Multilingual Performance Analysis of Large Language Models

Anonymous ACL submission

Abstract

The training process of Large Language Models (LLMs) requires extensive text corpus. However, these data are often unevenly distributed in different languages. As a result, LLMs perform well on common languages, such as English, German, and French, but perform poorly on low-resource languages. However, currently, there is no work to quantitatively measure the performance of LLMs in low-resource languages. To fill this gap, we proposed the Language Ranker that aims to benchmark and rank different languages according to the performance of LLMs on those languages. We employ the LLM's performance on the English corpus as a baseline to compare the performances of different languages and English. We have the following three findings: 1. The performance rankings of different LLMs in all languages are roughly the same. 2. LLMs with different sizes have the same partial order of performance. 3. There is a strong correlation between LlaMa2's performance in different languages and the proportion of the pre-training corpus. These findings illustrate that the Language Ranker can be used as an indicator to measure the performance of LLMs with different languages.

1 Introduction

Large Language Models (LLMs), such as ChatGPT and GPT-4, have demonstrated surprising performance in various NLP tasks (Achiam et al., 2023; Ouyang et al., 2022; Touvron et al., 2023; Team et al., 2024; Jiang et al., 2023; Bai et al., 2023). However, the majority of the text datasets are presented in high-resource languages such as English (Xie et al., 2024). According to the statistics, for GPT-3 model approximately 92.65% of the training tokens are English and all other languages share the remaining 7.35% training tokens (OpenAI, 2023). Similarly, English accounts for 89.70% of data for pre-training LlaMa 2 (Touvron et al., 2023). Thus,

this imbalanced token distribution will cause bias towards English (Blasi et al., 2021). As a result, the excellent performance of LLM is often limited to some common languages, such as English.

This imbalanced distribution makes the LLM less capable of understanding low-resource languages. For example, LLM cannot understand the true meaning of some slang terms with specific cultural backgrounds, such as Chinese idioms (Zhang et al., 2023). Moreover, recent research has shown that the pre-trained model often underperforms in language with insufficient training data (Lankford et al., 2024). Above phenomena illustrate the importance of training data for LLM. However, it is often not released by leading companies and it does not take the inner representations of LLMs into account. Therefore, it is necessary to propose a metric to measure different language proportions in the LLM's pre-training corpus and further implicitly measure the language ability for different languages, especially low-source languages.

In this paper, we propose to utilize internal representations to quantitatively measure the multilingual abilities of LLMs. Specifically, we employ the representation of LLMs on the English corpus as the baseline. Then, we measure the similarity between the representations on the corpus of low-resource languages and those of English. We take this similarity value as the performance score of the model in each language. In experiments, we found that the ranking results obtained by our method are roughly the same as the ranking results of different language proportions in the LLM's pre-training corpus. It shows that our proposed method can effectively measure the performance of LLMs in different languages.

2 Analysis Method

In this section, we will give an introduction to our analysis method. First, we will introduce the

1

dataset we used in our experiment. Then, we will introduce how to obtain the similarity between English and other languages as well as how to compare different LLMs' performances.

2.1 Probing Datasets

We use OPUS-100 (Zhang et al., 2020) as our evaluation datasets. OPUS-100 is an English-centric multilingual corpus that covers 100 languages. Each sample consists of text in a non-English language as the original data, with its English translation serving as the target data. For example, {"German": "Ich wollte dir erst noch etwas zeigen.","English": "I wanted to show you something first."}. After filtering, there are 94 subsets containing English, including high-resource languages such as German, French, and Chinese, as well as low-resource languages such as Oriya, Kannada, and Kazakh. Each subset contains 2000 samples.

2.2 Similarity Measurement

We employ cosine similarity to measure the LLMs' performance gap between the target language and English. Specifically, given two sentences $X = \{x_i\}_{i=1}^{n}$ and $Y = \{y_i\}_{i=1}^{m}$ representing the text in English and the text in the target language. We use the representation obtained after LLM mapping of the last token x_n and y_m as the representation of the text and calculate the similarity between them. As we know, LLM consists of several layers of a Transformer block (Vaswani et al., 2017). Therefore, after each layer of mapping by the transformer block, we can get a representation vector x_n^l and y_m^l, $l = 1...H$, where H represents the number of the layer of LLMs. According to (Li et al., 2024), the intermediate representation can be briefly summarized by the following equations:

$$x^{l+1} = \text{MLP}(x^l + \text{MHA}(x^l)) \quad l = 1...H, \quad (1)$$

where MHA means multi-head attention or multi-group attention, and MLP means standard multi-layer perceptron layer. Next, we take x_n^l and y_m^l to calculate the similarity. To implement a more robust similiarity measure, we use the average similarity obtained by several intermediate layers as the final similarity. This process can be described as follows:

$$Sim = \frac{1}{|l_{sub}|} \sum_{i=1}^{|l_{sub}|} Sim_i, \text{ where } Sim_i = \frac{x_n^i y_m^i}{||x_n^i|| ||y_m^i||}, \quad (2)$$

where $l_{sub} = \{5, 10, 15, 20, 25\}$ is the subset of the layers we selected. Finally, we use Sim to evaluate the performance gap between English and Non-English corpus.

2.3 Rank Correlation Measurement

When we get the similarity between each non-English representation and the English representation, we sort them according to the similarity to get a sorted ranking list of all languages. To measure the similarity of the sorted ranking lists of two LLMs, we use the longest common partial order sublist to measure. It can be defined as follows: For two sorted lists A and B, find a sublist C which is a subset of A and B such that for any number of index $i_1 \leq i_2 \leq ... \leq i_n$, Index($C_{i_1}$)≤Index($C_{i_2}$)≤...≤Index($C_{i_n}$) is true for both A and B, and the longest sublist C that makes it true is called the longest common partial order sublist of A and B. We use the ratio of the length of the longest common partial order sublist of two LLMs to the total length of the ranking list as a metric to measure the correlation.

3 Experiments

3.1 Open-source Models

We use four popular open-source large models as our analysis baselines: **LlaMa2** (Touvron et al., 2023), **Qwen** (Bai et al., 2023), **Mistral-v0.1** (Jiang et al., 2023), and **Gemma** (Team et al., 2024). In Section 3.2, we concentrate on the 7B version of these models. The performance of models of various sizes will be discussed in Section 3.3.

3.2 Comparison of Different Models

To visualize the performance of different LLMs in these languages, we selected 10 representative languages to display their inference results. They consist of five high-resource languages, including German, Spanish, French, Indonesian, and Chinese, and five low-resource languages, including Igbo, Kazakh, Kannada, Oriya, and Turkmen. Figure 1 shows detailed results, where the X-axis represents different layers of LLMs, while the Y-axis represents the similarity between the target language and English for each layer. From Figure 1, we can derive the following key observations:

(1) **High-resource languages have representations more similar to English, whereas low-resource languages show less similarity.** Although the exact proportion of high-resource and
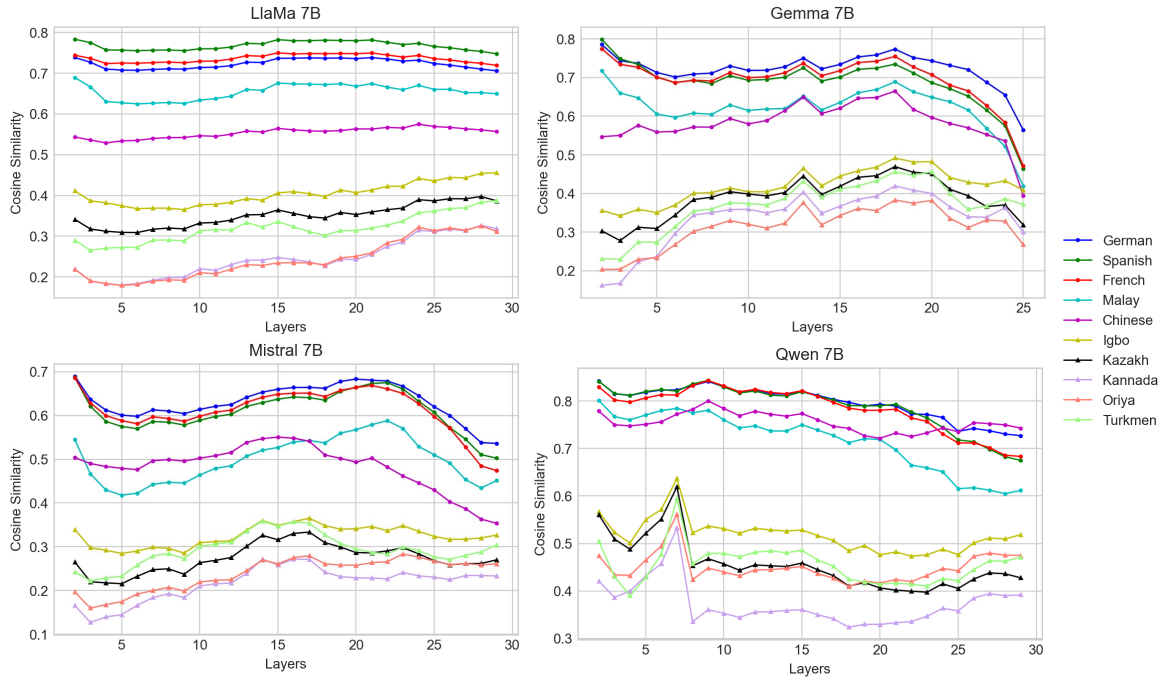
Figure 1: Performance of different LLMs for ten kinds of language, German, Spanish, French, Indonesian and Chinese are five high-source languages; Igbo, Kazakh, Kannada, Oriya and Turkmen are five low-source languages.
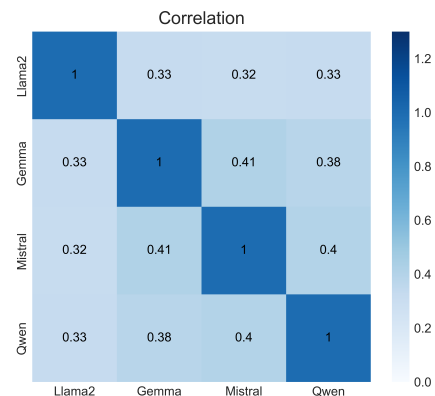

