CRANE: Causal Relevance Analysis of Language-Specific Neurons in Multilingual Large Language Models

Anonymous ACL submission

Abstract

Multilingual large language models (LLMs) achieve strong performance across languages, yet how language capabilities are organized at the neuron level remains poorly understood. Prior work has identified language-related neurons mainly through activation-based heuristics, which conflate language preference with functional importance. Prior work has identified language-related neurons mainly through activation-based heuristics, which conflate language preference with functional importance. We propose **CRANE**, a relevance-based analysis framework that *redefines language specificity in terms of functional necessity*, identifying language-specific neurons through targeted neuron-level interventions. CRANE characterizes neuron specialization by their contribution to language-conditioned predictions rather than activation magnitude. Our implementation will be made publicly available. Neuron-level interventions reveal a consistent asymmetric pattern: masking neurons relevant to a target language selectively degrades performance on that language while preserving performance on other languages to a substantial extent, indicating language-selective but non-exclusive neuron specializations. Experiments on English, Chinese, and Vietnamese across multiple benchmarks, together with a dedicated relevance-based metric and base-to-chat model transfer analysis, show that CRANE isolates language-specific components more precisely than activation-based methods.

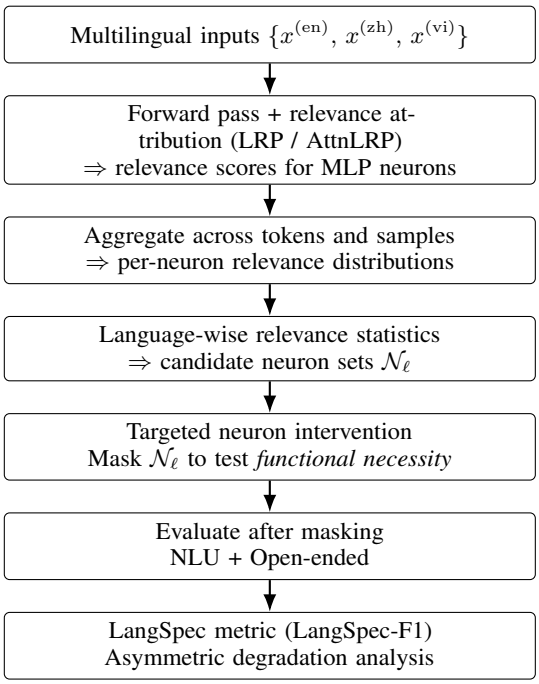

Figure 1: **Overview of CRANE.** CRANE identifies language-specific neurons via relevance attribution and language-wise relevance statistics, and validates *functional necessity* through targeted neuron-level interventions with unified evaluation across NLU and open-ended benchmarks.

1 Introduction

Large Language Models (LLMs) have achieved remarkable progress in multilingual natural language processing, enabling a single model to perform understanding and generation across many languages. Foundational multilingual models such as Meta's LLaMA2 (Touvron et al., 2023), Google DeepMind's Gemini 1.5 (Google DeepMind Team, 2024), and emerging open-source GPT families such as GPT-OSS (OpenAI, 2025) have demonstrated strong performance across diverse linguistic settings. Despite these empirical advances, how language capabilities are organized and specialized at the neuron level remains poorly understood.

A growing line of work has attempted to identify *language-specific neurons*, often based on activation statistics or language-conditioned probing. Earlier studies on recurrent neural networks examined gated units as mechanisms for information storage and control (Hochreiter and Schmidhuber, 1997; Elman, 1990), while later work analyzed intrinsic linguistic knowledge in sequential models (Lu et al., 2018). More recently, neuron-level analyses in LLMs have focused on activation dis-

tributions and activation probabilities (Tang et al., 2024), as well as finer-grained neuron identification approaches (Zhang et al., 2025; Sachan et al., 2025). These studies reveal statistical regularities associated with language, but largely assume that statistical correlation reflects functional importance.

