

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

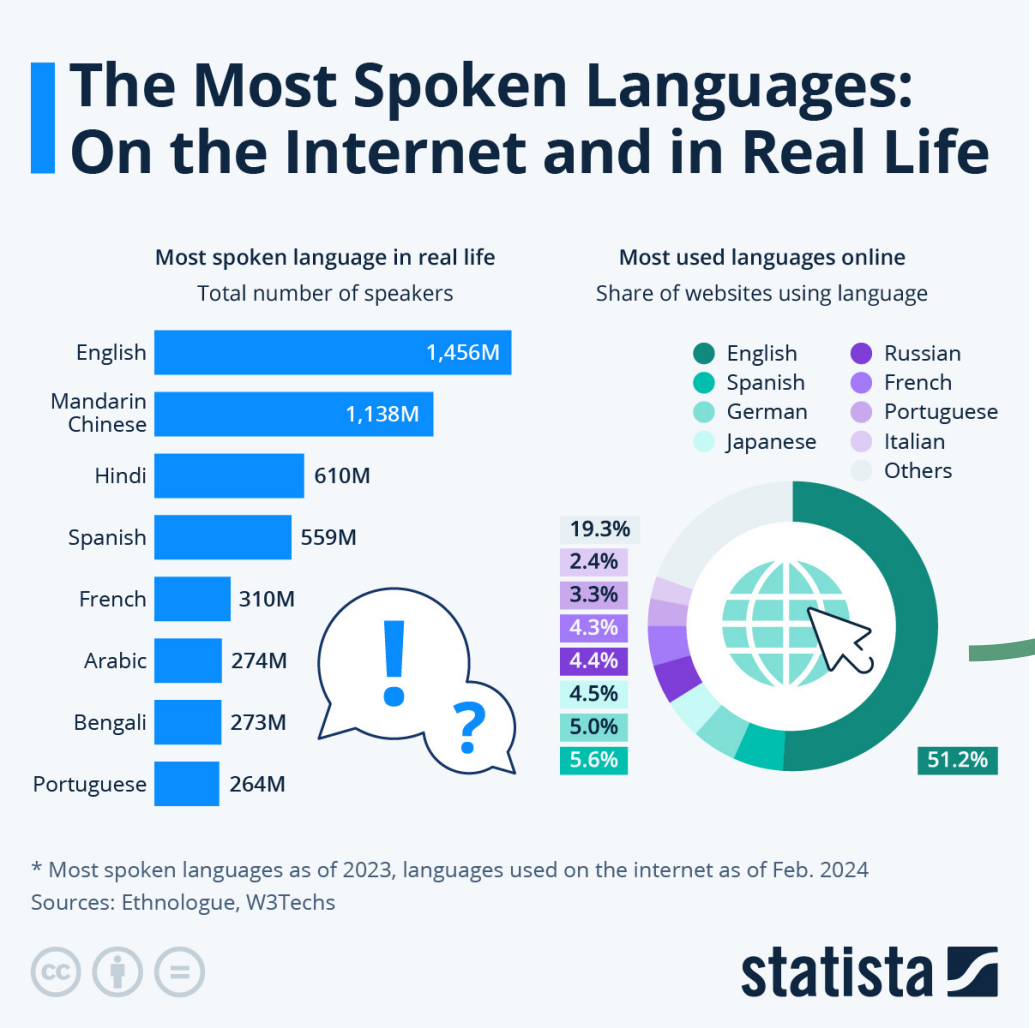
Language Ranker: A Metric for Quantifying LLM Performance Across High and Low-Resource Languages

