

Lang-Prune: Unlocking Fair and Powerful Pruning for Multilingual Large Language Models

Juhao Liang^{1,*}, Shiqi Zhang^{1,2,*}, Min Zhang³, Hao Yang³, Benyou Wang^{1,2,†}

¹The Chinese University of Hong Kong, Shenzhen

²Shenzhen Loop Area Institute

³Huawei

Abstract

Multilingual large language models (LLMs) are increasingly deployed across languages under tight memory and latency budgets, yet pruning them with mixed-language reference data can induce *cross-lingual interference*: decisions that look good on average disproportionately damage particular languages. We introduce **Lang-Prune**, a drop-in extension to structured pruning that computes importance per language on small reference sets and aggregates via a Max operator—preserving any structure critical to at least one language. On aya-expanse-8b across nine typologically diverse languages, Lang-Prune reduces average perplexity by 62% at 70% sparsity relative to mixed-data pruning, and outperforms monolingual pruning on average under the same reference-data budget. Lang-Prune generalizes across model families and scales: on Qwen3-8B, it reduces average IF from $9.00\times$ to $0.66\times$ at 50% sparsity and remains effective up to Qwen3-14B. Post-training experiments further show preserved adaptation capacity, with pruned models reaching 59.6% vs. 50.6% accuracy on Belebele. Lang-Prune modifies only the importance estimation and aggregation step, preserving LLM-Pruner’s coupled-structure mechanics and requiring no additional parameters or retraining.

1 Introduction

With the rapid advancement of large language models (LLMs) (OpenAI et al., 2024; Touvron et al., 2023), people worldwide are increasingly benefiting from this technology. To further broaden its impact, researchers have devoted substantial effort to collecting low-resource language data and developing multilingual LLMs with larger parameter scales and stronger capabilities (Chen et al., 2023; Yang et al., 2025). However, the massive size of these models imposes heavy computational and memory demands, restricting their deployment in resource-constrained environments such as mobile devices and causing significant latency in client-server interactions. Model pruning has emerged as a practical solution to alleviate these challenges by removing redundant parameters while striving to maintain performance, particularly when adapting an existing, well-aligned model to resource-constrained deployments (Wang et al., 2020; Xia et al., 2022; Muralidharan et al., 2024; Xia et al.; Kong et al., 2025).

Currently, most existing pruning methods are evaluated primarily on monolingual or high-resource languages, often neglecting the cross-lingual variability inherent in multilingual LLMs (Sun et al., 2024; Frantar & Alistarh, 2023). Our pilot study reveals a central obstacle: *cross-lingual interference*. When pruning with mixed-language reference data, decisions optimized for average case disproportionately harm certain languages. Using the *interference factor* (IF), defined as the ratio of mixed vs. monolingual perplexity per language, we find strong degradation under structured pruning (LLM-Pruner (Ma et al., 2023)) and a smaller

*Equal contribution; †Corresponding to: wangbenyou@cuhk.edu.cn. All code and data associated with the project can be accessed at <https://github.com/Juhaoliang1997/LangPrune>

high-sparsity effect under unstructured pruning (SparseGPT (Frantar & Alistarh, 2023)), confirming that structured coupling amplifies cross-lingual interference.

We introduce *Lang-Prune*, a language-aware extension to LLM-Pruner that estimates importance per language on small reference sets and aggregates these scores via a Max operator to protect units critical to any language. *Lang-Prune* preserves LLM-Pruner’s coupled-structure mechanics and sparsity schedules, modifying only the scoring and aggregation process. We conduct primary evaluation on aya-expense-8b, a purpose-built multilingual model, and further evaluate on the Qwen3 family (0.6B–14B) to assess generalization across scales and model types, and to measure downstream task performance. Experiments across nine languages and multiple sparsity ratios show that *Lang-Prune* consistently improves both average and worst-case performance, often surpassing monolingual pruning even when using the same total reference-data budget. Interpretability analyses indicate that it better retains language-specific capacity. Extensive ablation studies further demonstrate that *Lang-Prune* (1) *adapts to multiple model families and scales, with advantages amplifying from mid-size to large models (0.6B–14B)*, (2) *preserves the potential of pruned LLMs for post-training*, and (3) *exhibits strong transfer to out-of-distribution languages*. Overall, this work presents a practical framework for multilingual LLM compression while preserving language coverage.

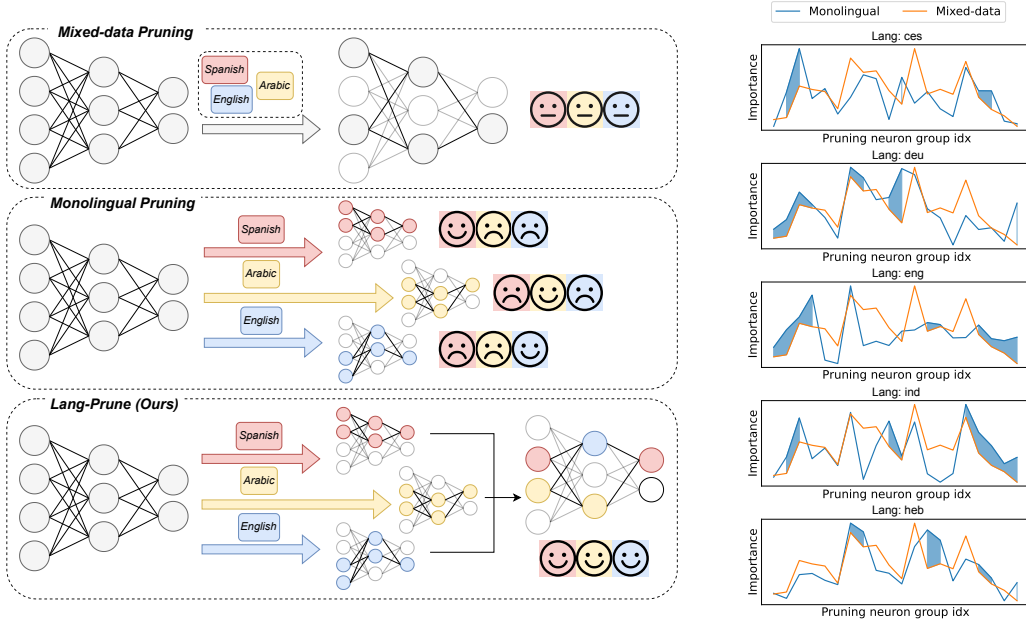


Figure 1: **Left:** Illustration of three pruning strategies—*Mixed-data Pruning*, *Monolingual Pruning*, and *Lang-Prune*. **Right:** Language-specific neuron-group importance under monolingual vs. mixed-language reference data. Shaded areas indicate neurons important for a single language but less salient under multilingual pruning.

This paper makes the following contributions:

1. We introduce *Lang-Prune*, a language-aware pruning framework that computes per-language importance and aggregates it via a Max operator to mitigate cross-lingual interference while preserving a deployment-friendly structure.
2. We demonstrate consistent improvements across nine typologically diverse languages, multiple sparsity levels, and model families (0.6B–14B), with gains amplifying at larger scales (e.g., avg. IF $8.04\times \rightarrow 0.46\times$ at Qwen3–14B).
3. We provide interpretability analyses revealing *Lang-Prune* retains language-specific neuron groups, offering insight into multilingual capacity preservation during compression.

LLM-Pruner under 70% Sparsity										
PPL score ↓	ara	ces	deu	eng	spa	ind	heb	rus	zho	avg.
Monolingual	42.71	81.99	82.82	236.44	85.19	77.52	46.87	57.73	36.47	83.08
Mixed-data	80.99	164.12	245.65	378.39	227.26	173.63	71.97	127.51	226.88	188.49
IF score	1.90×	2.00×	2.97×	1.60×	2.67×	2.24×	1.54×	2.21×	6.22×	2.27×
Base Model	8.10	10.46	10.66	10.50	9.58	12.77	11.19	11.13	10.27	10.52

SparseGPT under 70% Sparsity										
PPL score ↓	ara	ces	deu	eng	spa	ind	heb	rus	zho	avg.
Monolingual	50.03	94.52	65.40	63.72	28.47	66.64	63.52	44.57	100.19	64.12
Mixed-data	67.63	113.97	78.55	64.09	28.51	71.18	85.78	49.53	140.54	77.75
IF score	1.35×	1.21×	1.20×	1.01×	1.00×	1.07×	1.35×	1.11×	1.40×	1.21×
Base Model	8.10	10.46	10.66	10.50	9.58	12.77	11.19	11.13	10.27	10.52

Table 1: Cross-lingual Interference Factor (IF) for multilingual pruning. Values greater than 1 indicate **performance degradation** due to cross-lingual interference. The avg. IF cell is computed from average PPL, i.e., $\text{AvgPPL}_{\text{mixed}} / \text{AvgPPL}_{\text{mono}}$.

2 Lang-Prune: A Multilingual Pruning Framework

2.1 Pilot Study: Cross-lingual Interference on Pruning

We conduct a pilot study to quantify how existing pruning methods behave in multilingual settings. We evaluate a structured pruning method, LLM-Pruner (Ma et al., 2023), and an unstructured pruning method, SparseGPT (Frantar & Alistarh, 2023), on aya-expanse-8b (Dang et al., 2024). The experimental settings match those in the main experiments (see Section 3). In the Mixed-data Pruning setting, we pool all nine languages (900 sequences total, uniformly sampled per language); in the Monolingual Pruning setting, pruning is performed separately for each language using 100 sequences, yielding nine language-specific pruned checkpoints. Note that both settings use the same total reference-data budget (900 sequences), ensuring that observed differences reflect reference-data strategy rather than data quantity. To quantify cross-lingual interference, we report the *Cross-lingual Interference Factor* (IF) defined as $\text{IF}(l) = \text{PPL}_{\text{mixed}}(l) / \text{PPL}_{\text{mono}}(l)$, where values above 1 indicate performance degradation due to mixed-language reference data.

As shown in Table 1, Mixed-data Pruning consistently underperforms Monolingual Pruning across all nine languages and both pruning paradigms. For structured pruning (LLM-Pruner), the average IF is $2.27\times$, with particularly severe degradation for zho ($6.22\times$), deu ($2.97\times$), and spa ($2.67\times$); unstructured pruning (SparseGPT) shows a smaller but still visible effect at 70% sparsity (avg-column IF $1.21\times$). Notably, zho shows the largest IF under both paradigms yet differs by more than $4\times$ between structured ($6.22\times$) and unstructured ($1.40\times$) pruning—direct evidence that structured coupling concentrates and amplifies language-specific interference. Figure 1 (right) reveals the mechanism: under mixed-language reference data, language-specific importance peaks become attenuated or misaligned, causing critical structures to appear less salient—a phenomenon we term *cross-lingual interference* in pruning.

2.2 Overview of Lang-Prune

To address cross-lingual interference and achieve balanced pruning across languages, we introduce **Lang-Prune**, a multilingual-aware extension to LLM-Pruner guided by *language-specific importance scores*. As shown in the left panel of Figure 1, unlike conventional approaches that estimate neuron importance from a single reference set, Lang-Prune computes importance scores per language and aggregates them using fairness-aware policies, ensuring that structures critical to any language are preserved during pruning.

