# **Multilingual Brain Surgeon:** Large Language Models Can be Compressed Leaving No Language Behind

Hongchuan ZENG, Hongshen Xu, Lu Chen, Kai Yu

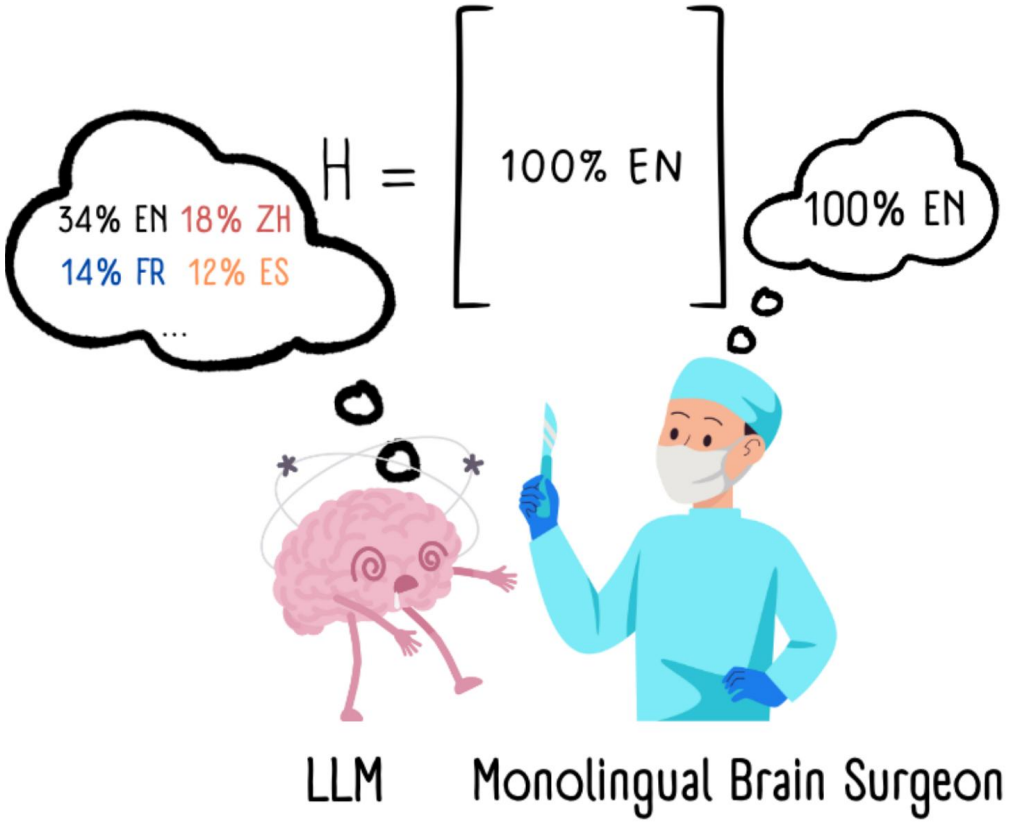
{charlie68, xuhongshen, chenlusz, kai.yu}@sjtu.edu.cn

# Introduction

- Large size and computational demands of Large Language Models(LLMs) necessitate effective **Model Compression** (MC) techniques

- SOTA MC methods typically rely on a **calibration set** that overlooks the multilingual context

