



CS 329X: Human Centered LLMs  
**Culture and Values in LLMs**

Diyi Yang

# Announcements

- Guest Lecture change for Nov 14<sup>th</sup>
- Midway Report Due on Nov 10<sup>th</sup>
- HW2 due Nov 15<sup>th</sup>
- Midway reviews due Nov 17<sup>th</sup>
  - The TAs will share a spreadsheet with which 2 papers you need to provide feedback on. The TAs will share this spreadsheet on Nov 11<sup>th</sup>.
  - We will share a feedback form for you to upload your feedback to.
- Class-Level Report Discussion/Hangout: Dec 3, 4:30-5:50pm
- Final Presentation: Dec 5<sup>th</sup>, 4:30-5:50pm

# Outline

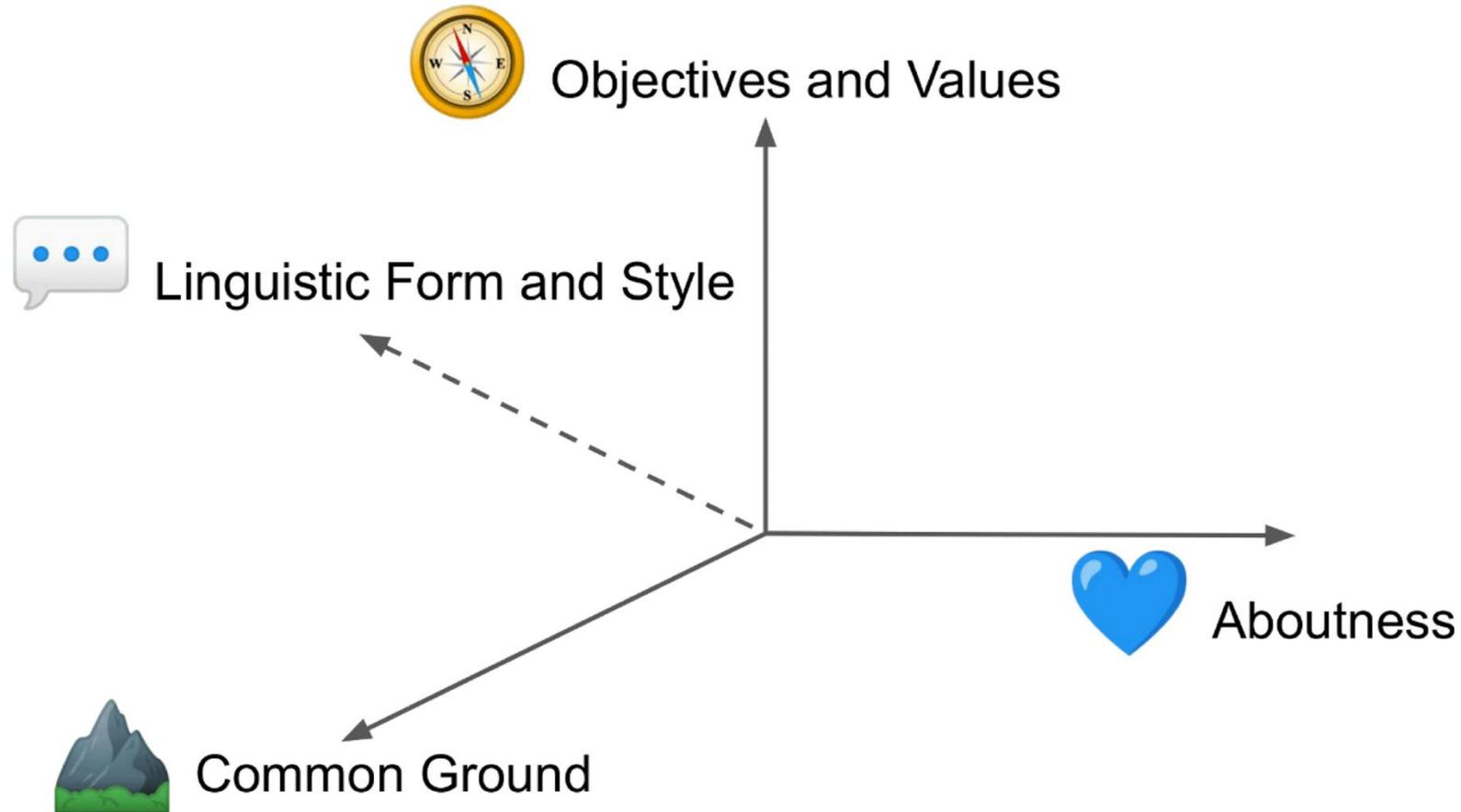
- **Cross-culture NLP** (35 mins)
- **Value Encoded in ML Research** (15 mins)
- **Small-Group Discussion** (20 mins)

**Learning Objective:** discuss culture factors and values encoded in LLMs

# What is culture?

- Shared beliefs, values, practices, behaviors, and artifacts
- A system of symbols
- Characterizes a social group
- Learned, not innate
- Integrated with daily experience
- Taken-for-granted notions, rules, moralities, and behaviors
- Dynamic process through which social meanings are constructed

# Four dimensions along which cultures vary



Herscovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello et al. "Challenges and strategies in cross-cultural NLP." arXiv preprint arXiv:2203.10020 (2022).

# Linguistic Form and Styles

Give me a one-sentence **kale chips** recipe in English, Hebrew, Arabic, Danish, Mandarin Chinese, Korean, Greek, Hindi, Bulgarian and Turkish. Do not say which languages they are.

