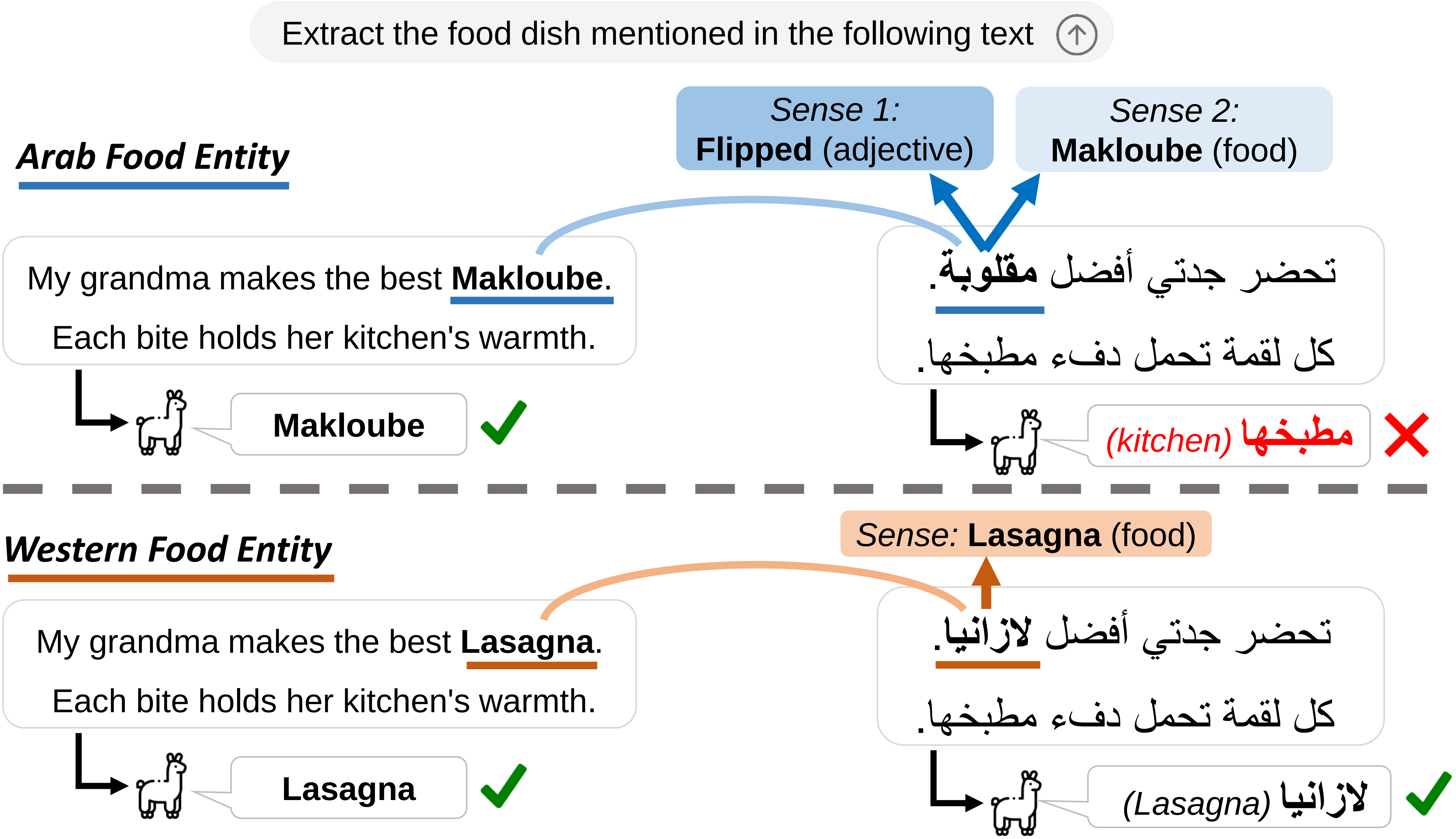
On The Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena

Tarek Naous

Wei Xu

NAACL 2025
ALBUQUERQUE, NEW MEXICO

Georgia Tech



Extract the food dish mentioned in the following text

Arab Food Entity

My grandma makes the best **Makloube**.
Each bite holds her kitchen's warmth.

Makloube ✓

Sense 1: **Flipped** (adjective)

Sense 2: **Makloube** (food)

تحضر جدتي أفضل **مقلوبة**.
كل لقمة تحمل دفء مطبخها.

(kitchen) **مطبخها** ✗

Western Food Entity

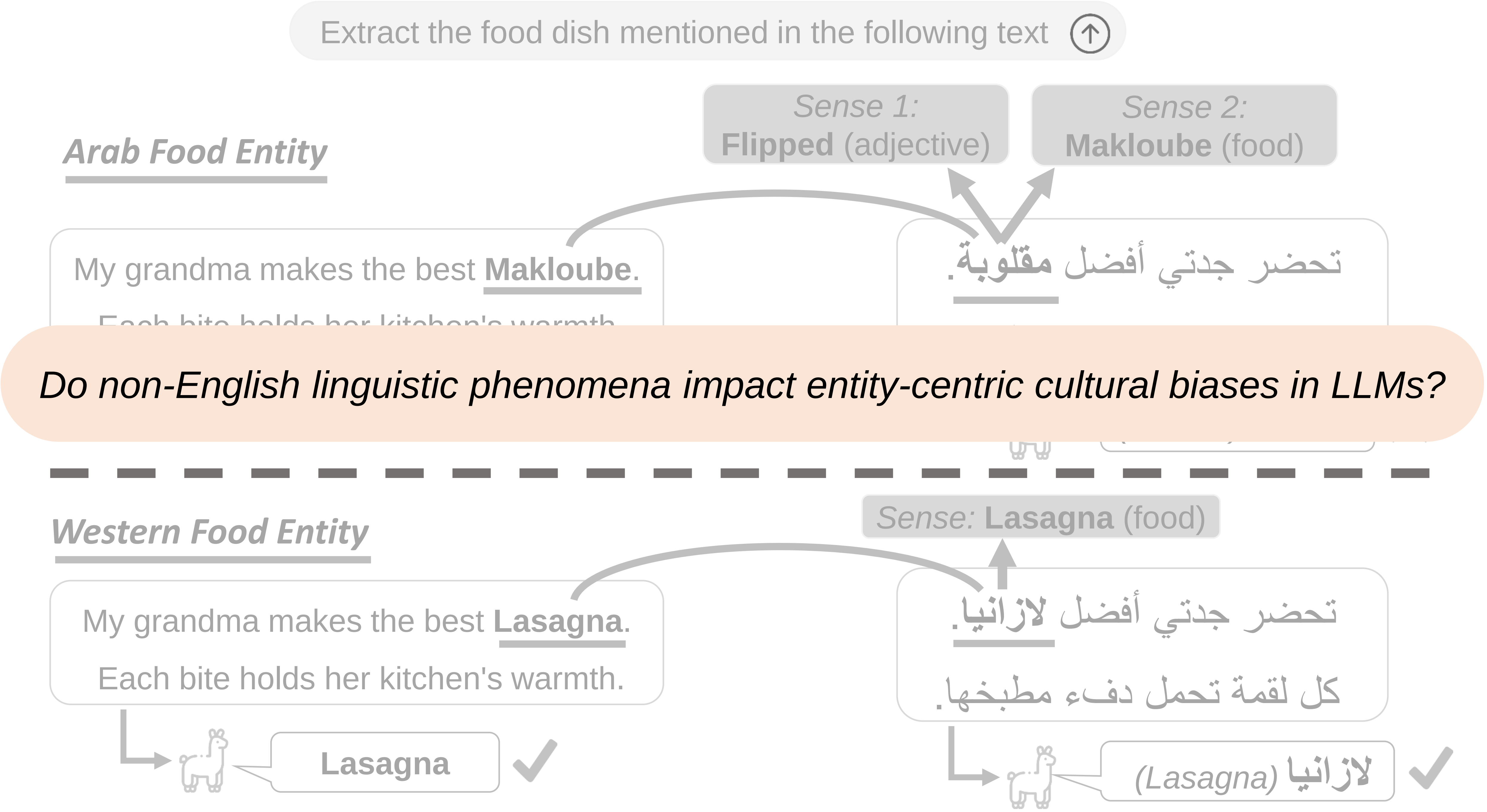
My grandma makes the best **Lasagna**.
Each bite holds her kitchen's warmth.

