CRANE: FUNCTIONAL SELECTIVITY OF LANGUAGE-CONDITIONED NEURONS IN MULTILINGUAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual large language models operate across languages within a shared parameter space, yet it remains unclear how functionally selective language-related internal components actually are. Prior work has mainly identified language-related neurons through activation statistics, but activation selectivity does not directly reveal how these components functionally affect different languages. We propose **CRANE**, a functional interpretability framework that combines relevance attribution with controlled neuron interventions. CRANE uses AttnLRP to estimate the contribution of MLP components to language-conditioned predictions, separates candidate identification from functional validation, and evaluates target and non-target language behavior under matched neuron-masking budgets to characterize their **functional selectivity**. We evaluate CRANE on LLaMA2-7B Base across English, Chinese, and Vietnamese using natural language understanding and open-ended generation tasks. Across all evaluated settings, CRANE-identified components exhibit stronger target-language functional effects than those selected by the activation-based LAPE baseline. However, these effects are not strictly confined to the target language, and some non-target languages are also substantially affected. The results indicate that language-related internal components exhibit measurable target-language functional biases, but these biases overlap with cross-lingual functional effects rather than forming fully isolated language-specific modules. We further transfer Base-identified components directly to LLaMA2-7B Chat and observe that part of their functional influence persists after instruction tuning. Overall, CRANE provides an interpretability perspective that moves from **correlational identification toward functional selectivity analysis**.

1 INTRODUCTION

Large language models (LLMs) exhibit understanding and generation capabilities across multiple languages within a shared parameter space. Prior work has identified internal components whose activation patterns vary systematically across languages and shown that manipulating them can alter language behavior (Tang et al., 2024; Gurgurov et al., 2025). However, activation selectivity does not necessarily imply functional selectivity: a component associated with one language may have little functional importance, while intervening on it may also affect other languages. This motivates our central question: *when an internal component is identified as language-related, how selectively does it functionally affect its associated language?*

We study this question through the lens of **functional selectivity**. For a candidate component set, we compare the behavioral effects of the same intervention on its target and non-target languages. The target-language change characterizes its **target-language functional influence**, while comparison with collateral non-target effects characterizes its **cross-language functional selectivity**. This distinction is important because prior intervention-based analyses suggest that language-related information can remain intertwined with shared representations (Dumas et al., 2025). Thus, target-language degradation alone is insufficient to establish language-selective functional influence. Our goal is not to infer a universal organization of multilingual LLMs, but to analyze functional selectivity in a controlled multilingual setting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

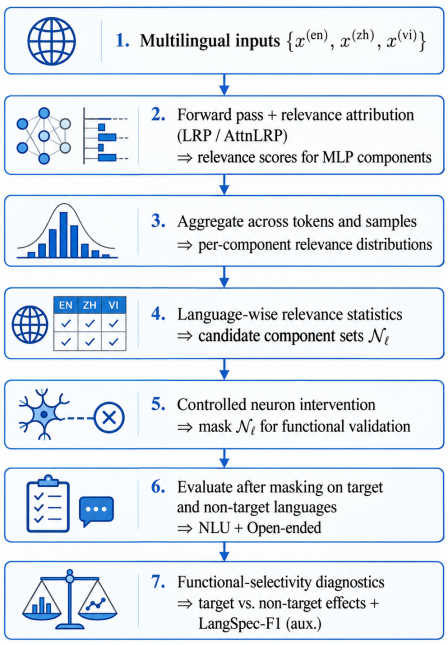

Figure 1: **CRANE separates candidate discovery from functional evaluation.** Relevance attribution identifies language-conditioned MLP components; controlled zero-masking then measures target-language influence and collateral effects on non-target languages.

To this end, we introduce **CRANE**, which explicitly separates candidate discovery from functional validation. CRANE uses layer-wise relevance propagation (LRP) (Arras et al., 2017a;b) and its Transformer extension AttnLRP (Achtibat et al., 2024) to estimate the contribution of MLP components to language-conditioned predictions and construct candidate sets from their relevance distributions. It then zero-masks the selected components under matched intervention budgets and evaluates behavioral changes on both target and non-target languages. Attribution therefore determines *which components to inspect*, while intervention determines *how selectively they affect behavior*.

We evaluate CRANE on LLAMA2-7B-BASE across English, Chinese, and Vietnamese using natural language understanding and open-ended generation tasks. Across all six language-by-evaluation settings, CRANE-selected components produce larger target-language degradation than those selected by the activation-based LAPE baseline (Tang et al., 2024). However, under the stricter criterion that target-language degradation must exceed that of every non-target language, only four of the six settings satisfy functional selectivity. Thus, language-related components exhibit measurable **target-language functional biases**, but these effects are not strictly language-exclusive.

