



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/239669/>

Version: Published Version

Article:

Kurz, S., Chen, J.-J., Flek, L. et al. (2026) On the limitations of language-targeted pruning: Investigating the calibration language impact in multilingual LLM pruning. Transactions of the Association for Computational Linguistics, 14. pp. 167-192. ISSN: 2307-387X

<https://doi.org/10.1162/tacl.a.599>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

On the Limitations of Language-targeted Pruning: Investigating the Calibration Language Impact in Multilingual LLM Pruning

Simon Kurz^{1,4} Jian-Jia Chen^{1,4} Lucie Flek^{2,4} Zhixue Zhao^{3*}

¹Department of Computer Science, TU Dortmund University, Germany

²Bonn-Aachen International Center for Information Technology, University of Bonn, Germany

³Computer Science School, University of Sheffield, United Kingdom

⁴Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

zhixue.zhao@sheffield.ac.uk

Abstract

Recent advances in large language model (LLM) pruning have shown state-of-the-art (SotA) compression results in post-training and retraining-free settings while maintaining high predictive performance. However, previous research mainly considered calibrating based on English text, despite the multilingual nature of modern LLMs and their frequent use in non-English languages. This analysis paper conducts an in-depth investigation of the performance and internal representation changes associated with pruning multilingual language models for monolingual applications. We present the first comprehensive empirical study, comparing different calibration languages for pruning multilingual models across diverse languages, tasks, models, and SotA pruning techniques. We further analyze the latent subspaces, pruning masks, and individual neurons within pruned models. Our results reveal that while calibration on the target language effectively retains perplexity and yields high signal-to-noise ratios, it does not consistently improve downstream task performance. Further analysis of internal representations at three different levels highlights broader limitations of current pruning approaches: While they effectively preserve dominant information like language-specific features, this is insufficient to counteract the loss of nuanced, language-agnostic features that are crucial for knowledge retention and reasoning.

1 Introduction

State-of-the-art (SotA) language models often rely on over-parameterization with billions of parameters, resulting in significant memory and computational demands (Zhang et al., 2017;

Allen-Zhu et al., 2019; Xu and Du, 2023). To address this, model compression methods such as quantization and pruning are commonly used (Gholami et al., 2022; Hoefler et al., 2021; Kuzmin et al., 2023; Frantar et al., 2023; Zhu et al., 2024; Liu et al., 2024). Established post-training pruning methods for language models, such as Sparse-GPT (Frantar and Alistarh, 2023) and Wanda (Sun et al., 2024), demonstrated competitive performance in a retraining-free setting (Zhu et al., 2024), using only a small calibration dataset instead of big training corpora (Zhu et al., 2024; Kuzmin et al., 2022; Frantar and Alistarh, 2023; Kuzmin et al., 2023).

However, current research on large language model (LLM) pruning mostly focuses on English calibration and evaluation, despite the widespread use of multilingual LLMs for non-English tasks (Touvron et al., 2023; OpenAI et al., 2024; Jiang et al., 2023). For example, if we want to prune a multilingual language model (LM) for tasks in French, should we use calibration data in French or English? This disconnect between research focus and real-world usage raises critical questions: **How does the choice of calibration language impact pruning outcomes for multilingual LLMs for monolingual applications? To what extent and why?**

We provide the first comprehensive empirical study of how the calibration language affects multilingual LLM pruning. We find that **calibrating on text in the target language minimizes pruning errors and best retains language modeling capabilities related to perplexity** (Section 5.1). **However, this does not always lead to better downstream task performance** (Section 5.2, Section 5.3). Further, calibration in a different language, despite its distinct linguistic features, can yield better results, suggesting that current

* Corresponding author.

pruning approaches give more importance to dominant language-specific features while neglecting subtle, language-agnostic features crucial for reasoning tasks.

Through multi-level analysis spanning latent subspaces, weight matrices, and individual neurons, we investigate the mechanisms behind our observations and derive the following findings.

- Calibrating on target languages consistently yields the lowest perplexity (Section 5.1) but will not guarantee optimal downstream performance (Section 5.2).
- The calibration language is unlikely to impact language-agnostic features associated with understanding, reasoning, and knowledge retrieval. In contrast, the preservation of language-specific features, such as language modeling capabilities related to perplexity, depends on the selection of calibration languages (Section 6.1).
- Pruning struggles to consistently identify essential weights in the attention output projections, partly responsible for the model’s language-agnostic reasoning capabilities (Section 6.2).
- The neuron activation frequency of language-specific neurons in the feed-forward network (FFN) layers is not reliably preserved after pruning (Section 6.3).

2 Background

2.1 Model Pruning

Pruning (LeCun et al., 1989) is a widely used compression approach, removing parameters that contribute least to model performance to reduce the computational and memory footprint. Unlike sparse training (Yuan et al., 2021; Hoang et al., 2023; Zhang et al., 2023) or pruning-aware training (Liu et al., 2021), which require iterative training to achieve sparsity, post-training pruning eliminates redundant weights directly based on their importance scores computed using a small calibration set, without requiring additional training. This retraining-free feature makes post-training pruning a more efficient approach for LLMs (Frantar and Alistarh, 2023; Sun et al., 2024).

Based on pruning granularity, pruning methods fall into two major categories (Zhu et al., 2024; Zafirir et al., 2021; Guo et al., 2025). **Structured pruning** removes entire model components such as neurons, attention heads, or even layers, enabling hardware efficiency but limiting flexibility. **Unstructured pruning** removes individual weights without pattern constraints, offering high performance at the cost of deployment complexity. Its interpretability and flexibility make it the basis for many analytical studies and recent methods.

Post-training pruning estimates weight importance by minimizing the local pruning error between original $Y = WX$ and pruned layer outputs $\hat{Y} = (M \odot W)X$ for weights $W \in R^{m \times n}$, calibration activations $X \in R^{n \times b}$ and binary pruning mask M . Because pruning preserves weights and activations that optimize $\operatorname{argmin}_M \|Y - \hat{Y}\|_F^2$ on the calibration set, the choice of calibration language directly affects which weights get retained. For instance, calibration in English biases pruning toward English-specific patterns. This issue becomes critical in multilingual models, where features vary between languages (Shamrai, 2024; Xie et al., 2022; Choenni and Titov, 2025).

There are two major unstructured pruning methods that still serve as strong benchmarks and foundations for extensions. **SparseGPT** (Frantar and Alistarh, 2023) iteratively prunes low-importance weights, estimated using local second-order derivative information via a diagonal Hessian approximation as:

$$S_{i,j}^{\text{SGPT}} = \frac{W_{i,j}^2}{H_{j,j}^{-1}}, \quad \text{with } H = XX^\top. \quad (1)$$

Weights are iteratively pruned by lowest pruning score $S_{i,j}^{\text{SGPT}}$, while using Hessian information H to update the remaining weights, compensating induced errors.

Wanda (Sun et al., 2024) uses a simpler one-shot heuristic based on weight magnitude and activation norms, omitting weight updates and expensive Hessian inversions:

$$S_{i,j}^{\text{Wanda}} = |W_{i,j}| \cdot \|X_j\|_2 \quad (2)$$

Although less accurate than SparseGPT in theory, Wanda is more efficient and easier to implement. Both methods are widely adopted and have inspired numerous extensions, such as

Wanda++ (Yang et al., 2025), and M-Wanda (Choenni and Titov, 2025), which adapt these core ideas to gradient-informed or multilingual settings.

2.2 Surface-level Evaluation Metrics

Pruned LLMs can be evaluated at the output-level and at the level of representational metrics that aim to quantify the internal consistency of the compressed model.

Perplexity (PPL) remains a fundamental metric in language modeling. It measures the next-token-prediction uncertainty of a model over a sequence of tokens, formally defined as:

$$\text{PPL} = \exp \left(-\frac{1}{T} \sum_{t=1}^T \log P(w_t | w_{<t}) \right), \quad (3)$$

where $P(w_t | w_{<t})$ is the probability of the next token predicted by the model. Lower perplexity indicates more confident predictions, typically interpreted as higher fluency or syntactic correctness. Despite its limitations in capturing semantic or reasoning quality (Meister and Cotterell, 2021; Jaiswal et al., 2024), PPL remains a standard first indicator of language modeling performance and is widely used to assess the effects of pruning and quantization (Yang et al., 2025; Frantar and Alistarh, 2023). In multilingual settings, PPL is typically reported per language to reflect pruning performance across language-specific features (Zeng et al., 2024; Shamrai, 2024). However, Poelman and de Lhoneux (2025) suggest that the performance of language modeling and language understanding does not necessarily reflect the downstream task performance.

A more direct measure of pruning effects is the **pruning error**, which quantifies the local reconstruction error at the activation level. It is defined as the L2 distance between layer outputs before and after pruning:

$$E = \|WX - (M \odot W)X\|_2^2. \quad (4)$$

This formulation aligns with the reconstruction error minimization paradigm used by many pruning methods, including SparseGPT and Wanda (Frantar and Alistarh, 2023; Sun et al., 2024). While one pruning error measurement alone only captures local, layer-wise deviations, it does not account for cumulative error propagation effects from one layer to the next. Therefore, some stud-

ies measure global pruning error over a test set by aggregating activation deviations across all layers, weighted by the number of activation feature elements (Shin et al., 2024; Li et al., 2025).

Signal-to-Noise Ratio (SNR) is a normalized version of the pruning error, eliminating the effect of different layer output magnitudes between layer components (Kuzmin et al., 2023), defined as

$$\text{SNR} = 10 \cdot \log \left(\frac{\|WX\|_2^2}{\|WX - (M \odot W)X\|_2^2} \right). \quad (5)$$

High SNR values indicate that the retained weights preserve the dominant signal, while low values suggest that pruning has disrupted key feature components.

While PPL, pruning error, and SNR do not reliably capture higher-order capabilities such as reasoning or knowledge retrieval, they are theoretically grounded and still correlated to model performance (Frantar and Alistarh, 2023; Kuzmin et al., 2023; Shin et al., 2024), making them surface-level indicators for assessing pruning quality.

3 Related Work

3.1 Multilingual Language Models

Most SotA LMs, such as Llama 3 (Meta, 2024), Phi 3 (Abdin et al., 2024), and DeepSeek-R1 (DeepSeek-AI et al., 2025), are trained on multilingual data, enabling them to process and generate text across multiple languages (Huang et al., 2023; Holmström et al., 2023; Xu et al., 2025; Meta, 2024). Although multilingual LMs follow the general training paradigm of monolingual LMs, they often exhibit distinct behaviors and unique characteristics (Xu et al., 2025). For example, Deng et al. (2024) reveal that multilingual LMs are prone to generate unsafe outputs on malicious instructions in non-English languages, i.e., multilingual jailbreak. Wang et al. (2023) report significantly higher rates of unsafe responses to non-English queries. Chen et al. (2024) find that instruction tuning in a multilingual setting performs on par with, or even surpasses, tuning a model individually for each language. Furthermore, previous work on model explanation finds that multilingual models have different process mechanisms from counterparts monolingual models (Jørgensen et al., 2022; 2024).

In other areas of multilingual LLM research, such as cross-lingual transfer learning and in-context learning, LLMs are often found to benefit from exposure to non-English languages. For example, Turc et al. (2021) find that pre-training one “pivot” language could be more effective than English for cross-lingual transfer. Similarly, Shi et al. (2023) and Tu et al. (2025) suggest that greater linguistic diversity in prompting languages improves performance, although Wang et al. (2025) note that multilingual prompting may lead to factual hallucination. However, in the context of pruning, it remains unknown whether calibrating on non-English or non-target languages will benefit the performance of pruned multilingual LLMs.

In short, these studies demonstrate that concepts derived from monolingual settings, particularly those focused on English, often fail to generalize to multilingual contexts involving non-English languages and indicate complex cross-linguistic patterns. Moreover, findings regarding the cross-lingual performance of LLMs are frequently inconsistent across different settings. Overall, it is difficult to generalize findings from English to non-English scenarios in the context of pruning LLMs. However, most existing work focuses on English-centric settings. Therefore, our work addresses this gap by examining the issue through the lens of language-targeted pruning.