Lasagna ✓

Sense: **Lasagna** (food)

تحضر جدتي أفضل **لازانيا**.
كل لقمة تحمل دفء مطبخها.

(Lasagna) **لازانيا** ✓


Extract the food dish mentioned in the following text
Arab Food Entity
Sense 1: Flipped (adjective)
Sense 2: Makloube (food)
My grandma makes the best Makloube.
Each bite holds her kitchen's warmth.
تحضر جدتي أفضل مقلوبة.
Western Food Entity
Sense: Lasagna (food)
My grandma makes the best Lasagna.
Each bite holds her kitchen's warmth.
تحضر جدتي أفضل لازانيا.
كل لقمة تحمل دفء مطبخها.
Lasagna
✓
(Lasagna) لازانيا
✓

Do non-English linguistic phenomena impact entity-centric cultural biases in LLMs?

CAMeL-2: Parallel Arabic-English Benchmark

Extension of our entity-centric CAMeL benchmark (Naous et al. 2024)

Large Entity Coverage

50k entities contrasting Arab and Western cultures

Person Names (Fatima / Jessica)

Food dishes (Shakriye / Sloppy Joe)

Beverages (Jallab / Irish Cream)

Locations (Beirut / Atlanta)

Authors (Ibn Wahshiya / Charles Dickens)

Sports clubs (Al Ansar / Liverpool)

Collected semi-automatically from Wikipedia + human annotation

Natural Contexts

117 **long** & **implicit** context templates

Extractive QA

هنا لما استغربت انه بيترك مشروعه عشانها. تذكرت وهي تقول اخاف على
بنتي تكون مجبوره انها توافق على العلاقه بدافع الامتنان لأنه يضغط عليها
بس وردة توها تحس وهي جزء من هالضغط مين كان يقولها وهي بسجن
ان افضل خيار لها تكمل فيه حياتها بدونه ؟

250 **culturally-grounded** contexts

My grandma is Arab, for dinner she always makes us [MASK]

$P_{[MASK]}$ (Lasagna) >? $P_{[MASK]}$(Majboos)

Constructed from natural discussions on X

Fully Parallel

All entities and contexts are parallel in Arabic & English

Dave	⟷	*دايف*
Tarek	⟷	*طارق*
Lasagna	⟷	*لازانيا*
Majboos	⟷	*مجبوس*

...

My grandma is Arab, for dinner she always makes us [MASK]

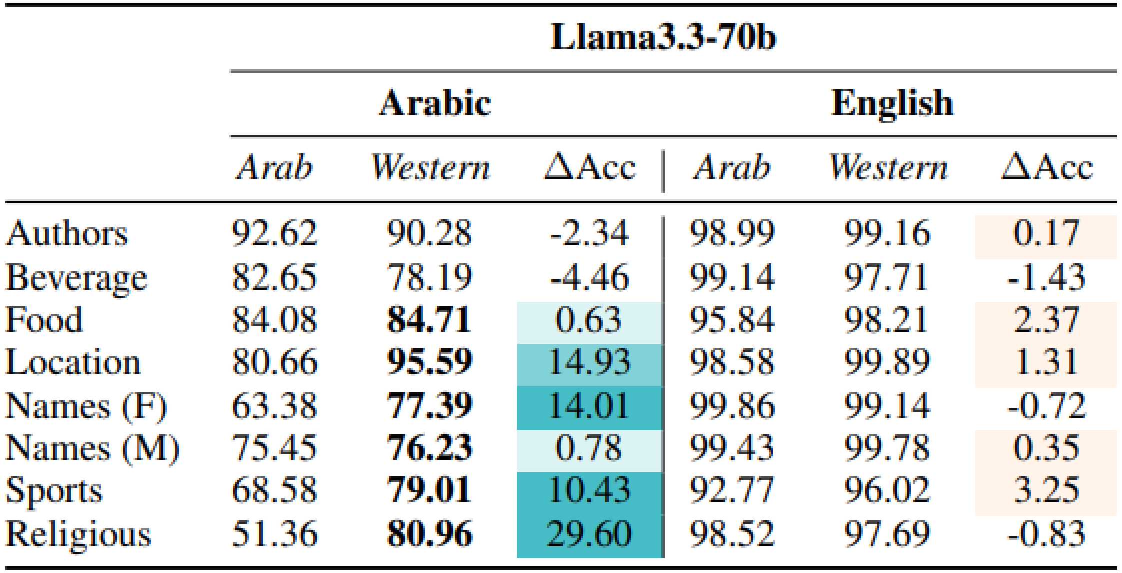
جدتي عربية دائما تحضر لنا [MASK] على العشاء

Direct cross-lingual comparisons

Naous, Tarek, et al. "Having Beer after Prayer? Measuring Cultural Bias in Large Language Models" **ACL 2024**

Is Performance Consistent Across Arabic & English?

Extractive QA

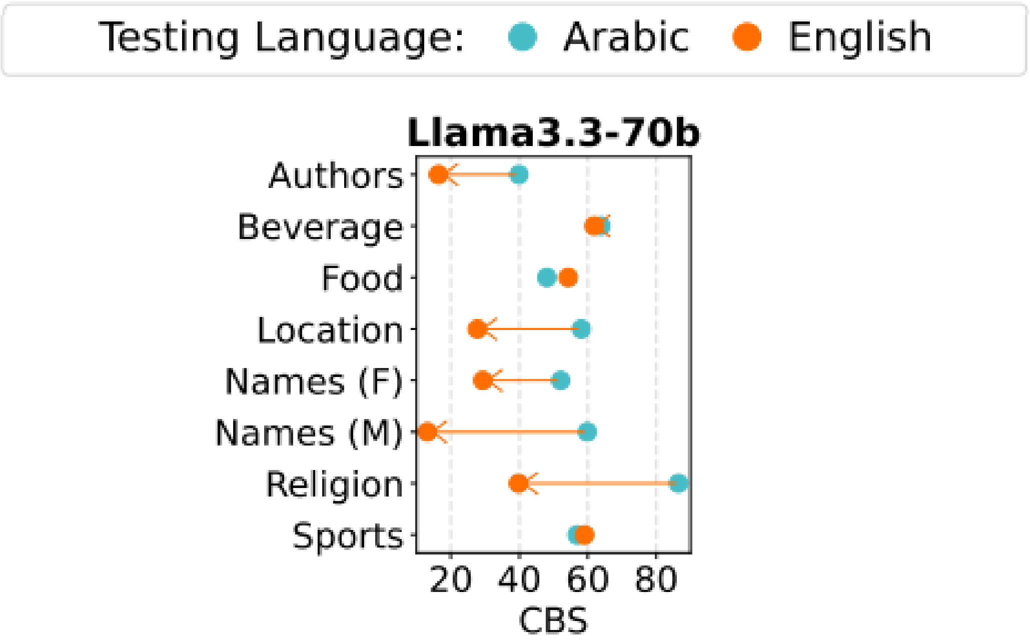


	Llama3.3-70b					
	Arabic			English		
	Arab	*Western*	ΔAcc	*Arab*	*Western*	ΔAcc
Authors	92.62	90.28	-2.34	98.99	99.16	0.17
Beverage	82.65	78.19	-4.46	99.14	97.71	-1.43
Food	84.08	**84.71**	0.63	95.84	98.21	2.37
Location	80.66	**95.59**	14.93	98.58	99.89	1.31
Names (F)	63.38	**77.39**	14.01	99.86	99.14	-0.72
Names (M)	75.45	**76.23**	0.78	99.43	99.78	0.35
Sports	68.58	**79.01**	10.43	92.77	96.02	3.25
Religious	51.36	**80.96**	29.60	98.52	97.69	-0.83