This design is motivated by evidence that multilingual LLMs contain both universal and language-specific mechanisms: some units encode language-agnostic patterns, while others capture script- or morphology-dependent features (Singh et al., 2019; Liu et al., 2019; Conneau et al., 2020; Zheng et al., 2025). Building on these insights, we extend LLM-Pruner’s structured pruning pipeline to explicitly respect per-language signals. The process consists of three stages: (1) coupled structure discovery; (2) language-aware importance estimation; (3) multilingual importance scores aggregation and pruning.

2.2.1 Coupled Structure Discovery in LLM-Pruner

The LLM-Pruner framework of Ma et al. (2023) discovers valid, shape-consistent pruning units by constructing a dependency graph over Transformer components and grouping parameters that must be pruned jointly (coupled structures). This enables structured pruning of attention heads and feed-forward (MLP) channels without breaking tensor shapes or deployment compatibility.

For example, in a feed-forward block, pruning the j -th hidden channel requires removing the j -th column of the input projection and the j -th row of the output projection simultaneously. LLM-Pruner treats this pair as a single coupled structure, denoted $\mathcal{C}_j$, and assigns it a score based on parameter magnitude and activation statistics collected on reference data. This coupled view generalizes across modules and forms the backbone of Lang-Prune.

Lang-Prune extends this framework by modifying how scores are computed and aggregated: instead of using global metrics, scores are calculated per language and then combined to guide multilingual pruning. Lang-Prune is designed for structured pruning methods that operate over functionally meaningful units and is not intended for unstructured methods such as Wanda (Sun et al., 2024), where individual weight removal does not preserve functional components (Appendix A.1). As that ablation confirms, preserving per-language importance at the level of functional components (e.g., attention heads, MLP channels) is far more impactful than focusing on isolated weights. Therefore, **the coupled structure discovery mechanism inherited from LLM-Pruner is essential for Lang-Prune**, as it ensures that language-specific importance scores meaningfully influence pruning decisions while maintaining model deployability.

2.2.2 Language-aware Importance Estimation

Given the set of coupled structures $\{\mathcal{C}_j\}$, we estimate their importance independently for each language using a loss-sensitivity criterion computed on language-specific reference subsets. Let $\mathcal{L}$ denote the set of languages and $\mathcal{D}_\ell$ the reference data for language ℓ . For a coupled structure $\mathcal{C}_j$, let $\mathcal{G}(\mathcal{C}_j)$ be the collection of learnable tensors (or parameter vectors) that constitute the structure (e.g., paired columns/rows in coupled projections). Following the Taylor-based importance estimator of Ma et al. (2023), we measure the first-order contribution of each parameter to the next-token prediction loss.

Concretely, let $L(x)$ denote the token-averaged negative log-likelihood (next-token prediction) for input x . For a scalar parameter $w \in \mathcal{C}_j$, the per-parameter, per-language importance is

$$s_\ell(w) = \mathbb{E}_{x \sim \mathcal{D}_\ell} \left| \frac{\partial L(x)}{\partial w} \cdot w \right|, \quad (1)$$

where the absolute value ensures non-negativity and robustness to sign cancellations. The expectation is approximated by the empirical mean over the reference subset, with the model in evaluation mode (no dropout) and gradients computed at the current weights without updating them.

When treating an entire tensor $W \in \mathcal{G}(\mathcal{C}_j)$ as a unit, we use the vectorized form

$$s_\ell(W) = \mathbb{E}_{x \sim \mathcal{D}_\ell} \left| \left\langle \frac{\partial L(x)}{\partial W}, W \right\rangle \right|, \quad (2)$$

$$\text{where } \langle A, B \rangle = \text{tr}(A^\top B).$$

i.e., the absolute value of the Frobenius inner product between the gradient and the parameter tensor.

We aggregate the parameter- or tensor-level scores into a structure-level importance using a group aggregator $\mathcal{A}$ over all elements of $\mathcal{G}(\mathcal{C}_j)$. By default, we adopt the sum aggregator:

$$I_\ell(\mathcal{C}_j) = \sum_{u \in \mathcal{G}(\mathcal{C}_j)} s_\ell(u), \quad (3)$$

where u denotes either a scalar weight (using Eq. 1) or a tensor (using Eq. 2). Other aggregators (e.g., max, mean, first) are supported and evaluated in ablations in Appendix A.2.

To make scores comparable across languages, we apply per-language min–max normalization over all coupled structures:

$$\tilde{I}_\ell(\mathcal{C}_j) = \frac{I_\ell(\mathcal{C}_j) - \min_k I_\ell(\mathcal{C}_k)}{\max_k I_\ell(\mathcal{C}_k) - \min_k I_\ell(\mathcal{C}_k) + \varepsilon}, \quad (4)$$

where the extrema are computed over all coupled structures indexed by k , and a small ε (e.g., 10^{-12}) ensures numerical stability. In the rare degenerate case where $\max_k I_\ell(\mathcal{C}_k) \approx \min_k I_\ell(\mathcal{C}_k)$, this normalization yields $\tilde{I}_\ell(\mathcal{C}_j) \approx 0$ for all j , effectively indicating no preference among structures for language ℓ . This language-wise normalization mitigates the effects of tokenization, script differences, and frequency imbalance, while preserving the importance ranking within each language.

2.2.3 Pruning LLMs with Multilingual Importance Scores

Lang-Prune merges per-language importances into a single score that guards against worst-case language degradation. Given $\{\tilde{I}_\ell(\mathcal{C}_j)\}_{\ell \in \mathcal{L}}$, we use the **Max** aggregator:

$$I_{\max}(\mathcal{C}_j) = \max_{\ell \in \mathcal{L}} \tilde{I}_\ell(\mathcal{C}_j). \quad (5)$$

This “any-language” criterion preserves structures that are important for at least one language, directly countering the dilution effect observed in mixed-language reference data. For comparison in ablations, we also consider the **Mean** and **Min** aggregators:

$$I_{\text{mean}}(\mathcal{C}_j) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \tilde{I}_\ell(\mathcal{C}_j), \quad I_{\min}(\mathcal{C}_j) = \min_{\ell \in \mathcal{L}} \tilde{I}_\ell(\mathcal{C}_j). \quad (6)$$

Structures are ranked by $I_{\max}$ (our default) and pruned (lowest first) until the target sparsity is reached, strictly respecting LLM-Pruner’s coupling constraints.

By explicitly estimating importance per language and aggregating with a Max policy, Lang-Prune avoids pruning decisions that are optimal on average yet harmful to minority or script-diverse languages. The framework is compute-efficient—importance estimation scales linearly with $|\mathcal{L}|$ using small reference subsets—and is a drop-in multilingual extension of LLM-Pruner: sparsity schedules and prunable units remain unchanged, only the scoring and aggregation are revised.

3 Experiments

Experiment Settings We conduct primary evaluation on aya-expanse-8b (Dang et al., 2024), a purpose-built multilingual model providing a controlled testbed for cross-lingual pruning behavior and interpretability analysis. We evaluate at 30%, 50%, and 70% global sparsity against LLM-Pruner under both monolingual and multilingual reference-data settings. We sample 100 sequences of length 128 per language from mC4 (Xue et al., 2021) across nine languages (ara, ces, deu, eng, spa, ind, heb, rus, zho), giving 900 sequences total—equally shared between multilingual and monolingual settings, so comparisons reflect reference-data *strategy* rather than data quantity. Pruning is one-shot with no recovery. We evaluate PPL on the mC4 validation split, computed as the exponentiated mean token-level negative log-likelihood using teacher forcing. Structured pruning at 70% sparsity reduces peak GPU memory by $\sim 49\%$ (15.10 GB $\rightarrow$ 7.77 GB) with no additional inference-time overhead relative to LLM-Pruner at the same sparsity; full efficiency numbers are in Appendix A.6.1. Further evaluation across model families and scales is in Section 4. All runs use a single NVIDIA A100 GPU, with each pruning instance requiring less than one GPU hour.

3.1 Results Analysis

Table 2 reports per-language PPL across three sparsity levels for aya-expanse-8b.¹ Since absolute PPL values can differ across languages due to tokenization, vocabulary, and corpus

¹Detailed monolingual pruning results in Table 12, Appendix A.3.1.

Method	Reference Data	ara	ces	deu	eng	spa	ind	heb	rus	zho	Avg. ↓
Original (8B)	None	8.10	10.46	10.66	10.50	9.58	12.77	11.19	11.13	10.27	10.52
<i>aya-expanses-8b with 30%</i>											
LLM-Pruner	monolingual	11.10	15.54	15.10	30.50	16.38	15.60	13.58	13.76	12.84	16.04
	mixed-data	11.57	16.93	16.11	30.36	18.51	17.54	14.28	14.69	16.97	17.44
SparseGPT	mixed-data	21.42	33.54	24.42	24.31	13.33	35.35	29.03	17.99	47.83	27.47
Lang-Prune-Avg	multilingual	11.47	15.80	14.76	24.78	16.74	16.25	14.25	14.13	15.71	15.99
Lang-Prune-Min	multilingual	354.12	612.46	284.30	66.85	182.67	262.45	1925.36	675.20	562.30	547.30
Lang-Prune-Max	multilingual	11.22	15.15	13.68	21.52	15.73	15.35	13.98	14.09	13.35	14.90
<i>aya-expanses-8b with 50%</i>											
LLM-Pruner	monolingual	17.48	28.34	28.39	70.76	29.64	26.71	21.10	22.09	17.93	29.16
	mixed-data	21.56	37.84	43.30	103.49	49.98	40.83	24.73	22.55	43.85	43.90
SparseGPT	mixed-data	24.88	38.97	29.91	27.44	14.78	39.25	34.34	20.80	56.01	31.82
Lang-Prune-Avg	multilingual	20.73	31.11	32.68	71.03	35.65	31.97	24.34	24.74	38.61	34.54
Lang-Prune-Min	multilingual	757.21	1606.93	864.86	229.03	666.20	1053.59	26134.73	2618.83	1924.89	3984.03
Lang-Prune-Max	multilingual	17.99	25.63	23.02	49.20	26.52	25.58	22.22	22.78	20.30	25.91
<i>aya-expanses-8b with 70%</i>											
LLM-Pruner	monolingual	42.71	81.99	82.82	236.44	85.19	77.52	46.87	57.73	36.47	83.08
	mixed-data	80.99	164.12	245.65	378.39	227.26	173.63	71.97	127.51	226.88	188.49
SparseGPT	mixed-data	67.63	113.97	78.55	64.09	28.51	71.18	85.78	49.53	140.54	77.75
Lang-Prune-Avg	multilingual	78.57	126.19	185.68	292.57	179.81	144.15	66.51	95.47	251.11	157.78
Lang-Prune-Min	multilingual	1588.60	2890.65	1973.42	579.72	1334.47	5097.47	56508.07	6862.74	3904.55	8960.00
Lang-Prune-Max	multilingual	44.03	69.71	64.85	158.33	72.53	70.08	51.62	60.29	46.18	70.85

Table 2: Per-language perplexity (PPL, lower is better) after pruning *aya-expanses-8b*. Lang-Prune variants merge per-language importance via Max (**Lang-Prune-Max**), Mean (**Lang-Prune-Avg**), or Min (**Lang-Prune-Min**) aggregator. For comparison, LLM-Pruner is evaluated under monolingual and multilingual (mixed-data) reference-data settings, and SparseGPT is included as an unstructured mixed-data baseline.

statistics, we use raw PPL primarily for within-language comparisons and rely on IF or relative degradation for cross-language interpretation. We analyze results along three dimensions:

1. Overall performance improvement. Lang-Prune-Max consistently outperforms LLM-Pruner under multilingual reference data for every language and sparsity. Relative to LLM-Pruner mixed-data, it reduces average PPL by 14.6% (30%), 41.0% (50%), and 62.4% (70%). Compared to LLM-Pruner monolingual under the same total reference-data budget, Lang-Prune-Max achieves lower average PPL at all sparsities (14.90 vs 16.04 at 30%; 25.91 vs 29.16 at 50%; 70.85 vs 83.08 at 70%). SparseGPT exhibits milder interference at high sparsity but remains an unstructured baseline: its mixed-data average PPL is 27.47, 31.82, and 77.75 at 30%, 50%, and 70% sparsity, while standard dense kernels do not automatically realize proportional memory or throughput gains from unstructured masks. We therefore use it as an informative comparison, while focusing Lang-Prune on deployment-friendly structured pruning.