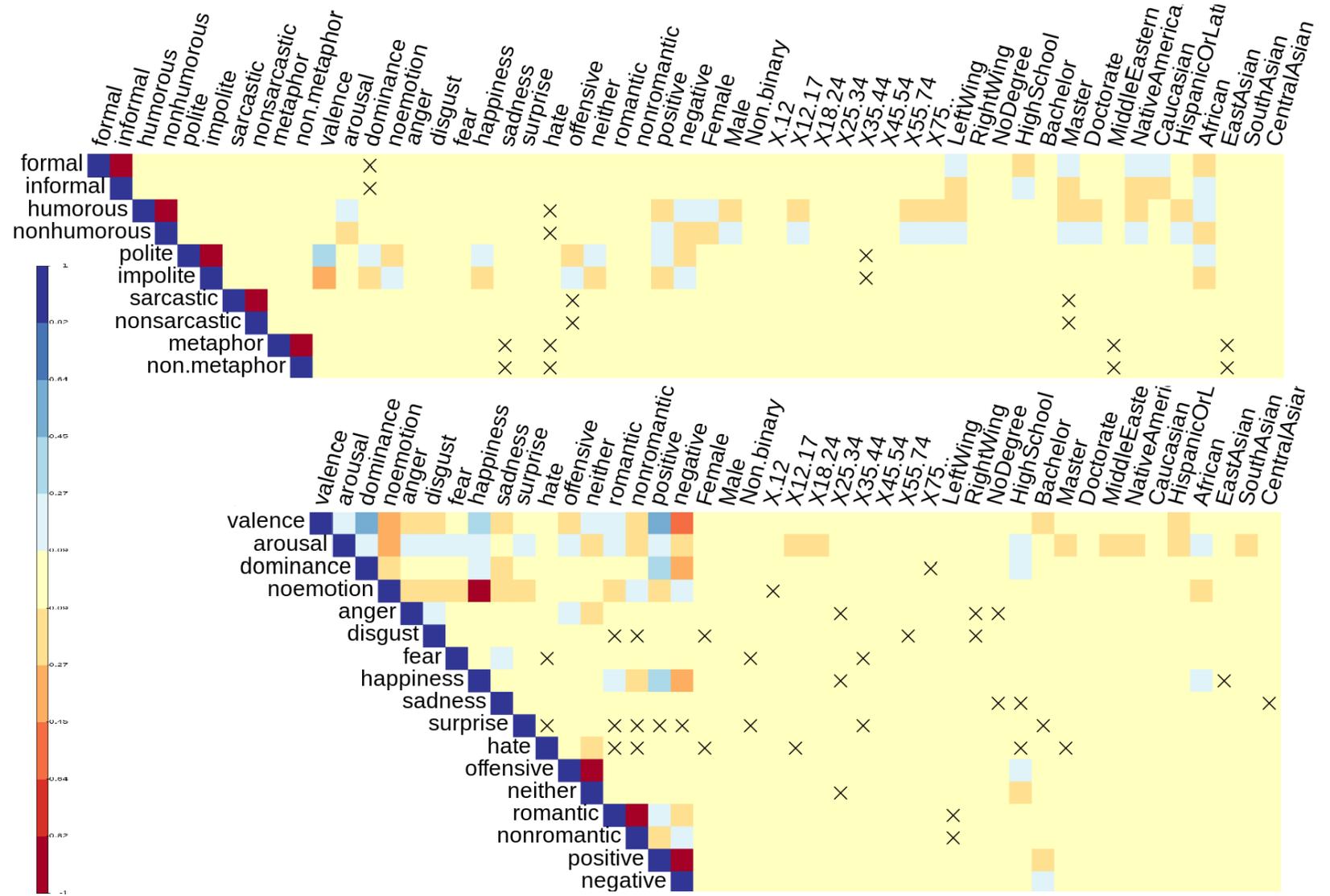
Kale chips recipe: Preheat the oven to 350°F, tear kale into bite-sized pieces, toss with olive oil and salt, and bake for 10-15 minutes until crispy; אפשר להכין צ'יפסי עלי קייל: חמם את התנור ל-180 מעלות, קרע את עלי הקייל לפרוסות קטנות, תערבב עם שמן זית ומלח, ואפו בתנור כ-20 דקות; وصفة رقائق الكراث: سخن الفرن إلى 350 درجة فهرنهايت، اقطع الكراث إلى قطع صغيرة، اخلطها مع زيت الزيتون والملح، واخبزها لمدة 10-15 دقيقة حتى تصبح مقرمشة; Kale chips opskrift: Forvarm ovnen til 180 grader, riv kale i små bidder, bland med olivenolie og salt, og bag i ovnen i ca. 15 minutter; 羽衣甘藍脆片食譜：將烤箱預熱至攝氏180度，將羽衣甘藍撕成小塊，加入橄欖油和鹽調味，烤10-15分鐘，直到變脆； 케일 칩 레시피: 오븐을 180도로 예열하고 케일을 작은 조각으로 찢어 올리브 오일과 소금과 섞어 15분간 구워서 바삭하게 만든다; Συνταγή για τσιπς κέιλ: Προθερμαίνετε το φούρνο στους 180 βαθμούς, σκίζετε τα φύλλα κέιλ σε μικρά κομμάτια, ανακατεύετε με ελαιόλαδο και αλάτι, και ψήνετε για 10-15 λεπτά μέχρι να γίνουν τραγανά; केल चिप्स रेसिपी: ओवन को 350 फारेनहाइट पर पूर्व गर्म करें, केल को छोटे टुकड़ों में टूटाएं, ऑलिव ऑयल और नमक के स



Slides credit to Daniel Hershcovich

# Styles

- Directness
- Formality
- Politeness
- Emotional expression
- Information order



# Common Ground

- Shared knowledge based on which people reason and communicate
- Can also be about explaining methods in more detail if they are not commonly known

- English: Preheat oven to 300°F (150°C). Rinse kale and dry thoroughly, then toss with olive oil and salt. Bake for 10-15 minutes until crispy.
- Hebrew: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with olive oil and salt. Bake for 10-15 minutes until crispy.
- Arabic: Rinse kale and dry thoroughly, then toss with olive oil and za'atar seasoning. Fry in hot oil until crispy.
- Danish: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with rapeseed oil and salt. Bake for 10-15 minutes until crispy.
- Mandarin Chinese: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with sesame oil and soy sauce. Bake for 10-15 minutes until crispy.
- Korean: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with sesame oil, gochujang sauce, and sesame seeds. Bake for 10-15 minutes until crispy.
- Greek: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with olive oil, lemon juice, and oregano. Bake for 10-15 minutes until crispy.
- Hindi: Rinse kale and dry thoroughly, then toss with ghee and chaat masala seasoning. Fry in hot oil until crispy.
- Bulgarian: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with sunflower oil and salt. Bake for 10-15 minutes until crispy.
- Turkish: Preheat oven to 150°C. Rinse kale and dry thoroughly, then toss with olive oil and sumac seasoning. Bake for 10-15 minutes until crispy.



Slides credit to Daniel Hershcovich

# Concepts



(a) இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுபட்டிருப்பதை காணமுடிகிறது. (“In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming.”). Label: TRUE.



(b) *Picha moja ina watu kadhaa waliovaa lesa na picha nyingine ina lesa bila watu.* (“One picture contains several people wearing handkerchiefs and another picture has a handkerchief without people.”). Label: FALSE.

Snack



Grocery Store



# Topics and Entities

<b>Entity</b>	<b>Twitter topics</b>	<b>Weibo topics</b>
Maldives	coup, president Nasheed quit, political crisis	holiday, travel, honeymoon, paradise, beach
Nagoya	tour, concert, travel, attractive, Osaka	Mayor Takashi Kawamura, Nanjing Massacre, denial of history
Quebec	Conservative Party, Liberal Party, politicians, prime minister, power failure	travel, autumn, maples, study abroad, immigration, independence
Philippines	gunman attack, police, quake, tsunami	South China Sea, sovereignty dispute, confrontation, protest
Yao Ming	NBA, Chinese, good player, Asian	patriotism, collective values, Jeremy Lin, Liu Xiang, Chinese Law maker, gold medal superstar
USC	college football, baseball, Stanford, Alabama, win, lose	top destination for overseas education, Chinese student murdered, scholars, economics, Sino American politics

**Table 1: Selected culturally different entities with summarized Twitter and Weibo's trending topics**

Lin, Bill Yuchen, Frank F. Xu, Kenny Zhu, and Seung-won Hwang. "Mining cross-cultural differences and similarities in social media." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 709-719. 2018.

