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

Multilingual and Multimodal Culturally Grounded Spoken and Visual Question Answering

Shammur Absar Chowdhury

Qatar Computing Research Institute



Beyond Language: What Does Cultural Grounding Mean?

- Language proficiency ≠ cultural understanding.
- Culture shapes questions, expressions, meanings, and expectations.
- Trustworthy AI must be grounded in people's actual experiences.

How can we make AI that is representative and reliable for Communities they are build to serve?

Towards Culturally Grounded AI

- **Are we evaluating what matters to people?**
  - Do the benchmark reflects everyday questions, local knowledge and diverse cultural experience?
  - Are different communities adequately represented in this models?

- **What happens when communities interact with these AI?**
  - Does the benchmark translates into reliable measure for user-interaction?
  - How do language, dialects and cultural context affect user experience?



Multimodal Interaction

(Culturally-Grounded Everyday Knowledge)

EveryDayMMQA

Everyday Multimodal Interaction

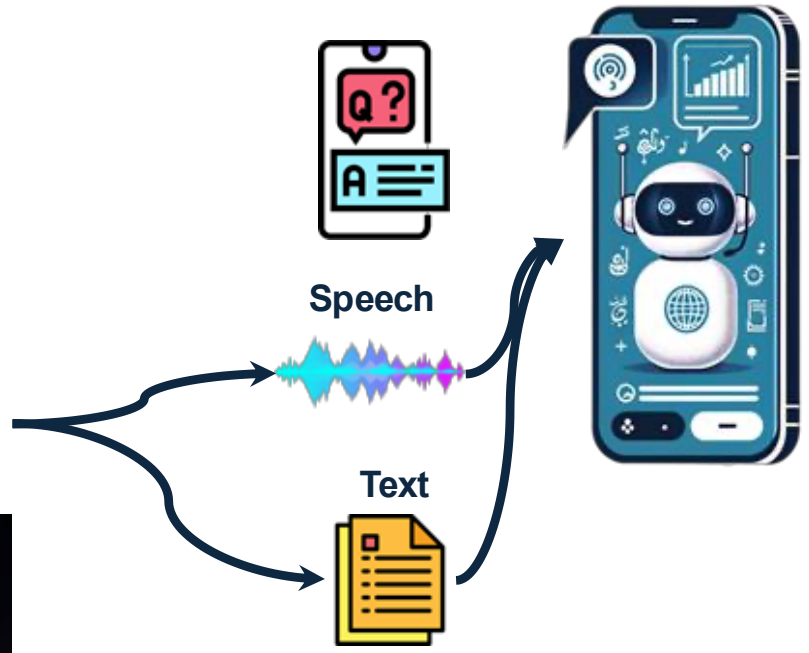
- **Cover real-life topics** like food, fashion, social customs, daily activities.



- Enable model development by incorporating **tri-modality**:
  - **Images** – Capturing real-world and cultural contexts.
  - **Text-Based Questions** – In MSA, dialects, and English.
  - **Spoken Queries** – Testing AI's ability to handle voice interactions.

Everyday Multimodal Interaction

Develop a **specialized AI assistant** focused on **everyday knowledge in the Arab world**.



