QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Native, Local and Cultural Knowledge in LLMs

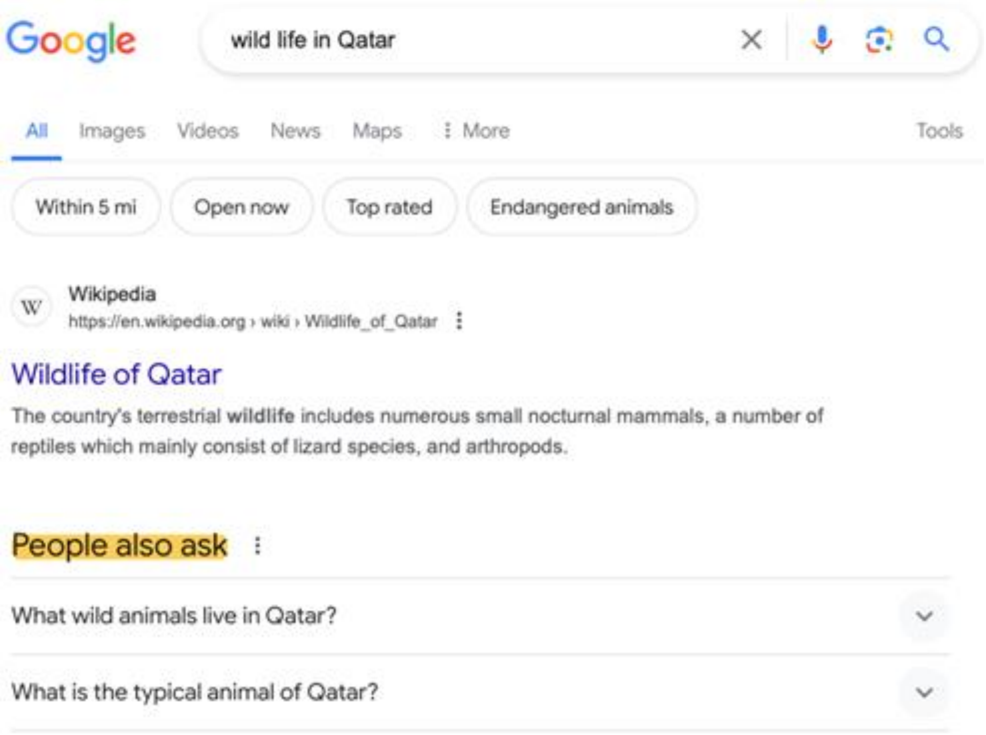
Firoj Alam

Qatar Computing Research Institute

BoF Session: Multilingual and Multimodal Cultural Inclusivity in LLMs

23 January, 2025

Native, Local and Cultural Knowledge



Wikipedia
https://en.wikipedia.org › wiki › Wildlife_of_Qatar

Wildlife of Qatar

The country's terrestrial **wildlife** includes numerous small nocturnal mammals, a number of reptiles which mainly consist of lizard species, and arthropods.

People also ask

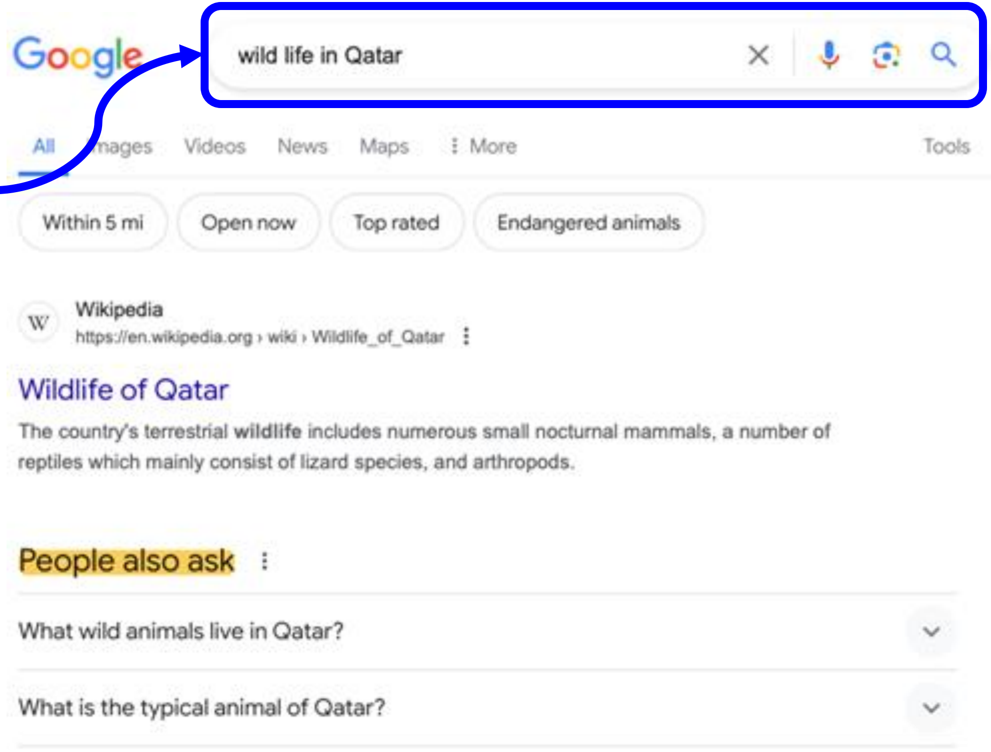
What wild animals live in Qatar?

What is the typical animal of Qatar?

NativQA: Multilingual Culturally-Aligned Natural Queries for LLMs
Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, Firoj Alam

Native, Local and Cultural Knowledge

Local information asked by **native/non-native** speaker



wild life in Qatar

Wikipedia
https://en.wikipedia.org › wiki › Wildlife_of_Qatar

Wildlife of Qatar

The country's terrestrial **wildlife** includes numerous small nocturnal mammals, a number of reptiles which mainly consist of lizard species, and arthropods.