3.2 Calibration of Post-training Pruning

Prior research has primarily focused on improving weight importance estimation to optimize pruning performance. Few studies have examined the role of calibration data, focusing on factors like data quantity (Zhang et al., 2024) and source (Williams and Aletras, 2024; Bandari et al., 2024), but these efforts have been confined to English.

In the multilingual setting, approaches such as multilingual brain surgeon (Zeng et al., 2024) aim to retain broad multilingual capabilities by mixing languages in the calibration data according to their pre-training distribution. Kim et al. (2024) leverage the inherent semantic latent space alignment between the pre-training dominant English and non-English languages to improve performance in non-English languages by preserving large-magnitude features that emerge during translation. Similarly, Choenni and Titov (2025) introduce M-Wanda, a multilingual-aware variant of

the Wanda algorithm that adjusts pruning based on per-language activation correlations. However, little is known about how the calibration language influences performance when optimizing for a specific target language rather than preserving general performance across languages.

Much previous work on multilingual pruning still uses PPL as a first indicator for pruning and calibration performance (Zeng et al., 2024; Shamrai, 2024), despite the limitations of such surface-level metrics in capturing nontrivial reasoning or semantic transfer (Meister and Cotterell, 2021; Jaiswal et al., 2024). These studies fall short in methodically investigating the extent of these limitations for the language-targeted pruning case we examine.

Our work sheds light on this issue through a multi-view analysis on the calibration language impact on model intrinsics. Unlike previous studies on multilingual pruning, this work examines language-targeted pruning calibration through a systematic evaluation across multiple LLMs, pruning methods, and downstream tasks.

4 Methodology

We first compare how calibration languages affect model performance after pruning: Using seven distinct calibration languages, we prune a full-sized model to create seven variants. We compare their performance across multiple metrics: pruning error, signal-to-noise ratio, perplexity, and downstream task performance. Implementation details are given in Appendix A.

Models Our experiments use two SotA open-source LLM families: Llama 3 (Meta, 2024), the leader at the time of experiments, and Aya 23 (Aryabumi et al., 2024), renowned for its multilingual pre-training. Given our focus on instructed generation tasks, we use their instruction-tuned versions (Chrysostomou et al., 2024). Our experimental setup includes Llama 3 Instruct models in 8B and 70B parameter sizes, alongside Aya 23 instruction-tuned models in 8B and 35B parameter sizes, with the latter moved to the Appendix for space constraints.

Languages We study seven languages: Arabic (AR), German (DE), English (EN), Spanish (ES), Russian (RU), Swahili (SW), and Chinese (ZH). This selection spans six language families and four writing systems and encompasses both high and mid-resource languages. To ensure consistent

model support, we focus on well-supported languages, while including Swahili as a low-resource outlier, which is not reported to be part of the pre-training data for Llama 3 or Aya 23. Due to the closed-source nature of pre-training and fine-tuning datasets, language-specific data properties cannot be reliably assessed. A summary of languages used in our paper is given in Table 5 in Appendix C.

Pruning and Calibration We construct calibration sets for each of the previously listed seven languages from mC4 (Raffel et al., 2020). Specifically, following Frantar and Alistarh (2023) and Sun et al. (2024), we randomly draw 512 sequences of 1,024 tokens for each language from the mC4 test split for that language, ensuring no duplicates by sampling without replacement. The original Wanda and SparseGPT implementations use 262K calibration tokens. We extend this to 1M tokens for greater sample diversity and potentially more stable pruning, though prior work suggests diminishing returns beyond this size (Sun et al., 2024). For mixing different calibration languages, we mix in equal shares and keep the budget of 512 samples.

We focus on two post-training *pruning methods*, Wanda (Sun et al., 2024) and SparseGPT (Frantar and Alistarh, 2023), and prune for 50% unstructured sparsity, a common setting for LLMs to better observe performance differences while maintaining practical usability (Jaiswal et al., 2024). We keep all other hyperparameters as in the original paper.

Evaluation Downstream Tasks We compare the performance of pruned models calibrated on different languages using *perplexity*, *SNR*, and *pruning error* on a subset of the mC4 validation set, and a selection of downstream tasks in different target languages: MKQA (Longpre et al., 2021), and Belebele (Bandarkar et al., 2024). Additionally, we evaluate on the multilingual mARC, mMMLU, and mHellaSwag (Lai et al., 2023) datasets as well as their original versions, ARC (Clark et al., 2018), MMLU (Hendrycks et al., 2021) and HellaSwag (Zellers et al., 2019) in English, due to their lower sensitivity to the choice of calibration samples (Williams and Aletras, 2024). These tasks primarily assess commonsense reasoning, reading comprehension, and question answering using multiple choice questions. We evaluate in a zero-shot setting. Further details of each task are presented in Appendix D.

Uncertainty Estimation To estimate uncertainty, we apply non-parametric bootstrapping to compute 95% confidence intervals (Efron and Tibshirani, 1993; Berg-Kirkpatrick et al., 2012). For each evaluation configuration, we draw 10,000 bootstrap samples by resampling prediction outcomes with replacement from all test samples across five pruning runs. Each sample includes at least 1,000 instances or 5% of the total, ensuring sufficient variability. This approach captures empirical uncertainty without assuming any particular distributional form.

5 Results

5.1 Pruning Results

Table 1 evaluates 50% sparsity pruning performance for calibration on different languages using three common surface-level pruning metrics. PPL reflects the general language modeling capability, while SNR and pruning error estimate the deviation of activations from the full-sized model. Following Kuzmin et al. (2023), we compute both over all linear sublayers and report size-weighted averages. PPL and SNR, including confidence intervals, are computed globally by aggregating token-level negative log-likelihoods and squared activation differences across five seeds, each using 1,000 samples of 1,024 tokens.

Our findings reveal that no single calibration language consistently outperforms others in perplexity, pruning error, or SNR metrics, despite English-dominated pre-training and unbalanced representation in tokenizer vocabulary. The optimal choice depends on the target evaluation language, with *calibrating in the target language yielding the lowest magnitude deviation from the full-sized model and lowest perplexity*.

5.2 Downstream Task Performance

How to select the calibration language to optimize performance in downstream tasks?

Table 2 shows the models’ performance on downstream tasks. Given a downstream task in a specific language, performance is analyzed column-wise among all calibration languages, where lighter colors indicate better results.

First, for both pruning methods, the calibration language affects downstream performance. For example, when evaluating Llama 3 8B model on Belebele in Spanish, pruning with Wanda results in performance ranging from 47.8 with