Activation does not imply functional necessity. Prior work identifies language-specific neurons based on activation statistics, but typically lacks direct functional validation (Tang et al., 2024). Consequently, neurons correlated with a language may not be required for its performance. We instead define language specificity in terms of *functional necessity*.

We propose **CRANE** (*Causal Relevance-based Analysis of Neuron Specialization*), a relevance-based framework that operationalizes this definition through targeted neuron-level interventions. CRANE builds on layer-wise relevance propagation (LRP) (Arras et al., 2016, 2017) and its Transformer extension AttnLRP (Achtibat et al., 2024) to attribute language-conditioned predictions to individual neurons. By identifying neurons based on output-level relevance rather than activation magnitude, CRANE distinguishes language preference from functional contribution.

Using CRANE, we uncover a consistent asymmetric specialization pattern across languages. Masking neurons relevant to a target language leads to substantially larger degradation on that language while generally preserving performance on others. This pattern supports *language-selective but non-exclusive* specialization: neurons contribute disproportionately to specific languages while remaining part of shared multilingual computation. We validate these findings on English, Chinese, and Vietnamese across multiple multilingual benchmarks.

In summary, this work makes four main contributions.

(1) We redefine language specificity at the neuron level by shifting from activation-based correlation to functional necessity, and provide a concrete operationalization via CRANE.

(2) We introduce a relevance-based evaluation metric that quantifies language-selective functional effects under targeted neuron intervention, enabling systematic comparison across languages and models.

(3) We present functional evidence of asymmetric, non-exclusive language specialization in multilingual LLMs, where neurons contribute disproportionately to specific languages while remaining involved in multilingual computation.

(4) We conduct a controlled transfer analysis from pretrained Base models to post-trained Chat models without re-identification, providing empirical insight into how language-selective neuron effects persist or shift after instruction tuning.

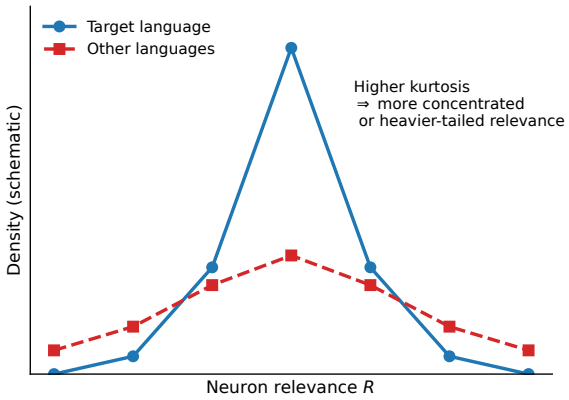

Figure 2: Schematic intuition of kurtosis-based language contrast. Under a target language condition, a neuron may exhibit a more concentrated or heavy-tailed relevance distribution than under other languages. Kurtosis is used as one example statistic to quantify such concentration differences.

2 Related Work

Correlation-based explanations for LLMs. Many interpretability methods for LLMs extract correlational explanatory signals. Feature attribution techniques such as Integrated Gradients (Sundararajan et al., 2017) and perturbation-based approaches (Li et al., 2016) estimate input contributions, while attention-based visualizations (Vig, 2019) provide intuitive but potentially unfaithful explanations (Jain and Wallace, 2019). Probing-based analyses (Belinkov, 2022) and neuron- or concept-level studies (Dalvi et al., 2019; Kim et al., 2018) reveal representational patterns but may conflate selectivity with importance (Hewitt and Liang, 2019). Overall, these approaches characterize statistical or descriptive signals, without establishing functional necessity through intervention.

Multilingual structure and language-selective components. For multilingual models, prior work studies cross-lingual alignment and shared representations using probing and representation similarity analysis (Karthikeyan et al., 2020; Blevins et al., 2022). More recent studies investigate language-specific or language-selective neurons by analyzing activation statistics or frequency,

exemplified by LAPE (Tang et al., 2024) and related neuron-level methods (Zhang et al., 2025; Sachan et al., 2025). While these methods uncover language-related structure, they typically equate activation selectivity with functional importance. As noted in prior work, such selectivity does not guarantee that intervening on the identified neurons induces targeted language degradation (Tang et al., 2024).