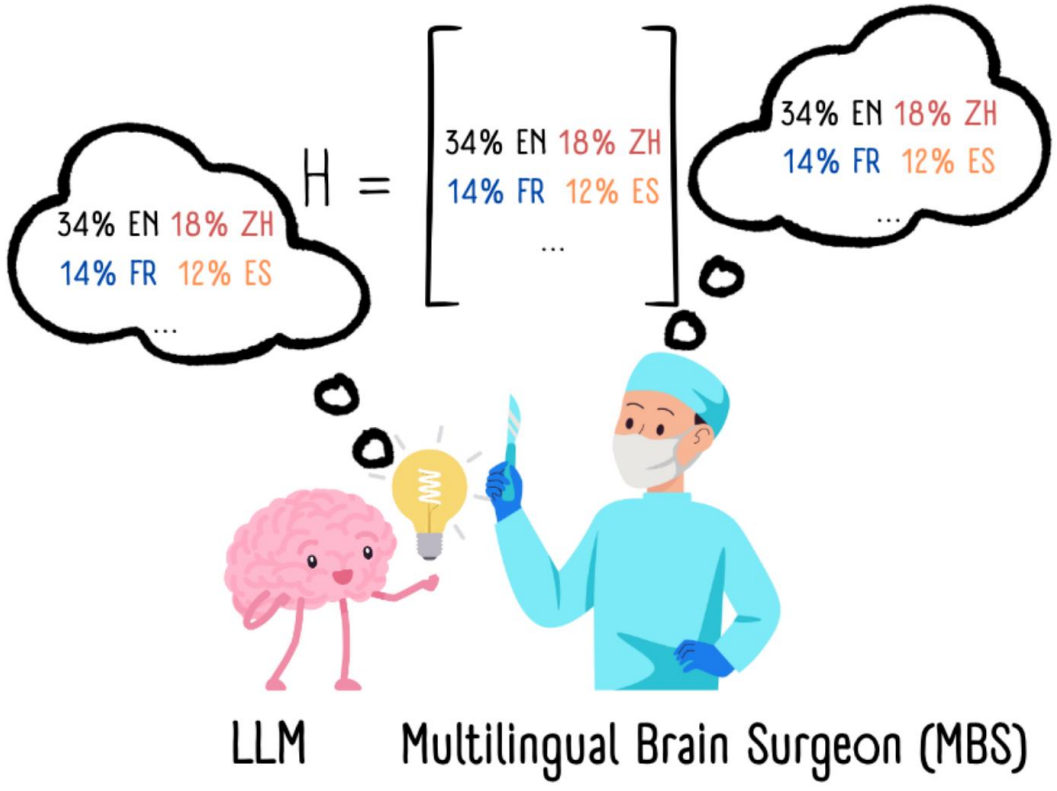
- **English-centric calibration** leads to performance degradation in multilingual models, especially in low-resource languages.



# Introduction

- **Multilingual Brain Surgeon(MBS)**: Sample calibration data proportionally to language distribution in training datasets, retaining the performance of low-resource languages.

- **Dynamics of language interaction during compression:**
- The larger the proportion of a language in the training set and the more similar the language is to the calibration language, the better performance the language retains after compression.



# Background

- Optimal Brain Surgeon (OBS): A network pruning framework(Hassibi et al., 1993)
  - Assume that a network's error converges to a local minima.
  - Calculate the second-order derivatives (Hessian matrix H) of the error (E) with respect to each parameter (w) to determine which connections can be safely pruned without significantly affecting performance.
  - The increase in error ($L_j$) when a parameter ($w_j$) is set to zero: $$L_j = \frac{1}{2}\frac{{w_j}^2}{[\mathbf{H}^{-1}]_{jj}}$$
  - The optimal adjustment (δw) of the remaining weights to compensate for the removal are given by: $$\delta w = -\frac{w_j}{[\mathbf{H}^{-1}]_{jj}}\mathbf{H}^{-1}_{:,j}$$
  - SparseGPT, Wanda and GPTQ are model compression methods based on OBS.
- Error Measurement
  - Given inputs X (the training dataset), the original weights W, the updated weights $\hat{\mathbf{W}}$ , and a sparsity mask M of the same size as W, the error is defined as:

$$E = ||\mathbf{WX} - (\mathbf{M} \odot \hat{\mathbf{W}})\mathbf{X}||_2^2$$

# What happens when calibration set is monilingual?

- Totol error of model : $E$
- Error on language $\mathrm{m}$ : $E_m$
- Total number of languages: $\mathrm{N}$

