One concept map for four models

Will Sackfield, Open Interpretability · https://open-interpretability.com/research/shared-concept-space

Abstract

We placed 1,998 studied neuron poles of four language models in one shared map, by the meaning of the text each pole fires on. The map is fitted once over all four, and a new pole is placed among its neighbours without moving the old ones. Each model's poles reach every region of the map, and every region of a finer cut into 29. Pooled over the corpora they were scanned on, most pairs of models spread over the regions less alike than shuffled labels would. But the scan corpus explains most of that difference. One model's poles from two corpora overlap 62–69% by region, against 84–92% for shuffled labels. When both models were scanned on one corpus, pile-10k, no pair differs from its null. On this evidence the registry's studied poles of the four models cover the same concepts, and the map can't yet tell the models apart once the corpus is fixed.

Status: Preprint. Work left: #5425 (one fixed corpus and an untrained control)

1 Introduction

The registry explains neurons one pole at a time, in four models. Each model page used to draw its own concept map, fitted to that model's poles alone, so a place on one map meant nothing on another. Huh et al. (2024) argue that the representations of different models converge "to a shared statistical model of reality". A shared map asks a narrower question of the registry's own units. Do different models' explained neurons fall in the same regions of meaning, and does a new model fill in the space that the others made? It bears on research question 4, whether explanations carry across models (research questions). If two models' poles fall in different regions, their explanations can't carry.

The paper's contributions are these:

1. **One shared map that new poles extend.** All four models' poles share one fit. A new pole is placed among its nearest old ones, and no old point moves until a deliberate refit. A re-run on the same records moved none of the 1,998 points. *Limit:* the map places a pole by its top examples, so it is a map of text.
2. **No model keeps a region to itself.** Each model has poles in every region at 7, 14 and 29 regions, and in 57 to 60 of 60 (Figure 2). *Limit:* the registry's studied poles are a sample of each model.
3. **Pooled over corpora, the models spread differently.** At the map's 7 regions, all six pairs overlap less than their shuffled nulls, $p \leq 0.025$ (Table 3).
4. **The scan corpus explains most of the difference.** One model's poles from pile-10k and from the mixed corpus overlap 62–69%, against nulls of 84–92%. On pile-10k alone, no pair of models differs from its null, with p from 0.48 to 0.995 (Table 4). *Limit:* the untrained-network control is not yet run (#5425).

These results don't show that the four models learned the same concepts. They show that the registry's evidence can't yet tell them apart.

2 Related work

Huh et al. (2024) argue that representations in different models converge. They measure how alike two models are with a mutual nearest-neighbour metric: the average overlap of each input's nearest neighbours in the two models' representation spaces. This paper compares where units fall, not how inputs are arranged, and it uses the registry's explained poles only. UMAP (McInnes et al., 2018) is a manifold-learning method for dimension reduction, and the map uses it for the layout. Leiden (Traag et al., 2019) finds communities that are guaranteed to be connected, and the map uses it for the regions. Sentence-BERT (Reimers & Gurevych, 2019) gives sentence embeddings that can be compared with cosine similarity. The embedder here, `all-MiniLM-L6-v2`, is one of them: a 6-layer MiniLM model (Wang et al., 2020), fine-tuned on over a billion sentence pairs. The registry's earlier study of concept maps found that layouts keep only about half of each point's neighbours (the study). The measure here therefore uses regions, not distances.

3 Method

The map. `open-interpretability map` (`openinterpretability/concept_map.py`) reads the newest record of every studied MLP neuron pole of the four configured models: Pythia-70m, GPT-2 small, Qwen2.5-0.5B and Qwen2.5-1.5B. A pole's texts are its record's top examples, each cut to 20 tokens on each side of its peak. Lines of the package's 40-line sample corpus are left out, because a scan on it finds the same few lines for every neuron. A pole with no other text isn't placed, which leaves 1,998 poles. Each text is embedded with `all-MiniLM-L6-v2`. A pole's vector is the mean of its texts' embeddings. All poles of all models are centred on one pooled mean and L2-normalised. The layout is a two-dimensional UMAP, cosine, seed 0. The regions are Leiden communities of the 10-nearest-neighbour graph. The resolution is searched for the count nearest 8, which gave 7. Each region is named only from its VERIFIED members' explanations.

