Do explanations carry across models?

Will Sackfield, Open Interpretability · https://open-interpretability.com/research/explanations-across-models

Abstract

When a neuron pole in one model has a VERIFIED explanation, does the same explanation describe a matched pole in another model? The registry has 80 sister studies that test this. Each audits a pole of one model with another model's VERIFIED explanation, unchanged. The explanation verifies on 8 of 50 sisters (16%), 1 of 26 other close matches and 0 of 4 random poles. Most sisters fail only in part: their median balanced accuracy is 0.78, against 0.55 for other matches and 0.45 for random poles. The explanation is near chance (REJECTED) on 24% of sisters and 54% of other matches. So matching finds poles where an explanation largely carries over, but it seldom carries over whole. Sisters across tokenizers, GPT-2 to Qwen, carry about as well as sisters within the Qwen family.

Status: Preprint. Work left: #3817 (a random-pole control and one protocol).

1 Introduction

Interpretability research often explains a trained network one unit at a time. Such an explanation describes one model. A second model, trained on other data or at another size, gets its own audit from the start. If units in different models learned the same features, an explanation written for one model could describe its counterpart in another. The cost of explaining a new model would then fall, and a shared explanation would be evidence that the feature is universal.

Earlier work asks whether units recur across networks, but it measures recurrence by correlated activity. Li et al. (2016) match the units of two vision networks, trained from different random starts, by the correlation of their activations. Gurnee et al. (2024) do the same for GPT-2 models trained from five random seeds, against a baseline from a random rotation of each layer. Both studies find that some units have a close counterpart and others do not. Neither study asks whether a natural-language explanation of one unit, written and tested on its own model, predicts what its counterpart does. Tigges et al. (2026) name this gap for their own results on neuron poles: "We must also establish how far these findings generalize across models and behaviors."

The registry that this paper draws on is in a position to measure this. It holds explanations of single neuron poles, each tested by an adversarial review, in GPT-2 small, Qwen2.5-0.5B and Qwen2.5-1.5B. A search matches each VERIFIED pole with the poles of another model. A *sister study* then audits a matched pole with the source pole's explanation, unchanged, under the same review. By 2026-11-14 the registry held 89 such studies, on 80 distinct poles, which had not yet been read together. If an explanation carried over, the registry could explain new models without auditing every pole again.

This paper asks whether a VERIFIED explanation describes a matched pole in another model. It also asks how much of an explanation carries when the whole of it does not, and whether a change of tokenizer matters. It is research question 4, *Do explanations carry across models?* (research questions).

The paper's contributions are these:

- A measurement of how often a VERIFIED explanation passes the same review on another model's pole. It compares the source pole's sister, its other close matches and random poles.
- Evidence that an explanation mostly carries in part. On a sister, it predicts most probes but seldom all of them.
- A comparison of sisters within one tokenizer (Qwen2.5-0.5B and Qwen2.5-1.5B) and across two tokenizers (GPT-2 and Qwen2.5).
- Evidence that a higher correlation between the two poles goes with an explanation that carries over.
- A per-study table of the 80 tested poles, with each one's class, correlation, verdict and balanced accuracy (Appendix B).

2 Related work

Matching units across networks. Li et al. (2016, Section 3.1) compute the correlation of every unit in one network with every unit in another. They match units in two ways. A *semi-matching* pairs each unit with the most correlated unit of the other network. A *bipartite matching* pairs the units one to one, so that each unit has a unique partner. They find that, "for many units a one-to-one assignment is possible", and that for other units no good match exists. This paper's *mutual best match* is a strict form of their one-to-one pairing. A pair is mutual only when each unit is the other's semi-matching partner, so mutual pairs are always one to one. The matching step is therefore not new. What this paper adds is the use of a match to choose poles on which to test an explanation.

Universal neurons. Gurnee et al. (2024, Section 4.1) cite Li et al. as inspiration, and compute the maximum correlation of each GPT-2 neuron with every neuron of four other seeds. For a baseline, they correlate each neuron with the elements of "a random (approximate) rotation of a layer of neurons". They call the gap between the two the *excess correlation*, and they find 1–5% of neurons universal. The registry's search follows them in both steps, the maximum correlation and the rotated baseline. It differs in two ways. It compares models of different families, sizes and tokenizers, not seeds of one model. And it tests explanations, where Gurnee et al. interpret the universal neurons from their activations and weights.

The neuron basis. Shabalin and Piotrowski (2026) split each gated MLP neuron by the sign of its activation, and "treat each neuron as two features — one per sign". The registry's pole is the same unit. They also train two transcoders on one layer of one model, with different initialization seeds. Among the features that recur in both runs, most are aligned with a single neuron (their Section 3.5). Their runs are two dictionaries for one model, not two models, so their result does not test transfer between models. It does suggest that features aligned with neurons are the stable ones, which makes the neuron basis a sensible place to look for matches. Tigges et al. (2026) also read each pole apart, and report that a pole's positive and negative values "usually (but not always)"

have unrelated roles. They split each pole's function into a few *modes*, and find that most neurons have one mode, or one that dominates. They tested one model, and name the question across models as open work.