	PPL								SNR _{dB}								Pruning Error × 10 ³							
	AR	DE	EN	ES	RU	SW	ZH		AR	DE	EN	ES	RU	SW	ZH		AR	DE	EN	ES	RU	SW	ZH	
--	13.7 _{2.1}	10.7 _{0.4}	10.2 _{0.4}	12.9 _{2.2}	9.2 _{0.4}	11.6 _{0.5}	14.2 _{1.2}																	
Wanda	AR	30.1 _{1.2}	19.1 _{0.8}	18.6 _{0.6}	23.1 _{1.2}	16.6 _{0.6}	37.9 _{2.1}	33.8 _{1.9}	98.5 _{9.3}	95.3 _{0.2}	95.8 _{0.5}	99.2 _{0.3}	98.2 _{0.2}	84.0 _{0.3}	100.2 _{0.7}		40.1 _{0.3}	43.9 _{0.2}	42.5 _{0.5}	40.3 _{0.3}	40.8 _{0.2}	54.8 _{0.4}	37.6 _{0.6}	
	DE	39.1 _{1.7}	16.3 _{0.7}	17.1 _{0.8}	21.0 _{1.1}	16.2 _{0.6}	39.3 _{2.1}	34.2 _{1.9}	91.3 _{0.3}	99.1 _{0.3}	97.2 _{0.5}	100.0 _{0.3}	98.5 _{0.2}	80.9 _{0.3}	99.0 _{0.5}		47.3 _{0.3}	40.3 _{0.2}	41.2 _{0.4}	39.5 _{0.3}	40.5 _{0.2}	58.8 _{0.5}	38.7 _{0.5}	
	EN	44.5 _{2.2}	19.3 _{0.7}	16.5 _{0.8}	21.3 _{1.1}	18.3 _{0.7}	42.1 _{1.8}	35.6 _{0.9}	86.3 _{0.3}	93.2 _{0.2}	98.3 _{0.5}	97.2 _{0.3}	94.1 _{0.2}	78.5 _{0.4}	97.5 _{0.5}		53.1 _{0.4}	46.1 _{0.3}	40.2 _{0.4}	42.2 _{0.3}	44.9 _{0.2}	62.1 _{0.6}	40.0 _{0.5}	
	ES	39.5 _{1.7}	18.2 _{0.7}	17.1 _{0.8}	21.2 _{1.0}	16.6 _{0.6}	36.8 _{2.1}	34.7 _{1.9}	91.8 _{0.3}	95.8 _{0.2}	97.2 _{0.5}	102.3 _{0.3}	97.9 _{0.2}	83.0 _{0.3}	98.3 _{0.5}		46.8 _{0.3}	43.4 _{0.2}	41.2 _{0.4}	37.4 _{0.4}	41.1 _{0.2}	56.0 _{0.5}	39.3 _{0.5}	
	RU	34.6 _{1.4}	18.0 _{0.7}	17.5 _{0.8}	19.2 _{1.1}	15.0 _{0.5}	39.7 _{2.1}	32.9 _{1.9}	93.7 _{0.3}	96.7 _{0.2}	97.3 _{0.5}	100.2 _{0.3}	102.0 _{0.2}	79.4 _{0.4}	100.7 _{0.6}		44.8 _{0.3}	42.5 _{0.2}	41.1 _{0.4}	39.4 _{0.3}	37.4 _{0.2}	60.3 _{0.6}	37.1 _{0.5}	
	SW	39.0 _{1.6}	19.9 _{0.8}	18.8 _{0.6}	22.3 _{1.2}	18.2 _{0.7}	28.7 _{1.7}	38.4 _{2.2}	90.1 _{0.3}	91.5 _{0.2}	91.5 _{0.3}	95.0 _{0.3}	93.4 _{0.2}	92.3 _{0.3}	94.4 _{0.3}		48.7 _{0.3}	48.0 _{0.2}	47.0 _{0.3}	44.3 _{0.3}	45.9 _{0.4}	45.2 _{0.3}	43.0 _{0.6}	
	ZH	39.3 _{1.7}	20.0 _{0.8}	18.2 _{0.8}	22.8 _{1.2}	17.9 _{0.7}	40.8 _{2.1}	31.6 _{1.8}	90.7 _{0.3}	92.7 _{0.2}	95.3 _{0.5}	96.6 _{0.3}	95.6 _{0.2}	80.4 _{0.3}	103.3 _{0.7}		48.0 _{0.3}	46.7 _{0.3}	43.0 _{0.5}	42.7 _{0.3}	43.4 _{0.2}	59.4 _{0.6}	35.0 _{0.6}	
	AR	19.1 _{0.7}	17.4 _{0.7}	16.5 _{0.7}	20.5 _{1.0}	14.8 _{0.5}	29.4 _{1.8}	31.8 _{1.8}	129.4 _{0.7}	101.4 _{0.2}	102.6 _{0.6}	107.2 _{0.3}	107.0 _{0.2}	93.6 _{0.3}	105.7 _{0.7}		19.7 _{0.3}	38.2 _{0.2}	36.3 _{0.5}	33.5 _{0.2}	33.3 _{0.2}	43.9 _{0.4}	33.1 _{0.6}	
Llama 3-8B-Instruct	DE	36.5 _{1.5}	13.6 _{0.5}	15.7 _{0.7}	18.9 _{0.9}	14.6 _{0.5}	32.1 _{2.0}	33.3 _{1.9}	96.2 _{0.3}	121.4 _{0.4}	105.3 _{0.6}	110.2 _{0.3}	108.4 _{0.2}	89.4 _{0.3}	103.7 _{0.7}		42.3 _{0.3}	24.1 _{0.2}	34.2 _{0.5}	31.2 _{0.2}	32.3 _{0.2}	48.4 _{0.4}	34.7 _{0.5}	
	EN	39.8 _{1.7}	17.0 _{0.7}	14.9 _{0.6}	19.8 _{1.0}	16.0 _{0.6}	33.6 _{2.1}	32.4 _{1.8}	92.2 _{0.3}	102.4 _{0.3}	110.9 _{0.7}	105.7 _{0.3}	102.0 _{0.2}	86.9 _{0.4}	104.1 _{0.6}		46.4 _{0.3}	37.3 _{0.3}	30.1 _{0.4}	34.7 _{0.3}	37.4 _{0.2}	51.3 _{0.5}	34.4 _{0.5}	
	ES	36.2 _{1.5}	16.2 _{0.6}	15.8 _{0.7}	16.4 _{0.8}	14.7 _{0.5}	31.9 _{2.0}	33.5 _{1.9}	97.7 _{0.3}	106.6 _{0.3}	105.9 _{0.6}	123.6 _{0.6}	107.6 _{0.2}	90.6 _{0.3}	103.2 _{0.7}		40.8 _{0.3}	33.9 _{0.2}	33.7 _{0.5}	23.0 _{0.3}	32.9 _{0.2}	47.1 _{0.4}	35.1 _{0.6}	
	RU	30.9 _{1.3}	16.3 _{0.6}	16.1 _{0.7}	19.5 _{1.0}	11.8 _{0.4}	33.2 _{2.1}	32.3 _{1.8}	99.6 _{0.3}	105.0 _{0.3}	103.7 _{0.6}	107.7 _{0.3}	126.1 _{0.4}	88.2 _{0.4}	104.9 _{0.7}		39.1 _{0.3}	35.2 _{0.2}	35.5 _{0.5}	33.1 _{0.2}	21.5 _{0.2}	49.7 _{0.5}	33.7 _{0.6}	
	SW	29.2 _{1.2}	17.6 _{0.7}	16.4 _{0.7}	19.7 _{1.0}	16.2 _{0.6}	16.1 _{1.0}	31.7 _{2.0}	102.0 _{0.3}	99.9 _{0.2}	101.1 _{0.6}	105.4 _{0.3}	101.4 _{0.2}	134.2 _{0.5}	100.3 _{0.6}		37.0 _{0.3}	39.5 _{0.2}	37.6 _{0.5}	34.9 _{0.3}	37.9 _{0.2}	17.2 _{0.2}	37.5 _{0.6}	
	ZH	34.4 _{1.4}	18.1 _{0.8}	16.4 _{0.7}	20.8 _{1.0}	15.6 _{0.6}	33.5 _{2.1}	22.3 _{1.1}	97.5 _{0.3}	98.3 _{0.2}	102.5 _{0.5}	102.3 _{0.3}	103.2 _{0.2}	88.7 _{0.4}	133.8 _{0.8}		41.0 _{0.3}	41.0 _{0.2}	36.5 _{0.4}	37.5 _{0.2}	36.4 _{0.2}	49.2 _{0.4}	17.3 _{0.3}	
	--	8.2 _{3.3}	10.0 _{0.5}	9.0 _{0.5}	12.6 _{0.2}	12.0 _{0.6}	38.3 _{1.5}	14.5 _{1.7}																
	Wanda	AR	12.9 _{0.5}	15.4 _{0.7}	13.5 _{0.6}	20.0 _{1.0}	19.5 _{0.9}	78.1 _{6.7}	28.8 _{2.1}	90.7 _{0.6}	89.0 _{0.3}	94.2 _{0.4}	92.4 _{0.5}	91.7 _{0.4}	85.0 _{0.2}	85.5 _{0.5}		74.7 _{1.0}	84.6 _{0.7}	74.5 _{0.7}	79.5 _{1.0}	80.3 _{0.9}	85.1 _{0.6}	82.5 _{1.0}
DE		15.6 _{0.7}	14.2 _{0.6}	13.3 _{0.6}	19.8 _{1.0}	19.9 _{0.9}	80.2 _{7.1}	30.3 _{2.2}	83.5 _{0.5}	93.6 _{0.3}	95.7 _{0.4}	93.9 _{0.5}	92.6 _{0.5}	84.1 _{0.5}	84.1 _{0.5}		88.1 _{1.0}	76.1 _{0.6}	72.0 _{0.7}	76.9 _{0.8}	78.8 _{1.0}	87.1 _{0.6}	85.3 _{1.1}	
EN		16.5 _{0.7}	15.5 _{0.6}	13.1 _{0.6}	20.2 _{1.1}	21.2 _{1.0}	80.2 _{7.1}	30.4 _{2.2}	81.5 _{0.5}	89.7 _{0.3}	97.0 _{0.4}	92.9 _{0.5}	89.9 _{0.4}	83.7 _{0.3}	83.9 _{0.5}		92.5 _{1.0}	83.2 _{0.6}	69.9 _{0.7}	78.6 _{0.9}	83.0 _{1.0}	87.7 _{0.7}	85.6 _{1.1}	
ES		15.6 _{0.7}	15.2 _{0.6}	13.3 _{0.6}	18.8 _{0.9}	20.1 _{0.9}	78.0 _{6.8}	30.3 _{2.2}	84.0 _{0.5}	91.0 _{0.3}	95.9 _{0.4}	96.6 _{0.5}	89.5 _{0.4}	85.3 _{0.2}	83.9 _{0.6}		87.3 _{1.0}	80.9 _{0.6}	71.7 _{0.7}	72.2 _{0.9}	79.0 _{0.8}	84.6 _{0.6}	85.6 _{1.1}	
RU		14.9 _{0.7}	15.0 _{0.6}	13.4 _{0.6}	19.8 _{1.0}	17.7 _{0.7}	80.3 _{7.1}	29.4 _{2.1}	85.5 _{0.5}	90.9 _{0.3}	95.2 _{0.4}	93.9 _{0.5}	96.5 _{0.5}	84.4 _{0.2}	84.8 _{0.6}		84.3 _{1.0}	81.1 _{0.7}	72.8 _{0.7}	76.9 _{0.9}	72.0 _{0.8}	86.2 _{0.6}	83.9 _{1.1}	
SW		16.0 _{0.7}	16.4 _{0.7}	14.1 _{0.7}	21.3 _{1.1}	22.8 _{1.1}	69.0 _{8.2}	32.0 _{2.4}	81.2 _{0.4}	85.6 _{0.3}	90.7 _{0.4}	88.6 _{0.5}	86.7 _{0.4}	91.3 _{0.2}	81.4 _{0.6}		92.9 _{1.0}	81.6 _{0.7}	80.7 _{0.8}	86.8 _{1.0}	80.4 _{1.0}	73.6 _{0.5}	90.6 _{1.1}	
ZH		15.4 _{0.7}	16.0 _{0.7}	13.7 _{0.7}	21.0 _{1.1}	21.1 _{1.0}	82.0 _{7.1}	26.3 _{1.9}	83.5 _{0.5}	87.8 _{0.3}	94.1 _{0.4}	91.1 _{0.4}	89.6 _{0.4}	83.3 _{0.2}	90.5 _{0.5}		88.1 _{0.9}	87.0 _{0.6}	74.7 _{0.7}	81.9 _{0.9}	84.4 _{0.9}	88.4 _{0.6}	73.5 _{0.9}	
AR		10.7 _{0.4}	15.5 _{0.6}	13.5 _{0.6}	20.1 _{1.0}	19.5 _{0.8}	76.4 _{6.6}	29.0 _{2.1}	120.9 _{0.6}	98.6 _{0.3}	105.7 _{0.5}	104.2 _{0.6}	102.7 _{0.3}	97.7 _{0.3}	91.0 _{0.5}		37.3 _{0.5}	67.9 _{0.4}	57.1 _{0.6}	60.6 _{0.8}	62.4 _{0.4}	63.6 _{0.4}	72.7 _{0.6}	
Aya 23-8B	DE	15.6 _{0.7}	12.7 _{0.4}	12.9 _{0.6}	19.1 _{1.0}	19.3 _{0.8}	79.4 _{6.8}	30.3 _{2.2}	89.0 _{0.6}	119.7 _{0.5}	110.2 _{0.6}	108.8 _{0.4}	106.7 _{0.3}	94.9 _{0.3}	88.3 _{0.6}		77.8 _{0.7}	41.8 _{0.4}	51.5 _{0.6}	54.5 _{0.5}	57.0 _{0.4}	67.8 _{0.5}	77.3 _{0.4}	
	EN	16.2 _{0.7}	14.9 _{0.6}	12.4 _{0.6}	19.6 _{1.0}	20.1 _{0.9}	77.9 _{6.7}	30.6 _{2.2}	86.5 _{0.4}	103.6 _{0.3}	117.8 _{0.6}	107.0 _{0.5}	102.0 _{0.3}	95.9 _{0.3}	88.2 _{0.6}		82.3 _{0.6}	61.4 _{0.4}	43.3 _{0.5}	56.8 _{0.6}	63.5 _{0.6}	66.1 _{0.5}	77.5 _{1.1}	
	ES	15.6 _{0.7}	14.6 _{0.6}	12.9 _{0.6}	16.9 _{0.8}	19.5 _{0.8}	77.5 _{6.8}	30.5 _{2.2}	89.0 _{0.6}	104.6 _{0.3}	110.5 _{0.6}	123.1 _{0.6}	105.5 _{0.3}	97.1 _{0.3}	88.3 _{0.6}		76.2 _{0.6}	59.1 _{0.4}	51.1 _{0.6}	39.3 _{0.5}	58.5 _{0.4}	64.4 _{0.5}	77.3 _{0.4}	
	RU	14.0 _{0.6}	14.7 _{0.6}	13.2 _{0.6}	19.5 _{1.0}	15.2 _{0.6}	82.6 _{7.1}	29.4 _{2.1}	91.9 _{0.4}	103.6 _{0.3}	107.6 _{0.5}	106.8 _{0.5}	126.8 _{0.4}	93.9 _{0.2}	89.2 _{0.6}		72.8 _{0.7}	60.5 _{0.4}	54.7 _{0.6}	57.1 _{0.6}	35.8 _{0.3}	69.3 _{0.5}	75.7 _{1.1}	
	SW	14.0 _{0.6}	15.9 _{0.6}	13.5 _{0.6}	20.0 _{1.0}	20.6 _{0.9}	49.9 _{8.3}	31.0 _{2.2}	92.7 _{0.4}	97.3 _{0.3}	104.6 _{0.5}	102.7 _{0.4}	99.1 _{0.3}	143.7 _{0.6}	86.4 _{0.6}		71.5 _{0.6}	69.9 _{0.5}	58.7 _{0.6}	62.7 _{0.6}	67.8 _{0.4}	22.0 _{0.3}	80.8 _{1.1}	
	ZH	15.2 _{0.7}	16.1 _{0.7}	13.6 _{0.7}	21.7 _{1.1}	20.7 _{0.9}	81.9 _{7.1}	21.6 _{1.5}	89.2 _{0.4}	96.0 _{0.3}	103.8 _{0.4}	99.2 _{0.5}	99.2 _{0.3}	94.8 _{0.6}	121.4 _{0.6}		77.3 _{0.6}	72.1 _{0.4}	59.8 _{0.5}	68.0 _{0.8}	67.6 _{0.4}	67.9 _{0.5}	36.1 _{0.6}	

Table 1: Language-specific perplexity (PPL), signal-to-noise ratio (SNR), and pruning error at 50% unstructured sparsity over five pruning runs. The leftmost columns show the model, pruning method, and calibration language (“--” for unpruned baseline). Columns report performance per evaluation language. The lighter the **column-wise color-coding**, the better. Sub- and superscripts show distances to bootstrapped 95% confidence intervals. We omit both if negligible or one of them if equal.

	ARC _[acc]						Belebele _[acc]						MMLU _[acc]						HellaSwag _[acc]					
	AR	DE	EN	ES	RU	Z																		

on MMLU, the pruned Llama 3 8B model mainly achieves higher accuracy on evaluation languages other than on its calibration languages. For Llama 3 (Aya 23), in 58.3% (45.8%) comparison cases, calibration with the target language yields the best performance, 54.2% (29.2%) for Wanda and 62.5% (62.5%) for SparseGPT (e.g. Russian on ARC and MMLU). Comparing calibration with the target language against all other calibrations per pruning configuration and evaluation language yields a global win rate of 66.8% with 69.9% (63.8%) per model, 65.2% (63%) for Wanda and 74.6% (64.5%) for SparseGPT. Therefore, calibrating using the target language often results in acceptable, though not consistently the best performance. ***Calibration with the target language benefits but does not guarantee peak performance for reasoning downstream tasks.***

Downstream performance for baseline models varies by language due to pertaining distribution and an unbalanced tokenizer. English performs best, followed by other Latin-based languages, then Russian. Arabic and Chinese downstream tasks are the most challenging. However, despite the dominance of English in pretraining, it is not the optimal calibration language. Pruning can alter the original ranking of languages observed in the baseline models. For example, on the Belebele benchmark, the Llama 3 8B model scores 66.1 for German and 61.0 for Russian, but the pruned models reverse this trend and achieve a peak accuracy of 47.5 for German and 51.1 for Russian. That is, ***pruning can shift which languages the model performs best or worst on.***

Does calibrating on an outlier language or a similar one benefit downstream tasks in non-English?

