Cultural Awareness in Language Models: A Multilingual Perspective

Jose Camacho-Collados

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Copenhagen, Sprogteknologisk Konference, 14 November 2025

About me

Professor at Cardiff NLP (Cardiff University, Wales, UK)

Research interests. Mainly NLP, and in particular:

- ➔ **Semantics** and **language understanding**
- ➔ **Resources** and **evaluation**
- ➔ **Multilinguality** and **cultural awareness**
- ➔ **Computational social science**

Today’s talk

- **NLP and Multilingual Language Models**
  - Subjectivity and disagreements
  - Examples
    - **Cultural awareness**
      - Why we should care
      - Multi-lingual and multi-cultural evaluation

There is no a single gold standard

This is true for many (subjective) NLP tasks.

Basis of perspectivism and **humal label variation** -> Annotators disagree

Need to be considered in both **design settings and evaluation**.

Example: Sentiment analysis

“The movie wasn’t too terrible”

Is this statement positive/negative/neutral?

Many such examples on *offensive language identification*, *hate speech detection*, *emotion recognition*, *political classification*, etc.

Causes of annotator disagreements

Human errors -> **Noise** ❌

Subjectivity of the task

Ambiguity/**underspecification**

Annotator **background** and perspective

Causes of annotator disagreements

Human errors -> **Noise** ❌

Subjectivity of the task

Ambiguity/**underspecification**

Annotator **background** and perspective

➢ **Culture**

Why is cultural awareness relevant for LLMs?

LLMs are **widely used** for variety of tasks and settings.

Answers may be subjective, and vary given different contexts, including user-dependent contexts such as **cultural background**.

Can lead to obvious errors, and to **unfair** scenarios.

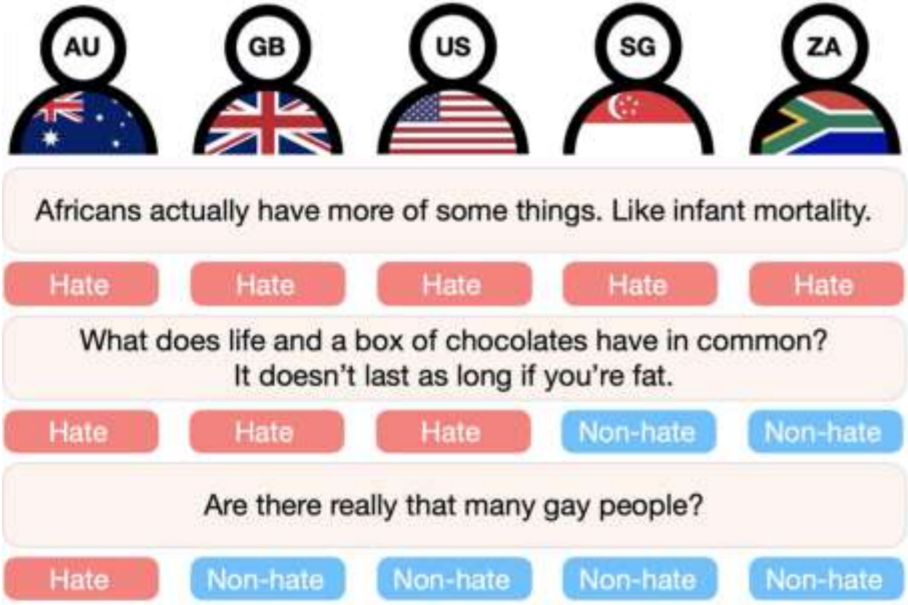
Connected with **language diversity** -> **language is not universal**.

Cross-cultural differences in English hate speech

(Lee, Jung and Myung et al. NAACL 2024)

KAIST

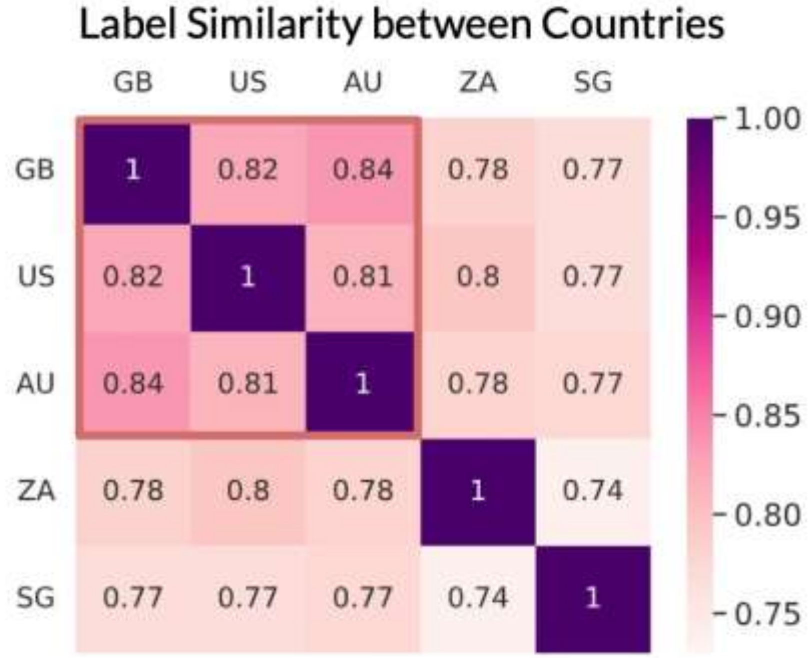
Hate speech dataset annotated by people from 5 different countries



	AU	GB	US	SG	ZA
Africans actually have more of some things. Like infant mortality.	Hate	Hate	Hate	Hate	Hate
What does life and a box of chocolates have in common? It doesn't last as long if you're fat.	Hate	Hate	Hate	Non-hate	Non-hate
Are there really that many gay people?	Hate	Non-hate	Non-hate	Non-hate	Non-hate

Annotator disagreement across countries

UK, US and Australia annotations are similar.



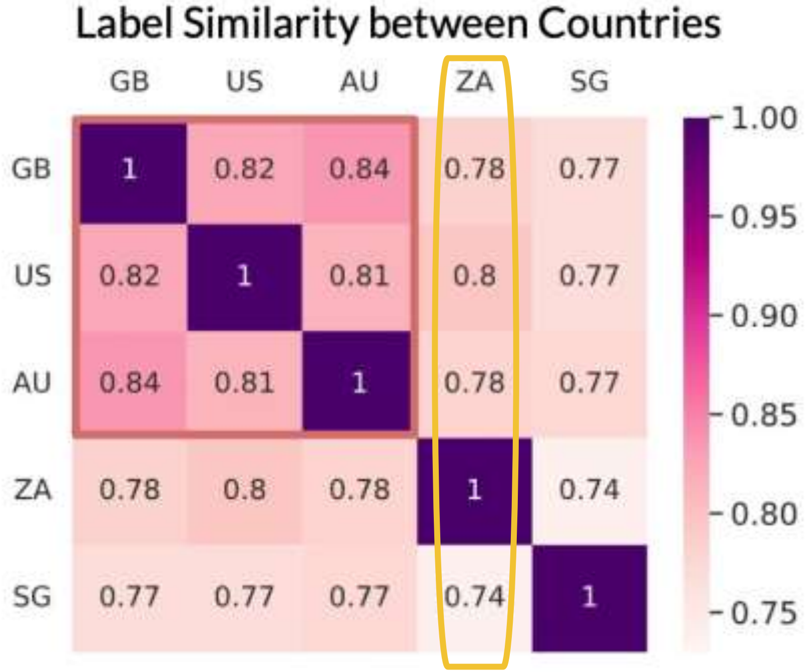
Label Similarity between Countries

	GB	US	AU	ZA	SG
GB	1	0.82	0.84	0.78	0.77
US	0.82	1	0.81	0.8	0.77
AU	0.84	0.81	1	0.78	0.77
ZA	0.78	0.8	0.78	1	0.74
SG	0.77	0.77	0.77	0.74	1

Annotator disagreement across countries

UK, US and Australia annotations are similar.