2. Worst-case language improvement. At 70% sparsity, the worst-language PPL decreases from 378.39 (LLM-Pruner mixed-data) and 236.44 (LLM-Pruner monolingual) to 158.33 (Lang-Prune-Max), with similar trends at lower sparsities. Notably, English consistently exhibits the highest post-pruning PPL despite being the dominant pretraining language; cross-corpus evaluation on Wikitext-2 confirms this is a genuine model property, with Lang-Prune-Max achieving 23.5% lower English PPL (Appendix A.3.2).

3. Lang-Prune-Min (negative control). With the Min aggregator, most languages suffer severe degradation while English remains robust, indicating that even language-agnostic neurons carry residual bias toward dominant languages and highlighting the necessity of a language-aware aggregation strategy.

These findings are further corroborated across model families, scales, and downstream tasks in Section 4.

3.2 Analysis of Pruning Neuron Groups

To compare the inner mechanisms of LLM-Pruner and Lang-Prune, we analyze the retention of language-specific capacity at high sparsity. We first identify, on the unpruned model, the set of strong language-related neuron groups per language and then measure their recall

Strategy	rus	heb	ind	spa	eng	deu	ces	ara	zho	Avg.
Mixed-data	67.47%	69.20%	67.43%	65.06%	60.06%	66.22%	69.68%	66.79%	61.74%	65.96%
Lang-Prune (ours)	81.94%	82.86%	81.62%	80.68%	75.41%	81.71%	83.11%	81.27%	80.78%	81.04%

Table 3: Recall ratio of strong language-related neuron groups (Top- $p\%$ = 30%) under different pruning strategies at 70% sparsity. Higher is better.

under different pruning strategies. Concretely, let $\{\mathcal{C}_j\}$ denote the coupled MLP channels, where j uniquely identifies each candidate structure, and $\tilde{I}_\ell(\mathcal{C}_j)$ the per-language, min-max normalized importance from Section 2. For each language ℓ , we define a contrastive specificity score for structure $\mathcal{C}_j$:

$$S_\ell(\mathcal{C}_j) = \tilde{I}_\ell(\mathcal{C}_j) - \frac{1}{|\mathcal{L}| - 1} \sum_{\ell' \neq \ell} \tilde{I}_{\ell'}(\mathcal{C}_j).$$

We construct the set of strong language-related groups by selecting the top- $p\%$ structures per language according to S_ℓ (with a fixed p across languages). The recall ratio for language ℓ under a pruning strategy is defined as

$$\text{Recall}_\ell = \frac{|\{\mathcal{C}_j \in \text{Top-}p\% \text{ for } \ell\} \cap \{\mathcal{C}_j \text{ retained after pruning}\}|}{|\{\mathcal{C}_j \in \text{Top-}p\% \text{ for } \ell\}|}.$$

This metric quantifies how well a pruning method preserves neuron groups most specialized to each language, independent of the method’s own scoring.

Table 3 reports the recall ratio at 70% sparsity with Top- $p\%$ = 30% across nine languages. Compared with multilingual mixed-data pruning (LLM-Pruner mixed-data), Lang-Prune consistently retains a larger fraction of language-specific neurons, improving the average recall from 65.96% to 81.04%. Gains are broad (e.g., zho: 61.74% $\rightarrow$ 80.78%; ces: 69.68% $\rightarrow$ 83.11%), indicating that Max aggregation protects structures that are critical to any language rather than optimizing for average-case activation. These retention improvements align with the perplexity results in Table 2, and are most pronounced in languages that showed higher interference under mixed-language reference data.

4 Ablation Studies

4.1 Extended Evaluation on the Qwen3 Family

To assess generalization beyond the primary model, we conduct a comprehensive evaluation on Qwen3 (Yang et al., 2025). Unlike aya-expanse-8b, Qwen3 is a general-purpose family spanning 0.6B–14B, enabling scale analysis across architectures. We replicate the pruning setup from Section 3 at 30%, 50%, and 70% sparsity with identical reference-data and evaluation protocols. In addition to PPL, we evaluate downstream task performance via few-shot and post-training benchmarks in Section 4.2.

	30% sparsity		50% sparsity		70% sparsity	
	Avg. PPL	Avg. IF	Avg. PPL	Avg. IF	Avg. PPL	Avg. IF
LLM-Pruner (monolingual)	15.76	—	36.94	—	388.69	—
LLM-Pruner (mixed-data)	16.00	1.06×	310.69	9.00×	53567.15	137.85×
Lang-Prune (ours)	14.12	0.90×	24.30	0.66×	174.74	0.45×

Table 4: Average PPL and Interference Factor (IF) on Qwen3-8B. Lang-Prune consistently outperforms both baselines and reduces IF below 1 at all sparsities, while LLM-Pruner (mixed-data) degrades catastrophically at high sparsity (avg IF 137.85× at 70%).

As shown in Table 4, (i) at 30% sparsity, Lang-Prune achieves avg IF 0.90, reducing PPL below the monolingual baseline in most languages, while LLM-Pruner (mixed-data) slightly degrades (avg IF 1.06). (ii) At 50%, LLM-Pruner (mixed-data) suffers severe interference (avg IF 9.00), whereas Lang-Prune maintains IF well below 1 (0.66). (iii) At 70%, LLM-Pruner

Model	LLM-Pruner (mixed-data)		Lang-Prune	
	Avg. PPL	Avg. IF	Avg. PPL	Avg. IF
Qwen3-0.6B	52.38	1.10×	56.07	1.18×
Qwen3-1.7B	80.96	2.22×	45.68	1.25×
Qwen3-4B	40.72	1.24×	31.43	0.96×
Qwen3-8B	310.69	9.00×	24.30	0.66×
Qwen3-14B	555.36	8.04×	31.60	0.46×

Table 5: Scaling behavior on the Qwen3 family at 50% sparsity. Avg. IF is computed against the corresponding LLM-Pruner monolingual baseline for each model size, using the same compact format as Table 4.

(mixed-data) collapses catastrophically (avg IF 137.85 $\times$), while Lang-Prune remains stable (avg IF 0.45). IF values below 1 do not imply that pruning improves over the unpruned model; they indicate that Qwen monolingual pruning is unstable at high sparsity, so Lang-Prune yields lower PPL than that monolingual pruning baseline. Table 5 reports the same IF metric across the 0.6B–14B scale range at 50% sparsity. Small models (0.6B–1.7B) are a transitional regime where structured redundancy is limited, but from 4B upward Lang-Prune reduces both average PPL and IF relative to mixed-data pruning. At 8B and 14B, mixed-data pruning reaches avg IF 9.00 $\times$ and 8.04 $\times$, whereas Lang-Prune reduces them to 0.66 $\times$ and 0.46 $\times$. Full per-language results and the 14B breakdown are provided in Appendix A.4. The Max aggregator also remains stable as reference-language count grows from 9 to 15 languages; at 25 languages the all-language average remains comparable, while the original core-9 set degrades only mildly (Appendix A.6.2).

Key Finding 1. *Lang-Prune generalizes effectively across model families, architectural designs, and scales (0.6B–14B), with its advantage over mixed-data pruning amplifying as model size increases.*

4.2 Post-Training Potential After Pruning

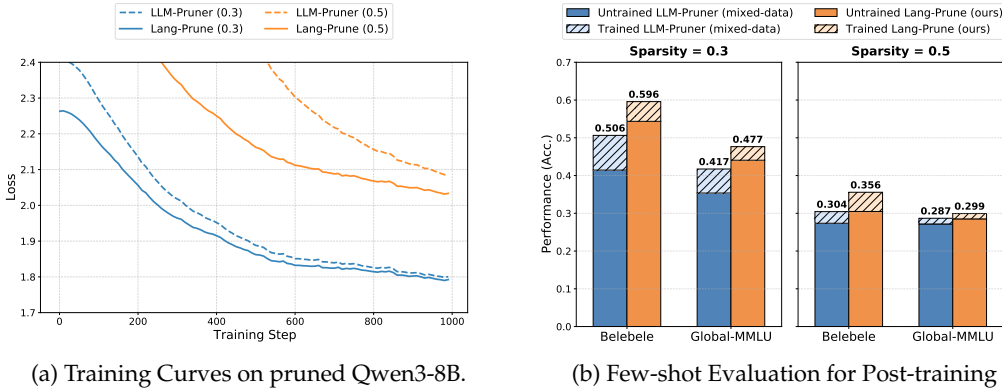


Figure 2: Post-training recovery of pruned models: (a) Training loss during continued pre-training of Qwen3-8B. Lang-Prune consistently achieves lower loss than LLM-Pruner across steps. (b) Few-shot accuracy before and after post-training for models pruned by LLM-Pruner and Lang-Prune. Lang-Prune yields higher pre- and post-training accuracy and larger recovery across languages.

Complementing the PPL analysis in Section 4.1, we evaluate whether Lang-Prune preserves post-training capacity on pruned Qwen3-8B models. Per-language few-shot results on raw and fine-tuned Qwen3-8B models are in Appendix A.5.1. We apply an identical fine-tuning budget (LoRA) to both methods and measure language modeling recovery and 3-shot downstream performance on Belebele (Bandarkar et al., 2023) and Global-MMLU (Singh et al., 2024) (Figure 2).

Language	Family / Script	Best mono	PPL (Mixed)	PPL (Mono)	PPL (Lang-Prune)	GR ↓
urd (Urdu)	Indo-Aryan / Arabic	heb	273.27	107.91	31.08	0.11
amh (Amharic)	Semitic / Ethiopic	zho	258.83	12.25	11.90	0.05
bul (Bulgarian)	Slavic / Cyrillic	rus	128.64	91.21	71.05	0.55
pol (Polish)	Slavic / Latin	ces	199.85	104.12	68.24	0.34
nld (Dutch)	Germanic / Latin	deu	410.68	178.78	104.86	0.26
fra (French)	Romance / Latin	spa	318.40	143.33	78.00	0.25
msa (Malay)	Austronesian / Latin	zho	14.46	8.06	7.64	0.53
jpn (Japanese)	Japonic / CJK	zho	608.95	158.51	155.66	0.26
kor (Korean)	Koreanic / Hangul	zho	346.27	89.62	76.57	0.22
vie (Vietnamese)	Austroasiatic / Latin	zho	303.86	149.50	76.76	0.25

Table 6: OOD generalization at 70% sparsity on aya-expanse-8b, evaluated on the mC4 validation split (representative languages; full 16-language table in Appendix A.5.2). $GR = PPL_{Lang-Prune} / PPL_{mixed}$; values below 1 favor Lang-Prune. The PPL(Mono) reports the best monolingual proxy among the nine in-distribution reference languages; Appendix A.5.2 provides the full 16-language table and proxy rankings.