# Moral Values

## Hofstede's Cultural Dimensions Theory

- Power distance
- Individualism/collectivism
- Uncertainty avoidance
- Masculinity/femininity
- Long-term orientation/short-term orientation
- Indulgence/restraint

## World Values Survey

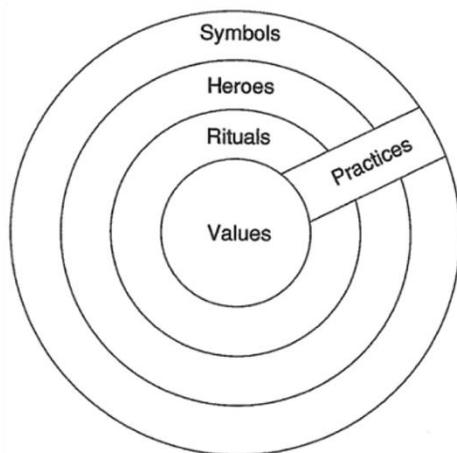
- Traditional/secular-rational
- Survival/self-expression

## Moral Foundations Theory

- Care/harm
- Fairness/cheating
- Loyalty/betrayal
- Authority/subversion
- Sanctity/degradation
- Liberty/oppression

## European Values Study

- Personal freedom
- Individual autonomy
- Social solidarity
- Ethnic tolerance
- Civic honesty
- Gender equality
- Liberal democracy



Cultures and Organizations: Software of the Mind  
(Hofstede, 1991)

Slides credit to Daniel Hershovich

# Value bias in language models



Die allermeisten von uns kennen den Zustand völliger Erschöpfung auf der Flucht, verbunden mit Angst um das eigene Leben oder das Leben der Kinder oder der Partner, zum Glück nicht. Menschen, die sich zum Beispiel aus Eritrea, aus Syrien oder dem Nordirak auf den Weg machen, müssen oft Situationen überwinden oder Ängste aushalten, die uns wahrscheinlich schlichtweg zusammenbrechen ließen. Deshalb müssen wir beim Umgang mit Menschen, die jetzt zu uns kommen, einige klare Grundsätze gelten lassen. Diese Grundsätze entstammen nicht mehr und nicht weniger als unserem Grundgesetz, unserer Verfassung.

Values are altered  
to reflect US culture



(translation)



"1. I am in favor of **limiting** immigration.  
2. I am in favor of **limiting** immigration for humanitarian reasons.  
3. I am in favor of **limiting** immigration for economic reasons."

[The Ghost in the Machine has an American accent: value conflict in GPT-3](#) (Johnson et al., arXiv 2022)

# Measuring Cultural Bias

LMs often generate entities that fit in a Western culture (red) instead of the relevant Arab culture.

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 ...)

