REASONING ACROSS LANGUAGES: ACTIVATION AGREEMENT AND THE LIMITS OF LANGUAGE HEURISTICS

Junjie Nian
College of Computer Science and Artificial Intelligence
Fudan University
Shanghai, China
github.com/JunjieNian/multilingual_neuron_analysis

ABSTRACT

Multilingual reasoning evaluations usually compare final-answer accuracy by language, leaving open a harder test-time question: when the same problem has candidate trajectories in many languages, can internal model structure select the stronger trajectory without seeing its answer? We study 2,250 mathematical reasoning trajectories arranged as 125 PolyMath-Top problems in 18 languages. Sparse activated-neuron sets define pairwise Jaccard geometry, and thirteen selectors choose one trajectory per problem using activation count, centrality, consensus, or language-aware heuristics. Accuracy varies from 11.2% to 22.4% across languages in this run, while English reaches 16.8%, showing that no fixed language is an obvious default. The strongest observed selector, activation-based ConsensusMin, solves 32 of 125 problems (25.6%), compared with 26 (20.8%) for a fixed-first baseline and 28 (22.4%) for the best single language. Medoid centrality and language diversity each reach 24.8%, whereas cross-language consistency reaches 16.0% and a low-entropy preference 13.6%. Auxiliary analyses associate moderate output-language mixing with higher accuracy and reveal broad cross-language activation overlap, but these diagnostics are not causal and are limited by script-level language detection. The central conclusion is therefore modest: activation agreement offers useful information beyond a fixed language choice, while language identity, apparent consistency, and monolinguality are unreliable proxies for correctness. We release the analysis code, paper source, and compact documentation, and explicitly separate auditable results from exploratory observations whose large activation caches cannot be redistributed.

1 INTRODUCTION

Chain-of-thought reasoning exposes intermediate text that can improve difficult problem solving (Wei et al., 2022), yet most studies remain English-centric. Multilingual evaluations show that reasoning quality, output-language consistency, and trace length can change substantially with the prompt language (Wang et al., 2025; Chen et al., 2024). Cross-lingual prompting can sometimes recover performance by moving information or reasoning through a stronger language (Qin et al., 2023). These results establish that language matters, but they do not tell us how to select among multiple trajectories for the same underlying problem.

Best-of-N decoding normally compares samples in one language through answer voting, likelihood, or a learned verifier. Neuron Agreement Decoding (NAD) instead represents each response through its sparse activated-neuron set and selects candidates that are both economical and consistent with other samples (Chen et al., 2025). Its key premise is that correct reasoning trajectories tend to occupy a more coherent activation region than divergent mistakes. This raises a natural multilingual question: does activation agreement remain useful when the candidates are not paraphrases in one language, but solutions expressed across 18 languages and writing systems?

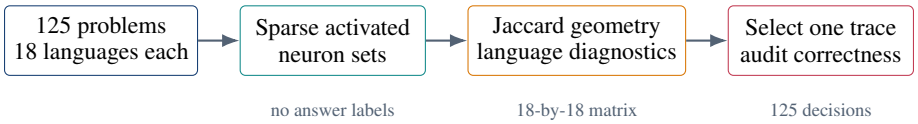

Figure 1: Experimental design. Correctness labels enter only after each selector has chosen one trajectory per problem.

We answer this question with a controlled cross-language selection experiment. For each of 125 PolyMath-Top problems, we gather one trajectory in each of 18 languages and ask a selector to choose exactly one. The selectors do not observe correctness at decision time. Eight reuse activation-space strategies from NAD, while five test common multilingual intuitions: prefer historically accurate languages, maximize language diversity, seek cross-language consistency, prefer low language entropy, or combine language and activation signals.

The experiment has two main lessons. First, accuracy differs by language but language identity is not a stable shortcut. English is not best in this run, and the historically strongest-language selector does not match the strongest activation-based selector. Second, consensus and centrality are useful but imperfect. ConsensusMin yields the best observed accuracy, yet the evaluation has only 125 independent problems, and even a central trajectory can be a shared mistake. Our contribution is therefore an audited empirical comparison and a set of failure boundaries, not a claim that activation geometry verifies mathematical truth.

