

NLP for Low-Resource Languages

Fred Philippy

European Language Data Space Country Workshop
Luxembourg
19th June 2024

zortify

About Me

- Data Scientist at *Zortify*
- PhD Student at the *University of Luxembourg (SnT)*
- Research Interest:
Multilinguality and Cross-Linguality in Large Language Models (LLMs)

Battle of (English) LLMs

Google's Gemini Pro Beats GPT-4

Meanwhile, Meta's Llama 3 is on the horizon and has the potential to surpass GPT-4.

AI

Anthropic claims its new AI chatbot models beat OpenAI's GPT-4

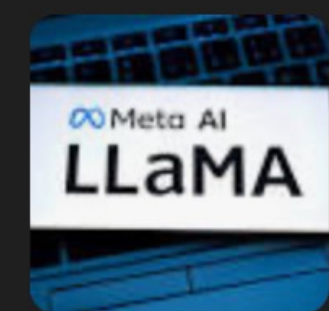
[Home](#) > [AI](#) > [Meet Groq, a Lightning Fast AI Accelerator that Beats ChatGPT and Gemini](#)

Meet Groq, a Lightning Fast AI Accelerator that Beats ChatGPT and Gemini

BZ Benzinga

Mark Zuckerberg's Meta Says Llama 3 Beats Google's Gemini, Mistral And Jeff Bezos-backed Anthropic's Clau

Meta Platforms revealed that Llama 3 has surpassed other AI models in benchmark tests, but OpenAI's GPT-4 is missing from its comparison.



Capability	Benchmark	Description	GEMINI 1.0 PRO	GEMINI 1.0 ULTRA	GEMINI 1.5 PRO (Feb 2024)	GEMINI 1.5 FLASH	GEMINI 1.5 PRO (May 2024)
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	71.8%	83.7%	81.9%	78.9%	85.9%
Code	Natural2Code	Python code generation. Held out dataset HumanEval-like, not leaked on the web	69.6%	74.9%	77.7%	77.2%	82.6%
Math	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	32.6%	53.2%	58.5%	54.9%	67.7%
Reasoning	GPQA (main)	Challenging dataset of questions written by domain experts in biology, physics, and chemistry	27.9%	35.7%	41.5%	39.5%	46.2%
	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	75.0%	83.6%	84.0%	85.5%	89.2%

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

Languages of Europe

- 24 official EU languages
- 7164 living languages in the world
 - of which 291 are European
 - of which many are endangered (~42% worldwide)