Everyday knowledge

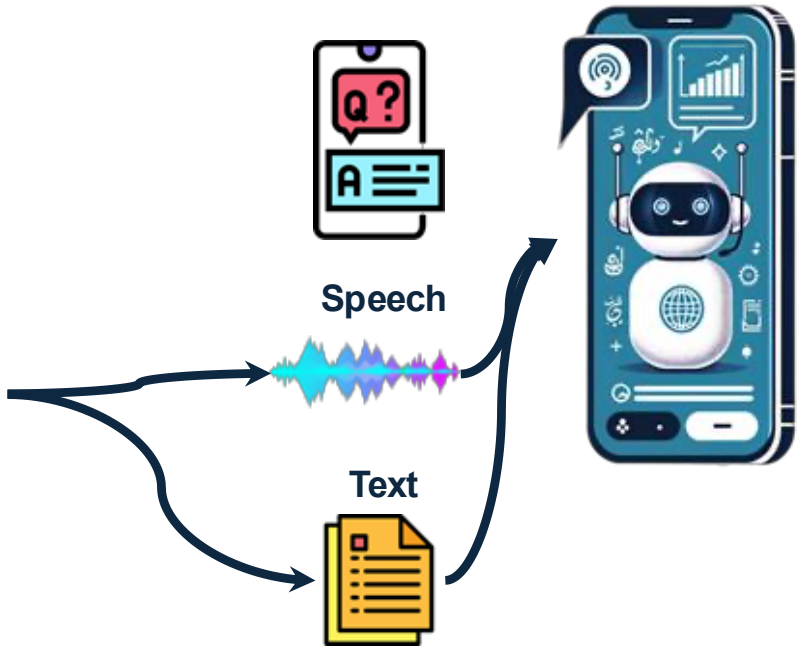
What is the significance of this dance in Arabic culture?



EveryDayMMQA

Everyday Multimodal Interaction

Develop a **specialized AI assistant** focused on **everyday knowledge in the Arab world**.



Everyday knowledge

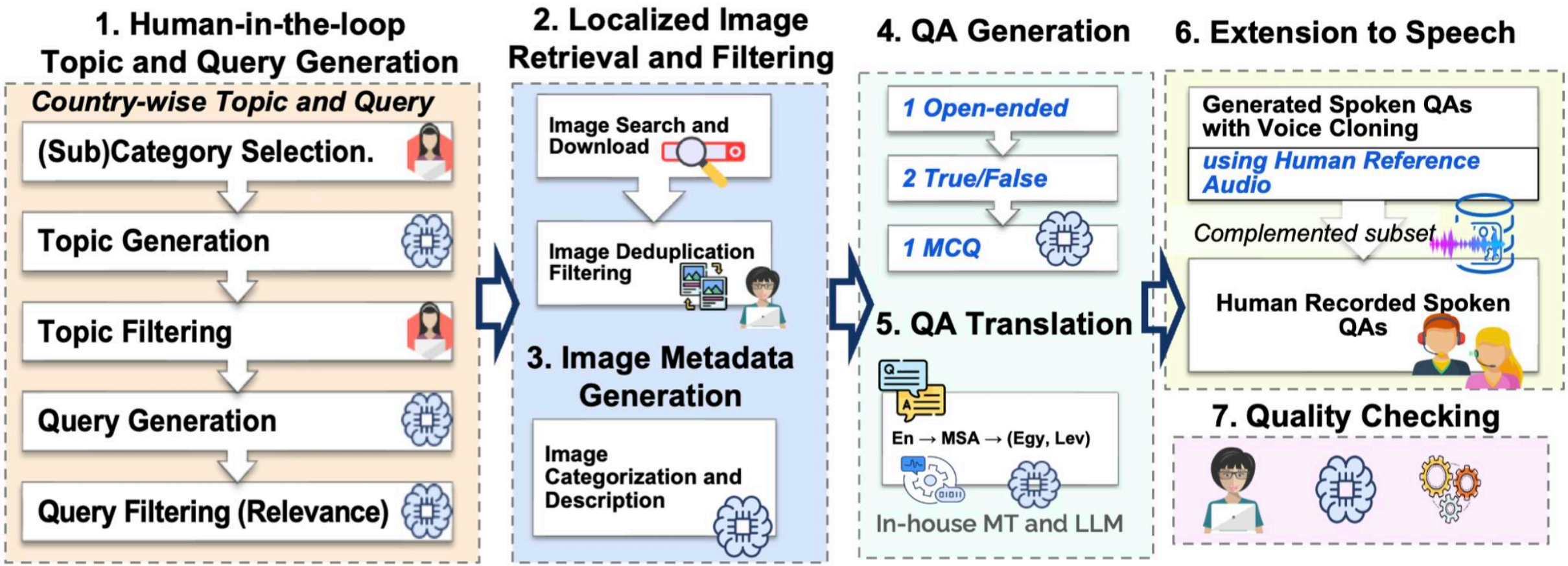
What is the significance of falcon here?