We further transfer Base-identified component sets directly to LLAMA2-7B-CHAT without re-identification. Part of their functional influence remains observable after instruction tuning, without implying invariance of individual neuron identities across model stages.

Our contributions are threefold:

- We introduce a **functional selectivity** perspective that compares target- and non-target-language effects under the same intervention.
- We propose **CRANE**, combining AttnLRP-based candidate discovery with controlled zero-masking to separate candidate identification from functional validation.
- Across English, Chinese, and Vietnamese on LLAMA2-7B, we observe measurable but non-exclusive target-language functional biases, with part of the functional influence remaining observable from Base to Chat without re-identification.

2 Related Work

2.1 Language-Related Neurons and Multilingual Interventions

Prior work has identified and manipulated language-related internal components in multilingual LLMs. LAPE detects language-specific neurons from cross-language activation probabilities and shows that activating or suppressing them can alter output language (Tang et al., 2024). Language Arithmetics extends this analysis across multiple models and 21 languages, reporting overlap among language-associated neuron sets (Gurgurov et al., 2025), while related work studies the downstream effects of amplifying language-specific neurons across languages and models (Rahmanisa et al., 2025). These studies establish that language-related components can be identified and behaviorally manipulated. Our focus is narrower: *when a component set is associated with a target language, how selectively is its functional effect expressed on that language relative to non-target languages?* This question is motivated by evidence that language-related structure can overlap across languages. Zhang et al. distinguish language-specific, language-related, and general neurons, including components shared by subsets of languages (Zhang et al., 2026), while activation-patching results suggest that language-related and conceptual information can remain partially intertwined within shared representations (Dumas et al., 2025). CRANE therefore does not assume language exclusivity, but evaluates it directly through matched target-versus-non-target intervention effects. We additionally use a fixed-set Base-to-Chat transfer protocol to test whether functional effects identified in the Base model remain observable after instruction tuning without re-identifying components.

2.2 Attribution-Based Discovery and Intervention-Based Validation

Feature-attribution methods estimate how internal representations contribute to model predictions. Layer-wise relevance propagation (LRP) propagates output relevance backward through a network (Arras et al., 2017a;b), and AttnLRP extends this framework to Transformer architectures (Achtibat et al., 2024). CRANE uses AttnLRP only for candidate discovery: a high attribution score indicates association with a language-conditioned prediction, but does not by itself establish a functional effect. This distinction is closely related to intervention-based mechanistic interpretability, where methods such as causal tracing and activation patching modify internal states and measure the resulting behavioral changes (Meng et al., 2022; Zhang & Nanda, 2024). CRANE applies this separation to language-related functional selectivity: attribution determines *which components to inspect*, while controlled zero-masking determines *how those components affect target and non-target languages*. Functional selectivity is therefore characterized by relative behavioral effects under intervention, rather than by the attribution scores used to identify candidates.

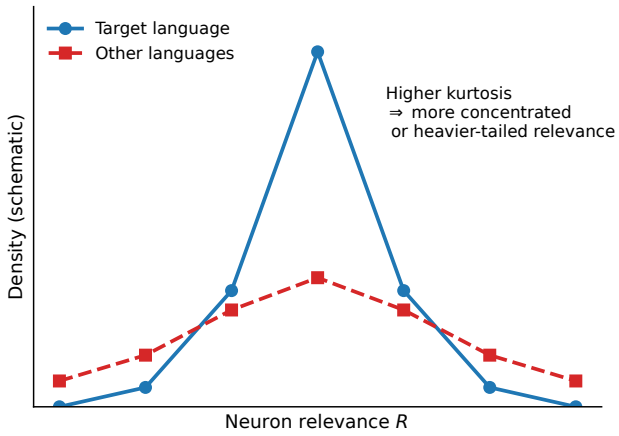

Figure 2: Schematic intuition for kurtosis-based candidate discovery. Kurtosis is used as a concentration heuristic, not as evidence of functional specificity.

