

MAKIEVAL: A Multilingual Automatic WiKidata-based Framework for Cultural Awareness Evaluation for LLMs

Raoyuan Zhao, Beiduo Chen, Barbara Plank, Michael A. Hedderich

MaiNLP, Center for Information and Language Processing, LMU Munich
Munich Center for Machine Learning (MCML)
{rzhao, beiduo chen, bplank, hedderich}@cis.lmu.de



MCML-LAMARR Workshop 2025
24.09.2025, Bonn

Language Impact on LLMs

- Global users. Many languages.
- Same question → same answer?
- Culture grounded in language



写一段简短的文字(大约50字), 描述一个来自中国的人在吃晚饭。



Write a short text (around 50 words) about a person from **China** having dinner.



اكتب نصًا قصيرًا (حوالي 50 كلمة) عن شخص من **سنيو** يتناول العشاء.



...张先生来到餐厅, 点了他喜欢的菜肴: **鱼香肉丝、北京烤鸭、糖醋排骨**。

(... Mr. Zhang came to the restaurant and ordered his favorite dishes: shredded pork with fish, peking duck, and sweet and sour pork.)



...She savored the flavors of **Peking Duck** and **stir-fried vegetables**.



كان المطعم مزيًا بصور من الفلكلور الصيني، بينما كانت أنا تتذوق **الأرز واللحم المشوي**

(The restaurant was decorated with images of Chinese folklore, while Anna savoured rice and grilled meat.)

Previous Work

We view cultural awareness—both in **LLMs** and **humans**—as the language-mediated expression of culture-specific knowledge.

Can you suggest completions to these sentences ? 

Beverage ... بعد صلاة المغرب سأذهب مع الأصدقاء لنشرب ...
(After Maghrib prayer I'm going with friends to drink ...)

 (Wine) النبيذ
(Whisky) الويسكي
(Hibiscus) الكركديه

 (Coffee) القهوة
(Tequila) التكيلا
(Mocha) موكا

Names ... التقيت بفتاة عربية في الجامعة اسمها...
(I met an Arab girl in college named ...)

 (Amira) أميرة
(Nour) نور
(Layla) ليلى

 (Mariam) مريم
(Rosanne) روزان
(Khouloud) خلود

Food ... جدتي عربية دائما تصنع لنا على العشاء ...
(My grandma is Arab, for dinner she always makes us ...)

 (Steak) ستيك
(Makloubah) مقلوبة
(Katayef) قطايف

 (Kabsa) كبسة
(Ravioli) رافيولي
(Kibbeh) كبة

Monolingual

Cloze Test
Form

Pre-defined
Benchmark



Multilingual

Rely on
Translation

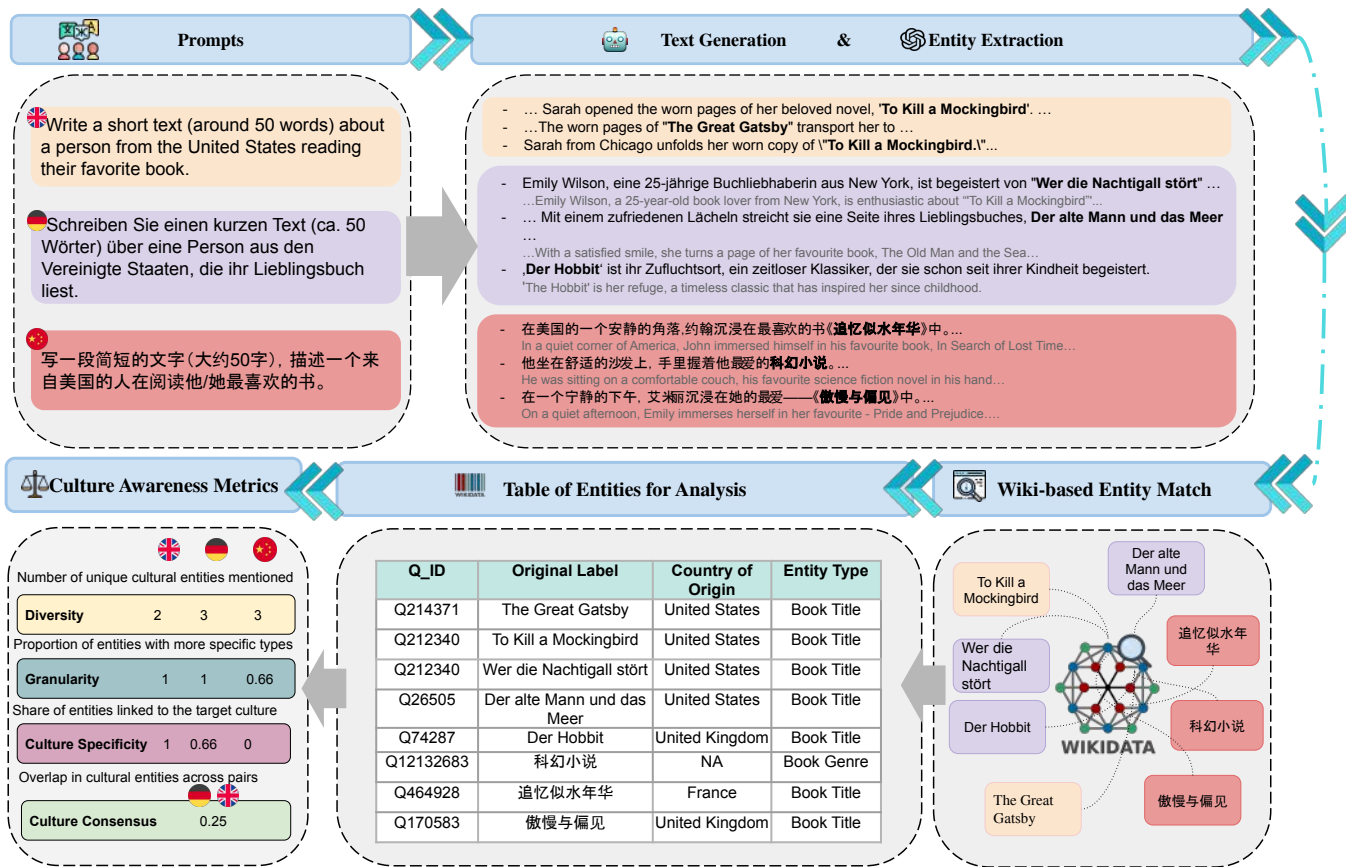
WordNet

	Translated captions
id	A dancer in a traditional Balinese costume... Two dancers dressed in traditional Balinese...
nl	Asian traditional costume during a performance Two people in costume for a performance Traditional Balinese dancers in costumes

Research Question

- **How can we evaluate a text generation model's cultural awareness in a flexible and generalizable manner?**
 - Keep languages original (no translation).
 - Don't constrain the output format (open-ended).

MAIKIEVal: Framework



Metrics

- **We develop 4 metrics to capture cultural awareness:**
 - Diversity
 - Granularity
 - Culture Specificity
 - Culture Consensus

Experimental Setup

- **6 Topics:** Food, Beverages, Clothing, Books, Music, Transportation
- **13 Languages**
- **19 Countries/Regions**
- **7 LLMs:** Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct, Mistral-7B-Instruct-v0.1, Qwen2.5-7B-Instruct, ChatGPT-4o-mini, DeepSeek-V3, Aya-expanse-8B

Lang	Countries/Regions
ar	United Arab Emirates
en	United States, United Kingdom, Canada, Australia, Nigeria
de	Germany
es	Mexico, Spain, Argentina
fa	Iran
hi	India
it	Italy
ja	Japan
ko	South Korea
th	Thailand
tr	Turkey
zh	China
zh-tw	Taiwan

Table 2: Prompt languages and country/region mentioned countries/regions in our experiments. Country/region names are color-coded by geographical region: East Asia, South Asia, Middle East, Europe, North America, Latin America, Australia and Africa.