Figure 2: Rank correlation between different LLMs

low-resource languages in each LLM's pre-training corpus is unknown, high-resource languages are generally more prevalent, and the results in the figure show that our similarity-based measurement method can effectively measure the proportion of each language in the LLM's pre-training corpus.

(2) **Different models display similar results across languages**. Their performance is that the high-resource language similarity is higher than the low-resource language similarity. Figure 2 further illustrates this conclusion, we can see from the fig-

ure that for each LLM, the ranking result is used as the baseline, and the remaining three LLMs are roughly similar to the baseline.

(3) **Fine-tuning on specific languages will improve its performance**. From the result of Qwen, we can observe that the performance of the Chinese improves as the number of layers increases. In the last few layers, it surpasses other high-resource languages in the figure. According to the technical report of Qwen(Bai et al., 2023), Qwen has additional fine-tuning on the Chinese corpus, which leads to better performance in Chinese.

3.3 Comparison of LLMs of Different Sizes

We also conducted analytical experiments on the same model of different sizes. The result is shown in Figure 3. We found that the results of low-resource languages fluctuated greatly, so we defined the layer depth as dividing the interval [0,1] equally by the number of layers and selected a specific layer depth interval [0.4,0.6] to display the results of low-resource languages. We can observe two phenomena:

(1) There is a modest positive correlation between the size of an LLM and its performance on low-resource languages. As shown in the figure, for Kannada, Occitan, and Western Frisian, the perfor-
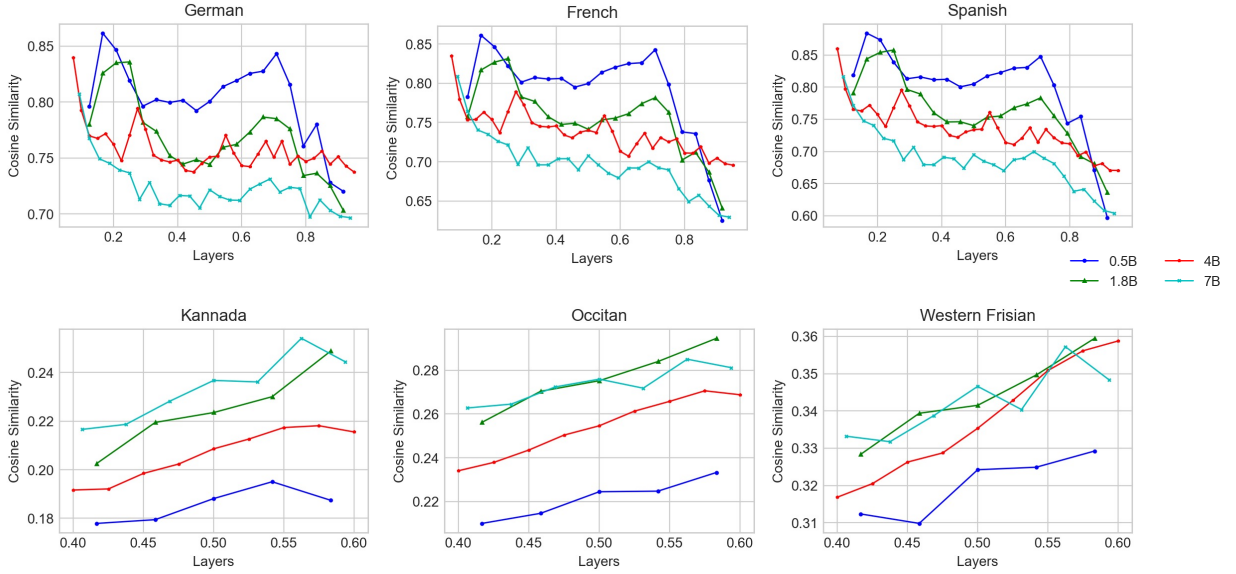
3