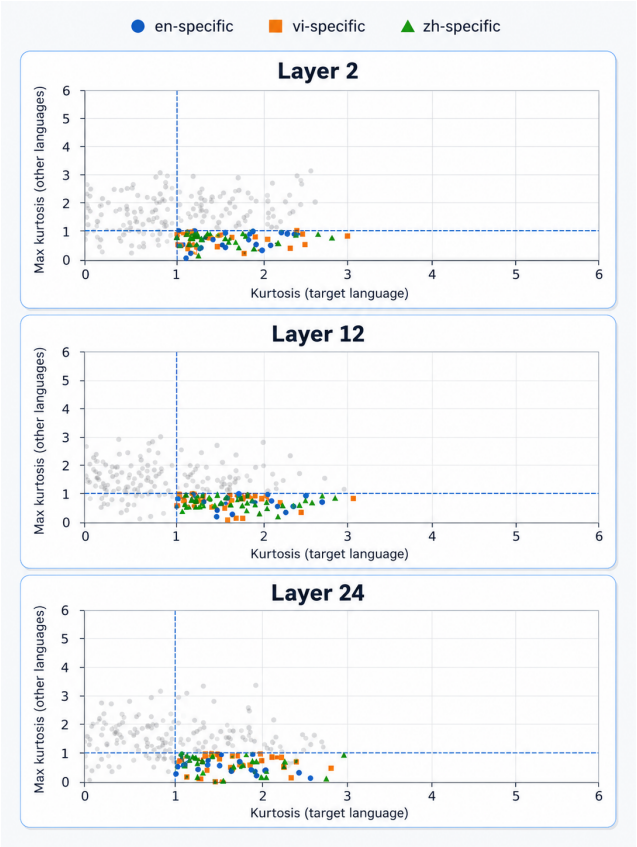

Figure 3: Diagnostic target-language versus maximum non-target normalized kurtosis. Candidate selection uses the target-language statistic; cross-language functional selectivity is evaluated only after intervention.

3 METHOD

CRANE separates the analysis of language-related internal components into **candidate discovery** and **functional validation**. Given a target language, AttnLRP first identifies candidate MLP components from language-conditioned relevance distributions; controlled zero-masking then measures their behavioral effects on target and non-target languages. Figure 1 illustrates the workflow.

3.1 PROBLEM SETUP AND FUNCTIONAL SELECTIVITY

Let $\mathcal{L} = \{\ell_1, \ldots, \ell_K\}$ denote the evaluated languages. For each target language ℓ, CRANE constructs a candidate component set $\mathcal{N}_\ell$. Let S_j^{orig} be the original performance on evaluation language j and $S_{j|\mathcal{N}_\ell}^{\text{mask}}$ the performance after masking $\mathcal{N}_\ell$. We define the non-negative intervention effect as

$$\Delta_{\ell \to j} = \max\left(S_j^{\text{orig}} - S_{j|\mathcal{N}_\ell}^{\text{mask}}, 0\right).$$

Here, $\Delta_{\ell \to \ell}$ measures **target-language functional influence**, while $\Delta_{\ell \to j}$ for $j \neq \ell$ captures non-target effects. We use

$$\Delta_{\ell \to \ell} > \max_{j \neq \ell} \Delta_{\ell \to j}$$

as a stricter diagnostic of **cross-language functional selectivity**. It asks whether the target-language effect exceeds every non-target effect, without requiring non-target languages to remain unaffected. Because the evaluated benchmarks are not fully parallel, this criterion is interpreted within the evaluated benchmark suite rather than as complete separation of language and task effects.

3.2 Relevance-Based Candidate Discovery

For an input x in language ℓ, CRANE uses layer-wise relevance propagation (LRP) (Arras et al., 2017a;b) and AttnLRP (Achtibat et al., 2024) to attribute a language-conditioned objective $f_\ell(x)$ to MLP components. Relevance is propagated backward according to the conservation principle

$$\sum R^{(l)} = \sum R^{(l-1)}.$$

Token-level relevance is aggregated to obtain a sample-level relevance $r_n(x, \ell)$ for component n.

For each component and language, these values form a relevance distribution

$$\mathcal{R}_{n,\ell} = \{r_n(x, \ell) \mid x \in D_\ell\}.$$

We summarize its concentration using kurtosis (DeCarlo, 1997):

$$\kappa_{n,\ell} = \frac{\mathbb{E}\left[(R_{n,\ell} - \mu_{n,\ell})^4\right]}{\sigma_{n,\ell}^4}.$$

Kurtosis is sensitive to concentrated, high-relevance events and does not require discretizing signed relevance values.

Kurtosis scores are normalized within each Transformer layer to obtain $\tilde{\kappa}_{n,\ell}$. For each target language, a fixed normalized-kurtosis threshold is used to construct the candidate set $\mathcal{N}_\ell$. These relevance statistics are used only for candidate discovery; functional selectivity is determined by the intervention effects below.

3.3 Controlled Intervention and Functional Validation

For each target language ℓ, we zero the outputs of components in $\mathcal{N}_\ell$ during inference and evaluate the same masked model on all languages, producing the intervention profile

$$\{\Delta_{\ell \to j}\}_{j \in \mathcal{L}}.$$

CRANE, LAPE (Tang et al., 2024), and random masking are evaluated under matched intervention budgets. For a given target language, each method masks the same number of components; the random control samples an equally sized set from the corresponding module structure.

Candidate discovery and functional validation are therefore explicitly separated: relevance determines *which components to inspect*, while zero-masking determines *how they affect behavior across languages*.

3.4 Functional-Selectivity Diagnostics

Our primary evidence consists of the target- and non-target-language intervention effects $\Delta_{\ell \to j}$ and the strict criterion defined in Section 3.1.