EveryDayMMQA

EverydayMMQA: Framework

EveryDayMMQA



EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA
Firoj Alam, Ali Ezzat Shahroor, Md. Arid Hasan, Zien Sheikh Ali, Hunzalah Hassan Bhatti, Mohamed Bayan Kmainasi, Shammur Absar Chowdhury, Basel Mousi, Fahim Dalvi, Nadir Durrani, Natasa Milic-Frayling

EverydayMMQA: OASIS Dataset

EveryDayMMQA



Metadata

Image Description:
MSA": "التصميم مستوحى من تشكيلات …",
"Eng": "A modern architectural building featuring …",
"ARZ" : ….,
"AJP": ….

Category: *"Geography, Buildings & Landmarks"*
Subcategory: *"Famous Landmarks"*
Topic: *"Futuristic Design of the National Museum of Qatar"*
Cognitive focus: *"knowledge-based"*
Semantic focus: *["Architectural Features and Functions, "Cultural Significance and Heritage"]*
Country: *Qatar*

Open-ended:
Q (MSA): ما اسم المبنى الموضح في الصورة، وما الذي ألهم تصميمه؟
(What is the name of the building shown in the image, and what inspired its design?)

Ans (MSA): المبنى هو متحف قطر الوطني، وتصميمه مستوحى من تشكيلات الورود الصحراوية
(The building is the National Museum of Qatar, and its design is inspired by desert rose formations.)
Rationale (MSA): تُظهر الصورة متحف قطر الوطني الشهير، المعروف بهياكله على شكل قرص تشبه تشكيلات وردة الصحراء، وهي ظاهرة جيولوجية فريدة من نوعها
("The image depicts the iconic National Museum of Qatar, known for its disk-shaped structures resembling desert rose formations, a unique geological phenomenon.")

Multiple Choice:
Q (MSA): ما هو المعلم الظاهر في الصورة وأين يقع؟
("Which landmark is shown in the image, and where is it located?")
Options:
A. "متحف اللوفر أبوظبي، الإمارات العربية المتحدة"
("The Louvre Abu Dhabi, UAE")
B. "متحف قطر الوطني، قطر"
("National Museum of Qatar, Qatar")
C. "دار أوبرا سيدني"
("Sydney Opera House, Australia")

Ans (MSA): "متحف قطر الوطني، قطر"
("National Museum of Qatar, Qatar")
Rationale (MSA): تصميم المبنى المميز على شكل قرص متشابك هو سمة من سمات المتحف الوطني في قطر، وتقع في الدوحة، قطر
("The building's distinctive interlocking disk-shaped design is characteristic of the National Museum of Qatar, located in Doha.")