Intervention-based and causal analyses. A growing line of mechanistic interpretability emphasizes causal validation via intervention. Methods such as activation patching and causal tracing intervene on internal activations to localize computations responsible for model behavior (Meng et al., 2022; Zhang et al., 2024). These approaches motivate defining interpretability in terms of functional necessity rather than correlational signals. Our work aligns with this perspective and applies intervention-based validation to study language specialization in multilingual LLMs.

3 Method

CRANE operationalizes *functional necessity* as the definition of language specificity at the neuron level. Rather than inferring language specificity from activation-based correlation, CRANE evaluates whether intervening on a neuron set induces a disproportionately larger performance degradation on a target language than on others. This definition-driven criterion guides all components of the framework, from neuron attribution to intervention-based validation. Figure 1 illustrates the overall workflow.

3.1 Problem Setup and Neuron-level Representation

Let the input token sequence be $\mathbf{x} = \{x_i\}_{i=1}^{N}$ and the model output logits be $\mathbf{y} \in \mathbb{R}^{|V|}$, where $|V|$ denotes the vocabulary size. For layer l, the hidden state at token position i is $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$.

We analyze standard Transformer architectures without architectural modification. Attention layers are treated as contextual mixing modules, while neuron-level analysis is conducted on MLP components, where each **column** of the linear projection matrices is treated as an individual neuron, consistent with prior neuron-level analyses. Given a set of languages $\mathcal{L}$, our goal is to identify, for each language $\ell \in \mathcal{L}$, a neuron subset $\mathcal{N}_\ell$ such that intervening on $\mathcal{N}_\ell$ results in a larger relative performance degradation on language ℓ than on non-target languages under the same intervention budget.

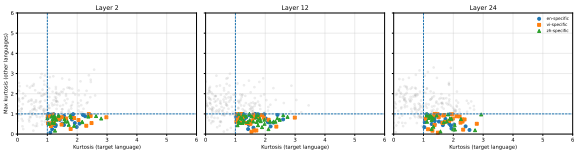

Figure 3: Kurtosis-based contrast for identifying language-specific neurons. Each subplot corresponds to one Transformer layer. Each point represents a neuron, plotted by its normalized kurtosis under a target language (x-axis) versus the maximum kurtosis under other languages (y-axis). Dashed lines indicate the kurtosis threshold (set to 1). Neurons with high kurtosis for the target language but low kurtosis for others (bottom-right region) exhibit stronger language-specific relevance concentration.

3.2 Neuron-level Relevance Attribution

To estimate the contribution of individual neurons to model outputs, CRANE employs *layer-wise relevance propagation* (LRP), a relevance attribution technique that redistributes output relevance backward through the network while preserving relevance conservation (Arras et al., 2016, 2017). We adopt existing extensions of LRP for Transformer architectures (Achtibat et al., 2024).

Given an input $\mathbf{x}$ in language ℓ and a language-conditioned output objective $f_\ell(\mathbf{x})$, relevance is initialized at the output layer and propagated backward according to the conservation principle:

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

Relevance is propagated to MLP neurons, yielding token-level relevance scores. These scores are aggregated across tokens to obtain a sample-level relevance value per neuron. We emphasize that relevance attribution serves as a *tool* for estimating neuron contributions within CRANE, rather than constituting the method itself.

3.3 Language-conditioned Relevance Distributions

CRANE characterizes language specificity at the *distributional* level across samples. For each neuron n and language ℓ, aggregating relevance values over a large set of inputs induces a language-conditioned relevance distribution.

Intuitively, neurons that are functionally necessary for a target language tend to exhibit more con-

centrated or heavy-tailed relevance distributions under that language compared to others. Figure 2 provides a schematic illustration of this intuition. To quantify such concentration differences, we adopt kurtosis, a fourth-order moment statistic that captures distributional peakedness and tail behavior (DeCarlo, 1997).

