Quantifying Multilingual Performance of Large Language Models Across Languages

Anonymous ACL submission

Abstract

The development of Large Language Models (LLMs) relies on extensive text corpora, which are often unevenly distributed across languages. This imbalance results in LLMs performing significantly better on high-resource languages like English, German, and French, while their capabilities in low-resource languages remain inadequate. Currently, there is a lack of quantitative methods to evaluate the performance of LLMs in these low-resource languages. To address this gap, we propose the **Language Ranker**, an intrinsic metric designed to benchmark and rank languages based on LLM performance using internal representations. By comparing the LLM's internal representation of various languages against a baseline derived from English, we can assess the model's multilingual capabilities in a robust and language-agnostic manner. Our analysis reveals that high-resource languages exhibit higher similarity scores with English, demonstrating superior performance, while low-resource languages show lower similarity scores, underscoring the effectiveness of our metric in assessing language-specific capabilities. Besides, the experiments show that there is a strong correlation between the LLM's performance in different languages and the proportion of those languages in its pre-training corpus. These insights underscore the efficacy of the Language Ranker as a tool for evaluating LLM performance across different languages, particularly those with limited resources.

1 Introduction

Large Language Models (LLMs), such as GPT-4, Claude-3 and LlaMa-3, 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 LlaMa2 (Touvron et al., 2023). This disparity leads to a bias where LLMs exhibit superior performance in high-resource languages like English, German, and French but struggle significantly with low-resource languages.

The imbalance in language representation during the training phase of LLMs means these models are less proficient in low-resource languages. For example, LLM cannot understand the true meaning of some slang terms with specific cultural background, such as Chinese idioms (Zhang et al., 2023). Additionally, recent studies have shown that pre-trained models perform poorly in languages with insufficient training data (Lankford et al., 2024). These observations highlight the need for a metric that quantitatively assesses LLM performance across different languages, especially for those languages with limited resources.

In this paper, we introduce **Language Ranker**, a novel method that leverages internal representations to quantitatively evaluate the multilingual capabilities of LLMs, particularly focusing on low-resource languages. We establish the representation of LLMs on the English corpus as a baseline and measure the similarity between this baseline and representations from other languages. This similarity metric serves as the performance score for each language. We validate our approach by applying the Language Ranker to five state-of-the-art LLMs: LlaMa2 (Touvron et al., 2023), LlaMa3 (Meta-AI, 2024), Qwen (Bai et al., 2023), Mistral-v0.1 (Jiang et al., 2023), and Gemma (Team et al., 2024). We also compare the Language Ranker's performance with the proportion of each language in the training datasets and against other established benchmarks. Through comprehensive experiments, our major observations can be summarized as follows:

- We demonstrate that high-resource languages show higher similarity scores with English, whereas low-resource languages exhibit lower similarity scores, validating the effectiveness of the Language Ranker in measuring language-specific performance.
- We uncover a strong correlation between the LLM performance and the proportion of languages in the pre-training corpus, providing insights into the impact of training data distribution on LLM performance.
- We show that high-resource languages are more evenly distributed in the embedding space, while low-resource languages tend to be narrowly clustered. This distribution analysis further supports our intrinsic metric's reliability in assessing language performance.

2 The Proposed Method

In this section, we will give an introduction to our analysis method. First, we will introduce the dataset that 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 Transformer blocks (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:

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

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 similarity measure, we use the average similarity obtained by several intermediate layers as the final similarity. This process can be described as follows:

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

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 that is a subset of A and B such that for any number of index $i_1 \le i_2 \le ... \le i_n$, 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

In our experiments, we utilize five prominent open-source large models: LlaMa2 (Touvron et al., 2023), LlaMa3 (Meta-AI, 2024), Qwen (Bai et al., 2023), Mistral-v0.1 (Jiang et al., 2023), and Gemma (Team et al., 2024).