As shown in Figure 2, Lang-Prune exhibits faster convergence and lower final perplexity than LLM-Pruner at both sparsity levels, indicating better preservation of language modeling capacity under limited post-training budgets. For downstream tasks, Lang-Prune also achieves consistently higher few-shot accuracy and larger post-training gains. For example, at 30% sparsity on Belebele, Lang-Prune + Training reaches 59.6% accuracy compared to 50.6% for LLM-Pruner after identical LoRA fine-tuning, reflecting a substantial improvement in a practical deployment regime.

Key Finding 2. *Lang-Prune not only mitigates cross-lingual interference during pruning, but also preserves the model’s capacity for effective post-training in both language modeling and downstream multilingual tasks.*

4.3 Generalization to Out-of-Distribution Languages

We evaluate zero-shot generalization to languages absent from the reference set (OOD). Using the 70% structured pruning setting, we compare: (i) LLM-Pruner with mixed-language reference data (mixed-data), (ii) the best monolingual proxy among the nine in-distribution languages (Best-mono), and (iii) Lang-Prune. We report per-language PPL and the generalization ratio $GR(l) = PPL_{Lang-Prune}(l) / PPL_{Data-mixed}(l)$, where values below 1 favor Lang-Prune.

As shown in Table 6, **(i) Strong OOD gains:** Lang-Prune consistently outperforms the mixed-data baseline across all OOD languages, with the largest gains for urd ($0.11\times$), amh ($0.05\times$), kor ($0.22\times$), and vie ($0.25\times$). **(ii) Outperforming monolingual proxies:** On average, Lang-Prune achieves $\sim 32\%$ lower PPL than the best monolingual proxy, suggesting that preserving structures shared across in-distribution languages transfers better than committing to a single source. **(iii) Script affinity:** OOD languages often align with proxies sharing script and tokenization rather than genealogical family—Cyrillic with rus, CJK with zho, Arabic-script with ara/heb—while the full proxy rankings in Appendix A.5.2 show where genealogically closer alternatives are competitive. Full results for all 16 OOD languages are in Appendix A.5.2.

Key Finding 3. *Lang-Prune improves OOD generalization over mixed and monolingual pruning, with transfer driven more by script and tokenization overlap than by language family.*

4.4 Conclusion on Ablation Studies

Beyond the primary ablation findings, we also observe the following. **(4)** Applying Lang-Prune in multi-task or multi-domain settings yields only marginal improvements, which we attribute to much weaker structural separation between tasks than between languages (see Appendix A.7). **(5)** Lang-Prune achieves consistent gains across different reference-data sizes. Specifically, increasing the monolingual reference set from 10 to 200 sequences improves performance but still falls short of Lang-Prune (see Appendix A.8). **(6)** The Max aggregator remains stable as the reference-language count grows from 9 to 15 languages;

at 25 languages, the all-language average remains comparable while the original core-9 subset shows only mild and non-catastrophic degradation, confirming the 9-language paper setting is not artificially narrow (see Appendix A.6.2).

5 Related Works

LLM Pruning and Multilingual Extensions

Pruning reduces inference cost by removing parameters while preserving accuracy, spanning unstructured methods (Sun et al., 2024; Frantar & Alistarh, 2023) that achieve high sparsity via activation-aware criteria, and structured methods (Michel et al., 2019; Lagunas et al., 2021; Fan et al., 2019; Sajjad et al., 2023) that remove entire components for deployment-friendly speedups, with the LLM-Pruner framework of Ma et al. (2023) formalizing the latter via dependency graphs. Most methods are language-agnostic. Prior work reports heterogeneous pruning effects across languages (Ogueji et al., 2022). Language-balanced reference data can help at moderate sparsity but typically retains a shared importance criterion: the Multilingual Brain Surgeon (MBS) approach of Zeng et al. (2024) varies data mixture ratios, and Kurz et al. (2026) analyze how reference-language choice affects multilingual pruning outcomes. We include an MBS-style mixture sweep in Table 28, which shows that rebalancing data helps but remains far behind per-language importance aggregation. M-Wanda (Choenni & Titov, 2025) is a related multilingual unstructured pruning method; it operates at weight level, whereas Lang-Prune targets coupled functional structures that produce directly smaller dense models. Alignment-informed approaches introduce bilingual supervision to guide pruning (Kim et al., 2024); however, such methods operate in the unstructured paradigm on different model families, making direct comparison with structured methods like Lang-Prune architecturally incompatible. Lang-Prune differs from these approaches by maintaining per-language importance estimates and aggregating them via a Max operator, directly targeting cross-lingual interference within a deployment-friendly structured pruning framework.

Language-Specific vs. Shared Structure in Multilingual Models Multilingual transformers exhibit both shared and language-specific representations: lower and middle layers capture form and morpho-syntax, while higher layers are more semantic and language-agnostic (Belingkov & Glass, 2019; Tenney et al., 2019; Pires et al., 2019; Conneau et al., 2020). Neurons and subspaces often align with scripts or linguistic features (Singh et al., 2019; Liu et al., 2019), and preserving language-specific components benefits cross-lingual transfer (Choenni et al., 2022). Recent work further confirms that neurons in multilingual LLMs exhibit clear task- and language-specific specialization (Tan et al., 2024; Wang et al., 2025), providing empirical grounding for Lang-Prune’s design principle of protecting units critical to any individual language. Lang-Prune operationalizes this insight by protecting units critical to any language while pruning universally unimportant structures, outperforming uniform or MBS-style reference-data mixtures in one-shot structured pruning (Appendix A.9).

6 Conclusion

In this paper, we identified cross-lingual interference in pruning multilingual LLMs and proposed *Lang-Prune*, which computes per-language importance scores and aggregates them via a Max operator to protect parameters critical to any individual language. On aya-expansive-8b, Lang-Prune outperforms mixed-reference-data pruning and surpasses monolingual pruning under the same reference-data budget, while retaining higher recall of language-specific neurons and generalizing to out-of-distribution languages. Extended evaluation on the Qwen3 family (0.6B–14B) confirms that its advantage amplifies with model size and that pruned models preserve greater post-training adaptation capacity. Lang-Prune modifies only the importance estimation and aggregation step, requiring no architectural changes or additional parameters, making it a practical drop-in solution for multilingual LLM compression.

Acknowledgments

This work was supported by Major Frontier Exploration Program (Grant No. C10120250085) from the Shenzhen Loop Area Institute under grant 2026SFP004, Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen Medical Research Fund (B2503005), NSFC grant 72495131, the 1+1+1 CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund, Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001), and the International Science and Technology Cooperation Center, Ministry of Science and Technology of China (under grant 2024YFE0203000).

References

- 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, 2023.
- Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey, 2019. URL <https://arxiv.org/abs/1812.08951>.
- Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. Phoenix: Democratizing chatgpt across languages, 2023. URL <https://arxiv.org/abs/2304.10453>.
- Rochelle Choenni and Ivan Titov. M-wanda: Improving one-shot pruning for multilingual llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26939–26952, 2025.
- Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. Data-efficient cross-lingual transfer with language-specific subnetworks, 2022. URL <https://arxiv.org/abs/2211.00106>.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL <https://aclanthology.org/2020.acl-main.536/>.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermiş, Ahmet Üstün, and Sara Hooker. Aya expand: Combining research breakthroughs for a new multilingual frontier, 2024. URL <https://arxiv.org/abs/2412.04261>.
- Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019.
- Wikimedia Foundation. Wikimedia downloads. URL <https://dumps.wikimedia.org>.
- Elias Frantar and Dan Alistarh. Sparsegpt: massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML’23*, pp. 10323–10337. PMLR, 2023.
- Hwichan Kim, Jun Suzuki, Tosho Hirasawa, and Mamoru Komachi. Pruning multilingual large language models for multilingual inference. In Yaser Al-Onaizan, Mohit Bansal, and

- Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9921–9942, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.580. URL <https://aclanthology.org/2024.findings-emnlp.580/>.
- Jun Kong, Xinge Ma, Jin Wang, and Xuejie Zhang. Sample-aware adaptive structured pruning for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17938–17946, 2025.
- Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. On the limitations of language-targeted pruning: Investigating the calibration language impact in multilingual llm pruning. *Transactions of the Association for Computational Linguistics*, 14:167–192, 2026. doi: 10.1162/TACL.a.599.
- François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. Block pruning for faster transformers. *arXiv preprint arXiv:2109.04838*, 2021.
- Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In Jill Burstein, Christy Doran, and Tamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1073–1094, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1112. URL <https://aclanthology.org/N19-1112/>.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, 2023.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.
- Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024.
- Kelechi Ogueji, Orevaoghene Ahia, Gbemileke Onilude, Sebastian Gehrmann, Sara Hooker, and Julia Kreutzer. Intriguing properties of compression on multilingual models. 2022. doi: 10.48550/ARXIV.2211.02738. URL <https://arxiv.org/abs/2211.02738>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florenzia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis,

- Kyle Kotic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL <https://aclanthology.org/P19-1493/>.
- Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429, 2023.
- Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. BERT is not an interlingua and the bias of tokenization. In Colin Cherry, Greg Durrett, George Foster, Reza Haffari, Shahram Khadivi, Nanyun Peng, Xiang Ren, and Swabha Swayamdipta (eds.), *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 47–55, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6106. URL <https://aclanthology.org/D19-6106/>.
- Shivalika Singh, Angelika Romanou, Clémentine Fourier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermiş, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024. URL <https://arxiv.org/abs/2412.03304>.
- Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models, 2024. URL <https://arxiv.org/abs/2306.11695>.
- Shaomu Tan, Di Wu, and Christof Monz. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation, 2024. URL <https://arxiv.org/abs/2404.11201>.

- Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019. URL <https://arxiv.org/abs/1905.05950>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. Sharing matters: Analysing neurons across languages and tasks in llms, 2025. URL <https://arxiv.org/abs/2406.09265>.
- Ziheng Wang, Jeremy Wohlwend, and Tao Lei. Structured pruning of large language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.496. URL <http://dx.doi.org/10.18653/v1/2020.emnlp-main.496>.
- Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*.
- Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models, 2022. URL <https://arxiv.org/abs/2204.00408>.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 483–498, 2021.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 11794–11812, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1030/>.
- Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts, 2025. URL <https://arxiv.org/abs/2410.10626>.

A Appendix

A.1 Scope of Lang-Prune: Structured vs. Unstructured Pruning

We conducted additional experiments to evaluate the behavior of Lang-Prune when applied to Wanda (Sun et al., 2024), an unstructured pruning method that removes weights or rows/columns without preserving structured computational units. Under 30% sparsity, we compared three strategies: **Mixed-data Pruning** (mixed 9-language reference data), **Monolingual Pruning** (per-language reference data), and **Lang-Prune (max aggregation)** applied on top of Wanda.

Lang-Prune performs poorly under Wanda, which is consistent with the design of multilingual importance estimation. Wanda operates at the level of individual weights or rows/columns, which do not correspond to coherent functional components such as MLP channels or attention heads. Preserving individual weights, even if important for a specific language, does not preserve the functional behavior of the network.

Wanda	Base Model	Mixed-data	Monolingual	Lang-Prune
Avg. PPL↓	11.02	11.26	11.25	16.56

Table 7: Average perplexity (PPL) on the mC4 validation split across the nine reference languages for Wanda under 30% sparsity. Lang-Prune shows poor performance in this unstructured setting, consistent with its reliance on structured pruning units.