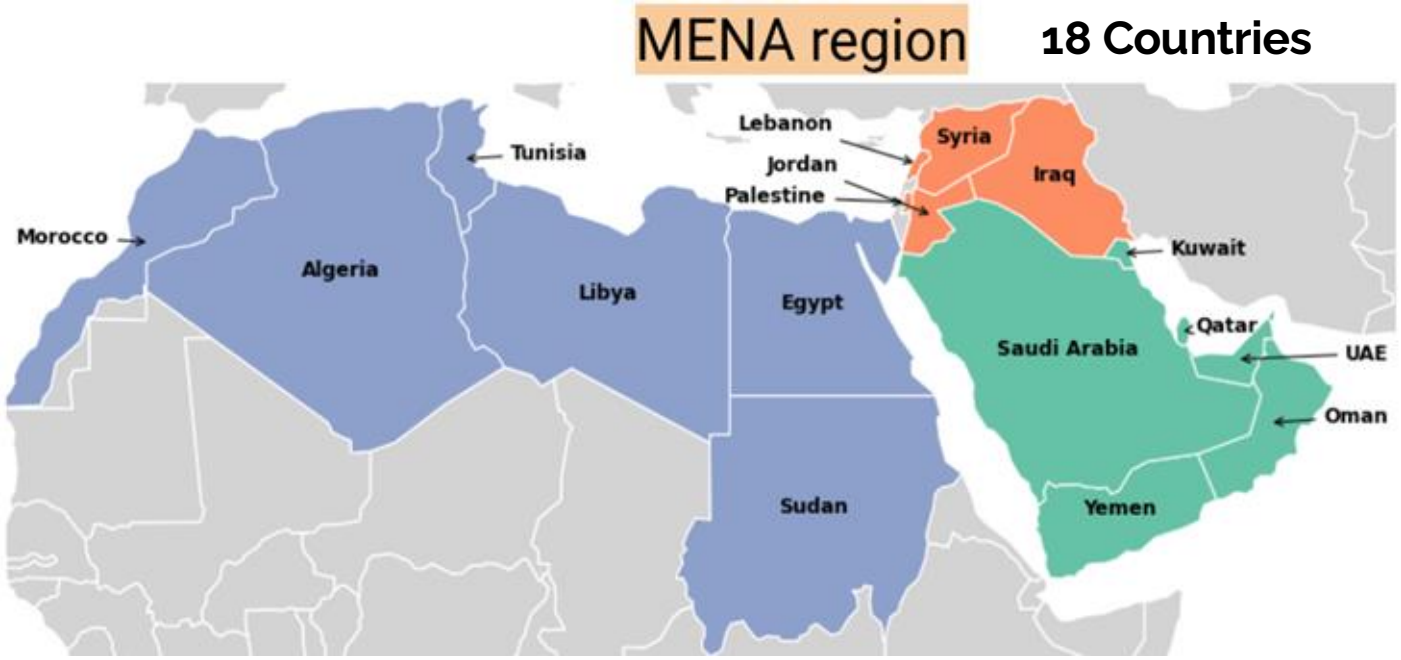
True/False:
Q (MSA): يقع المبنى الذي يظهر في الصورة في دولة الإمارات العربية المتحدة
("The building shown in the image is located in the United Arab Emirates.")
Options:
A. True
B. False

Ans (MSA): False
Rationale (MSA): المبنى يقع في قطر وليس في الإمارات العربية المتحدة. هو متحف قطر الوطني
("The building is located in Qatar, not the United Arab Emirates. It is the National Museum of Qatar.")

EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA
Firoj Alam, Ali Ezzat Shahroor, Md. Arid Hasan, Zien Sheikh Ali, Hunzalah Hassan Bhatti, Mohamed Bayan Kmainasi, Shammur Absar Chowdhury, Basel Mousi, Fahim Dalvi, Nadir Durrani, Natasa Milic-Frayling