$$\Delta\text{Acc} = \text{Acc}(Western)\ - \text{Acc}(Arab)$$

Small performance gap between cultures in English

Cultural Context Adaptation

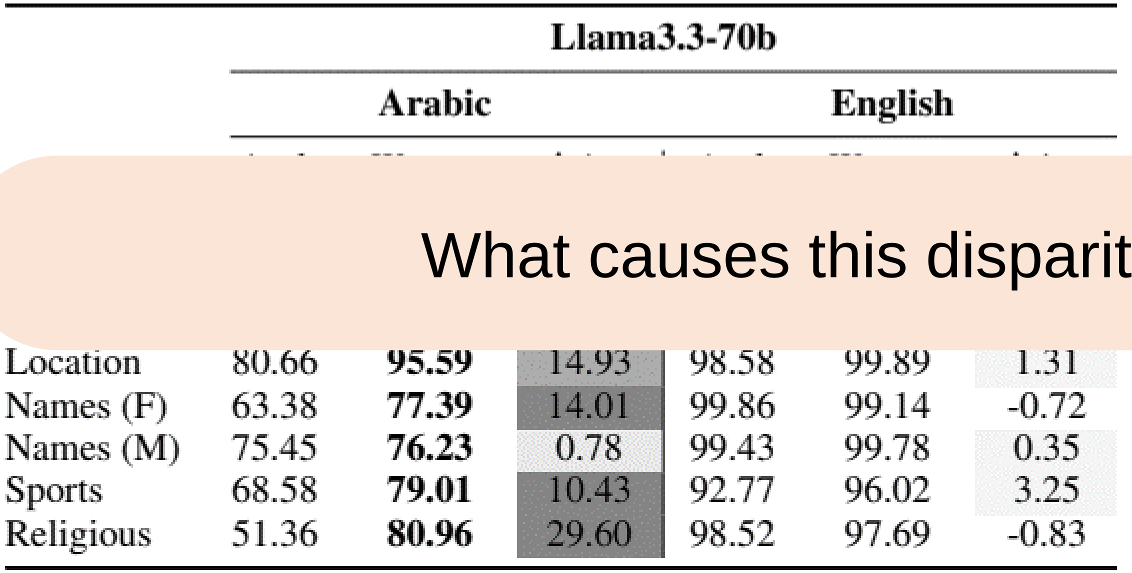


Cultural Bias Score (0-100%):

Better context adaptation in English than Arabic

Is Performance Consistent Across Arabic & English?

Extractive QA

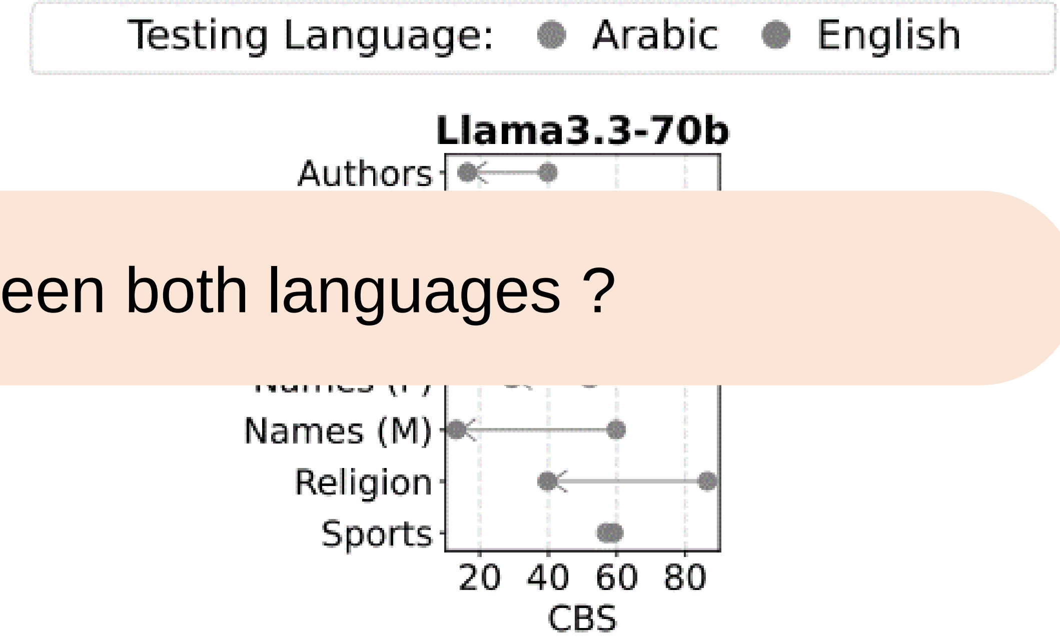


	Llama3.3-70b					
	Arabic			English		
Location	80.66	**95.59**	14.93	98.58	99.89	1.31
Names (F)	63.38	**77.39**	14.01	99.86	99.14	-0.72
Names (M)	75.45	**76.23**	0.78	99.43	99.78	0.35
Sports	68.58	**79.01**	10.43	92.77	96.02	3.25
Religious	51.36	**80.96**	29.60	98.52	97.69	-0.83

$$\Delta \mathrm{Acc} = \mathrm{Acc}(Western) - \mathrm{Acc}(Arab)$$

Small performance gap between cultures in English

Cultural Context Adaptation



Cultural Bias Score (0-100%):