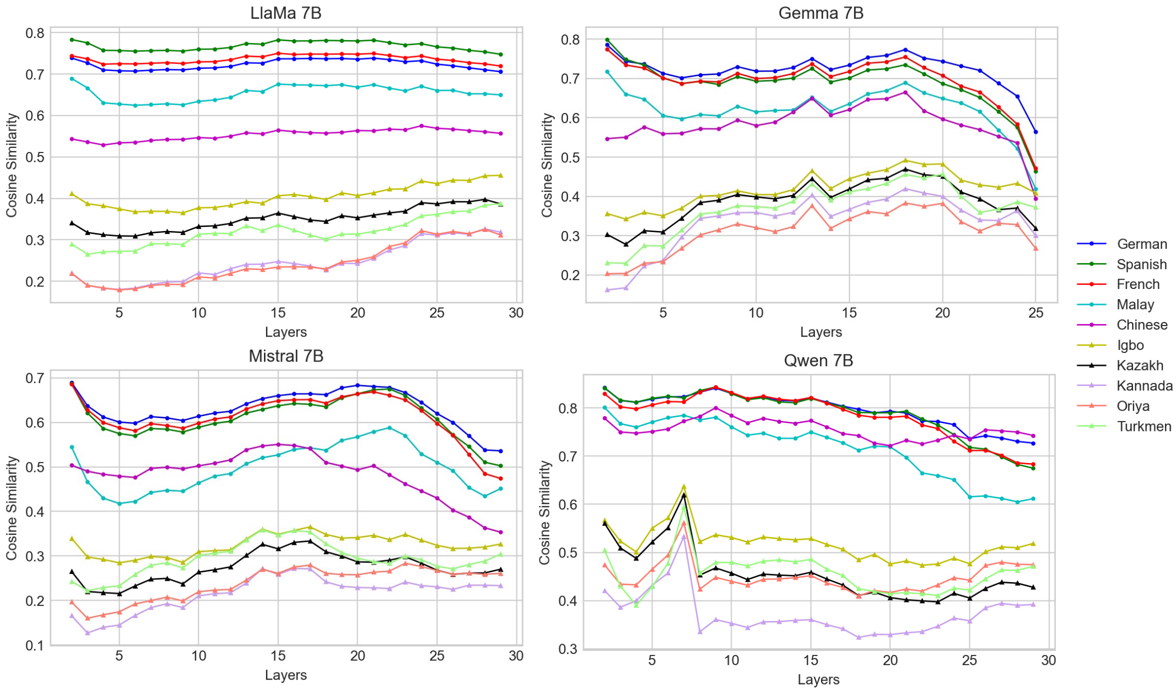

Figure 1: Performance of different LLMs for ten kinds of language. High-resource languages: German, Spanish, French, Indonesian and Chinese; and low-resource languages: Igbo, Kazakh, Kannada, Oriya and Turkmen.

We conduct experiments to answer the following research questions: RQ1: Can the Language Ranker effectively quantify the performance of LLMs across multiple languages? (Section 3.1) RQ2: How consistent are the performance rankings of different LLMs when evaluated across a diverse set of languages? (Section 3.2) RQ3: Is the proposed cosine similarity metric correlated with the proportion of a language in the LLMs' pre-training corpus? (Section 3.3) RQ4: Is the proposed cosine similarity metric correlated with performance on other benchmark tasks for quantifying the multilingual capabilities of LLMs? (Section 3.4)

3.1 Can Language Ranker Quantify LLM Performance Across Languages?

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 observe that high-resource languages have representations more similar to English, whereas low-resource languages show less similarity. Specifically, German, Spanish, French, and Malay generally maintain cosine similarity scores above 0.6, with Spanish and French often showing the highest scores, indicating that these languages are better represented in the models' embeddings. In contrast, low-resource languages, such as Igbo, Kazakh, Kannada, Oriya, and Turkmen, display significantly lower cosine similarity scores, often below 0.4. These results show the disparities in performance across languages and highlights the utility of the Language Ranker in quantifying these differences robustly.

3.2 Comparison Across Different LLMs

In this section, we analyze how various LLMs perform across multiple languages using the Language Ranker. We focus on understanding the consistency of language performance among different LLMs, including models with varying architectures and training specifics. We have the following findings:

(1) **Different models display similar results across languages**. Figure 1 presents the cosine similarity scores across various layers for four different 7B parameter LLMs: LLaMa2, Gemma, Mistral, and Qwen. Four LLMs display a similar trend where high-resource languages (i.e., German, Spanish, French, Malay, and Chinese) consistently exhibit higher cosine similarity scores compared to

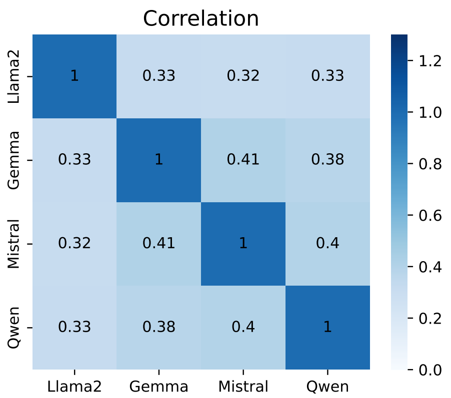

Figure 2: Rank correlation between different LLMs. This is calculated using metric introduced in Section 2.3. It shows high correlations across LLMs.