EverydayMMQA: OASIS

EveryDayMMQA

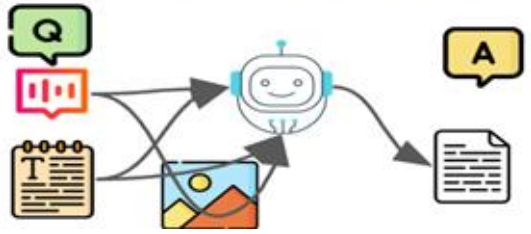


18 Countries

Lang/Dialects

Eng , MSA, Egy, Lev

Modality Setups



9 Categories and 31 subcategories

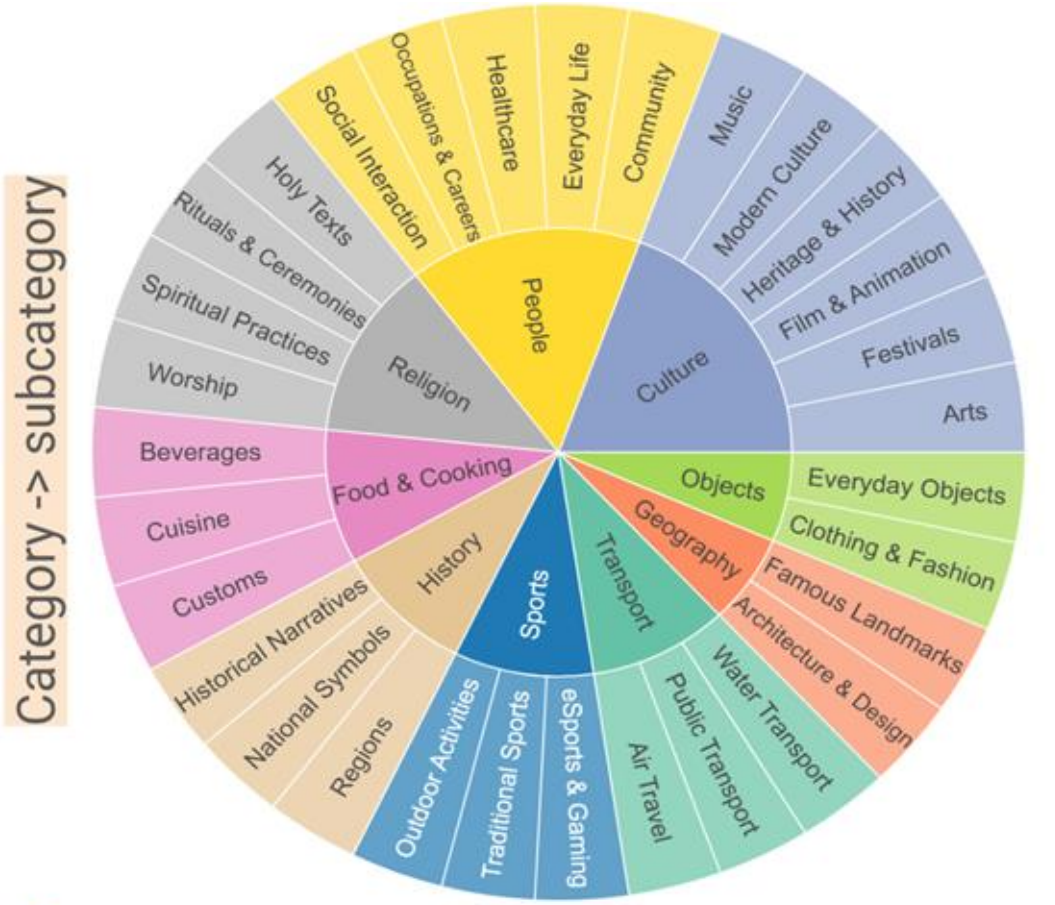


Culture, Arts & Entertainment **People**, Society & Education **Religion** & Spirituality **History**, Geography & National Identity **Food & Cooking** **Objects**, Materials & Clothing **Sports** & Recreation **Geography**, Buildings & Landmarks Vehicles & **Transport**ation

EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA

EverydayMMQA: OASIS

EveryDayMMQA



Culture, Arts & Entertainment **People**, Society & Education **Religion** & Spirituality **History**, Geography & National Identity **Food & Cooking** **Objects**, Materials & Clothing **Sports** & Recreation **Geography**, Buildings & Landmarks Vehicles & **Transport**ation

QA: ~3.7Mx4=~14.8M

- Open-ended: ~0.92M
- MCQ: ~0.92M
- True/False: ~1.85M

Audio Stats

Generated audio (20,279 h)
- en: ~14M audio (14,509 h), 35 ref speakers
- msa: ~3.7M audio (5,770 h), 47 ref speakers
- avg dur: 5s (en), 6s (msa)
- avg audio token: 52