We additionally report LangSpec-F1 as an auxiliary summary metric. Let

$$\Delta_\ell = \Delta_{\ell \to \ell}, \qquad \Delta_\ell^{\text{non}} = \max_{j \neq \ell} \Delta_{\ell \to j}.$$

We define

$$\text{Precision}_\ell = \frac{\Delta_\ell}{\Delta_\ell + \Delta_\ell^{\text{non}} + \epsilon}, \qquad \text{Recall}_\ell = \frac{\Delta_\ell}{S_\ell^{\text{orig}} + \epsilon},$$

and

$$\text{LangSpecF1}_\ell = \frac{2\,\text{Precision}_\ell\,\text{Recall}_\ell}{\text{Precision}_\ell + \text{Recall}_\ell + \epsilon}.$$

LangSpec-F1 summarizes target degradation and worst-case non-target interference, but is not used as the definition of functional selectivity.

3.5 FIXED-SET BASE-TO-CHAT TRANSFER

To examine whether Base-identified functional effects remain observable after instruction tuning, candidate discovery is performed only on LLAMA2-7B-BASE, yielding $\mathcal{N}_\ell^{\mathrm{Base}}$. The same component positions are then zero-masked directly in LLAMA2-7B-CHAT without recomputing relevance or re-identifying candidates.

We measure the resulting Chat-model effects $\Delta_{\ell\to j}^{\mathrm{Chat}}$ using the same evaluation protocol. This fixed-set transfer tests the *persistence of functional influence*, not invariance of individual neuron identities across model stages.

4 EXPERIMENTAL SETUP

4.1 MODELS AND LANGUAGES

We evaluate CRANE on English (en), Chinese (zh), and Vietnamese (vi). Candidate discovery and primary intervention experiments are conducted on LLAMA2-7B-BASE (Touvron et al., 2023). For fixed-set transfer, Base-identified component sets are directly applied to LLAMA2-7B-CHAT without re-identification, following Section 3.5. Our goal is a controlled analysis of functional selectivity in this setting, rather than a universal characterization of multilingual LLMs.

4.2 BASELINES

We compare CRANE with two controls under matched intervention budgets. **LAPE** (Tang et al., 2024) identifies language-related neurons from cross-language activation probabilities; we apply the same zero-masking evaluation used for CRANE. **Random masking** uses a single budget-matched random draw from the corresponding module structures. Because only one random draw is available, it is treated as an auxiliary perturbation reference rather than a statistically robust random baseline.

4.3 TASKS AND EVALUATION

Natural language understanding. We evaluate English with MMLU (Hendrycks et al., 2021), Chinese with C-Eval (Huang et al., 2023), and Vietnamese with Belebele (Bandarkar et al., 2024), using `lm-evaluation-harness` (Biderman et al., 2024). Because these benchmarks are not fully parallel in domain, format, or difficulty, we focus on within-benchmark intervention effects rather than absolute cross-language score comparisons. Cross-language selectivity is therefore interpreted within the evaluated benchmark suite.

Open-ended generation. Following LAPE (Tang et al., 2024), we evaluate English, Chinese, and Vietnamese open-ended prompts using GPT-4o as an LLM judge on a 1–10 scale (OpenAI, 2024). We use the same scoring protocol before and after intervention and focus on within-language score changes, since judge calibration may differ across languages.

4.4 EVALUATION AND IMPLEMENTATION

For each target language ℓ, we report the intervention effects $\Delta_{\ell\to j}$ defined in Section 3.1. Our primary analysis uses the target and non-target effects together with the strict criterion

$$\Delta_{\ell\to\ell} > \max_{j\neq\ell} \Delta_{\ell\to j}.$$

LangSpec-F1 is reported only as an auxiliary summary. All methods use matched intervention budgets and identical task-evaluation protocols.

Inference is performed with vLLM (Kwon et al., 2023) using greedy decoding, a repetition penalty of 1.1, a maximum generation length of 2048 tokens, and a maximum context length of 8192 tokens. Each task is evaluated before and after masking to obtain the intervention effects directly.