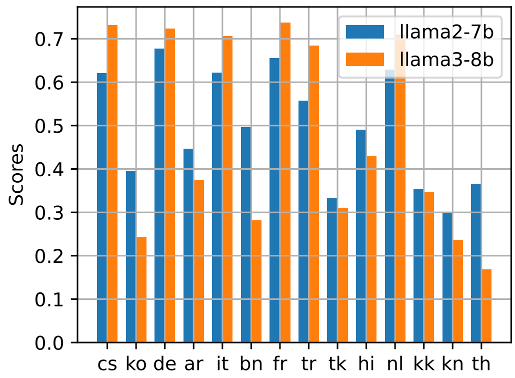

Figure 3: Similarity scores of LlaMa2 7B and LlaMa3 8B. From left to right, the languages are Czech (cs), Korean (ko), German (de), Arabic (ar), Italian (it), Bengali (bn), French (fr), Turkish (tr), Turkmen (tk), Hindi (hi), Dutch (nl), Kazakh (kk), Kannada (kn), and Thai (th).

low-resource languages (i.e., Igbo, Kazakh, Kannada, Oriya, and Turkmen). Figure 2 further corroborates these findings by comparing the rank correlation of the similarity scores across the LLMs. Each LLM's ranking is used as a baseline, and the remaining three models exhibit ranking patterns that are largely similar to this baseline. This similarity indicates that despite differences in model architecture or training specifics, the relative performance of languages remains consistent across these four models.

(2) **Fine-tuning on specific languages will improve its performance**. 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. In Figure 1, we observe that for LlaMa2, Gemma, and Mistral, the performance of Chinese is slightly lower than that of other high-resource languages. However, for Qwen, the performance of Chinese is roughly comparable to other high-resource languages and even shows a gradual improvement in the last few layers. This improvement is more clear in Qwen, mainly due to additional fine-tuning of the Qwen model family on the Chinese corpus, as noted in the technical report of Qwen.

(3) **Comparison of LlaMa2 and LlaMa3.** We also explore the performance of LlaMa2 7B and LlaMa3 8B. From Figure 3, we can observe that in some high-resource languages, such as German, French, and Dutch, both LlaMa2 and LlaMa3 perform well, and LlaMa3 performs better than LlaMa2, for low-resource languages, both perform poorly, and LlaMa3 performs worse than LlaMa2. This phenomenon suggests that there is a certain consistency in the performance of high-resource languages and low-resource languages in similarity metrics, which can be used to distinguish high-resource and low-resource languages.

Language	Proportion	Similarity	Language	Proportion	Similarity
German	0.17%	0.723	Welsh	$\leq 0.01\%$	0.396
French	0.16%	0.737	Persian	$\leq 0.01\%$	0.300
Swedish	0.15%	0.662	Urdu	$\leq 0.01\%$	0.275
Chinese	0.13%	0.552	Kannada	$\leq 0.01\%$	0.236

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%.

3.3 Relationship to Ratio of Training Corpus?

In this section, we explore the relationship between the proportion of each language in the LLaMa2 pre-training corpus and their corresponding performance as measured by the similarity metric. According to the technical report of LlaMa2 (Touvron et al., 2023), we obtain the proportion of the pre-training corpus of some languages. Table 1 illustrates this relationship by listing a selection of languages with their proportion in the training data and their similarity scores relative to English. The table is divided into two parts: the left side lists high-resource languages with relatively higher proportions in the LLaMa2 pre-training corpus, and the right side lists low-resource languages with very low proportions ($\leq 0.01\%$). For example, German, with a proportion of 0.17%, has a high similarity score of 0.723, indicating strong performance in comparison to English. This trend suggests that lan-

Language	ARC				MMLU			
	LlaMa2 7B	Gemma 7B	Mistral 7B	Qwen 7B	LlaMa2 7B	Gemma 7B	Mistral 7B	Qwen 7B
Chinese	27%	71%	57%	66%	32%	54%	37%	44%
German	27%	68%	63%	32%	25%	57%	47%	27%
French	31%	76%	59%	42%	24%	58%	48%	28%
Spanish	31%	77%	60%	46%	29%	56%	52%	33%
Italian	29%	77%	67%	44%	23%	56%	44%	32%
Kannada	24%	48%	27%	21%	21%	40%	22%	19%
Hindi	28%	60%	42%	22%	25%	45%	32%	23%
Armenian	19%	40%	36%	20%	20%	36%	30%	25%
Marathi	28%	46%	26%	25%	27%	42%	30%	26%
Telugu	30%	42%	30%	30%	24%	33%	34%	23%