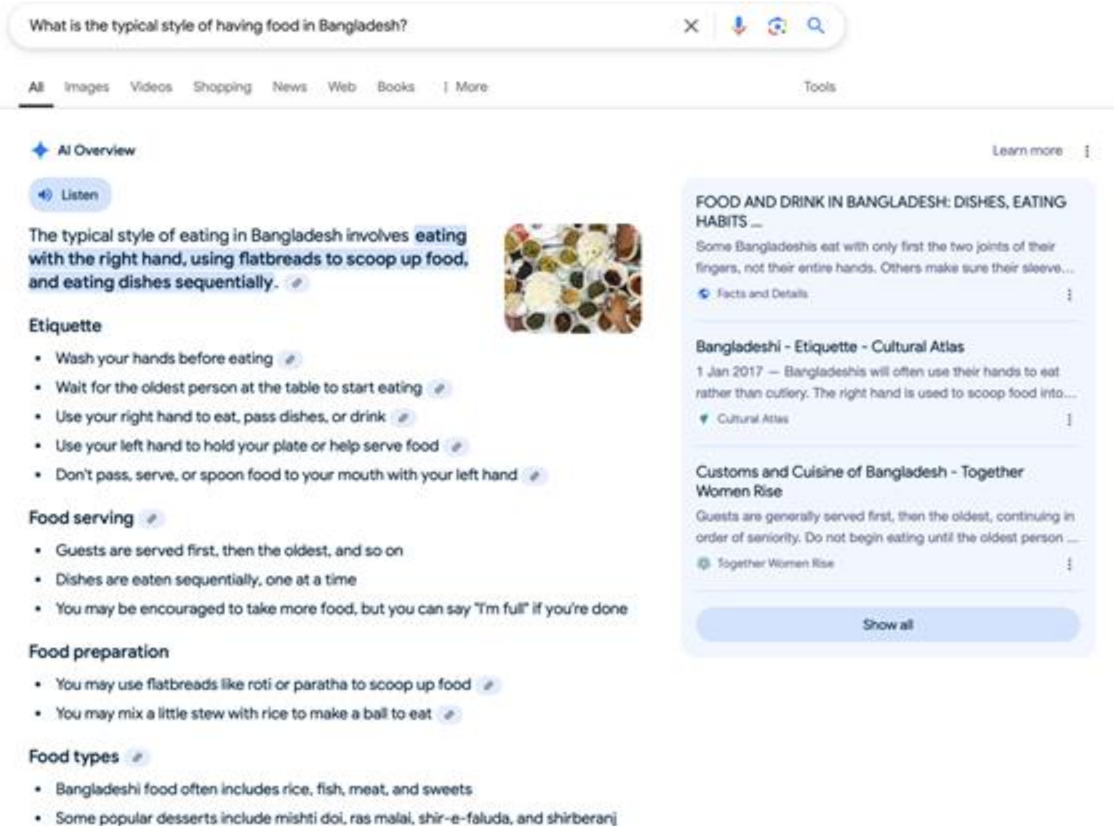
What wild animals live in Qatar?

What is the typical animal of Qatar?

NativQA: Multilingual Culturally-Aligned Natural Queries for LLMs
Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, Firoj Alam

Native, Local and Cultural Knowledge

What is the typical style of having food in Bangladesh?



The typical style of eating in Bangladesh involves **eating with the right hand, using flatbreads to scoop up food, and eating dishes sequentially**.

Etiquette

- Wash your hands before eating
- Wait for the oldest person at the table to start eating
- Use your right hand to eat, pass dishes, or drink
- Use your left hand to hold your plate or help serve food
- Don't pass, serve, or spoon food to your mouth with your left hand

Food serving

- Guests are served first, then the oldest, and so on
- Dishes are eaten sequentially, one at a time
- You may be encouraged to take more food, but you can say "I'm full" if you're done

Food preparation

- You may use flatbreads like roti or paratha to scoop up food
- You may mix a little stew with rice to make a ball to eat

Food types

- Bangladeshi food often includes rice, fish, meat, and sweets
- Some popular desserts include mishti doi, ras malai, shir-e-faluda, and shirberanj

FOOD AND DRINK IN BANGLADESH: DISHES, EATING HABITS ...

Some Bangladeshis eat with only first the two joints of their fingers, not their entire hands. Others make sure their sleeve...

Facts and Details

Bangladeshi - Etiquette - Cultural Atlas

1 Jan 2017 — Bangladeshis will often use their hands to eat rather than cutlery. The right hand is used to scoop food into...

Cultural Atlas

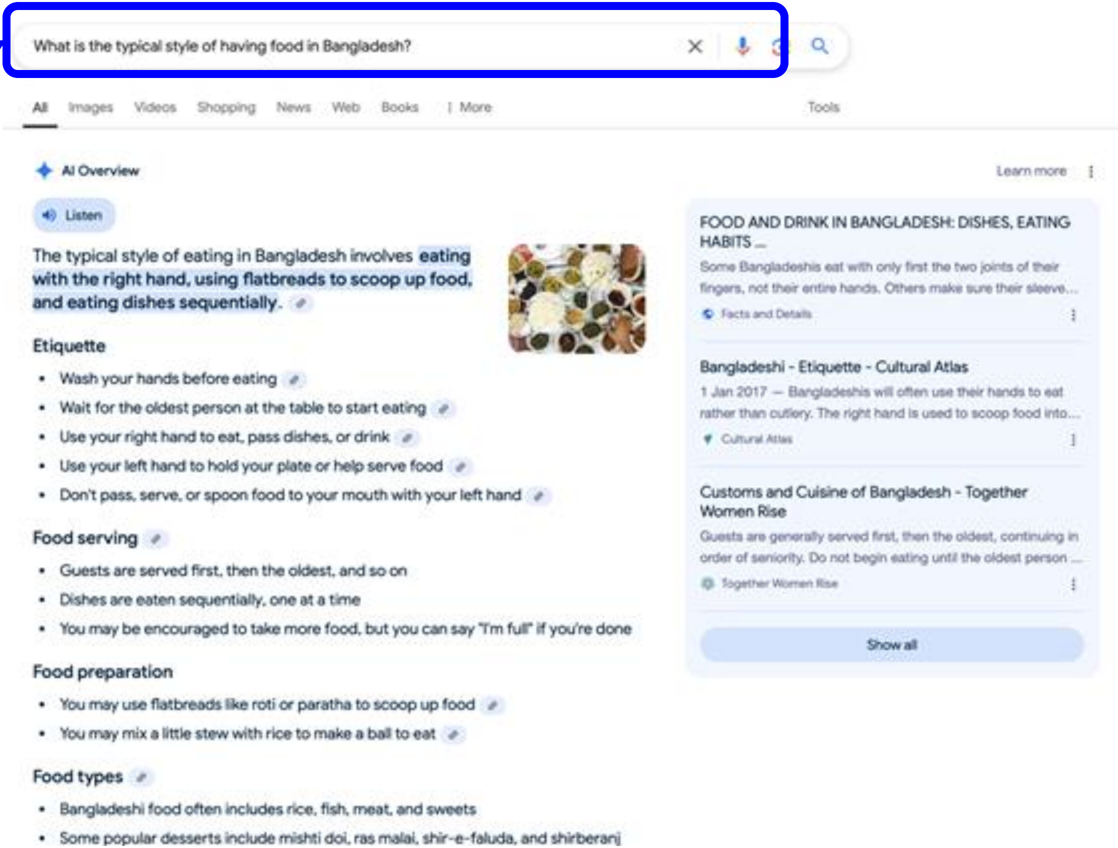
Customs and Cuisine of Bangladesh - Together Women Rise

Guests are generally served first, then the oldest, continuing in order of seniority. Do not begin eating until the oldest person ...