Figure 3: The performance of Qwen 1.5 (0.5B, 1.8B, 4B, and 7B) in various languages. German, French, and Spanish are high-resource languages, Kannada, Occitan, and Western Frisian are low-resource languages. For low-resource languages, to make the results clearer, we selected the intermediate layer representation (layers depth 0.4-0.6) results that change relatively smoothly.

mance of Qwen 1.5 on three sizes of 0.5B, 4B, and 7B gradually improves as the size increases.

(2) For high-resource languages, there is a strong negative correlation between the size of the LLMs and the performance of high-resource languages.

In the figure 3, high-resource languages and low-resource languages show completely opposite trends. The possible reason for this phenomenon is that as the size of the LLM increases, the complexity of the high-resource training corpus also increases, leading to interference from lower-quality data. On the contrary, low-resource language corpus is relatively scarce, so the feature distribution is relatively uniform. In such cases, the size of the model is positively correlated with the performance in the language.

3.4 Relationship to Pre-training Corpus

According to the technical report of LlaMa2 (Touvron et al., 2023), we get the proportion of the pre-training corpus of some languages. Table 1 shows the relationship between the proportion of pre-training corpus for some languages and the similarity metric. We can observe that from French to Swedish to Finnish to Norwegian, as the pro-

portion of corpus decreases, the similarity metric also decreases. It does not hold for all languages, because it is not only the proportion that affects the performance of LLM in a certain language but also factors such as the grammatical similarity between the language and English.

Language	Proportion	Similarity	Language	Proportion	Similarity
German	0.17%	0.581	Polish	0.09%	0.534
French	0.16%	0.591	Vietnamese	0.08%	0.529
Swedish	0.15%	0.531	Finnish	0.03%	0.516
Chinese	0.13%	0.446	Norwegian	0.03%	0.501

Table 1: The proportion of different languages in the LlaMa2 pre-training corpus and the similarity metric we proposed. The English language ratio is 89.7%.

4   Conclusions and Future Work

In this work, we propose a similarity-based evaluation method to measure the LLMs' performance in various languages quantitatively. The results show that this similarity metric has a clear correlation with the proportion of each language in the pre-training corpus, and can roughly measure the performance ability of the model in each language. In the future, we plan to design more detailed evaluation criteria to measure LLM's capabilities in each language.

Limitations

The proposed Language Ranker approach provides an initial quantitative way to analyze LLM performance across languages. However, it has several limitations. First, the similarity metric based on English representations may not fully capture the nuances and complexities of each language's linguistic properties. Additionally, low-resource languages are likely to exhibit more noise and variance in the similarity scores due to the smaller dataset sizes used for pre-training these languages in LLMs. Furthermore, the method does not account for potential biases or skews that could be present in the multilingual evaluation datasets themselves. The existence of such biases can also introduce noise in the resulting rankings of language abilities for different LLMs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Seamus Lankford, Haithem Afli, and Andy Way. 2024. Transformers for low-resource languages: Is f\'eidir linn! *arXiv preprint arXiv:2403.01985*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

OpenAI. 2023. Gpt-3 dataset statistics. https://github.com/openai/gpt-3/tree/master/dataset_statistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. 2024. Language imbalance can boost cross-lingual generalisation. *arXiv preprint arXiv:2404.07982*.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yangchen Xie, Xinyuan Chen, Hongjian Zhan, Palaiahnakote Shivakumara, Bing Yin, Cong Liu, and Yue Lu. 2024. Weakly supervised scene text generation for low-resource languages. *Expert Systems with Applications*, 237:121622.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.