Table 2: Performance on two inference tasks

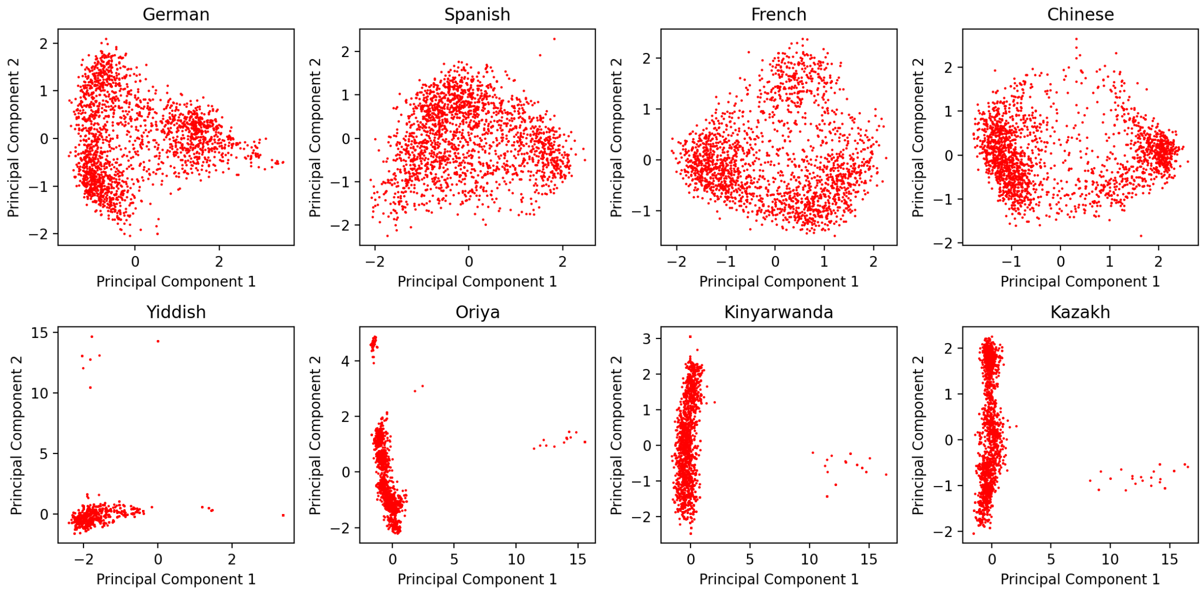

Figure 4: Visualization of the embedding space of Gemma 7B for eight languages. Four figures at the top are high-resource languages, Four figures at the bottom are low-resource languages.

guages with a higher proportion in the pre-training corpus tend to have higher similarity scores, reflecting better model performance. In contrast, low-resource languages like Kannada and Urdu, each with proportions of less than 0.01%, have much lower similarity scores (i.e., 0.236 and 0.275).

3.4 Correlation with Other Inference Tasks?

To more comprehensively reflect the performance of LLMs in various languages, we evaluate the multilingual reasoning ability of these LLMs.

We use MLMM-evaluation[1] as our benchmark dataset to evaluate LLMs' performances on reasoning tasks in various languages. The benchmark dataset can be used to evaluate the LLM across 26 different languages. It consists of three datasets: ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2020). We chose ARC and MMLU for evaluation, and both of them are multiple-choice datasets. The ARC dataset consists of 7,787 multiple-choice science questions drawn from a variety of sources. The MMLU dataset contains multiple-choice questions derived from diverse fields of knowledge. We selected five high-resource languages (Chinese, German, French, Spanish, Italian) and five low-resource languages (Kannada, Hindi, Armenian, Marathi, Telugu) for evaluation, randomly selected 100 samples from each language for 4-shot learning prediction, and used accuracy as the metric. The LLMs evaluated are consistent with Figure 1: LlaMa2 7B, Gemma 7B, Mistral 7B, and Qwen 7B.

The predicted result is shown in Table 2. From

[1] https://github.com/nlp-uoregon/mlmm-evaluation

High-High	Similarity Score	High-Low	Similarity Score	Low-Low	Similarity Score
English-German	0.72	German-Silesian	0.48	Azerbaijani-Turkmen	0.51
Italian-French	0.68	French-Erzya	0.35	Hungarian-Yiddish	0.24
German-French	0.67	Italian-Romany	0.32	Kab-SMT	0.36
French-Chinese	0.59	Italian-Uighur	0.27	Mari-Tatar	0.48

Table 3: Similarity score of different language pairs of Gemma 7B.

High-High	Similarity Score	High-Low	Similarity Score	Low-Low	Similarity Score
English-German	0.72	German-Silesian	0.44	Azerbaijani-Turkmen	0.51
Italian-French	0.69	French-Erzya	0.31	Hungarian-Yiddish	0.19
German-French	0.68	Italian-Romany	0.15	Kab-SMT	0.40
French-Chinese	0.56	Italian-Uighur	0.20	Mari-Tatar	0.42

Table 4: Similarity score of different language pairs of LlaMa2 7B.

the result, we find that for Gemma, Mistral, and Qwen, the performance of high-resource languages is significantly better than that of low-resource languages, and Gemma performs best. For the LlaMa2, the performance in all languages is generally not as good as the first three LLMs. This result shows that LLM reasoning ability in low-resource languages is worse than that in high-resource languages. This result proves that there are differences in performance between high-resource and low-resource languages in reasoning tasks, illustrating the effectiveness of the proposed cosine similarity metric.