Matching layers by depth. Faumi and Sumanathilaka (2026) compare GPT-2 with DistilGPT-2, its six-layer distilled student. They pair the two models' layers either by index or by relative depth, which they call proportional mapping. The similarity of the residual streams, measured as CKA, is 0.921 under proportional mapping and 0.752 under the same index. At the student's last layer, it is 0.314 under the same index and 0.790 under proportional mapping. The registry's search does not assume a layer, because it compares every layer. Its sisters nonetheless sit at about the same relative depth (sister neurons). The registry's review of their paper proposes DistilGPT-2, which shares GPT-2's tokenizer, as a further sibling for the search (#3853, review).

Evaluating explanations. Oikarinen et al. (2025) frame an explanation's score as a metric over two vectors, the unit's activations and the concept's presence. They test each metric with two sanity tests. One removes positive labels at random, and the other adds extra positive labels. Balanced accuracy fails the extra-labels test on rare concepts, because it cannot tell an exact concept from a too generic one. In their theoretical experiments, all metrics except recall pass that test on concepts active on more than 10% of inputs. This paper reports balanced accuracy, so it relies on that condition (*Method*). Yan et al. (2025) ask whether an explanation is stable, that is, "consistent across different probing datasets". They resample the probing set to measure it. Each sister study in this paper reads one probe set from one reviewer. So their kind of instability limits the trust that one verdict deserves (*Limitations*).

3 Method

Terms. A neuron in a transformer's MLP layer can carry a different meaning when it is strongly positive and when it is strongly negative. The registry studies one side at a time, and calls it a *pole*. We write L5/N3833+ for the positive pole of neuron 3833 in layer 5, and N3833− for its negative pole. An *explanation* of a pole is a sentence that says on which text the pole fires. It must be precise enough that a third party can label new text as active or inactive.

The review. A second model, the *reviewer*, writes probes meant to break the explanation, without access to the pole's activations. *Hard negatives* are near misses that the explanation says should not activate the pole. *Holdouts* are new texts that it says should activate the pole. A probe activates the pole when the pole's peak on it reaches 0.25 of the pole's maximum, which we call the *activation line*. Near-duplicate probes are clustered, and each cluster counts as one observation. *Specificity* is the share of hard negatives that stay below the line. *Sensitivity* is the share of holdouts that reach it.

Verdicts. An explanation is *VERIFIED* when every gate of the review passes. The main gates are specificity of at least 0.80 and sensitivity of at least 0.60, with enough distinct probes of each kind. Other gates ask that the proposer's predicted counterfactuals reproduce, and that the explanation beats trivial baselines such as token identity. A last gate asks that the pole is not on for most tokens. An explanation is *REJECTED* when the probe set is adequate and its balanced accuracy is below 0.60, which is close to chance. Every other outcome is *NEEDS_REVISION*. The explanation then passes many probes and fails some, or fails a gate other than the probes.

Balanced accuracy. The balanced accuracy of an explanation on a probe set is the mean of its sensitivity and its specificity. It is thus the mean of the true-positive and true-negative rates. A random guess scores 0.5, whatever the share of each kind of probe. Oikarinen et al. (2025) show that balanced accuracy fails one of their sanity tests when the concept is rare. In every one of the 80 studies, holdouts make up 30–37% of the hard negatives and holdouts together. The probe sets are therefore balanced enough for the condition that Oikarinen et al. give, which is a concept active on more than 10% of inputs.

The models. We use three models. GPT-2 small (Radford et al., 2019) has 12 layers, and its byte-level vocabulary "is expanded to 50,257" tokens (their Section 2.3). Qwen2.5-0.5B and Qwen2.5-1.5B have 24 and 28 layers, and a SwiGLU MLP (Yang et al., 2024, Section 2). Both use Qwen's byte-level tokenizer, with 151,643 regular tokens. Yang et al. (2024) state that this "establishes a unified vocabulary across all Qwen2.5 models". So the two Qwen models share a tokenizer, and GPT-2 to Qwen crosses tokenizers.

Matching. A search matches each VERIFIED pole with the poles of another model. Both models read the same corpus of 500 documents. They are compared at *shared boundaries*, the character offsets where a token ends in both tokenizations. Both models are causal, so at a shared boundary each has read the same text. The search correlates the pole's activation, by Pearson's r, with every pole of every layer of the other model, and keeps the 5 best matches. A second pass correlates each match with every pole of the first model.

- **Sister.** A match whose own best partner is the source pole is a *mutual* best match. A mutual best match is a *sister* when its correlation also beats the best that a random rotation of the matched layer reaches.
- **Our steps.** The maximum correlation across models and the rotated baseline follow Gurnee et al. (2024). The mutual step is a strict form of the one-to-one matching of Li et al. (2016). This registry's own step is the use of the match to pick poles for a sister study.

