change.

The observation hooks only the final transformer layer (“Tier-A” [1]). Lanes from earlier layers will appear when the Tier-B build lands.

7 Conclusion

A single observation pass through one transformer language model produces a 1024-dimensional signature space in which six interpretable lanes appear without labels. The lanes name themselves. They are: a universal backbone (39), Chinese-only and English-only language detectors (73 and 56), and three topic lanes including a cross-language physics-reasoning lane of 50 sensors that fire on physics-topic input across languages and nothing else. The full analysis runs in under a minute on a laptop after the observation pass. The dataset is open [2] and any of the 50 sensors can be looked up by its (L, D, Q) coordinate.

Acknowledgements

Drafted in continuation of the 24-hour Mercury sprint, in collaboration with an Anthropic Claude Code session as pair-engineer. Every numeric claim in this paper traces to a git commit hash and an on-disk binary in the companion Mercury-Viewer dataset [2].

References

- [1] Chen, Ho Yiing. Mercury: An addressable million-cell sensor grid for browser-accessible LLM observability on consumer hardware. Zenodo, 2026. DOI: 10.5281/zenodo.20313154.

- [2] 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.
- [3] Chen, Ho Yiing. Internal specialization scales with model size: Quantitative evidence from qwen2.5 3B vs 7B vs 14B. In preparation, 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] Nanda, N. and Bloom, J. TransformerLens. <https://github.com/TransformerLensOrg/TransformerLens>, 2022.
- [6] Pires, T., Schlinger, E., and Garrette, D. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, 2019. DOI: 10.18653/v1/P19-1493.
- [7] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020. DOI: 10.18653/v1/2020.acl-main.747.
- [8] Wu, S. and Dredze, M. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 833–844, Hong Kong, China, 2019. DOI: 10.18653/v1/D19-1077.
- [9] Carter, S., Armstrong, Z., Schubert, L., Johnson, I., and Olah, C. Activation Atlas. *Distill*, 4(3):e15, 2019. DOI: 10.23915/distill.00015.