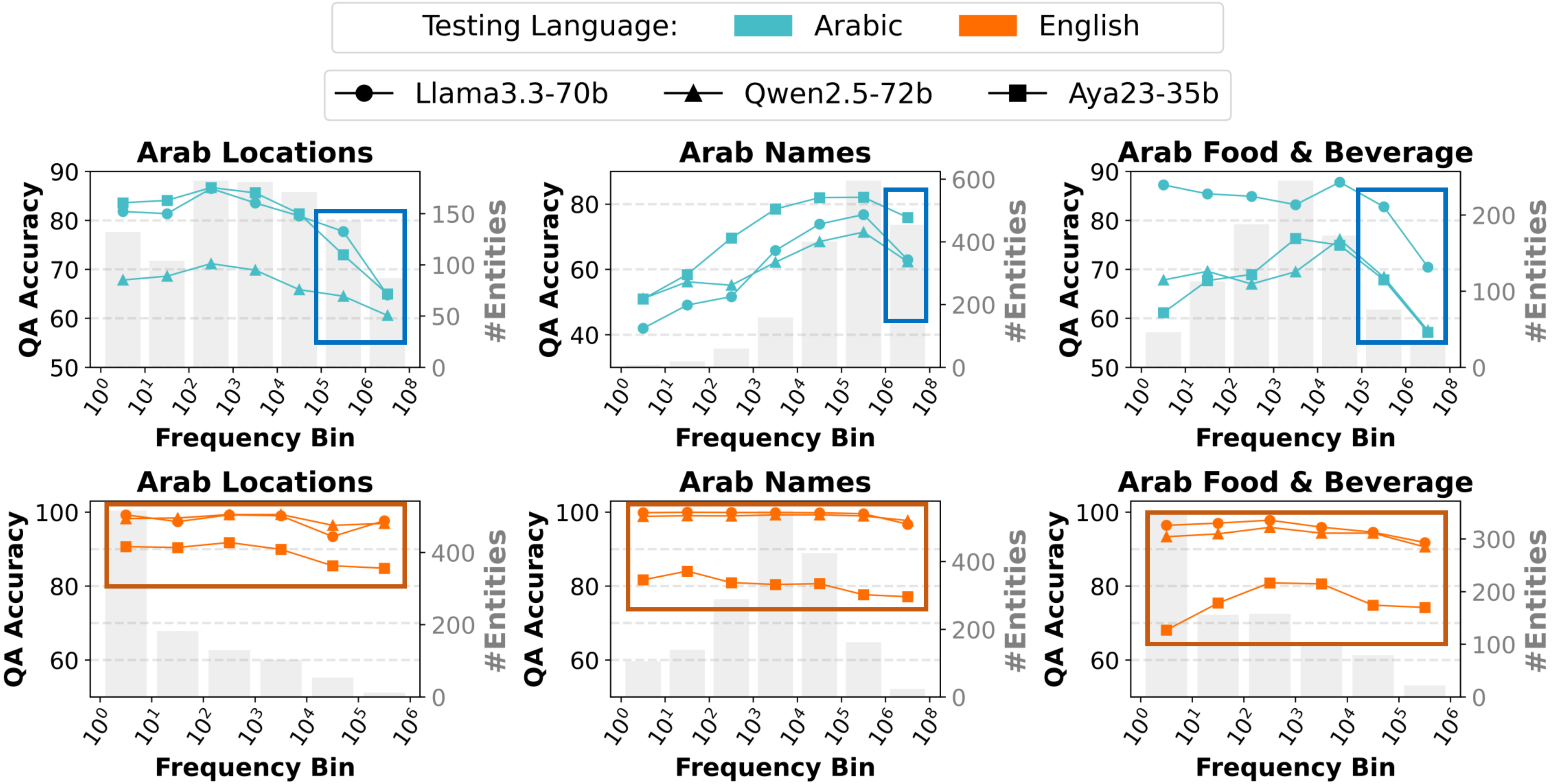
Better context adaptation in English than Arabic

What causes this disparity between both languages ?

On the Origin of Biases

1 **Frequency in Pre-training Data**

Do we perform better on higher frequency entities?



Performance drops in Arabic on entities that appear at very high frequencies (>1M times)

Performance is much more stable in English where we don’t see similar steep drops at high frequencies

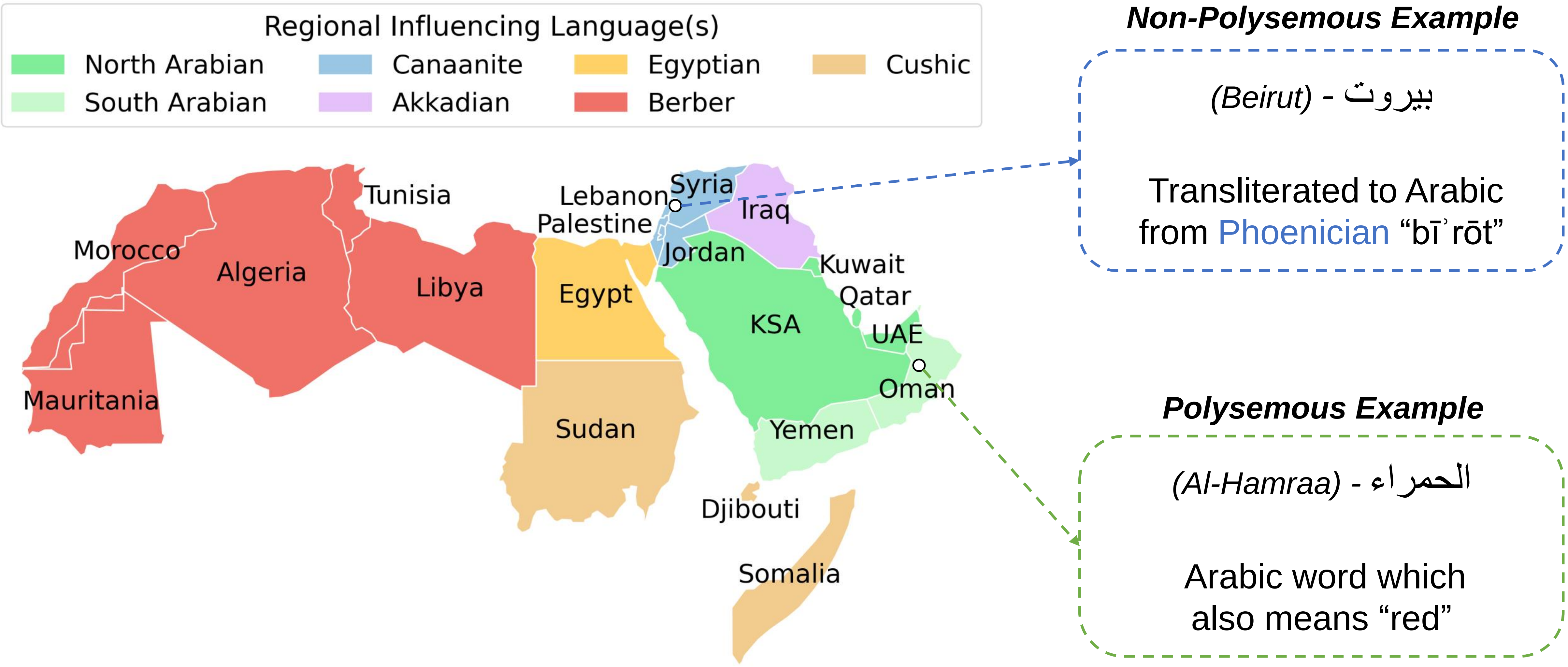
On the Origin of Biases

1 Frequency in Pre-training Data

Do we perform better on higher frequency entities?

2 Impact of Entity Word Polysemy

What happens when entities exhibit polysemy?



Non-Polysemous Example

(Beirut) - بيروت

Transliterated to Arabic from Phoenician “bīʾrōt”

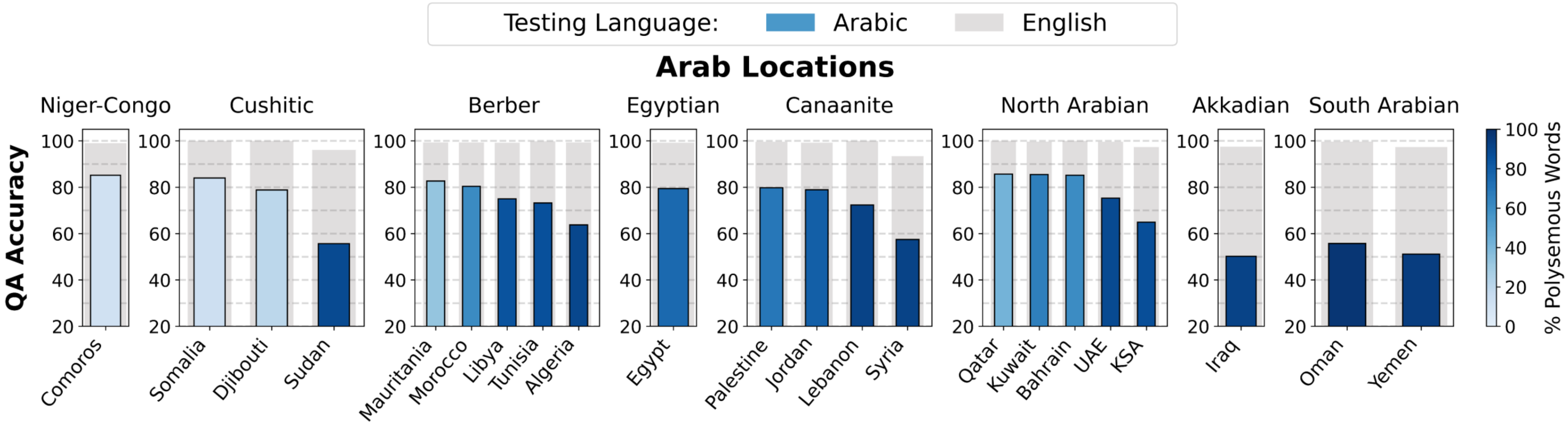
Polysemous Example

(Al-Hamraa) - الحمراء

Arabic word which also means “red”