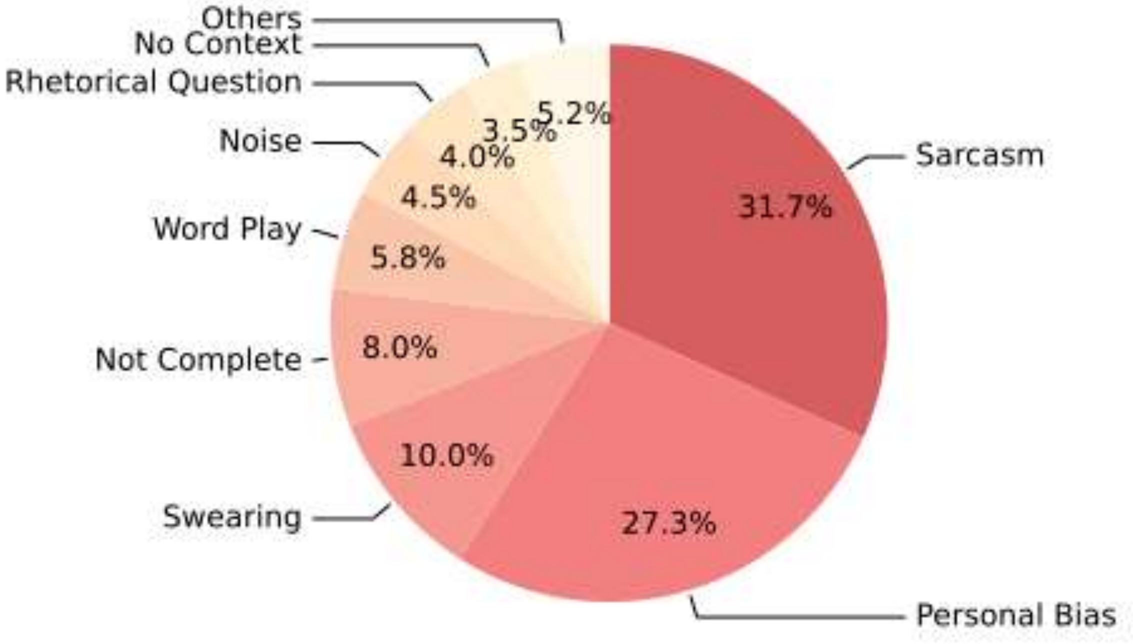
Singapore and South Africa differ.



Label Similarity between Countries

	GB	US	AU	ZA	SG
GB	1	0.82	0.84	0.78	0.77
US	0.82	1	0.81	0.8	0.77
AU	0.84	0.81	1	0.78	0.77
ZA	0.78	0.8	0.78	1	0.74
SG	0.77	0.77	0.77	0.74	1

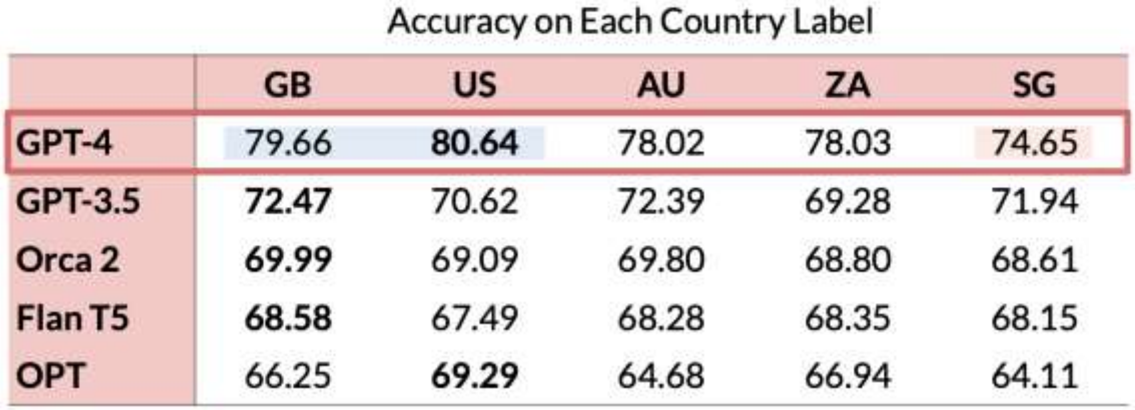
Causes of annotation disagreements



Classification based on disagreement taxonomy from Sandri et al. (EACL 2023)

Results of LLMs prompted to detect “hate speech”

Significant differences between Western countries and Singapore



Accuracy on Each Country Label

	GB	US	AU	ZA	SG
GPT-4	79.66	**80.64**	78.02	78.03	74.65
GPT-3.5	**72.47**	70.62	72.39	69.28	71.94
Orca 2	**69.99**	69.09	69.80	68.80	68.61
Flan T5	**68.58**	67.49	68.28	68.35	68.15
OPT	66.25	**69.29**	64.68	66.94	64.11

OK, this was for English, what about for other languages?

The problem is even more marked when it comes to **different languages** (and especially low-resource languages!)

LLMs are nowadays being used for many languages and types of user.

Most LLMs nowadays are (sort of) multilingual

GPT-4 understands most major languages (50+).

Mistral “supports dozens of languages including French, German, Spanish, Italian, Portuguese, Arabic, Hindi, Russian, Chinese, Japanese, and Korean”.

LLaMA 3 supports “Arabic, English, French, German, Hindi, Indonesian, Italian, Portuguese, Spanish, Tagalog, Thai and Vietnamese”.

Claude supports a “wide array of languages, including but not limited to English, French, German, Portuguese, Spanish, Japanese, Italian, Mandarin, Russian, Arabic, Hindi, and Korean”.

Qwen has a “multilingual support for over 29 languages”.

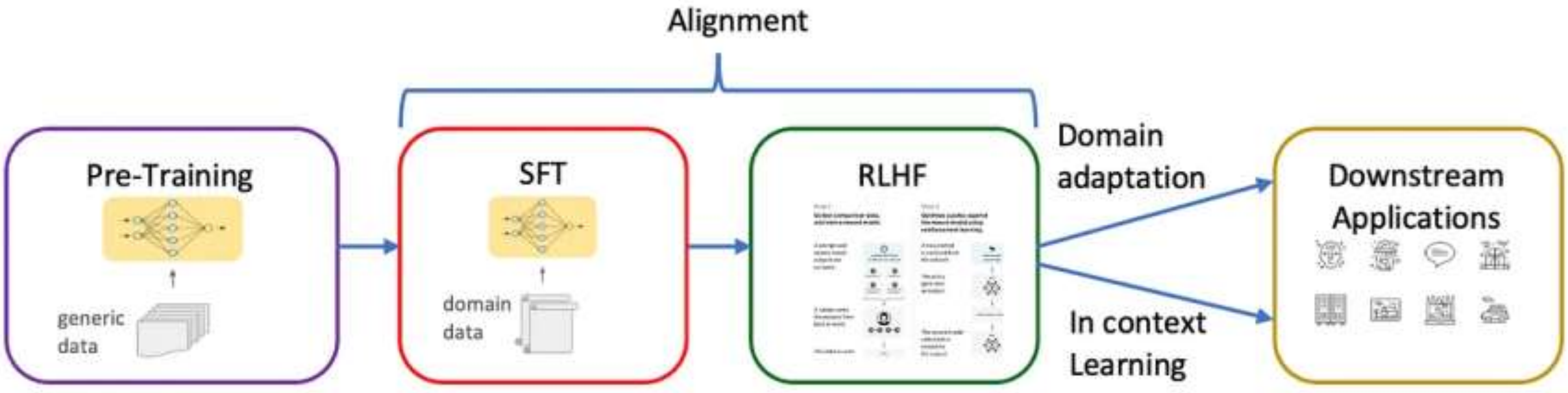
Deepseek-R1 is “currently optimized for Chinese and English”.

So, what are Multilingual Language Models?

Language models that can interpret (and generate) text in different languages.