Formally, the kurtosis of a neuron's relevance distribution R_n is defined as:

$$\text{kurtosis}(R_n) = \frac{\mathbb{E}[(R_n - \mu)^4]}{\sigma^4}, \tag{2}$$

where μ and σ denote the mean and standard deviation, respectively.

We compare language-conditioned kurtosis values across languages and rank neurons by their relative concentration under the target language. In practice, kurtosis scores are normalized within each layer, and neurons whose normalized scores exceed a fixed threshold are selected to form candidate sets. This threshold is used to control the intervention budget rather than to define language specificity itself. We show in experiments that the resulting findings are robust across a reasonable range of thresholds.

3.4 Neuron-level Intervention and Validation

Given a candidate neuron set $\mathcal{N}_\ell$, CRANE validates functional necessity through targeted neuron-level intervention. Specifically, neurons in $\mathcal{N}_\ell$ are masked by setting their outputs to zero during inference, while all other components remain unchanged.

Rather than assuming strict language exclusivity, we evaluate whether masking $\mathcal{N}_\ell$ induces a larger relative performance degradation on the target language than on non-target languages under identical intervention budgets. This relative degradation criterion provides functional evidence for *language-selective but non-exclusive* neuron specialization.

We evaluate intervention effects on both natural language understanding (NLU) and open-ended generation benchmarks. The complete CRANE procedure is summarized in Algorithm 1.

3.5 Language-specificity Metric

To quantify language-specific functional necessity under neuron-level intervention, we introduce **LangSpec-F1**, a composite metric that balances targeted performance degradation on the masked language with stability on non-target languages.

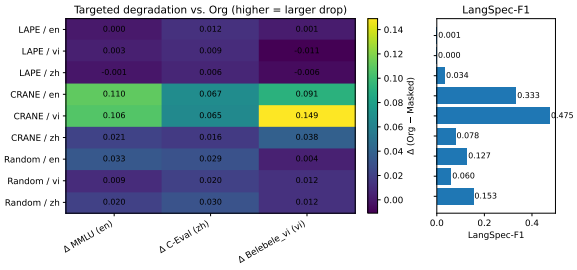

Figure 4: Targeted NLU degradation under neuron-level intervention on LLAMA2-7B-BASE. The heatmap reports absolute performance drops (Δ = Org − Masked) on each evaluation benchmark (columns) when masking neuron sets identified for a given target language and method (rows). Higher values indicate stronger degradation. The bar plot summarizes language-specific functional effects using LangSpec-F1.

LangSpec-F1 is a metric to quantify functional selectivity, not strict exclusivity; high LangSpec-F1 indicates that neuron interventions disproportionately affect the target language relative to others.

Let Δ_ℓ denote the performance drop on the target language ℓ after masking a neuron set, and let $\max_{\ell' \neq \ell} \Delta_{\ell'}$ denote the maximum performance drop observed on any non-target language. We consider only negative performance changes by defining $\Delta = \max(\text{drop}, 0)$, such that performance improvements are not counted as degradation.

We define precision and recall as:

$$\text{Precision} = \frac{\Delta_\ell}{\Delta_\ell + \max_{\ell' \neq \ell} \Delta_{\ell'} + \epsilon}, \tag{3}$$

$$\text{Recall} = \frac{\Delta_\ell}{S_\ell + \epsilon}, \tag{4}$$

where S_ℓ denotes the original (unmasked) performance score on language ℓ, and ϵ is a small constant for numerical stability.

The final LangSpec-F1 score is computed as the harmonic mean of precision and recall:

$$\text{LangSpec-F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \epsilon}. \tag{5}$$

We emphasize that LangSpec-F1 is an operational metric for evaluating functional necessity, and that these design choices reflect a conservative assessment of non-target language interference rather than a definition of language specificity itself.

Intuitively, LangSpec-F1 assigns high scores to neuron sets whose intervention causes substantial degradation on the target language while inducing minimal degradation on non-target languages, relative to the original performance level.