$$\mathrm{E} = E_1 + E_2 + E_3 + \ldots + E_N$$

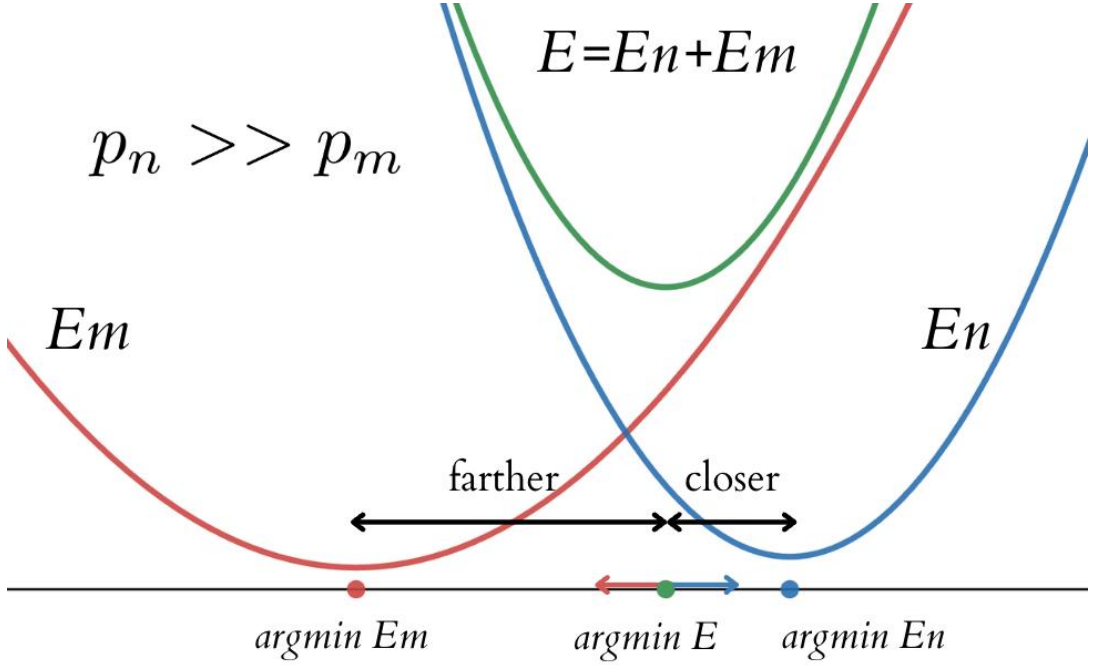
- Model trained to convergence $\Rightarrow$ $\mathrm{E}$ resides in a local minimum

- Factor 1: Proportion in training data
- Factor 2: Similarity between languages

# Factor 1: Proportion in training data

- We consider two languages in model: m and n
- Propotion of language m, n in training set: $p_m$, $p_n$
- If the proportion of *n* in training set is larger than *m* (i.e. $p_n >> p_m$), language *n* has a greater power to influence total error *E* than language *m*.
- ⇒ The local minima of E is closer to the local minima of $E_n$ than to the local minima of $E_m$
- Compressing with calibration data of language *n* "pushes" the minina of *E* towards $E_n$

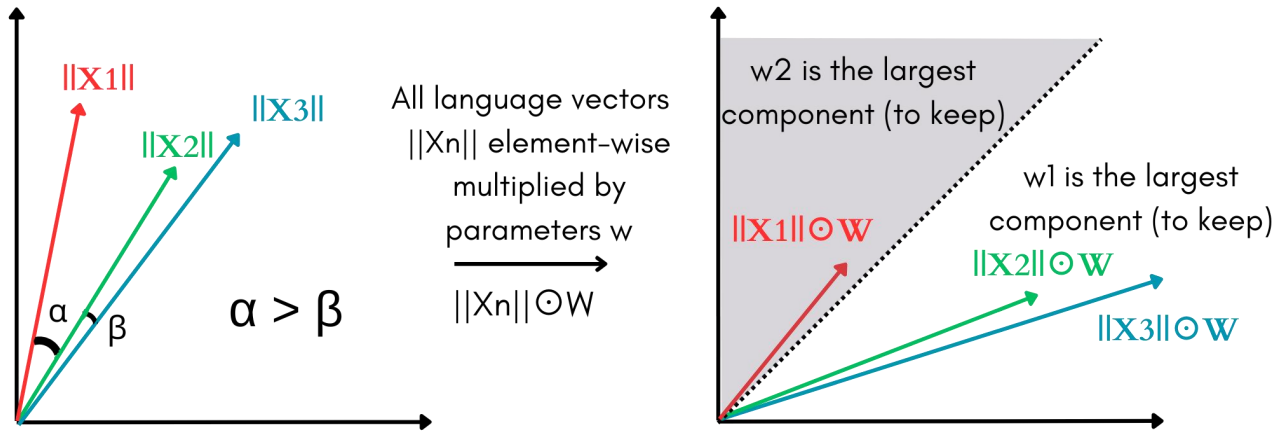
⇒ When compressing models with only the calibration data of higher-resource language n, it has a significant impact on the performance of lower-resource language m (as it push the model even further away from the local minima of $E_m$).
However, when compressing models with only the calibration data of lower-resource language m, it does not impact much the performance of higher-resource language n (as the model is still close to the local minima of $E_n$ even though pushed).