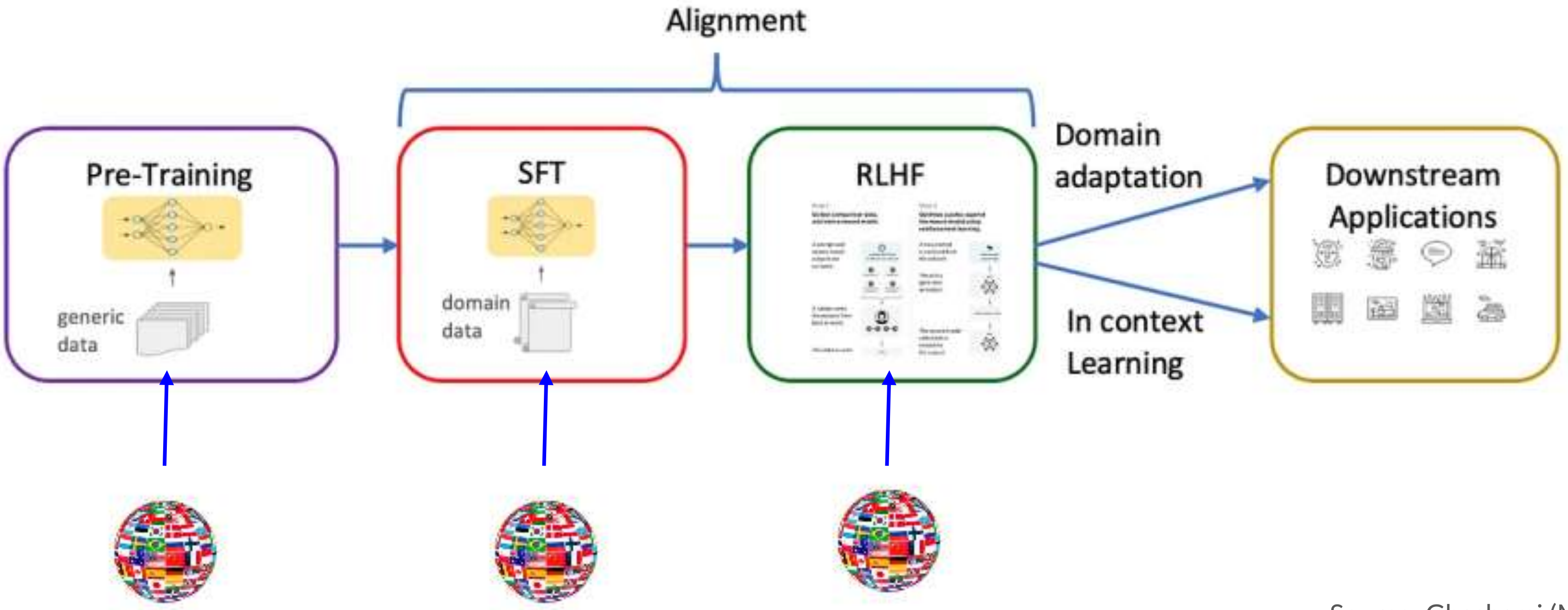
Source image: Vermont Public

LLM training pipeline



Source: Ghashami (Medium)

LLM training pipeline (multilingual)



Source: Ghashami (Medium)

Issue: Language coverage

Data is mostly available in English and high-resource languages.

Current LMs are incredibly data-hungry, so this leads to obvious **performance variation across languages**.

Also, some languages are then more “multilingual” than others!

Solution? No obvious solution other than creating data for low-resource languages and develop models less dependant on data (hard)

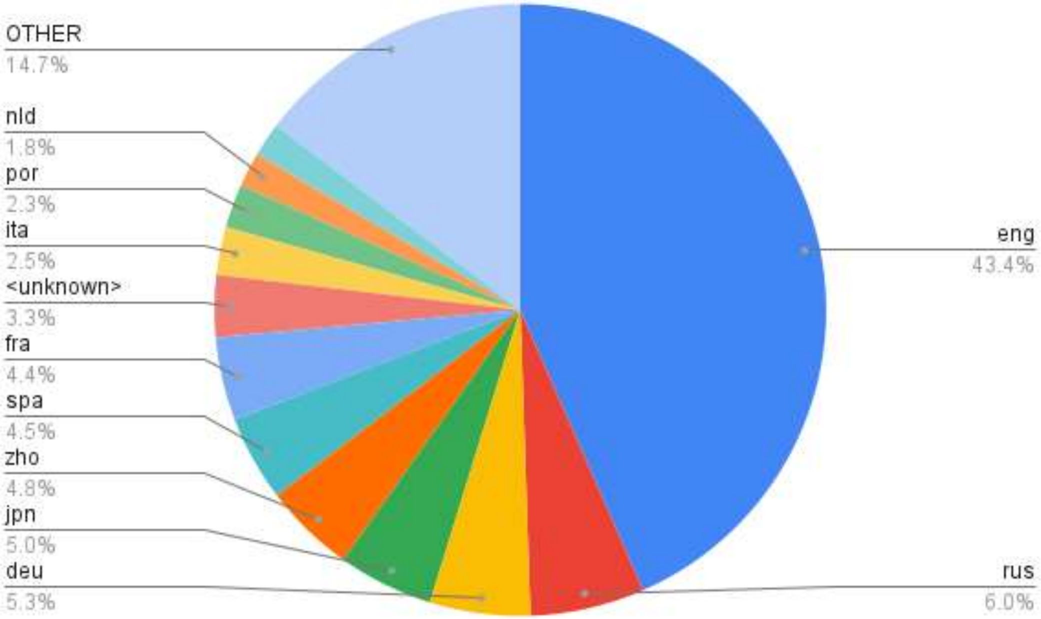
Common Crawl language distribution

Main source of pre-training data for (multilingual) LMs.

English: 43.4%

Top 10 languages: >82%

Rest of ~7,000 languages in the world: <15%



Interesting behaviour: English as pivot

Multilingual LMs use **English as a pivot language** (Wendler et al. 2024, Saji et al. 2025).

This happens because English is the dominant language in these models.

Effect of this manifests in **translation** but also other (non-multilingual) tasks.

Issue: Of course, this comes with unintended biases.

Cultural sensitivity and awareness in multilingual LLMs

Are multilingual LMs sensitive to different **cultures and contexts**?

For instance, common traditions are different across countries.

While there are many “objective” usages, in many cases LLMs need to **adapt to the context of the user** (e.g. their region/country, and others).

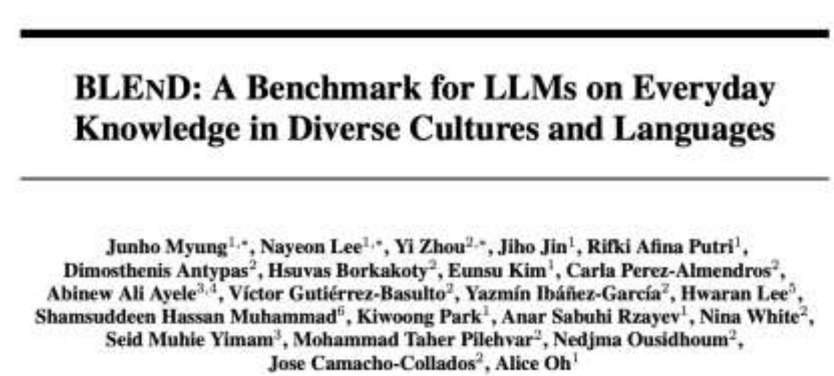
Hard to evaluate: how to get relevant data for many languages and countries?

BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages

(Myung, Lee and Zhou et al. NeurIPs D&B 2024)

Cardiff NLP

KAIST



BLEND: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages

Junho Myung[1,*], Nayeon Lee[1,*], Yi Zhou[2,*], Jiho Jin[1], Rifki Afina Putri[1], Dimosthenis Antypas[2], Hsuvas Borkakoty[2], Eunsu Kim[1], Carla Perez-Almendros[2], Abinew Ali Ayele[3,4], Víctor Gutiérrez-Basulto[2], Yazmín Ibáñez-García[2], Hwaran Lee[5], Shamsuddeen Hassan Muhammad[6], Kiwoong Park[1], Anar Sabuhi Rzayev[1], Nina White[2], Seid Muhie Yimam[3], Mohammad Taher Pilehvar[2], Nedjma Ousidhoum[2], Jose Camacho-Collados[2], Alice Oh[1]

Motivation

Most previous work on cultural adaptation focused on **individual languages or regions/countries** - *now more variety, including at EMNLP!*

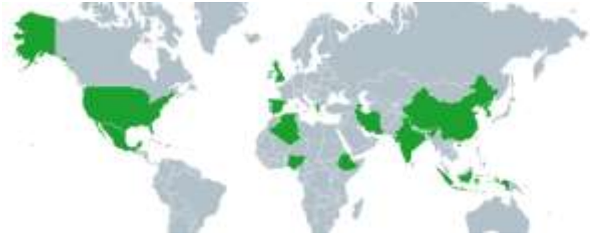
Since nowadays **most LLMs are multilingual**, we should look at this problem holistically.

BLEnD is a (very small) **step in this direction**!