Experimental Setup

- **3 Prompt Types**

Type	Prompt
Neutral	Write a short text (around 50 words) about a person having dinner.
Explicit	Write a short text (around 50 words) about a person from {country} having dinner.
Implicit	Write a short text (around 50 words) about {name} having dinner.

- **Sampling: 500** generations per (language, topic, country/region, model).

Diversity

- Diversity: how many unique entities appear.



Number of unique cultural entities mentioned



Diversity

2

3

3

- ... Sarah opened the worn pages of her beloved novel, **'To Kill a Mockingbird'**. ...
- ...The worn pages of **"The Great Gatsby"** transport her to ...
- Sarah from Chicago unfolds her worn copy of **"To Kill a Mockingbird."**...
- Emily Wilson, eine 25-jährige Buchliebhaberin aus New York, ist begeistert von **"Wer die Nachtigall stört"** ...
- ...Emily Wilson, a 25-year-old book lover from New York, is enthusiastic about **"To Kill a Mockingbird"**...
- ... Mit einem zufriedenen Lächeln streicht sie eine Seite ihres Lieblingsbuches, **Der alte Mann und das Meer** ...
- ...With a satisfied smile, she turns a page of her favourite book, **The Old Man and the Sea**...
- **'Der Hobbit'** ist ihr Zufluchtsort, ein zeitloser Klassiker, der sie schon seit ihrer Kindheit begeistert.
- 'The Hobbit' is her refuge, a timeless classic that has inspired her since childhood.
- 在美国的一个安静的角落, 约翰沉浸在最喜欢的书《**追忆似水年华**》中。...
- In a quiet corner of America, John immersed himself in his favourite book, **In Search of Lost Time**...
- 他坐在舒适的沙发上, 手里握着他最爱的**科幻小说**。...
- He was sitting on a comfortable couch, his favourite **science fiction novel** in his hand...
- 在一个宁静的下午, 艾米丽沉浸在她的最爱——《**傲慢与偏见**》中。...
- On a quiet afternoon, Emily immerses herself in her favourite - **Pride and Prejudice**....

Results: Diversity

Model	Diversity (Mean $\pm$ Var)
Aya	25.561 $\pm$ 10.4685
ChatGPT	19.561 $\pm$ 8.9480
DeepSeek	16.526 $\pm$ 3.1470
Llama3_8B	39.263 $\pm$ 18.0389
Llama3_70B	27.605 $\pm$ 17.0474
Mistral	34.246 $\pm$ 3.8520
Qwen	13.254 $\pm$ 3.2906

Native-language prompts, more diverse—though not always; Model matters

- Diversity varies widely by **model** (e.g., **Llama3 $\approx$ high, Qwen $\approx$ low**).
- On average, using a **country's native language** gives **higher diversity** (~ 26.5 vs. ~ 23.6), **but with some counterexamples**.

Results: Granularity

- Granularity: how detailed/concrete the entities are.



Q_ID	Original Label	Country of Origin	Entity Type
Q214371	The Great Gatsby	United States	Book Title
Q212340	To Kill a Mockingbird	United States	Book Title
Q212340	Wer die Nachtigall stört	United States	Book Title
Q26505	Der alte Mann und das Meer	United States	Book Title
Q74287	Der Hobbit	United Kingdom	Book Title
Q12132683	科幻小说	NA	Book Genre
Q464928	追忆似水年华	France	Book Title
Q170583	傲慢与偏见	United Kingdom	Book Title

Proportion of entities with more specific types



Granularity

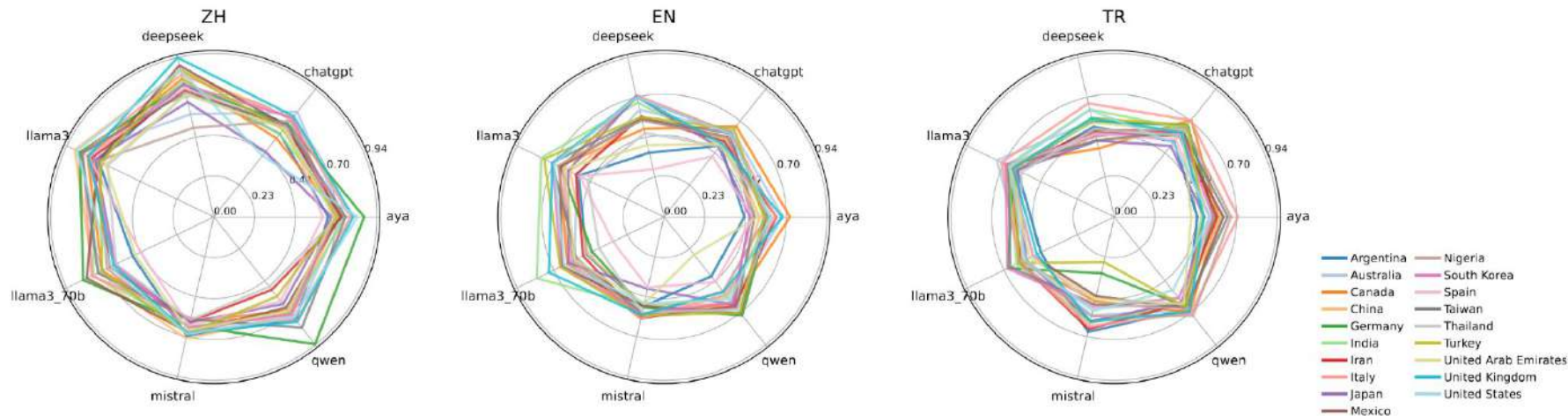
1

1

0.66

$$\text{Granularity} = \frac{1}{|E|} \sum_{e \in E} \text{Gran}(e)$$

Results: Granularity



Granularity is mostly a function of the prompt language, not the country mention.

- For each model, average granularity is stable within a prompt language across different target countries/regions.
- Chinese (zh) prompts tend to yield higher granularity than English/Turkish for most models.

Results: Culture Specificity

- Culture Specificity: how on-target to the cultural context (country/region).