Sister studies. A sister study takes a source pole's VERIFIED explanation as the hypothesis for a pole of another model. The explanation is not changed. A reviewer writes new probes against it for the tested pole, and the pole is judged by the protocol of the time. The registry's planner draws a source's sister, its other top matches and, at a low weight, random poles of the other model as controls.

Classes. Each tested pole falls in one of three classes, by its relation to the source pole:

- **Sister:** the source's sister in the tested model.
- **Another top match:** one of the source's 5 best matches that is not its sister.
- **Random pole:** any other pole of the tested model.

Each tested pole counts once, by its first study. That gives 80 poles from 89 studies. For a sister or another match, r is the correlation between the two poles that the search recorded. We call a study *within a tokenizer* when both models are Qwen2.5, and *across tokenizers* when one is GPT-2.

Intervals and tests. For each share, we give a 95% Wilson score interval. For k successes in n trials, with $\widehat{p} = k \,/\, n$ and $z = 1.96$, the interval is

$$\frac{\widehat{p}+\frac{z^2}{2n}}{1+\frac{z^2}{n}} \pm \frac{z}{1+\frac{z^2}{n}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}+\frac{z^2}{4n^2}}$$

Unlike the normal interval, it stays within 0 and 1 and does not collapse to a point at $k = 0$. To compare two shares, we use the two-sided Fisher exact test on the 2×2 table of the two classes and the two outcomes. With the margins of the table fixed, the counts follow a hypergeometric distribution. The p-value is the sum of the probabilities of all tables no more likely than the one observed.

Setup. The search is `open-interpretability sisters` (`docs/PROTOCOL.md` §4, *Sister neurons*), and its rows are in `docs/analyses/sister-neurons.jsonl`. A sister study is `open-interpretability audit --sister-of`, with the source record as its argument. Draws of the kind `sister` come from `open-interpretability next`. `scripts/explanation_transfer.py` reads every study that names a source and assigns its class. The intervals and tests are in `openinterpretability/builder/learned_baseline.py`. Appendix A gives the commands.

Figure. Figure 1 moves. It reveals the three classes in turn, each with its median, so that a reader sees the contrast build from sisters to random poles.

4 How often does an explanation verify on another model's pole?

The first question is whether a VERIFIED explanation passes the full review on a matched pole. If poles in two models learned the same feature, a sister would often verify. Table 1 gives the counts. The explanation verifies on 8 of 50 sisters, which is 16% (95% CI 8–29%). It verifies on 1 of 26 other top matches, which is 4% (1–19%), and on none of 4 random poles (0–49%).

Tested pole	Studies	VERIFIED (95% CI)	REJECTED (95% CI)	Median balanced accuracy
The source's sister	50	8 (8–29%)	12 (14–37%)	0.78
Another top match	26	1 (1–19%)	14 (35–71%)	0.55
A random pole	4	0 (0–49%)	4 (51–100%)	0.45

Table 1. How far each explanation carries, by the tested pole's relation to the source pole. Each tested pole counts once, by its first study. The intervals are 95% Wilson score intervals. A REJECTED study has balanced accuracy below 0.60 on an adequate probe set.

Sisters verify four times as often as other matches, but the difference is not significant (Fisher $p = 0.15$). The counts are small. Both intervals are wide, and they overlap from 8% to 19%. So the data show that a whole explanation seldom carries over, even to a sister. They do not show that a sister is a better target than another close match, on this measure.

The 80 studies were not a designed experiment. The registry's planner drew them, and each ran with the proposer and reviewer models that its draw gave. This rate is therefore a rate over the studies that the planner chose to run, not over all matched poles.

5 How much of an explanation carries over?

A verdict of VERIFIED needs every gate to pass, so it hides how much of an explanation holds. The second question is how well the explanation predicts the probes when it does not verify. Balanced accuracy answers it, and so does the share REJECTED.