BLEnD: a multi-lingual and multi-cultural benchmark

Most cultural datasets rely heavily on social media or Wikipedia, which often overlook the **mundane everyday lifestyles of underrepresented cultures.**

In BLEND, we **manually** collect questions about everyday life from people from **16 countries and regions, in 13 different languages**



Languages Included:

- ○ English
- ○ Chinese
- ○ Spanish
- ○ Indonesian
- ○ Korean
- ○ Greek
- ○ Persian
- ○ Arabic
- ○ Azerbaijani
- ○ Sundanese
- ○ Assamese
- ○ Hausa
- ○ Amharic

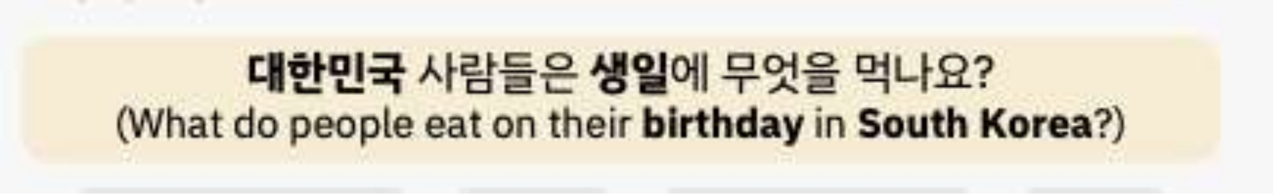
BLEnD: Main team structure

- **Lead authors** were in charge of day to day activities, data preparation and experiments.
- **Language leads** oversaw and supervised anything happening for their particular language/region, from question creation to annotation/aggregation. All language leads were native speakers.
- **Annotators.** Each question for each language/region was answered by at least five paid annotators who lived in the region at least half their lives.
- **Advisory team** provided leading ideas and guidance for each step, as well as oversight.

Topics in BLEnD

- ➢ Food 🥧
- ➢ Sports ⚽
- ➢ Family 👨‍👩‍👦
- ➢ Education 🏫
- ➢ Holidays/celebrations 🥳
- ➢ Work-life 💼

Questions for every topic by each language lead in their native language



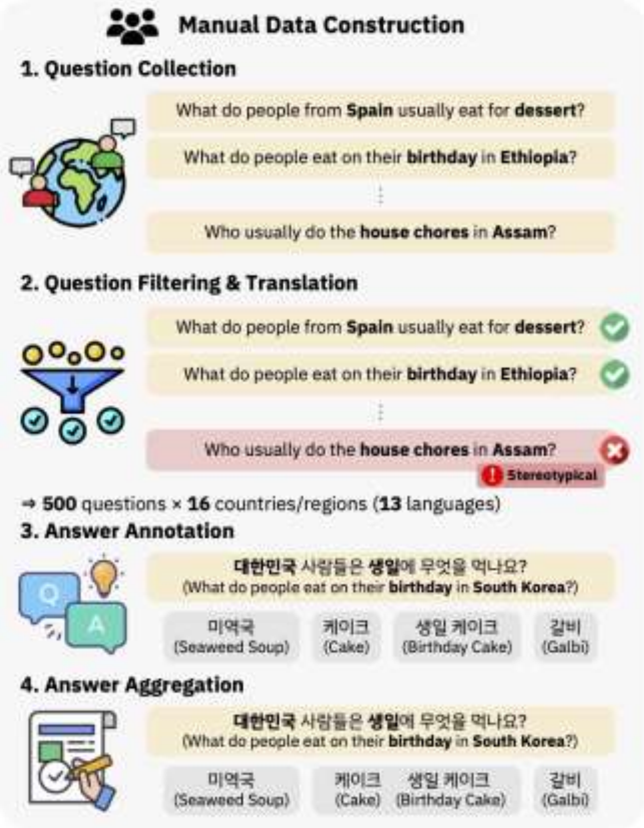
대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

Construction of BLEnD

Manual collection of question and answers from **native annotators in each country/region**

Filtering and aggregation steps are done to remove any duplicates and to ensure high quality

Two tasks (manually validated): **Short Answer Questions (SAQ), and Multiple Choice Questions (MCQ)**



Manual Data Construction

1. Question Collection

What do people from **Spain** usually eat for **dessert**?

What do people eat on their **birthday** in **Ethiopia**?

Who usually do the **house chores** in **Assam**?

2. Question Filtering & Translation

What do people from **Spain** usually eat for **dessert**? ✓

What do people eat on their **birthday** in **Ethiopia**? ✓

Who usually do the **house chores** in **Assam**? ✗ ! Stereotypical

→ **500** questions × **16** countries/regions (**13** languages)

3. Answer Annotation

대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

미역국 (Seaweed Soup) | 케이크 (Cake) | 생일 케이크 (Birthday Cake) | 갈비 (Galbi)

4. Answer Aggregation

대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

미역국 (Seaweed Soup) | 케이크 (Cake) 생일 케이크 (Birthday Cake) | 갈비 (Galbi)

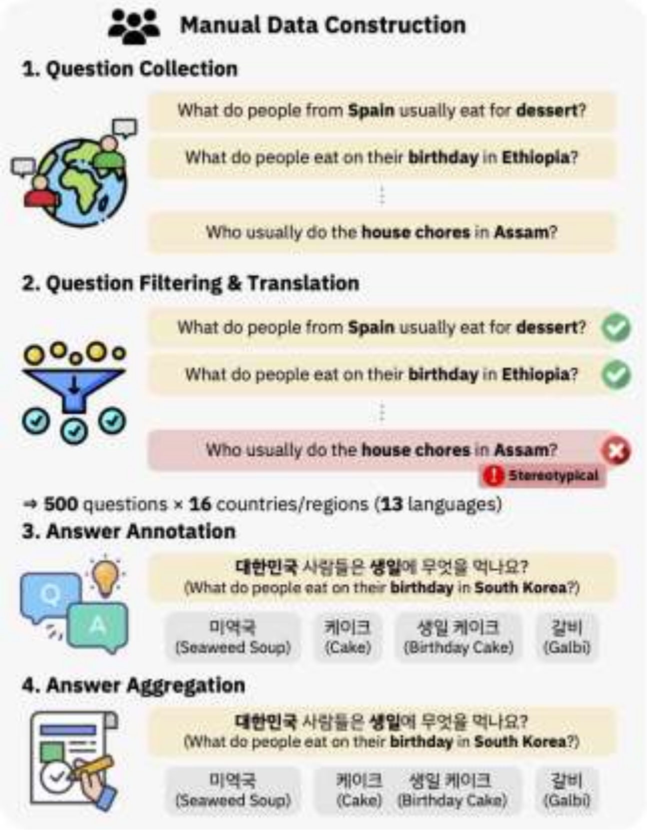
Construction of BLEnD

Manual collection of question and answers from **native annotators in each country/region**

Filtering and aggregation steps are done to remove any duplicates and to ensure high quality

Two tasks (manually validated): **Short Answer Questions (SAQ), and Multiple Choice Questions (MCQ)**

Key feature: avoid use of LLMs for dataset construction and evaluation



Manual Data Construction

1. Question Collection

What do people from **Spain** usually eat for **dessert**?

What do people eat on their **birthday** in **Ethiopia**?

⋮

Who usually do the **house chores** in **Assam**?

2. Question Filtering & Translation

What do people from **Spain** usually eat for **dessert**? ✓

What do people eat on their **birthday** in **Ethiopia**? ✓

⋮

Who usually do the **house chores** in **Assam**? ✗

Stereotypical

→ **500** questions × **16** countries/regions (**13** languages)

3. Answer Annotation

대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

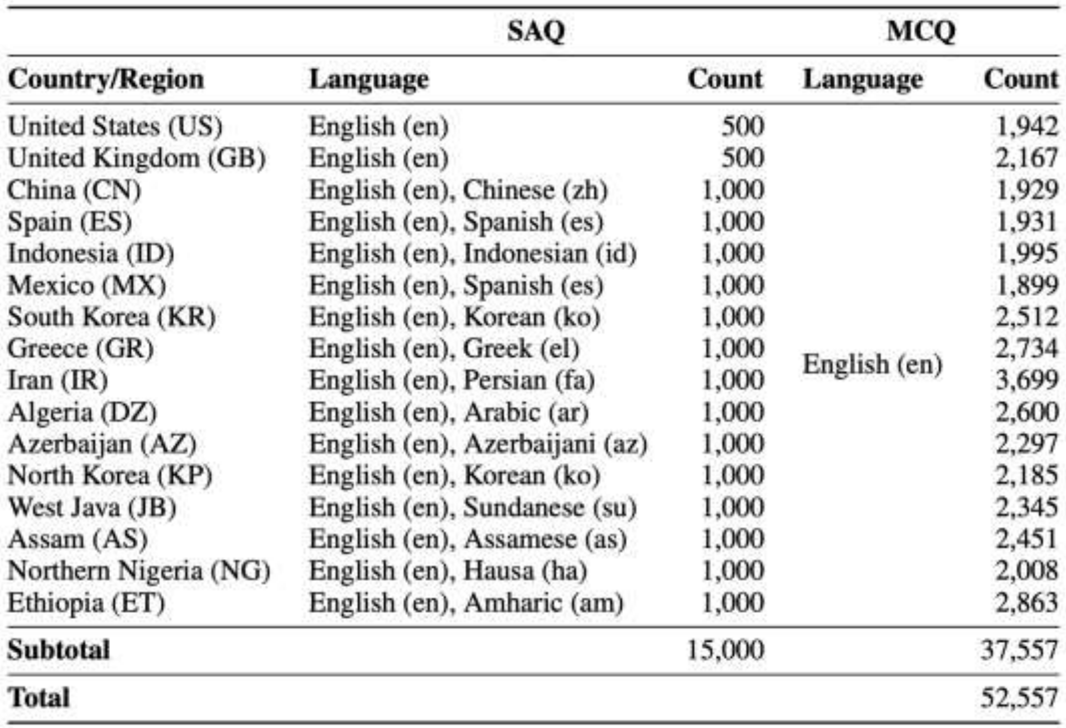
미역국 (Seaweed Soup) | 케이크 (Cake) | 생일 케이크 (Birthday Cake) | 갈비 (Galbi)