Together Women Rise

Native, Local and Cultural Knowledge

Local and **Cultural** information asked by **native/non-native** speaker



The typical style of eating in Bangladesh involves **eating with the right hand, using flatbreads to scoop up food, and eating dishes sequentially.**

Etiquette

- Wash your hands before eating
- Wait for the oldest person at the table to start eating
- Use your right hand to eat, pass dishes, or drink
- Use your left hand to hold your plate or help serve food
- Don't pass, serve, or spoon food to your mouth with your left hand

Food serving

- Guests are served first, then the oldest, and so on
- Dishes are eaten sequentially, one at a time
- You may be encouraged to take more food, but you can say "I'm full" if you're done

Food preparation

- You may use flatbreads like roti or paratha to scoop up food
- You may mix a little stew with rice to make a ball to eat

Food types

- Bangladeshi food often includes rice, fish, meat, and sweets
- Some popular desserts include mishti doi, ras malai, shir-e-faluda, and shirberanj

FOOD AND DRINK IN BANGLADESH: DISHES, EATING HABITS ...

Some Bangladeshis eat with only first the two joints of their fingers, not their entire hands. Others make sure their sleeve...

Facts and Details

Bangladeshi - Etiquette - Cultural Atlas

1 Jan 2017 — Bangladeshis will often use their hands to eat rather than cutlery. The right hand is used to scoop food into...

Cultural Atlas

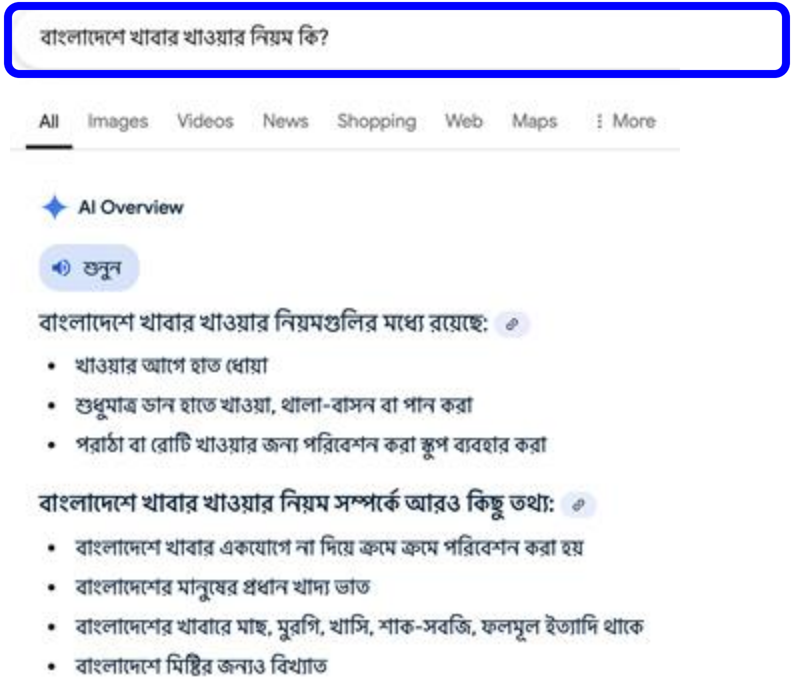
Customs and Cuisine of Bangladesh - Together Women Rise

Guests are generally served first, then the oldest, continuing in order of seniority. Do not begin eating until the oldest person ...

Together Women Rise

NativQA: Multilingual Culturally-Aligned Natural Queries for LLMs
Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, Firoj Alam

Native, Local and Cultural Knowledge



বাংলাদেশে খাবার খাওয়ার নিয়ম কি?

বাংলাদেশে খাবার খাওয়ার নিয়মগুলির মধ্যে রয়েছে:

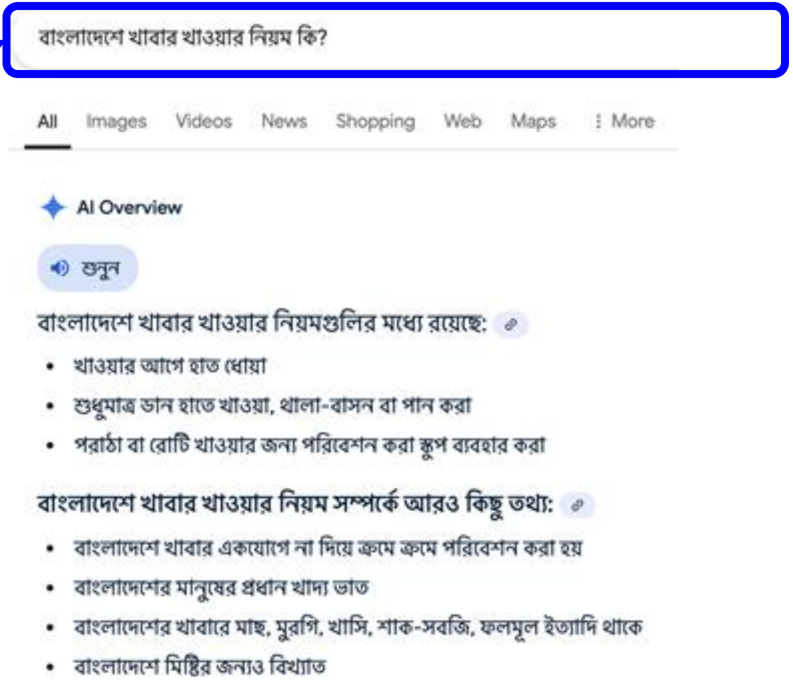
- খাওয়ার আগে হাত ধোয়া
- শুধুমাত্র ডান হাতে খাওয়া, থালা-বাসন বা পান করা
- পরাঠা বা রোটি খাওয়ার জন্য পরিবেশন করা স্কুপ ব্যবহার করা

বাংলাদেশে খাবার খাওয়ার নিয়ম সম্পর্কে আরও কিছু তথ্য:

- বাংলাদেশে খাবার একযোগে না দিয়ে ক্রমে ক্রমে পরিবেশন করা হয়
- বাংলাদেশের মানুষের প্রধান খাদ্য ভাত
- বাংলাদেশের খাবারে মাছ, মুরগি, খাসি, শাক-সবজি, ফলমূল ইত্যাদি থাকে
- বাংলাদেশে মিষ্টির জন্যও বিখ্যাত

Native, Local and Cultural Knowledge

Local and **Cultural** information asked by **native** speaker



বাংলাদেশে খাবার খাওয়ার নিয়ম কি?