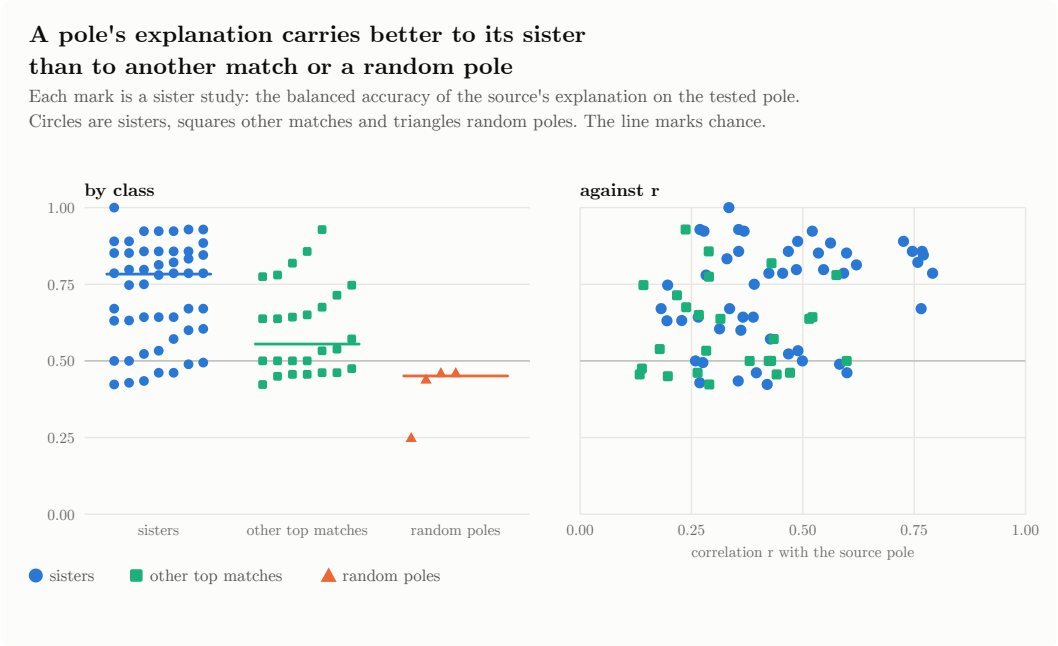

Figure 1. Left: the balanced accuracy of each of the 80 studies, in one column per class, with each class's median as a short line. The line across the panel marks chance, 0.5. Right: balanced accuracy against the correlation r between the source pole and the tested pole, for sisters and other matches. Random poles have no recorded r, so the right panel leaves them out. Each class has its own shape and colour: circles for sisters, squares for other matches and triangles for random poles. The classes appear in turn, sisters first, in both panels.

On sisters, the explanation mostly carries in part. The median balanced accuracy is 0.78 on sisters, 0.55 on other matches and 0.45 on random poles (Table 1). In Figure 1, left, sisters crowd between 0.75 and 0.95, and other matches spread from 0.4 to 0.9. The middle half of the sister studies lies between 0.59 and 0.86. Sisters are REJECTED 12 of 50 times (24%), and other matches 14 of 26 times (54%). That difference is significant (Fisher $p = 0.012$).

Most sister studies end NEEDS_REVISION: 30 of the 50. In these, the explanation passes many probes and misses some. In 13 of the 30, its balanced accuracy is 0.80 or more, and no sister study that ends NEEDS_REVISION scores below 0.60. One sister study scores 1.00 and still ends NEEDS_REVISION, because a gate other than the probes failed (Table 3). So the review's other gates, as well as the probes, keep sisters from verifying.

All 4 random poles are REJECTED, with balanced accuracies from 0.25 to 0.46. The random control thus behaves as expected, but it is only 4 studies. The interval for its share REJECTED runs from 51% to 100%. The contrast between sisters and random poles is therefore the least measured part of this paper.

Taken together, the results suggest that a sister often shares the core of its source's feature, but not every detail. The search finds poles where an explanation largely holds. A whole explanation seldom holds.

6 Does crossing tokenizers matter?

The third question is whether an explanation carries less well when the two models split text into different tokens. Many explanations name tokens or parts of words. A change of tokenizer could therefore break an explanation even when the two poles track the same feature. Table 2 splits each class by tokenizer.

Tested pole	Tokenizer	Studies	VERIFIED	REJECTED	Median balanced accuracy
Sister	the same (Qwen ↔ Qwen)	28	4	6	0.75
Sister	across (GPT-2 ↔ Qwen)	22	4	6	0.79
Other match	the same	13	1	6	0.64
Other match	across	13	0	8	0.53

Table 2. Studies within one tokenizer and across two, by class. The same tokenizer means Qwen2.5-0.5B and Qwen2.5-1.5B. Across means GPT-2 small and either Qwen2.5 model. The random poles are too few to split.

For sisters, the tokenizer matters little. 4 of 28 sisters verify within the Qwen family, which shares a tokenizer. 4 of 22 verify across tokenizers, GPT-2 to Qwen. Each group has 6 REJECTED, and their median balanced accuracies are 0.75 and 0.79. A sister found across tokenizers is as good a target as one within a family.

Other matches across tokenizers do worst. None of 13 verify, 8 are REJECTED, and their median balanced accuracy is 0.53. Within the Qwen family, other matches reach a median of 0.64. A plausible reading is that the mutual step filters out matches that rest on shared tokens. A sister across tokenizers must then track something that both tokenizations express. Each cell holds only 13 to 28 studies, though, so this reading is a hint and not a result.

7 Does a higher correlation predict transfer?

The fourth question is whether the correlation between two poles predicts that an explanation carries over. If the correlation measures a shared feature, studies with a higher r should verify more often. Of the 80 studies, 76 have a recorded r: the sisters and the other matches.