Mengnan Du

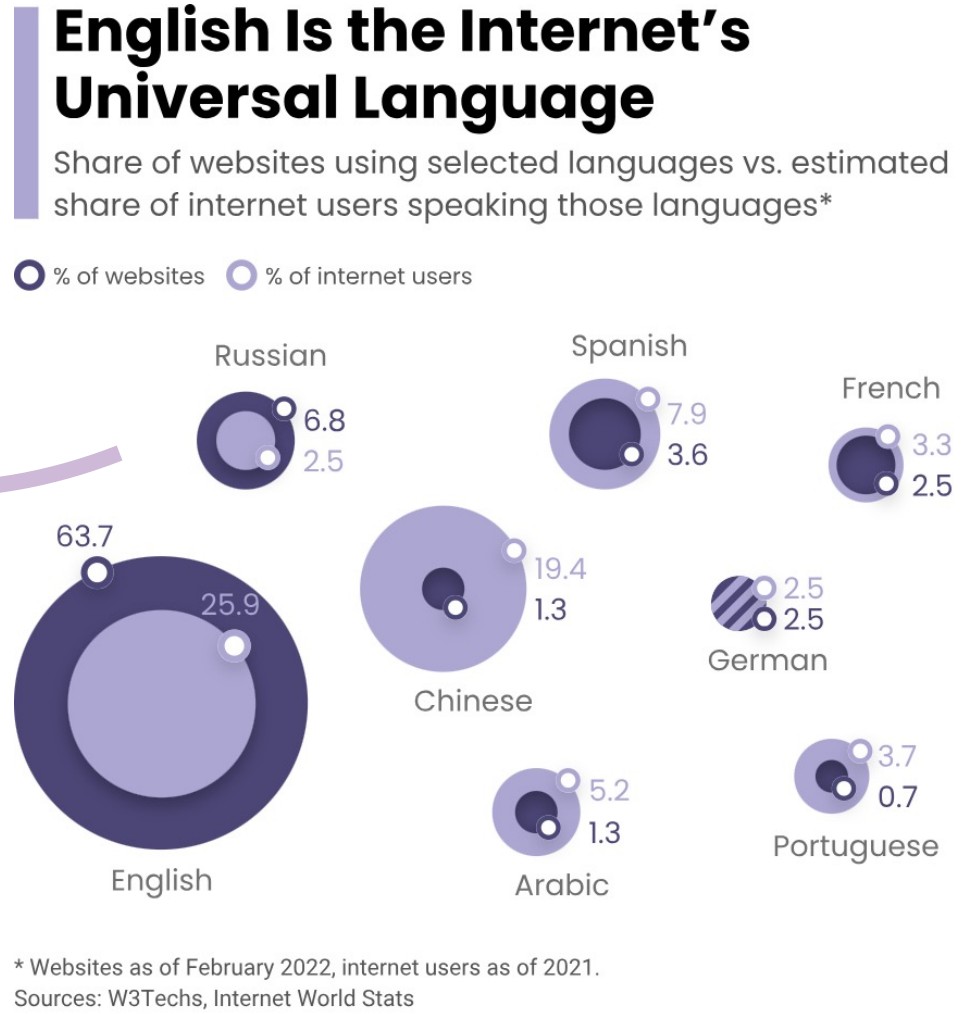
The Chinese University of Hong Kong, Shenzhen

mengnandu@cuhk.edu.cn

Languages Used in Real Life



"More than 60% of online content is in ***English***"



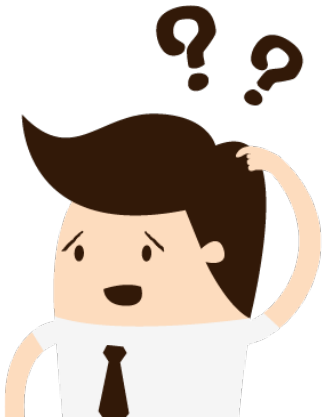
Share of websites using selected languages vs. estimated share of internet users speaking those languages | **Statista**

Language Distribution in Pre-training Data

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Llama 2: Open Foundation and Fine-Tuned Chat Models