4 Further Analysis of Proposed Metric

In the last section, we introduced and evaluated the Language Ranker, demonstrating its ability to quantify the multilingual capabilities of LLMs by comparing their internal representations against an English baseline. This provided a robust measure of how LLMs perform across different languages, especially highlighting the disparities between high-resource and low-resource languages.

Building on these insights, in this section, we delve deeper into the proposed metric to explore its credibility and reliability further. Specifically, we aim to answer the following questions: RQ5: Is choosing English as the benchmark a wise choice (Section 4.1)? RQ6: What does the subspace of each language look like (Section 4.2)? RQ7: Is choosing cosine similarity a wise choice (Section 4.3)?

4.1 Why Using English as Baseline?

In the above sections, we choose English as a baseline. This is based on the a priori assumption that low-resource languages generally perform worse than high-resource languages. But if we choose other high-resource languages as baselines, will we get the same performance? In other words, how can we ensure that our metric is not affected by the English language itself? To answer this question, we divided our probing datasets into three types: High Resource-High Resource (**H-H**), High Resource-Low Resource (**H-L**), and Low Resource-Low Resource (**L-L**). To fulfill our requirement, we utilize Tatoeba-Challenge (Tiedemann, 2020) as our dataset instead of opus-100 because the latter is an English-centric dataset which means there is no Low Resource-Low Resource language pair. Tatoeba-Challenge is a challenge set for machine translation that contains 32G translation units in 2,539 bitexts. The whole data set covers 487 languages linked to each other in 4,024 language pairs. We select four language pairs for each group, English-German (en-de), English-French (en-fr), German-French (de-fr), and Italian-German (it-de) represent **H-H**; German - Silesian (de-szl), French-Erzya (fr-myv), Italian - Romany (it-ro) and Italian - Uighur (it-uig) represent **H-L**; Azerbaijani-Turkmen (az-tr), Hungarian-Yiddish (hu-yi), Kabyle-Standard Moroccan Tamazight (kab-SMT) and Mari-Tatar (ma-ta) represent **L-L**. The results are shown in Table 3 and Table 4.

From the results, we can observe that the score of High-High is higher than the score of High-Low and Low-Low universally. An obvious inference is that the distribution of high-resource languages is relatively close to each other, while the distribution of low-resource languages varies greatly, neither being close to each other nor to high-resource languages. Therefore, the distribution of high-resource languages is relatively consistent, while the distribution of low-resource languages varies greatly. Choosing English as the

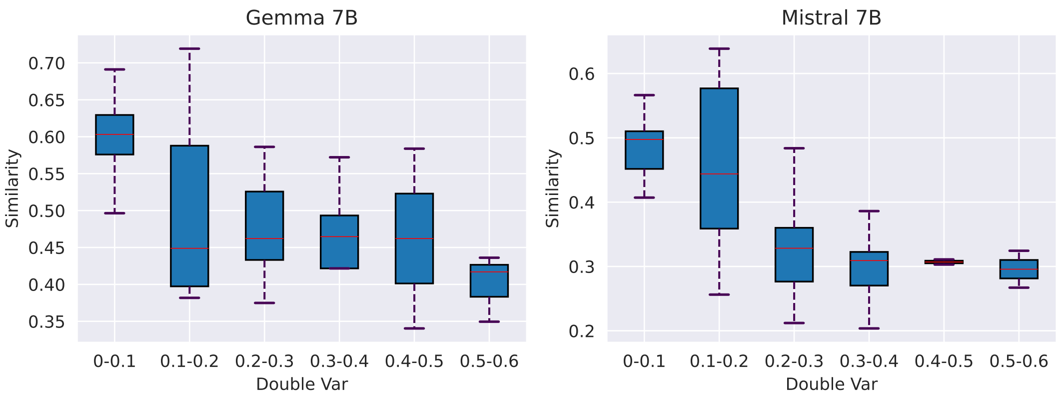

Figure 5: Box plot of the relationship between the double variance and similarity of Gemma 7B and Mistral 7B.

baseline is a convenient choice. Thus, we can also choose other high-resource languages such as German and French as the baseline.