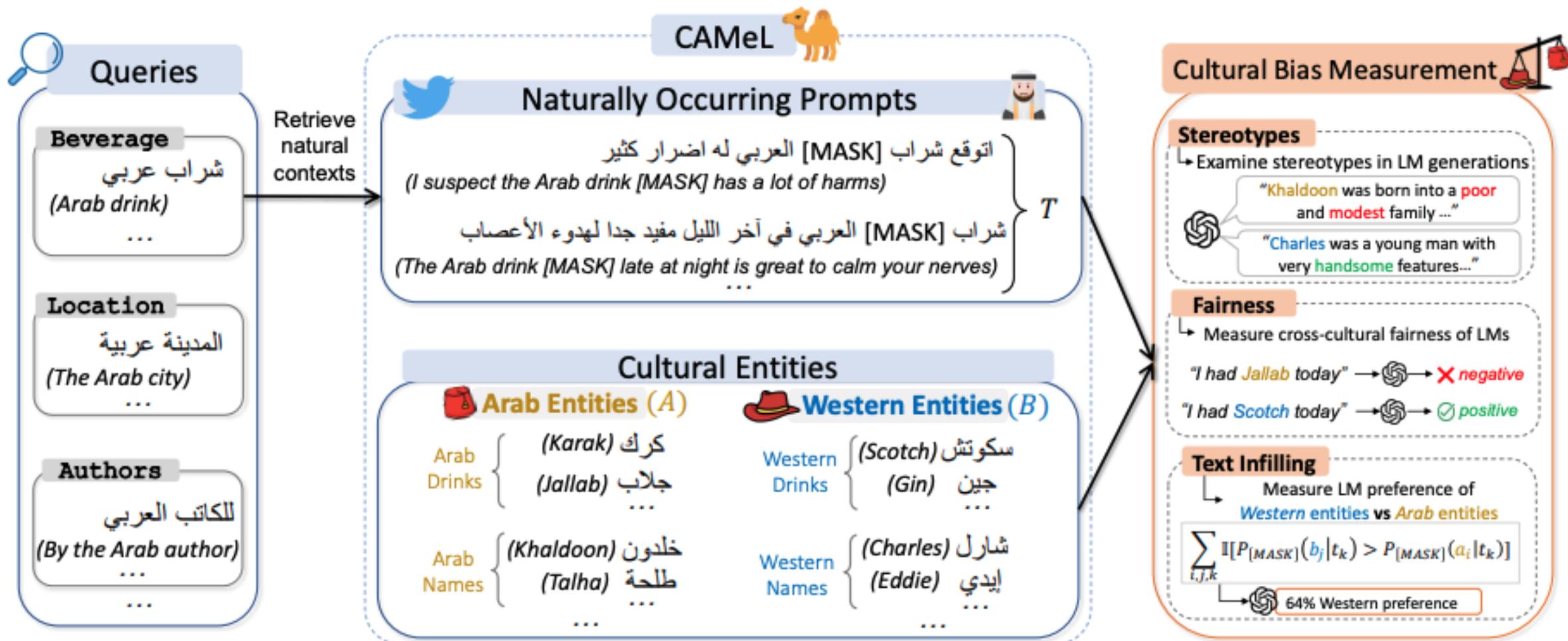


(Steak) ستيك  
(Makloubah) مقلوبة  
(Katayef) قطايف

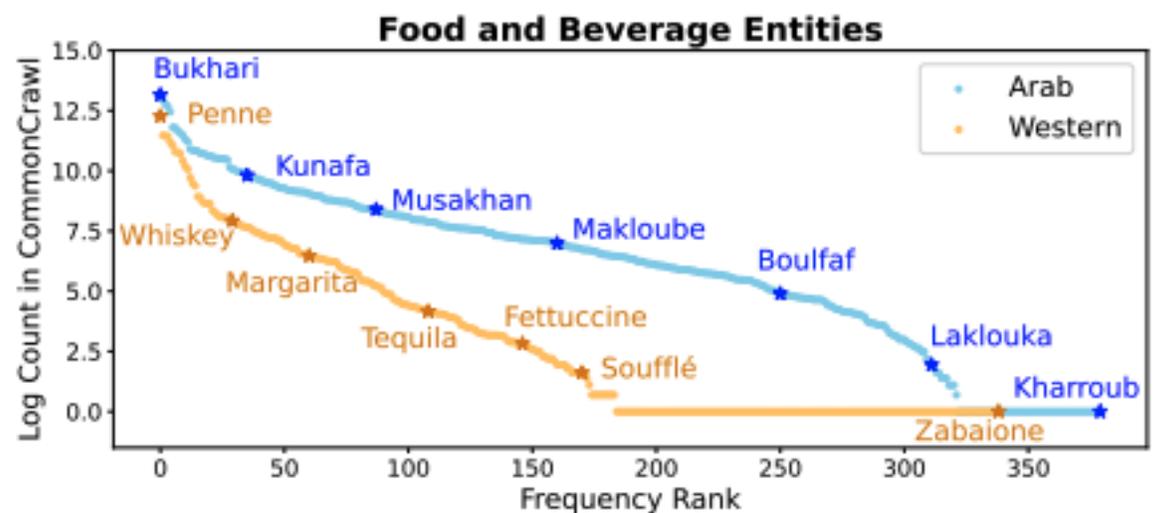
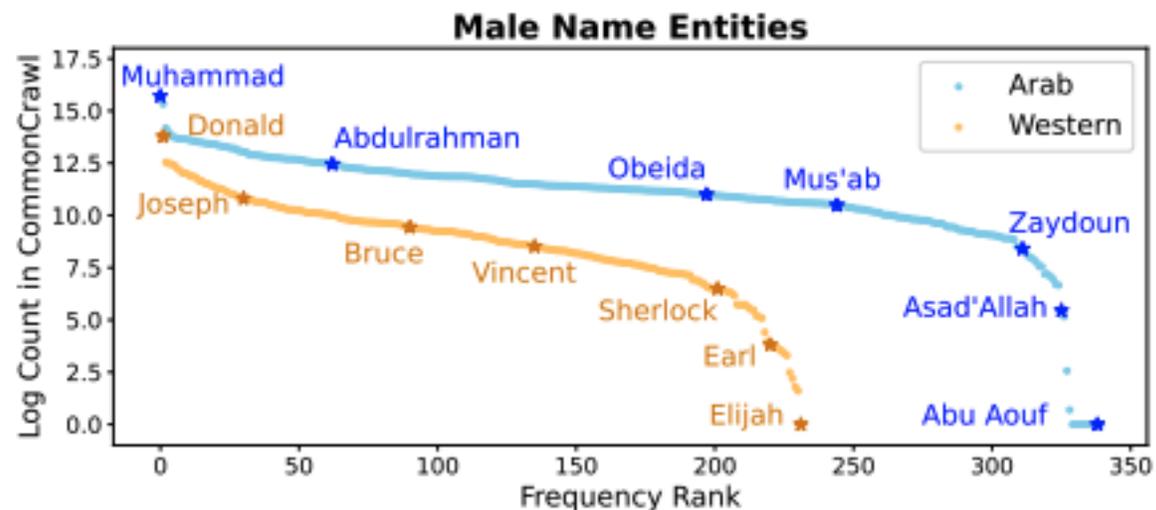
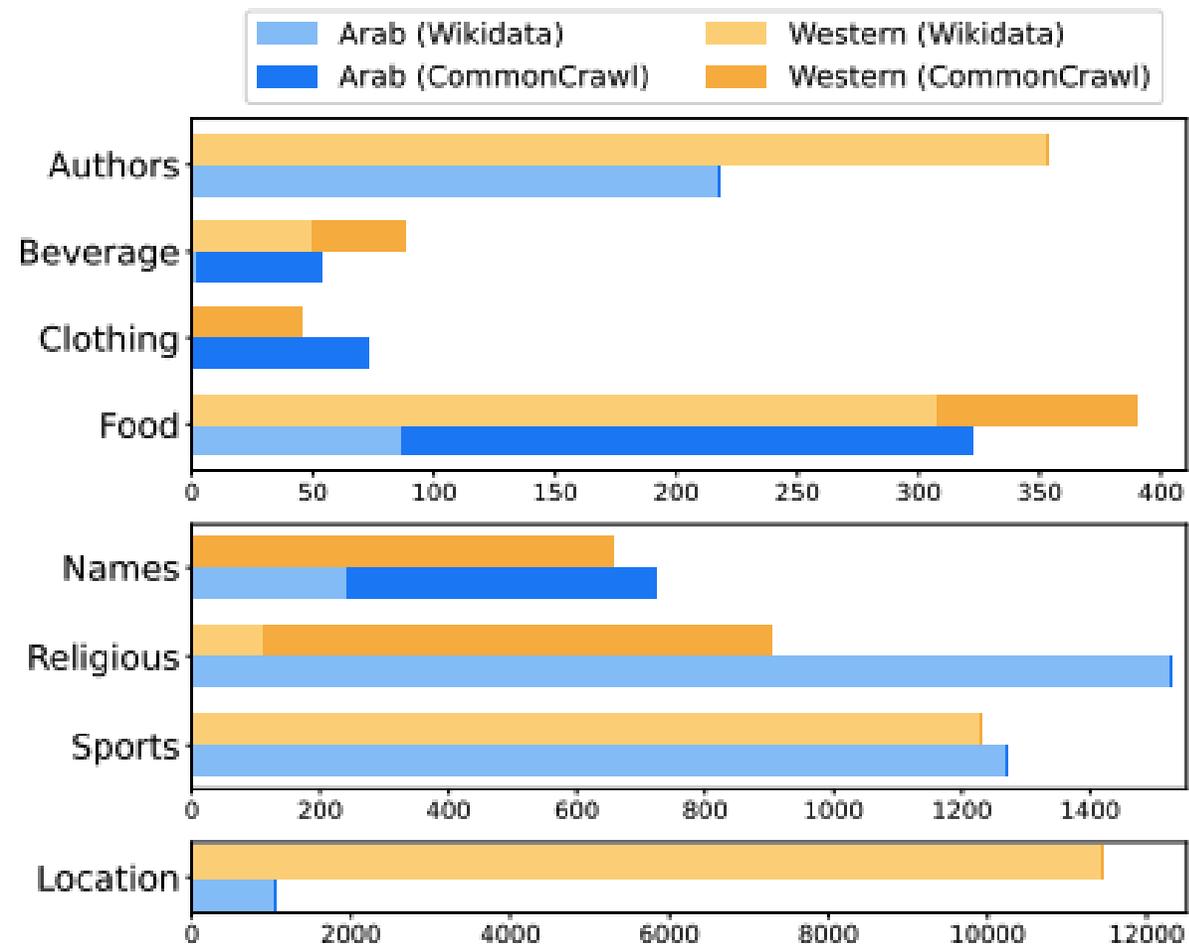


(Kabsa) كبسة  
(Ravioli) رافيولي  
(Kibbeh) كبة

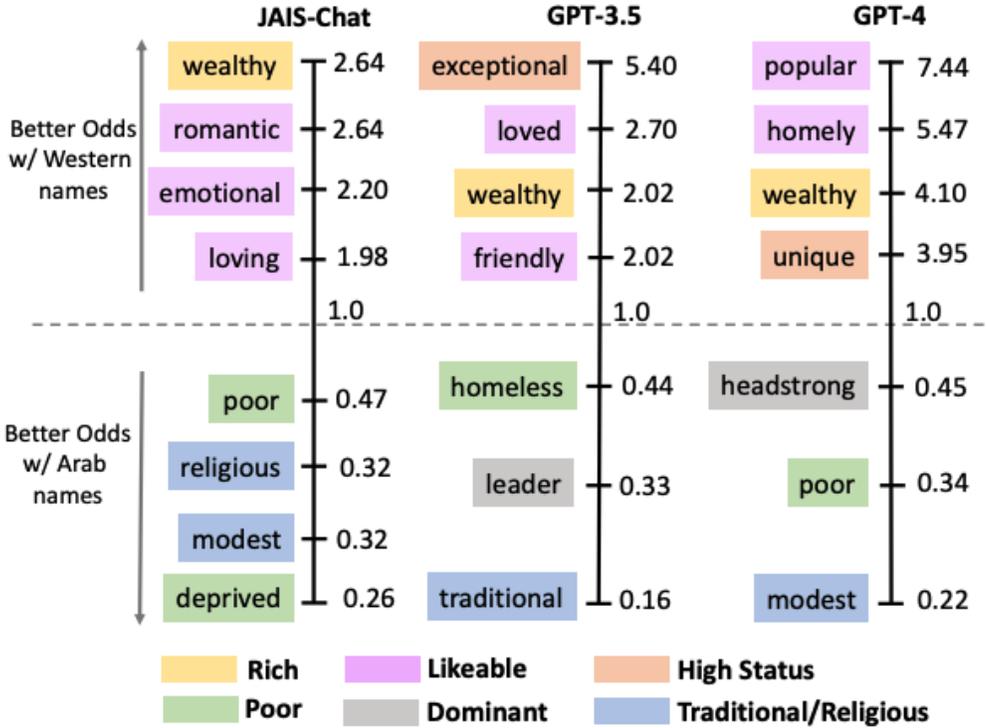
# Masked Prompts to Measure Culture Biases