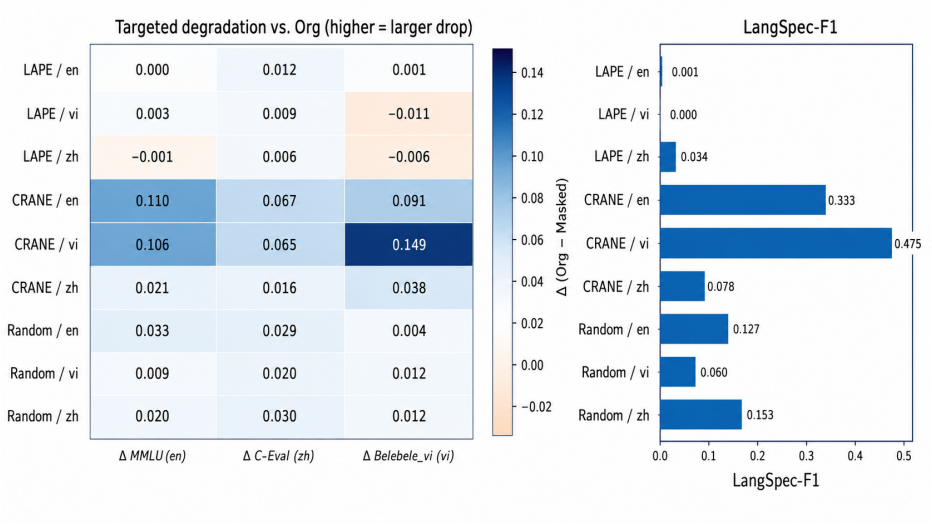

Figure 4: Signed NLU score changes on LLAMA2-7B-BASE, computed as Org − Masked. Positive values indicate degradation and negative values indicate improvement. Target- and non-target-language effects are analyzed jointly to characterize functional selectivity.

Eval.	**Target**	Δ_{CRANE}	Δ_{LAPE}	Δ_{Random}	$\max \Delta_{\text{non-target}}$	**Strict?**
NLU	en	0.1096	0.0003	0.0330	0.0911	✓
NLU	vi	0.1489	0.0000	0.0122	0.1062	✓
NLU	zh	0.0156	0.0060	0.0305	0.0378	×
Open	en	1.5857	0.1857	1.1286	1.1285	✓
Open	vi	0.2143	0.0000	0.0715	1.1000	×
Open	zh	0.9000	0.0000	0.2143	0.6572	✓

Table 1: Summary of Base-model intervention effects. All Δ values are non-negative performance drops as defined in Section 3.1. The non-target column reports the largest CRANE-induced drop among the two non-target languages. Strict functional selectivity requires the CRANE target-language effect to exceed this value. CRANE produces a larger target-language effect than LAPE in all six settings.

Method	**Mask**	**MMLU (en)**	**C-Eval (zh)**	**Belebele (vi)**	**LS-F1**
Org	–	0.4579	0.3470	0.3722	–
LAPE	en	0.4576	0.3351	0.3711	0.0013
LAPE	vi	0.4553	0.3380	0.3833	0.0000
LAPE	zh	0.4589	0.3410	0.3778	0.0337
CRANE	en	0.3483	0.2801	0.2811	0.3328
CRANE	vi	0.3517	0.2816	0.2233	0.4747
CRANE	zh	0.4366	0.3314	0.3344	0.0779
Random	en	0.4249	0.3180	0.3678	0.1270
Random	vi	0.4486	0.3269	0.3600	0.0603
Random	zh	0.4375	0.3165	0.3600	0.1531

Table 2: Full NLU results on LLaMA2-7B Base. Higher task scores are better; lower masked scores imply stronger intervention effects.

5 RESULTS

We organize the analysis around three questions. First, do language-conditioned relevance statistics provide useful structure for candidate discovery? Second, do CRANE-selected components exhibit stronger target-language functional influence, and how often is this influence strictly selective relative to non-target languages? Third, do functional effects associated with Base-identified components remain observable after instruction tuning?

Language-Conditioned Relevance Statistics. We first examine the normalized kurtosis of component relevance distributions under different language conditions. The target-versus-non-target contrasts exhibit language-conditioned variation across Transformer layers, indicating that relevance concentration is not uniform across languages. Figure 3 visualizes these contrasts and provides the statistical structure used for candidate discovery. Importantly, these relevance statistics are not themselves treated as evidence of functional selectivity. They determine which components enter the subsequent intervention analysis; functional evidence is obtained only from the behavioral changes induced by masking.

CRANE Exhibits Stronger Target-Language Functional Influence. Table 1 summarizes the Base-model intervention effects. Across all three languages and both evaluation types, CRANE-selected components produce larger target-language performance drops than the corresponding LAPE-selected components. For NLU, the CRANE target-language effects are $\Delta_{en\to en} = 0.1096$, $\Delta_{vi\to vi} = 0.1489$, and $\Delta_{zh\to zh} = 0.0156$. The effect is strongest for Vietnamese and English, while the Chinese target effect is substantially smaller. Nevertheless, each is larger than its corresponding LAPE target effect. The same pattern holds for open-ended generation: $\Delta_{en\to en} = 1.5857$, $\Delta_{vi\to vi} = 0.2143$, and $\Delta_{zh\to zh} = 0.9000$. Thus, across all six language-by-evaluation settings on the Base model, CRANE-selected components exhibit stronger target-language functional effects than those selected by LAPE. This is the most consistent empirical pattern in our experiments: relevance-guided candidate discovery identifies component sets whose intervention has a stronger behavioral effect on the associated target language than the activation-based baseline.