Method	Paradigm	30%	50%	70%
Lang-Prune (ours)	Structured	14.90	25.91	70.85
LLM-Pruner (mixed)	Structured	17.44	43.90	188.49
SparseGPT (mixed)	Unstructured	27.47	31.82	77.75
Wanda (mixed)	Unstructured	28.20	36.30	250.80
M-Wanda	Unstructured + multilingual	31.18	130.47	unstable ($> 10k$)

Table 8: Average PPL across the nine reference languages on aya-expense-8b for structured and unstructured pruning baselines. M-Wanda is run with the official default setting ($\lambda = 0.2$ for cross-layer weighting). The comparison shows that the language-aware signal is most effective when applied to functional structured units rather than individual weights.

In contrast, Lang-Prune is intended for settings where pruning is applied over semantically meaningful structures. Methods such as LLM-Pruner, head-pruning, channel-pruning, or block-level structured pruning expose units that correspond to functional submodules. In these cases, per-language importance can be translated into actionable preservation across languages. Table 8 further shows that Lang-Prune outperforms unstructured baselines on average across sparsity levels while producing dense compressed models.

Overall, these results demonstrate that Lang-Prune is compatible with structured pruning methods, but unstructured methods like Wanda fall outside this scope by design. Effective multilingual pruning requires preserving importance at the level of functional components rather than individual weights, which explains the performance difference observed with Wanda. M-Wanda (Choenni & Titov, 2025) is therefore complementary: it improves multilingual unstructured pruning, whereas Lang-Prune targets structured units that change dense tensor shapes and realize memory/throughput gains on standard kernels. Similarly, quantization is orthogonal to pruning because it reduces numerical precision rather than removing functional structures; Table 9 shows that GPTQ INT8 introduces negligible cross-lingual interference compared with pruning, and a Lang-Pruned model can still be quantized afterward in a compression pipeline.

A.2 Ablation Studies on Group Aggregator

We investigate how the choice of group aggregator $\mathcal{A}$ affects pruning behavior. Recall that $\mathcal{A}$ reduces parameter- or tensor-level importance scores within each coupled structure. In addition to our default sum aggregator, we evaluate $\mathcal{A} = \{\text{mean}, \text{max}, \text{first}\}$ (see Table 10 for details).

Method	Mixed PPL	Mono PPL	Avg. IF	Max per-lang IF
GPTQ INT8	26.63	26.59	$1.000\times$	$1.001\times$ (ces)
SparseGPT 70%	77.75	64.12	$1.213\times$	$1.403\times$ (zho)
LLM-Pruner 70%	188.49	83.08	$2.27\times$	$6.22\times$ (zho)

Table 9: Average PPL and IF comparison for quantization and pruning on aya-expense-8b. INT8 quantization shows negligible mixed-vs.-monolingual interference, while weight-removal methods, especially structured pruning, induce much larger cross-lingual interference.

Aggregator	Description
sum	accumulates the total contribution of all parameters.
mean	normalizes by group size to reduce the effect of large structures.
max	highlights structures where a single parameter dominates.
first	uses the first entry as a simplified representative value.

Table 10: Group Aggregator Description

Table 11 reports average perplexity across pruning strategies (Monolingual, Mixed-data, and Lang-Prune) using Qwen3-8B with 50% sparsity. All other settings follow Section 3.

Avg PPL ↓	sum	mean	max	first
Monolingual	36.72	33.28	20.74	381.29
Mixed-data	310.69	126.83	29.29	6122.22
Lang-Prune (ours)	24.30	24.24	22.53	45.23

Table 11: Ablation on group aggregators. Lower is better.

These results show that although alternative aggregators may work reasonably in certain settings, sum and mean provide the most stable performance across multilingual scenarios. max offers more aggressive sparsity behavior, while first produces unstable behavior and is therefore not recommended.

A.3 Supplementary Experiment Results

A.3.1 Per-language Monolingual Pruning Detail

Table 12 reports per-language PPL for aya-expanse-8b at 70% sparsity under all nine monolingual reference-data settings, alongside multilingual baselines and Lang-Prune variants. Each monolingual row uses only the indicated language as reference data, illustrating severe cross-language degradation when the reference language is mismatched.

Method	Reference Data	ara	ces	deu	eng	spa	ind	heb	rus	zho	Avg. ↓
Original (8B)	None	8.10	10.46	10.66	10.50	9.58	12.77	11.19	11.13	10.27	10.52
LLM-Pruner (70%)	ara	42.71	1933.60	1045.78	602.23	651.80	1361.45	297.98	1038.27	1121.41	899.47
	ces	347.90	81.99	236.12	339.14	244.47	272.87	883.65	130.36	998.25	392.75
	deu	497.99	503.61	82.82	365.89	309.53	342.55	2629.72	697.91	1426.27	761.81
	eng	1166.51	2090.46	1165.79	236.44	705.24	2062.33	64235.58	4802.09	3246.88	8856.81
	spa	404.32	532.11	422.02	340.05	85.19	354.68	1764.65	631.21	1376.49	656.75
	ind	333.15	482.84	328.63	293.57	252.14	77.52	1418.63	594.34	713.24	499.34
	heb	84.05	628.21	481.31	360.53	326.31	414.81	46.87	252.18	372.41	329.63
	rus	252.89	165.36	337.70	336.58	261.16	323.61	506.07	57.73	450.77	299.10
	zho	271.20	878.06	591.66	368.54	495.26	371.47	629.59	705.54	36.47	483.09
	monolingual	42.71	81.99	82.82	236.44	85.19	77.52	46.87	57.73	36.47	83.08
	mixed-data	80.99	164.12	245.65	378.39	227.26	173.63	71.97	127.51	226.88	188.49
Lang-Prune-Avg (70%)	multilingual	78.57	126.19	185.68	292.57	179.81	144.15	66.51	95.47	251.11	157.78
Lang-Prune-Min (70%)	multilingual	1588.60	2890.65	1973.42	579.72	1334.47	5097.47	56508.07	6862.74	3904.55	8960.00
Lang-Prune-Max (70%)	multilingual	44.03	69.71	64.85	158.33	72.53	70.08	51.62	60.29	46.18	70.85

Table 12: Per-language perplexity after 70% structured pruning of aya-expanse-8b. Each monolingual row uses only the indicated language as reference data, illustrating severe cross-language degradation under mismatched reference languages. Lang-Prune-Max uses multilingual reference data with Max aggregation.

A.3.2 Cross-corpus PPL Investigation

A potential concern is that the high post-pruning PPL for certain languages (notably English) reflects an mC4 evaluation artifact rather than genuine model degradation. We address this by evaluating pruned aya-expanse-8b checkpoints on multiple independent corpora per language.

English. Table 13 reports English PPL across two independent mC4 random seeds and Wikitext-2 (book and Wikipedia text, entirely different domain from mC4 web crawl).

Checkpoint	mC4-eng (seed 0)	mC4-eng (seed 1)	Wikitext-2
Unpruned aya-expanse-8b	23.74	29.93	24.51
LLM-Pruner (mixed) SP70	336.83	412.89	341.68
Lang-Prune-Max SP70	266.78	322.03	261.43

Table 13: English PPL across corpora at 70% sparsity. Wikitext-2 reproduces mC4 magnitudes within $<2\%$, confirming the degradation is a genuine model property.

Wikitext-2 reproduces mC4 seed 0 magnitudes within 2% for both pruned models (LLM-Pruner: $\Delta = +1.4\%$; Lang-Prune-Max: $\Delta = -2.0\%$), confirming that the degradation reflects genuine loss of English language capacity rather than a corpus artifact. Lang-Prune-Max achieves 23.5% lower English PPL than LLM-Pruner on Wikitext-2 (261.43 vs. 341.68).

Arabic and Chinese. Table 14 extends the investigation to Arabic and Chinese, using language-specific Wikipedia as the independent corpus.

Checkpoint	Arabic		Chinese	
	mC4-ara	Arabic Wiki	mC4-zho	Chinese Wiki
Unpruned aya-expanse-8b	20.61	15.53	46.46	35.94
LLM-Pruner (mixed) SP70	175.87	151.60	665.71	794.91
Lang-Prune-Max SP70	129.87	112.81	293.39	390.11

Table 14: Arabic and Chinese PPL across corpora at 70% sparsity. Arabic degradation is consistent across corpora ($<14\%$ deviation). Chinese Wikipedia is harder than mC4 for pruned models, reflecting disproportionate loss of formal written Chinese capacity; Lang-Prune-Max reduces this gap by $\sim 50\%$.

Degradation ratios across languages. Table 15 summarizes the degradation ratio (pruned PPL / unpruned PPL) across all three languages investigated, using mC4 seed 0.

Language	Unpruned PPL	LLM-Pruner SP70	LangP-Max SP70	LLM-P ratio
English	23.74	336.83	266.78	$14.2\times$
Arabic	20.61	175.87	129.87	$8.5\times$
Chinese	46.46	665.71	293.39	$14.3\times$

Table 15: Degradation ratio (pruned / unpruned PPL) at 70% sparsity across three languages. Arabic is most resilient; Chinese and English degrade most severely for distinct structural reasons (see text).

Arabic degrades the least ($8.5\times$ for LLM-Pruner), likely because Arabic morphological structure is highly concentrated in specific network components. Chinese and English degrade most severely ($\sim 14\times$) but for different reasons: English because its representations are distributed broadly across the network (each component removal has multiplicative fluency cost), and Chinese because of the extreme character-coverage demands of a logographic script. Lang-Prune-Max reduces Chinese degradation to $6.3\times$ – matching Arabic’s resilience – by providing explicit per-language Chinese reference data that concentrates and protects Chinese-specific components during pruning.