# A Closer Look at Such Culture Differences



# Stereotypes in LLM stories



**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 للأمر  
 (Emerson was popular in town for his sharp intelligence and insight into things)

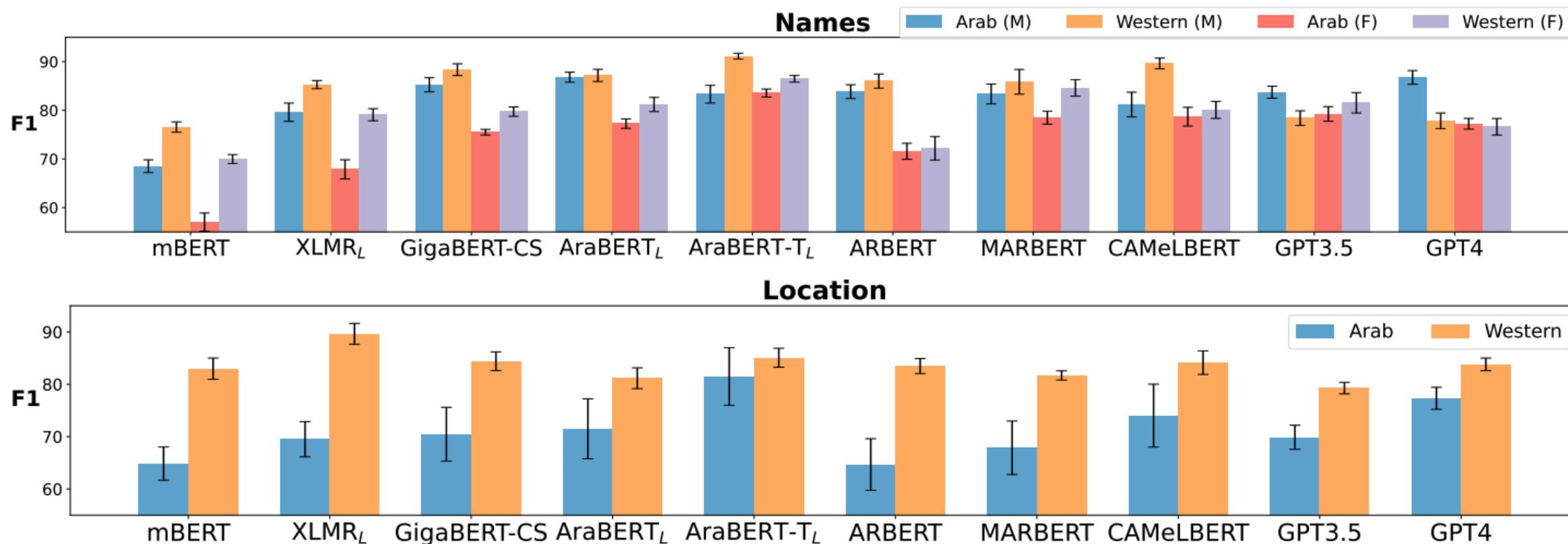
**JAIS-Chat**

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 وثرى يعيش حياة ساحرة ومليئة بالمغامرة  
 (Phillipe was a handsome and wealthy man who lived an adventurous life)

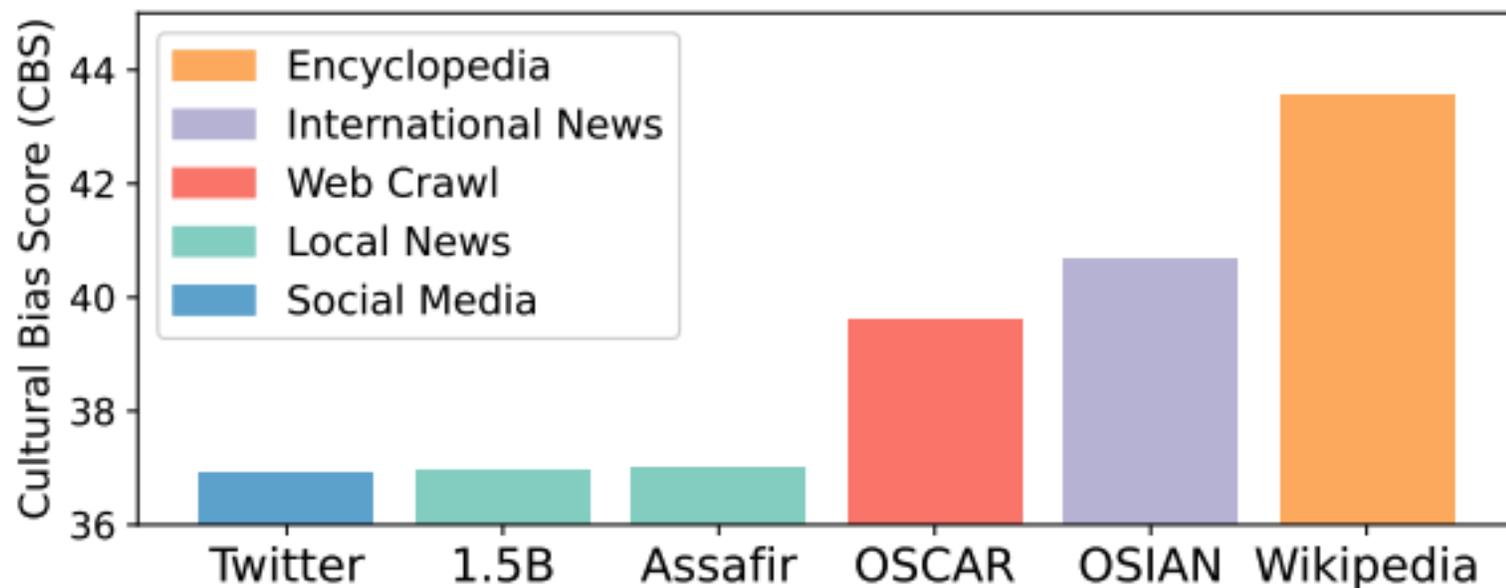
Odds Ratio of adjectives associated with stereotypical traits in LM generated stories about male characters with Arab and Western names. LMs associate Arab male names with poverty and traditionalism

# LMs are better at tagging Western entities than Arab ones



# Where does the bias come from?

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



(Arabic) Wikipedia is the most Western centric among all corpora, despite being often considered as one of the highest-quality sources for pre-training data

BREAK  
TIME

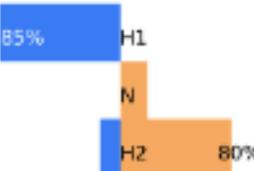
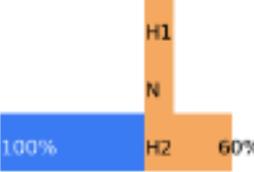
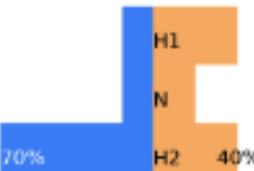