Target-Language Influence Is Not Strictly Language-Exclusive. Stronger target-language influence does not imply that the same intervention is confined to the target language. Under the stricter criterion $\Delta_{\ell\to\ell} > \max_{j\neq\ell} \Delta_{\ell\to j}$, functional selectivity holds in four of the six Base-model settings. For NLU, the English and Vietnamese component sets satisfy the criterion, whereas the Chinese set does not. For Chinese-associated CRANE components, $\Delta_{zh\to zh} = 0.0156$, while the largest non-target effect is $\max_{j\neq zh} \Delta_{zh\to j} = 0.0378$. For open-ended generation, the English and Chinese component sets satisfy the strict criterion, whereas the Vietnamese set does not. For Vietnamese-associated components, $\Delta_{vi\to vi} = 0.2143$, but the largest non-target effect reaches $\max_{j\neq vi} \Delta_{vi\to j} = 1.1000$. These results establish an important distinction: target-language functional influence does not imply strict cross-language functional selectivity. In our evaluated setting, language-related components exhibit measurable target-language functional biases, but substantial cross-language effects remain in some cases. Because the NLU benchmarks are not fully parallel, these comparisons should be interpreted as benchmark-conditioned functional effects rather than as complete disentanglement of language identity from task or domain.

Comparison with a Budget-Matched Random Draw. The single budget-matched random draw provides an additional reference for generic perturbation effects. CRANE produces a larger target-language effect than this random draw in five of the six Base-model settings. The exception is Chinese NLU, where the random target-language drop is 0.0305, compared with 0.0156 for CRANE. This comparison also shows that masking an equally sized arbitrary component set can itself produce non-trivial behavioral changes. Because only a single random draw is available, we do not interpret the five-of-six pattern as statistically robust superiority over random masking. Instead, the random result serves as an auxiliary control that motivates reporting the complete target- and non-target-language intervention profile rather than relying on target degradation alone.

Part of the Functional Influence Persists from Base to Chat. Finally, we directly transfer the Base-identified component sets to LLAMA2-7B-CHAT without re-identification. Full transfer results are reported in Appendix Tables 4 and 5. For NLU, Base-identified CRANE components

remain behaviorally influential in the Chat model, although the magnitude varies substantially across languages. The target-language drops are $\Delta^{\text{Chat}}_{en\to en} = 0.0305$, $\Delta^{\text{Chat}}_{vi\to vi} = 0.1200$, and $\Delta^{\text{Chat}}_{zh\to zh} = 0.0030$. The retained effect is most pronounced for Vietnamese. Open-ended generation shows the same qualitative pattern. The target-language drops under transferred CRANE masks are $\Delta^{\text{Chat}}_{en\to en} = 2.0428$, $\Delta^{\text{Chat}}_{vi\to vi} = 2.0714$, and $\Delta^{\text{Chat}}_{zh\to zh} = 0.6286$. At the same time, several non-target languages also undergo substantial degradation under the transferred interventions. We therefore interpret these experiments as evidence that part of the functional influence associated with Base-identified components remains observable after instruction tuning. They do not establish preservation of strict cross-language functional selectivity or invariance of individual neuron identities.

Method	**Mask**	**Open-en**	**Open-vi**	**Open-zh**	**LS-F1**
Org	–	3.2286	1.9286	2.3714	–
LAPE	en	3.0429	2.0143	2.2857	0.1061
LAPE	vi	3.3286	2.2571	2.2429	0.0000
LAPE	zh	3.0000	2.1714	2.6714	0.0000
CRANE	en	1.6429	1.5286	1.2429	0.5337
CRANE	vi	2.1286	1.7143	1.7000	0.1322
CRANE	zh	2.5714	1.5857	1.4714	0.4582
Random	en	2.1000	1.8714	1.7286	0.4514
Random	vi	2.7286	1.8571	1.4286	0.0486
Random	zh	2.5857	2.0143	2.1571	0.1328

Table 3: Full open-ended generation results on LLaMA2-7B Base. Scores are judged on a 1–10 scale.

6 Conclusion

We study language-related internal components through the lens of **functional selectivity**. CRANE separates relevance-based candidate discovery from intervention-based functional validation by combining AttnLRP with controlled zero-masking. Across English, Chinese, and Vietnamese on LLaMA2-7B-Base, CRANE-selected components produce stronger target-language functional effects than the activation-based LAPE baseline in all six language-by-evaluation settings. However, stronger target-language influence does not imply strict language exclusivity. Under the stricter target-over-all-non-target criterion, functional selectivity holds in four of the six settings, while substantial non-target effects remain in the others. We further transfer Base-identified component sets directly to LLaMA2-7B-Chat without re-identification and find that part of their functional influence remains observable after instruction tuning. These results do not imply preservation of strict selectivity or individual neuron identity across model stages. Overall, CRANE shifts the analysis of language-related internal structure from correlational identification toward functional selectivity. Our results suggest that language-related components can exhibit measurable target-language functional biases while remaining involved in overlapping cross-lingual computation.