4. Answer Aggregation

대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

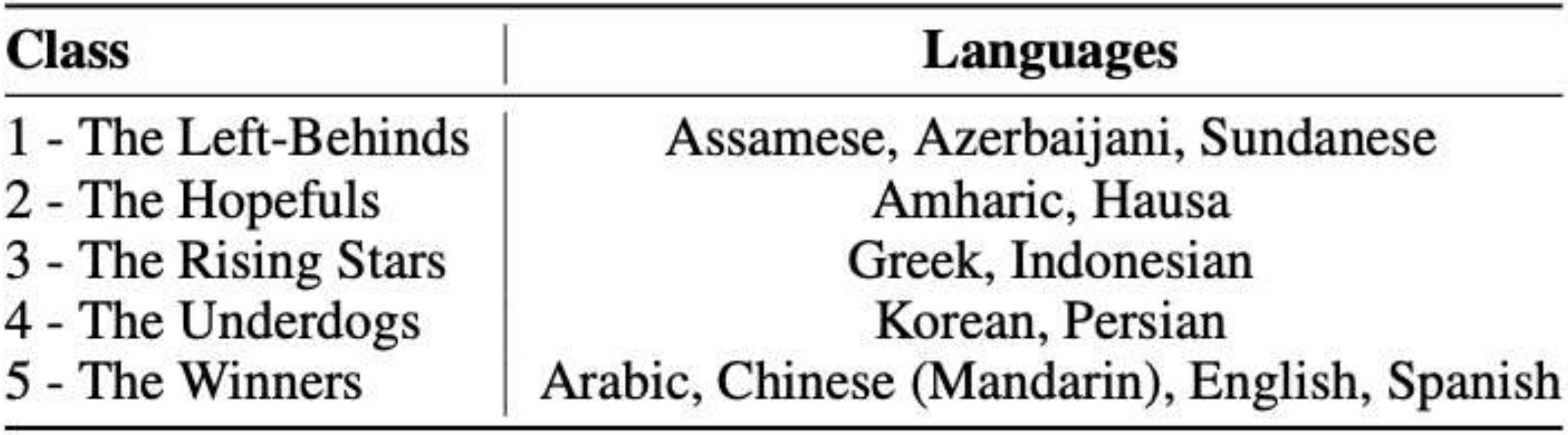
미역국 (Seaweed Soup) | 케이크 (Cake) 생일 케이크 (Birthday Cake) | 갈비 (Galbi)

BLEnD: Statistics



	SAQ		MCQ	
Country/Region	**Language**	**Count**	**Language**	**Count**
United States (US)	English (en)	500	English (en)	1,942
United Kingdom (GB)	English (en)	500		2,167
China (CN)	English (en), Chinese (zh)	1,000		1,929
Spain (ES)	English (en), Spanish (es)	1,000		1,931
Indonesia (ID)	English (en), Indonesian (id)	1,000		1,995
Mexico (MX)	English (en), Spanish (es)	1,000		1,899
South Korea (KR)	English (en), Korean (ko)	1,000		2,512
Greece (GR)	English (en), Greek (el)	1,000		2,734
Iran (IR)	English (en), Persian (fa)	1,000		3,699
Algeria (DZ)	English (en), Arabic (ar)	1,000		2,600
Azerbaijan (AZ)	English (en), Azerbaijani (az)	1,000		2,297
North Korea (KP)	English (en), Korean (ko)	1,000		2,185
West Java (JB)	English (en), Sundanese (su)	1,000		2,345
Assam (AS)	English (en), Assamese (as)	1,000		2,451
Northern Nigeria (NG)	English (en), Hausa (ha)	1,000		2,008
Ethiopia (ET)	English (en), Amharic (am)	1,000		2,863
Subtotal		15,000		37,557
Total				52,557

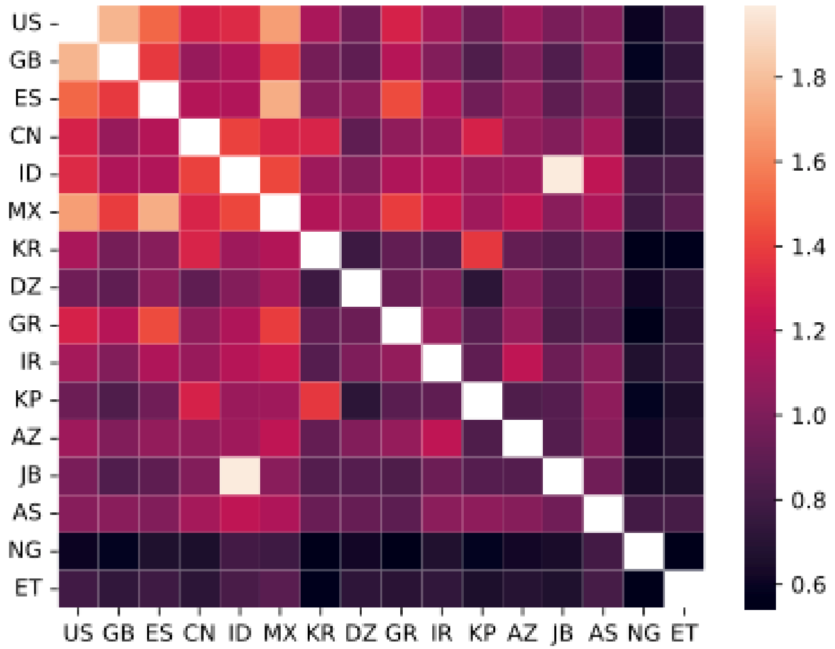
Languages in terms of resource availability



Class	Languages
1 - The Left-Behinds	Assamese, Azerbaijani, Sundanese
2 - The Hopefuls	Amharic, Hausa
3 - The Rising Stars	Greek, Indonesian
4 - The Underdogs	Korean, Persian
5 - The Winners	Arabic, Chinese (Mandarin), English, Spanish

Classification based on Joshi et al. (ACL 2020)

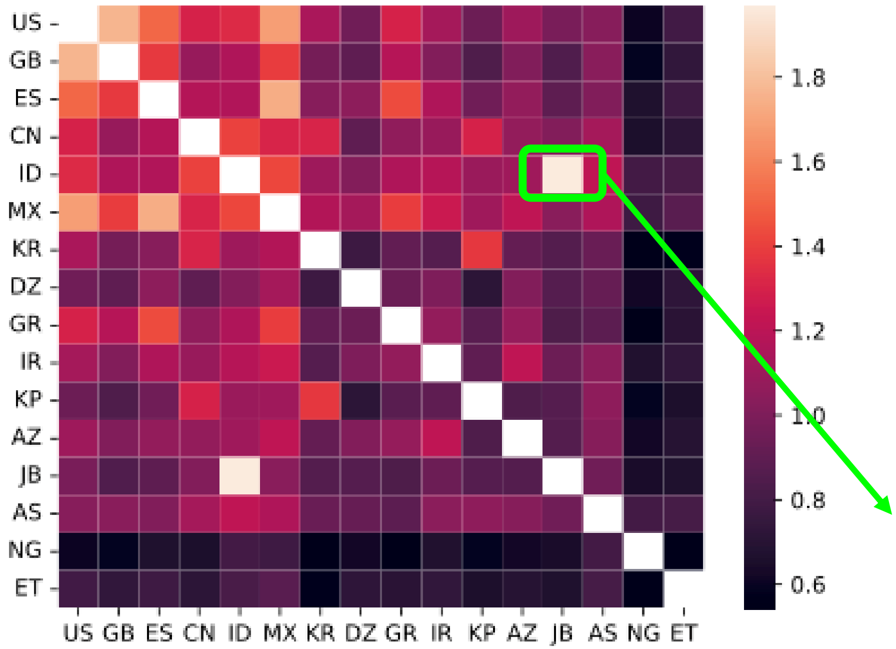
Annotator disagreements across countries