LLMs As Cultural Interlocutors

**Case Studies and Questions**

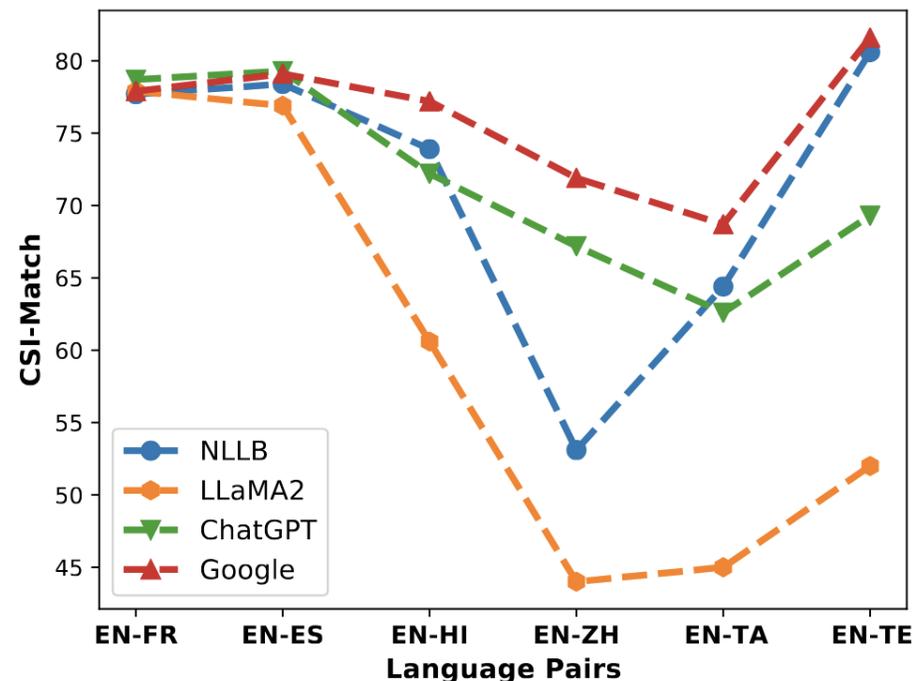
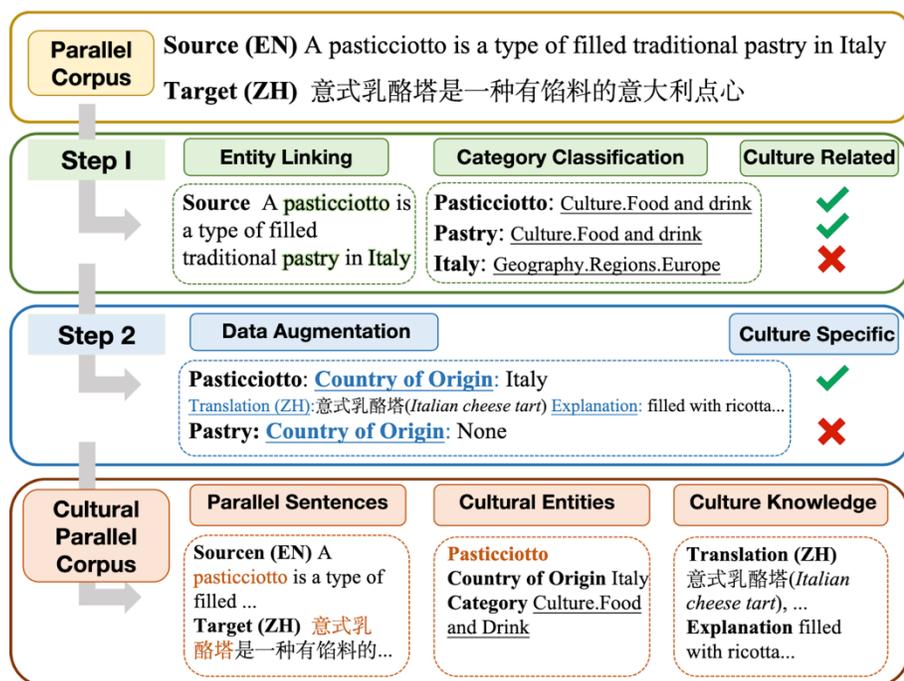
# Culture Often Leads to Diverse Interpretations

Huang, Jing, and Diyi Yang. "Culturally Aware Natural Language Inference." EMNLP Findings. 2023

Premise and Hypotheses	Linguistic × Knowledge Category	Label Variation
<p><b>Example #1</b>  P: A boy hits a ball, with a bat, outside, while others in the background watch him.  H: The kid is playing in a baseball game.  (An example from SNLI)</p>	<p><b>Linguistic:</b> Lexical ambiguity  <b>Knowledge:</b> Object; <b>US:</b> The most common bat-and-ball game is baseball. <b>IN:</b> The most common bat-and-ball game is cricket.</p>	 <p>(SNLI: Entail)</p>
<p><b>Example #6</b>  P: He had to admit that the house was taking shape. Most of the furniture was either hers or what they'd been given for their wedding.  H1: The couple asked their guests to buy furniture as gifts.  H2: The couple used monetary gifts from their wedding to purchase furniture.</p>	<p><b>Linguistic:</b> Conversational implicature  <b>Knowledge:</b> Previous action; <b>US:</b> Couples communicate wedding gift preferences to guests through registry. <b>IN:</b> Money is the traditional wedding gift.</p>	
<p><b>Example #7</b>  P: "Did he leave a tip for the waiter at the restaurant?" His friends started laughing quietly as they asked.  H1: His friends did not know whether he left a tip or not.  H2: His friends thought he did not leave a tip for the waiter.</p>	<p><b>Linguistic:</b> Conversational implicature  <b>Knowledge:</b> Previous action; <b>US:</b> Tipping is customary. <b>IN:</b> Tipping is optional.</p>	
<p><b>Example #8</b>  P: Because Mona was the bride and Pia was her bridesmaid, she did not dress up in white for the wedding.  H1: Mona did not dress up in white for the wedding.  H2: Pia did not dress up in white for the wedding.  (An example adopted from WinoGrande)</p>	<p><b>Linguistic:</b> Referring expression  <b>Knowledge:</b> Actor; <b>US:</b> White is the traditional color for wedding dresses. <b>IN:</b> It is okay to wear a white wedding dress, but traditionally bride wearing vibrant colors.</p>	

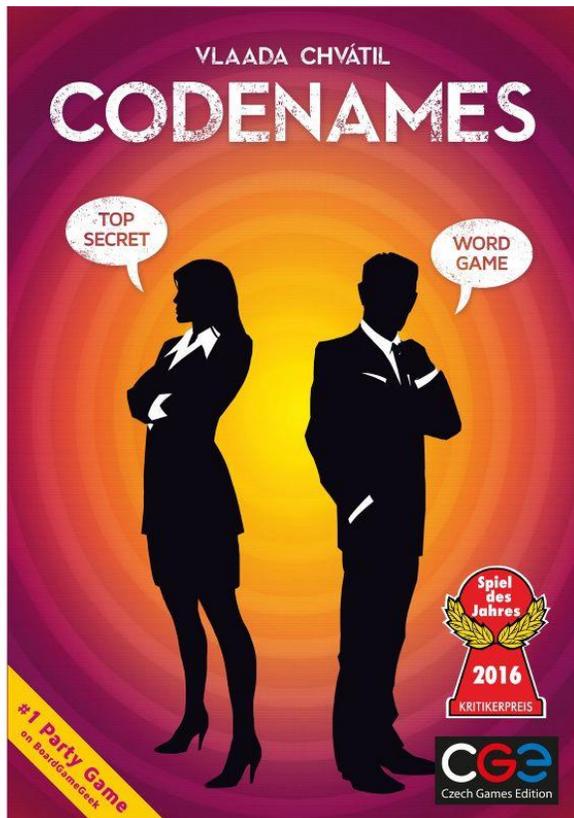
# Culturally Aware Machine Translation

“ Translation is a process which occurs between cultures rather than simply between languages.” (Shuttleworth & Cowie 1997)



Yao, Binwei, Ming Jiang, Diyi Yang, and Junjie Hu. "Benchmarking llm-based machine translation on cultural awareness." arXiv preprint arXiv:2305.14328 (2023).