বাংলাদেশে খাবার খাওয়ার নিয়মগুলির মধ্যে রয়েছে:

- খাওয়ার আগে হাত ধোয়া
- শুধুমাত্র ডান হাতে খাওয়া, থালা-বাসন বা পান করা
- পরাঠা বা রোটি খাওয়ার জন্য পরিবেশন করা স্কুপ ব্যবহার করা

বাংলাদেশে খাবার খাওয়ার নিয়ম সম্পর্কে আরও কিছু তথ্য:

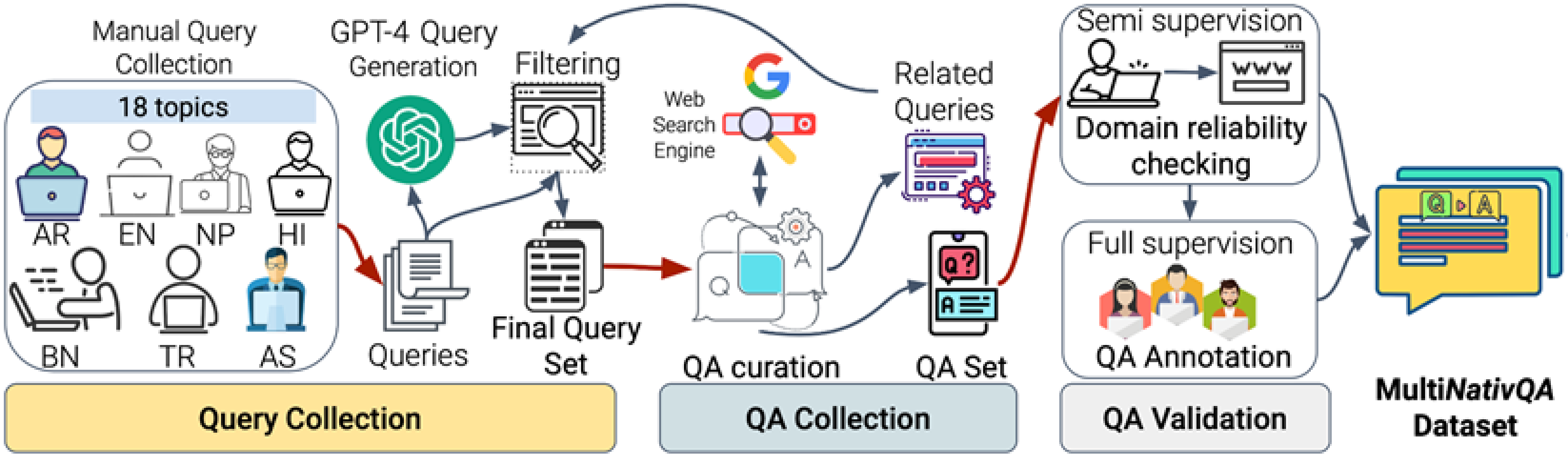
- বাংলাদেশে খাবার একযোগে না দিয়ে ক্রমে ক্রমে পরিবেশন করা হয়
- বাংলাদেশের মানুষের প্রধান খাদ্য ভাত
- বাংলাদেশের খাবারে মাছ, মুরগি, খাসি, শাক-সবজি, ফলমূল ইত্যাদি থাকে
- বাংলাদেশে মিষ্টির জন্যও বিখ্যাত

Native, Local, and Cultural Knowledge

LLMs should be capable of:

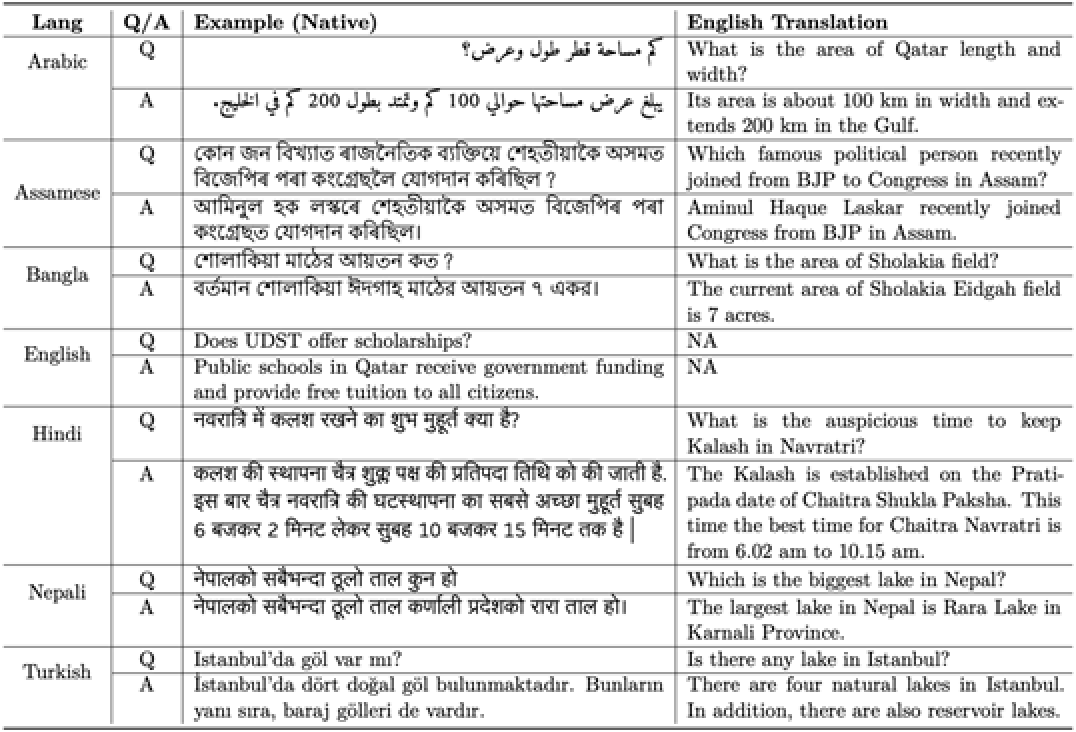
- Answering questions posed by native speakers in their own language.
- Providing information tailored to specific regions or locations.

NativQA Framework



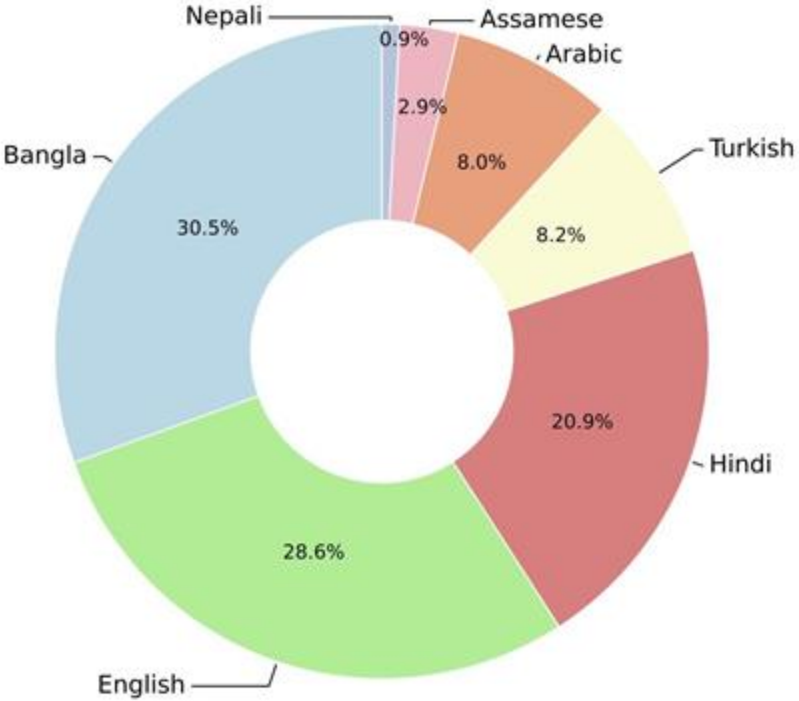
NativQA: Multilingual Culturally-Aligned Natural Queries for LLMs
Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, Firoj Alam