2 Related Work

Multilingual mathematical reasoning. PolyMath provides 9,000 carefully translated mathematical questions covering 18 languages and four difficulty levels (Wang et al., 2025). Its original evaluation finds large cross-language performance variation and imperfect agreement between input and output language. Related work shows that multilingual mathematical reasoning benefits from cross-lingual transfer but remains sensitive to prompting, data balance, and language consistency (Qin et al., 2023; Chen et al., 2024). Our work differs in task: instead of scoring one response per language, we select one response from a multilingual candidate set.

Internal signals for reasoning selection. NAD finds that correct trajectories activate fewer unique neurons and show stronger cross-sample activation agreement, enabling unsupervised best-of-N selection (Chen et al., 2025). We preserve its activation-set representation and centrality intuition but change the candidate distribution. Cross-language candidates can differ lexically and by script even when they implement the same mathematical plan, providing a stress test for whether the internal geometry captures more than surface-form similarity.

Language mixing. Reasoning models often drift toward a dominant language during long reasoning. Script switching can reflect transfer, generation habits, or tokenization effects rather than a deliberate cognitive strategy. We use Shannon entropy (Shannon, 1948) only as a descriptive measure of detected-script mixture. We do not interpret output code-switching as a causal intervention.

3 Data and Method

3.1 Cross-language problem groups

The primary evaluation uses the Top difficulty slice of PolyMath: 125 underlying problems, each represented in 18 languages, for 2,250 trajectories total. Problem identifiers follow `top-{language}-{number}`. The activation cache stores sparse neuron keys for each complete model run; a separate evaluation file supplies binary answer correctness.

The 125 problem groups are the independent selection units. Language-level percentages contain 125 Bernoulli outcomes each. Selector accuracy is the fraction of problem groups in which the

chosen trajectory is correct. Because all selectors operate on the same 125 groups, a difference of one selected answer corresponds to 0.8 percentage points.

3.2 Activation geometry

Let K_i be the activated-neuron set for candidate i in a problem group. Pairwise similarity and distance are

$$s(i,j) = \frac{|K_i \cap K_j|}{|K_i \cup K_j|}, \qquad d(i,j) = 1 - s(i,j). \tag{1}$$

The full 18×18 distance matrix supports medoid, neighborhood, density-cluster, and consensus selection. Activation count is recorded separately. The representation is language-agnostic in form, but not guaranteed language-invariant in value.

3.3 Selectors

Eight built-in selectors test different activation priors. *MinActivation* and *MaxActivation* choose extrema of activation count. *Medoid* minimizes mean distance to all candidates; *KNNMedoid* restricts centrality to a local neighborhood; *DBSCANMedoid* selects the medoid of a density cluster. *ConsensusMin* and *ConsensusMax* combine activation-count and centrality preferences. The fixed-first baseline always selects the first candidate in group order, which is Arabic in this cache.

Five additional selectors test language-aware intuitions. *HighAccuracyLang* favors languages with stronger aggregate history; *LanguageDiversity* and *CrossLangConsistency* favor broad or stable cross-language structure; *LowEntropyPref* favors the lowest detected language mixture; and *Hybrid-Multilingual* combines language preference with activation consensus. These variants are exploratory because their design was informed by the same research setting.

3.4 Auxiliary language diagnostics

Character ranges identify Arabic, Bengali, Chinese, Japanese, Korean, Cyrillic, Thai, Telugu, and Latin-script characters in text enclosed by `<think>` tags. For proportions p_ℓ over detected script groups, mixing entropy is

$$H = -\sum_{\ell} p_\ell \log_2 p_\ell. \tag{2}$$

This detector is reproducible but coarse. It cannot reliably distinguish English from Spanish, French, German, Italian, Portuguese, Indonesian, Malay, Swahili, or Vietnamese when unaccented Latin characters dominate. We therefore refer to *detected-script mixing* when precision matters.