Method	Mask lang	MMLU (en)	C-Eval (zh)	Belebele_vi (vi)	LangSpec-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 1: NLU results on LLaMA2-7B-Base with neuron masking. Higher is better for task metrics. LangSpec-F1 reflects targeted degradation and non-target stability.

4 Experimental Setup

4.1 Languages and Models

We study three typologically diverse languages: English (en), Chinese (zh), and Vietnamese (vi). Experiments are conducted at two levels. First, we identify and validate language-specific neuron sets on **LLaMA2-7B-Base**. Second, we transfer neuron sets discovered on the base model to **LLaMA2-7B-Chat** to examine whether language-specific functional necessity is preserved after post-training. This setting allows us to probe the stability of neuron-level specialization across pretraining and post-training stages.

4.2 Baselines

We compare CRANE against the following baselines, all evaluated under identical intervention budgets:

- **LAPE** (Tang et al., 2024): an activation-based approach that identifies language-related neurons by measuring neuron activation likelihood across language-specific corpora.
- **Random masking (budget-matched)**: randomly sampling the same number of neurons from identical module structures, serving as a control to distinguish targeted language effects from generic perturbations.

4.3 Tasks and Evaluation

Natural language understanding (NLU). We evaluate language understanding using standard language-specific benchmarks: **MMLU** for English (Hendrycks et al., 2021), **C-Eval** for Chinese (Huang et al., 2023), and **Belebele** for Vietnamese (Bandarkar et al., 2023). All NLU evaluations are conducted using a unified evaluation framework, `lm-evaluation-harness` (Biderman et al., 2024), with consistent prompting and decoding settings across languages and methods.

Open-ended generation. We evaluate open-ended generation using an English/Chinese/Vietnamese question set following the LAPE protocol. Model outputs are scored on a 1–10 scale by an LLM-as-a-judge based on GPT-4o(Hurst et al., 2024). The same evaluation protocol is applied before and after neuron-level intervention to ensure direct comparability.

Language-specificity metric. We report **LangSpec-F1**, a composite metric that captures (i) performance degradation on the target language after neuron-level intervention and (ii) relative performance retention on non-target languages under the same intervention budget. The formal definition and computation of LangSpec-F1 are provided in Section 3.5.

4.4 Inference and Implementation Details

Inference and evaluation are performed using **vLLM**(Kwon et al., 2023). We use greedy decoding with repetition_penalty=1.1, max_new_tokens=2048, and a maximum context length of 8192. For each identified neuron set, we evaluate all tasks both **before** and **after** masking to ensure direct comparison of intervention effects.

5 Results

We report results on LLaMA2-7B-Base and LLaMA2-7B-Chat. We analyze distributional relevance statistics to characterize language-related structure, validate functional effects via neuron-level intervention on NLU and open-ended generation, and examine transferability from the Base

model to the Chat model.

5.1 Normalized Kurtosis Reveals Language-related Structure

After computing neuron-level relevance scores, we analyze the *normalized kurtosis* of relevance distributions under different language conditions. For each neuron, we compare its language-conditioned kurtosis values to assess how strongly its relevance concentrates for one language relative to others.

We observe clear separation patterns across languages, indicating that normalized kurtosis captures systematic differences in how neurons participate in computation under different language inputs. Importantly, this analysis reflects *language-related structure* at the distributional level rather than functional language specificity. Accordingly, kurtosis-based selection serves solely as a candidate identification step, with functional necessity established only through subsequent intervention.

Figure 3 visualizes this structure by plotting neurons in a two-dimensional contrast space defined by target-language versus non-target normalized kurtosis. Neurons exhibiting high target-language kurtosis and low non-target kurtosis form a distinct region, motivating our candidate selection criterion.