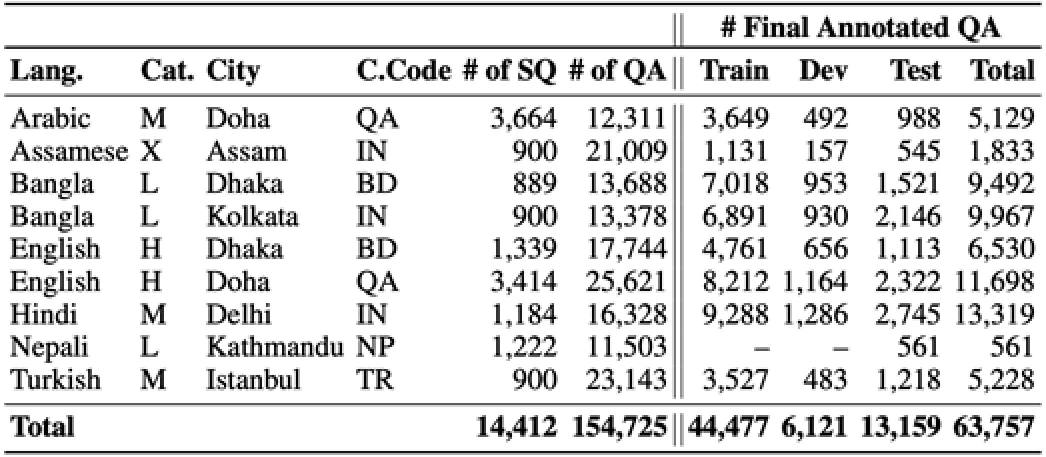
MultiNativQA Dataset



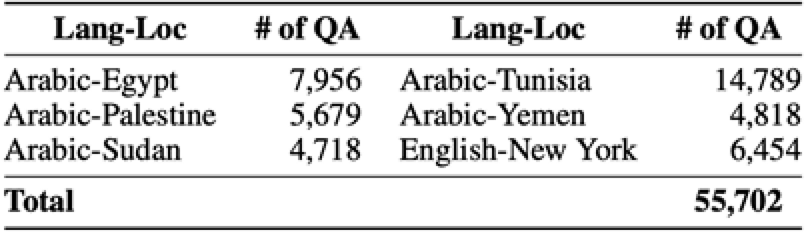
Lang	Q/A	Example (Native)	English Translation
Arabic	Q	كم مساحة قطر طول وعرض؟	What is the area of Qatar length and width?
	A	يبلغ عرض مساحتها حوالي 100 كم وتمتد بطول 200 كم في الخليج.	Its area is about 100 km in width and extends 200 km in the Gulf.
Assamese	Q	কোন জন বিখ্যাত ৰাজনৈতিক ব্যক্তিয়ে শেহতীয়াকৈ অসমত বিজেপিৰ পৰা কংগ্ৰেছলৈ যোগদান কৰিছিল ?	Which famous political person recently joined from BJP to Congress in Assam?
	A	আমিনুল হক লস্কৰে শেহতীয়াকৈ অসমত বিজেপিৰ পৰা কংগ্ৰেছত যোগদান কৰিছিল।	Aminul Haque Laskar recently joined Congress from BJP in Assam.
Bangla	Q	শোলাকিয়া মাঠের আয়তন কত ?	What is the area of Sholakia field?
	A	বর্তমান শোলাকিয়া ঈদগাহ্ মাঠের আয়তন ৭ একর।	The current area of Sholakia Eidgah field is 7 acres.
English	Q	Does UDST offer scholarships?	NA
	A	Public schools in Qatar receive government funding and provide free tuition to all citizens.	NA
Hindi	Q	नवरात्रि में कलश रखने का शुभ मुहूर्त क्या है?	What is the auspicious time to keep Kalash in Navratri?
	A	कलश की स्थापना चैत्र शुक्ल पक्ष की प्रतिपदा तिथि को की जाती है. इस बार चैत्र नवरात्रि की घटस्थापना का सबसे अच्छा मुहूर्त सुबह 6 बजकर 2 मिनट लेकर सुबह 10 बजकर 15 मिनट तक है |	The Kalash is established on the Pratipada date of Chaitra Shukla Paksha. This time the best time for Chaitra Navratri is from 6.02 am to 10.15 am.
Nepali	Q	नेपालको सबैभन्दा ठूलो ताल कुन हो	Which is the biggest lake in Nepal?
	A	नेपालको सबैभन्दा ठूलो ताल कर्णाली प्रदेशको रारा ताल हो।	The largest lake in Nepal is Rara Lake in Karnali Province.
Turkish	Q	Istanbul'da göl var mı?	Is there any lake in Istanbul?
	A	İstanbul'da dört doğal göl bulunmaktadır. Bunların yanı sıra, baraj gölleri de vardır.	There are four natural lakes in Istanbul. In addition, there are also reservoir lakes.

MultiNativQA Dataset





						# Final Annotated QA			
Lang.	**Cat.**	**City**	**C.Code**	**# of SQ**	**# of QA**	**Train**	**Dev**	**Test**	**Total**
Arabic	M	Doha	QA	3,664	12,311	3,649	492	988	5,129
Assamese	X	Assam	IN	900	21,009	1,131	157	545	1,833
Bangla	L	Dhaka	BD	889	13,688	7,018	953	1,521	9,492
Bangla	L	Kolkata	IN	900	13,378	6,891	930	2,146	9,967
English	H	Dhaka	BD	1,339	17,744	4,761	656	1,113	6,530
English	H	Doha	QA	3,414	25,621	8,212	1,164	2,322	11,698
Hindi	M	Delhi	IN	1,184	16,328	9,288	1,286	2,745	13,319
Nepali	L	Kathmandu	NP	1,222	11,503	–	–	561	561
Turkish	M	Istanbul	TR	900	23,143	3,527	483	1,218	5,228
Total				**14,412**	**154,725**	**44,477**	**6,121**	**13,159**	**63,757**