# Factor 2: Similarity between languages

- Optimal Brain Surgeon (OBS) tells us that the priority of compression is fully determined by H.
- We may suppose the non-diagonal elements are trivial (Le Cun et al., 1989) to calculate the inverse of H.
  - ⇒ The metric is thus simplified to $S = |W| \cdot ||X||_2$ ($X$ represents the training data for language n)
  - ⇒ Use the cosine similarity between $||X||_2$ as the similarity metrics between languages.
- Why cosine similarity?
  - ⇒ We need to compare two vectors based on the likelihood that their largest components remain consistent after undergoing the same element wise multiplication with unknown vectors (model parameters)

When $||X_m||_2$ and $||X_n||_2$ are similar, using only data of language *m* as calibration data will introduce little performance drop in language *n*, and vice versa.
That is to say, when two languages are very different, employing data from just one of the two languages as calibration data will lead to a significant performance decrease in the other.



# Multilingual Brain Surgeon (MBS)

$$\mathrm{E} = E_1 + E_2 + E_3 + \ldots + E_N$$

The Hessian matrix of *E* :

$$\mathrm{H} = H_1 + H_2 + H_3 + \ldots + H_N \qquad \text{where } \mathbf{H_n} = \mathbf{X_n}^{\mathbf{T}}\mathbf{X_n}$$

$X_n$ represents the inputs (training data) for language n, with a shape of q × $p_n$, where q is the total number of network parameters, and $p_n$ is the total number of training samples for language n.

Let's denote a subset of training data as $X_n^{[k]}$. We have : $\mathbf{H_n} = \mathbf{X_n}^{\mathbf{T}}\mathbf{X_n} = \sum_{k=1}^{p_n} X_n^{[k]^T} X_n^{[k]}$

which leads to:

$$\mathbf{H} = \sum_{k=1}^{p_1} X_1^{[k]^T} X_1^{[k]} + \sum_{k=1}^{p_2} X_2^{[k]^T} X_2^{[k]} + \ldots + \sum_{k=1}^{p_n} X_n^{[k]^T} X_n^{[k]}.$$

When selecting calibration data, it's essential to choose samples from each language in proportion to its presence in the training set.

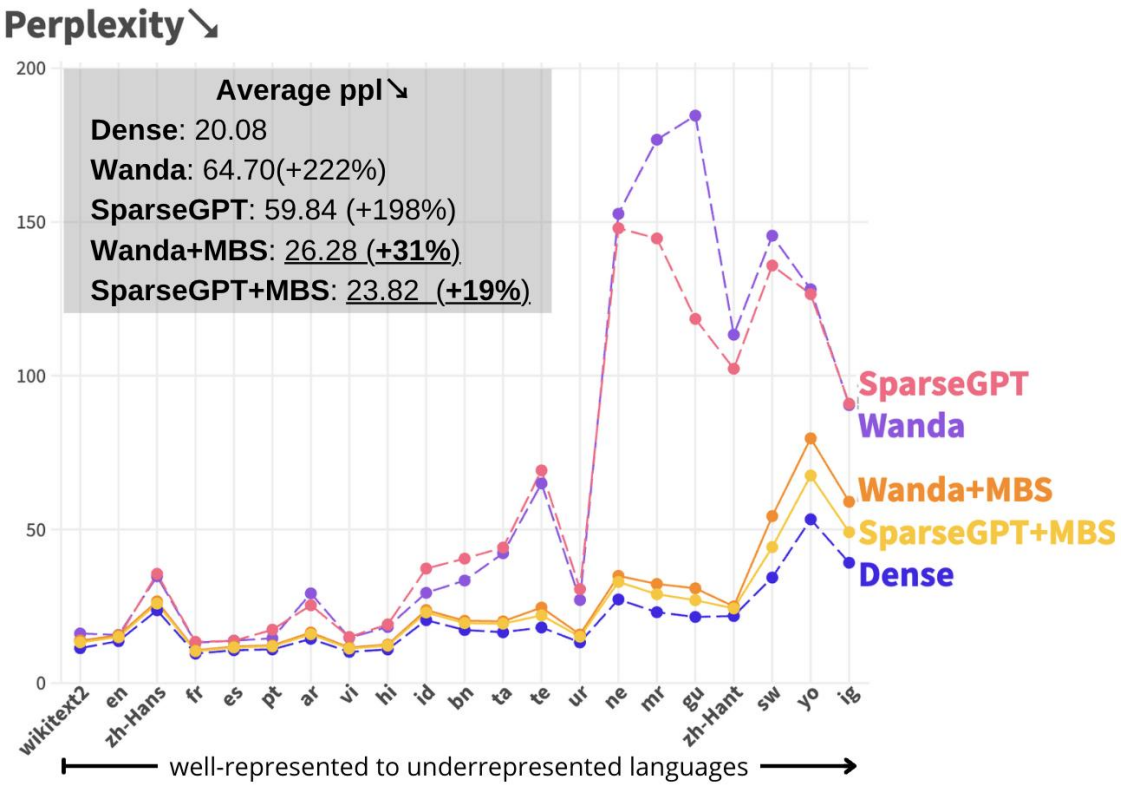
E.g. 50% English, 30% Chinese, 20% French in training set
⇒ 50% English, 30% Chinese, 20% French in calibration dataset