5.2 NLU: CRANE Induces Targeted Functional Degradation

Table 1 reports NLU results on LLAMA2-7B-BASE under different neuron masking strategies. Activation-based baselines such as LAPE induce small and relatively uniform performance changes across languages, yielding LangSpec-F1 values close to zero. In contrast, neuron sets selected by CRANE produce substantially larger degradation on the intended target language under identical intervention budgets.

For example, when targeting Vietnamese, BELEBELE_VI accuracy decreases from 0.3722 (unmasked) to 0.2233 after masking CRANE-selected neurons, yielding a LangSpec-F1 of 0.4747. Across all language and benchmark settings, CRANE consistently yields higher LangSpec-F1 than activation-based and random baselines, indicating robust target-aligned degradation while preserving non-target performance.

Figure 4 summarizes absolute performance drops on NLU benchmarks relative to the unmasked model. Compared to baselines, CRANE yields larger drops on target-language benchmarks and smaller changes on non-target languages, consistent with language-selective but non-exclusive functional effects.

Algorithm 1 CRANE: Causal Relevance-based Analysis of Neuron Specialization

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

// Stage 1: Neuron relevance attribution
for each language $\ell \in \mathcal{L}$ **do**
  for each sample $x \in \mathcal{D}_\ell$ **do**
    Run forward pass and compute objective $s = f(x)$
    Apply LRP/AttnLRP to propagate relevance to MLP neurons
    Obtain neuron relevance vector $\mathbf{r}(x, \ell) \in \mathbb{R}^N$ {N: number of MLP neurons}
  end for
  Aggregate $\{\mathbf{r}(x, \ell)\}_{x \in \mathcal{D}_\ell}$ to form per-neuron relevance distributions $\{\mathcal{R}_{n,\ell}\}_{n=1}^{N}$
end for
// Stage 2: Language-conditioned relevance concentration
for each language $\ell \in \mathcal{L}$ **do**
  for each neuron $n = 1, \ldots, N$ **do**
    Compute kurtosis $\kappa_{n,\ell}$ of $\mathcal{R}_{n,\ell}$
  end for
  Normalize $\{\kappa_{n,\ell}\}_{n=1}^{N}$ within each layer to obtain normalized scores $\{\tilde{\kappa}_{n,\ell}\}$
end for
// Stage 3: Candidate neuron selection (budget control)
for each target language $\ell \in \mathcal{L}$ **do**
  Select $\mathcal{N}_\ell \leftarrow \textsc{BudgetSelect}(\{\tilde{\kappa}_{n,\ell}\}_{n=1}^{N}, B)$ {e.g., threshold on $\tilde{\kappa}$ or Top-B ranking}
end for
// Stage 4: Functional necessity validation (intervention)
for each target language $\ell \in \mathcal{L}$ **do**
  Mask neuron outputs for $\mathcal{N}_\ell$ (set outputs to 0) during inference
  Evaluate NLU and open-ended benchmarks across languages under the same budget B
  Compute LangSpec-F1 from relative target-language degradation vs. non-target retention
end for
return $\{\mathcal{N}_\ell\}_{\ell \in \mathcal{L}}$ and evaluation metrics

5.3 Open-ended Generation: Base Model

Table 2 reports open-ended generation results on LLAMA2-7B-BASE. Activation-based baselines yield weak or inconsistent intervention effects, whereas CRANE induces clearer target-language degradation and achieves higher LangSpec-F1 in multiple cases. Given the inherent variability of open-ended evaluation, these results are presented as supportive evidence of functional influence.

5.4 Transferability of Base-identified Neurons to the Chat Model

We examine whether neuron sets identified on the pretrained Base model retain functional influence

Method	Mask lang	Open_en	Open_vi	Open_zh	LangSpec-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	0.4406
Random	vi	2.7286	1.8571	1.4286	0.0751
Random	zh	2.5857	2.0143	2.1571	0.1422

Table 2: Open-ended generation results on LLaMA2-7B-Base (LLM-judge scores; higher is better).

after post-training. Under a strict transfer setting, neuron sets are identified only on LLAMA2-7B-BASE and directly transferred to LLAMA2-7B-CHAT without re-identification. Observed target-language degradation under this setting indicates that a subset of Base-identified neurons continues to contribute functionally after post-training.