Swahili is an out-of-domain language for Llama 3 and Aya 23 as it is not included in the pre-training corpora. In column-wise comparison, the SW cells are among the darkest, indicating that Swahili calibration yields performance among the worst across tasks.

No clear pattern emerges regarding the similarity of calibration-evaluation language pairs. Latin language pairs such as EN-ES (calibrating in English and evaluating in Spanish) or pairs from the same language family, like EN-DE, do not always yield optimal performance. Conversely, pairs with different writing systems, such as RU-EN or ES-AR, do not consistently perform poorly. However, calibration with a highly

dissimilar language, i.e., Chinese, often results in particularly low accuracy across many tasks and evaluation languages, as demonstrated by the darker row of ZH. In summary, ***for the limited amount of calibration languages compared, we cannot observe an immediate benefit from calibrating with an outlier language or a similar language to the target language.***

Does the model or the pruning method matter?

The results discussed above apply to both Llama 3 and Aya 23 model types as evidenced in Table 10 in Appendix F and Table 2.

However, between Llama 3 8B and Aya 23 8B, despite their similar decoder-only architecture, there are distinct performance patterns. Aya 23 8B generally outperforms the Llama 3 8B model in most evaluation languages and tasks, both before and after pruning. Notably, Aya 23 8B experiences less performance drop after pruning but shows less stable results, often performing better on languages other than the one used for calibration.

Between Wanda and SparseGPT, Llama 3 8B’s and Aya 23 8B’s performance degrades less after SparseGPT pruning. Our findings highlight the future work required for more robust pruning methods across downstream tasks in different languages.

5.3 Open Domain Question Answering without Context

How does pruning impact the knowledge of LLMs?

MKQA is a ‘closed-book’ question-answering task that requires the model to generate answers based solely on its internal knowledge, without external context.¹ As shown in Table 3, full-size models exhibit significant performance differences across evaluation languages, with Latin languages performing best and Arabic and Chinese performing worst. Pruning leads to a notable accuracy drop across languages, even for English performance. In summary, ***pruning substantially impacts the storage and retrieval of knowledge in a multilingual model across different languages.***

5.4 Multiple Calibration Languages

Will more calibration languages benefit the downstream?

¹To ensure fair cross-languages comparisons, the MKQA dataset is fully parallel and primarily consists of entity-based and structured ‘atomic’ answer types. See Appendix D for details.

		MKQA _[f1]						
		AR	DE	EN	ES	RU	ZH	
Llama 3-8B-Instruct	Wanda	-	9.1 _{1.6}	27.0 _{2.6} ^{0.6}	38.6 _{2.7}	27.4 _{2.5} ^{0.6}	16.6 _{2.0} ^{0.1}	2.6 _{0.9} ^{1.0}
		AR	0.5 _{0.2}	5.7 _{0.6} ^{0.7}	18.3 _{1.6} ^{1.8}	12.1 _{1.3} ^{1.4}	5.3 _{0.9}	2.0 _{0.5} ^{0.6}
		DE	0.3 _{0.1} ^{0.2}	6.0 _{0.7}	19.8 _{1.7} ^{1.8}	11.7 _{1.4}	5.9 _{1.0}	2.1 _{0.5} ^{0.6}
		EN	0.1	5.4 _{0.7}	20.4 _{1.7}	12.0 _{1.4}	5.2 _{0.9}	1.5 _{0.4}
		ES	0.4 _{0.2}	6.1 _{0.7} ^{0.8}	20.2 _{1.7}	12.6 _{1.4}	5.6 _{1.0} ^{0.9}	0.9 _{0.3} ^{0.4}
		RU	0.2	5.8 _{0.7}	19.2 _{1.6} ^{1.7}	12.4 _{1.3} ^{1.4}	5.7 _{1.0}	1.7 _{0.5}
		ZH	0.5 _{0.2}	6.5 _{0.8}	18.5 _{1.6} ^{1.7}	11.2 _{1.3}	4.5 _{0.8} ^{0.9}	1.6 _{0.5}
	SparseGPT	AR	3.9 _{0.8}	7.6 _{1.0}	20.8 _{1.8} ^{1.9}	12.1 _{1.4} ^{1.5}	6.8 _{1.1}	1.4 _{0.5}
		DE	0.0	10.3 _{1.2}	20.8 _{1.7} ^{1.8}	12.3 _{1.4}	6.5 _{1.0}	1.6 _{0.5}
		EN	0.4 _{0.2}	8.6 _{1.1}	22.4 _{1.8}	13.2 _{1.4} ^{1.5}	6.8 _{1.1} ^{1.0}	1.8 _{0.5} ^{0.6}
		ES	0.4 _{0.2}	8.8 _{1.1}	21.3 _{1.7} ^{1.8}	13.2 _{1.4} ^{1.5}	6.7 _{1.1}	1.5 _{0.5}
		RU	0.4 _{0.2}	8.6 _{1.0} ^{1.1}	20.4 _{1.7}	11.8 _{1.4}	7.4 _{1.1}	1.5 _{0.5}
		SW	1.3 _{0.3}	7.6 _{0.9} ^{1.0}	20.7 _{1.7} ^{1.8}	11.9 _{1.3} ^{1.4}	6.2 _{1.0}	1.1 _{0.4}
		ZH	0.9 _{0.2} ^{0.3}	10.4 _{1.2} ^{1.3}	20.3 _{1.7}	11.2 _{1.3} ^{1.4}	6.5 _{1.1} ^{1.0}	1.4 _{0.4} ^{0.5}
Aya 23-8B	Wanda	-	11.2 _{1.5}	20.5 _{1.9}	32.1 _{2.3} ^{2.4}	17.9 _{1.7} ^{1.8}	13.7 _{1.5} ^{1.6}	0.0
		AR	6.9 _{0.9}	9.3 _{1.0}	16.4 _{1.3}	10.2 _{1.0}	7.0 _{0.9}	0.1
		DE	6.2 _{0.8} ^{0.9}	8.4 _{0.9}	17.7 _{1.4}	10.6 _{1.0}	7.2 _{0.9}	0.1
		EN	5.3 _{0.7} ^{0.8}	8.4 _{0.9}	17.0 _{1.3} ^{1.4}	10.9 _{1.0} ^{1.1}	7.6 _{0.9} ^{1.0}	0.1
		ES	5.8 _{0.8}	8.3 _{0.9}	17.2 _{1.3} ^{1.4}	11.1 _{1.1}	7.9 _{0.9} ^{1.0}	0.1
		RU	5.9 _{0.8}	7.7 _{0.8} ^{0.9}	16.2 _{1.3} ^{1.4}	10.6 _{1.0} ^{1.1}	7.3 _{0.9}	0.1
		ZH	5.9 _{0.9}	8.4 _{0.9} ^{1.0}	16.9 _{1.4}	9.7 _{1.0}	6.7 _{0.9}	0.2 _{0.1}
	SparseGPT	AR	5.8 _{0.8}	8.6 _{0.9} ^{1.0}	19.3 _{1.5} ^{1.6}	10.3 _{1.0}	6.8 _{0.9}	0.1 _{0.1} ^{0.2}
		DE	4.9 _{0.7}	8.9 _{0.9} ^{1.0}	19.3 _{1.5}	10.3 _{1.0}	6.8 _{0.9}	0.3 _{0.3} ^{0.3}
		EN	4.7 _{0.7}	9.1 _{0.9} ^{1.0}	17.1 _{1.4}	10.4 _{1.0}	7.2 _{0.9}	0.1 _{0.1} ^{0.2}
		ES	5.6 _{0.8}	8.9 _{1.0}	19.6 _{1.6}	9.8 _{0.9} ^{1.0}	6.9 _{0.9} ^{1.0}	0.2 _{0.1} ^{0.2}
		RU	5.2 _{0.7} ^{0.8}	8.6 _{1.0}	19.2 _{1.5} ^{1.6}	10.8 _{1.0} ^{1.1}	7.3 _{0.9}	0.2 _{0.1} ^{0.2}
		SW	4.7 _{0.7}	8.5 _{1.0}	18.7 _{1.5}	10.0 _{1.0} ^{1.1}	7.2 _{1.0}	0.3 _{0.2} ^{0.3}
		ZH	4.4 _{0.7}	8.4 _{0.9} ^{1.0}	18.1 _{1.5}	9.7 _{1.0}	5.9 _{0.8} ^{0.9}	0.1

Table 3: Averaged MKQA F1 scores over five pruning runs for the Llama 3 8B and Aya 23 8B models pruned with Wanda and SparseGPT for 50% unstructured sparsity. Sub- and superscripts show the distance to bootstrapped 95% confidence intervals. We omit them if negligible or equal.

We repeat the experiments but include more languages in the calibration set for diversification. We experiment with bilingual calibration as well as including all seven languages in the calibration, i.e., multilingual setup. For all setups, the total calibration sample number remains the same, i.e., 512.

A comparison of Table 4 with Table 2 shows that downstream task performance remains similarly unpredictable for bi- or multilingual calibration as it does for monolingual calibration. Certain language combinations used for calibration yield good performance among a wider range of target languages, such as AR-DE-EN-ES-RU-SW-ZH on Belebele. On the other hand, calibration sets, such as EN-RU, lead to poor performance across nearly all scenarios.

In summary, *multilingual calibration can help retain performance on a wider range of languages on downstream tasks. However, there is no clear pattern identifying which specific language combinations are most effective for a specific downstream language.*

5.5 Impact of Model Sizes

To investigate the scaling impact towards pruning behavior, we repeat experiments in Table 10 on the larger Llama 3 70B and Aya 23 35B models, the results of which are reported in Appendix F. Overall, pruning performance increases with higher baseline accuracy of the full-sized models. However, we observe that *the performance patterns and findings from the smaller models, mentioned above, do not consistently hold true on their bigger counterparts*. For example, while Llama 3 8B exhibits a more predictable diagonal pattern for pruning with SparseGPT, pruning performance for the larger Llama 3 70B depends more on the task than the pruning technique.

5.6 Quantization

We further explore the impact of the calibration language in quantization on downstream performance. We use GPTQ (Frantar et al., 2023) to quantize weights to 4 bits with a group size of 128 and 8 bits with a group size of 128 (equivalent to 50% sparsity in pruning) on LLaMA-3-8B. We follow our pruning setup for downstream tasks and languages for calibration and evaluation. The results are present in Table 11 and 12 in Appendix G, revealing several key findings that are consistent with our findings on pruning: (1) calibrating with the target language yields reasonable although not consistently the best downstream performance; (2) pruning can alter which languages the model performs best or worst on; and (3) calibrating with an outlier or a linguistically similar language does not provide any notable advantage.

6 Internal Representation Analysis

We review Table 1 and Table 2 together and find it an interesting pattern that pruning with target-language calibration effectively prunes the less impactful weights, yielding highest SNR and lowest pruning error. This benefits simple linguistic tasks like general language modeling captured by PPL, but it is less effective for downstream tasks that require knowledge or reasoning.