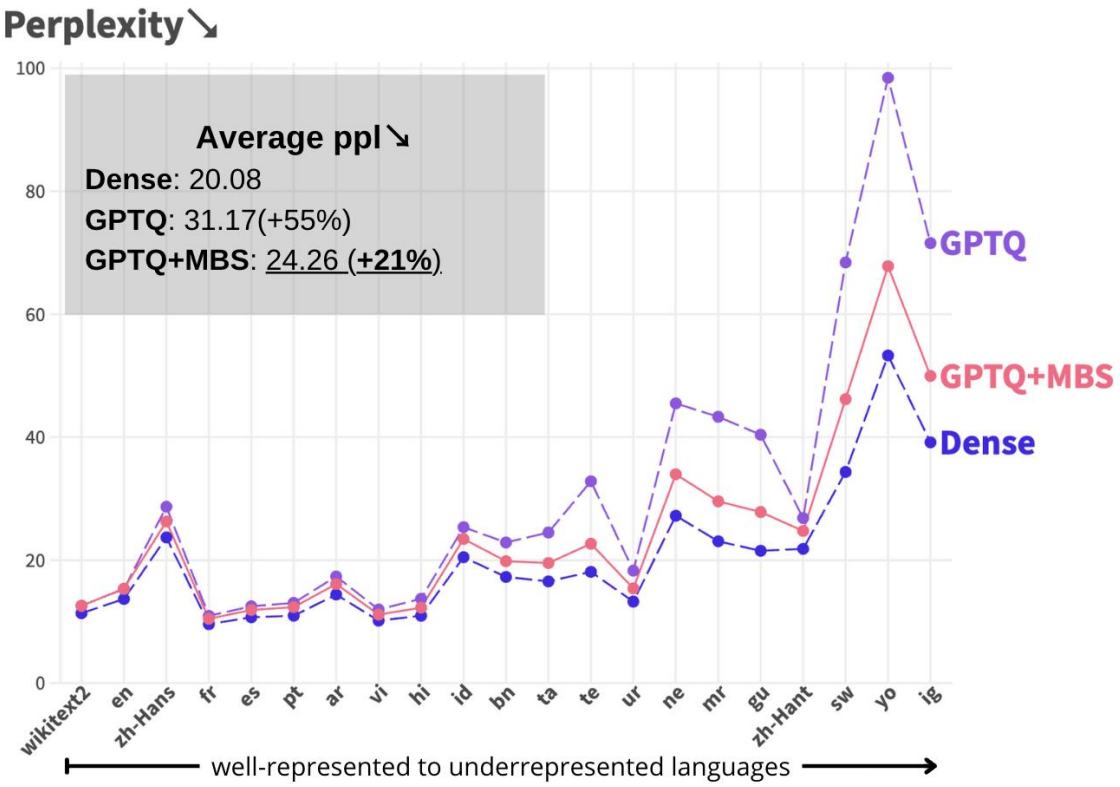
# Experimental Setup

- Models : BLOOM-560m and BLOOM-7b1.
- Datasets: CC-100 for calibration, XL-Sum for perplexity measurement.
- Evaluation: perplexity and zero-shot tasks(EleutherAI eval-harness framework).
- Language Case Study: performed monolingual compression using English(high-resource), Igbo(low-resource), Urdu(most similar languages to the others) and Tamil(least similar languages to the others).