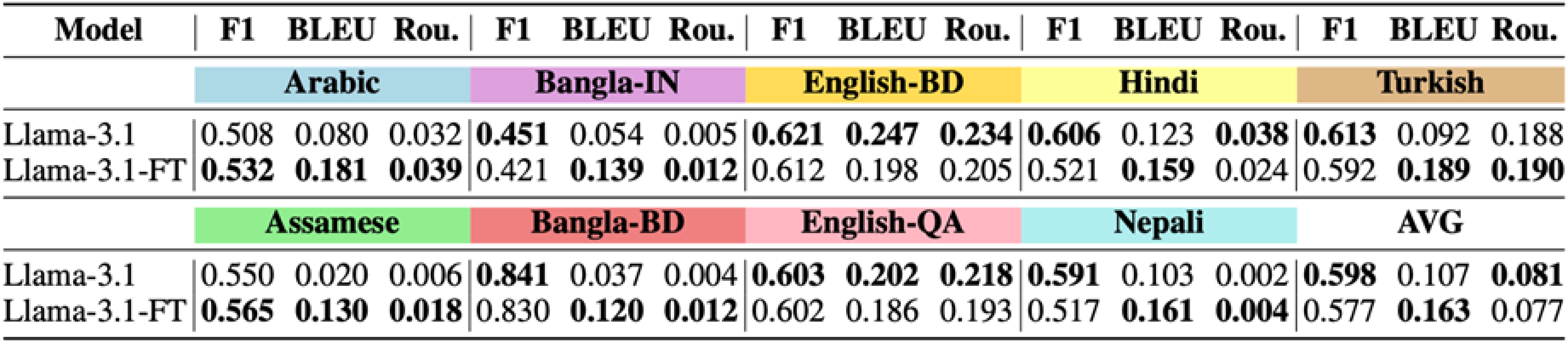
Lang-Loc	**# of QA**	**Lang-Loc**	**# of QA**
Arabic-Egypt	7,956	Arabic-Tunisia	14,789
Arabic-Palestine	5,679	Arabic-Yemen	4,818
Arabic-Sudan	4,718	English-New York	6,454
Total			**55,702**

NativQA: Multilingual Culturally-Aligned Natural Queries for LLMs
Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, Firoj Alam

MultiNativQA Dataset

https://gitlab.com/nativqa/nativqa-framework/
https://huggingface.co/datasets/QCRI/MultiNativQA
https://arxiv.org/pdf/2407.09823

MultiNativQA: Results

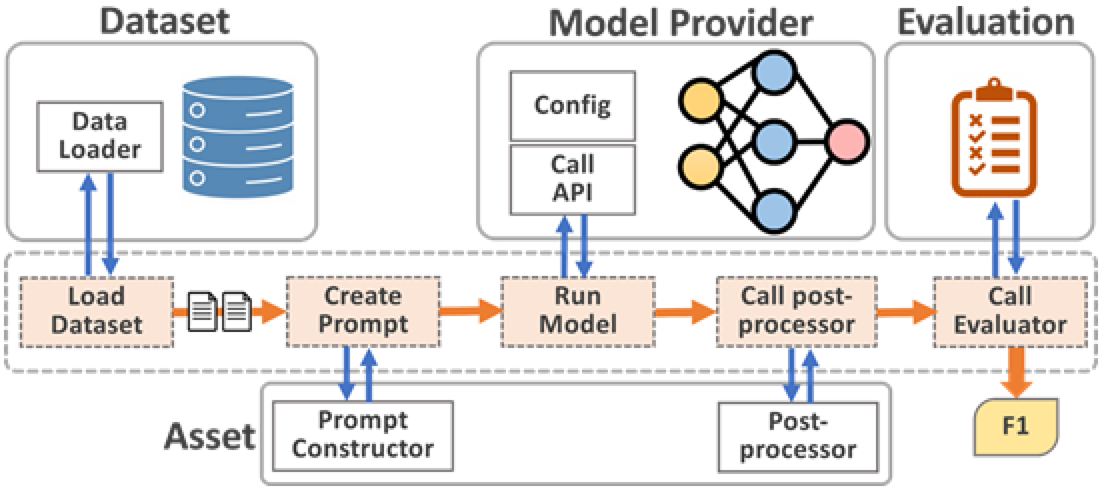


Model	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.	F1	BLEU	Rou.
	Arabic			**Bangla-IN**			**English-BD**			**Hindi**			**Turkish**		
Llama-3.1	0.508	0.080	0.032	**0.451**	0.054	0.005	**0.621**	**0.247**	**0.234**	**0.606**	0.123	**0.038**	**0.613**	0.092	0.188
Llama-3.1-FT	**0.532**	**0.181**	**0.039**	0.421	**0.139**	**0.012**	0.612	0.198	0.205	0.521	**0.159**	0.024	0.592	**0.189**	**0.190**
	Assamese			**Bangla-BD**			**English-QA**			**Nepali**			**AVG**		
Llama-3.1	0.550	0.020	0.006	**0.841**	0.037	0.004	**0.603**	**0.202**	**0.218**	**0.591**	0.103	0.002	**0.598**	0.107	**0.081**
Llama-3.1-FT	**0.565**	**0.130**	**0.018**	0.830	**0.120**	**0.012**	0.602	0.186	0.193	0.517	**0.161**	**0.004**	0.577	**0.163**	0.077

MultiNativQA Experimental Resources

LLMeBench: A Flexible Framework for Accelerating LLMs Benchmarking

(Dalvi et al., EACL 2024)



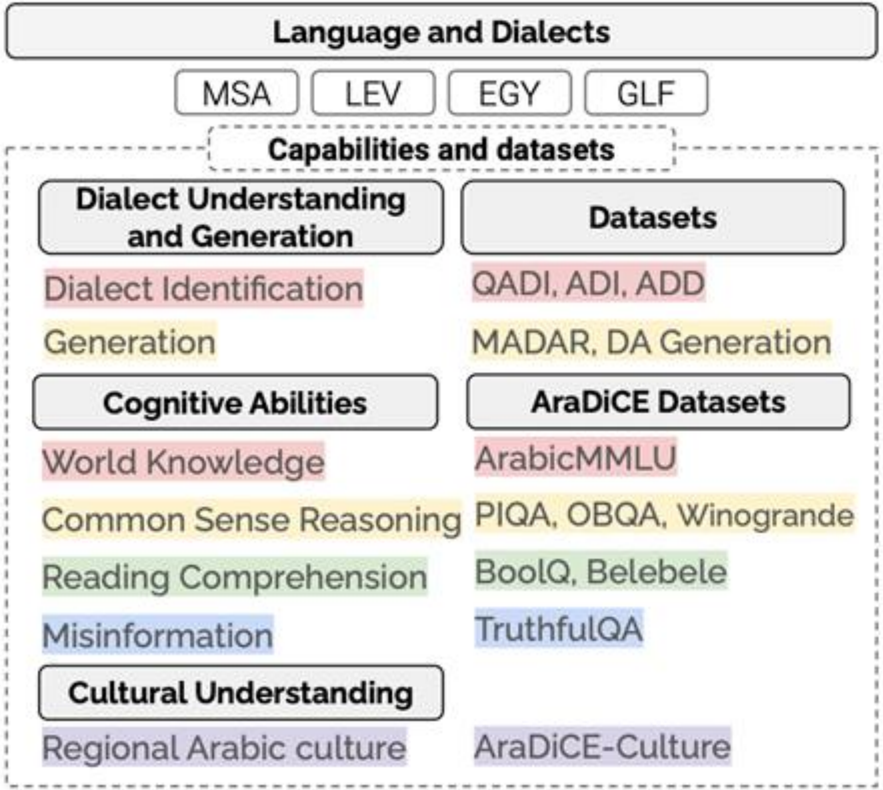
Getting started

LLMeBench Repo

https://llmebench.qcri.org

Check out the docs, examples, tutorials, source code and more!