4 Results

4.1 Language accuracy varies by more than twofold

Table 1 reports every language. Accuracy ranges from 11.2% to 22.4%, with 417 correct trajectories overall (18.5%). Korean and Russian tie at the top; Swahili and Telugu are lowest. English reaches 16.8% and ranks below ten other languages in this particular run.

Table 1: Final-answer accuracy by input language. Each language contains the same 125 underlying problems.

Language	Correct	Accuracy	Language	Correct	Accuracy
Korean	28	22.4%	Vietnamese	24	19.2%
Russian	28	22.4%	Bengali	22	17.6%
Italian	27	21.6%	Indonesian	22	17.6%
Spanish	27	21.6%	English	21	16.8%
Arabic	26	20.8%	Thai	21	16.8%
Chinese	26	20.8%	German	20	16.0%
French	25	20.0%	Japanese	20	16.0%
Malay	25	20.0%	Telugu	16	12.8%
Portuguese	25	20.0%	Swahili	14	11.2%

These ranks should not be read as properties of the languages. They conflate model pretraining, tokenization, prompting, generation success, and sampling noise. Their relevance to selection is narrower: no fixed language dominates enough to make trajectory comparison unnecessary.

4.2 Consensus and centrality outperform simple heuristics

ConsensusMin selects 32 correct trajectories (25.6%), the strongest observed result (Table 2). It gains six correct problems over the fixed-first baseline and four over the best fixed language. Medoid and LanguageDiversity each select 31 correct trajectories; their tie illustrates that a language-aware construction need not add information beyond activation centrality. HybridMultilingual and HighAccuracyLang follow at 24.0% and 23.2%.

Table 2: One-of-18 trajectory selection on 125 problem groups. The "signal" column identifies the primary design prior.

Selector	Signal	Correct	Accuracy
ConsensusMin	activation consensus	**32**	**25.6%**
Medoid	global activation centrality	31	24.8%
LanguageDiversity	language-aware diversity	31	24.8%
HybridMultilingual	activation + language	30	24.0%
HighAccuracyLang	historical language accuracy	29	23.2%
DBSCANMedoid	density-cluster centrality	26	20.8%
Fixed-first baseline	fixed Arabic candidate	26	20.8%
KNNMedoid	local activation centrality	23	18.4%
ConsensusMax	activation consensus	22	17.6%
MinActivation	activation count	20	16.0%
CrossLangConsistency	cross-language consistency	20	16.0%
LowEntropyPref	low detected-script entropy	17	13.6%
MaxActivation	activation count	15	12.0%

The results also reject three tempting simplifications. First, activation count alone is not a stable quality score: the minimum and maximum selectors reach 16.0% and 12.0%. Second, local centrality is weaker than global medoid centrality in this setting. Third, cross-language agreement is not synonymous with correctness; multiple candidates can converge on the same incorrect plan.

4.3 Language-aware design does not reliably improve selection

Across the eight built-in selectors, mean accuracy is 19.5% with a wide range from 12.0% to 25.6%. The five language-aware selectors average 20.3% and range from 13.6% to 24.8%. Category means are not meaningful significance tests: methods are heterogeneous, share the same problem groups, and were not preregistered. The more useful comparison is qualitative. Language diversity can match a strong centrality baseline, but neither history, consistency, nor low entropy displaces activation consensus.

4.4 Moderate mixing is an exploratory correlate

Auxiliary analyses split completed traces into low, middle, and high detected-script entropy groups. In two evaluation runs, the middle group has the highest observed accuracy (Table 3). The same inverted-U shape appears with a separate code-switch score. This recurrence is suggestive, but the analysis is conditional on successful generation, uses post hoc terciles, and has weak identification for Latin-script languages.

Table 3: Exploratory accuracy among completed traces by detected-script entropy group. These descriptive rates come from two evaluation runs and are not directly comparable to Table 1.