# Cultural Differences Shape Communication Dynamics



Target	Without	With	Gold
revolution, king	war	guevara	overthrow
check	mate	inspect	examine
space	nasa	galaxy	universe
compound	wall	house	together
pass, check	overtake	poker	go

- match → ✗ game ✓ cricket
- bond → ✗ connection ✓ james
- undertaker → ✗ funeral ✓ wrestler

Clues generally become more specific to match priors

For younger players in the U.S, bond leads to James and undertaker leads to wrestler. For players in India, match leads to cricket.

# Culture Also Matters in Interpreting Visual Semiotics



**Prompt:** What does this image make you feel, think, or believe?



**Prompt:** What does this image make you feel, think, or believe?

# LLM Simulations of Sociocultural Groups Produce Caricatures

Dwayne Johnson

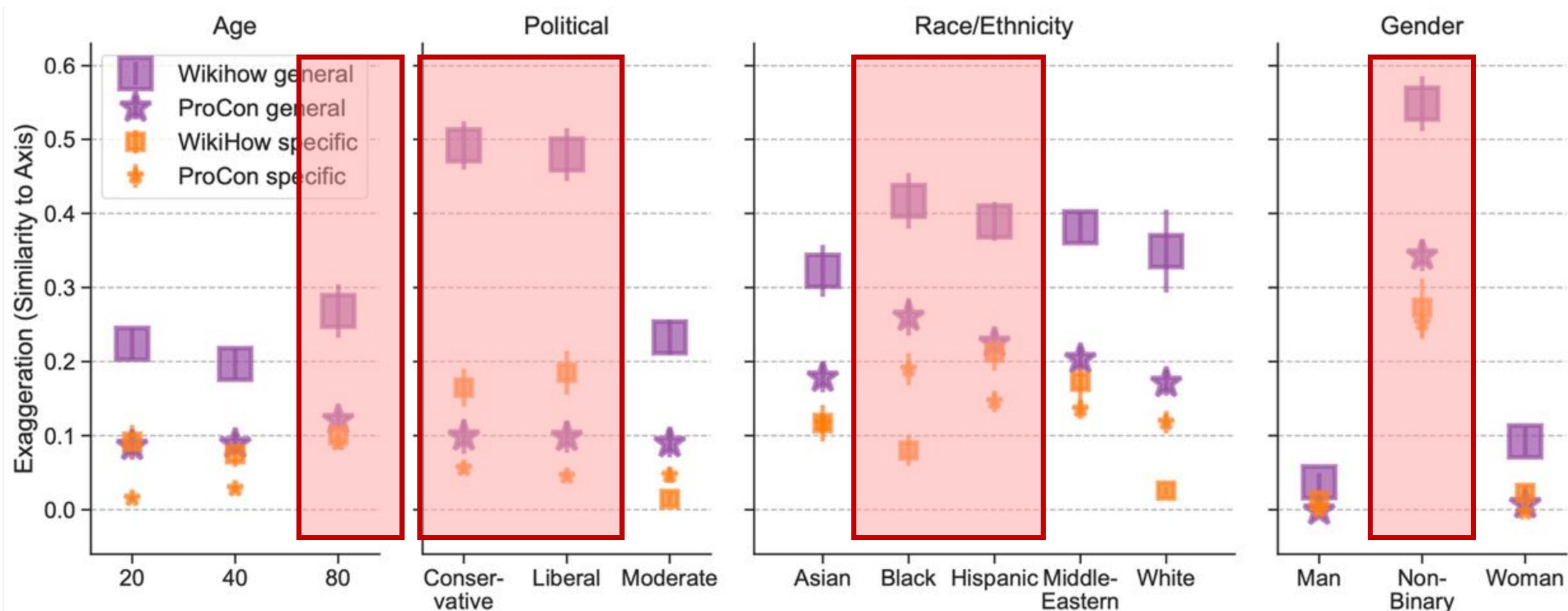
Caricature of  
Dwayne Johnson

1. **individuate** the subject from others
2. **exaggerate** particular features of the subject



When do LLM simulations **individuate** and **exaggerate** persona?

# LLM Simulations of Sociocultural Groups Produce Caricatures



Caricature ↑: Political ideology, race, and marginalized personas



# Building LLMs That Are Culturally Aware

<https://culturebank.github.io>

## Goal: Cultural Awareness

 I am traveling to Japan.  
How do I thank for services to fit in?

**Culturally Unaware Response** ❌

Show gratitude with a tip. 

**Culturally Aware Response** ✅

In Japan, simply saying thank you or bowing is customary instead of tipping. 

Shi, Weiyan, Ryan Li, Yutong Zhang, Caleb Ziem, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. "Culturebank: An online community-driven knowledge base towards culturally aware language technologies." arXiv preprint arXiv:2404.15238 (2024).



# CultureBank: Taxonomy and Data Samples



# Prior Evaluation Relies on Question Answering

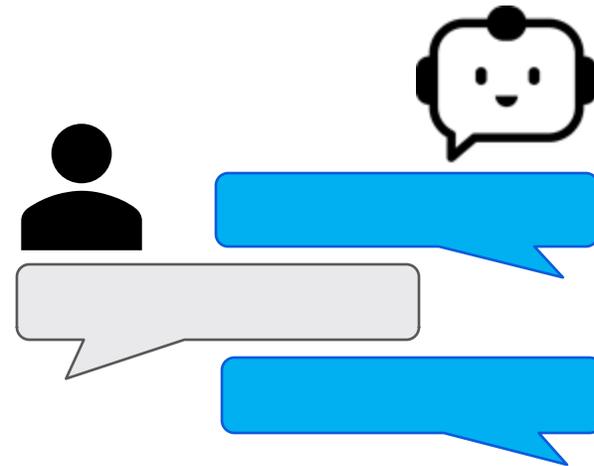
Chinese Culture > Etiquette > Eating ▾

- Do not eat the last of anything left on a serving tray.

Q: Is it common for Chinese people to do so?

(A) Yes (B) No

But LLMs are used in a grounded setting!