A Appendix

A.1 Related Work

Multilingual Language Model. The imbalance distribution of training corpus in different languages leads to the bias of LLM towards some common languages such as English (Blasi et al., 2021). Some approaches employ multilingual language modeling to alleviate the phenomenon (Shen et al., 2024; Kalyan et al., 2021; Conneau et al., 2019). These studies show the importance of strengthening the cross-lingual capabilities of the pretrained model. (Schäfer et al., 2024) found that the presence of a primary language in the training process of LLMs can improve the performance of low-resource languages and lead to a more consistent representation of LLMs in different languages. (Liu et al., 2024) found that for English-centric LLMs, although translation into English helps improve the performance of NLP tasks, it is not the best choice for all situations.

A.2 Ranking Result For LLMs

We give the similarity scores of the four LLMs used in the experiment on 18 common languages. Results are shown in following tables.

Language	Similarity Score	Language	Similarity Score
German	0.581	Polish	0.534
French	0.592	Portuguese	0.598
Swedish	0.531	Vietnamese	0.529
Chinese	0.446	Ukrainian	0.551
Spanish	0.616	Korean	0.199
Russian	0.589	Catalan	0.582
Dutch	0.569	Serbian	0.555
Italian	0.567	Indonesian	0.577
Japanese	0.194	Czech	0.587

Table 2: The similarity score of LlaMa2 7B.

Language	Similarity Score	Language	Similarity Score
German	0.571	Polish	0.487
French	0.546	Portuguese	0.535
Swedish	0.494	Vietnamese	0.456
Chinese	0.471	Ukrainian	0.484
Spanish	0.537	Korean	0.338
Russian	0.531	Catalan	0.492
Dutch	0.516	Serbian	0.472
Italian	0.522	Indonesian	0.499
Japanese	0.328	Czech	0.512

Table 3: The similarity score of Gemma 7B.

A.3 Performance Comparison of Qwen1.5 of Different Sizes

The following two figures show the performance of Qwen1.5 in different size for different languages.

Language	Similarity Score	Language	Similarity Score
German	0.516	Polish	0.421
French	0.503	Portuguese	0.482
Swedish	0.435	Vietnamese	0.392
Chinese	0.399	Ukrainian	0.437
Spanish	0.499	Korean	0.261
Russian	0.494	Catalan	0.445
Dutch	0.460	Serbian	0.408
Italian	0.466	Indonesian	0.421
Japanese	0.248	Czech	0.459

Table 4: The similarity score of Mistral 7B.

Language	Similarity Score	Language	Similarity Score
German	0.642	Polish	0.596
French	0.634	Portuguese	0.625
Swedish	0.603	Vietnamese	0.584
Chinese	0.608	Ukrainian	0.597
Spanish	0.638	Korean	0.481
Russian	0.634	Catalan	0.601
Dutch	0.612	Serbian	0.588
Italian	0.615	Indonesian	0.597
Japanese	0.457	Czech	0.611

Table 5: The similarity score of Qwen 7B.

Figure 4 shows the result of five high-resource languages. Figure 5 shows the result of five low-resource languages. From figure 4, we observe that the performance of 0.5B is the best, while 7B performs the worst. Figure 5 shows the opposite result. It also can be found that the performance variance in low-resource languages is much greater than the performance in high-resource languages.
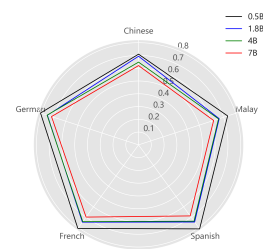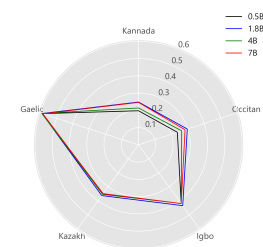

Figure 4: Result for five high-resource languages.

Figure 5: Result for five low-resource languages.