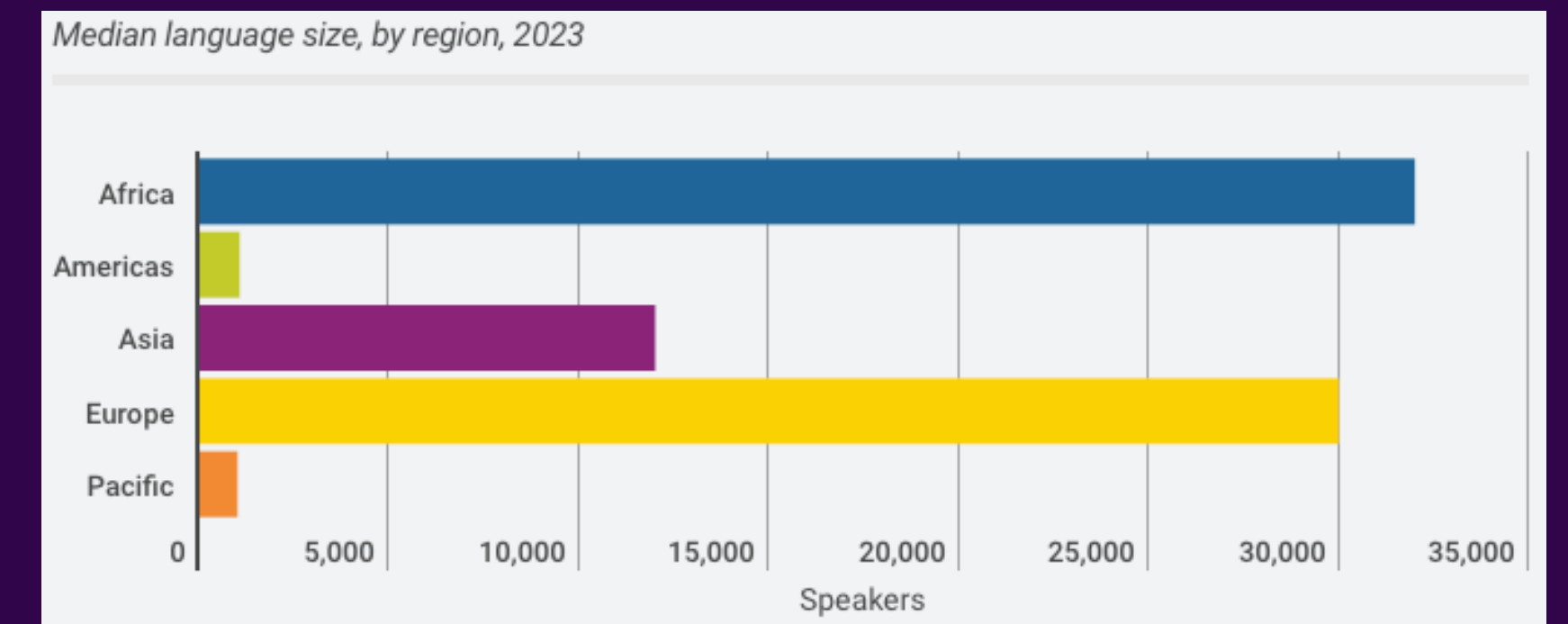
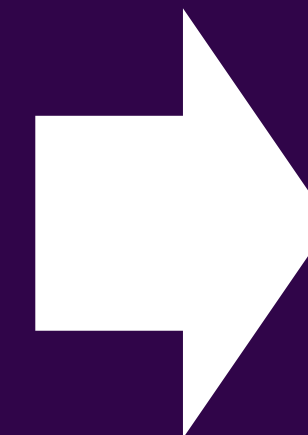
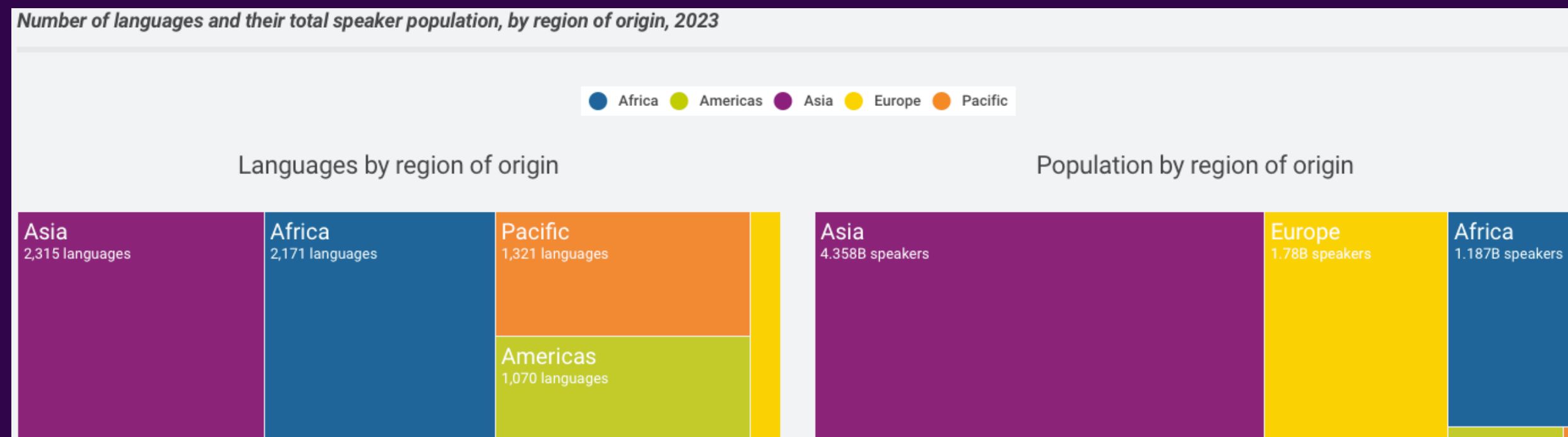
Source: *Ethnologue*



UNESCO Atlas of the World's Languages in Danger. UNESCO. 2010.