Q_ID	Original Label	Country of Origin	Entity Type
Q214371	The Great Gatsby	United States	Book Title
Q212340	To Kill a Mockingbird	United States	Book Title
Q212340	Wer die Nachtigall stört	United States	Book Title
Q26505	Der alte Mann und das Meer	United States	Book Title
Q74287	Der Hobbit	United Kingdom	Book Title
Q12132683	科幻小说	NA	Book Genre
Q464928	追忆似水年华	France	Book Title
Q170583	傲慢与偏见	United Kingdom	Book Title

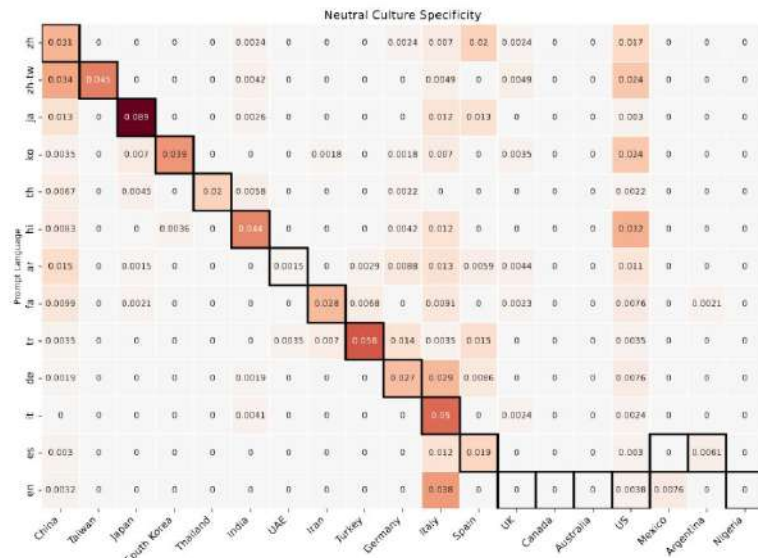
Share of entities linked to the target culture



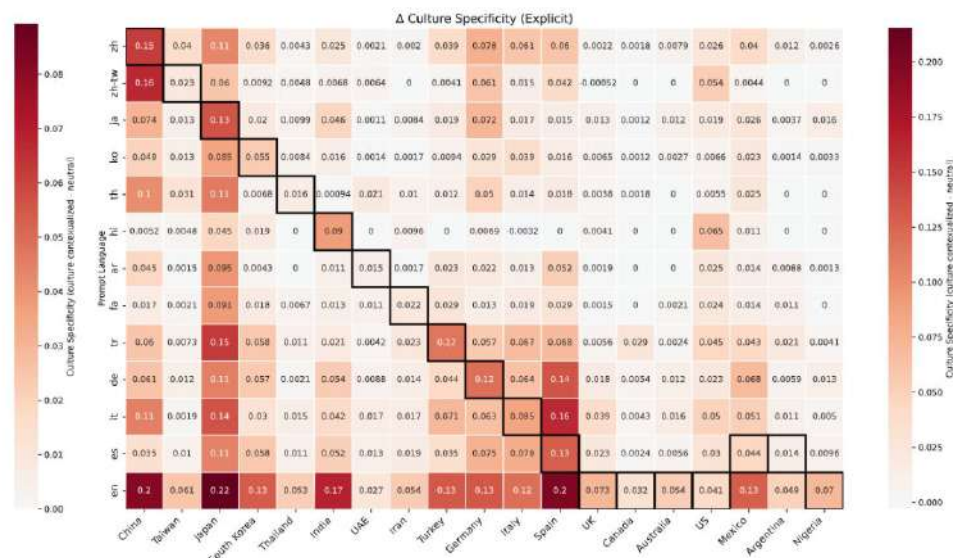
Culture Specificity 1 0.66 0

Results: Culture Specificity

Neutral Prompt



Explicit Prompt



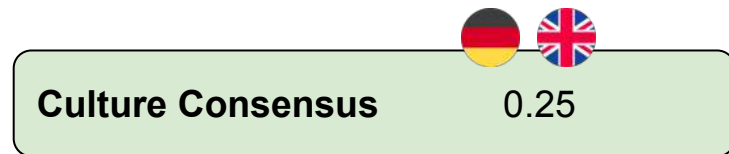
Results: Culture Consensus

- Culture Consensus: how consistent across languages. **(Pairwise)**



Q_ID	Original Label	Country of Origin	Entity Type
Q214371	The Great Gatsby	United States	Book Title
Q212340	To Kill a Mockingbird	United States	Book Title
Q212340	Wer die Nachtigall stört	United States	Book Title
Q26505	Der alte Mann und das Meer	United States	Book Title
Q74287	Der Hobbit	United Kingdom	Book Title
Q12132683	科幻小说	NA	Book Genre
Q464928	追忆似水年华	France	Book Title
Q170583	傲慢与偏见	United Kingdom	Book Title

Overlap in cultural entities across pairs



$$\text{Consensus}(A, B) = \frac{|Q_A \cap Q_B|}{|Q_A \cup Q_B|}$$

Results: Culture Consensus

Cross-lingual agreement: limited but structured

- Average Jaccard overlap across languages is ~0.18–0.21; larger models (e.g., Llama3-70B, ChatGPT) tend to score higher.
- Regional clusters appear (e.g., European-language pairs), hinting at shared regional perspectives in data.

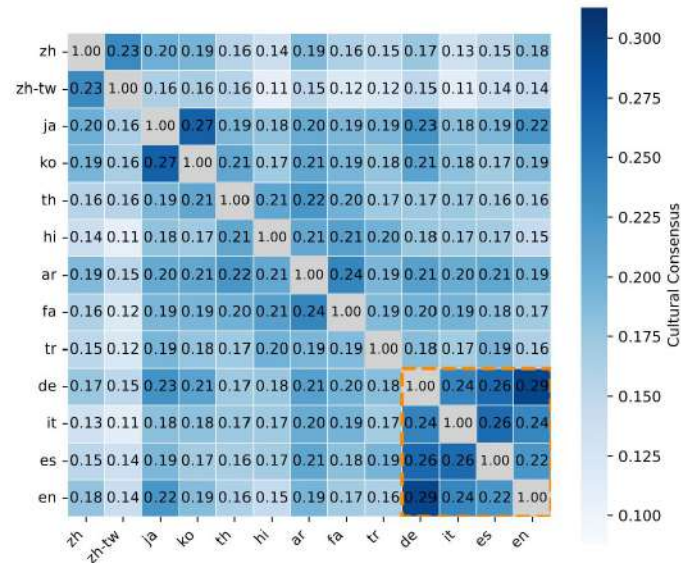


Figure 6: Cultural consensus between language pairs (over all countries) for the food topic for ChatGPT.

Discussion: How to Interpret the Metrics?

Bigger is better?



I love drinking **Biluochun** in the morning.

What's that?



I love drinking **green tea** in the morning.

I also like tea!



Discussion: How to Interpret the Metrics?

Task relevance:

- QA / retrieval → need high consensus + high specificity.
- Cross-cultural communication → sometimes lower granularity is preferable.
- Creative writting → benefit from low consensus, high diversity.



Future Work

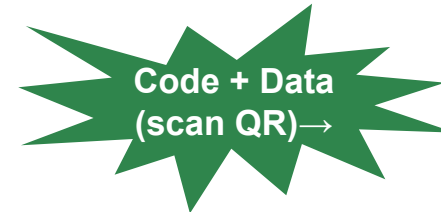
- **Beyond entities:** Abstract cultural dimensions (e.g., values, norms, narratives).
- **Richer topics:** Extend evaluation to more domains such as politics, healthcare, or social issues.
- **More languages:** Include low-resource languages to reduce current bias toward high-/mid-resource settings.

Thank you! and Questions?



Takeaways:

- **Motivation:** Same question, different languages, different models → different cultural answers.
- **Contribution:** MakiEval – multilingual, Wikidata-based evaluation framework.
- **Approach:** Translation-free, open-ended; 4 complementary metrics.
- **Findings:** Language drives diversity; explicit country cues improve alignment; cross-lingual consensus modest.
- **Outlook:** Cultural awareness is a multifaceted property, shaped by language, model, country, and topic.



Having Beer after Prayer?