Evaluation run	Low	Middle	High
MUI cache	27.2%	**39.1%**	29.6%
R1–Qwen3 cache	7.3%	**12.3%**	8.5%

The table supports a practical warning rather than a prescription: a low-entropy or monolinguality preference is not justified by these data. It does not establish that encouraging code-switching would improve accuracy.

5 Discussion

5.1 What activation agreement contributes

Language identity is a coarse property shared by every trajectory written in that language. Activation agreement is problem-conditional: it asks which candidate lies near the common internal structure induced by the current problem. This difference explains why a consensus selector can outperform a fixed language choice even when language-level performance varies substantially.

The strongest selectors combine centrality with an economy prior rather than using either alone. A representative trajectory need not have the fewest activated neurons, and the smallest activation set can omit relevant computation. Conversely, a dense trajectory can reflect drift. ConsensusMin operationalizes a balance between these signals, consistent with NAD's broader finding that sparse activation and cross-sample agreement are complementary (Chen et al., 2025).

5.2 Why agreement remains insufficient

Activation centrality measures typicality, not mathematical validity. If most multilingual trajectories take the same wrong shortcut, their shared mistake can form the medoid. This is visible in the modest absolute accuracy: even the best selector is wrong on 93 of 125 problems. The method should therefore be paired with answer verification, execution, or external evidence when available.

Language also affects the representation itself. Script, tokenization, output length, and generation truncation can change which neurons are included in a full-trajectory set. A high Jaccard similarity can mix mathematical and surface-form effects. Controlled translations and within-step alignment would be needed to separate them.

5.3 Reproducibility and claim tiers

The public repository contains all selection scripts, configuration templates, methodological notes, and the manuscript source. Large activation caches and evaluation dumps are excluded. The paper therefore distinguishes three evidence tiers:

1. exact tables backed by compact saved JSON from the original analysis;
2. rerunnable scripts that require private or large caches;
3. exploratory observations documented in the repository but lacking compact public per-neuron outputs.

The 99.8% universality claim for a top-500 quality-neuron set belongs to the third tier and is intentionally not a headline result here. The public evidence supports broad activation overlap; it does not support presenting that exact percentage as independently auditable.

6 LIMITATIONS

Statistical power. There are only 125 independent problem groups. A six-answer improvement over baseline is meaningful enough to motivate replication, but too small for a broad claim of selector dominance. Future work should report paired bootstrap intervals and test on all PolyMath difficulty levels.

Model and data scope. The results come from one activation-cache pipeline and one model configuration. They may change with architecture, scale, instruction tuning, sampling temperature, or tokenizer. PolyMath translation quality is high, but translated mathematics does not cover culturally grounded multilingual reasoning.

Detector validity. The language-mixing measure is script-based. It detects Chinese, Arabic, Cyrillic, and several other scripts reasonably directly, but conflates unaccented Latin-script languages with English. The entropy analysis is therefore weakest exactly where code-switching can be most subtle.

Post hoc design. The language-aware selectors and several auxiliary analyses were designed after observing the same setting. Their performance is descriptive and susceptible to researcher degrees of freedom. A clean follow-up should freeze the selector family before evaluating new models and problems.

Representation. Full-trajectory activation sets discard temporal order. Two solutions can activate similar neurons in a different sequence. Slice-level or step-aligned representations could reveal whether a candidate returns to the same mathematical state for the same reason.

7 CONCLUSION

Across 18 languages, no fixed language, consistency heuristic, or monolinguality preference provides a dependable shortcut to correct reasoning. Activation agreement contributes a more problem-specific signal: the strongest consensus selector reaches 25.6%, compared with 20.8% for a fixed-first baseline and 22.4% for the best single language. The gain is promising but small, and centrality cannot certify truth. The professional use of multilingual internal signals therefore lies in ranking and diagnosis, not in replacing verification. Larger preregistered evaluations, compact auditable artifacts, and temporally aligned activation representations are the next steps.