Stable positions. A later run keeps each old pole's position and region. A new pole goes at the similarity-weighted mean of its 10 nearest old poles, in their majority region. No fitted reducer is stored. Only a deliberate refit (`map --refit`) moves old points, and it starts from their old positions.

The measure (`openinterpretability/concept_regions.py`). A model's coverage is the share of regions that hold at least one of its poles. The overlap of two models is the histogram intersection of their region shares, 100% when the two spread alike. The null shuffles the model labels across all poles 200 times (seed 4586). It is the overlap that two models of those sizes would show if they drew from one spread. The one-sided p counts the shuffles at or below the overlap seen, plus one, over 201, so the smallest p is 0.005. A model's split-half overlap is the overlap of two random halves of its own poles. The partitions finer than the map's are Leiden cuts of the same graph, searched for 15, 30 and 60 regions (`scripts/concept_space_study.py`).

The corpus control. Each point records the family of the corpus its texts come from: `mixed` for every version of the mixed blend, else the dataset's name. Two families hold at least 40 poles of every model, pile-10k and the mixed corpus (Table 1).

Table 1. Each model's mapped poles, and how many come from each of the two shared corpus families.

Model	Poles	On pile-10k	On the mixed corpus
Pythia-70m	293	41	251
GPT-2 small	724	251	302
Qwen2.5-0.5B	698	145	133
Qwen2.5-1.5B	283	104	161

Agents wrote this study's code and text, and the maintainer directed it. The explanations that name the regions come from the registry's audits.

4 Every model reaches every region

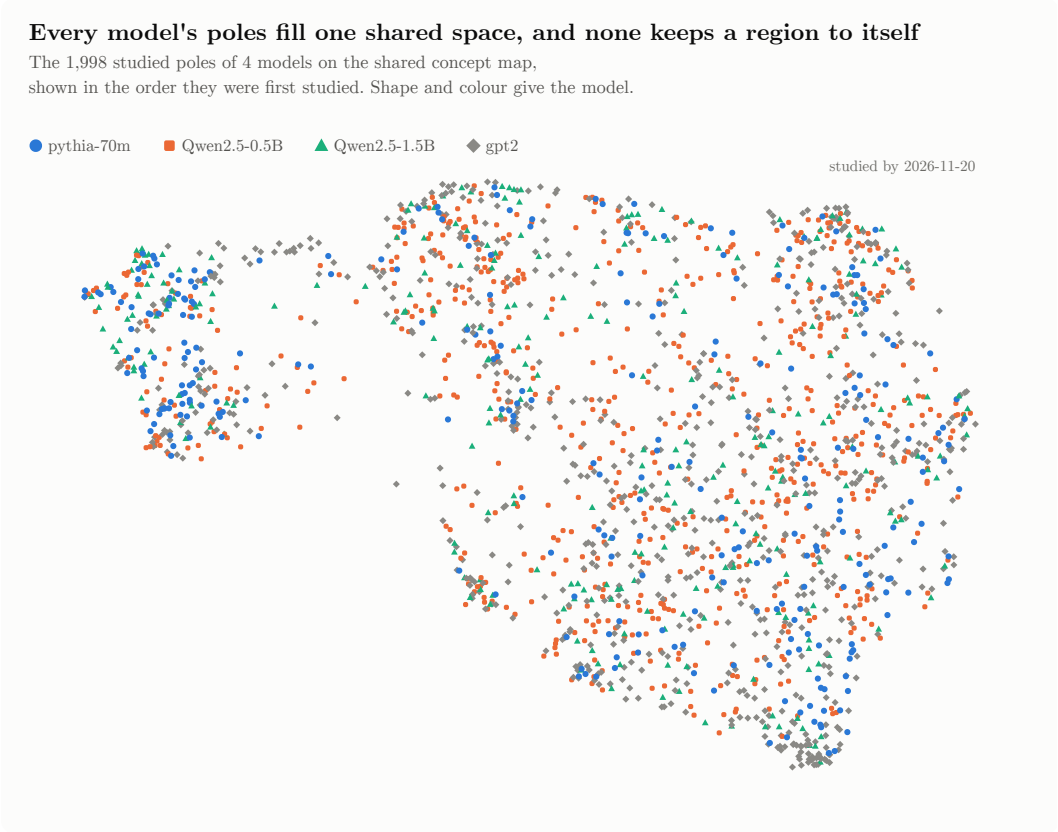

Figure 1. The 1,998 poles on the shared map, appearing in the order they were first studied. Shape and colour give the model.