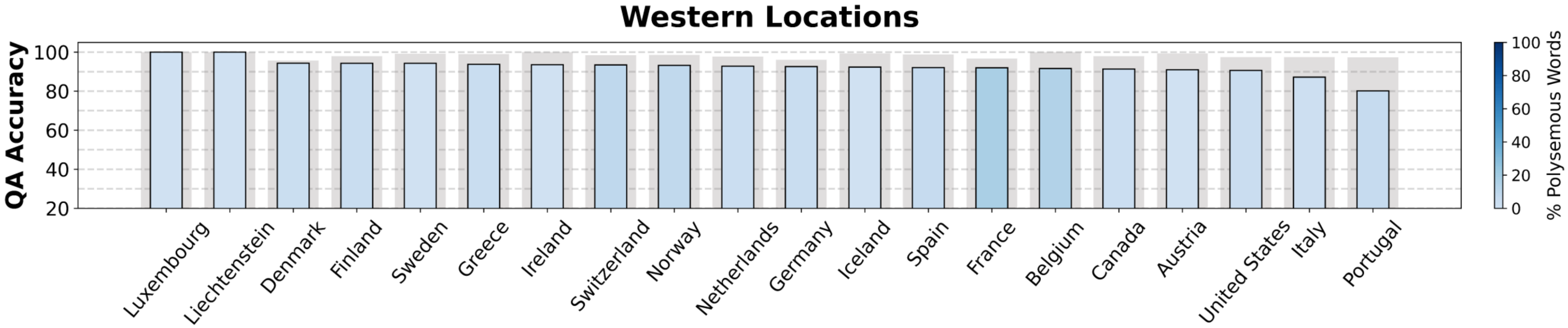
Locations in Arab countries can be ***non-polysemous transliterations*** or ***polysemous Arabic words***

Great testing setup to analyze the robustness of models to word polysemy when recognizing entities



Darker blue color reflect higher percentage of polysemous entities

Performance drops as more Arab entities exhibit Arabic polysemy



Performance is stable for Western entities in Arabic since they are transliterations with no other sense

On the Origin of Biases

1 Frequency in Pre-training Data

Do we perform better on higher frequency entities?

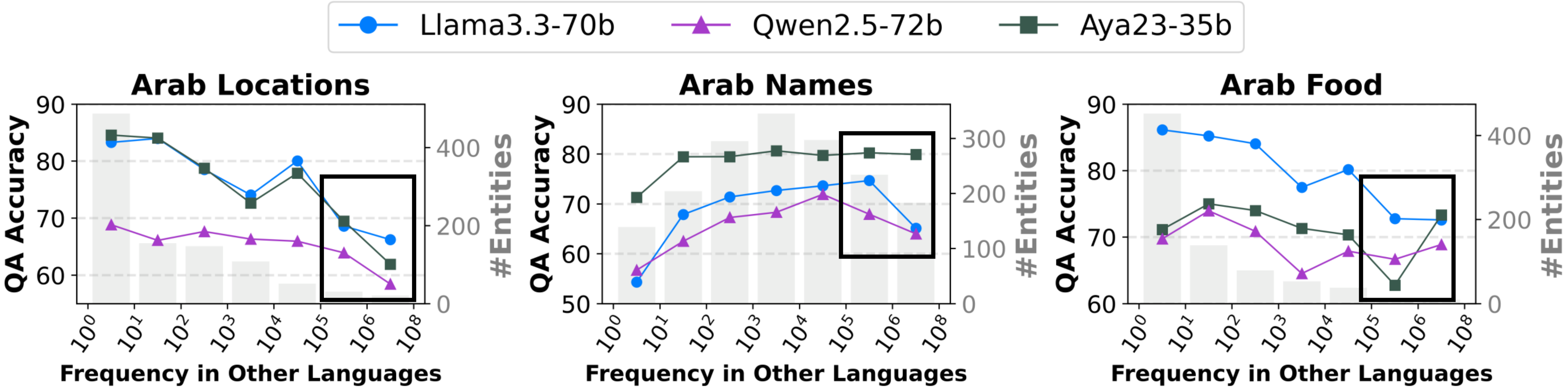
2 Impact of Entity Word Polysemy

What happens when entities exhibit polysemy?

3 Overlaps with other Languages

Does overlap of entities with other languages matter?

We analyze the impact of overlaps between Arab entities and words in languages that use Arabic script:
Farsi, *Urdu*, *Tajik*, *Kurdish*, *Pashto*



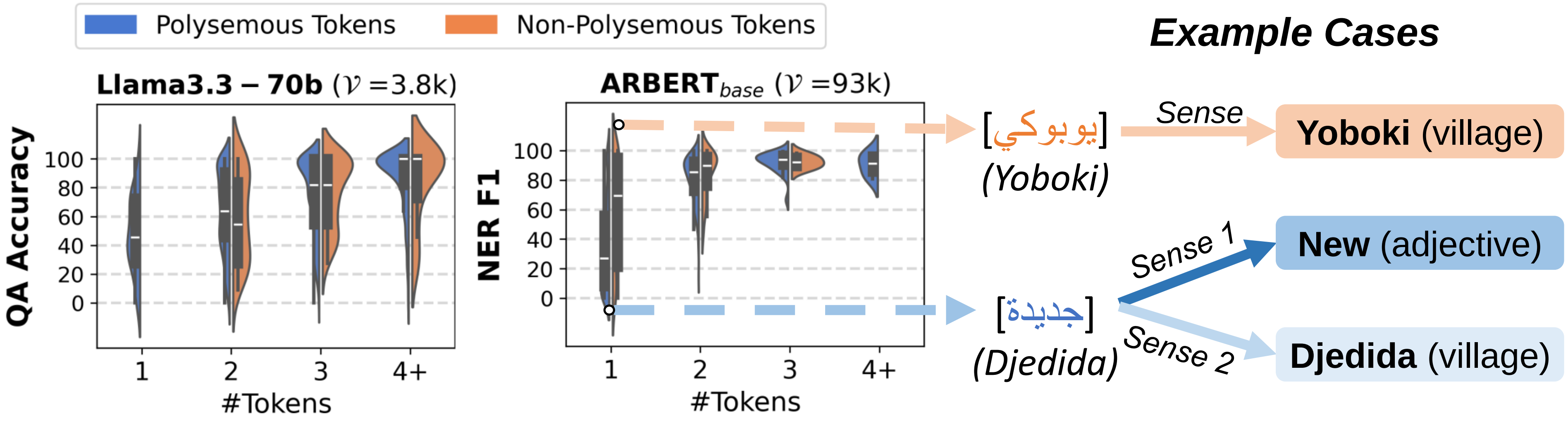
Performance drops as more Arab entities appear in high frequency in other languages