NLU transfer. Table 3 reports NLU results under transferred neuron masks. Activation-based baselines induce small and inconsistent changes, whereas CRANE-selected neuron sets produce larger degradation on the target language on the Chat model, yielding higher LangSpec-F1 in several settings, most notably for Vietnamese. These findings indicate partial preservation of functional influence rather than invariance of neuron identity across training stages.

Figure 6 presents raw target-language benchmark scores under transferred neuron masks, illustrating the absolute performance changes underlying the aggregated metrics.

Open-ended transfer. Table 4 reports open-ended generation results on the Chat model under transferred neuron masks. Consistent with NLU transfer, CRANE generally produces stronger target-language degradation than activation-based baselines under comparable intervention budgets. Given evaluation noise, these results are interpreted as supportive evidence of retained functional influence.

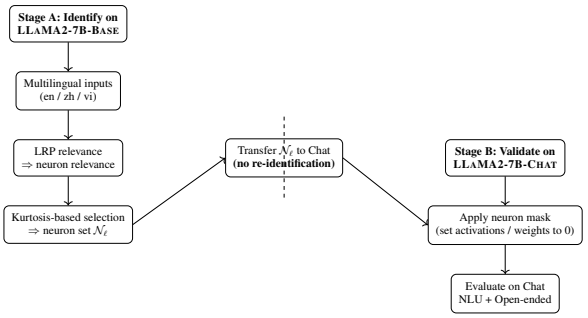

Figure 5: **Transfer setting from Base to Chat.** We first identify language-related neuron sets $\mathcal{N}_\ell$ on LLAMA2-7B-BASE using relevance attribution and kurtosis-based statistics. The same neuron sets are then directly transferred and masked on LLAMA2-7B-CHAT *without re-identification* to evaluate whether their functional influence persists after post-training.

6 Conclusion

We introduced **CRANE**, a relevance-based framework for analyzing language-specific functional contributions in multilingual large language models. By combining relevance attribution with neuron-level intervention, CRANE operationalizes language specificity in terms of functional necessity rather than activation-based correlation, and we further propose **LangSpec-F1**, a metric to quantify language-selective functional effects under targeted neuron interventions. Across multiple benchmarks and languages, CRANE consistently induces stronger and more target-aligned degradation than activation-based baselines under matched intervention budgets. Under a strict transfer setting from a pretrained Base model to a post-trained Chat model without re-identification, we find that a subset of Base-identified neurons retains measurable functional influence after post-training, while others shift. More broadly, this work highlights the importance of distinguishing statistical language correlation from functional language contribution at the neuron level, and positions CRANE, together with the proposed metric, as a general framework for studying multilingual representations and their evolution across training stages.