| Language | Size in Bytes in BLOOM training data | MBS sampling | Equal sampling |
|---|---|---|---|
| en | 4.85E+11 | 87 | 13 |
| zh-Hans | 2.61E+11 | 47 | 13 |
| fr | 2.08E+11 | 37 | 13 |
| es | 1.75E+11 | 31 | 13 |
| pt | 7.93E+10 | 14 | 13 |
| ar | 7.49E+10 | 13 | 13 |
| vi | 4.37E+10 | 7 | 13 |
| hi | 2.46E+10 | 4 | 13 |
| id | 2.00E+10 | 3 | 13 |
| bn | 1.86E+10 | 3 | 13 |
| ta | 7.99E+09 | 1 | 13 |
| te | 2.99E+09 | 1 | 13 |
| ur | 2.78E+09 | 1 | 13 |
| ne | 2.55E+09 | 1 | 13 |
| mr | 1.78E+09 | 1 | 13 |
| gu | 1.20E+09 | 1 | 13 |
| zh-Hant | 7.62E+08 | 1 | 12 |
| sw | 2.36E+08 | 1 | 12 |
| yo | 8.97E+07 | 1 | 12 |
| ig | 1.41E+07 | 1 | 12 |

# Results of MBS

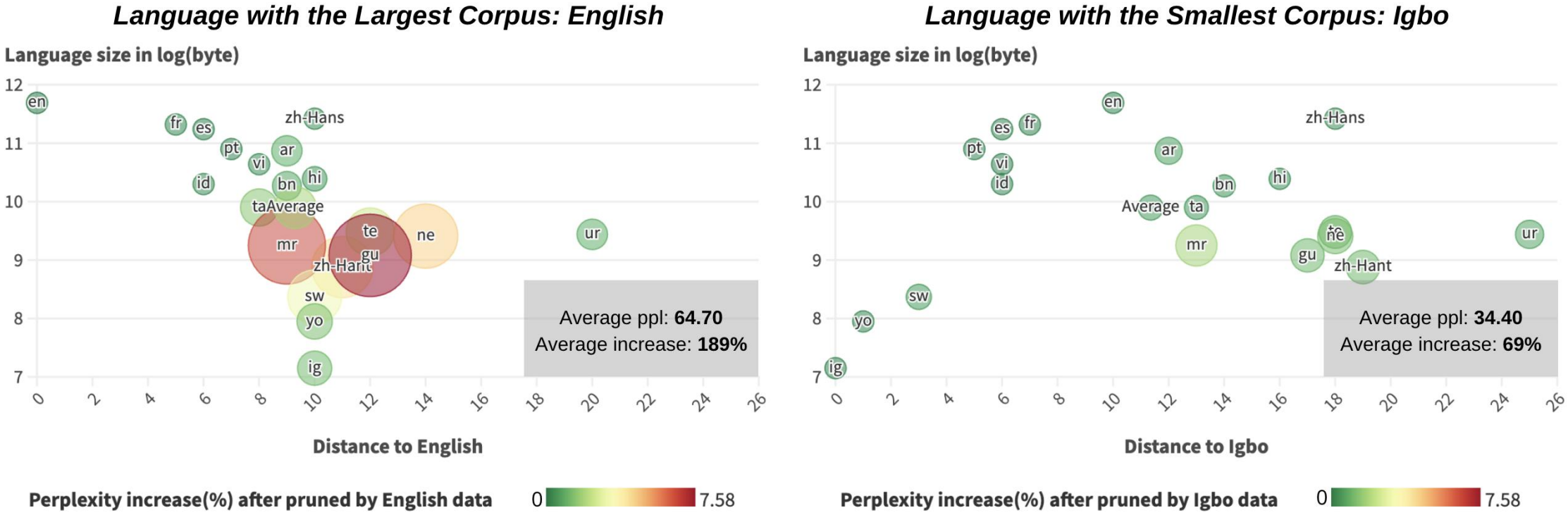



| Accuracy of 0-shot task | Dense | Wanda | Wanda +MBS | SparseGPT | SparseGPT +MBS | GPTQ | GPTQ +MBS |
|---|---|---|---|---|---|---|---|
| Average↑ | 57.63% | 55.36% | ***55.49%*** | 55.38% | ***56.13%*** | 55.59% | ***57.08%*** |