The poles of all four models spread across the whole map, and each model's poles appear across it as the dates advance (Figure 1). Table 2 gives the map's 7 regions, by size, with each model's share of its poles in each. The names are the words that set a region's VERIFIED explanations apart, so they are clues, not labels. Pythia-70m has 28% of its poles in the region of scripts and languages, against 11–20% for the others. Most of Pythia-70m's poles were scanned on the mixed corpus, which holds Wikipedia in ten languages. The section after next tests whether that is the model or the corpus.

Table 2. The map's regions, and each model's share of its poles in each.

Region	Poles	Pythia-70m	GPT-2	Qwen2.5-0.5B	Qwen2.5-1.5B
guilty · penalty · penal	443	16%	23%	24%	22%
abbey · cyrillic · korean	299	28%	12%	11%	20%
sixty · land · feed	297	14%	14%	16%	16%
brackets · django · operator	278	9%	12%	17%	15%
clinical · disease · micro	269	10%	12%	17%	11%
effort · feminine · got	262	15%	19%	8%	9%
permission · gnu · unspaced	150	7%	8%	8%	7%

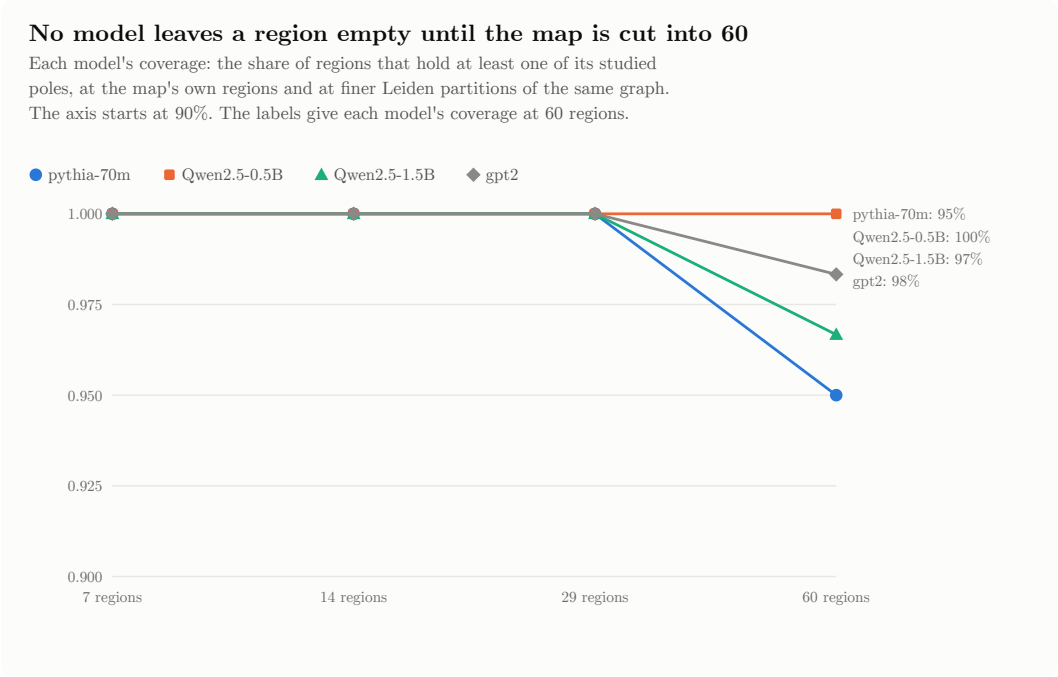

Figure 2. Each model's coverage at the map's 7 regions and at Leiden cuts of 14, 29 and 60 regions. The axis starts at 90%.