A.4 Lang-Prune Performance Across Model Scales

The per-language breakdown in Table 16 confirms that Lang-Prune’s interference suppression is consistent across all nine languages, not driven by any single language, and that

<i>Qwen3-8B at 30% sparsity</i>										
<i>PPL ↓</i>	ara	ces	deu	eng	spa	ind	heb	rus	zho	Avg.
LLM-Pruner (monolingual)	13.53	9.05	13.62	31.19	13.93	9.41	24.22	8.50	14.37	15.76
LLM-Pruner (mixed-data)	15.76	9.82	13.05	24.31	14.87	9.99	31.06	8.93	16.18	16.00
<i>IF vs monolingual</i>	<i>1.16×</i>	<i>1.09×</i>	<i>0.96×</i>	<i>0.78×</i>	<i>1.07×</i>	<i>1.06×</i>	<i>1.28×</i>	<i>1.05×</i>	<i>1.13×</i>	<i>1.06×</i>
Lang-Prune	14.19	9.24	12.00	21.24	13.39	9.20	26.23	8.46	13.13	14.12
<i>IF vs monolingual</i>	<i>1.05×</i>	<i>1.02×</i>	<i>0.88×</i>	<i>0.68×</i>	<i>0.96×</i>	<i>0.98×</i>	<i>1.08×</i>	<i>0.99×</i>	<i>0.91×</i>	<i>0.90×</i>
<i>Qwen3-8B at 50% sparsity</i>										
LLM-Pruner (monolingual)	20.28	14.77	35.92	118.76	29.96	20.69	37.05	17.93	35.11	36.94
LLM-Pruner (mixed-data)	208.99	102.79	207.90	565.88	205.65	144.37	723.06	61.84	575.77	310.69
<i>IF vs monolingual</i>	<i>10.31×</i>	<i>6.96×</i>	<i>5.79×</i>	<i>4.77×</i>	<i>6.86×</i>	<i>6.98×</i>	<i>19.52×</i>	<i>3.45×</i>	<i>16.40×</i>	<i>9.00×</i>
Lang-Prune	22.87	14.57	19.62	44.50	22.02	14.65	44.87	13.25	22.35	24.30
<i>IF vs monolingual</i>	<i>1.13×</i>	<i>0.99×</i>	<i>0.55×</i>	<i>0.38×</i>	<i>0.73×</i>	<i>0.71×</i>	<i>1.21×</i>	<i>0.74×</i>	<i>0.64×</i>	<i>0.66×</i>
<i>Qwen3-8B at 70% sparsity</i>										
LLM-Pruner (monolingual)	73.10	134.00	752.05	948.83	391.40	344.82	117.63	179.41	559.01	388.69
LLM-Pruner (mixed-data)	99441.44	35822.38	46402.90	15611.48	35874.89	28035.05	180241.13	62441.93	28241.14	53567.15
<i>IF vs monolingual</i>	<i>1360.35×</i>	<i>267.33×</i>	<i>61.70×</i>	<i>16.45×</i>	<i>91.66×</i>	<i>81.30×</i>	<i>1532.27×</i>	<i>348.04×</i>	<i>50.52×</i>	<i>137.85×</i>
Lang-Prune	126.94	57.09	122.29	400.39	151.12	86.70	339.80	56.10	230.26	174.74
<i>IF vs monolingual</i>	<i>1.74×</i>	<i>0.43×</i>	<i>0.16×</i>	<i>0.42×</i>	<i>0.39×</i>	<i>0.25×</i>	<i>2.89×</i>	<i>0.31×</i>	<i>0.41×</i>	<i>0.45×</i>

Table 16: Full per-language PPL and IF on Qwen3-8B across sparsities, summarized in Table 4 of the main text. Lang-Prune suppresses interference across all languages and sparsities; LLM-Pruner (mixed-data) degrades catastrophically at 70% sparsity.

LLM-Pruner (mixed-data) degradation at high sparsity is widespread rather than language-specific. We further evaluate Lang-Prune across the full Qwen3 size range (0.6B–14B) under 50% sparsity; Table 5 reports the main IF-based scaling summary, while Table 17 provides the corresponding average PPL values including the unpruned and monolingual references.

Model Size	0.6B	1.7B	4B	8B	14B
Base Model	23.95	16.52	14.28	11.02	9.50
Mixed-data	52.38	80.96	40.72	310.69	555.36
Monolingual	47.59	36.47	32.72	36.72	69.10
Lang-Prune	56.07	45.68	31.43	24.30	31.60

Table 17: Average PPL across nine languages for Qwen3 model sizes under 50% sparsity. This table provides the PPL values underlying the IF-based scaling summary in Table 5. Lang-Prune consistently performs best for mid-to-large models ($\geq 4B$); its advantage amplifies with scale.

Lang-Prune achieves the lowest PPL at 4B, 8B, and 14B scales, with the mixed-data baseline degrading catastrophically at larger sizes (555.36 avg at 14B). At 0.6B and 1.7B, Lang-Prune underperforms, indicating a transitional regime where models lack sufficient structured redundancy for language-aware aggregation to exploit.

Qwen3-14B detailed results. Table 18 reports the full per-language breakdown at 14B, including an English-only monolingual reference-data baseline that tests the lower bound of reference-data quality.

Method	ara	heb	ces	rus	deu	eng	spa	ind	zho	Avg.
Lang-Prune-Max	22.77	57.18	13.49	13.43	22.45	75.54	21.94	15.52	42.08	31.60
LLM-Pruner (mixed)	153.27	1084.39	282.74	87.33	307.06	1733.69	452.70	173.60	723.42	555.36
LLM-Pruner (mono-eng)	>10k	>10k	2,888	>10k	978	309	1,050	535	>10k	$\gg 1000$

Table 18: Qwen3-14B per-language PPL at 50% sparsity. Lang-Prune-Max achieves $17.6\times$ lower average PPL than LLM-Pruner (mixed). English-only reference data causes catastrophic multilingual failure, confirming that multilingual reference data is not optional at this scale.

Lang-Prune-Max achieves $17.6\times$ lower average PPL than LLM-Pruner (mixed) at 14B (31.60 vs. 555.36), with the largest per-language gaps for Hebrew ($19.0\times$: 57.18 vs. 1084.39), English ($22.9\times$: 75.54 vs. 1733.69), and Chinese ($17.2\times$: 42.08 vs. 723.42). English-only reference data

(mono-eng) destroys all non-Western languages ($>10k$ PPL for Arabic, Hebrew, Chinese, and Russian), confirming that multilingual reference data is not optional at larger scales. The 14B advantage over 8B ($17.6\times$ vs. $12.8\times$ for mixed-data) indicates that larger models develop richer cross-language dependencies that amplify the benefit of per-language importance scoring.

These findings also contextualize the relationship between pruning and training smaller dense models. In a zero-shot setting, a dense model trained from scratch at a given parameter scale typically achieves lower perplexity than a larger model pruned to the same effective size. This difference largely stems from data regimes: dense models are trained with full-scale curated corpora, whereas one-shot pruning focuses purely on structural compression using minimal reference data. Consequently, **pruned models should be viewed as strong initializations** that preserve the parent model’s tokenizer, alignment, and behavioral characteristics, **rather than direct substitutes for fully trained dense models**. When further post-training is applied (Section 4.2), pruned variants recover substantial capability, supporting their practical role in adapting a single high-quality parent model to multiple deployment constraints.

A.5 Downstream Task Evaluation

We evaluate pruned Qwen3-8B models on multilingual downstream benchmarks to complement the PPL results in the main paper, reporting few-shot performance before and after LoRA post-training across all sparsity levels.

A.5.1 Few-shot Results on Qwen3-8B

Setup. (1) Parameter-efficient tuning: LoRA with rank $r = 16$ and scaling factor $\alpha = 32$, applied to all MLP and attention modules while freezing base model weights. **(2) Training protocol:** Models are trained on Wikipedia (Foundation) with a sequence length of 256. An effective batch size of 1M tokens per step is achieved via gradient accumulation. Training runs for 1,000 steps (1B tokens total) using the HuggingFace Trainer with identical optimization settings. All evaluations are performed under 3-shot settings.

Benchmark	Languages
Belebele	Arabic, Hebrew, Czech, Russian, German, Spanish, Indonesian, Chinese
Global-MMLU	Arabic, Czech, Russian, German, Spanish, Indonesian, Chinese

Table 19: Languages for few-shot benchmarks.

Table 20 summarizes average accuracy across languages; Tables 21 and 22 provide per-language breakdowns.

Method	Sparsity 0.3		Sparsity 0.5		Sparsity 0.7	
	Belebele	G-MMLU	Belebele	G-MMLU	Belebele	G-MMLU
LLM-Pruner (Mixed-data)	0.4144	0.3536	0.2736	0.2717	0.2233	0.2665
LLM-Pruner + Training	0.5064	0.4173	0.3049	0.2852	0.2629	0.2716
Lang-Prune (Ours)	0.5436	0.4406	0.3044	0.2868	0.2658	0.2717
Lang-Prune + Training (Ours)	0.5958	0.4765	0.3557	0.2994	0.2624	0.2665
Qwen3-8B (Unpruned)	0.8796	0.6777	–	–	–	–

Table 20: Few-shot accuracy on multilingual benchmarks across sparsity ratios (Qwen3-8B). “+Training” denotes LoRA post-training recovery. Best pruned results are **bolded**.

The per-language results are consistent with the aggregate trends in Table 20. On Belebele, Lang-Prune outperforms LLM-Pruner (mixed) at every sparsity level both before and after post-training, with the largest gains at 30% sparsity (e.g., zho: 0.604 vs. 0.410 untrained; 0.664 vs. 0.593 after training). Gains are broad across script-diverse and morphologically rich languages alike. On Global-MMLU, Lang-Prune similarly leads at all sparsities, though

Method	Sparsity	ara	heb	ces	rus	deu	spa	ind	zho	Avg
Unpruned baseline	0%	0.863	0.831	0.882	0.892	0.913	0.887	0.876	0.892	0.880
LLM-Pruner (mixed)	30%	0.414	0.362	0.387	0.443	0.428	0.432	0.439	0.410	0.414
LLM-Pruner + Training	30%	0.502	0.427	0.463	0.526	0.481	0.529	0.530	0.593	0.506
Lang-Prune (Ours)	30%	0.548	0.468	0.542	0.558	0.543	0.541	0.544	0.604	0.544
Lang-Prune + Training (Ours)	30%	0.597	0.534	0.587	0.583	0.616	0.601	0.584	0.664	0.596
LLM-Pruner (mixed)	50%	0.286	0.249	0.281	0.279	0.250	0.286	0.279	0.280	0.274
LLM-Pruner + Training	50%	0.287	0.301	0.303	0.306	0.336	0.307	0.303	0.297	0.305
Lang-Prune (Ours)	50%	0.302	0.290	0.314	0.324	0.290	0.314	0.286	0.314	0.304
Lang-Prune + Training (Ours)	50%	0.348	0.310	0.322	0.326	0.361	0.376	0.366	0.438	0.356
LLM-Pruner (mixed)	70%	0.219	0.219	0.219	0.219	0.219	0.219	0.219	0.254	0.223
LLM-Pruner + Training	70%	0.260	0.292	0.284	0.261	0.218	0.277	0.261	0.250	0.263
Lang-Prune (Ours)	70%	0.278	0.276	0.242	0.274	0.218	0.280	0.280	0.279	0.266
Lang-Prune + Training (Ours)	70%	0.253	0.287	0.268	0.243	0.241	0.274	0.263	0.269	0.262

Table 21: Per-language Belebele 3-shot accuracy on Qwen3-8B before and after LoRA post-training. Random chance = 0.25. Lang-Prune consistently outperforms LLM-Pruner (mixed) at all sparsities, with the largest gains at 30% sparsity across all languages.

Method	Sparsity	ara	ces	rus	deu	spa	ind	zho	Avg
Unpruned baseline	0%	0.626	0.659	0.683	0.693	0.702	0.682	0.700	0.678
LLM-Pruner (mixed)	30%	0.336	0.343	0.349	0.367	0.373	0.355	0.354	0.354
LLM-Pruner + Training	30%	0.394	0.393	0.416	0.428	0.432	0.428	0.432	0.417
Lang-Prune (Ours)	30%	0.406	0.422	0.439	0.455	0.453	0.446	0.462	0.441
Lang-Prune + Training (Ours)	30%	0.443	0.458	0.474	0.494	0.490	0.479	0.497	0.477
LLM-Pruner (mixed)	50%	0.270	0.276	0.276	0.272	0.267	0.275	0.266	0.272
LLM-Pruner + Training	50%	0.287	0.281	0.287	0.286	0.284	0.291	0.280	0.285
Lang-Prune (Ours)	50%	0.289	0.280	0.285	0.290	0.286	0.282	0.295	0.287
Lang-Prune + Training (Ours)	50%	0.298	0.294	0.297	0.305	0.294	0.303	0.306	0.299
LLM-Pruner (mixed)	70%	0.263	0.267	0.257	0.264	0.273	0.275	0.266	0.267
LLM-Pruner + Training	70%	0.271	0.276	0.270	0.271	0.271	0.267	0.277	0.272
Lang-Prune (Ours)	70%	0.274	0.270	0.269	0.281	0.271	0.266	0.273	0.272
Lang-Prune + Training (Ours)	70%	0.270	0.264	0.268	0.262	0.268	0.261	0.273	0.267

Table 22: Per-language Global-MMLU 3-shot accuracy on Qwen3-8B before and after LoRA post-training. Random chance = 0.25. Lang-Prune consistently outperforms the mixed-data baseline; at 70% sparsity both methods converge near random chance, consistent with the severe PPL degradation reported in Table 16.

margins narrow at higher sparsity as both methods approach random chance. Post-training with LoRA yields consistent recovery at 30% and 50% sparsity, with Lang-Prune achieving larger absolute gains on Belebele and comparable gains on Global-MMLU.