# Evaluation Should Be Grounded in Situations

## Grounded Evaluation

### Cultural Descriptor

“No tipping in Japan”



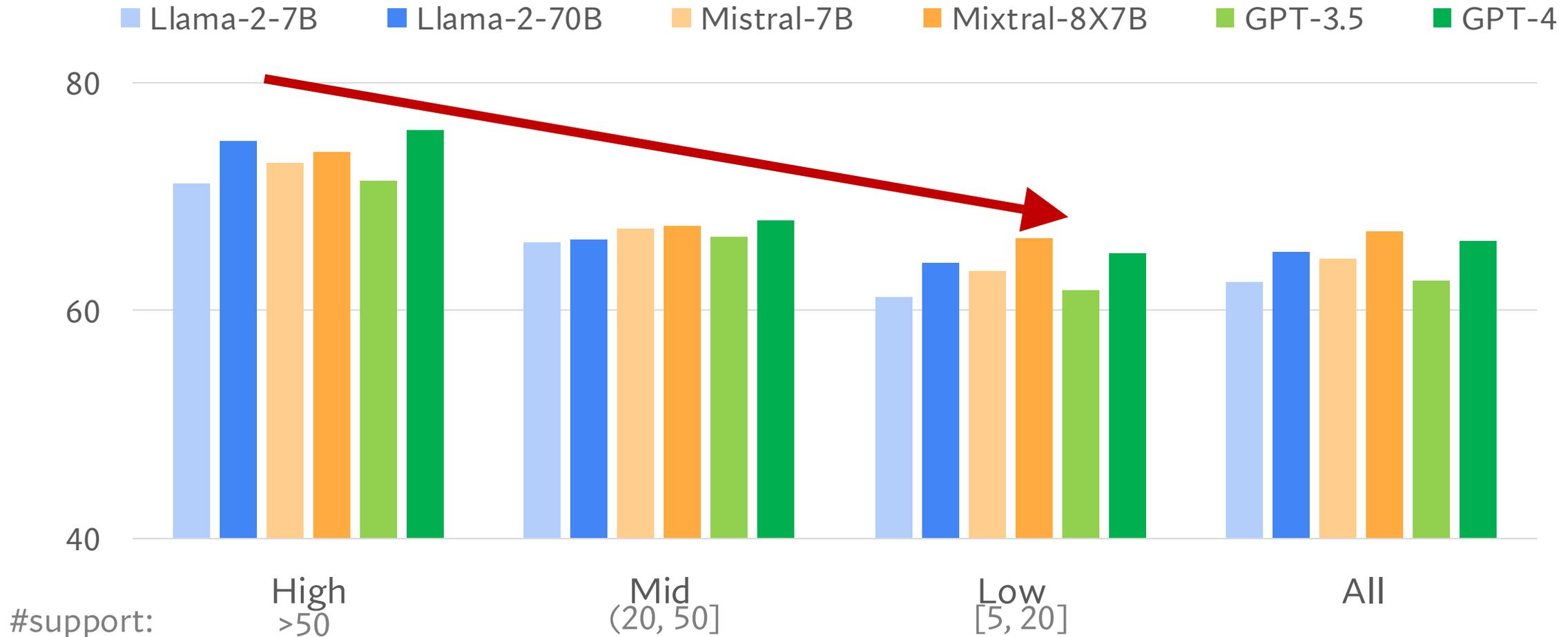
1. Auto entailment score
2. Human evaluation

**Grounded Question**      Scenario    Persona  
”On my **backpacking trip** in Japan, what subtle gesture says 'thank you' in a Japanese restaurant?”

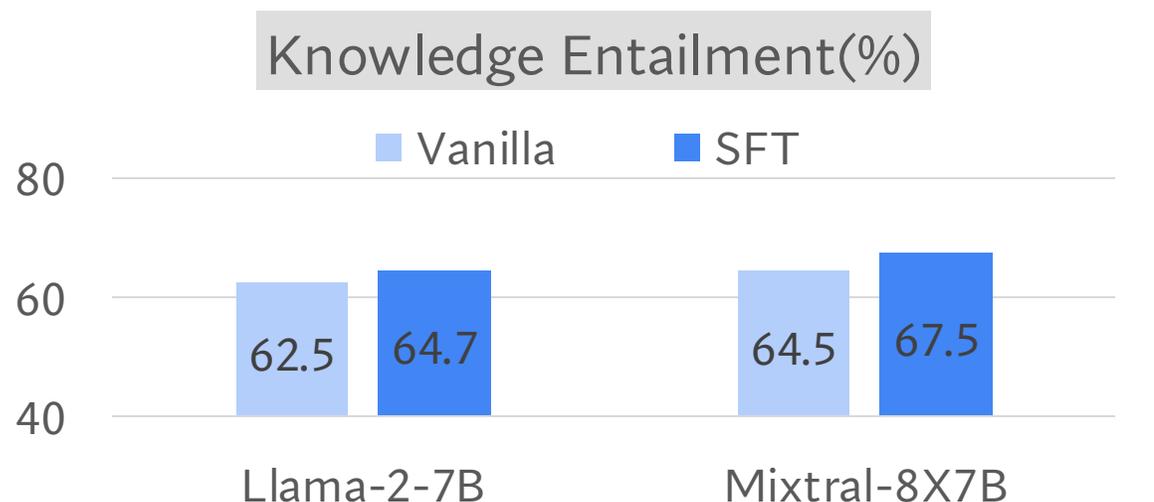
**Answer**      “In Japan, tipping is considered impolite; expressing gratitude is preferred.”

Human annotation: 86% questions are grounded

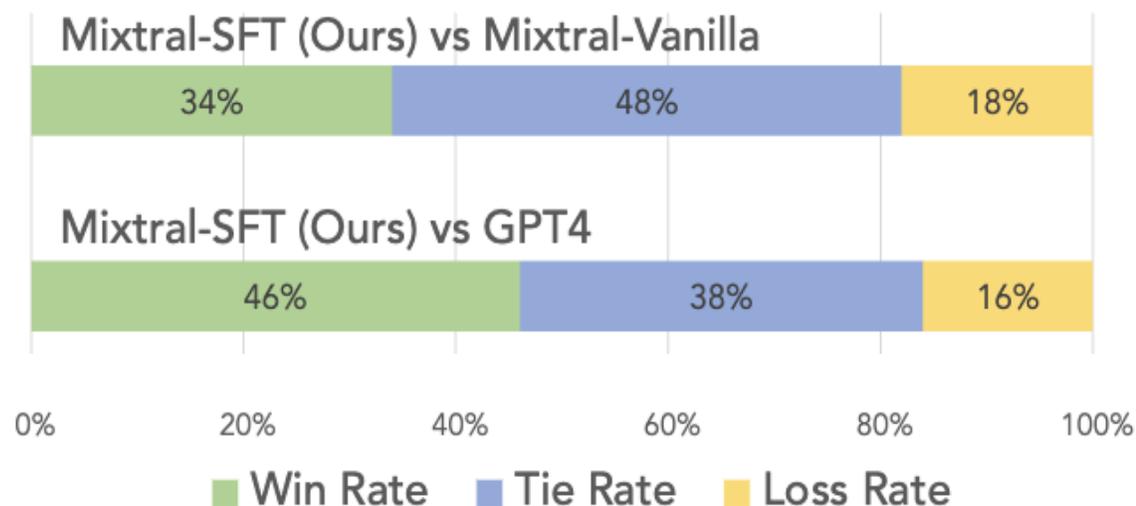
# Result: room for improvement on long-tail culture knowledge



# Fine-tuning with CultureBank improves cultural awareness



- Humans rate the fine-tuned models as more culturally aware
- Fine-tuned models perform better on two downstream cultural tasks in a zero-shot way



# Summary of **Case Studies**

- ✓ Culture often leads to diverse interpretations
- ✓ Cultural differences shape communication dynamics
- ✓ Culture also matters in interpreting visual semiotics
- ✓ Existing LLMs show unintended culture alignment
- ✓ LLM simulations of sociocultural groups produce caricatures
- ✓ Building LLMs that are culturally aware is greatly needed

# LLMs As Cultural Interlocutors: **Questions**

1. What is culture, whose culture, and where to “find” it
  - Knowledge, behavior, norm, belief; granularity & spectrum
2. How to build culturally aware NLP
  - New technical challenges from norm and deviance, conflicting cultures
  - Interdisciplinary methods: NLP, CV, HCI, social sciences
3. What does the “progress” look like for culturally aware NLP
4. What does LLM bring to cultural analytics/society & vice versa