Every model reaches every region up to 29 regions (Figure 2). At 60 regions, Pythia-70m reaches 57, GPT-2 59, Qwen2.5-0.5B 60 and Qwen2.5-1.5B 58. The models with fewer poles miss more regions, as sampling alone would make them: Pythia-70m and Qwen2.5-1.5B have fewer than 300 poles each, against about 700 for the other two.

5 Pooled over corpora, the models spread differently

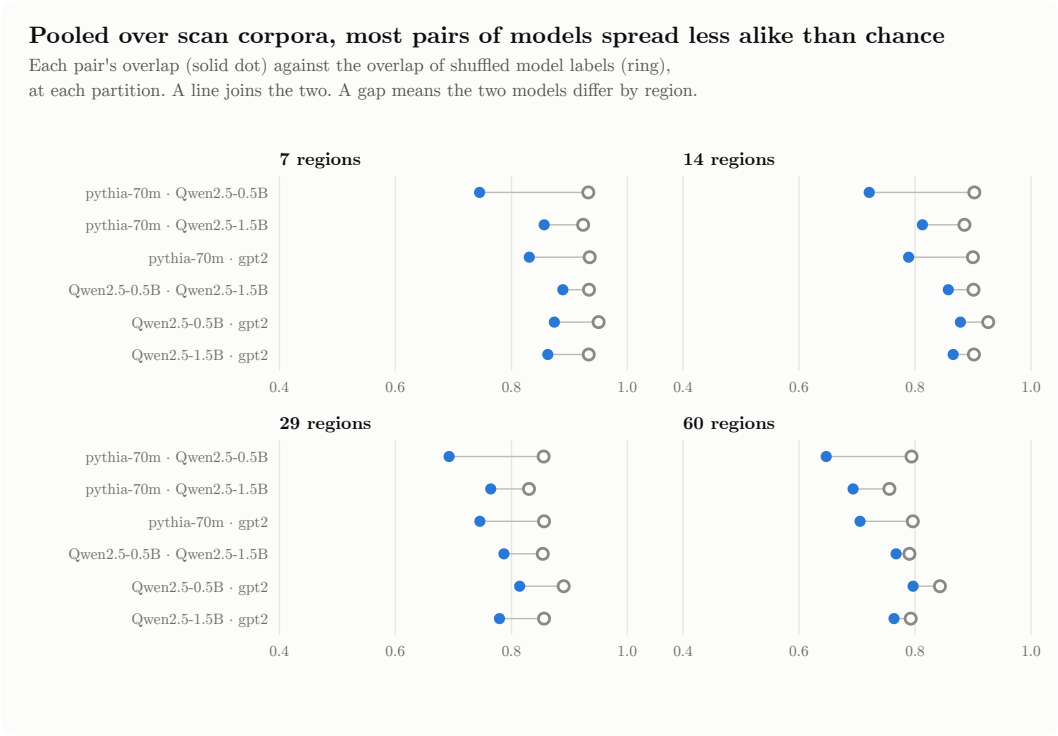

Figure 3. Each pair's overlap (solid dot) against its shuffled null (ring), at four partitions, with poles from every scan corpus.

Table 3. Each pair's overlap and shuffled null, at the map's 7 regions and at 60, with poles from every scan corpus.

Pair	7 regions	Null	p	60 regions	Null	p
Pythia-70m · Qwen2.5-0.5B	75%	93%	0.005	65%	79%	0.005
Pythia-70m · Qwen2.5-1.5B	86%	92%	0.005	69%	76%	0.010
Pythia-70m · GPT-2	83%	93%	0.005	70%	80%	0.005
Qwen2.5-0.5B · Qwen2.5-1.5B	89%	93%	0.025	77%	79%	0.129
Qwen2.5-0.5B · GPT-2	87%	95%	0.005	80%	84%	0.005
Qwen2.5-1.5B · GPT-2	86%	93%	0.005	76%	79%	0.075

At 7 regions every pair falls below its null (Table 3). At 60 regions four pairs do, and the two Qwen models come closest to theirs. Pythia-70m and Qwen2.5-0.5B differ most at every partition (Figure 3). Each model's split-half overlap at 7 regions is 89% for Pythia-70m and Qwen2.5-1.5B, and 93% for the two larger samples. Every pair with Pythia-70m falls below Pythia-70m's own split-half overlap. Halves are smaller samples than whole models, so the split-half overlap is a loose ceiling, and the shuffled null is the test.