A.5.2 Full OOD Generalization Results

Table 23 reports zero-shot OOD generalization for all 16 languages at 70% sparsity on aya-expense-8b, extending the representative subset in Table 6 of the main text. We include the unpruned/base PPL for each OOD language and report the top three monolingual proxies among the nine reference languages; values in parentheses are PPL gaps relative to the best proxy.

A.6 Computational Analysis

A.6.1 Efficiency Benchmark

We report concrete inference efficiency metrics for structured pruning at multiple sparsity levels. All measurements use sequence length 128, with 5 warmup and 50 measurement forward passes. Attention is evaluated in eager mode (not Flash Attention) to reflect true reduced dimensions.

Language	Family Script	Base PPL	Best mono (source)	PPL (Mixed)	PPL (Mono)	PPL (Lang-Prune)	GR ↓	2nd mono proxy (gap)	3rd mono proxy (gap)
fas (Persian)	Indo-Iranian / Arabic	27.73	ara	154.37	92.90	62.26	0.40	heb (+30.59)	rus (+147.47)
urd (Urdu)	Indo-Iranian / Arabic	32.30	heb	273.27	107.91	31.08	0.11	zho (+38.06)	ara (+165.33)
amh (Amharic)	Semitic / Ethiopic	12.93	zho	258.83	12.25	11.90	0.05	heb (+180.02)	rus (+326.15)
bul (Bulgarian)	Slavic / Cyrillic	65.31	rus	128.64	91.21	71.05	0.55	heb (+59.20)	ces (+196.05)
ukr (Ukrainian)	Slavic / Cyrillic	18.63	rus	152.68	67.09	57.91	0.38	ces (+66.66)	heb (+188.99)
pol (Polish)	Slavic / Latin	24.43	ces	199.85	104.12	68.24	0.34	rus (+15.19)	heb (+211.76)
nld (Dutch)	Germanic / Latin	25.98	deu	410.68	178.78	104.86	0.26	ces (+96.97)	ind (+162.65)
swe (Swedish)	Germanic / Latin	61.69	deu	404.02	291.75	144.52	0.36	ces (+42.42)	ind (+44.30)
dan (Danish)	Germanic / Latin	55.18	deu	440.81	280.81	160.90	0.37	ces (+33.46)	ind (+41.81)
fra (French)	Romance / Latin	16.47	spa	318.40	143.33	78.00	0.25	ces (+73.80)	deu (+106.29)
ita (Italian)	Romance / Latin	14.80	spa	268.33	148.46	78.04	0.29	ces (+77.14)	ind (+96.37)
por (Portuguese)	Romance / Latin	16.94	spa	333.11	127.94	81.01	0.24	ces (+125.91)	rus (+165.99)
msa (Malay)	Austronesian / Latin	11.27	zho	14.46	8.06	7.64	0.53	rus (+29.43)	ind (+56.26)
jpn (Japanese)	Japonic / Japanese	19.75	zho	608.95	158.51	155.66	0.26	heb (+366.31)	rus (+462.68)
kor (Korean)	Koreanic / Hangul	19.23	zho	346.27	89.62	76.57	0.22	heb (+170.30)	rus (+249.67)
vie (Vietnamese)	Austroasiatic / Latin	21.54	zho	303.86	149.50	76.76	0.25	heb (+85.92)	ind (+260.53)

Table 23: Full zero-shot OOD generalization at 70% sparsity on aya-expense-8b, evaluated on the mC4 validation split. Base PPL is the unpruned aya-expense-8b score on the same language. GR (generalization ratio) = $\text{PPL}_{\text{Lang-Prune}} / \text{PPL}_{\text{mixed}}$; values below 1 indicate Lang-Prune outperforms mixed-data pruning. Lang-Prune outperforms the mixed-data baseline across all 16 languages and achieves $\sim 32\%$ lower PPL than the best monolingual proxy on average.

Model	Method / Sparsity	Params (B)	Reduction	Tok/s (bs=1)	Tok/s (bs=8)	Peak Mem (GB)
aya-expense-8b	Unpruned (0%)	8.028	—	3,954	10,138	15.10
	LLM-Pruner (30%)	6.339	−21.1%	4,396	7,418	11.98
	Lang-Prune-Max (30%)	6.343	−21.0%	4,319	7,537	11.98
	LLM-Pruner (50%)	5.212	−35.1%	4,991	10,148	9.87
	Lang-Prune-Max (50%)	5.209	−35.1%	4,755	9,156	9.88
	LLM-Pruner (70%)	4.080	−49.2%	4,893	12,073	7.76
	Lang-Prune-Max (70%)	4.080	−49.2%	4,839	12,082	7.77
Qwen3-8B	Unpruned (0%)	8.191	—	3,560	10,925	15.32
	LLM-Pruner (30%)	6.556	−20.0%	3,465	8,241	12.31
	Lang-Prune-Max (30%)	6.560	−19.9%	3,333	7,991	12.32
	LLM-Pruner (50%)	6.060	−26.0%	3,460	10,150	11.35
	Lang-Prune-Max (50%)	6.074	−25.9%	3,447	10,031	11.40
	LLM-Pruner (70%)	4.386	−46.5%	3,335	12,422	8.26
	Lang-Prune-Max (70%)	4.384	−46.5%	3,413	12,548	8.25

Table 24: Inference efficiency at varying sparsity levels for both primary models. At 70% sparsity, structured pruning reduces parameters by $\sim 49\%$ and peak memory from 15.1 GB to 7.77 GB for aya-expense-8b. Lang-Prune-Max and LLM-Pruner are efficiency-equivalent at equal sparsity – efficiency gains derive from pruning itself, not from the importance scoring strategy.

At 70% sparsity, structured pruning reduces parameters by $\sim 49\%$ and peak memory by $\sim 49\%$ (15.10 $\rightarrow$ 7.77 GB for aya-expense-8b; 15.32 $\rightarrow$ 8.25 GB for Qwen3-8B). Batch-8 throughput increases by $\sim 19\%$ for aya-expense-8b. Crucially, Lang-Prune-Max and LLM-Pruner are efficiency-equivalent at equal sparsity: the efficiency gains come from structured pruning itself, not from the choice of importance scoring strategy. This confirms that Lang-Prune’s multilingual improvements carry zero inference-time overhead relative to the standard baseline.

Qwen3-8B shows a non-monotone throughput pattern at low sparsity due to its Grouped Query Attention (GQA) architecture: at 30% sparsity, MLP channels are reduced but the KV cache bandwidth bottleneck for attention at large batches remains, creating an imbalanced speedup profile. This architecture-specific effect resolves at 70% sparsity, where batch-8 throughput exceeds the unpruned baseline by +14.9%.

A.6.2 Reference-Language Scaling

A natural question is whether Lang-Prune’s Max aggregator remains stable as the number of reference languages grows. We evaluate on Qwen3-8B at 50% sparsity with 9, 15, and 25 reference languages (adding OOD languages beyond the paper’s default 9).

Panel A: Core-9 Languages

N refs	ara	heb	ces	rus	deu	eng	spa	ind	zho	Avg	Worst
Unpruned	11.00	19.66	7.64	6.96	9.78	15.42	10.75	7.45	10.51	11.02	19.66
9 refs	18.04	34.62	11.55	10.50	15.18	29.74	17.03	11.53	16.89	18.34	34.62
15 refs	17.97	34.86	11.54	10.34	15.00	29.40	16.76	11.40	17.04	18.26	34.86
25 refs	18.17	35.69	11.78	10.72	15.44	31.27	17.11	11.63	17.23	18.78	35.69

Panel B: Added-6 Languages

N refs	fas	ukr	pol	nld	fra	kor	Avg	Worst
Unpruned	5.46	4.60	9.72	9.66	8.69	9.78	7.98	9.78
9 refs	8.03	6.63	15.16	15.52	13.41	17.14	12.65	17.14
15 refs	7.81	6.56	15.02	15.16	12.97	15.60	12.18	15.60
25 refs	7.84	6.66	15.57	15.54	13.46	15.91	12.50	15.91

Panel C: Extra-10 Languages

N refs	swe	dan	ita	por	msa	jpn	vie	bul	urd	amh	Avg	Worst
Unpruned	11.96	10.66	8.40	9.05	2.25	11.30	9.28	6.29	4.64	6.70	8.05	11.96
9 refs	19.82	17.23	13.65	14.68	9.61	23.00	15.33	9.02	6.88	10.24	13.95	23.00
15 refs	19.43	16.75	13.18	14.28	7.60	21.06	14.98	8.94	6.55	9.56	13.23	21.06
25 refs	19.45	17.43	13.48	14.93	2.82	21.09	15.02	8.99	6.51	8.65	12.84	21.09

Panel D: All-25 Global Summary

N refs	All-25 Avg	All-25 Worst
Unpruned	9.10	19.66
9 refs	15.22	34.62
15 refs	14.79	34.86
25 refs	14.90	35.69

Table 25: Lang-Prune-Max PPL on all 25 evaluated languages as the number of reference languages grows (Qwen3-8B, 50% sparsity). The unpruned row reports base-model PPL for the same language groups, making the pruned results directly interpretable against the uncompressed model. The 15-language setting gives the best all-language average; using the full 25-language reference set mainly benefits the extra low-resource/OOD group while mildly degrading the original core-9 languages.

The Max aggregator is stable across the tested range. Moving from 9 to 15 reference languages improves the all-language average (15.22 $\rightarrow$ 14.79) while leaving the core-9 average nearly unchanged (18.34 $\rightarrow$ 18.26). At 25 reference languages, the all-language average remains close (14.90) and the extra-10 group improves further (13.23 $\rightarrow$ 12.84), but the core-9 average degrades mildly (18.26 $\rightarrow$ 18.78). Reference-data processing time scales approximately linearly: $\sim$ 2 min 53 sec (9 langs), $\sim$ 4 min 15 sec (15 langs), $\sim$ 5–6 min (25 langs), all negligible for a one-time pruning operation.

The mild core-language degradation at 25 languages is consistent with the Max aggregator’s conservatism: with more languages, more components are classified as “important for at least one language,” leaving less room for the pruner to remove genuinely unimportant structure. This suggests that for very large language sets, Mean, Sum, or grouped-Max aggregators may be preferable. Within the practical range tested here (9–15 languages), the paper’s reference set is not artificially narrow.