# Results of monolingual compression

- **Factor 1: Proportion in training data**

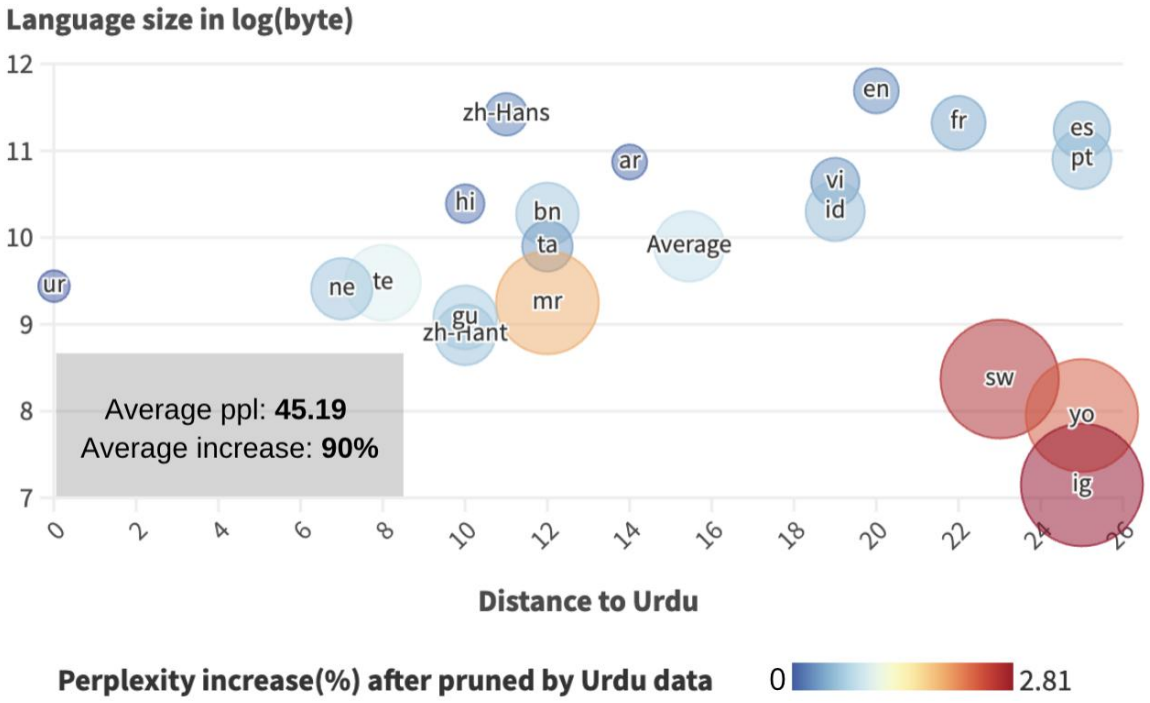



Monolingual pruning results using Wanda with calibration data in English or Igbo. The size of each bubble corresponds to the magnitude of the increase in perplexity for the model in that particular language, while the vertical axis represents the size of training data in log(bytes) from the language in the training set of BLOOM. Using only a language with higher proportion in the training set as calibration data has a greater impact on model's performance.