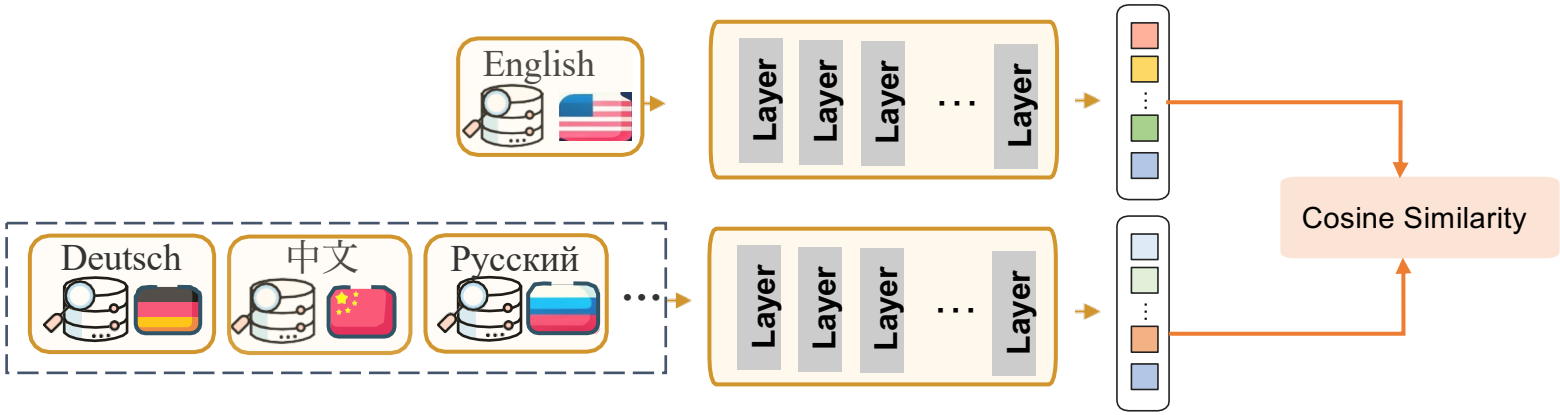
"There are around 7,000 language whole world. Language distribution in pre-training data is extremely imbalanced."



So, what about LLM's multilingual ability? Especially the performance on low resource languages?