Heatmap on the question answer overlap between countries

(based on English translations)

Annotator disagreements across countries



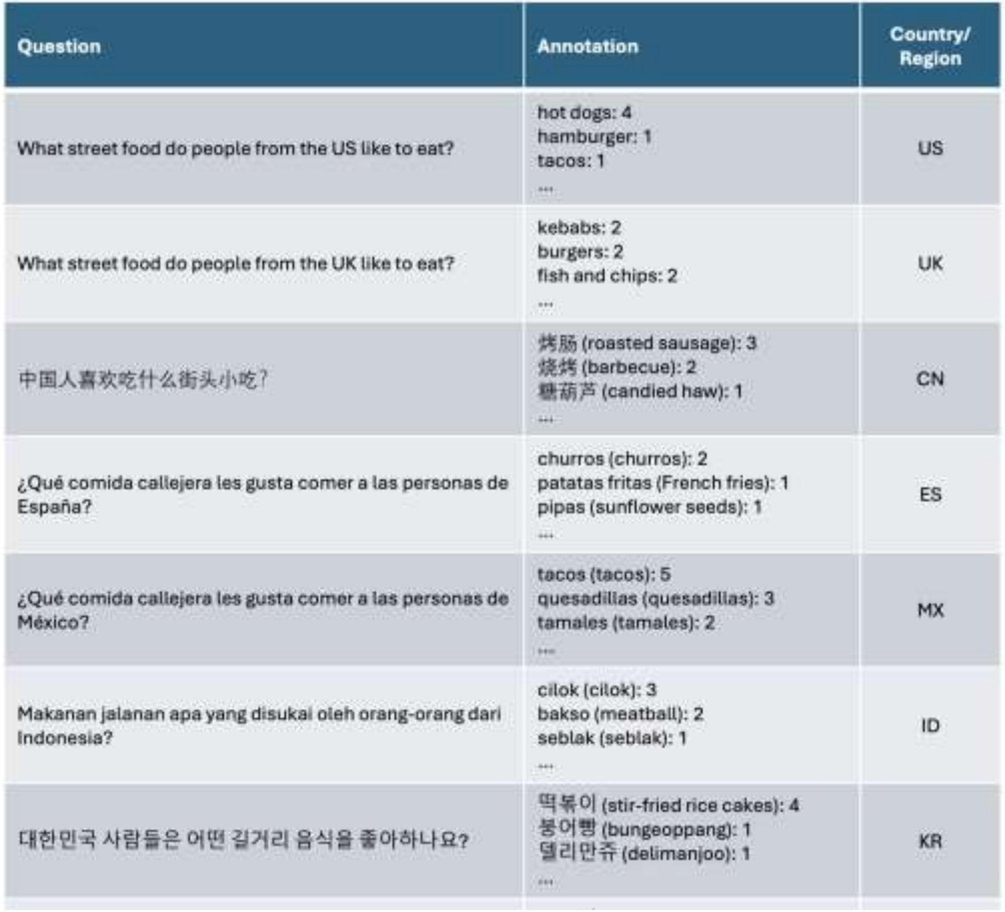
Heatmap on the question answer overlap between countries

(based on English translations)

For instance, Indonesian and West Java answers are quite similar

Example in BLEnD: *“What street food do people like to eat?”*

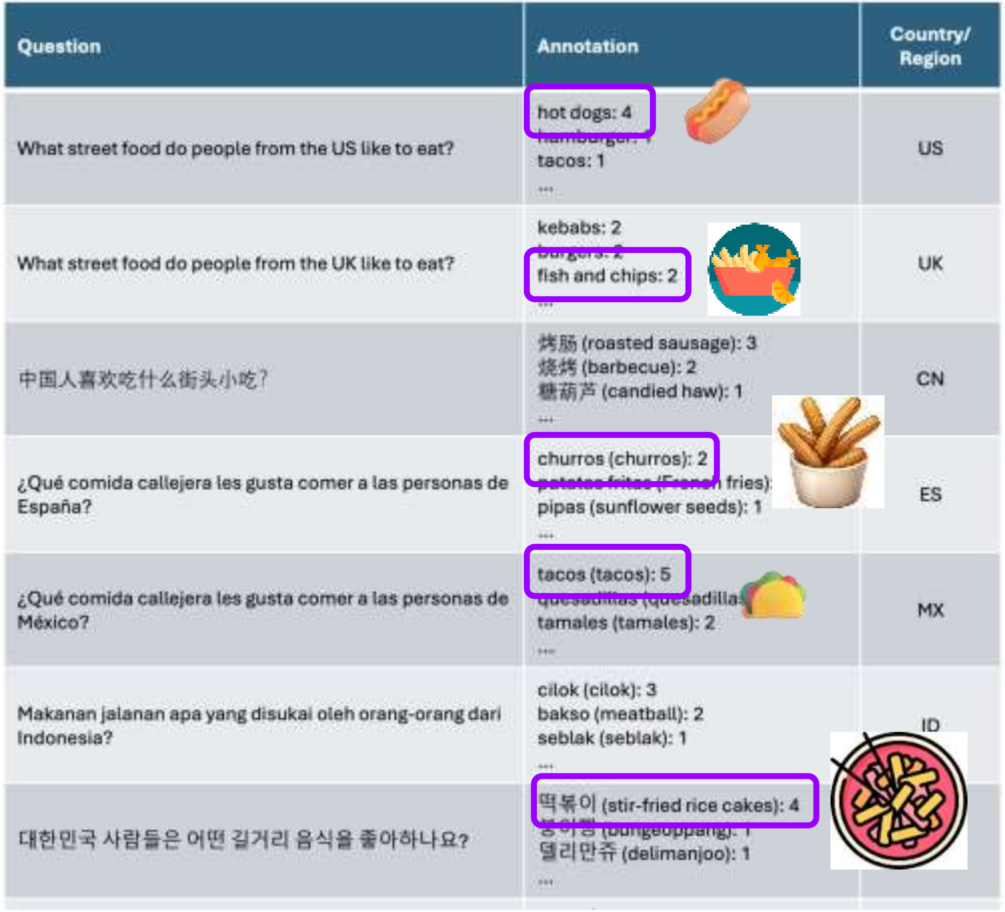
Answers for this simple question vary a lot across countries!



Question	Annotation	Country/ Region
What street food do people from the US like to eat?	hot dogs: 4 hamburger: 1 tacos: 1 ...	US
What street food do people from the UK like to eat?	kebabs: 2 burgers: 2 fish and chips: 2 ...	UK
中国人喜欢吃什么街头小吃？	烤肠 (roasted sausage): 3 烧烤 (barbecue): 2 糖葫芦 (candied haw): 1 ...	CN
¿Qué comida callejera les gusta comer a las personas de España?	churros (churros): 2 patatas fritas (French fries): 1 pipas (sunflower seeds): 1 ...	ES
¿Qué comida callejera les gusta comer a las personas de México?	tacos (tacos): 5 quesadillas (quesadillas): 3 tamales (tamales): 2 ...	MX
Makanan jalanan apa yang disukai oleh orang-orang dari Indonesia?	cilok (cilok): 3 bakso (meatball): 2 seblak (seblak): 1 ...	ID
대한민국 사람들은 어떤 길거리 음식을 좋아하나요?	떡볶이 (stir-fried rice cakes): 4 붕어빵 (bungeoppang): 1 델리만쥬 (delimanjoo): 1 ...	KR