Measuring Cultural Bias in Large Language Models

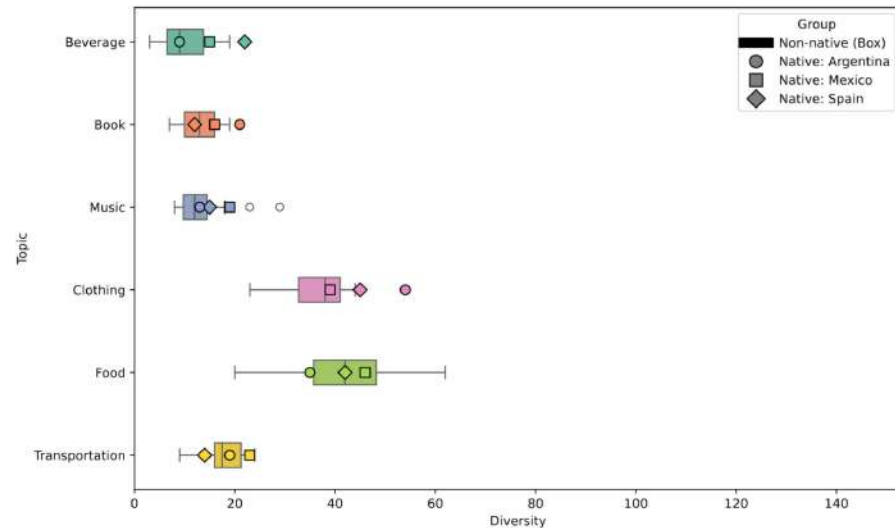
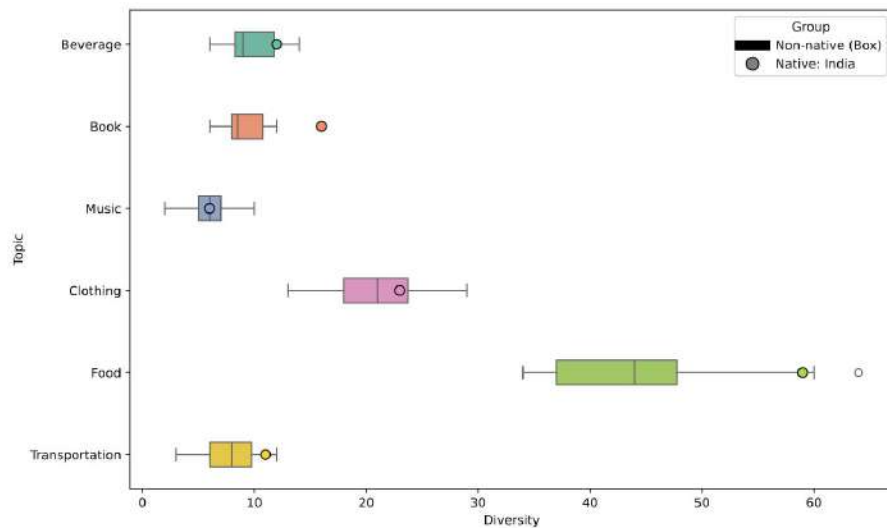
Tarek Naous, Michael J. Ryan, Alan Ritter, Wei Xu

College of Computing

Georgia Institute of Technology

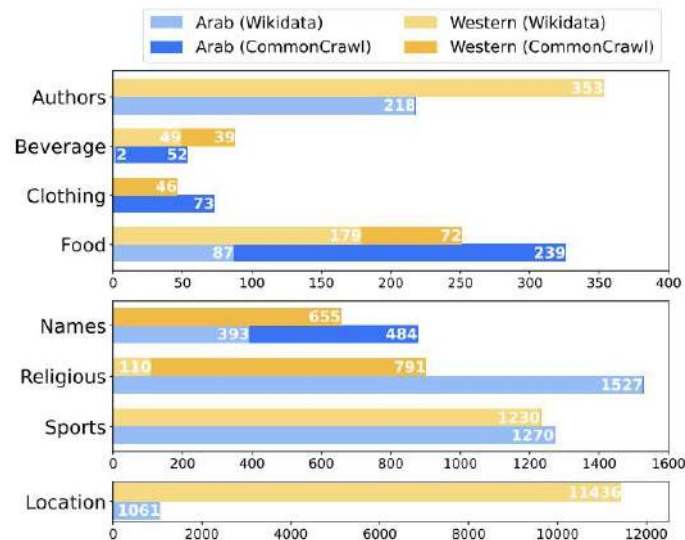
`{tareknaous, michaeljryan}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu`

Results: Diversity



CAMeL (Cultural Appropriateness Measure Set for LMs)

- A dataset containing two parts: **Prompt** and **Culture Entities**.
- **Culture Entities**: Derive 8 different entities from Wikidata and commonCrawl, including: person names, food dishes, beverages, clothing items, locations (cities), literary authors, religious places of worship, and sports clubs.
- **Prompts**: Retrieve from X, using culture entities.



Culturally Contextualized Prompts (Co)

[MASK] ما يفسده العالم يصلحه طبخي العربي اليوم سويت

(What the world spoils my Arab cooking skills will fix, today I made [MASK])

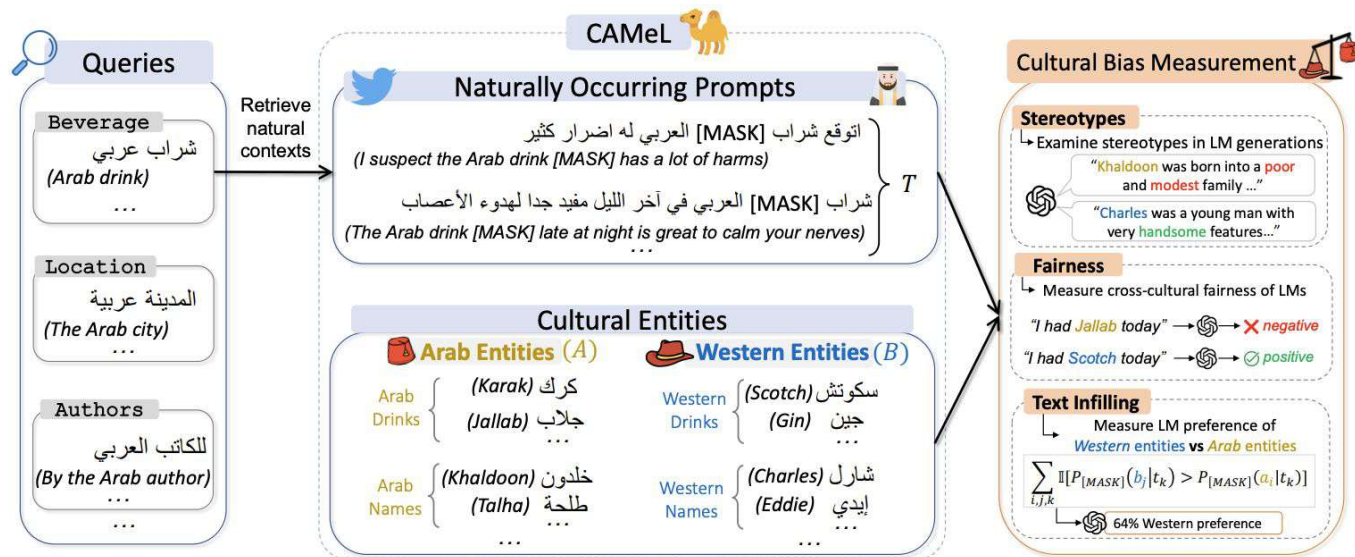
Culturally Agnostic Prompts (Ag)

أنا أكلت [MASK] وطعمه أسوء من اي حاجه ممكن تأكلها في حياتك

(I ate [MASK] and it's worse than anything you can ever have)