AraDiCE:
Benchmarks for **Dialectal and Cultural Capabilities** in LLMs



Language and Dialects

MSA | LEV | EGY | GLF

Capabilities and datasets

Dialect Understanding and Generation	Datasets
Dialect Identification	QADI, ADI, ADD
Generation	MADAR, DA Generation

Cognitive Abilities	AraDiCE Datasets
World Knowledge	ArabicMMLU
Common Sense Reasoning	PIQA, OBQA, Winogrande
Reading Comprehension	BoolQ, Belebele
Misinformation	TruthfulQA

Cultural Understanding	
Regional Arabic culture	AraDiCE-Culture

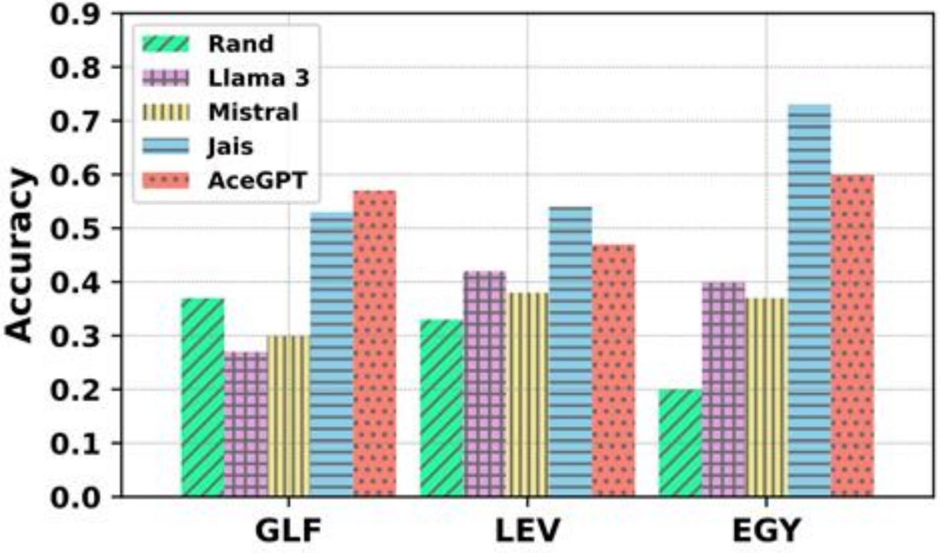
COLING-2025 AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs
Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, Firoj Alam

AraDiCE:
Benchmarks for **Dialectal and Cultural Capabilities** in LLMs

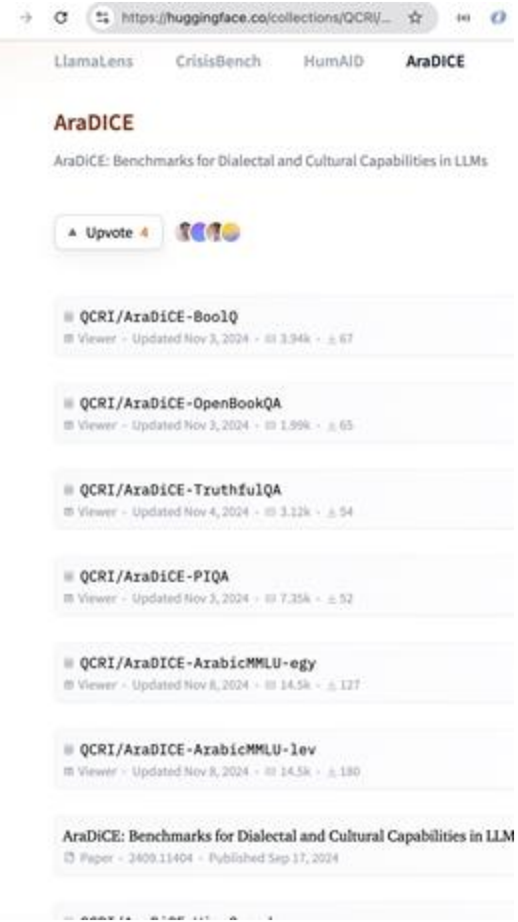
Cultural Understanding

- Results on the cultural evaluation show that Jais is the most culturally aligned model followed by AceGPT

- **Llama3** and **Mistral** generally show performances close to the random baseline, suggesting limited awareness of Arabic culture



AraDiCE:
Benchmarks for **Dialectal and Cultural Capabilities** in LLMs



AraDICE

AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs

QCRI/AraDiCE-BoolQ
Viewer · Updated Nov 3, 2024 · 3.94k · 67

QCRI/AraDiCE-OpenBookQA
Viewer · Updated Nov 3, 2024 · 1.99k · 65

QCRI/AraDiCE-TruthfulQA
Viewer · Updated Nov 4, 2024 · 3.12k · 54

QCRI/AraDiCE-PIQA
Viewer · Updated Nov 3, 2024 · 7.35k · 52

QCRI/AraDICE-ArabicMMLU-egy
Viewer · Updated Nov 8, 2024 · 14.5k · 127

QCRI/AraDICE-ArabicMMLU-lev
Viewer · Updated Nov 8, 2024 · 14.5k · 180

AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLM
Paper · 2409.11404 · Published Sep 17, 2024

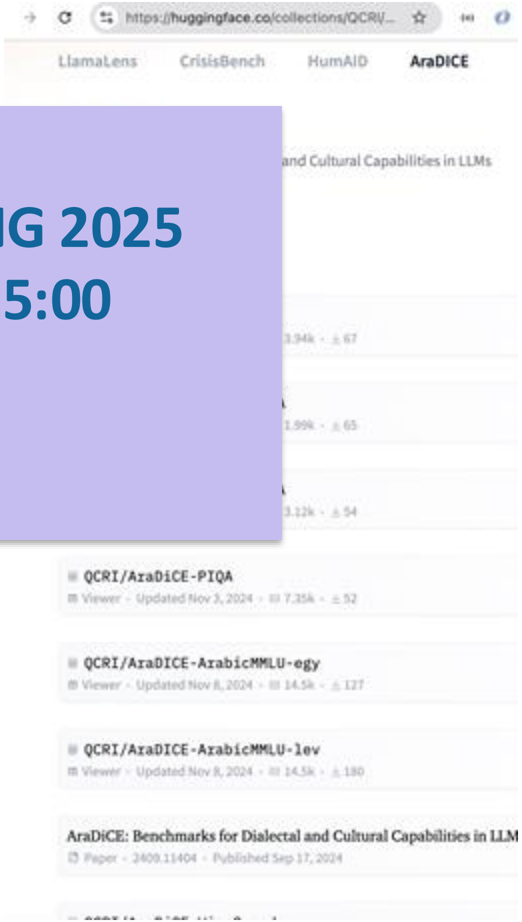
https://huggingface.co/collections/QCRI/aradice-6727765839bf89aa78e9f132
https://arxiv.org/pdf/2409.11404