4.2 Deeper Analysis of the Embedding Space

We explained why we chose English as the baseline in the above section. Choosing English is actually choosing a high-resource language as our baseline. A question naturally arises: How can we make sure that the performance of high-resource languages is better than the performance of low-resource languages? The result of Table 2 has confirmed the answer to the question by the reasoning task. To answer this question more deeply, we need to analyze the distribution of the embedding of different languages. As is shown in Figure 4, the top four sub-figures are embedding spaces of high-resource languages, while the bottom four sub-figures are embedding spaces of low-resource languages. It is obvious that high-resource languages are more evenly distributed throughout the space, while low-resource languages are more narrowly distributed, and compressed into a near straight line. Therefore, the performance of low-resource languages is worse than that of high-resource languages, which means it is suitable to choose a high-resource language like English as the baseline.

4.3 Why Using Cosine Similarity?

Recent research (Steck et al., 2024) has shown that cosine similarity is not always a reliable metric. Inspired by Section 4.2, the quality of the performance of various languages can be clearly judged from the subspace distribution. Therefore, we decided to quantify the performance of these languages from a distribution perspective.

Back to Figure 4, the projection distributions of high-resource languages in different directions are relatively consistent, while the distribution of low-resource languages is compressed approximately into a straight line. The projection distributions outside the straight line, such as the projections perpendicular to the straight line, are crowded in a smaller area. This suggests that we can use the projection variance to approximately measure the quality of the distribution.

According to PCA, we assume the embedding vectors are $\{X_i\}_{i=1}^{n}$(being centralized), the projection direction is ω, the project variance $Var(X, \omega)$ can be calculated as follows:

$$\begin{aligned} Var(X,\omega) &= \frac{1}{n}\sum_{i=1}^{n}(X_i^T\omega)^2 = \frac{1}{n}\sum_{i=1}^{n}\omega^T X_i X_i^T \omega \\ &= \omega^T Cov(X)\omega \quad s.t. \quad \omega^T\omega = 1 \end{aligned} \tag{3}$$

It is obvious that $Var(X, \omega)$ is the eigenvalue of the $Cov(X)$. For high-resource languages, projection variance in different directions should be close to each other so that it can be evenly distributed in all directions. The opposite is true for low-resource languages. Therefore, we can extract the first K eigenvalues $\{Var(X, \omega_i)\}_{i=1}^{k}$ and calculate their variance. The variance of the eigenvalues can be used to measure the differences in the distribution in each direction, which is called double variance. This metric can be used to specifically measure the quality of the distribution. The higher the double variance, the more unbalanced the distribution and the worse the performance, and the vice versa.

We employ the box plot to show the relationship between the proposed cosine similarity metric and the double variance metric more clearly. From the results of Table 5, we can observe that for each LLM,

Gemma 7B						Mistral 7B					
High-resource			Low-resource			High-resource			Low-resource		
Language	Double Var	Similarity	Language	Double Var	Similarity	Language	Double Var	Similarity	Language	Double Var	Similarity
Italian	0.04	0.66	Nepali	0.75	0.42	Italian	0.12	0.59	Nepali	0.60	0.28
French	0.09	0.69	Kazakh	0.85	0.38	French	0.17	0.65	Kazakh	0.32	0.32
Spanish	0.06	0.68	Burmese	0.36	0.30	Spanish	0.17	0.64	Burmese	0.40	0.20
German	0.10	0.72	Pashto	0.72	0.36	German	0.19	0.66	Pashto	0.73	0.29

Table 5: Similarity scores and double variance results for some languages on Gemma 7B and Mistral 7B.

the languages in the left part are some common high-resource languages, which have higher similarity and lower double variance, while the right part is the opposite for low-resource languages. The second observation is that as the variance increases, the similarity score also tends to decrease. The increase in variance means that the distribution of the subspace becomes uneven and the similarity score decreases accordingly. This shows that the proposed cosine similarity metric can be utilized to roughly measure the quality of distribution of the subspace, which can thus measure the performance of LLM in different languages.

5 Related Work

Representation Engineering. Representation engineering has emerged as an effective approach to enhance the interpretability and transparency of LLMs. Researchers have been leveraging internal representations to tackle various challenges. Zou et al. (2023) summarizes the application of representation engineering in bias, fairness, model editing, and other areas. Schwenk (2007) found that the internal representation of LLM has a certain correlation with time and space, and the internal representation can be employed to represent time and space. Li et al. (2024) found that the representation of the attention head inside the LLM can be used to indicate the correct reasoning direction, probe analysis is further used to correct the internal representation direction to improve the LLM's performance. Marks and Tegmark (2023) study the structure of LLM representations of true/false statements, proved that language models linearly represent the true/false of factual statements. Ju et al. (2024) used probe technique to detect how LLM stores knowledge layer by layer.