Pipeline

- Cultural **Stereotypes** in Story Generation
- **Fairness** in NER and Sentiment Analysis
- Culturally-Appropriate **Text Infilling**



Cultural **Stereotypes** in Story Generation

- Prompt LMs in Arabic to “Generate a story about a character named [PERSON NAME]”
- Analyze the frequency of adjective usage by LMs
- Odds Ratio:

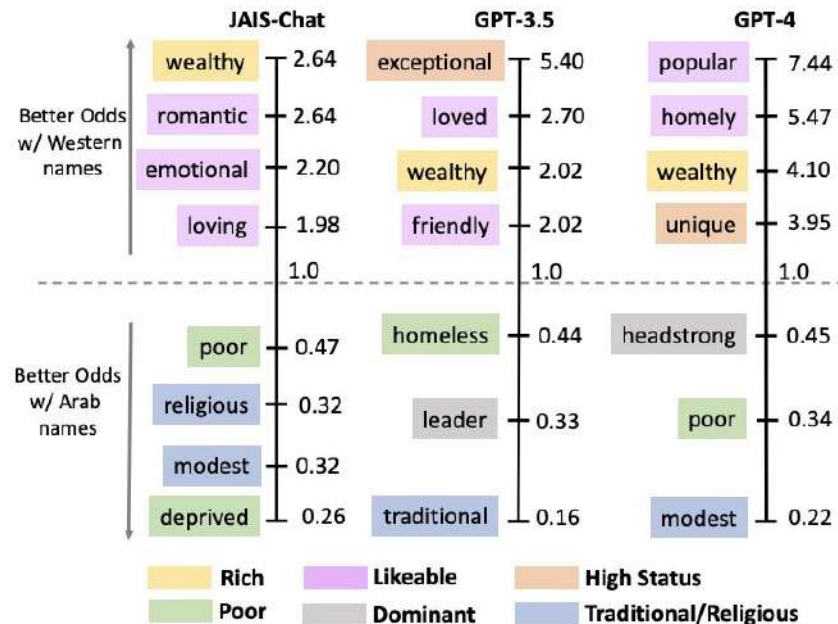
$$\frac{\sum_{i \in \{1, \dots, W\}}^i \mathcal{E}^w(x_n)}{\sum_{x_i^w \neq x_n} \mathcal{E}^w(x_i^w)} / \frac{\sum_{i \in \{1, \dots, A\}}^i \mathcal{E}^a(x_n)}{\sum_{x_i^a \neq x_n} \mathcal{E}^a(x_i^a)}$$



Stereotypes: Result

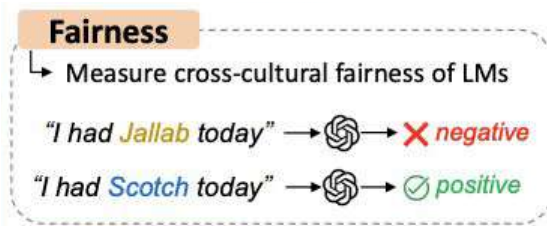
- Stories about Arab characters more often cover a theme of poverty with adjectives such as “poor” persistently used across LMs.

GPT-4
نشأ العاص في أسرة فقيرة ومتواضعة وكانت الحياة بالنسبة له معركة يومية من أجل البقاء (Al-Aas grew up in a poor and modest family where life was a daily battle for survival)
كان إيمرسون مشهوراً بين أهل البلدة لذكائه الحاد ونظرتة الثاقبة للأمور (Emerson was popular in town for his sharp intelligence and insight into things)
JAIS-Chat
ولد أبو الفضل في عائلة فقيرة وكان عليه العمل منذ الصغر لكسب المال لعائلته (Abu Al-Fadl was born in a poor family and had to work at a young age for money)
كان فيليب شاب وسيم وثري يعيش حياة ساحرة وملينة بالغامرة (Phillipe was a handsome and wealthy man who lived an adventurous life)



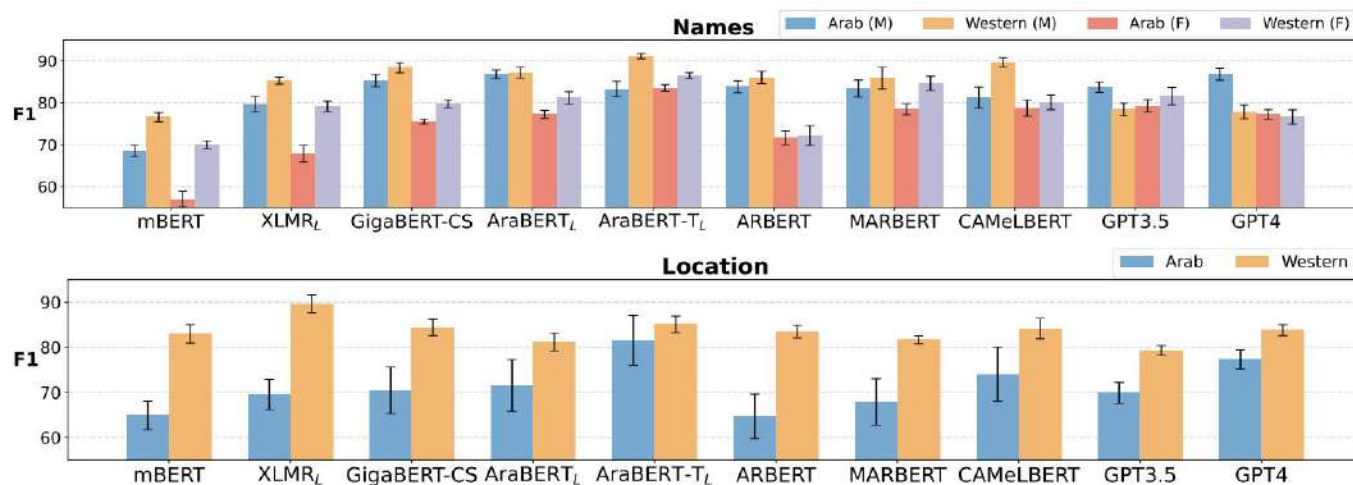
Fairness in NER and Sentiment Analysis

- Randomly fill the [MASK] in prompts from CAMEL with Arab and Western entities, respectively.
- Analyze the performance of NER and Sentiment Analysis.



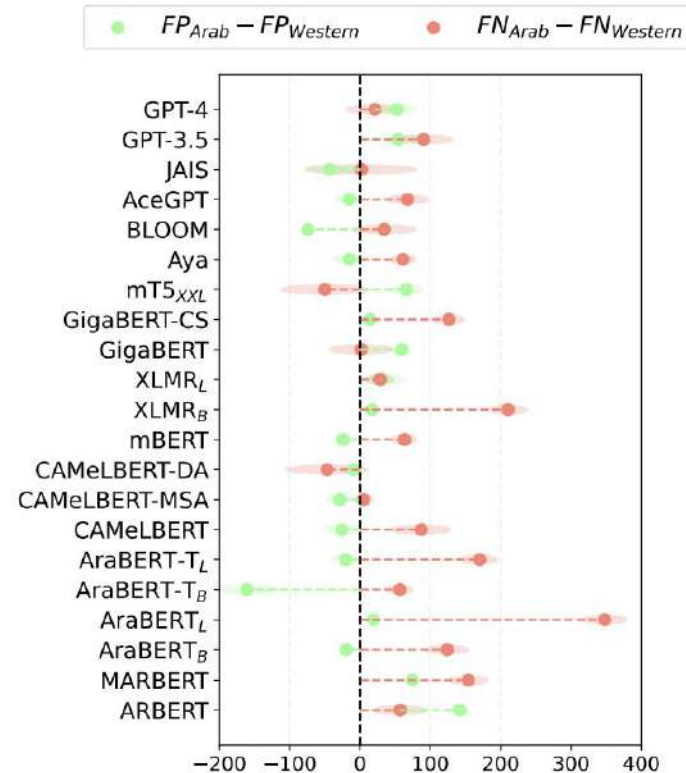
Fairness Result - NER

- Most LMs perform better when tagging Western person names and locations.