A.7 Multilingual vs. Multi-domain Pruning

While the high-level idea behind language-aware pruning resembles task- or domain-aware pruning, the underlying structure differs fundamentally. Human languages are highly distinct in vocabulary, syntax, morphology, and script, whereas closely related tasks often share significant latent structure, vocabulary, and reasoning patterns. Consequently, multilingual models naturally organize **discrete, language-specific neuron groups**, as observed in prior work (Tan et al., 2024; Wang et al., 2025). These neuron clusters are highly separable, providing a clear signal for structure-preserving pruning strategies such as Lang-Prune. In contrast, multi-task or multi-domain settings typically involve subtle differences. Tasks or domains that share vocabulary and reasoning patterns produce overlapping activations, making per-task neuron importance less stable.

To empirically validate this, we applied a similar *per-domain importance + max aggregation* strategy to a multi-domain scenario using Qwen3-8B, treating three MMLU subfields (Law, Medical, Finance) as domains. Reference data uses the respective train/dev splits; evaluation is on the test split at 50% sparsity.

Avg Acc $\uparrow$ (stderr)	Law	Medical	Finance
Base Model	73.1% $\pm$ 13.8%	76.4% $\pm$ 8.3%	78.9% $\pm$ 12.7%
Mixed-data	37.6% $\pm$ 6.6%	35.3% $\pm$ 6.7%	37.5% $\pm$ 5.3%
Lang-Prune (ours)	40.1% $\pm$ 9.4%	37.7% $\pm$ 7.7%	37.8% $\pm$ 7.1%

Table 26: Domain-aware pruning results on Qwen3-8B (50% sparsity). Unlike multilingual pruning, domain-aware pruning provides only marginal improvements.

Domain-aware pruning achieves only marginal improvements over mixed-data pruning (<3 percentage points) and remains far below the base model. The differences between domains are not sufficiently large to yield stable per-domain importance rankings, leading to limited benefits from Max aggregation. These observations highlight that **language-aware pruning is fundamentally different from task-aware pruning**: the strong, discrete separation of neuron groups in multilingual LLMs enables reliable structured pruning, which cannot be trivially generalized to multi-domain or multi-task scenarios.

A.8 Influence of Reference-Data Size and Importance Dilution

Method	Sparsity	N=10	N=30	N=50	N=80	N=100	N=150	N=200
Mixed-data (9N total)	0.3	14.55	14.29	15.02	14.91	16.00	14.58	15.19
Monolingual (N each)	0.3	18.39	16.95	15.07	14.96	15.76	14.70	15.23
Lang-Prune (ours, 9N total)	0.3	13.57	13.53	13.39	13.35	14.12	13.44	13.44
Mixed-data (9N total)	0.5	34.83	35.70	44.72	76.68	310.69	331.62	384.20
Monolingual (N each)	0.5	77.40	54.17	36.70	34.12	36.94	32.70	36.73
Lang-Prune (ours, 9N total)	0.5	23.22	21.54	21.00	20.60	24.30	21.36	21.33
Mixed-data (9N total)	0.7	736.09	707.91	2406.78	19716.96	53567.15	16444.23	91970.55
Monolingual (N each)	0.7	1668.59	1009.15	432.54	348.63	388.69	254.56	247.01
Lang-Prune (ours, 9N total)	0.7	200.12	146.17	103.90	90.31	174.74	85.28	80.13

Table 27: Average PPL vs. reference-data size for Qwen3-8B. “N each” denotes sequences per language. Lang-Prune uses the same total budget as mixed-data (9N total). Lang-Prune consistently outperforms both baselines across all sizes and sparsity levels.

To evaluate the effect of reference-data size, we compare monolingual pruning at varying per-language budgets against Lang-Prune using the same total reference-data size (9N sequences across all languages). Monolingual pruning benefits from larger reference sets but still consistently underperforms Lang-Prune. At moderate sparsity (30–50%), increasing monolingual reference data reduces PPL, yet performance plateaus and remains notably

worse than Lang-Prune. At 70% sparsity, monolingual pruning becomes highly unstable (PPL $3\text{--}5\times$ higher than Lang-Prune), while mixed-data pruning shows catastrophic degradation beyond $N=50$. These results confirm that Lang-Prune’s improvement is due to its design principle of protecting language-specific structures, not merely to a larger total reference-data budget.

A.8.1 Importance Distribution under Varying Reference-Data Size

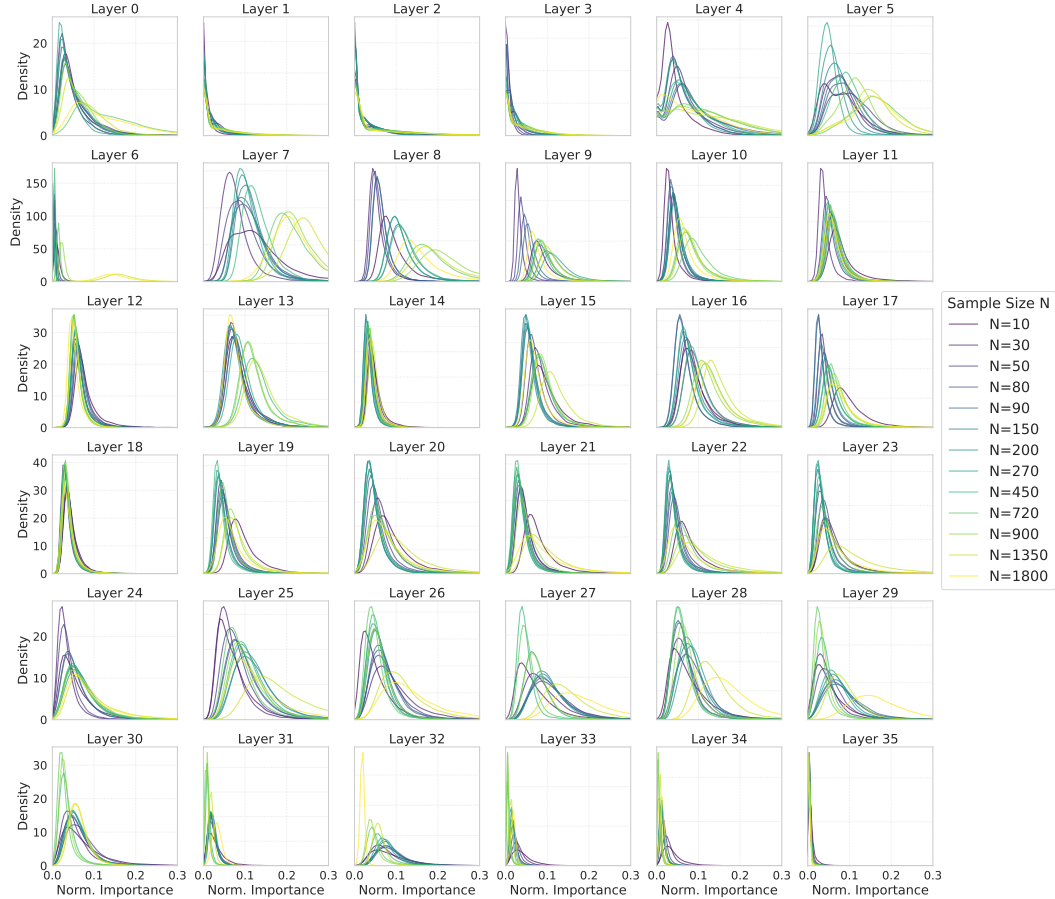


Figure 3: Impact of reference-data size (N) on normalized importance score distribution across 36 layers of Qwen3-8B (zoomed to 0–0.3). Larger N increases density in the mid-importance region, reducing score separability – a phenomenon we term **importance dilution**.

A counter-intuitive result in Table 27 is that increasing reference-data size eventually *harms* pruning quality, especially at high sparsity. Figure 3 reveals the mechanism: as N increases, importance scores become more uniformly distributed, with a clear density increase in the mid-importance region (0.1–0.3). This reduces the contrast between critical and non-critical structures, making ranking-based pruning less discriminative – a phenomenon we term **importance dilution**.

Under high sparsity, this dilution makes pruning unstable: if many components appear moderately important, aggressive pruning may remove genuinely critical structures while retaining mediocre ones, resulting in sudden performance collapse. Mixed-data pruning is even more vulnerable due to higher topic and domain diversity, which further spreads residual importance. In contrast, Lang-Prune remains robust because its Max aggregation selectively preserves peak language-specific signals, maintaining importance separability even as reference-data size increases.

We also checked whether the large values in Table 27 are caused by a single unlucky sentence sample. A three-seed study on Qwen3-8B shows that at 50% sparsity, LLM-Pruner mixed-data reaches average PPL 44.0 with 2.4% relative standard deviation, while Lang-Prune reaches 22.4 with 1.5% relative standard deviation. At 70% sparsity, LLM-Pruner mixed-data collapses to average PPL 2647 with 22.5% relative standard deviation, whereas Lang-Prune remains much lower at 104.5 with 11.7% relative standard deviation. This indicates that mixed-data collapse is a stable high-sparsity failure mode rather than a one-run artifact.

A.9 MBS-Style Reference-Data Mixture Comparison

To evaluate the impact of multilingual reference-data strategies, we approximate the Multilingual Brain Surgeon (MBS) approach of Zeng et al. (2024) by sampling reference data with various language mixture ratios; exact MBS recipes are not available for the Qwen and Aya model families. We compare these against uniform mixtures and Lang-Prune’s language-aware Max aggregation on Qwen3-8B at 50% sparsity.

Lang	EN:O 3:1	EN:O 2:1	EN:O 5:1	E:Z:E:O 3:2:2:1	E:Z:E:O 4:2:2:1	Unif.	Lang-Prune
ara	506.51	<u>110.02</u>	298.05	313.12	331.37	208.99	22.87
ces	211.97	<u>80.93</u>	386.84	182.11	159.93	102.79	14.57
deu	515.74	<u>157.60</u>	644.36	405.21	291.29	207.90	19.62
eng	1055.65	<u>364.47</u>	829.81	976.80	753.70	565.88	44.50
spa	379.91	<u>143.36</u>	411.79	326.87	289.16	205.65	22.02
ind	366.43	<u>100.47</u>	229.31	270.06	230.32	144.37	14.65
heb	1226.97	<u>539.44</u>	3167.02	1791.35	1333.17	723.06	44.87
rus	179.46	<u>51.07</u>	168.75	152.53	106.46	61.84	13.25
zho	722.71	617.56	1318.28	629.44	<u>528.75</u>	575.77	22.35
avg	573.93	<u>240.55</u>	828.24	560.83	447.13	310.69	24.30

Table 28: Per-language PPL↓ for different reference-data mixture strategies on Qwen3-8B (50% sparsity). Column headers denote EN:Others ratio (e.g., *EN:O 3:1* means English to all other languages at 3:1) or multi-language ratios (e.g., *E:Z:E:O 3:2:2:1* means EN:ZH:ES:Others at 3:2:2:1). **Bold**: best overall; underline: best among MBS-style mixtures. Lang-Prune’s per-language scoring with Max aggregation yields $\sim 10\times$ lower average PPL than the best MBS-style mixture (24.30 vs. 240.55).

MBS-style rebalancing helps versus uniform mixtures, but Lang-Prune’s per-language scoring with Max aggregation yields substantially better perplexities across all languages (avg 24.30 vs. best MBS 240.55, $\sim 10\times$ improvement). This demonstrates that varying reference-data mixture ratios – without changing how importance is computed – is fundamentally insufficient. The key ingredient is maintaining separate per-language importance estimates and aggregating them with a fairness-aware policy.