Comparison Across Continents

Despite making up only 4% (291) of the world's languages, European languages have the highest number of speakers after Asian languages

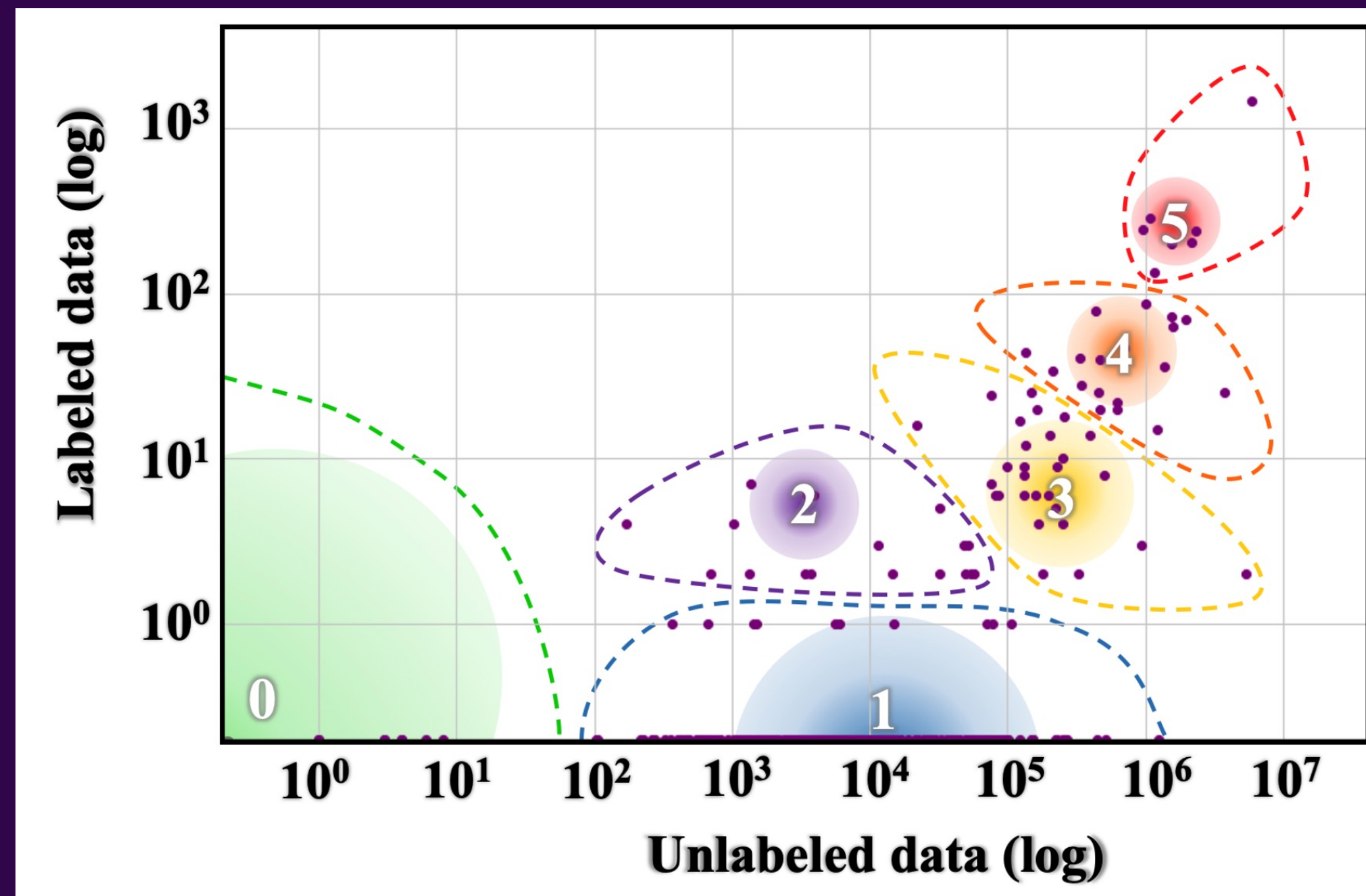


The English-Centric Nature of LLMs

Reasons Behind the English Dominance in LLMs

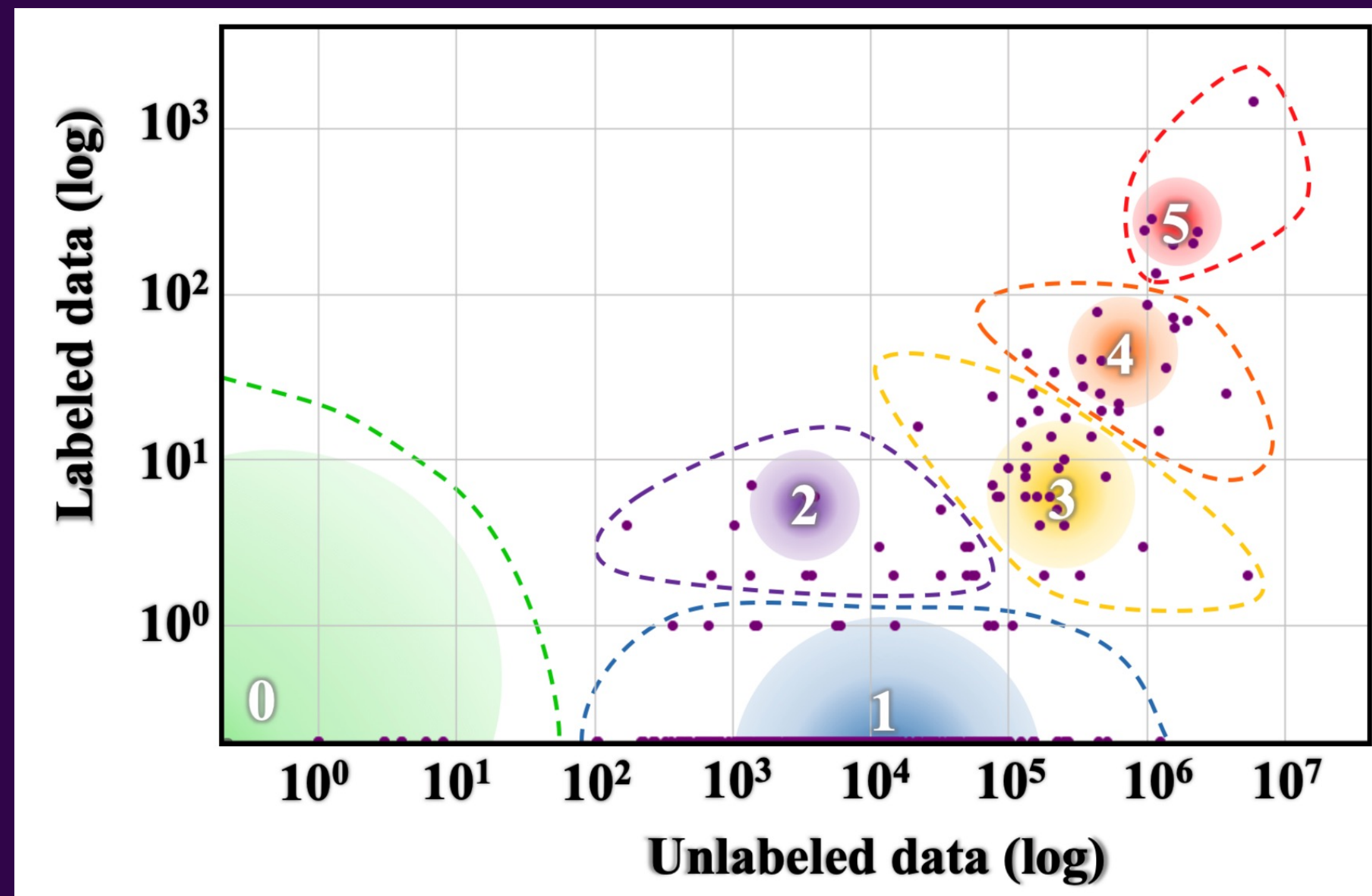
- Leading **NLP companies** (e.g., OpenAI, Meta, Google, Microsoft) are primarily **U.S.-based**
- English is the **world's most widely spoken language**, with 1.5 billion speakers (although only 400 million are native speakers)
- LLM performance heavily relies on the **amount** (and quality) **of available data**
- Languages exhibit significant **linguistic and cultural differences**

Unequal Data Availability Across Languages



The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., ACL 2020)