	ARC _(acc)						Belebele _(acc)						MMLU _(acc)						HellaSwag _(acc)																
	AR	DE	EN-C	ES	RU	ZH	AR	DE	EN	ES	RU	ZH	AR	DE	EN	ES	RU	ZH	AR	DE	EN	ES	RU												
-	29.91	7.38	11.8	48.31	9.39	01.8	36.41	8.36	21.8	45.02	1.66	12.0	42.12	1.72	21.9	61.02	1.32	92.1	25.02	6.36	12.9	31.42	8.43	83.0	29.12	7.24	22.6	36.02	8.43	62.9	52.63	0.45	92.9	41.62	9
ar-de	24.82	6.31	02.8	41.23	34.22	9.30	72.8	30.92	9	39.02	9.50	03.0	66.52	9.55	32.9	45.73	0.39	22.9	27.00	9.34	0.9	42.80	9.37	0.9	30.30	9.27	70.9	33.11	1.39	31.1	49.11	1.41	81.1	37.91	1
ar-en	25.52	4.29	02.6	41.62	81.72	30.12	6.31	12.6	36.82	6.44	42.7	61.22	7.52	42.8	41.52	7.37	02.6	27.21	27.21	5.31	4.6	40.91	6.34	3.1	28.81	5.27	11.9	33.11	38.21	49.41	9.41	6.1	37.11	3	
ar-es	25.32	4.29	42.6	41.02	34.52	6.80	32.6	30.82	38.42	6.45	42.6	60.82	7.52	12.7	45.62	8.37	72.7	26.21	5.30	1.1	39.21	6.34	7.1	28.91	5.26	8.1	33.21	38.31	48.81	42.21	6.37	72.0	0		
ar-ru	25.22	4.29	32.6	40.72	32.52	7.31	32.7	80.52	44.92	8.49	22.8	69.92	5.53	42.7	52.42	8.46	42.7	30.01	5.35	8.1	6.46	7.1	6.37	1.1	6.34	1.1	6.31	33.21	38.51	48.91	41.11	38.21	9		
de-en	23.42	31.62	6.42	62.6	92.72	30.02	31.62	7	31.92	6.51	52.7	69.92	5.53	82.7	46.52	8.44	02.7	25.81	3.35	1.1	6.45	1.1	7.36	1.1	6.30	5.1	29.51	31.41	39.61	49.51	41.51	37.41	9		
de-es	24.62	4.31	82.6	41.52	34.92	6.80	62.6	81.02	36.02	6.50	62.7	67.12	6.55	62.7	46.92	7.43	72.7	25.81	3.34	5.44	2.1	6.37	5.6	80.31	29.41	5.1	31.51	39.51	48.61	42.22	6.37	71.9	0		
de-ru	24.22	4.31	52.6	41.72	32.92	7.32	22.7	81.02	38.42	6.50	12.7	67.42	5.54	32.7	47.82	8.41	92.6	27.71	5.36	9.1	6.45	3.1	7.39	3.1	6.32	8.1	6.30	31.91	39.71	49.11	41.51	38.61	9		
en-es	23.52	4.29	22.7	42.12	35.22	6.29	42.7	31.72	31.32	6.43	12.7	65.82	6.55	62.8	43.32	7.39	72.7	25.01	2.9	9.1	4.1	3.1	6.36	1.1	6.28	7.1	28.21	31.31	38.41	49.41	42.41	6.37	31.9	0	
en-ru	23.92	4.29	12.6	41.42	32.92	7.31	82.6	80.82	31.82	5.42	42.7	61.82	6.49	42.7	41.92	6.34	52.7	26.41	5.32	7.1	5.40	6.1	6.36	5.6	30.21	5.28	4.1	31.81	38.51	49.41	41.31	38.21	9		
es-ru	23.82	4.29	82.6	41.82	35.62	7.30	92.7	80.52	38.62	7.47	92.8	67.82	5.55	62.8	47.72	8.42	12.7	27.01	5.34	2.1	6.43	3.1	6.38	3.1	6.32	3.1	6.30	21.51	31.91	38.91	48.91	42.42	6.38	51.9	0
Five langs.	24.82	6.30	72.8	41.52	35.22	8.30	92.8	81.22	40.22	9.51	43.0	67.82	7.58	12.9	47.62	9.40	12.9	27.21	3.34	3.1	4.3	6.1	2.38	4.1	30.91	1.28	1.1	33.01	39.21	49.21	42.01	4.38	11.4	0	
Seven langs.	25.02	4.30	42.6	41.22	34.62	7.30	72.6	31.52	42.12	6.55	52.7	69.62	5.59	72.7	50.62	8.42	52.7	28.01	5.35	7.1	6.45	5.1	7.39	3.1	7.31	8.1	6.29	11.51	32.91	39.11	49.11	41.92	6.38	11.9	0

Table 4: 50% unstructured sparsity SparseGPT-pruned, mC4-calibrated Llama 3 8B performance averaged over three pruning runs. The lighter, the better. In the row header, the ‘‘Five languages’’ refer to AR, DE, EN, ES, and RU. ‘‘Seven languages’’ include AR, DE, EN, ES, RU, SW, ZH. Sub- and superscripts show the distance to bootstrapped 95% confidence intervals. We omit them if negligible or equal.

We hypothesize that target-language calibration effectively preserves language-specific features, but not language-agnostic ones, such as knowledge retrieval and reasoning abilities for the pruned multilingual LLMs (see Appendix E for qualitative examples).

To test this, we investigate the internal changes of pruned models at three levels: feature subspace, matrix, and neuron level (columns in a matrix, followed by non-linearity). Prior work has separated language-specific features from language-agnostic features at the neuron level (Tang et al., 2024; Zhao et al., 2024; Wang et al., 2024) or via feature subspace extraction (Xie et al., 2022). We examine how these elements shift post-pruning to explain performance differences.

6.1 Language-specific Subspace Representations

We use the Low-rank Subspace for language-Agnostic Representations (LSAR) to separate dominant language-specific features from language-agnostic features (Xie et al., 2022). It constructs a mean language embedding matrix $\mathbf{M} \in \mathbb{R}^{d \times L}$ by concatenating averaged language embeddings $\mathbf{e}^{(l)} \in \mathbb{R}^d$ for L languages. Subsequently, LSAR decomposes \mathbf{M} into a vector $\boldsymbol{\mu}$ representing shared signals across languages and a matrix \mathbf{M}_s specifying a low-rank subspace on which different languages express different linguistic signals. This decomposition process is achieved via singular value decomposition on solving:

$$\min_{\boldsymbol{\mu}, \mathbf{M}_s, \Gamma} \left\| \mathbf{M} - \boldsymbol{\mu} \mathbf{1}^\top - \mathbf{M}_s \Gamma^\top \right\|_F^2 \quad (6)$$

s.t. $\boldsymbol{\mu} \perp \text{Span}(\mathbf{M}_s)$,

with the orthogonality constraint. $\Gamma \in \mathbb{R}^{L \times r}$ represents the coordinates of language-specific signals along the subspace’s r components, and $\mathbf{1} \in \mathbb{R}^d$ is a vector of all ones.

Since token-embeddings, through metrics like pruning error and SNR, poorly predict downstream task performance, we capture more high-level semantic and syntactic information through prompt-embeddings, layer output embeddings averaged per prompt, excluding special chat-template tokens. We extract language-specific features \mathbf{s} from a prompt-embedding \mathbf{e} using \mathbf{M}_s by projecting \mathbf{e} into and back from the low-rank, language signal retaining subspace and obtain the language-agnostic features \mathbf{a} through subtraction with $\mathbf{a} = \mathbf{e} - \mathbf{s} = \mathbf{e} - \mathbf{M}_s \mathbf{M}_s^\top \mathbf{e}$.

We use LSAR to decompose the output of each transformer layer of the full-sized and mC4-calibrated, SparseGPT-pruned Llama 3 8B model. To evaluate pruning-induced changes, we use all 900 fully-parallel Belebele samples per language, with each sample being semantically identical across languages (Bandarkar et al., 2024). For each model m , we compute a separate projection matrix $\mathbf{M}_s^{(m)}$ with Belebele samples from the six calibration languages, excluding SW. As we do not require generalization to unseen data as in Xie et al. (2022), we calculate $\mathbf{M}_s^{(m)}$ and evaluate pruning differences using the same 900 Belebele samples per language.

Figure 1 shows the layer-wise magnitude of differences (Δ magnitude) of (a) language-agnostic features and (b) language-specific features after pruning. A greater Δ magnitude suggests greater hidden state changes after pruning, correlating with increased pruning errors. Calibrating on the target language reduces pruning errors in

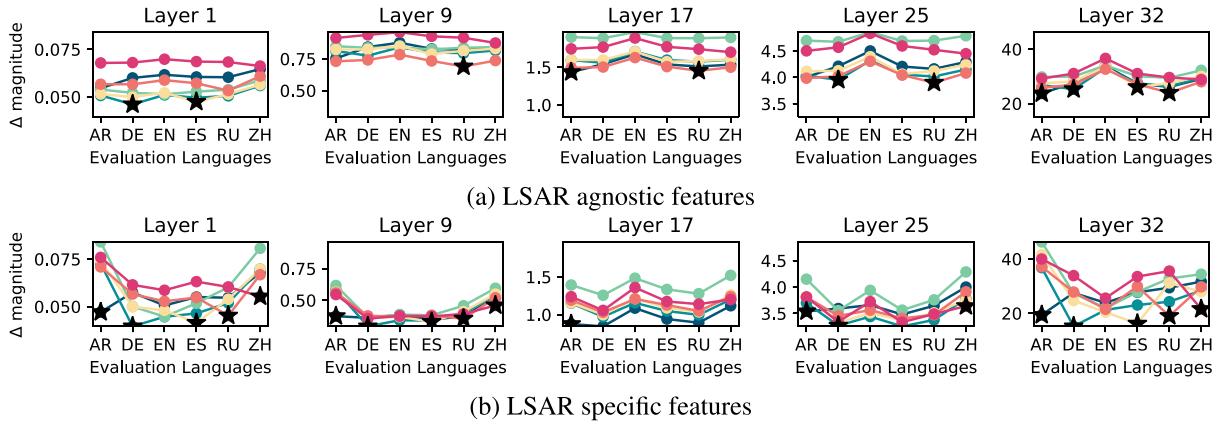


Figure 1: Post-pruning magnitude difference for language-agnostic (Figure 1a) and -specific features (Figure 1b), averaged over 900 Belebele samples per language. The x -axis indicates the evaluation language, the calibration language is color-coded: AR, DE, EN, ES, RU, ZH. Larger Δ means larger pruning error on the respective features. See Figure 4 and 5 in Appendix H.1 for the full plot over all layers. A star marks matching calibration and evaluation languages with the smallest post-pruning difference.

language-specific features (Figure 1b), particularly in early, final, and selected middle layers as indicated by stars. This potentially explains the findings in Section 5.1 that calibration on the target language leads to the lowest perplexity, which is associated with a robust language-specific linguistic modeling capability. On the other hand, as indicated by the relatively flat horizontal lines across languages and layers in Figure 1a, the pruning error on the language-agnostic features remains similar regardless of the calibration languages. This pattern helps explain the sub-optimal downstream task performance observed in Table 10, where no single calibration language consistently yields optimal performance across downstream tasks, including cases where the calibration is performed on the target language. That is, *the selection of calibration language is unlikely to benefit the language-agnostic features associated with understanding, reasoning, and knowledge retrieval. In contrast, the preservation of language-specific features, such as general language modeling capabilities in early and final layers related to perplexity, depends on the selection of calibration languages.*

Second, as an overall pruning trend independent of calibration language, the middle layers (as shown in the second to fourth columns in Figure 1) exhibit greater Δ magnitude on language-agnostic feature representations and smaller Δ magnitude on language-specific feature representations. This indicates that pruning errors can be predominantly attributed to the pruning errors on language-agnostic features, with less

pruning error arising from language-specific features. Therefore, we conclude *pruning affects language-agnostic features, potentially associated with reasoning and knowledge storage, more significantly than it impacts language-specific features.*

6.2 Pruning Mask Similarity

To better understand weight pruning decisions based on the calibration set, we conduct a matrix-level analysis by calculating the Intersection over Union (IoU) of pruning masks across different calibration sets, obtaining a measure of pruning mask similarity. We use the pruning masks from the Llama 3 8B model pruned with SparseGPT for 50% unstructured sparsity. To reduce calibration set-dependent noise as prevalent in the downstream tasks, we first compute the intersection M_l^l of pruned neuron indices M_i^l across three pruned models calibrated with different seeds i but in the same language l . This intersection represents more stable neuron indices.