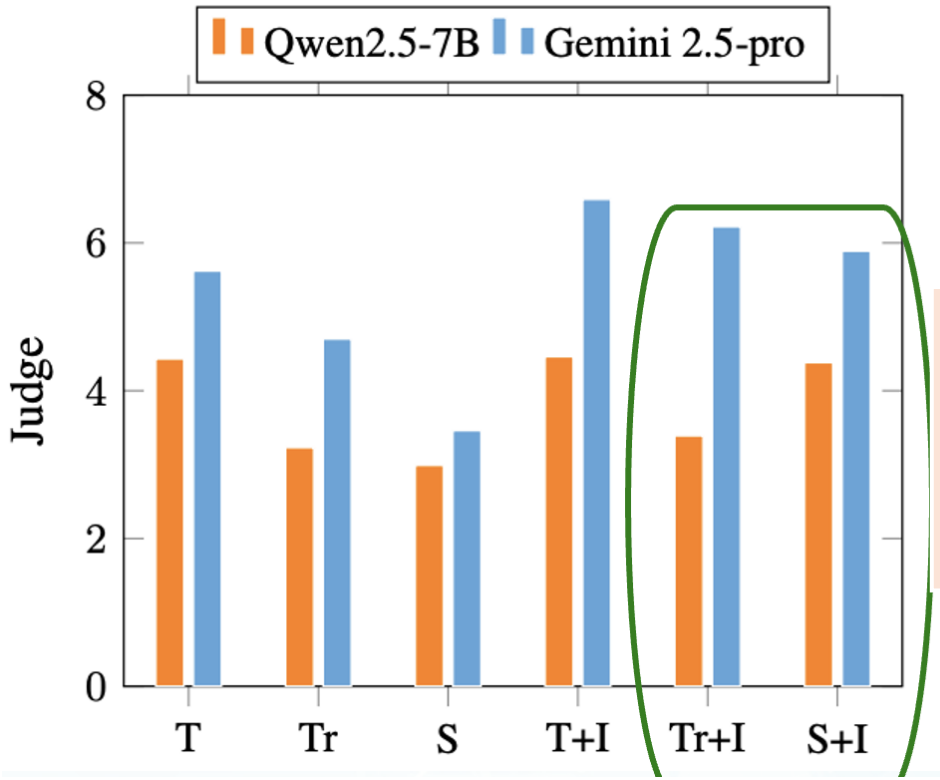
Recordings (141h)
- en: ~33K recordings (55 h); 13 speakers, 4 countries
- msa: ~61K recordings (86 h); 13 speakers, 6 countries