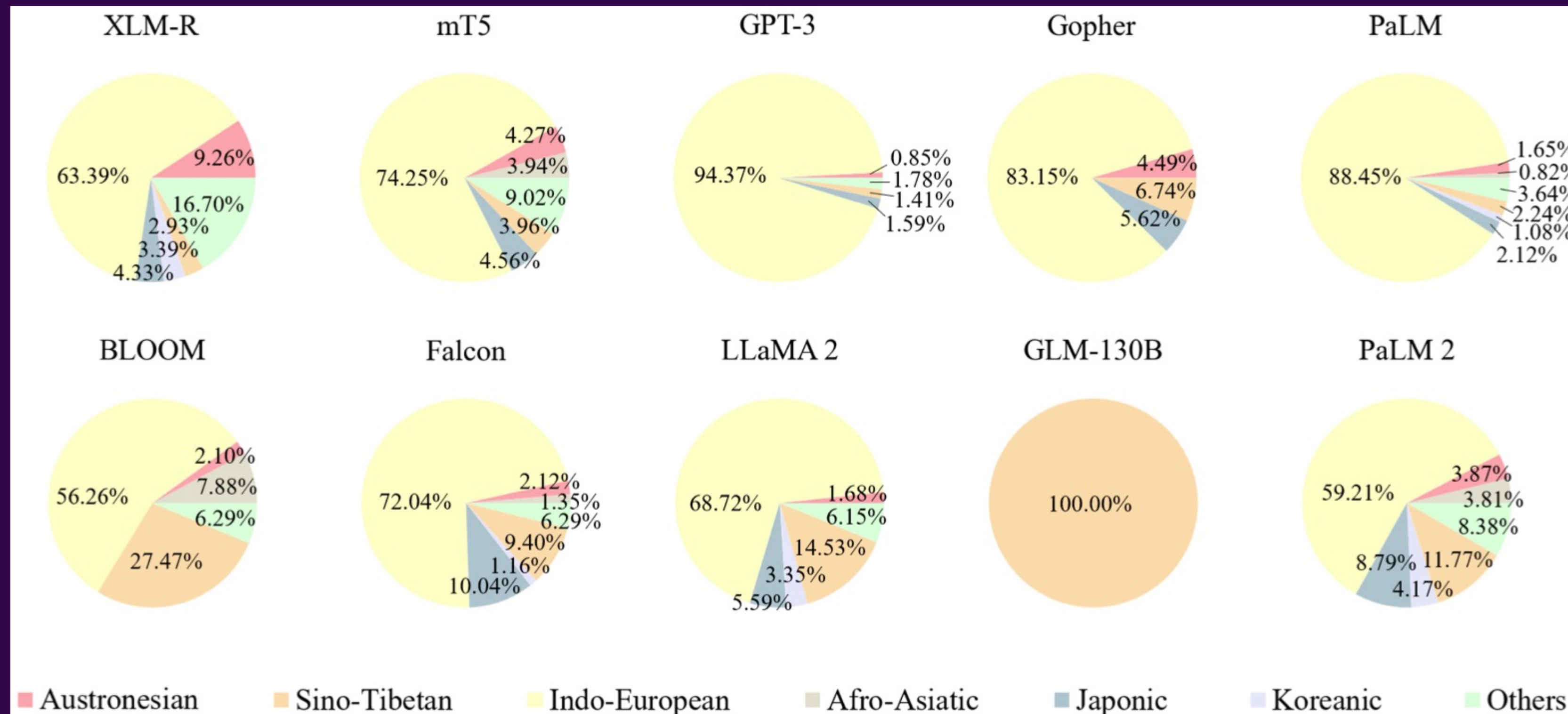
Unequal Data Availability Across Languages



Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al, ACL 2020)

Unequal Language Coverage



Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y., & Gu, H. (2024). A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *ArXiv, abs/2404.00929*.

Examples Where LLMs Fail (in Luxembourgish)

Examples Where LLMs Fail (in Luxembourgish)

PH

You

What are the colours of the Luxembourgish flag?



ChatGPT

The flag of Luxembourg consists of three horizontal stripes: red, white, and light blue, in that order from top to bottom.



Examples Where LLMs Fail (in Luxembourgish)

You
What are the colours of the Luxembourgish flag?

ChatGPT
The flag of
top to bott

You
Wat sinn d'Faarwe vum lëtzebuergesche Fändel?

ChatGPT
D'Lëtzebuergesch Fändel huet d'Reeg, Blo, an Giel, déi traditionell als d'Lëtzebuergesch Trikolore bekannt sinn. Déi dräi Faarwen huet eng symbolesch Bedeitung: d'Reeg representéiert d'Bléi vum Land, d'Blo ass fir d'Kraaft vum Vollek, an d'Giel symboliséiert d'Wuelstand. Des Faarwe kënnen verschidde Interpretatiounen hunn, awer si bleiwen eng wichteg Identitéitssymbol fir d'Land Lëtzebuerg.