The median r is 0.49 for the 9 studies that verified, and 0.38 for the other 67. A higher correlation thus goes with an explanation that carries over. Figure 1, right, shows the same trend across all studies, and the spread around it. Five sisters with r below 0.4 score above 0.9. Three studies with r near 0.6, two sisters and one other match, are REJECTED.

This result rests on only 9 VERIFIED studies. It also mixes two effects. Sisters have a higher r than other matches, with a median of 0.42 against 0.29 among the studies here. So r may predict transfer only because it predicts the class. A larger set of studies within each class would separate the two.

8 Discussion

The results fit one picture. A mutual best match is a good target for an explanation, and a better one than another close match. On sisters, the median balanced accuracy is 0.78, and only a quarter are REJECTED. On other matches, the median is 0.55, and over half are REJECTED. Yet the whole explanation seldom passes: 8 of 50 sisters verify. This suggests that a sister and its source tend to share the core of a feature, but differ at its edges.

Why partial transfer is the common outcome. There are at least three readings, and the data do not choose between them. First, the two poles may encode the same feature with different boundaries. One model may treat a token as an instance of the feature and the other may not. The first sister studies give an example (sister neurons). A Qwen explanation of a separator detector passed every hard negative on its GPT-2 sister, but two of its holdouts failed. Second, a pole may have more than one mode, in the sense of Tigges et al. (2026), and only some modes may recur. An explanation that covers two modes would then pass the probes of the shared mode and fail those of the other. This would show as NEEDS_REVISION, which is the most common verdict on sisters. Third, the review is noisy, and an explanation near a gate can fail by chance (*Limitations*).

Relation to universality. Gurnee et al. (2024) find 1–5% of GPT-2 neurons universal across seeds of one model. Li et al. (2016) find that "some features are learned reliably in multiple networks, yet other features are not consistently learned". Gurnee et al. chose seeds of one architecture, trained on the same data, as "the most favorable experimental conditions for measuring universality". Our models differ far more than seeds do: they differ in family, size, data and tokenizer. That sisters exist across them at all, and that their explanations carry in part, extends the picture of partial universality to a harder case. It also adds a test that correlation alone does not give. A high correlation over one corpus does not imply that an explanation transfers. Three studies with r near 0.6 are REJECTED (Figure 1, right).

Relation to the neuron basis. Shabalin and Piotrowski (2026) find that features aligned with neurons recur across training runs of a dictionary. Tigges et al. (2026) find that a pole's two signs usually have unrelated roles. Both support matching poles, not whole neurons. The search agrees: 24 of 53 sisters between the two Qwen models join a positive pole to a negative one (sister neurons). Our results add that the neuron basis also supports some transfer of explanations across models, though seldom of a whole explanation.

Relation to depth. Faumi and Sumanathilaka (2026) show that layers correspond better by relative depth than by index. The sister search compares every layer, so it does not need this assumption. It nonetheless finds sisters at about the same relative depth. DistilGPT-2 would test the search on a pair of models that share a tokenizer and a width, and differ in depth (#3853). If distillation keeps early structure and loses late structure, sisters of GPT-2's VERIFIED poles should be common early and rare late.

Balanced accuracy suits these probe sets. The headline medians are balanced accuracies. Oikarinen et al. (2025) show that balanced accuracy can mislead on rare concepts, where it cannot tell an exact explanation from a too generic one. The reviewer's probe sets are not of that kind. Holdouts make up 30–37% of each set, well above the 10% at which Oikarinen et al. find that the score behaves. On these probe sets, balanced accuracy therefore separates explanations that carry from those that do not.

What this does not show. Whether two poles that share an explanation represent the same feature in the same way. A sister study tests one explanation's predictions on new probes. It does not compare the two poles' roles in their models.

9 Limitations

- **The random control is four studies.** The planner draws random poles at a low weight, so few ran. The contrast that matters most, sisters against random poles, rests on 4 controls. Another match is a weaker control, because it already correlates with the source.
- **Not a designed experiment.** The 80 studies are those that agents ran when the planner drew them. The class sizes, the models and the reviewers were not balanced in advance.
- **Mostly protocol 0.8.** 88 studies ran under protocol 0.8 and one under 0.9, with the agents that each draw had.
- **One reviewer per study.** First readings of one hypothesis agree on VERIFIED about 71% of the time (registry statistics), so a study's verdict is noisy. Yan et al. (2025) frame this as the stability of an explanation across probing sets. Each study here reads one probe set, so it gives no measure of that stability.
- **One corpus.** The search correlates poles over one corpus of 500 documents, with one random rotation per layer. A match on that corpus may not hold on other text.
- **Few cells.** The tokenizer split has 13 to 28 studies per cell, and the correlation result rests on 9 VERIFIED studies.