REFERENCES

Kang Chen, Yaoning Wang, Kai Xiong, Zhuoka Feng, Wenhe Sun, Haotian Chen, and Yixin Cao. Do LLMs signal when they're right? evidence from neuron agreement. *arXiv preprint arXiv:2510.26277*, 2025. URL https://arxiv.org/abs/2510.26277.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://aclanthology.org/2024.findings-emnlp.411/.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2695–2709, 2023. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163/.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, and Jingren Zhou. PolyMath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*, 2025. URL https://arxiv.org/abs/2504.18428.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

A Activation Diagnostics

Figure 2 summarizes language-level activation count, overlap, clustering, and specificity generated by the repository's activation-analysis pipeline. It is included as a diagnostic map, not as an inferential result: several panels aggregate completed and truncated generations differently, and the large underlying cache is not distributed with the repository.

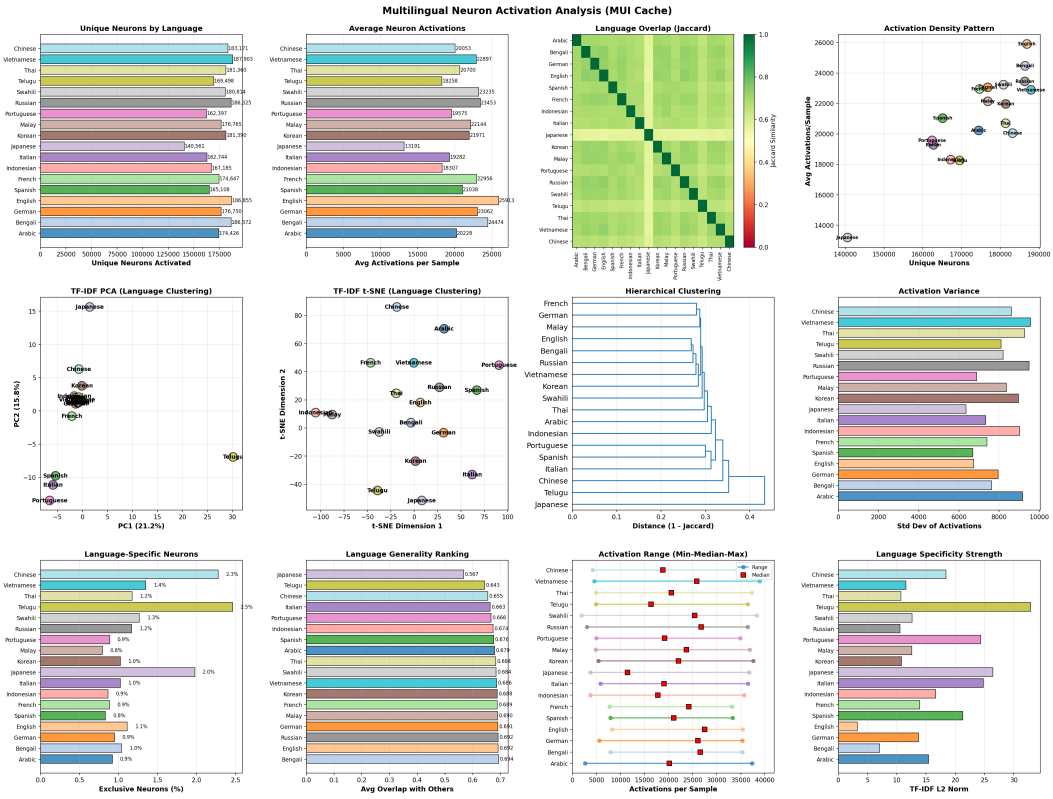

Figure 2: Exploratory multilingual activation diagnostics. Most languages show high average overlap with the others, while Japanese and Telugu form clearer deviations in several aggregate views.

B LLM Usage Disclosure

AI writing and coding assistants were used for repository restructuring, prose revision, LaTeX authoring, and consistency checks. The experiment design, scripts, saved numerical results, and interpretation originate from the author's project. During preparation of this manuscript, the author reconciled conflicting legacy documentation against the compact saved result files and retained the more conservative, auditable values.