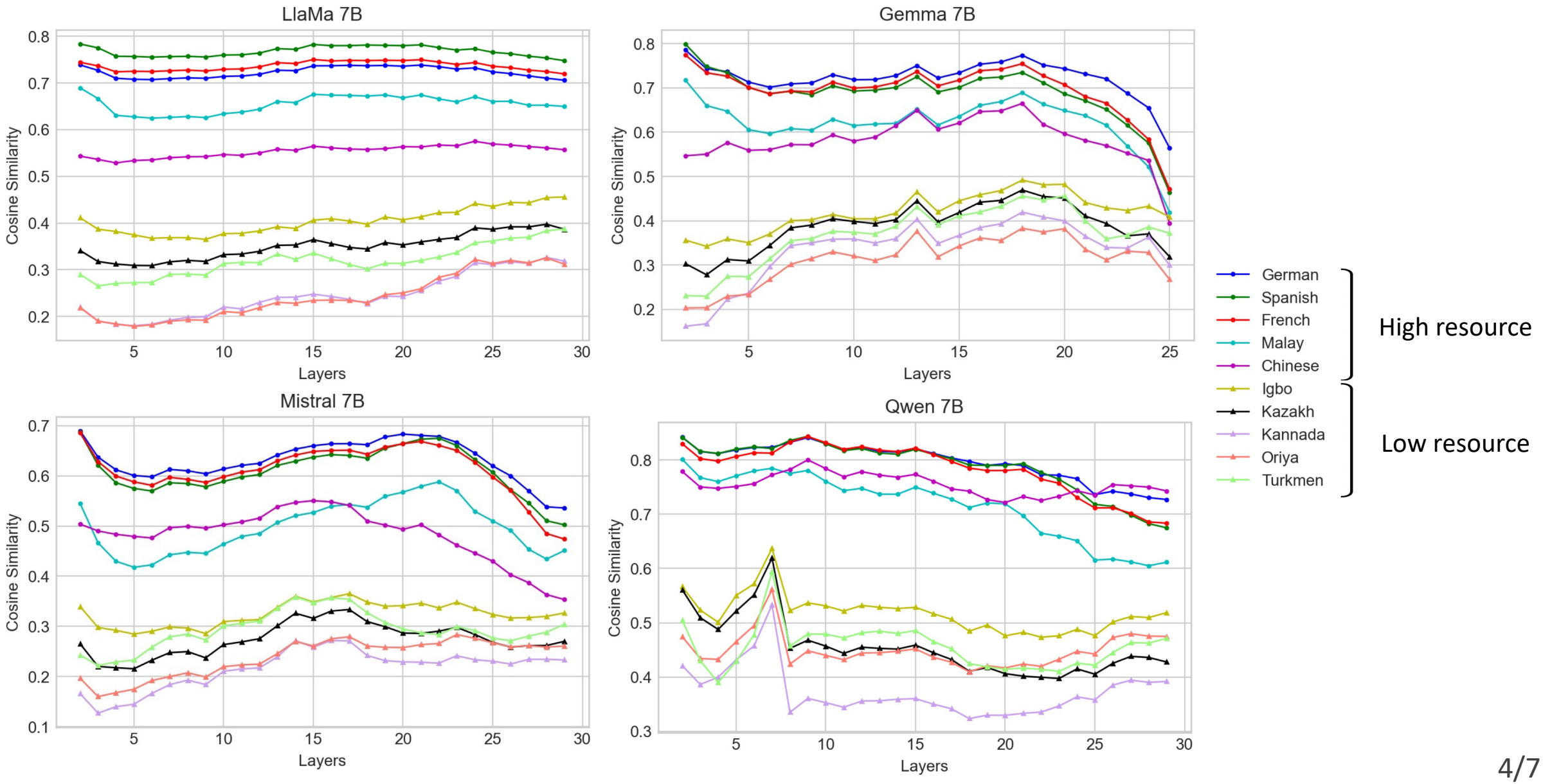
Proposed Language Ranker framework

Language Ranker: quantifying language ability of different languages



- Select the English language and its translation for the target language (e.g., Spanish)
  - Get the representation for Spanish and English respectively
  - Hidden representation: $\boldsymbol{h}_i^{\ell+1} = \boldsymbol{h}_i^{\ell} + \mathrm{MLP}^{\ell}(\boldsymbol{h}_i^{\ell} + \mathrm{Att}^{\ell}(\boldsymbol{h}_i^{l}))$
  - Calculate the cosine similarity for the two representations
  - Average the cosine similarity over all datasets as the metric for language ability

Language Ranker Results



Relationship to Ratio of Training Corpus?

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

- **LLaMa2 pre-training corpus ration vs. Language Ranker score**
  - Left side lists high-resource languages with relatively higher proportions.
  - Right side lists low-resource languages.

Correlation with Other Inference Tasks?

Language	ARC				MMLU			
	LlaMa7B	Gemma7B	Mistral7B	Qwen7B	LlaMa7B	Gemma7B	Mistral7B	Qwen7B
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%