Examples Where LLMs Fail (in Luxembourgish)

You
What are the colours of the Luxembourgish flag?

ChatGPT
The flag of Luxembourg is composed of three horizontal stripes of red, white, and light blue from top to bottom.

You
Wat sinn d'Faarwe vum lëtzebuergesche Fändel?

ChatGPT
D'Lëtzebuergesche Fändel ass bekannt d'Blo ass Interpret

You
Wat sinn d'Faarwe vum lëtzebuergesche Fändel?

ChatGPT
D'Fändel vu Lëtzebuerg, dëst gëllt als d'national Fändel vum Land, besteet aus dräi vertikale Sträifen. Vun der Stabmëtt aus geseit, sinn dës Faarwen vun uewen no ënnen: rou, wäiss, a blo. Déi rou Faarf repräsentéiert d'Couleur vun der Adolphe vun Nassau, de Grand-Duc vum Land. Déi wäiss Faarf steet fir Fridden, während déi blo Faarf de franséische Kulturen héichhält, well Lëtzebuerg eng laang Geschicht mat der franséischer Kultur an Politik huet.



Examples Where LLMs Fail (in Luxembourgish)

You have 10 apples and 7 bananas. If I take 3 bananas and 8 apples away from you, how many kiwis do you still have?



Let's break it down step by step:

You start with 10 apples and 7 bananas.

If you take away 8 apples, you're left with 2 apples.

If you take away 3 bananas, you're left with 4 bananas.

So, the number of kiwis you have is still 0, as there's no mention of kiwis in the original scenario.



Examples Where LLMs Fail (in Luxembourgish)

You have 10 apples and 7 bananas. If I take 3 bananas and 8 apples away from you, how many kiwis do you still have?



Let's break it d

You start with 1

If you take awa

If you take awa

So, the number



Fir ze berechnen, wéi vill Kiwien du nach hues, musst du ze éischten d'Gesamtzuel vun den Fruchtstécker wëssen. Du hues 10 Äppel a 7 Bananne, also 17 Fruchtstécker. Wann du 3 Banannen an 8 Äppel wechsels, gëtt et 7 Bananne (ursprénglech $7 - 3 + 8 = 12$ Bananne) a 2 Kiwien (0 Kiwien an 8 Äppel wechsele mat 3 Banannen ëm 3 Kiwien). Also hues du nach 2 Kiwien.



Why is NLP for Under-Resource Languages Important?