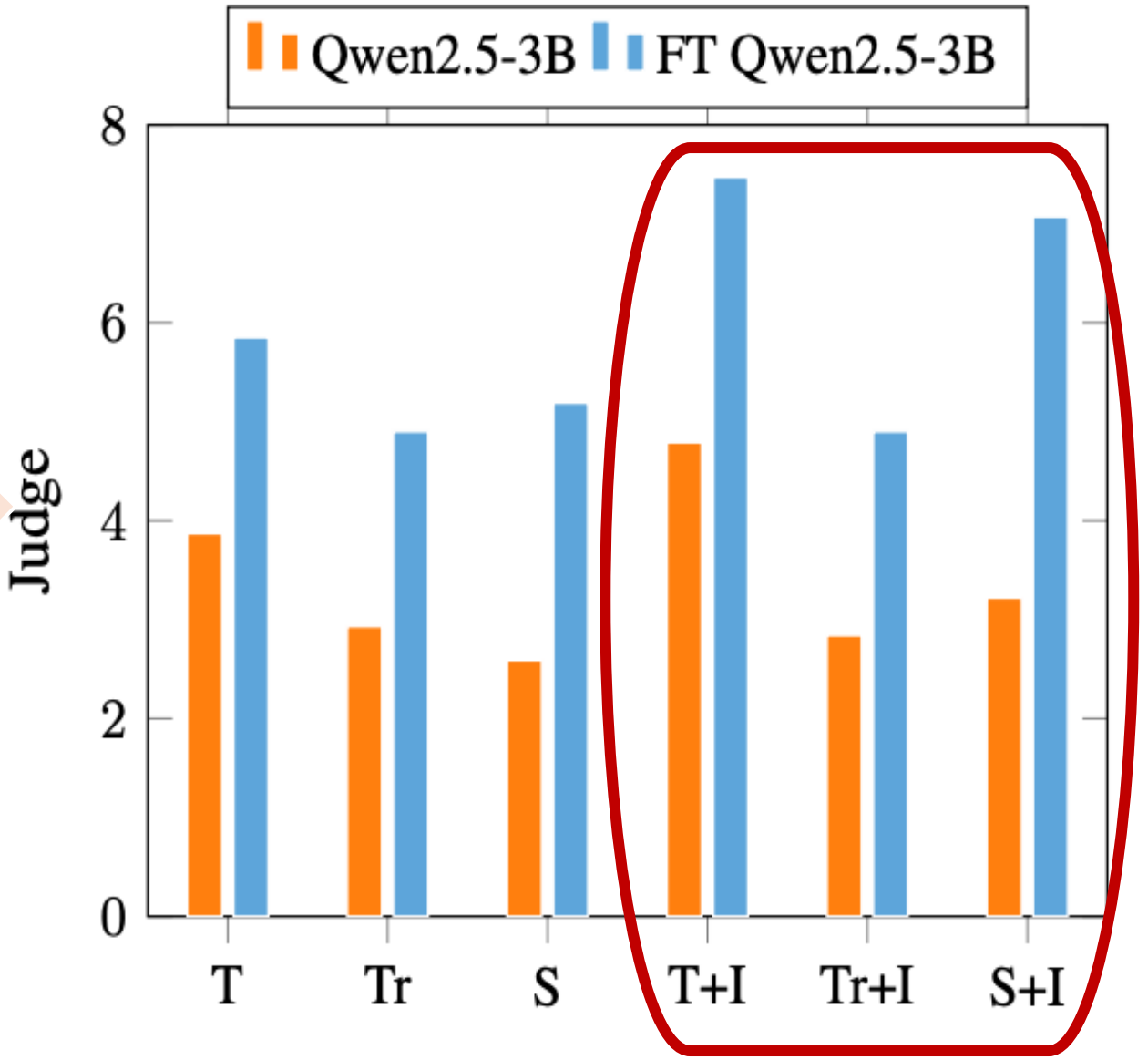
Image Stats

Total image: 0.92M
avg W x H: 1000 x 1226
avg image token: 1232

EverydayMMQA: A Multilingual and Multimodal Framework for Culturally Grounded Spoken Visual QA
Firoj Alam, Ali Ezzat Shahroor, Md. Arid Hasan, Zien Sheikh Ali, Hunzalah Hassan Bhatti, Mohamed Bayan Kmainasi, Shammur Absar Chowdhury, Basel Mousi, Fahim Dalvi, Nadir Durrani, Natasa Milic-Frayling