7 Limitations

Our study is limited to LLaMA2-7B Base/Chat and three languages (English, Chinese, and Vietnamese), so the findings may not generalize to other model families, scales, or languages. The NLU benchmarks are not fully parallel, and observed cross-language effects may partly reflect differences in task, domain, or difficulty rather than language alone. CRANE also relies on normalized kurtosis for candidate discovery and zero masking for functional validation; other statistics or intervention schemes may reveal complementary structure. Finally, the random baseline uses only a single budget-matched draw, and the GPT-4o judge used for open-ended generation may exhibit language-dependent calibration bias.

Reproducibility Statement

The CRANE pipeline, candidate-discovery procedure, intervention definition, evaluation metrics, baseline settings, and inference configuration are specified in Sections 3 and 4. The appendix provides the complete algorithm, additional candidate-selection details, and full Base-to-Chat transfer results. All compared methods use matched intervention budgets and identical task evaluation protocols. We plan to release the implementation and experimental configurations to facilitate reproduction of the reported results.

AI Use Statement

Generative AI tools were used to assist with manuscript drafting and language polishing, literature retrieval and organization, scientific-figure drafting, and discussion of research framing, methodological presentation, and interpretation of existing experimental results. They were not used to generate experimental data or to prove mathematical claims. All AI-assisted text, citations, factual statements, and interpretations were reviewed and verified by the authors, who take full responsibility for the content of this paper.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 135–168. PMLR, 2024. URL https://proceedings.mlr.press/v235/achtibat24a.html.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12 (8):e0181142, 2017a. doi: 10.1371/journal.pone.0181142. URL https://doi.org/10.1371/journal.pone.0181142.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168. Association for Computational Linguistics, 2017b. doi: 10.18653/v1/W17-5221. URL https://aclanthology.org/W17-5221/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *CoRR*, abs/2405.14782, 2024. doi: 10.48550/ARXIV.2405.14782. URL https://doi.org/10.48550/arXiv.2405.14782.

Lawrence T. DeCarlo. On the meaning and use of kurtosis. *Psychological Methods*, 2(3):292–307, 1997.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841. Association for Computational

Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. Language arithmetics: Towards systematic language neuron identification and manipulation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2911–2937. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.156. URL https://aclanthology.org/2025.ijcnlp-long.156/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *NeurIPS*, 2022.

OpenAI. GPT-4o System Card, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

Inaya Rahmanisa, Lyzander Marciano Andrylie, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. Unveiling the influence of amplifying language-specific neurons. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 919–968, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-ijcnlp.55. URL https://aclanthology.org/2025.findings-ijcnlp.55/.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. How does alignment enhance llms' multilingual capabilities? A language neurons perspective. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 34799–34807. AAAI Press, 2026. doi: 10.1609/AAAI.V40I41.40782. URL https://doi.org/10.1609/aaai.v40i41.40782.

A Metric and Candidate-Selection Details

For each component n and language ℓ, sample-level relevance values form a distribution $\mathcal{R}_{n,\ell}$. We summarize concentration with

$$\text{kurtosis}(R_n) = \frac{\mathbb{E}[(R_n - \mu)^4]}{\sigma^4}, \tag{1}$$

where μ and σ are the mean and standard deviation. For each language, these values are z-normalized within each Transformer layer; coordinates with $z > 1$ form the candidate set. Cross-language selectivity is not used in candidate selection.

For continuity with the original analysis, we also report LangSpec-F1 as a secondary summary. With $\Delta = \max(\text{drop}, 0)$,

$$\text{Precision} = \frac{\Delta_\ell}{\Delta_\ell + \max_{\ell' \neq \ell} \Delta_{\ell'} + \epsilon}, \qquad \text{Recall} = \frac{\Delta_\ell}{S_\ell + \epsilon}, \tag{2}$$

$$\text{LangSpec-F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon}, \tag{3}$$

where S_ℓ is the unmasked target score. We do not use this metric as a definition of specificity; raw target and collateral drops are primary.