6 The scan corpus explains most of the difference

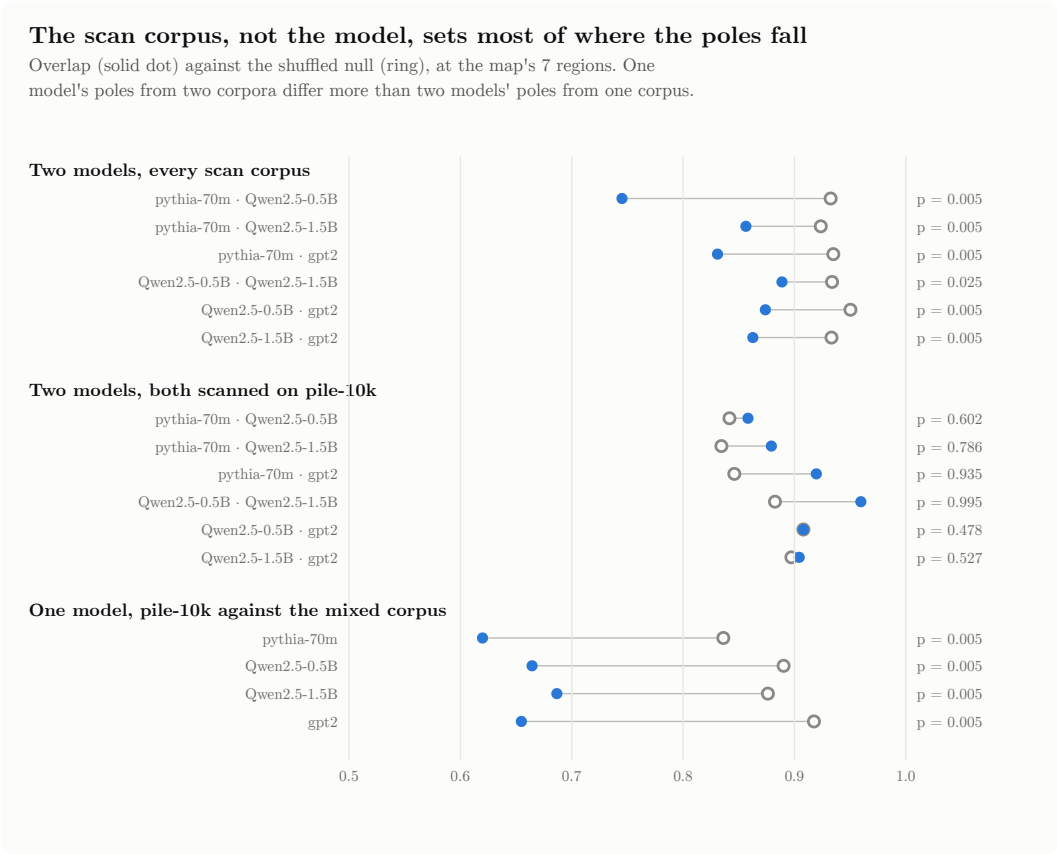

Figure 4. Overlap against the shuffled null at the map's 7 regions: pairs pooled over corpora, pairs on pile-10k alone, and each model against itself across two corpora.

Table 4. The corpus control at the map's 7 regions.

Comparison	Overlap	Null	p
Pythia-70m, pile-10k · mixed	62%	84%	0.005
GPT-2, pile-10k · mixed	65%	92%	0.005
Qwen2.5-0.5B, pile-10k · mixed	66%	89%	0.005
Qwen2.5-1.5B, pile-10k · mixed	69%	88%	0.005
Pythia-70m · Qwen2.5-0.5B, both on pile-10k	86%	84%	0.602
Pythia-70m · Qwen2.5-1.5B, both on pile-10k	88%	83%	0.786
Pythia-70m · GPT-2, both on pile-10k	92%	85%	0.935
Qwen2.5-0.5B · Qwen2.5-1.5B, both on pile-10k	96%	88%	0.995
Qwen2.5-0.5B · GPT-2, both on pile-10k	91%	91%	0.478
Qwen2.5-1.5B · GPT-2, both on pile-10k	90%	90%	0.527