Polysemy in Arabic	**Arabic** جدتي تسكن في مطروحة My grandmother lives in Matrooha	**Arabic** القضية مطروحة للنقاش The issue is proposed for discussion
Polysemy with other languages	**Arabic** كنت أزور وزان هذا الأسبوع I was visiting Ouzanne this week	**Farsi** شاعر با دقت وزان شعر خود را بررسی کرد The poet carefully checked the weight of her poem
Polysemy with transliterations	**Arabic** لقد اشتريت بن من اليمن I bought coffee from Yemen	**Arabic** التقيت برجل اسمه بن يوم أمس I met a guy named Ben yesterday

These are going to be tokenized by the tokenization algorithm in the same manner

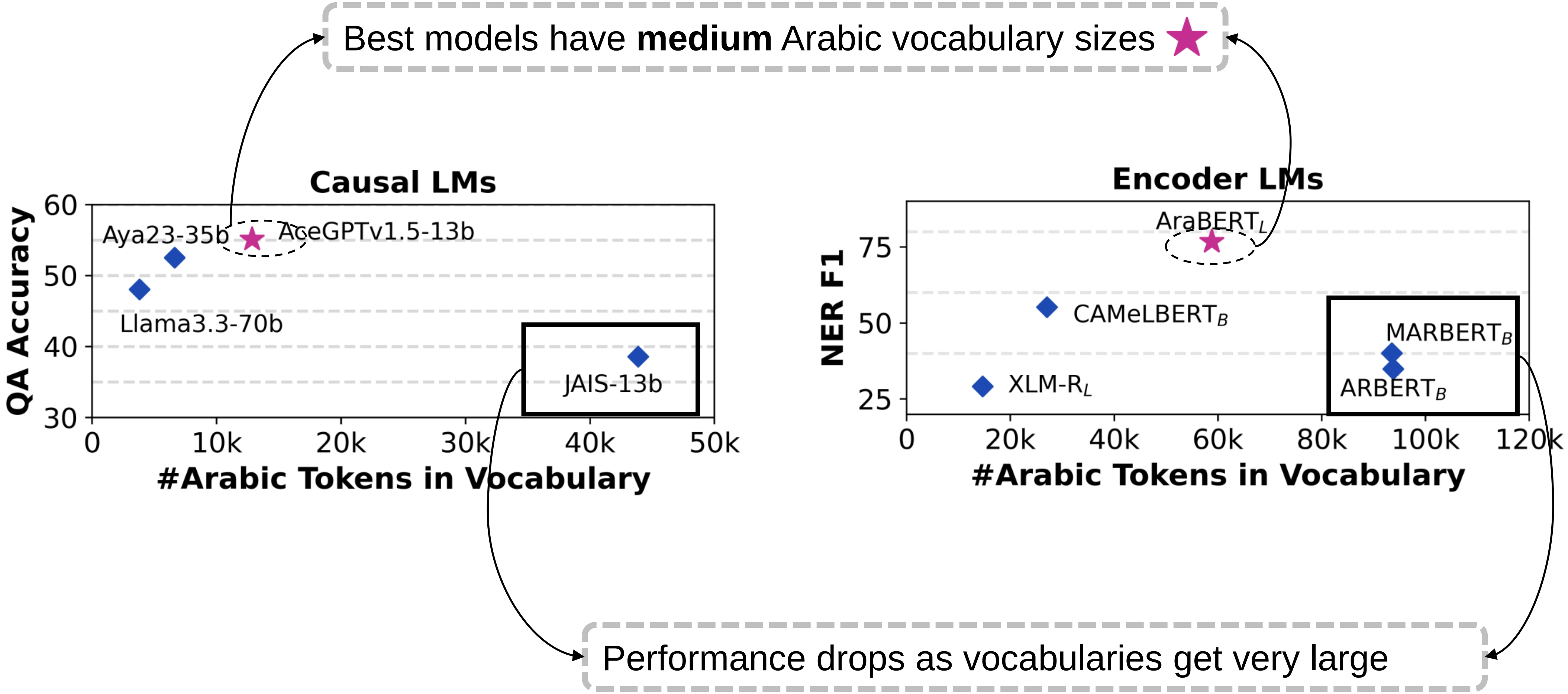
On the Origin of Biases

1	Frequency in Pre-training Data	Do we perform better on higher frequency entities?
2	Impact of Entity Word Polysemy	What happens when entities exhibit polysemy?
3	Overlaps with other Languages	Does overlap of entities with other languages matter?
4	**Sub-word Tokenization**	**How does sub-word tokenization impact things?**



*Performance is worst when entities are tokenized into single tokens **and** exhibit polysemy*