10 Conclusion

A VERIFIED explanation seldom describes a matched pole in another model in full. It verifies on 16% of sisters, 4% of other matches and none of 4 random poles. It carries in part much more often. On sisters, the median balanced accuracy is 0.78, and only 24% are REJECTED, against 54% of other matches. A sister across tokenizers is about as good a target as one within the Qwen family. So a mutual best match is a useful start for explaining a pole of a new model. Its explanation still needs a review of its own. Three studies would sharpen the result:

- About 30 random-pole sister studies, so that sisters can be compared with a real control.
- The same poles under one protocol and one agent pair.
- A study that compares what two sisters do, not only what fires them, for example by patching one model's pole and reading the other's.

References

Faumi, N., & Sumanathilaka, D. (2026). *What survives distillation? A mechanistic interpretability study of GPT-2 and DistilGPT-2*. Preprint. arXiv:2610.37157.

Gurnee, W., Horsley, T., Guo, Z. C., Kheirkhah, T. R., Sun, Q., Hathaway, W., Nanda, N., & Bertsimas, D. (2024). *Universal neurons in GPT2 language models*. Transactions on Machine Learning Research. OpenReview. arXiv:2401.12181.

Li, Y., Yosinski, J., Clune, J., Lipson, H., & Hopcroft, J. (2016). *Convergent learning: Do different neural networks learn the same representations?* ICLR 2016. arXiv:1511.07543.

Oikarinen, T., Yan, G., & Weng, T.-W. (2025). *Evaluating neuron explanations: A unified framework with sanity checks*. ICML 2025. arXiv:2506.05774.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI.

Shabalin, S., & Piotrowski, M. (2026). *Comparing sparsified MLPs and transcoders: Many transcoder features are neurons in disguise*. NeurIPS 2026 workshop on interpretability as a science. arXiv:2610.24521.

Tigges, C., Kowal, M., McGrath, T., Lewis, O., & Cammarata, N. (2026). *Neurons as an interpretable basis*. Goodfire Research. goodfire.com.

Yan, G., Oikarinen, T., & Weng, T.-W. (2025). *Faithful and stable neuron explanations for trustworthy mechanistic interpretability*. Mechanistic Interpretability Workshop at NeurIPS 2025. arXiv:2512.18092.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., . . . Qiu, Z. (2024). *Qwen2.5 technical report*. arXiv:2412.15115.

Appendices

A. Reproduce

The script reads the registry from a built site, so the first command builds it. It then reads every study that names a source pole, and assigns each tested pole its class from `docs/analyses/sister-neurons.jsonl`. No step draws random numbers. The rotated baseline of the sister search is seeded by model and layer, and its rows are already in that file.

```
uv run open-interpretability build --output-dir build/site
uv run python scripts/explanation_transfer.py --site-dir build/site \
    --out papers/explanations-across-models/results.jsonl
uv run python scripts/paper_figures.py
```

The classes depend on the sister search's rows and on the registry on the day the script runs. Each `study` row of `results.jsonl` names its experiment, so a reader can find each record in `wiki/experiments/`.

B. Every study

Table 3 lists the 80 tested poles, from `results.jsonl`. Each row is the first study of its tested pole. A pole is named by its model, its layer and its neuron, with + for the positive pole and − for the negative one. Qwen-0.5B and Qwen-1.5B stand for Qwen2.5-0.5B and Qwen2.5-1.5B. The column r is the correlation between the two poles that the sister search recorded. Random poles have none.

Tested pole	Class	Source pole	Tokenizer	r	Verdict	Balanced accuracy
Qwen-1.5B L16/N4458−	Sister	Qwen-0.5B L6/N1189+	same	0.33	NEEDS_REVISION	1.00
Qwen-1.5B L24/N6717−	Sister	Qwen-0.5B L21/N2536−	same	0.27	NEEDS_REVISION	0.93
Qwen-0.5B L1/N3190+	Sister	GPT-2 L3/N720+	cross	0.36	NEEDS_REVISION	0.93
Qwen-1.5B L19/N7272−	Sister	Qwen-0.5B L17/N1742+	same	0.52	NEEDS_REVISION	0.92
Qwen-1.5B L26/N6260−	Sister	Qwen-0.5B L22/N1068−	same	0.28	NEEDS_REVISION	0.92
Qwen-1.5B L19/N6120+	Sister	Qwen-0.5B L16/N1584−	same	0.37	NEEDS_REVISION	0.92
Qwen-0.5B L5/N3833+	Sister	GPT-2 L3/N1037+	cross	0.73	VERIFIED	0.89
Qwen-1.5B L1/N444−	Sister	GPT-2 L1/N1435+	cross	0.49	VERIFIED	0.89
GPT-2 L8/N1096+	Sister	Qwen-0.5B L15/N61−	cross	0.56	NEEDS_REVISION	0.88
GPT-2 L0/N2495+	Sister	Qwen-0.5B L0/N4325−	cross	0.77	NEEDS_REVISION	0.86
Qwen-0.5B L9/N4045−	Sister	Qwen-1.5B L13/N7836−	same	0.36	VERIFIED	0.86
Qwen-0.5B L20/N2392−	Sister	GPT-2 L7/N1101+	cross	0.47	NEEDS_REVISION	0.86
Qwen-1.5B L22/N7894+	Sister	Qwen-0.5B L16/N2128+	same	0.75	NEEDS_REVISION	0.86
GPT-2 L0/N754+	Sister	Qwen-0.5B L0/N3199+	cross	0.60	VERIFIED	0.85
Qwen-1.5B L5/N5356+	Sister	Qwen-0.5B L5/N151−	same	0.53	VERIFIED	0.85
Qwen-1.5B L21/N2033−	Sister	Qwen-0.5B L17/N4749−	same	0.77	VERIFIED	0.85
GPT-2 L6/N1314+	Sister	Qwen-0.5B L4/N2436−	cross	0.33	NEEDS_REVISION	0.83
Qwen-1.5B L0/N1370+	Sister	Qwen-0.5B L0/N869+	same	0.76	NEEDS_REVISION	0.82
Qwen-1.5B L27/N4346+	Sister	GPT-2 L8/N689+	cross	0.62	NEEDS_REVISION	0.81
Qwen-0.5B L3/N3573−	Sister	GPT-2 L1/N1435+	cross	0.55	NEEDS_REVISION	0.80
Qwen-1.5B L4/N6041+	Sister	Qwen-0.5B L5/N2106−	same	0.49	VERIFIED	0.80
Qwen-0.5B L7/N2313−	Sister	GPT-2 L6/N1828+	cross	0.45	NEEDS_REVISION	0.79
Qwen-0.5B L6/N794−	Sister	GPT-2 L6/N62+	cross	0.42	NEEDS_REVISION	0.79
Qwen-1.5B L6/N6675−	Sister	Qwen-0.5B L2/N414−	same	0.59	NEEDS_REVISION	0.79