AraDiCE:
Benchmarks for **Dialectal and Cultural Capabilities** in LLMs

Presentation at COLING 2025

Session 13: 13:30-15:00

24 January, 2025



...and Cultural Capabilities in LLMs

3.94k · 67

1.99k · 65

3.12k · 54

QCRI/AraDiCE-PIQA
Viewer · Updated Nov 3, 2024 · 7.35k · 52

QCRI/AraDICE-ArabicMMLU-egy
Viewer · Updated Nov 8, 2024 · 14.5k · 127

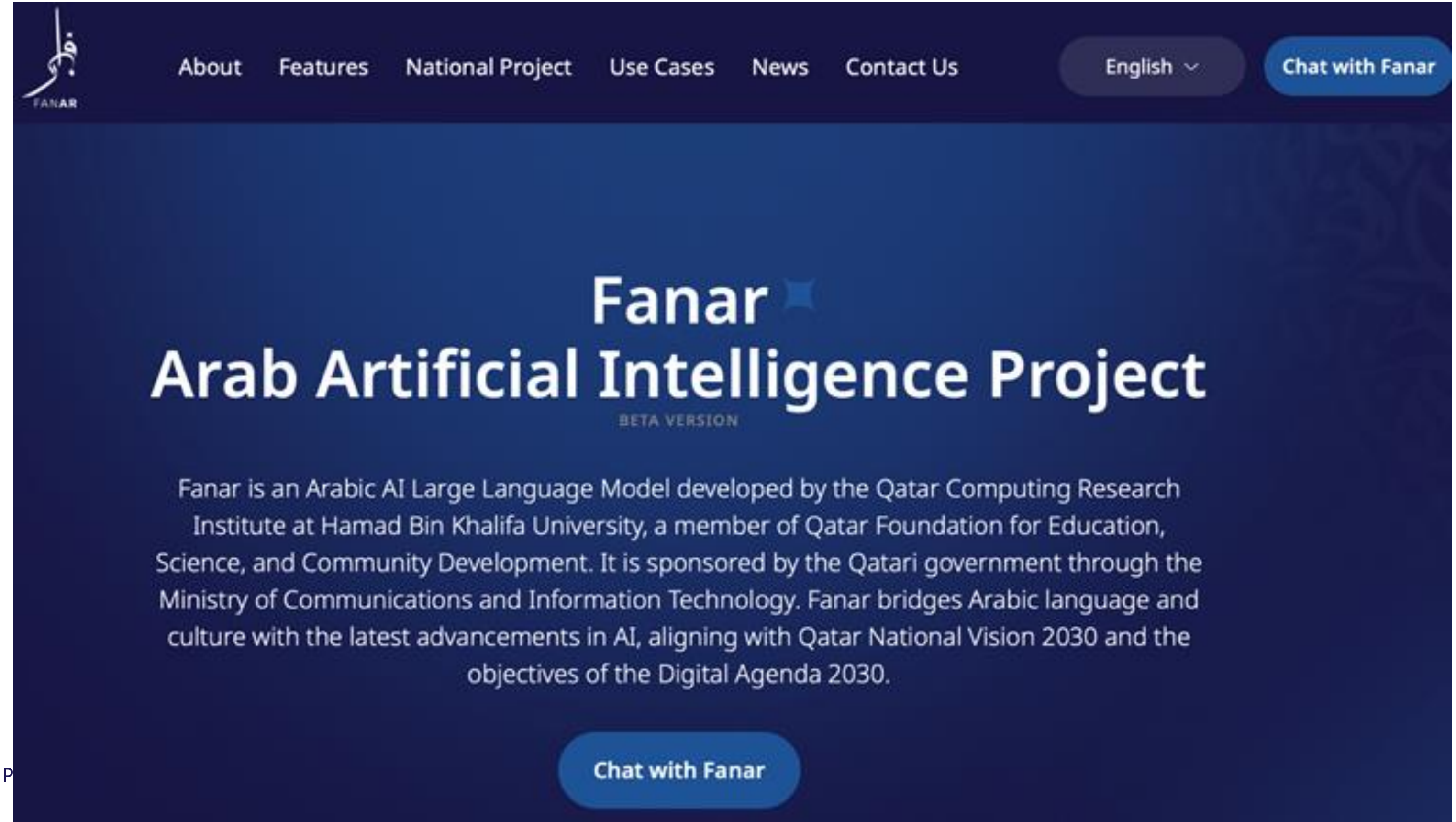
QCRI/AraDICE-ArabicMMLU-lev
Viewer · Updated Nov 8, 2024 · 14.5k · 180

AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLM
Paper · 2409.11404 · Published Sep 17, 2024

https://huggingface.co/collections/QCRI/aradice-6727765839bf89aa78e9f132
https://arxiv.org/pdf/2409.11404

Arabic LLM: Fanar

https://fanar.qa/



Fanar

Arab Artificial Intelligence Project

BETA VERSION

Fanar is an Arabic AI Large Language Model developed by the Qatar Computing Research Institute at Hamad Bin Khalifa University, a member of Qatar Foundation for Education, Science, and Community Development. It is sponsored by the Qatari government through the Ministry of Communications and Information Technology. Fanar bridges Arabic language and culture with the latest advancements in AI, aligning with Qatar National Vision 2030 and the objectives of the Digital Agenda 2030.

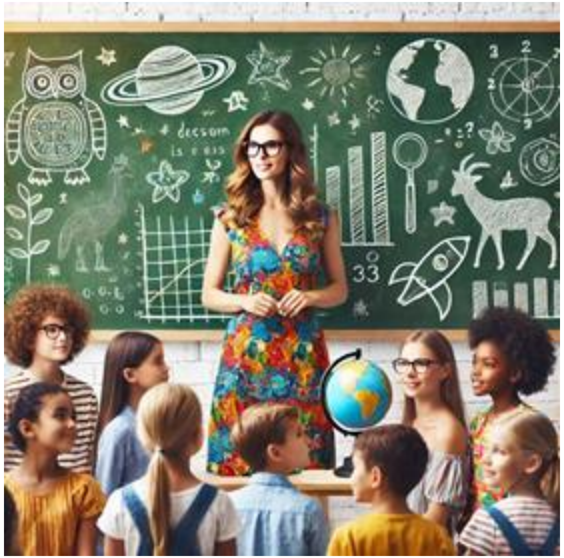
Arabic LLM: Fanar

https://fanar.qa/

Cultural Appropriateness

The LLM should reflect Arab culture for:

- Social practices
- Political issues
- Habits & manners
- Food
- Landmarks
- Names
- etc.



Arabic LLM: Fanar

https://fanar.qa/

Cultural Appropriateness: Story Writing

- The names should be more like **Hassan, Fadi, Marwa, & Fatimah!**
- They should be eating **matchbous** or **umm ali** and not brisket and donuts!
- It is likely that **grandparents would be in the story**
- They would **likely have a pet cat** and **less likely a dog** at home
- Etc.

Open Discussion

Questions:

- How can community-driven contributions be used to develop cultural datasets for underrepresented languages and cultures?
- For languages with limited digital resources, how can multimodal datasets be developed to enhance LLM training?
- How can we ensure LLMs remain inclusive and avoid reinforcing stereotypes in diverse languages?

Summary and Wrap-Up

Slack channel