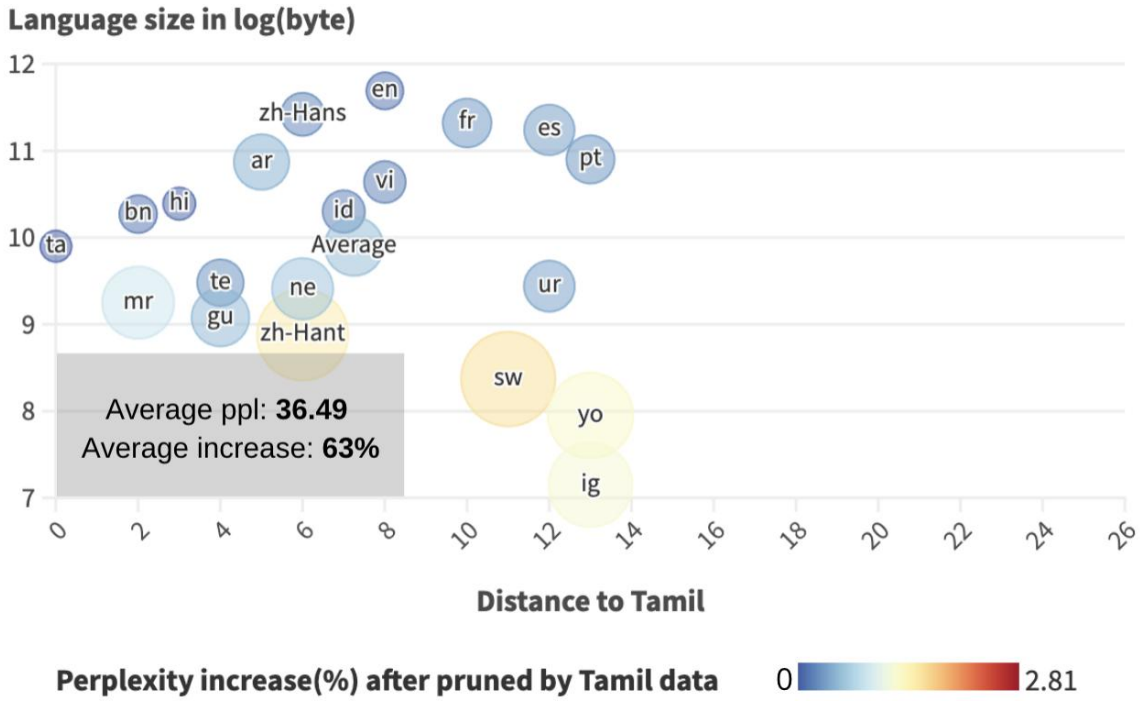
# Results of monolingual pruning

- **Factor 2: Similarity between languages**

***Least Similar Language to the Others: Urdu***




***Most Similar Language to the Others: Tamil***



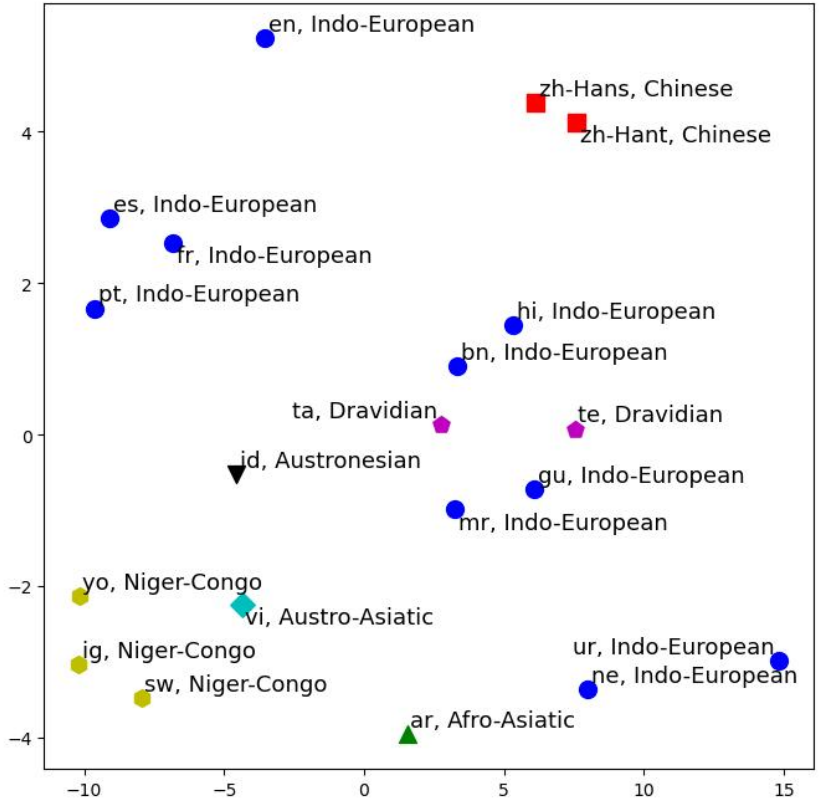

The languages less similar to the calibration language experience a greater increase in perplexity.

# Results of monolingual pruning

- **Factor 2: Similarity between languages**
  - Distance map of different languages associated with their corresponding language families. We can see that languages with the same family cluster together from this map.

Possible reason of clustering:

• Shared Grammar Structure: Languages within the same language family often share similar grammar structures.
• Shared Tokens: During the tokenization process, these languages frequently share tokens, including prefixes, suffixes, and other word building elements.



# Conclusion

- When performing model compression, we should sample the calibration data proportionally to language distribution in training datasets.

- Our experiments on the BLOOM model highlight the effectiveness of MBS, benefiting pruning and quantization methods like SparseGPT, Wanda, and GPTQ.

- The larger the proportion of a language in the training set and the more similar the language is to the calibration language, the better performance the language retains after compression.