Example in BLEnD: *"What street food do people like to eat?"*

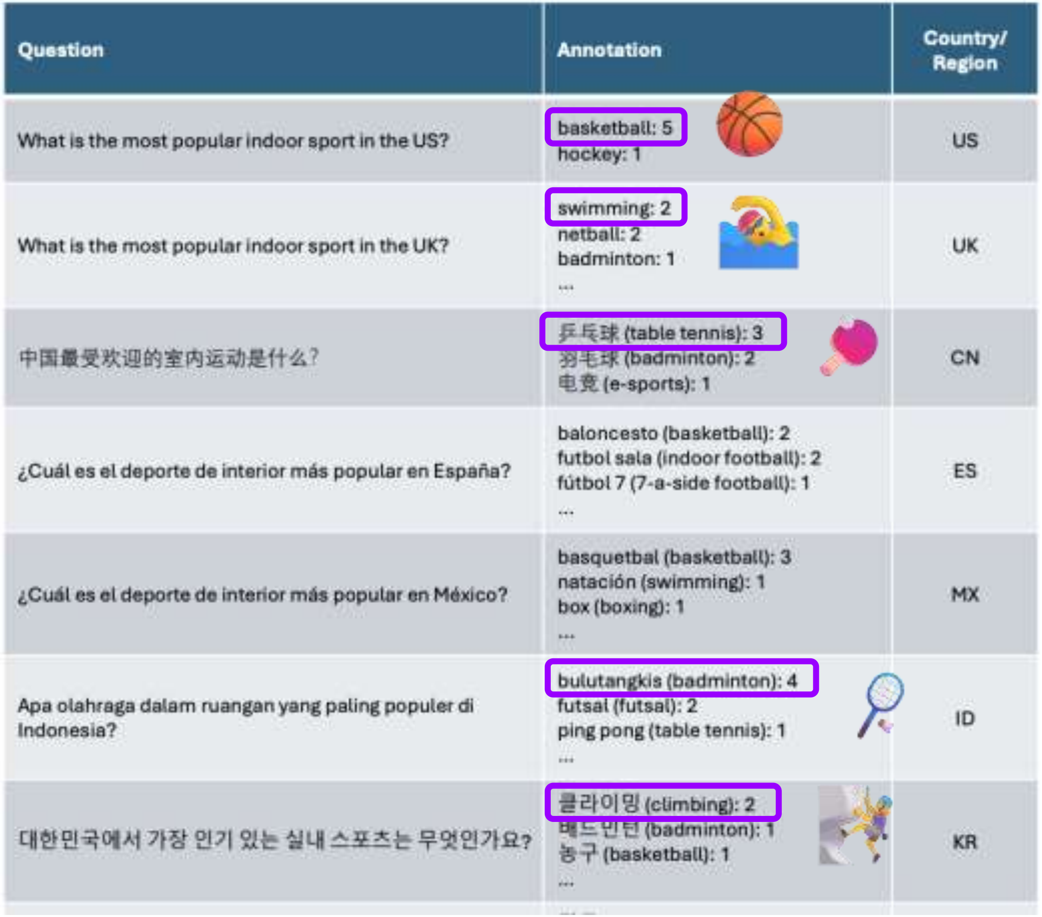
Answers for this simple question vary a lot across countries!



Question	Annotation	Country/ Region
What street food do people from the US like to eat?	hot dogs: 4 hamburger: 1 tacos: 1 ...	US
What street food do people from the UK like to eat?	kebabs: 2 burgers: 2 fish and chips: 2 ...	UK
中国人喜欢吃什么街头小吃？	烤肠 (roasted sausage): 3 烧烤 (barbecue): 2 糖葫芦 (candied haw): 1 ...	CN
¿Qué comida callejera les gusta comer a las personas de España?	churros (churros): 2 patatas fritas (French fries): pipas (sunflower seeds): 1 ...	ES
¿Qué comida callejera les gusta comer a las personas de México?	tacos (tacos): 5 quesadillas (quesadilla tamales (tamales): 2 ...	MX
Makanan jalanan apa yang disukai oleh orang-orang dari Indonesia?	cilok (cilok): 3 bakso (meatball): 2 seblak (seblak): 1 ...	ID
대한민국 사람들은 어떤 길거리 음식을 좋아하나요?	떡볶이 (stir-fried rice cakes): 4 붕어빵 (bungeoppang): 1 델리만쥬 (delimanjoo): 1 ...	

Example in BLEnD: *"What is the most popular indoor sport?"*

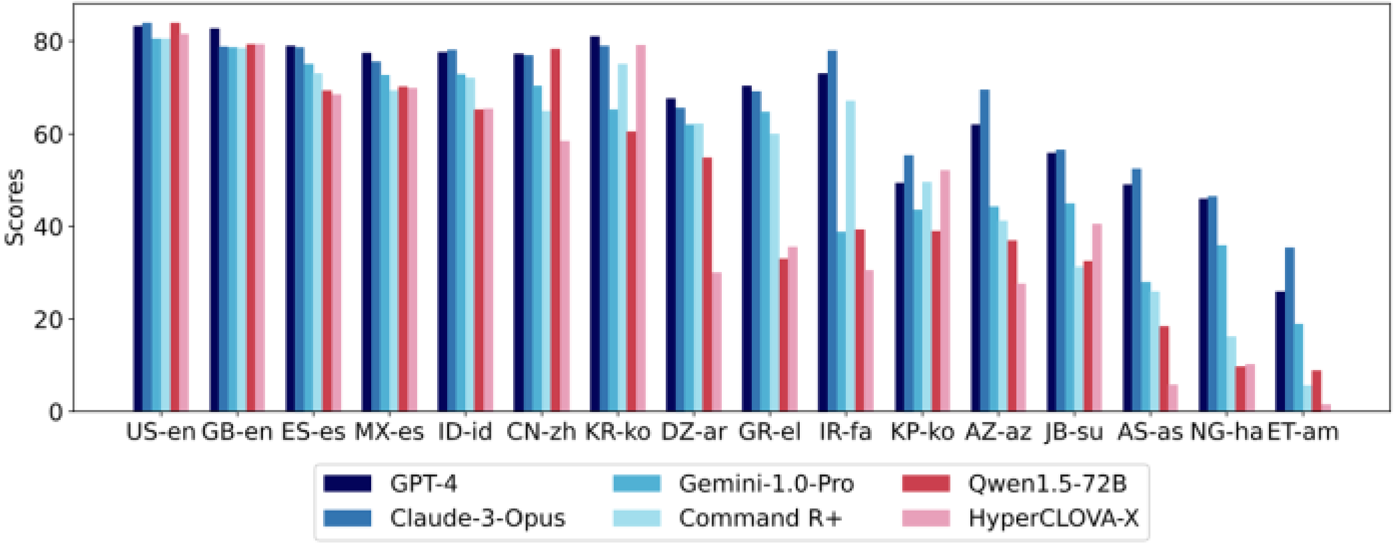
Answers for this simple question vary a lot across countries!



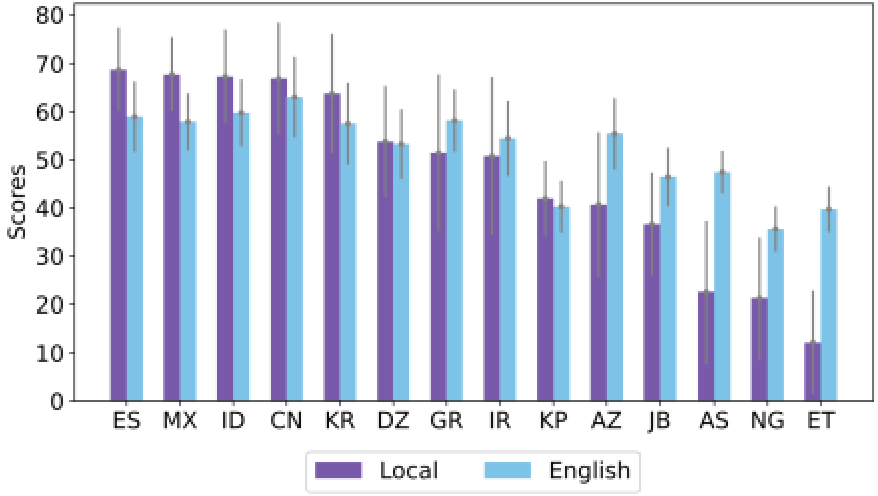
Question	Annotation	Country/ Region
What is the most popular indoor sport in the US?	basketball: 5 hockey: 1 🏀	US
What is the most popular indoor sport in the UK?	swimming: 2 netball: 2 badminton: 1 ... 🏊	UK
中国最受欢迎的室内运动是什么?	乒乓球 (table tennis): 3 羽毛球 (badminton): 2 电竞 (e-sports): 1 🏓	CN
¿Cuál es el deporte de interior más popular en España?	baloncesto (basketball): 2 futbol sala (indoor football): 2 fútbol 7 (7-a-side football): 1 ...	ES
¿Cuál es el deporte de interior más popular en México?	basquetbal (basketball): 3 natación (swimming): 1 box (boxing): 1 ...	MX
Apa olahraga dalam ruangan yang paling populer di Indonesia?	bulutangkis (badminton): 4 futsal (futsal): 2 ping pong (table tennis): 1 ... 🏸	ID
대한민국에서 가장 인기 있는 실내 스포츠는 무엇인가요?	클라이밍 (climbing): 2 배드민턴 (badminton): 1 농구 (basketball): 1 ... 🧗	KR