A Few Reasons

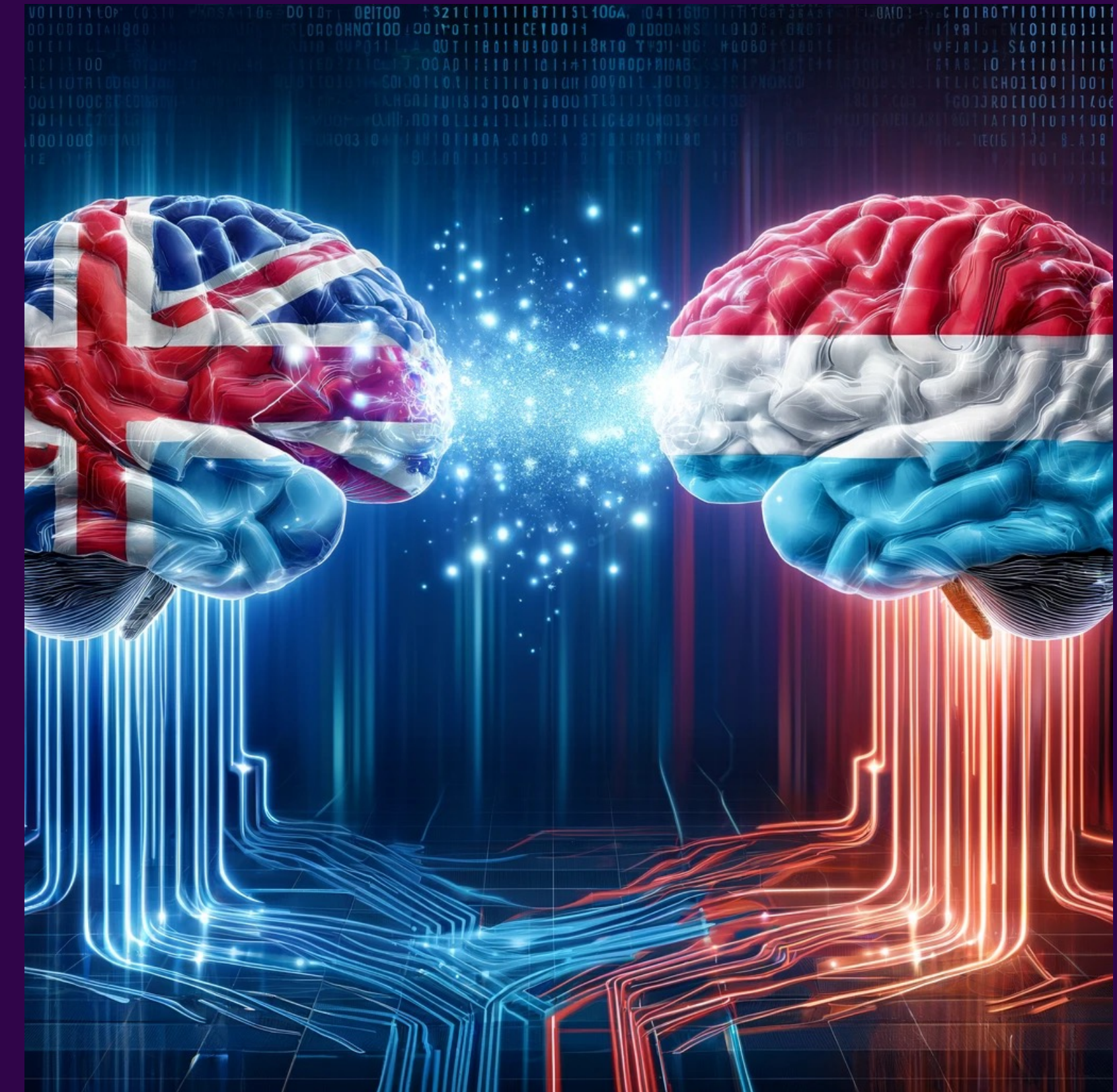
- Preservation of Linguistic Diversity
- Cultural Heritage
- Inclusion and Accessibility
- Economic Opportunities



Generated using DALL·E 3

Cross-Lingual Transfer

- **Definition**
 - **Cross-lingual transfer** is a technique where **knowledge from one language** (often a high-resource language like English) **is used to improve LLM performance in another language** (often a low-resource language).
- Straightforward for humans who speak more than one language, but more challenging for LLMs



Generated using DALL-E 3

Cross-Lingual Transfer

Benefits

- **Resource Efficiency**

Saves time and effort by leveraging existing data and models from high-resource languages.

- **Improved Performance**

Enhances the performance of NLP tasks (e.g., text classification, summarization) in languages with limited data.

Cross-Lingual Transfer

Typical Cross-Lingual Transfer Pipeline



Factors That Impact Cross-Lingual Transfer Performance

Linguistic Similarity

- Relatedness of the source and target languages in terms of grammar, syntax, morphology, etc.

Lexical Overlap

- Proportion of words or tokens that are shared between the source and target languages

Model Architecture

- Number of layers, number of attention heads, embedding dimension, etc.

Pre-Training Settings

- Pre-training objective, tokenizer quality

Pre-Training Data

- Source and size of the pre-training corpora

Parallel Data for Higher Cross-Linguality

- Parallel Data can increase the “**alignment**” of languages in a language model
- Examples of such datasets:
 - **Europarl** (Koehn, 2005), **The United Nations Parallel Corpus** (Ziemiński et al., 2016), **NLLB-200** (NLLB Team, 2022)
- **Limited availability** of high-quality parallel datasets **for many languages**

Future Research



Generated using DALL·E 3

Open Questions

- What other factors contribute to cross-lingual transfer?
- How to transfer to zero-resource languages?
- Should a single model cover ALL the languages? →
Curse of Multilinguality

Conclusion

- NLP is still a long way from supporting the world's 7,000+ languages
- Do not always trust LLMs, especially in under-resourced languages
- The largest and "best performing" LLMs might not always be the best for your language



Generated using DALL-E 3

Thank you

Fred Philippy

fred@zortify.com