Tested pole	Class	Source pole	Tokenizer	r	Verdict	Balanced accuracy
Qwen-1.5B L0/N935−	Sister	Qwen-0.5B L0/N3199+	same	0.79	NEEDS_REVISION	0.79
Qwen-1.5B L26/N884+	Sister	GPT-2 L9/N1357+	cross	0.28	VERIFIED	0.78
Qwen-1.5B L5/N4927−	Sister	Qwen-0.5B L5/N33−	same	0.39	NEEDS_REVISION	0.75
Qwen-1.5B L9/N6519+	Sister	Qwen-0.5B L11/N805+	same	0.20	NEEDS_REVISION	0.75
Qwen-1.5B L12/N4276+	Sister	GPT-2 L3/N2656+	cross	0.34	NEEDS_REVISION	0.67
Qwen-1.5B L2/N3369−	Sister	Qwen-0.5B L4/N429+	same	0.77	NEEDS_REVISION	0.67
Qwen-1.5B L20/N5477+	Sister	Qwen-0.5B L16/N671−	same	0.18	NEEDS_REVISION	0.67
Qwen-1.5B L17/N1567+	Sister	GPT-2 L5/N239+	cross	0.37	NEEDS_REVISION	0.64
Qwen-1.5B L12/N4721+	Sister	Qwen-0.5B L7/N3902+	same	0.39	NEEDS_REVISION	0.64
Qwen-1.5B L12/N5590+	Sister	Qwen-0.5B L11/N3129−	same	0.27	NEEDS_REVISION	0.64
Qwen-1.5B L4/N6041−	Sister	Qwen-0.5B L5/N2106+	same	0.23	NEEDS_REVISION	0.63
Qwen-1.5B L6/N8549+	Sister	Qwen-0.5B L4/N3740−	same	0.20	NEEDS_REVISION	0.63
Qwen-1.5B L2/N2198+	Sister	GPT-2 L3/N830+	cross	0.31	NEEDS_REVISION	0.60
Qwen-1.5B L20/N2500+	Sister	Qwen-0.5B L16/N4681+	same	0.36	NEEDS_REVISION	0.60
Qwen-1.5B L3/N1989−	Sister	Qwen-0.5B L4/N52−	same	0.43	REJECTED	0.57
Qwen-0.5B L9/N25+	Sister	GPT-2 L5/N428+	cross	0.49	REJECTED	0.53
Qwen-1.5B L23/N5941−	Sister	GPT-2 L7/N1101+	cross	0.47	REJECTED	0.52
Qwen-1.5B L7/N7802+	Sister	GPT-2 L8/N129+	cross	0.50	REJECTED	0.50
Qwen-1.5B L15/N8847+	Sister	Qwen-0.5B L9/N1458−	same	0.26	REJECTED	0.50
Qwen-1.5B L18/N8891−	Sister	Qwen-0.5B L15/N946−	same	0.28	REJECTED	0.49
Qwen-1.5B L25/N214+	Sister	Qwen-0.5B L21/N1368−	same	0.58	REJECTED	0.49
Qwen-0.5B L19/N1656+	Sister	GPT-2 L8/N129+	cross	0.40	REJECTED	0.46
Qwen-1.5B L25/N3013−	Sister	Qwen-0.5B L20/N3207−	same	0.60	REJECTED	0.46
Qwen-0.5B L3/N2347−	Sister	GPT-2 L6/N1711+	cross	0.35	REJECTED	0.43