Multilingual Language Model. Recent research such as (Qin et al., 2024) summarizes the recent progress and future trends in multilingual large language models. Ahuja et al. (2024) constructed a benchmark to evaluate LLM's multilingual ability comprising 22 datasets covering 83 languages. Huang et al. (2023); Qin et al. (2023) have proven that LLM performance varies substantially across different languages and they employ a prompt technique to improve task performance across languages. Wendler et al. (2024) explores how LlaMa2 works in multilingual tasks and what role English plays in these tasks. The imbalance distribution of training corpus in different languages leads to the bias of LLM towards some high-resource 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 pre-trained 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.

6 Conclusions and Future Work

In this work, we propose the Language Ranker to evaluate the performance of LLMs across diverse languages by comparing their internal representations to English. The results show that high-resource languages show higher similarity scores with English, while low-resource languages have lower scores, validating the effectiveness of our metric in assessing language performance. Besides, there is a strong correlation between the performance of LLMs in different languages and the proportion of those languages in the pre-training corpus. Further, results indicate that high-resource languages are more evenly distributed in the embedding space, whereas low-resource languages tend to be narrowly clustered. In the future, we plan to design more comprehensive benchmarks to measure LLM's capabilities in different languages.

Limitations

The proposed Language Ranker approach provides an initial quantitative way to analyze LLM performance across languages. We acknowledge that the language ranker method we proposed offers only a rough measurement. While our findings indicate a correlation between the similarity scores and the proportion of each language in the pre-training dataset, these scores alone are not sufficient to precisely measure the exact proportions. Our intent was to provide an initial quantitative approach to explore this relationship, and we recognize the need for more comprehensive methods and additional metrics to accurately assess the impact of pre-training data distribution across languages. 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*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks.

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*.

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 preprint arXiv:1803.05457*.

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne 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. *arXiv preprint arXiv:2305.07004*.

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*.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

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*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Meta-AI. 2024. Llama3. https://github.com/meta-llama/llama3. Accessed: 2024-06-14.

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.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and

Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

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

Holger Schwenk. 2007. Continuous space language models. *Computer Speech & Language*, 21(3):492–518.

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*.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM.

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*.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers.

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.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

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.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Ranking Result For LLMs

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

Language	Similarity Score	Language	Similarity Score
German	0.723	Western Frisian	0.378
French	0.737	Tamil	0.347
Spanish	0.768	Gujarati	0.313
Italian	0.706	Kurdish	0.308
Russian	0.734	Pashto	0.284
Dutch	0.709	Assamese	0.260
Polish	0.664	Central Khmer	0.240
Malay	0.651	Panjabi	0.218
Swedish	0.661	Amharic	0.202

Table 6: The similarity score of LlaMa2 7B.

Language	Similarity Score	Language	Similarity Score
German	0.719	Western Frisian	0.443
French	0.691	Tamil	0.420
Spanish	0.683	Gujarati	0.433
Italian	0.662	Kurdish	0.358
Russian	0.674	Pashto	0.362
Dutch	0.658	Assamese	0.396
Polish	0.618	Central Khmer	0.330
Malay	0.615	Panjabi	0.379
Swedish	0.629	Amharic	0.298

Table 7: The similarity score of Gemma 7B.

Language	Similarity Score	Language	Similarity Score
German	0.639	Western Frisian	0.346
French	0.623	Tamil	0.279
Spanish	0.616	Gujarati	0.270
Italian	0.571	Kurdish	0.262
Russian	0.611	Pashto	0.267
Dutch	0.566	Assamese	0.276
Polish	0.514	Central Khmer	0.252
Malay	0.497	Panjabi	0.213
Swedish	0.532	Amharic	0.191

Table 8: The similarity score of Mistral 7B.

Language	Similarity Score	Language	Similarity Score
German	0.805	Western Frisian	0.441
French	0.793	Tamil	0.510
Spanish	0.800	Gujarati	0.469
Italian	0.773	Kurdish	0.436
Russian	0.794	Pashto	0.448
Dutch	0.773	Assamese	0.507
Polish	0.752	Central Khmer	0.407
Malay	0.730	Panjabi	0.385
Swedish	0.759	Amharic	0.470

Table 9: The similarity score of Qwen 7B.

A.2 Details of experiment in Table 4 and Table 3

We selected data from the Tatoeba-Challenge repository[2]. Since the number of samples for some low-resource language pairs is small, we extracted 100 samples for each language pair. If there are less than 100, we extracted all samples and extracted them according to the test-dev-train priority.

High-High	Similarity Score	High-Low	Similarity Score	Low-Low	Similarity Score
English-German	0.64	German-Silesian	0.43	Azerbaijani-Turkmen	0.51
Italian-French	0.60	French-Erzya	0.30	Hungarian-Yiddish	0.18
German-French	0.62	Italian-Romany	0.25	Kab-SMT	0.39
French-Chinese	0.57	Italian-Uighur	0.24	Mari-Tatar	0.46

Table 10: Similarity score of different language pairs of mistral 7B.