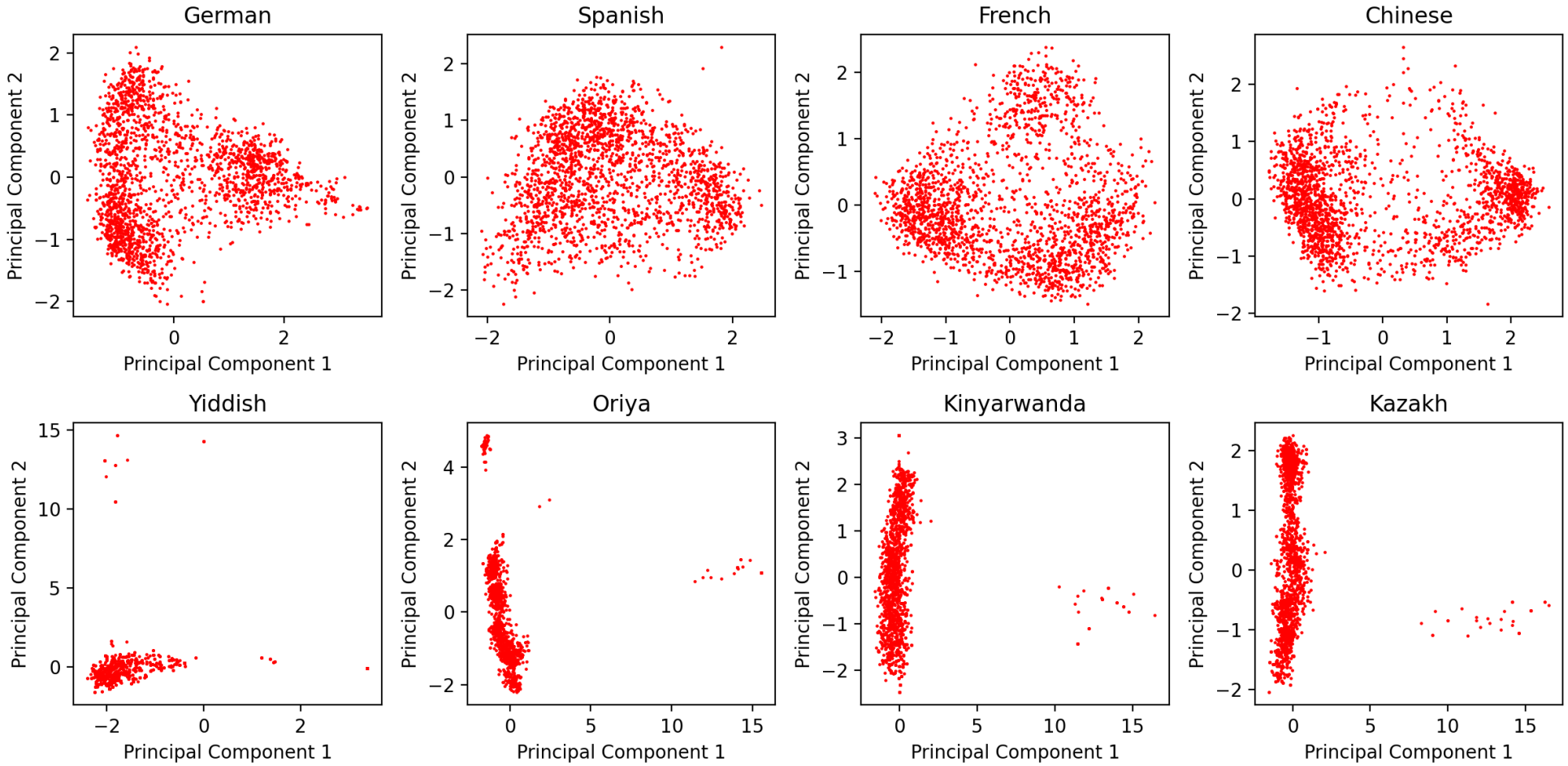
- **MLMM-evaluation as our benchmark**
  - The benchmark evaluates the LLM across 26 different languages.
  - ARC and MMLU multiple-choice datasets.
  - Five high-resource and five low-resource languages.
  - Results are consistent with Language Ranker performance.

https://github.com/nlp-uoregon/mlmm-evaluation

Deeper Analysis of the Embedding Space

High-resource

Low-resource



- High-resource languages are more evenly distributed throughout the space
- Low-resource languages are narrowly distributed, compressed into a near straight line