Tested pole	Class	Source pole	Tokenizer	r	Verdict	Balanced accuracy
Qwen-1.5B L4/N6256+	Sister	Qwen-0.5B L3/N3824+	same	0.27	REJECTED	0.43
Qwen-1.5B L0/N2596+	Sister	GPT-2 L1/N493+	cross	0.42	REJECTED	0.42
Qwen-1.5B L23/N7919−	Other match	Qwen-0.5B L20/N3989−	same	0.24	NEEDS_REVISION	0.93
Qwen-1.5B L17/N1630−	Other match	Qwen-0.5B L17/N2515−	same	0.29	VERIFIED	0.86
GPT-2 L11/N2215+	Other match	Qwen-0.5B L20/N4378−	cross	0.43	NEEDS_REVISION	0.82
GPT-2 L10/N657+	Other match	Qwen-0.5B L23/N3660+	cross	0.58	NEEDS_REVISION	0.78
Qwen-1.5B L25/N5907+	Other match	GPT-2 L4/N3028+	cross	0.29	NEEDS_REVISION	0.77
Qwen-1.5B L27/N1526−	Other match	Qwen-0.5B L23/N3967−	same	0.14	NEEDS_REVISION	0.75
Qwen-1.5B L18/N7662−	Other match	Qwen-0.5B L13/N3426+	same	0.22	NEEDS_REVISION	0.71
Qwen-1.5B L9/N7242+	Other match	Qwen-0.5B L16/N2641−	same	0.24	NEEDS_REVISION	0.68
Qwen-1.5B L4/N187−	Other match	Qwen-0.5B L7/N4652−	same	0.27	NEEDS_REVISION	0.65
Qwen-0.5B L23/N2135+	Other match	GPT-2 L11/N438+	cross	0.52	NEEDS_REVISION	0.64
Qwen-1.5B L23/N6981+	Other match	Qwen-0.5B L22/N337+	same	0.31	NEEDS_REVISION	0.64
GPT-2 L6/N2173+	Other match	Qwen-0.5B L5/N3833+	cross	0.51	NEEDS_REVISION	0.64
Qwen-1.5B L27/N5611+	Other match	GPT-2 L11/N830+	cross	0.43	REJECTED	0.57
Qwen-1.5B L17/N5841+	Other match	Qwen-0.5B L12/N2218−	same	0.18	REJECTED	0.54
Qwen-0.5B L17/N177+	Other match	GPT-2 L0/N1310+	cross	0.28	REJECTED	0.53
Qwen-0.5B L16/N4004−	Other match	GPT-2 L8/N1343+	cross	0.42	REJECTED	0.50
Qwen-1.5B L25/N3445−	Other match	Qwen-0.5B L12/N4099−	same	0.38	REJECTED	0.50
Qwen-1.5B L27/N6169−	Other match	GPT-2 L10/N849+	cross	0.43	REJECTED	0.50
Qwen-1.5B L27/N6147−	Other match	GPT-2 L11/N438+	cross	0.60	REJECTED	0.50
Qwen-0.5B L3/N1961+	Other match	Qwen-1.5B L6/N497−	same	0.14	REJECTED	0.47
Qwen-0.5B L23/N2817+	Other match	GPT-2 L11/N2411+	cross	0.26	REJECTED	0.46
Qwen-1.5B L26/N7101+	Other match	Qwen-0.5B L21/N953−	same	0.47	REJECTED	0.46

Tested pole	Class	Source pole	Tokenizer	r	Verdict	Balanced accuracy
Qwen-1.5B L10/N4725−	Other match	Qwen-0.5B L9/N618−	same	0.13	REJECTED	0.46
Qwen-1.5B L13/N7895+	Other match	GPT-2 L4/N2930+	cross	0.44	REJECTED	0.46
GPT-2 L2/N2673+	Other match	Qwen-1.5B L5/N5356+	cross	0.20	REJECTED	0.45
Qwen-1.5B L23/N6314−	Other match	Qwen-0.5B L22/N2126+	same	0.29	REJECTED	0.42
Qwen-1.5B L24/N3755−	Random	Qwen-0.5B L16/N706+	same	—	REJECTED	0.46
Qwen-0.5B L21/N1004+	Random	GPT-2 L3/N1242+	cross	—	REJECTED	0.46
Qwen-1.5B L17/N8865−	Random	GPT-2 L11/N2133+	cross	—	REJECTED	0.44
GPT-2 L4/N941+	Random	Qwen-0.5B L8/N4149−	cross	—	REJECTED	0.25

Table 3. The 80 tested poles, each by its first sister study, sorted by class and then by balanced accuracy. The tokenizer is *same* for two Qwen2.5 models and *cross* when one model is GPT-2 small.