One model against itself across two corpora is further from its null than any two models are (Figure 4). On pile-10k, every pair sits at or above its null (Table 4). The two Qwen models, which share a tokenizer and training data, overlap 96%, the most of any pair. At 29 regions on pile-10k, Pythia-70m with GPT-2 falls below its null ($p = 0.035$). With six pairs, a Bonferroni threshold is 0.008, which it doesn't pass.

On the mixed corpus, four pairs fall below their nulls at 7 regions: Qwen2.5-0.5B with GPT-2 ($p = 0.005$), Qwen2.5-1.5B with GPT-2 ($p = 0.010$), Pythia-70m with Qwen2.5-0.5B ($p = 0.020$) and Pythia-70m with GPT-2 ($p = 0.035$). The mixed corpus is six versions of a blend whose parts changed between versions, and the models were scanned on different versions. So it controls the corpus less well than pile-10k does.

7 Discussion

The pooled result looks like a difference between models. The control says most of it is a difference between corpora. Pythia-70m differs most from the others pooled, and it is also the model scanned almost only on the mixed corpus. Its large share in the region of scripts and languages matches that corpus, which holds ten Wikipedia languages. On pile-10k, the four models' spreads are as alike as random splits of one spread.

For research question 4 this is a modest result. The registry's explained poles of four models fall in the same regions of meaning, so there is room for their explanations to carry. It does not say that a given pole has a counterpart in another model. The sister search answers that (sister neurons). It also does not separate learned concepts from the corpus's own topics. An untrained network's neurons, scanned on the same corpus, might fall in the same regions too.

8 Limitations

- **Positions read the corpus, not the neuron.** A pole is placed by the text it fires on most, and that text comes from whatever corpus the scan used. The shared map is a map of top examples.
- **The studied poles are not a fixed sample.** The registry draws neurons at random, but it follows some up and stops on dead or default-on ones. Coverage therefore measures what has been studied.
- **The regions are coarse.** At 7 regions, coverage is 100% for every model and can't separate them. The finer cuts are Leiden partitions of the same graph, not separate fits.
- **Distances mean little.** Another layout keeps about half of each point's neighbours. The measure uses regions, not distances.
- **No untrained control.** #5425 holds that test.
- **No correction across tests.** Each p is per test. The text names the Bonferroni threshold where a result is near it.

9 Conclusion

One shared map now holds every model's studied poles, and new poles extend it without moving the old ones. Each model reaches every region. The differences between models that the pooled data show come mostly from the scan corpus: on one corpus, the models can't be told apart. These tests would go further:

- Place every pole from one fixed corpus, and add an untrained network as a control (#5425).
- Repeat the measure as each model's studied poles grow. A new model's poles fill in the same regions, because the map keeps old positions.

- Compare the map's regions with the mutual nearest-neighbour alignment of Huh et al. (2024), which measures how alike two models' representations are.

References

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). *Position: The Platonic Representation Hypothesis*. In Proceedings of the 41st International Conference on Machine Learning (PMLR 235, pp. [removed]). PMLR. Preprint: arXiv:2405.07987.

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Preprint. arXiv:1802.03426.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. EMNLP 2019. arXiv:1908.10084.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). *From Louvain to Leiden: Guaranteeing Well-Connected Communities*. Scientific Reports, 9, 5233. arXiv:1810.08473.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). NeurIPS proceedings. Preprint: arXiv:2002.10957.

Appendices

Reproduce

```
uv run --extra map open-interpretability map                    # data/concept-maps/shared.json, no refit
uv run --extra map python scripts/concept_space_study.py   # papers/shared-concept-space/results.jsonl
uv run python scripts/paper_figures.py                  # the figures
```

The seeds are fixed: UMAP `random_state=0`, Leiden seed 0 and the shuffles' seed 4586. The study embeds every pole again on CPU, in about ten minutes. The site's concept atlas draws the same map, and each model page draws it with that model's poles marked.