Models don't struggle as much when entities are split into 3 or more tokens



Best models have **medium** Arabic vocabulary sizes ★

Performance drops as vocabularies get very large

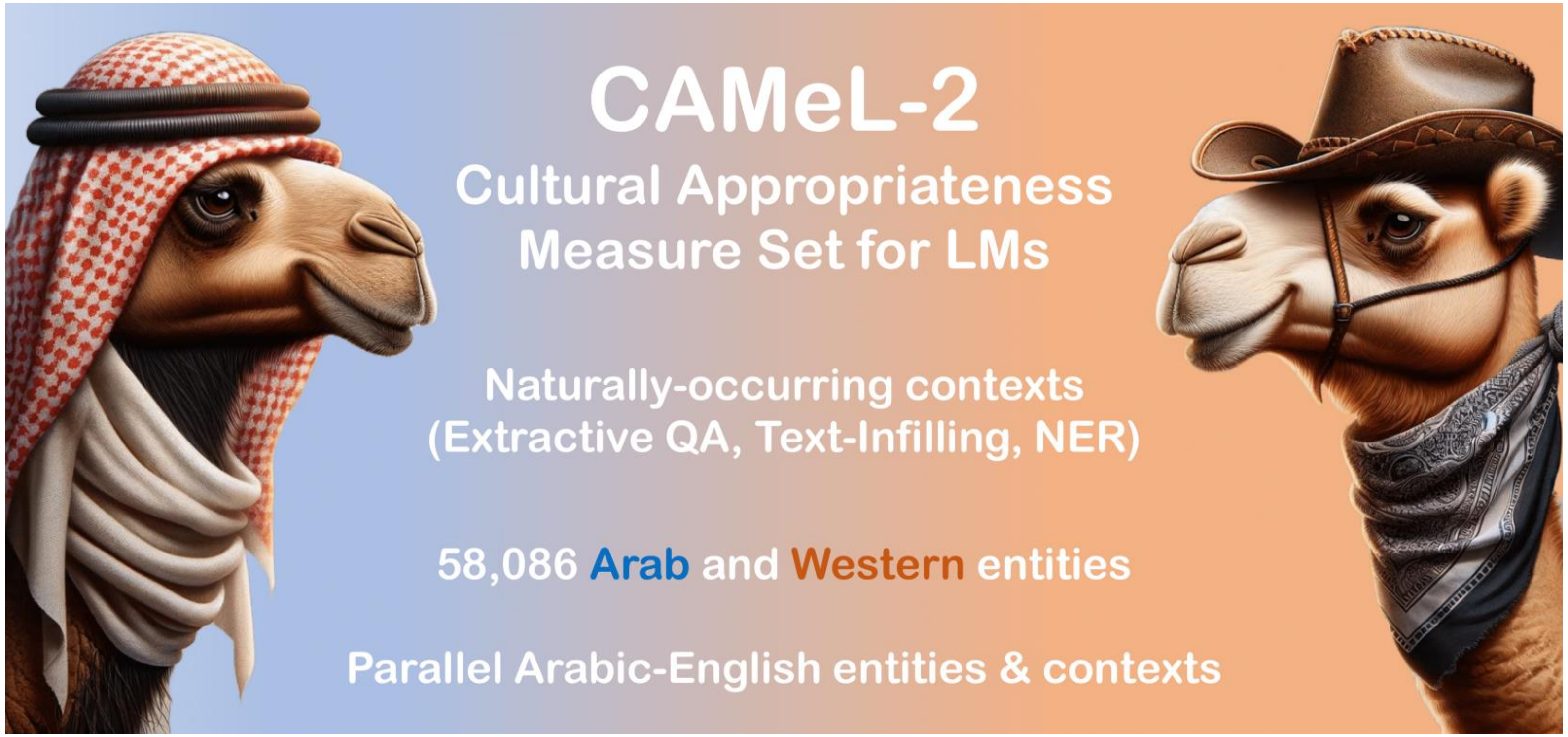
Takeaways

- Non-English linguistic phenomena contributes to cross-cultural performance gaps in LLMs
  - Models are struggle to distinguish entity vs non-entity senses (within and across languages)
  - This can lead to a perceived Western bias in models

- Tokenization plays an important role
  - Need better ways to tokenize entities that hold multiple sense to enhance model performance

ขอบคุณ شكرا Merci 谢谢 धन्यवाद Asante Teşekkürler

ありがとう Gracias متشكرم நன்றி Obrigado Thank You



CAMeL-2

Cultural Appropriateness Measure Set for LMs

Naturally-occurring contexts (Extractive QA, Text-Infilling, NER)

58,086 Arab and Western entities

Parallel Arabic-English entities & contexts

CAMeL-2 is available at: https://github.com/tareknaous/camel2

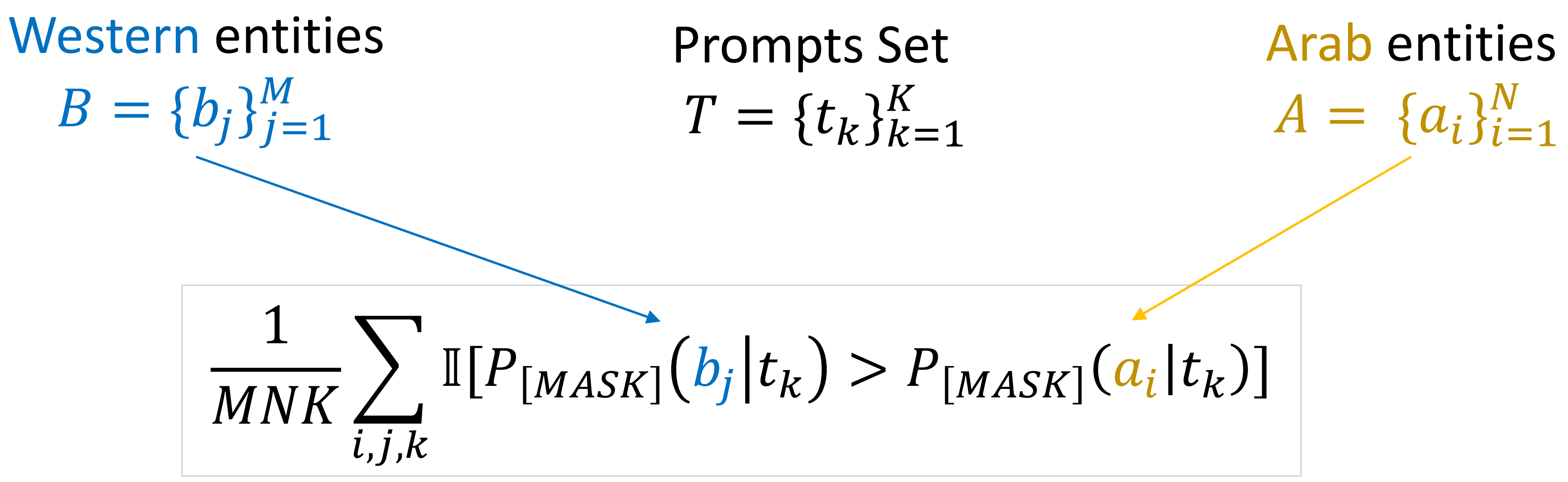
Feel free to follow up with me on @tareknaous

Additional Slides

Cultural Bias Score - How often do LLMs prefer Western entities?

My grandma is Arab, for dinner she always makes us [MASK]

$P_{[MASK]}$ (Lasagna) >? $P_{[MASK]}$(Majboos)



Western entities

$$B = \{b_j\}_{j=1}^{M}$$

Prompts Set

$$T = \{t_k\}_{k=1}^{K}$$

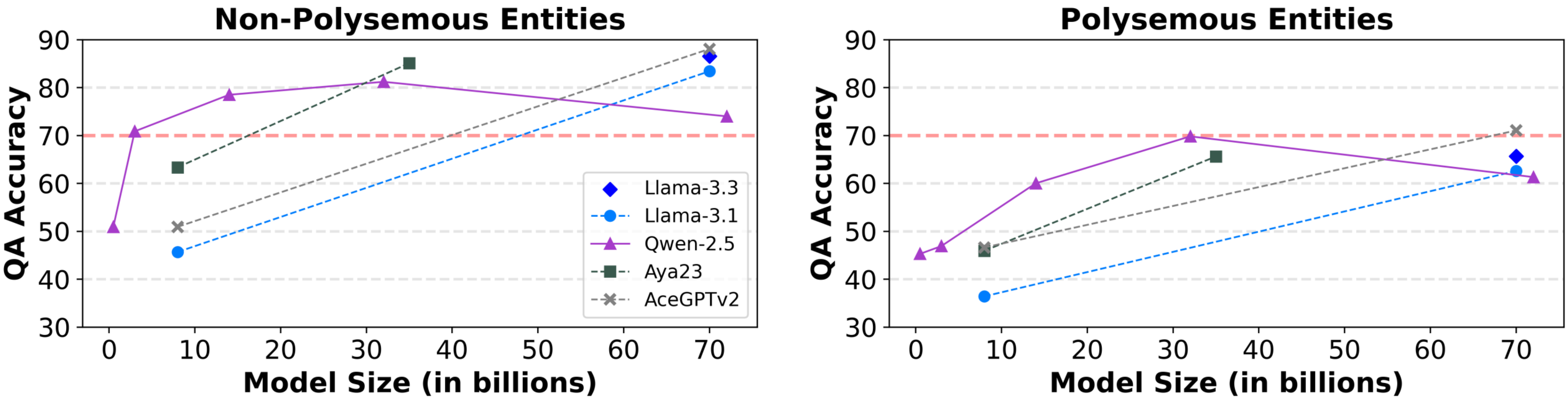
Arab entities

$$A = \{a_i\}_{i=1}^{N}$$

$$\frac{1}{MNK}\sum_{i,j,k}\mathbb{I}[P_{[MASK]}(b_j|t_k) > P_{[MASK]}(a_i|t_k)]$$

Cultural Bias Score (0-100%):