The IoUs in Figure 2a depict the proportion of M_l^l with respect to all pruned neuron indices. The figure reveals high pruning mask similarity in the attention query, key, and value in the first layer, while the attention output projection varies more significantly, especially after the 20th layer. This suggests that *pruning struggles to consistently identify essential neurons in the attention output projection, partly responsible for the models reasoning capabilities.*

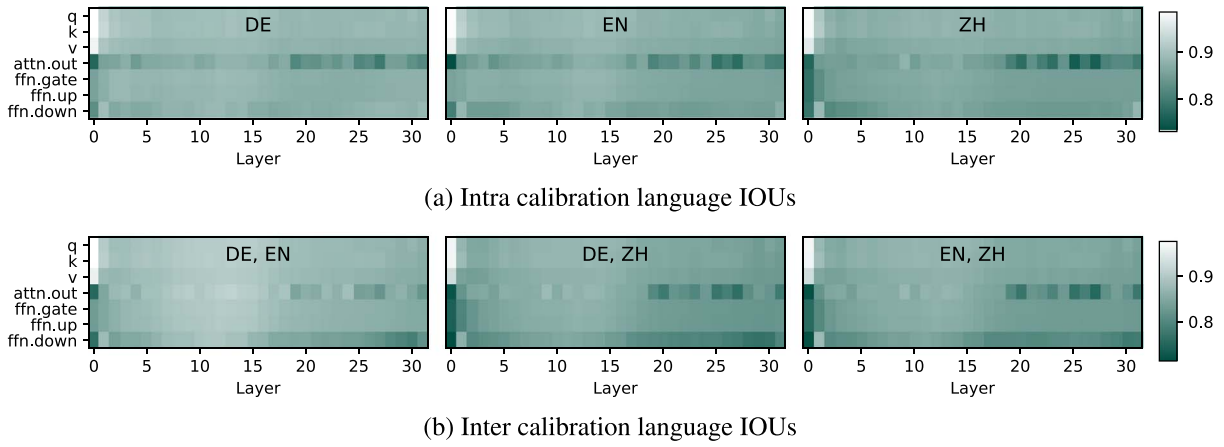


Figure 2: Pruning mask similarities (IoUs) between using different calibration languages for 50% SparseGPT-pruned Llama 3 8B models. 2a) IoUs of pruning masks for three calibration sets of the same language. 2b) IoUs between pruning masks for different calibration languages. The higher IoU (indicated as a lighter color), the more similar pruning masks between different calibration languages.

Figure 2b compares the IoU of intersected neuron indices M_j^l from Figure 2a across calibration in German, English, and Chinese. Notably, the attention query, key, and value in the first layer consistently achieve high IoUs above 0.95 across all languages, indicating that these components handle inputs similarly, irrespective of language differences. However, the attention output and FFN down projection show lower IoU, especially in early layers, with similarity peaking in middle layers (3rd to 15th) before decreasing again in later layers.

This pattern suggests that the attention output and FFN down-projection in early and late layers handle language-specific signals, while middle layers process language-agnostic signals, supporting Figure 1. In other words, **early layers focus on language comprehension, middle layers on language-independent reasoning, and later layers on generating language-specific predictions.** We make similar observations for the Aya 23 8B model as evidenced in Figure 6 of Appendix H.2. This aligns with Zhao et al. (2024), who propose that LLMs first comprehend queries by converting multilingual inputs into English in the early layers, reason in English in the intermediate layers, and then generate responses aligned with the input language in the final layers.

6.3 Language-specific Neurons

This section investigates neuron-level activation frequency changes after pruning using Language Activation Probability Entropy (LAPE) as intro-

duced by Tang et al. (2024). We focus on neurons in the up projection of the FFN layers, followed by the non-linearity. A neuron is considered activated when the non-linearity output is greater than zero. LAPE measures the likelihood of individual neurons i of a layer j activating across different language inputs, identifying neurons with high activation probability $p_{i,j}^{(k)}$ for a language k but low probabilities for all others (i.e., low LAPE score) as language-specific. Then, neuron-wise LAPE scores with L1 normalized $p_{i,j}^{(k)}$ are computed as:

$$LAPE_{i,j} = - \sum_{k=1}^L \tilde{p}_{i,j}^{(k)} \cdot \log(\tilde{p}_{i,j}^{(k)}) \quad (7)$$

$$\tilde{p}_{i,j} = \frac{(p_{i,j}^{(1)}, \dots, p_{i,j}^{(k)}, \dots, p_{i,j}^{(L)})}{\|(p_{i,j}^{(1)}, \dots, p_{i,j}^{(k)}, \dots, p_{i,j}^{(L)})\|_1}$$

Since LAPE scores originally summarize neuron activations across multiple languages, we adapt them into a language-specific metric by correlating them with activation probabilities in a single language. Specifically, we exclude neurons whose activation probability in the target language is lower than the average across all languages, removing those that are language-specific to another language. This correlation (below -0.7 Pearson correlation coefficient) not only refines LAPE as a language-specific metric but also enables linking post-pruning changes to language-specific and language-agnostic neuron activation patterns.

Figure 3 presents LAPE score statistics for the full-sized Llama 3 8B model and its DE-mC4

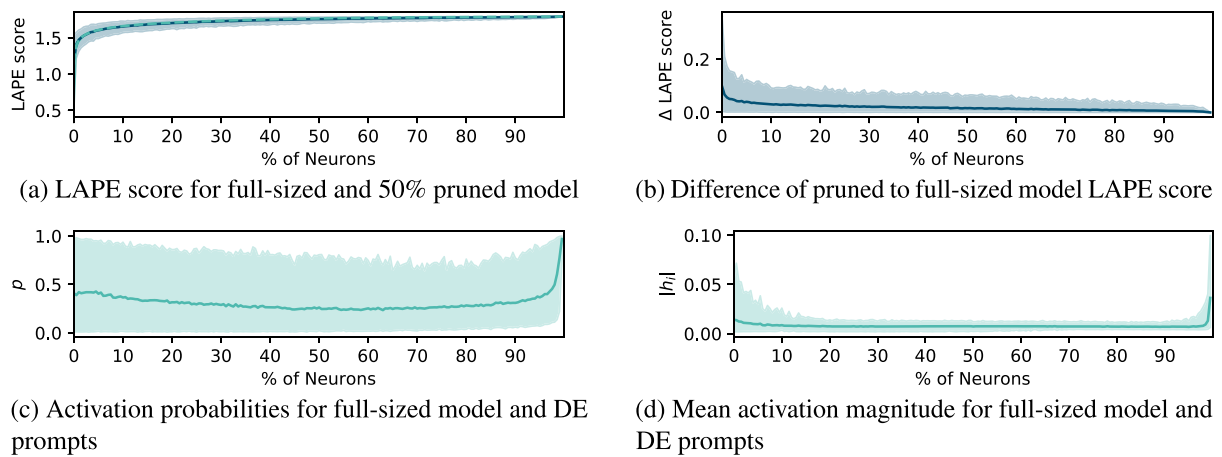


Figure 3: Statistics for FFN neurons of a **full-sized Llama 3 8B** and its **50% SparseGPT-pruned** version calibrated for DE. Plots show a 95% confidence interval and highlighted mean values. All neurons are ordered by ascending LAPE score of the full-sized model as shown by the dashed line in 3a. Additionally, LAPE score and activation probabilities are correlated by removing all neurons with an activation probability in DE that is less than the average activation probability among all languages. *The lower the LAPE score, the more specialized the neuron is for a particular language.*

calibrated, 50% SparseGPT-pruned version. Activation probabilities were computed from 128 samples per language, each containing 8192 tokens, drawn from the mC4 validation set across the seven previously used calibration languages. All neurons are sorted in ascending order based on LAPE scores in the full-sized model correlated with DE, placing language-specific neurons on the left and language-agnostic neurons, i.e., those firing across all languages, on the right. Additionally, all metrics are plotted with a 95% confidence interval and highlighted mean values.

Figure 3a shows a quickly diminishing variance for higher LAPE scores post-pruning. This suggests a lower impact of pruning on the activation frequency of high LAPE neurons, i.e., language-agnostic neurons. Isolating the LAPE score difference of the pruned to the full-sized model, Figure 3b specifically highlights a **strong pruning impact on low LAPE scores, shifting the LAPE score distribution and creating new language-specific (low LAPE) and agnostic (high LAPE) neurons**. In other words, pruning changes activation distributions, potentially causing neurons to activate more or less frequently for a given language. Such shift in FFN layers may contribute to performance degradation in downstream tasks. This hypothesis aligns with previous causal tracing studies, which have identified FFN layers are crucial for knowledge retention and retrieval (Meng et al., 2022, 2023).

Moreover, Figure 3c shows that neurons of lowest LAPE, those most affected by pruning, have low mean activation probabilities compared to neurons with the highest LAPE scores that activate more frequently. This indicates that **pruning struggles to retain the activation frequency of language-specific neurons of low activation probability**. We also examine the average neuron output magnitude for these neglected neurons in Figure 3d, depicting that despite the low activation probability, low LAPE neurons tend to have high output magnitudes. Pruned models show no significant differences in output magnitudes, leading us to conclude that pruning retains average neuron output magnitudes in FFN modules but may fail to preserve activation frequency. Figure 7 and 8 in Appendix H.3 confirm the previous findings for the Aya 23 8B model and Wanda-pruning.

7 Conclusion

This paper explores how calibration language choice affects pruned multilingual model performance, targeting a specific language for downstream tasks. We find that while calibrating on the target language maintains perplexity scores, it does not consistently benefit downstream task performance. In fact, calibration in a different language can sometimes yield better results, highlighting limitations in existing pruning strategies.

Our analysis of internal model representations shows that target language calibration provides limited benefits for features, layers and neurons that encode knowledge and reasoning. Current pruning methods prioritize dominant language-specific features, neglecting critical language-agnostic components that are essential for complex tasks. This calls for more effective pruning strategies that preserve nuanced language-agnostic information, more reliably identify important neurons across calibration sets with less variance and maintain language-specific activation patterns rather than just focusing on feature magnitudes.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

We thank Eryk Majoch,² then an undergraduate in Computer Science at the University of Sheffield, for his valuable contributions to experiments during his internship at the university early in the project.

References

- Marah I. Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norrick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.
- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. 2019. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032v2*.
- Abhinav Bandari, Lu Yin, Cheng-Yu Hsieh, Ajay Kumar Jaiswal, Tianlong Chen, Li Shen, Ranjay Krishna, and Shiwei Liu. 2024. Is C4 dataset optimal for pruning? An investigation of calibration data for LLM pruning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18089–18099, Miami, Florida, USA. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1004>
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.44>

²emajoch1@sheffield.ac.uk.

- Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.
- Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian’s, Malta. Association for Computational Linguistics.
- Rochelle Choenni and Ivan Titov. 2025. M-wanda: Improving one-shot pruning for multilingual LLMs. *CoRR*, abs/2505.21171. <https://doi.org/10.18653/v1/2025.emnlp-main.1369>
- George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181. <https://doi.org/10.1162/tacla.00695>
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948v1*.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.
- Bradley Efron and Ryan J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman &

- Hall. <https://doi.org/10.1007/978-1-4899-4541-9>
- Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. JMLR.org.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation. <https://doi.org/10.5281/zenodo.12608602>
- Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2022. A survey of quantization methods for efficient neural network inference. *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC. <https://doi.org/10.1201/9781003162810-13>
- Jialong Guo, Xinghao Chen, Yehui Tang, and Yunhe Wang. 2025. SlimLLM: Accurate structured pruning for large language models. In *Forty-second International Conference on Machine Learning*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Duc N.M. Hoang, Shiwei Liu, Radu Marculescu, and Zhangyang Wang. 2023. Revisiting pruning at initialization through the lens of Ramanujan graph. In *The Eleventh International Conference on Learning Representations*.
- Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. 2021. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(1):1–124.
- Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110, Tórshavn, the Faroe Islands. Association for Computational Linguistics.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.826>
- Ajay Kumar Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. 2024. Compressing LLMs: The truth is rarely pure and never simple. In *The Twelfth International Conference on Learning Representations*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Rasmus Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard. 2022. Are multilingual sentiment models equally right for the right reasons? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.blackboxnlp-1.11>
- Hwichan Kim, Jun Suzuki, Toshio Hirasawa, and Mamoru Komachi. 2024. Pruning multilingual large language models for multilingual

- inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9921–9942, Miami, Florida, USA. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.580>
- Andrey Kuzmin, Markus Nagel, Mart Van Baalen, Arash Behboodi, and Tijmen Blankevoort. 2023. Pruning vs quantization: Which is better? In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. 2022. Fp8 quantization: The power of the exponent. *Advances in Neural Information Processing Systems*, 35:14651–14662.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276
- Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-demo.28>
- Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.
- Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. 2025. GPTAQ: Efficient finetuning-free quantization for asymmetric calibration. In *Forty-second International Conference on Machine Learning*.
- Shiwei Liu, Tianlong Chen, Xiaohan Chen, Zahra Atashgahi, Lu Yin, Huanyu Kou, Li Shen, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. 2021. Sparse training via boosting pruning plasticity with neuroregeneration. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34 of *NIPS '21*, pages 9908–9922, Red Hook, NY, USA. Curran Associates Inc.
- Yifei Liu, Jicheng Wen, Yang Wang, Shengyu Ye, Li Lina Zhang, Ting Cao, Cheng Li, and Mao Yang. 2024. VPTQ: Extreme low-bit vector post-training quantization for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8181–8196, Miami, Florida, USA. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.467>
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406. https://doi.org/10.1162/tacl_a.00433
- Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.414>
- Kevin Meng, David Bau, Alex J. Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.
- Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.
- Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date—ai.meta.com. <https://ai.meta.com/blog/meta-llama-3/>. Accessed 15-07-2024.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia

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 Kosic, 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, C. J. 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. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774v6*.