High-High	Similarity Score	High-Low	Similarity Score	Low-Low	Similarity Score
English-German	0.81	German-Silesian	0.58	Azerbaijani-Turkmen	0.67
Italian-French	0.75	French-Erzya	0.51	Hungarian-Yiddish	0.39
German-French	0.80	Italian-Romany	0.53	Kab-SMT	0.55
French-Chinese	0.71	Italian-Uighur	0.49	Mari-Tatar	0.55

Table 11: Similarity score of different language pairs of qwen 7B.

[2] https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/data

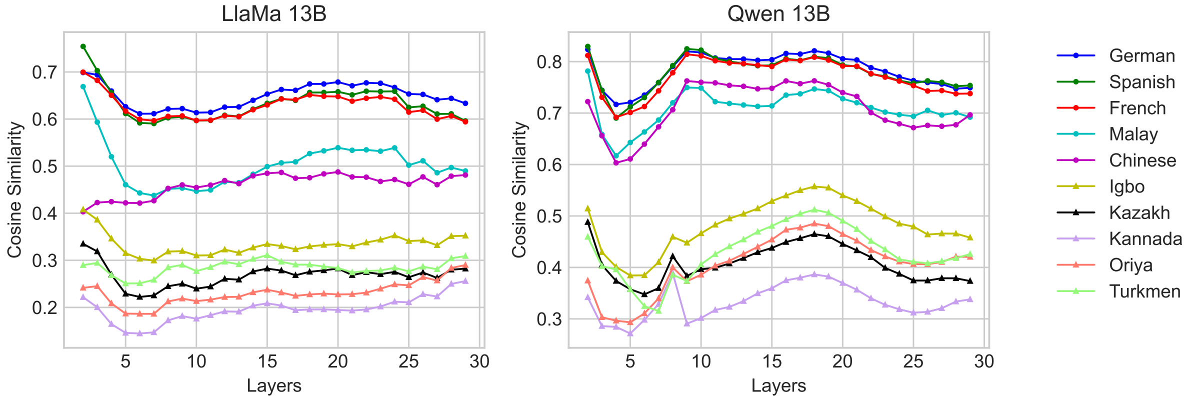

Figure 6: Simlarity scores curves of LlaMa2 13B and Qwen 13B.

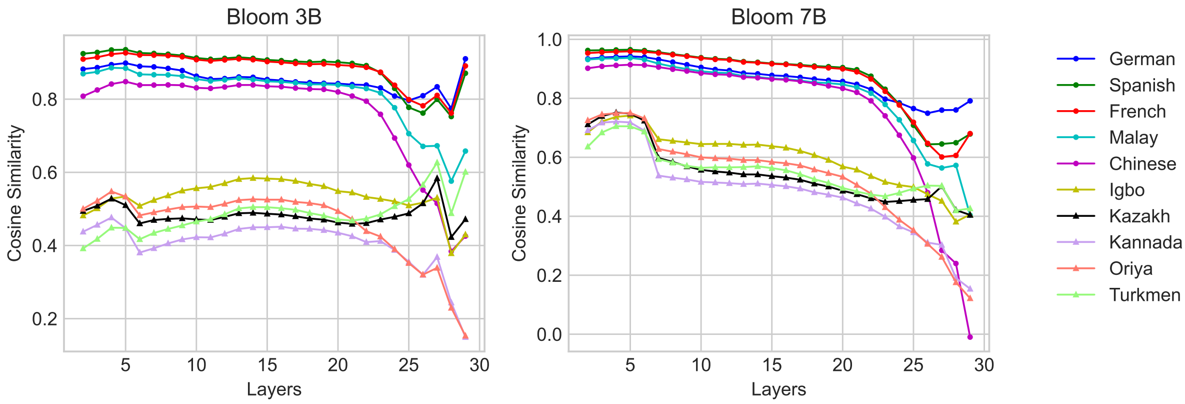

Figure 7: Simlarity scores curves of Bloom 3B and Bloom 7B.

A.3 Performance of LLMs with Other Sizes

A.3.1 LlaMa2 13B and Qwen 13B

We also explore the performance of the similarity metric in LLM of 13B parameters, We use LlaMa2 13B and Qwen 13B to display the results. From Figure 6 we can observe that the partial order results of LLM-13B are roughly the same as those of LLM 7B, and there is a clear gap between high-resource languages and low-resource languages.

A.3.2 Bloom 3B and Bloom 7B

Figure 7 shows the result of Bloom 3B and Bloom 7B. Except for the last few layers of the model, there are obvious differences between high-resource languages and low-resource languages which are similar to the above LLMs, while there are smaller differences within the same category of languages. The score is higher than LlaMa2, Gemma, Mistral, and Qwen.