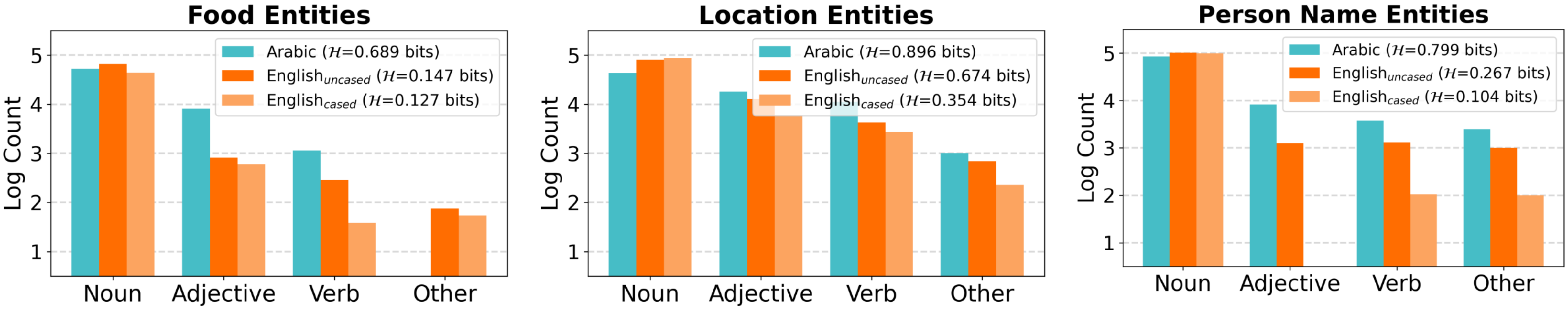
Scaling Trends



Non-Polysemous Entities

Polysemous Entities

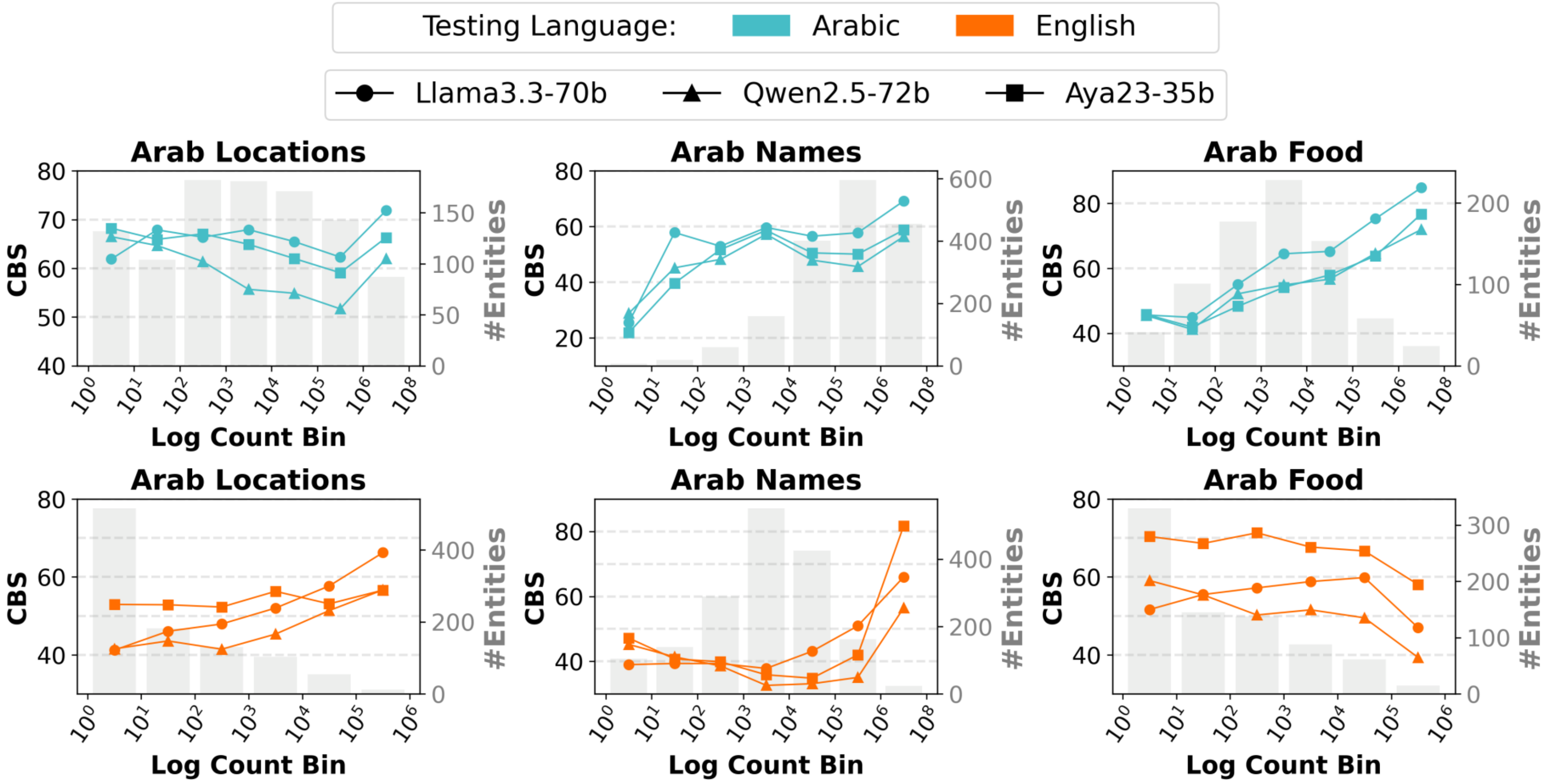
POS Tag Distributions of Entities in Arabic vs English



Entities in Arabic have high entropy in terms of their grammatical roles in natural language

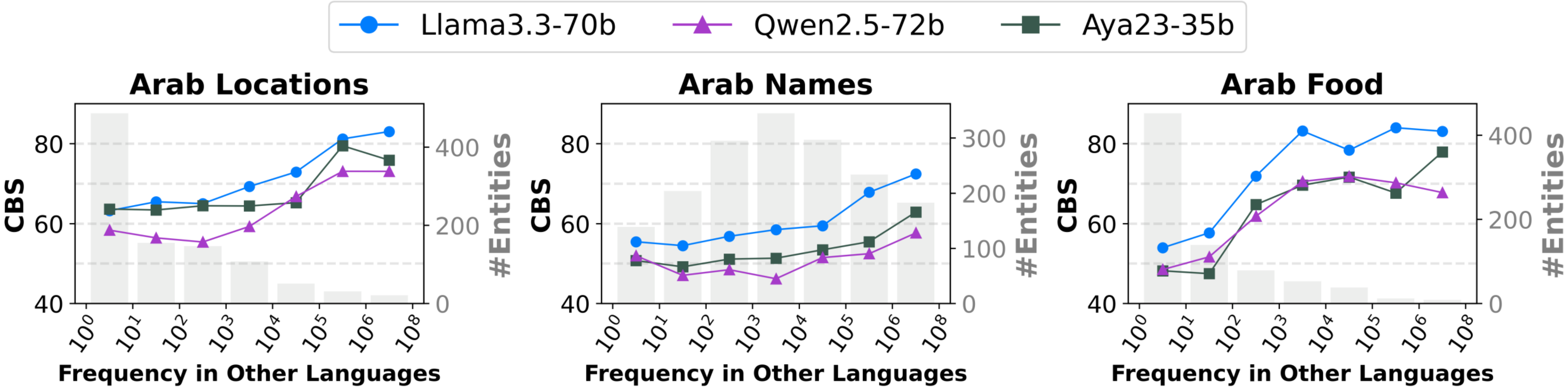
Entities in English have lower entropy, which decreases further when capitalization is used

Cultural Adaptation Results – Frequency in Pre-training



Similar trend to the extractive QA task, where the CBS increases for high frequency entities

Cultural Adaptation Results – Frequency in Other Languages



Similar trend to Extractive QA, where we find that the CBS increases with more overlap