Wessel Poelman and Miryam de Lhoneux. 2025. The roles of English in evaluating multilingual language models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic*

- Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 492–498, Tallinn, Estonia. University of Tartu Library.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).
- Maksym Shamrai. 2024. Language-specific pruning for efficient reduction of large language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 135–140, Torino, Italia. ELRA and ICCL.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.
- Sungbin Shin, Wonpyo Park, Jaeho Lee, and Namhoon Lee. 2024. Rethinking pruning large language models: Benefits and pitfalls of reconstruction error minimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1182–1191, Miami, Florida, USA. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.68>
- Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*.
- 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. <https://doi.org/10.18653/v1/2024.acl-long.309>
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971v1*.
- Yilei Tu, Andrew Xue, and Freda Shi. 2025. Blessing of multilinguality: A systematic analysis of multilingual in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6213–6248, Vienna, Austria. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.323>
- Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.
- Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12159–12173, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.724>
- Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.253>
- Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023. All languages matter: On the multilingual safety of large language models. *CoRR*, abs/2310.00905.
- Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings*

- of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.544>
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.379>
- Weihang Xu and Simon Du. 2023. Overparameterization exponentially slows down gradient descent for learning a single neuron. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1155–1198. PMLR.
- Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362. <https://doi.org/10.1007/s11704-024-40579-4>
- Yifan Yang, Kai Zhen, Bhavana Ganesh, Aram Galstyan, Goeric Huybrechts, Markus Müller, Jonas M. Kübler, Rupak Vignesh Swaminathan, Athanasios Mouchtaris, Sravan Babu Bodapati, Nathan Susanj, Zheng Zhang, Jack FitzGerald, and Abhishek Kumar. 2025. Wanda++: Pruning large language models via regional gradients. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4321–4333, Vienna, Austria. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.224>
- Geng Yuan, Xiaolong Ma, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, Yifan Gong, Zheng Zhan, Chaoyang He, Qing Jin, et al. 2021. Mest: Accurate and fast memory-economic sparse training framework on the edge. *Advances in Neural Information Processing Systems*, 34:20838–20850.
- Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, and Moshe Wasserblat. 2021. Prune once for all: Sparse pre-trained language models. *arXiv preprint arXiv:2111.05754v1*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1472>
- Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2024. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11794–11812, Torino, Italia. ELRA and ICCL.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.
- Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. 2024. Plug-and-play: An efficient post-training pruning method for large language models. In *The Twelfth International Conference on Learning Representations*. <https://doi.org/10.20944/preprints202310.1487.v2>
- Yingtao Zhang, Jialin Zhao, Wenjing Wu, Alessandro Muscoloni, and Carlo Vittorio Cannistraci. 2023. Epitopological sparse ultra-deep learning: A brain-network topological

theory carves communities in sparse and percolated hyperbolic ANNs. *Preprints*. <https://doi.org/10.20944/preprints202207.0139.v2>

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3226–3244, Mexico City, Mexico. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.178>

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577. https://doi.org/10.1162/tacl_a_00704

A Implementation Details

We adopt the code from Sun et al. (2024) for implementing model pruning. We use EleutherAI Evaluation Harness (Gao et al., 2024) for a robust and reproducible evaluation. We use HuggingFace (Wolf et al., 2020) for loading datasets and models. All experiments are conducted with at most two NVIDIA A100 GPUs.

B Limitations

Generality of Findings. Due to resource constraints, we predominantly experimented with the small versions of Llama 3 and Aya 23, and validated our findings with fewer pruning runs on their counterpart large version. Since our results translate between model families, and to bigger model sizes, we assume a certain degree of generalization. Nonetheless, other models trained with different techniques or for other tasks might show different behavior. Given the pace of this research field, it is also unclear whether these results translate to future models.

Underrepresented Languages. Our experiments focused on languages with sufficient support for models and downstream tasks. However, this selection does not encompass all languages of interest, particularly mid and low-resource languages that are underrepresented in the pre-training, and challenging to evaluate due to the lack of benchmark support. Future research could benefit from including more languages to explore the interplay between different language families or writing systems and performance after pruning.

C Calibration and Test Languages

Language (Code)	Language Family	Writing System	Script Direction	Morphological Type	Geographic Region
Modern Standard Arabic (AR)	Afro-Asiatic	Arabic	Right-to-left	Fusional	Middle East
German (DE)	Germanic	Latin	Left-to-right	Fusional	Western Europe
English (EN)	Germanic	Latin	Left-to-right	Analytic	Global
Spanish (ES)	Romance	Latin	Left-to-right	Fusional	Southern Europe/Latin America
Russian (RU)	Balto-Slavic	Cyrillic	Left-to-right	Fusional	Eastern Europe
Swahili (SW)	Atlantic-Congo	Latin	Left-to-right	Agglutinative	East Africa
Chinese (simplified) (ZH)	Sino-Tibetan	Simplified Han	Left-to-right	Isolating	East Asia

Table 5: Summary of languages included in this paper for calibration and evaluation.

D Downstream Datasets

Throughout the paper we used the following widely employed datasets for automated benchmarking. All evaluations were conducted in a zero-shot fashion and employ the chat-template of the respective instruction-tuned model.

ARC: The AI2 Reasoning Challenge (ARC) dataset introduced by Clark et al. (2018) tests the reasoning and knowledge capabilities through natural, grad-school multiple choice science questions originally authored for standardized human tests. The dataset comprises a total of 7787 questions in English divided into a Challenge set (ARC-C) of hard to answer questions and an Easy set (ARC-E) of questions.

For evaluation in English, we use the original datasets (e.g., ARC-c & ARC-e), for all other languages the translated version from Lai et al. (2023) is utilized.

Belebele: This carefully curated dataset evaluates 4-way multiple-choice machine reading comprehension among 122 language options, broadly focussing on high-, medium-, and low-resource languages (Bandarkar et al., 2024). Each of the 900 samples is based on an English FLORES-200 passage that has been translated into the respective target language by fluent expert speakers. Hence, the dataset is fully parallel, allowing direct performance comparison across all languages.

HellaSwag: The HellaSwag dataset by Zellers et al. (2019) comprises 10,042 English samples testing commonsense natural language inference on event descriptions that need to be continued/completed in a multiple-choice fashion. Though easily answerable by humans, such paragraph continuation questions still pose a challenge for SotA LLMs.

MKQA: Multilingual Knowledge Questions and Answers (MKQA) (Longpre et al., 2021) is an open-domain question-answering evaluation set of 10,000 samples aligned across 26 languages by human translators. Its question-answer pairs were filtered from the Google Natural Questions dataset (Kwiatkowski et al., 2019), annotating real Google search user questions with answers found on Wikipedia. Given a question, the task is to predict the correct answer or give no answer without any additional context provided. Hence, this dataset tests the knowledge retrieval capabilities of models. For our evaluation, we remove all unanswerable and questions requiring overly long answers for simplicity, yielding a total 6,758 remaining samples.

MMLU: The Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) is an English benchmark designed to evaluate a model’s ability to handle diverse subjects across multiple domains. It contains a total of 14,042 question-answer pairs covering 57 task categories, ranging from high school and college-level subjects to professional and specialized knowledge. Each task includes multiple-choice questions, and the dataset measures both the model’s factual knowledge and reasoning abilities.

Translated Datasets from Okapi: The Okapi framework, introduced by Lai et al. (2023), focuses on instruction tuning LLMs using reinforcement learning from human feedback (RLHF) across multiple languages. As part of its resources, it includes translated versions of the ARC, HellaSwag, and MMLU datasets, generated using ChatGPT. We leverage these translations to complement the evaluation of the original English datasets in multiple languages.

E Qualitative Generation Results of Pruned Models

Prompt: Bernice studied some organisms that live together in a field. In which way is a rabbit similar to the plant it eats?

Answer (EN-calibrated): A clever question! \n\nBernice studied organisms that live together in a field, and she found that a rabbit is similar to the plant it eats in the way that both are herbivores. Just like the plant, a rabbit is a herbivore, meaning it eats plants and vegetation.

Answer (DE-calibrated): Bernice is studying the organisms that live together in a field! \n\nIn this case, the rabbit is similar to the plant it eats in the following way:\n\nBoth the rabbit and the plant are heterotrophs, meaning they are organisms that obtain their nutrients from other organisms. Rabbits are heterotrophic because they eat plants and other animals, whereas plants are autotrophic, meaning they produce their own food from the energy they absorb from the sun. \n\nIn other words, both the rabbit and the plant are consumers that obtain their energy from sources outside themselves, rather than producing their own food.

Table 6: Example on the amount of detail retrieved and length of the answer for pruned models.

Prompt: During which stage of its life cycle does a butterfly spend the most time eating plant leaves?

Answer (EN-calibrated): –

Answer (DE-calibrated): Butterflies, also known as Lepidoptera, typically spend the most time eating plant leaves during the larval stage, also known as the caterpillar stage. \n\nAs a caterpillar, the butterfly spends most of its time feeding on plant leaves, leaves, and other plant material to accumulate energy and nutrients. This feeding period is crucial for the caterpillar’s growth and development. \n\nThe larval stage is typically the longest stage of the butterfly’s life cycle, and during this stage, the caterpillar feeds on various plant parts, such as leaves, stems, and flowers, to grow and develop. \n\nAfter the caterpillar has reached its full size, it forms a pupa or chrysalis, and then emerges as an adult butterfly, ready to start the next generation of its life cycle. \n\nSo, to answer your question, the stage where a butterfly spends the most time eating plant leaves is the larval stage, which is the caterpillar stage.

Table 7: Example of one pruned model skipping a prompt answer.

Prompt: Write me a funny story about a cat and a dog in five sentences.

Answer (EN-calibrated): Here’s a funny story about a cat and a dog: Whiskers the cat and Rufus the dog were the best of frenemies. They loved to play together, but their playful rivalry was often interrupted by their owners’ constant laughter at their silly antics. One day, they decided to team up and steal the family’s prized treats from the kitchen. In their excitement, they knocked over a bag of flour, covering themselves in a white cloud. As they ran around the house, they looked like two little snowballs, laughing and chasing each other in circles.

Answer (DE-calibrated): One day, a curious cat named Whiskers decided to play a prank on her dog friend, a goofy dog named Rufus. Whiskers snuck into Rufus’s doghouse and replaced his dog toys with cat toys, much to Rufus’s surprise. Rufus was confused and frustrated when he realized he couldn’t find his favorite ball. Whiskers laughed and laughed, thinking it was the best prank she had ever played. Rufus, on the other hand, was left wondering what was going on and why his dog toys were suddenly all cat-related!

Prompt: Write me a sweet little poem of Alice and her teddy bear in one stanza.

Answer (EN-calibrated): Here is a sweet little poem:\n Alice holds her teddy bear close,\n Her best friend, her heart’s gentle rose,\n Together they sit, a perfect pair,\n A love that’s strong, without a care.