LLMs' Performance in Local Languages

Models show a significant **drop in performance for underrepresented cultures**, with a maximum performance difference of 57.3 percentage points between the US and Ethiopia



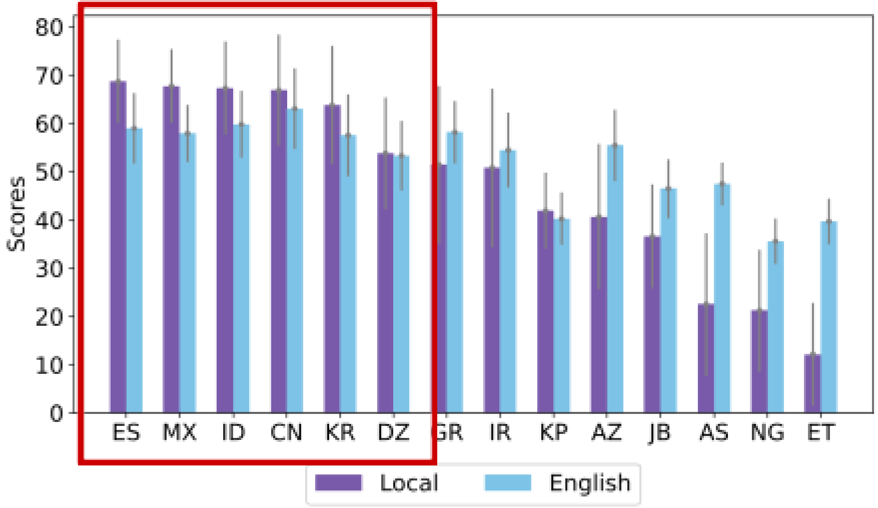
LLMs' Performance in Local Languages vs English



Average Score for All Models;
Models **prompted on English vs Local language** - same questions

LLMs’ Performance in Local Languages vs English

For **high-resource languages** like Spanish and Chinese, models showed **better performance when prompted with their local languages**

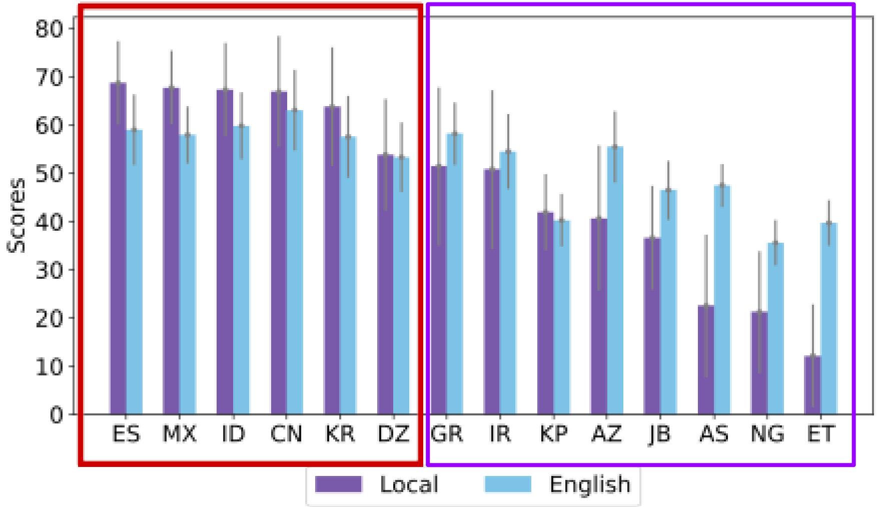


Average Score for All Models; Models **prompted on English vs Local language** - same questions

LLMs' Performance in Local Languages vs English

For **high-resource languages** like Spanish and Chinese, models showed **better performance when prompted with their local languages**

For **low-resource languages** like Azerbaijani, Sundanese, and Amharic, models generally showed **better performance when prompted in English**

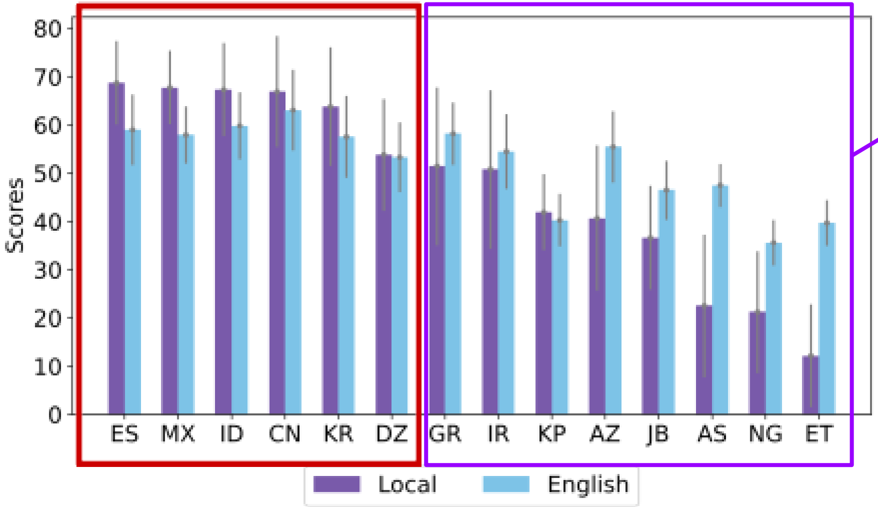


Average Score for All Models;
Models **prompted on English vs Local language** - same questions

LLMs’ Performance in Local Languages vs English

For **high-resource languages** like Spanish and Chinese, models showed **better performance when prompted with their local languages**

For **low-resource languages** like Azerbaijani, Sundanese, and Amharic, models generally showed **better performance when prompted in English**



💡 **Research idea!**

Analyse whether to prompt in one language or another for different tasks

Key Findings from Human Evaluation

Most **stereotypical responses** came from questions related to **food or festivals.**

LLMs often mentioned the **most famous** food item (e.g. Kimchi in Korea) or festival in response to completely unrelated questions.

Key Findings from Human Evaluation

Most **stereotypical responses** came from questions related to **food or festivals.**

LLMs often mentioned the **most famous** food item (e.g. Kimchi in Korea) or festival in response to completely unrelated questions.

Hallucinations were common for questions asking for a name or a title of an entity:

- ➢ For instance, the model answered 'Ruslan Cfrov' as the most famous basketball player in Azerbaijan, even though **no such player exists**
- ➢ Models occasionally **answered questions in a different language**, particularly for low-resource languages like Azerbaijani

LLM cultural awareness: What's next

SemEval 2026 Task 7:
Everyday Knowledge
Across Diverse Languages
and Cultures

New SemEval 2026 task based on BLEnD!

- ❖ We're doubling the number of languages and regions: **Japanese, Tamil, Basque, Bulgarian, Swedish** … and Danish soon?

LLM cultural awareness: What’s next

SemEval 2026 Task 7:
Everyday Knowledge
Across Diverse Languages
and Cultures

BLEnD

New **SemEval 2026 task** based on BLEnD!

- We’re doubling the number of languages and regions: **Japanese, Tamil, Basque, Bulgarian, Swedish** … and **Danish** soon?

Need to incorporate cultural perspectives in LLMs (same questions can be answered differently depending on culture/region).

Research on **LLM cultural alignment** - good progress recently: AlKhamissi et al. (ACL 2024), Li et al. (NeurIPS, 2024), Masoud et al. (COLING 2025) and others, including at EMNLP!

Also we need to better understand “hidden” implicit cultural biases of LLMs.

Open questions

How to **extend cultural framework** to other languages and cultures (beyond regions/countries and QA).

Open questions

How to **extend cultural framework** to other languages and cultures (beyond regions/countries and QA).

Other many **questions remain**, from the theoretical and practitioner perspectives:

- How to balance language abilities and **cultural awareness**?
- Do we need **multilingual or monolingual** LLMs for low-resource languages?
- Should we **prompt** in our native language or a high-resource one?
- How ***truly* multilingual** (and **multi-cultural**) are multilingual LMs?
- How to better **leverage cultural knowledge** in LLMs?

Open questions

How to **extend cultural framework** to other languages and cultures (beyond regions/countries and QA)

Other many **questions remain**, from the theoretical and practitioner perspectives:

- How to balance language abilities and **cultural awareness**?
- Do we need **multilingual or monolingual** LLMs for low-resource languages?
- Should we **prompt** in our native language or a high-resource one?
- How ***truly* multilingual** (and **multi-cultural**) are multilingual LMs?
- How to better **leverage cultural knowledge** in LLMs?

Interesting times for research in this area!

Gracias

Thank you!

Tak

@CamachoCollados