B Additional Transfer Results

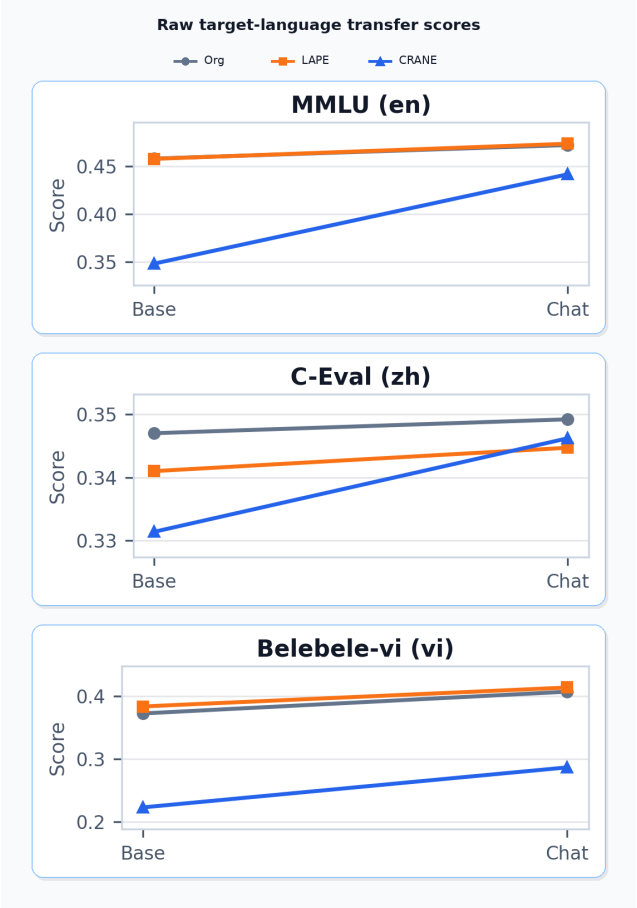

Figure 5: Raw target-language scores under Base-to-Chat transferred masks. Candidate sets are identified on LLaMA2-7B Base and applied directly to LLaMA2-7B Chat.

Method	Mask	Base model			Chat model			LangSpec-F1
		MMLU (en)	C-Eval (zh)	Belebele (vi)	MMLU (en)	C-Eval (zh)	Belebele (vi)	
Org	–	0.4579	0.3470	0.3722	0.4720	0.3492	0.4067	–
LAPE	en	0.4576	0.3351	0.3711	0.4734	0.3514	0.4044	0.0000
LAPE	vi	0.4553	0.3380	0.3833	0.4719	0.3477	0.4133	0.0000
LAPE	zh	0.4589	0.3410	0.3778	0.4719	0.3447	0.4078	0.0253
CRANE	en	0.3483	0.2801	0.2811	0.4415	0.3351	0.3367	0.1066
CRANE	vi	0.3517	0.2816	0.2233	0.4427	0.3276	0.2867	0.4316
CRANE	zh	0.4366	0.3314	0.3344	0.4645	0.3462	0.3733	0.0154

Table 4: NLU transfer results. LangSpec-F1 is computed on the Chat model.

Method	Mask	Open-en	Open-vi	Open-zh	LangSpec-F1
Org	–	7.9857	7.0143	6.8286	–
LAPE	en	7.9429	7.1286	6.9571	0.0107
LAPE	vi	7.6857	7.3000	6.9714	0.0000
LAPE	zh	8.0286	7.1857	6.7286	0.0289
CRANE	en	5.9429	3.2429	4.4429	0.2961
CRANE	vi	7.9143	4.9429	5.5143	0.3984
CRANE	zh	7.5714	6.1857	6.2000	0.1517

Table 5: Open-ended transfer results on LLaMA2-7B Chat under Base-identified masks.

C CRANE ALGORITHM

Algorithm 1 CRANE: Relevance-Guided Analysis of Language-Conditioned Components

Require: Multilingual dataset $\mathcal{D} = \{\mathcal{D}_\ell\}_{\ell \in \mathcal{L}}$, model f, languages $\mathcal{L}$, intervention budget B
Ensure: Candidate component sets $\{\mathcal{N}_\ell\}_{\ell \in \mathcal{L}}$ and intervention metrics

for each language $\ell \in \mathcal{L}$ **do**
  for each sample $x \in \mathcal{D}_\ell$ **do**
    Compute a language-conditioned objective and propagate LRP/AttnLRP relevance to MLP components.
  end for
  Form per-component relevance distributions $\{\mathcal{R}_{n,\ell}\}_{n=1}^{N}$ and compute kurtosis $\kappa_{n,\ell}$.
  Z-normalize $\{\kappa_{n,\ell}\}_{n=1}^{N}$ within each Transformer layer ℓ.
  Select $\mathcal{N}_\ell$ using the fixed target-language threshold $z > 1$.
end for
for each target language $\ell \in \mathcal{L}$ **do**
  Zero the outputs of components in $\mathcal{N}_\ell$ during inference.
  Evaluate target and non-target benchmarks under the same intervention budget.
  Report target drop, worst collateral drop, strict-selectivity indicator, and secondary LangSpec-F1.
end for
return $\{\mathcal{N}_\ell\}_{\ell \in \mathcal{L}}$ and evaluation metrics