Answer (ZH-calibrated): Here is a sweet little poem about Alice and her teddy bear:\n Alice’s teddy bear, her dearest friend,\n Lies on her lap, a perfect end.\n His button eyes, so bright and wide,\n Shining with love, as she snuggles inside.\n In his arms, she finds her nest,\n A world of dreams, where love is best.\n

Table 8: Examples of language modeling capabilities of 50% unstructured SparseGPT-pruned models.

Prompt: Who is the greatest ballerina of all time? Explain why.

Answer (EN-calibrated): According to many ballet enthusiasts, the greatest ballerina of all time is considered to be Natalia Makarova.

Answer (DE-calibrated): According to many experts and dance historians, the greatest ballerina of all time is Maria Callas (1923–1969). She was a Russian-born American ballet dancer and singer, known for her unique and expressive technique, and her iconic performances in roles like La Scala’s ‘La Bohème’ and ‘Eugene Onegin’.

Table 9: Example of pruned models retrieving different facts (disregarding the actual truth).

H Supplementary Analysis

H.1 Further Language-Subspace Results

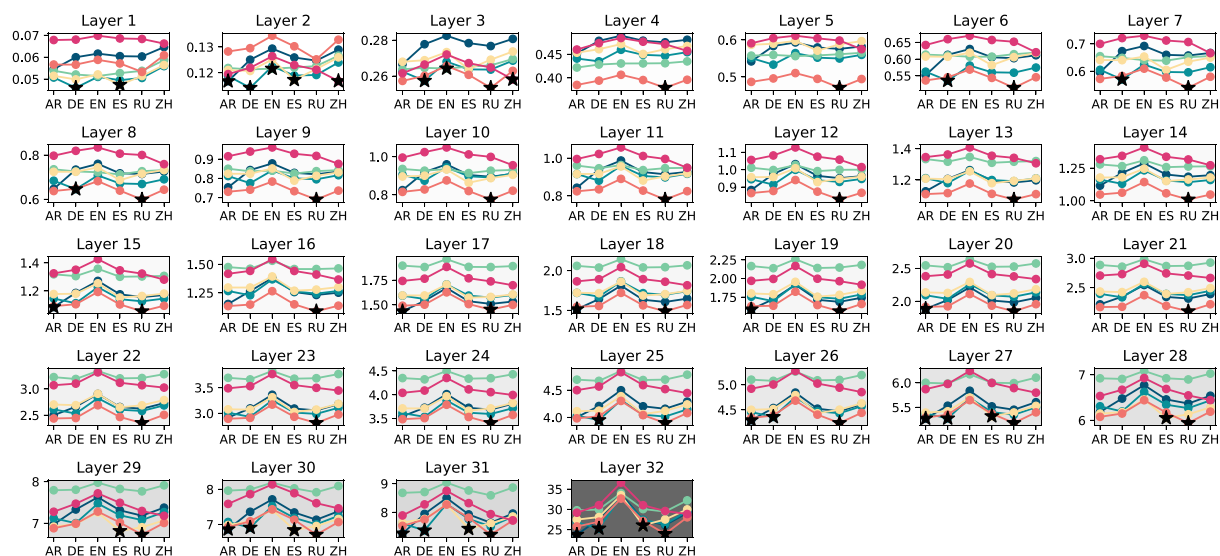


Figure 4: Language-wise mean magnitude of differences between the prompt-wise and layer-wise averaged **language-agnostic** features extracted with LSAR for a full-sized and 50% **SparseGPT**-pruned and **mC4-calibrated Llama 3 8b** model. Both, the LSAR projection matrix and the feature differences, were computed over 900 prompts from the **Belebele** dataset for the six calibration/test languages. The evaluation languages are shown on the x -axis, the calibration languages are color-coded (AR, DE, EN, ES, RU, ZH). The background color indicates the magnitude of the maximum deviation. A star marks the case where using the same language for calibration and evaluation results in the smallest difference after pruning.

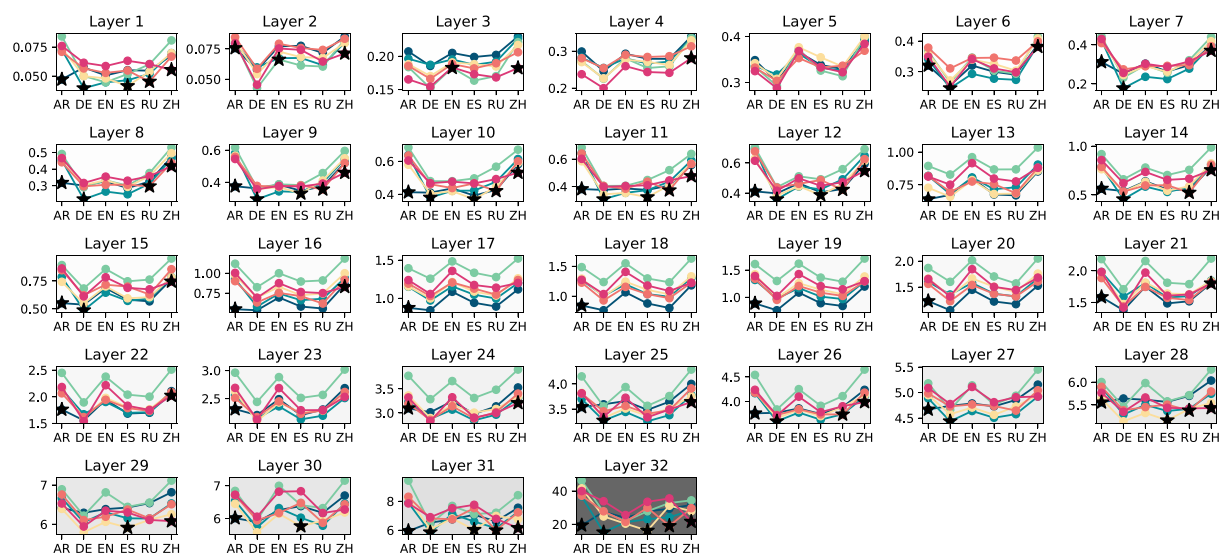


Figure 5: Language-wise mean magnitude of differences between the prompt-wise and layer-wise averaged **language-specific** features extracted with LSAR for a full-sized and 50% **SparseGPT**-pruned and **mC4-calibrated Llama 3 8b** model. Both, the LSAR projection matrix and the feature differences, were computed over 900 prompts from the **Belebele** dataset for the six calibration/test languages. The evaluation languages are shown on the x -axis, the calibration languages are color-coded (AR, DE, EN, ES, RU, ZH). The background color indicates the magnitude of the maximum deviation. A star marks the case where using the same language for calibration and evaluation results in the smallest difference after pruning.

H.2 Further Pruning Mask Similarity Results

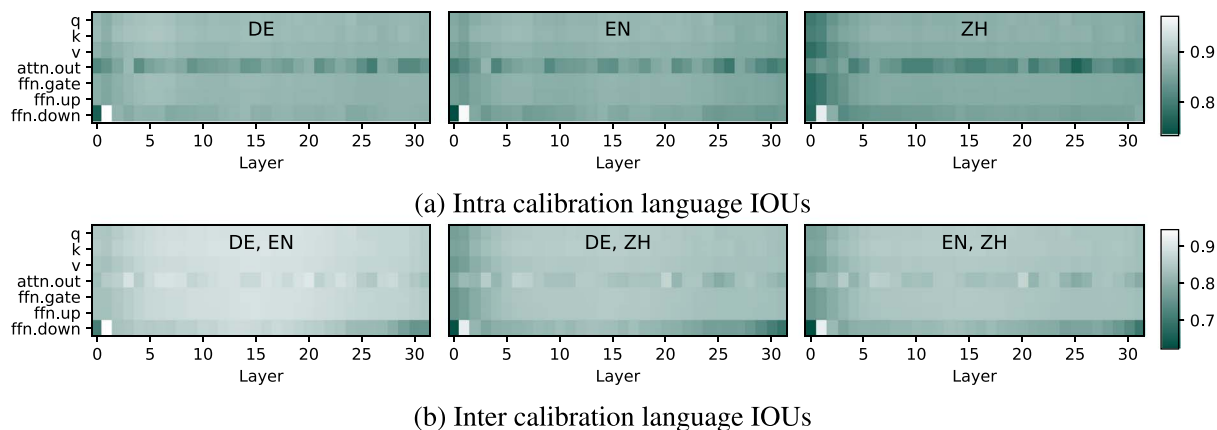


Figure 6: Pruning mask similarities (IoU) between EN, DE, and ZH calibrated on **mC4** and 50% sparsity **SparseGPT**-pruned **Aya 23 8B** models. 6a IoU of pruning masks for three calibration sets of the same language. 6b IoU between pruning masks for different calibration languages. The higher IoU (indicated as a lighter color), the more similar pruning masks between different calibration languages.

H.3 Further Language Entropy Results

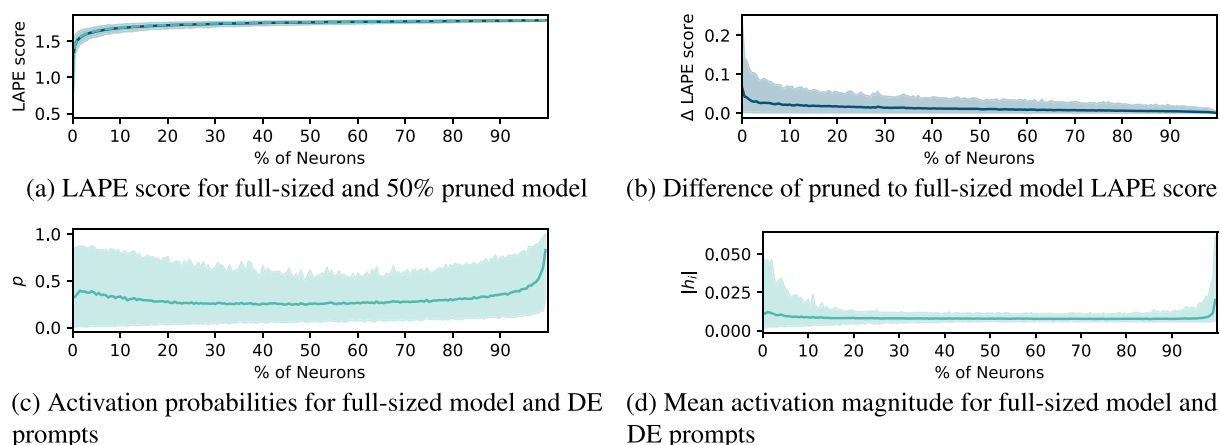


Figure 7: Statistics for FFN neurons of a **full-sized Aya 23 8B model** and its **50% SparseGPT-pruned** version calibrated for DE on **mC4**. Neurons are ordered by ascending LAPE score of the full-sized model as shown in 7a. Additionally, LAPE score and activation probabilities get correlated by removing all neurons with an activation probability in DE that is less than the average activation probability among all languages. *The lower the LAPE score, the more specialized the neuron is for a particular language.*

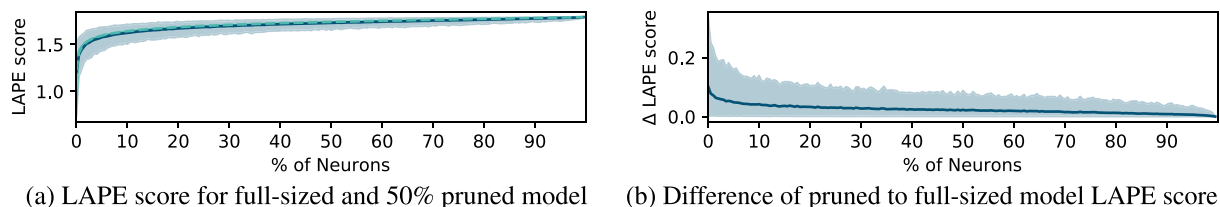


Figure 8: Statistics for FFN neurons of a **full-sized Llama 3 8B model** and its **50% Wanda-pruned** version calibrated for DE on **mC4**. Neurons are ordered by ascending LAPE score of the full-sized model (8a). LAPE score and activation probabilities get correlated by removing all neurons with an activation probability in DE that is less than the average activation probability among all languages.