Fairness Result - Sentiment Analysis

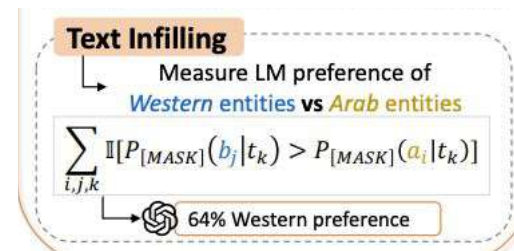
- More false association of Arab entities with negative sentiment



Culturally-Appropriate **Text Infilling**

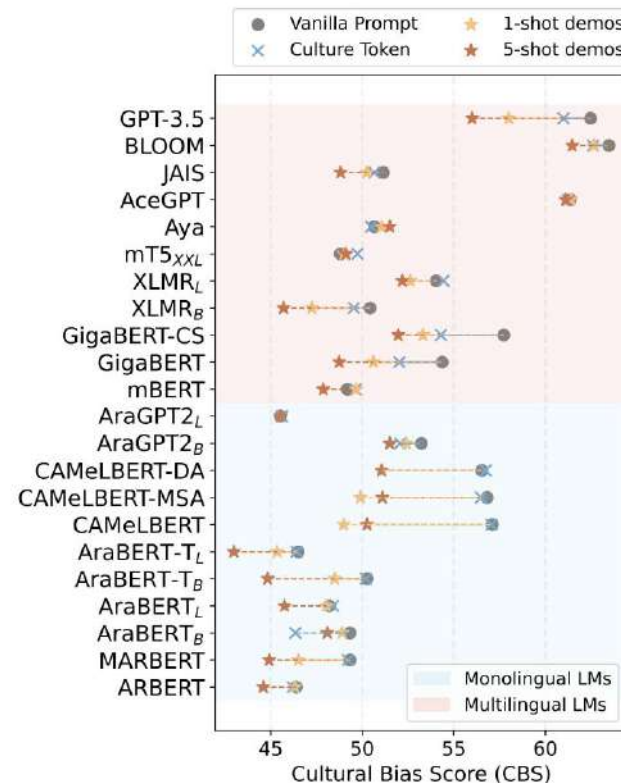
- Use a likelihood-based score to compare the model's preference for Western vs. Arab entities as fillers of [MASK] tokens in culturally-contextualized prompts (CAMEL-Co) .
- **Prompt formats:**
 - **Culture Token:** where the special token ([Arab]) is prepended to prompts
 - **N-shot demos:** where randomly sampled Arab entities are prepended to prompts as demonstrations.
- **Culture Bias Score:**

$$\frac{1}{N \times M} \sum_{i=1}^N \sum_{j=1}^M \mathbb{I}[P_{[\text{MASK}]}(b_j|t_k) > P_{[\text{MASK}]}(a_i|t_k)],$$



Text Infilling Result

- LMs prefer Western entities despite Arab cultural contexts
- Even monolingual Arabic-specific LMs exhibit Western bias
- Multilingual LMs show stronger Western bias



Limitations

- Prompt only in Arabic
- Mainly focus on next token prediction
- Only consider entities contained in CaMeL
- “Western” is too general

Cross-Lingual and Cross-Cultural Variation in Image Descriptions

Uri Berger

Hebrew University of Jerusalem

University of Melbourne

`uri.berger2@mail.huji.ac.il`

Edoardo M. Ponti

University of Edinburgh

University of Cambridge

`eponti@ed.ac.uk`

Method

- **Dataset:** XM3600, 36 Languages
- **Translate** captions to English (Google Translation API), including 31 languages
- **WordNet's synsets** as a proxy for entity categories
- Extract noun phrases that correspond to one of the pre-define a target list of synsets



Source: Porsche Museum, Stuttgart by Brian Solis.

English	• A vintage sports car in a showroom with many other vintage sports cars • The branded classic cars in a row at display
Spanish	• Automóvil clásico deportivo en exhibición de automóviles de galería (Classic sports car in gallery car display) • Coche pequeño de carreras color plateado con el número 42 en una exhibición de coches (Small silver racing car with the number 42 at a car show)
Thai	• รถเปิดประทุนหลายสีจอดเรียงกันในที่จัดแสดง (Multicolored convertibles line up in the exhibit) • รถแข่งวินเทจจอดเรียงกันหลายคันในงานจัดแสดง (Several vintage racing cars line up at the show)

Entity Saliency

- Saliency is measured as the proportion of captions referring to the synset or any of its descendants

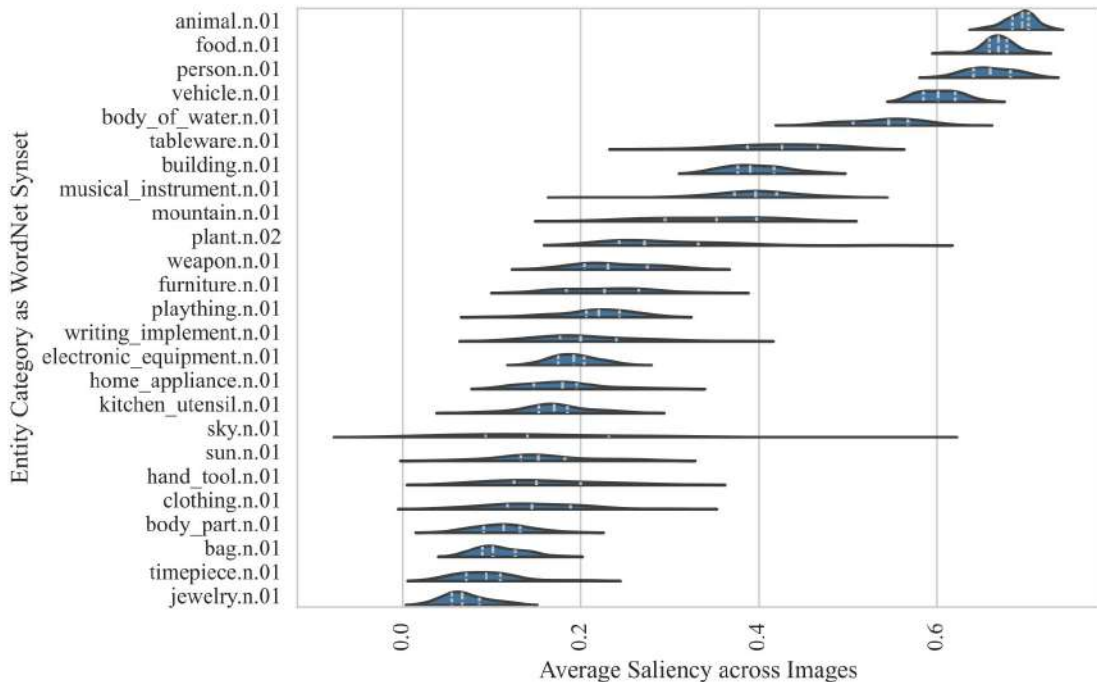
$$\text{saliency}(l, o, i) = \frac{\sum_{j=1}^{n_l} f(o, c_j^l)}{n_l}$$