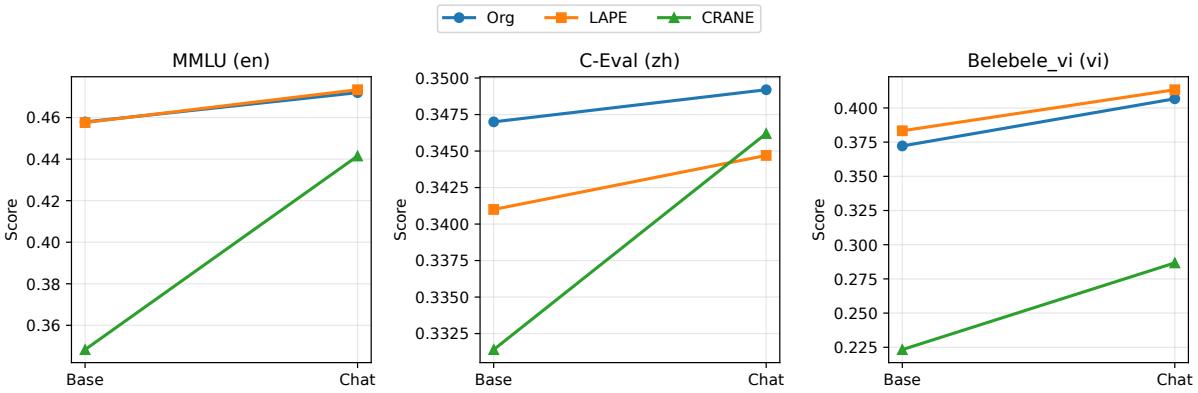

Figure 6: Raw target-language benchmark scores from Base to Chat under transferred neuron masks. Neuron sets are identified on LLAMA2-7B-BASE and directly applied to LLAMA2-7B-CHAT. Each panel corresponds to the target-language benchmark (MMLU / C-Eval / Belebele_vi).

Method	Mask	Base model			Chat model			LangSpec-F1
		MMLU (en)	C-Eval (zh)	Belebele_vi (vi)	MMLU (en)	C-Eval (zh)	Belebele_vi (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 3: NLU transfer results. Neuron sets are identified on LLAMA2-7B-BASE and directly applied (masked) on LLAMA2-7B-CHAT. LangSpec-F1 is computed for the Chat model to summarize language-targeted degradation under masking.

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 4: Open-ended transfer results on LLAMA2-7B-CHAT. We mask neuron sets identified from the Base model and evaluate on three language-specific open-ended benchmarks (scores in the 1–10 range).

7 Limitations

CRANE currently relies on normalized kurtosis to characterize concentration patterns in language-conditioned relevance distributions. While effective and interpretable, kurtosis is only one possible statistic, and other distributional measures may capture complementary aspects of language-related structure.

Functional validation in this work is performed via neuron masking, which constitutes a relatively coarse intervention. More fine-grained causal techniques, such as neuron re-weighting or path-specific analysis, may provide deeper insight into the mechanisms underlying language-specific functional effects.

Our experiments focus on three languages and a single model family. In addition, open-ended generation is evaluated using an LLM-as-a-judge protocol, which may introduce variance. Extending the analysis to broader language coverage, additional architectures, and alternative evaluation paradigms remains important future work.

Ethics Statement

This work analyzes internal mechanisms of multilingual language models and evaluates their behavior on public benchmarks and curated prompts. No personal or sensitive data are collected or used.

A potential risk is over-interpreting neuron-level findings as definitive causal explanations of language behavior. We mitigate this risk by grounding claims in controlled intervention, reporting results cautiously, and clearly delineating the empirical scope and limitations of our analysis.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. 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*, pages 135–168. PMLR.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. "what is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 11(12):e0168142.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, in conjunction with EMNLP 2017*, pages 159–168. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, and 1 others. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint*, abs/2308.16884.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, and 1 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Terra Blevins and 1 others. 2022. Analyzing and cross-lingual pretraining dynamics of multilingual transformers. In *EMNLP*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

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

Jeffrey L. Elman. 1990. Finding structure in time. In *Cognitive Science*, pages 179–211.

Google DeepMind Team. 2024. Introducing gemini 1.5: Efficient multimodal models with extended context. *arXiv preprint*.

Dan Hendrycks and 1 others. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural Computation*, pages 1735–1780. MIT Press.

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. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*.

Aaron Hurst and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Woosuk Kwon and 1 others. 2023. Efficient memory management for large language model inference. *arXiv preprint arXiv:2309.06180*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Y Lu and 1 others. 2018. Analyzing linguistic knowledge in sequential models of sentence. *arXiv preprint*.

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

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Devendra Sachan and 1 others. 2025. Knowledge neurons in pretrained transformers. *arXiv preprint*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. 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)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama2: Open foundation and fine-tuned chat models. *arXiv preprint*.

Jesse Vig. 2019. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR workshop: Debugging machine learning models*, volume 3.

F. Zhang and 1 others. 2024. Towards best practices of activation patching in language models. In *ICLR (OpenReview)*.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. 2025. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. *arXiv preprint*.