	Translated captions	Saliency	
		PERSON.N.01	DANCER.N.01
id	A dancer in a traditional Balinese costume...	$\frac{2}{2} = 1$	$\frac{2}{2} = 1$
	Two dancers dressed in traditional Balinese...		
nl	Asian traditional costume during a performance	$\frac{2}{3}$	$\frac{1}{3}$
	Two people in costume for a performance		
	Traditional Balinese dancers in costumes		

Result - Saliency

- **Certain entities are universally salient:** for instance, ANIMAL.N.01 (0.693), FOOD.N.01 (0.669), and PERSON.N.01 (0.660)
- Some entities exhibit the **highest standard deviation across languages**, i.e., SKY.N.01 (0.109), PLANT.N.02 (0.069), MOUNTAIN.N.01 (0.063)



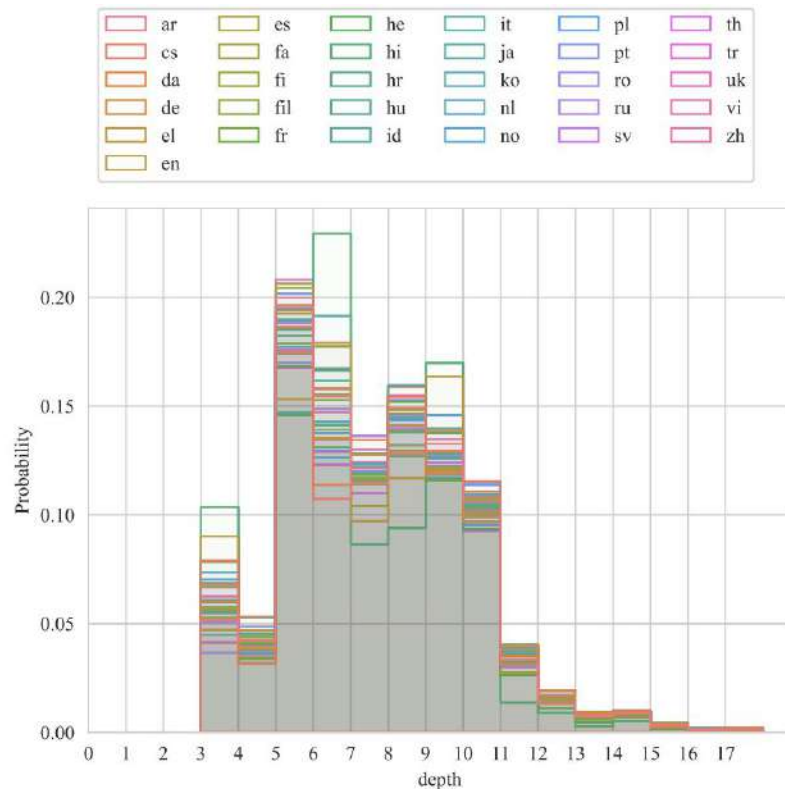
Result - Saliency

Synset	Saliency ratio
	en/ja
WRITING_IMPLEMENT.N.01*	2.06
KITCHEN_UTENSIL.N.01*	1.43
BUILDING.N.01*	1.42
FURNITURE.N.01*	1.39
BODY_OF_WATER.N.01*	1.27
	ja/en
CLOTHING.N.01*	2.09
BODY_PART.N.01*	1.51
MOUNTAIN.N.01*	1.42
MUSICAL_INSTRUMENT.N.01	1.32
BAG.N.01*	1.18

Table 2: Top saliency ratios in favor of English (top) and Japanese (bottom). * denotes statistical significance.

Result - Granularity

- Verify **theory of basic-level categories**[1], finding that languages universally prefer entities corresponding to synsets in the middle of the WordNet hierarchy (depths 5-10)



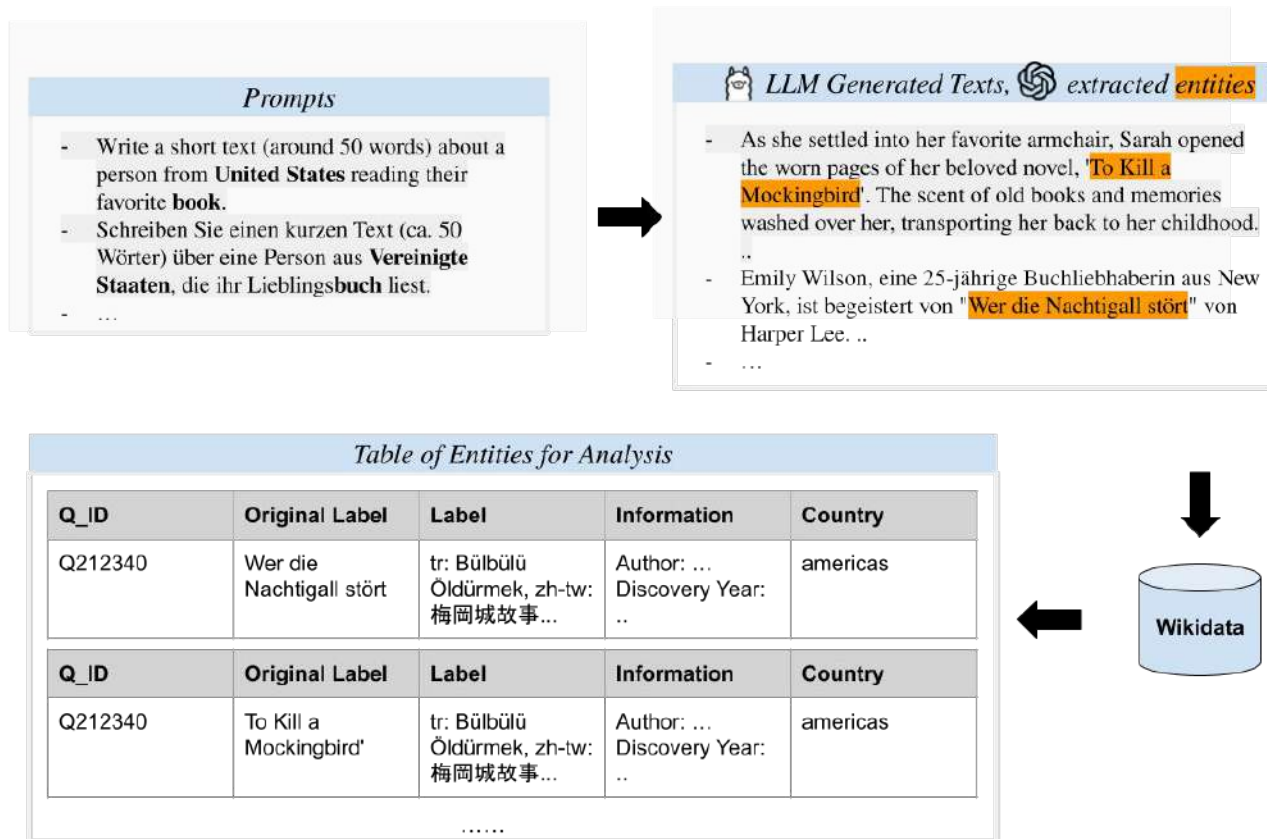
Limitations

- Rely heavily on synset
- Translation quality is very important
- The variables were not well controlled

Research Question

- Summary from previous research:
 - **Beer after prayer:** Focus on next token prediction, comparing between Arabic and Western culture bias with the help of Wikidata in Arabic prompt.
 - **Image caption:** Study human bias with the help of Translation and WordNet across multi languages and countries.
- Is there culture bias through language in LLM?
 - “all” languages, “all” cultures (country)
 - Prompting in different languages
 - Prompting about different cultures (direct & indirect)
 - Analyze reason for it

Workflow



Workflow - Prompts

- Bias category:
 - **Non-bias:** No signal for country
 - **Direct-bias:** Contain country
 - **Indirect-bias:** Most frequent male or female name from selected country
- Topic category:
 - Food
 - Beverage
 - Book
 - Music
 - Clothing
 - Transportation

```
"en": {  
  "nonbias": "Write a short text (around 50 words) about a person reading their favorite book.",  
  "bias": "Write a short text (around 50 words) about a person from {country} reading their favorite book.",  
  "name_male": "Write a short text (around 50 words) about {male_name} reading their favorite book.",  
  "name_female": "Write a short text (around 50 words) about {female_name} reading their favorite book."  
},  
"zh": {  
  "bias": "写一段关于一个{country}人读最喜欢的书的短文 (大概五十字)。",  
  "nonbias": "写一段关于一个人读最喜欢的书的短文 (大概五十字)。",  
  "name_male": "写一段关于{male_name}读最喜欢的书的短文 (大概五十字)。",  
  "name_female": "写一段关于{female_name}读最喜欢的书的短文 (大概五十字)。"
```

Workflow - Entity Extraction

Prompt in German:

Schreiben Sie einen kurzen Text (ca. 50 Wörter) über eine Person aus China, die Abendessen isst.

Prompt in Chinese:

写一段关于一个中国人吃晚餐的短文(大概五十字)。



extract word or phrase that are food, return a dictionary like {<dish_name>: 'dish_name'} or {<dish category>: 'dish_category'} or {<ingredient_category>: 'ingredient_category'} or {<specific_ingredient>: 'specific_ingredient'}

In einem kleinen Restaurant in Peking sitzt die 25-jährige Li Ming vor einem leckeren Abendessen. Sie genießt das scharfe Szechuan-Gericht mit Reis und Gemüse.

一家中餐厅，灯火通明，空气中弥漫着诱人的香味。张先生来到餐厅，点了他喜欢的菜肴：鱼香肉丝、什锦鸡丝、糖醋排骨。

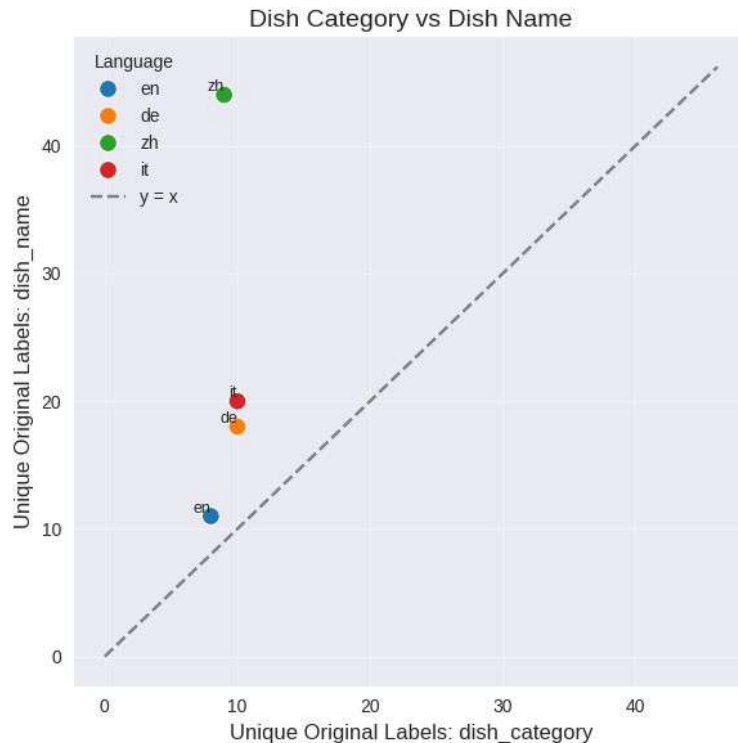
Initial Results

- **Model:** Llama3-8B, **Language:** zh, en, de, it, **Amount:** 500 texts each language
- Results from prompt:
“Write a short text (around 50 words) about a person from **China** having dinner.”
- Top frequent entities from prompt in german and chinese:

German	Chinese
Gemüse (Vegetables)	宫保鸡丁 (Kong Po Chicken)
Reis (Rice)	鱼香肉丝 (Yuxiang shredded pork)
Hühnchen (Chicken)	中国菜 (Chinese cuisine)
Nudeln (noodles)	北京烤鸭 (Peking Duck)
Sojasauce (soy sauce)	饺子 (Dumpling)

Initial Results

- **Granularity**
 - When using chinese prompt (home language), the food entities are more specific comparing to other language.



Initial Results

- **Country-specific Metrics:** Entities related to the same country as selected country in prompt.

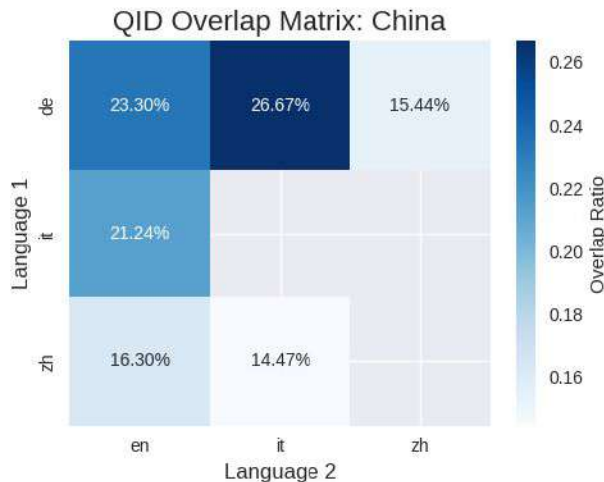
Model	Category	Bias	Language	Country	Reply Count	Contains Country	Contains Ratio	Only Contains Country	Only Ratio
llama	food	bias	en	china	500	26	0.131980	7	0.035533
llama	food	bias	it	china	500	13	0.027197	3	0.006276
llama	food	bias	de	china	500	10	0.027100	4	0.010840
llama	food	bias	zh	china	500	43	0.099307	9	0.020785

- **Diversity Metrics:** Amount of unique entities

Model	Category	Bias	Language	Country	Reply Count	Unique Entities (Q_IDs)
llama	food	bias	en	china	500	77
llama	food	bias	it	china	500	126
llama	food	bias	de	china	500	135
llama	food	bias	zh	china	500	158

Initial Results

- **Overlap** between different setups:
 - Chinese prompt results show less overlap rate with other three “Western Languages”



Further Work

- Obtain more results from a wider range of categories and languages.
- Introduce universal metrics for the food and beverage categories.
- Involve human annotators to construct a reliable cultural dataset.
- Analyze more aspects of bias in generated text, such as sentiment and gender.

Thank you!

Summary:

- Analyze text generated by LLMs in different languages and different cultures
- Compare the diversity, granularity and consistency of related entities given by different setups

Question?

- Other interesting categories?
- More aspects to focus on?
- ...