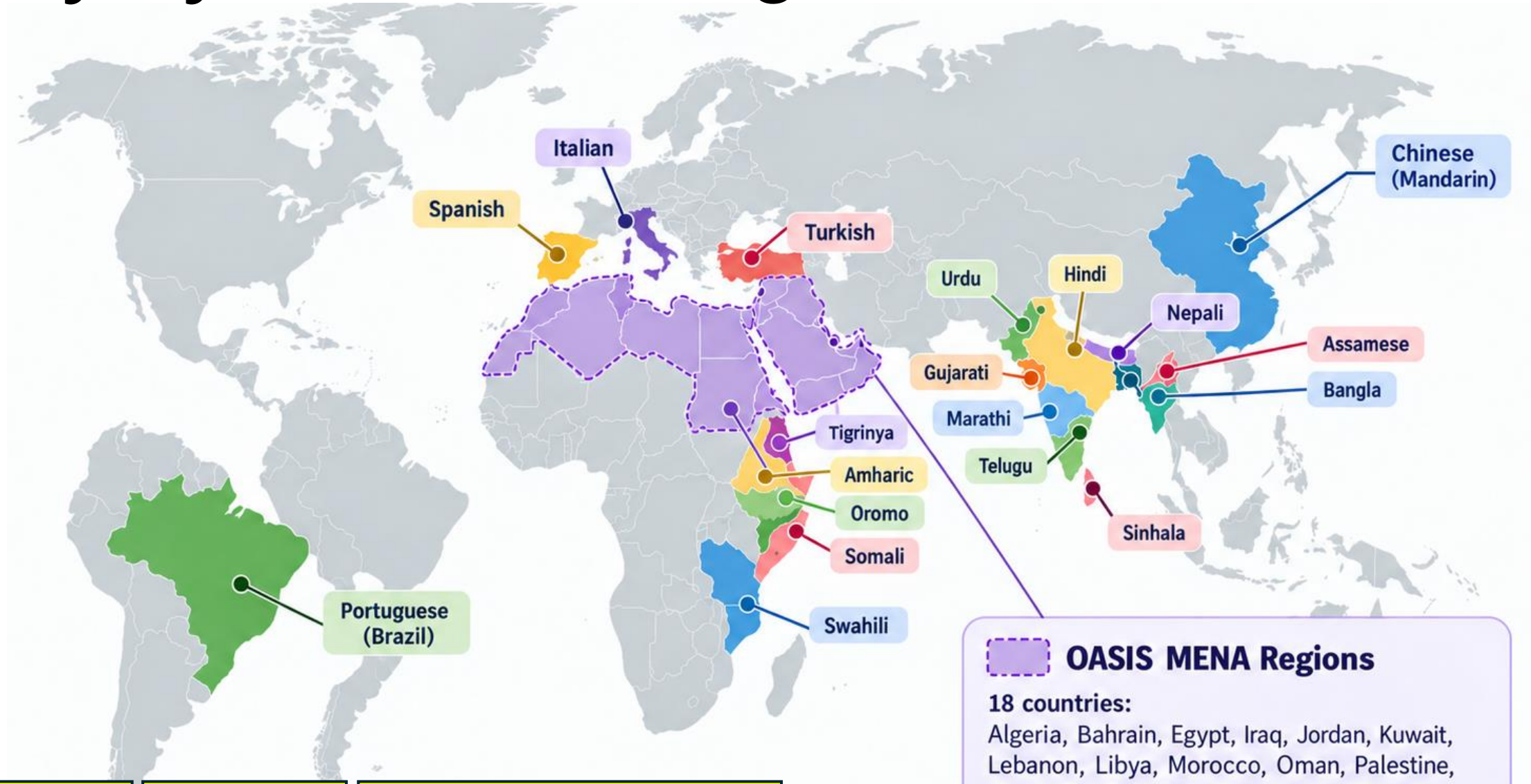
EverydayMMQA: OASIS Results





EverydayMMQA: Extending OASIS

EveryDayMMQA



33 Countries	21 Languages	2 Arabic Dialectal Regions

MMCultureQA | SEMEVAL 2027 | Shared Tasks



KEY DATES

Jul 15, 2026	Sep 23, 2026	Jan 10, 2027	Jan 31, 2027	Feb 2027	Summer 2027
Sample data	Training data	Eval opens	Eval closes	Papers due	Workshop

Towards Culturally Grounded AI

- **Are we evaluating what matters to people?**
  - Do the benchmark reflects everyday questions, local knowledge and diverse cultural experience?
  - Are different communities adequately represented in this models?

- **What happens when communities interact with these AI?**
  - Does the benchmark translates into reliable measure for user-interaction?
  - How do language, dialects and cultural context affect user experience?



In-the-Wild Arabic Spoken Interactions (with LLMs)

IN-THE-WILD

WASIL

IN-THE-WILD

WASIL

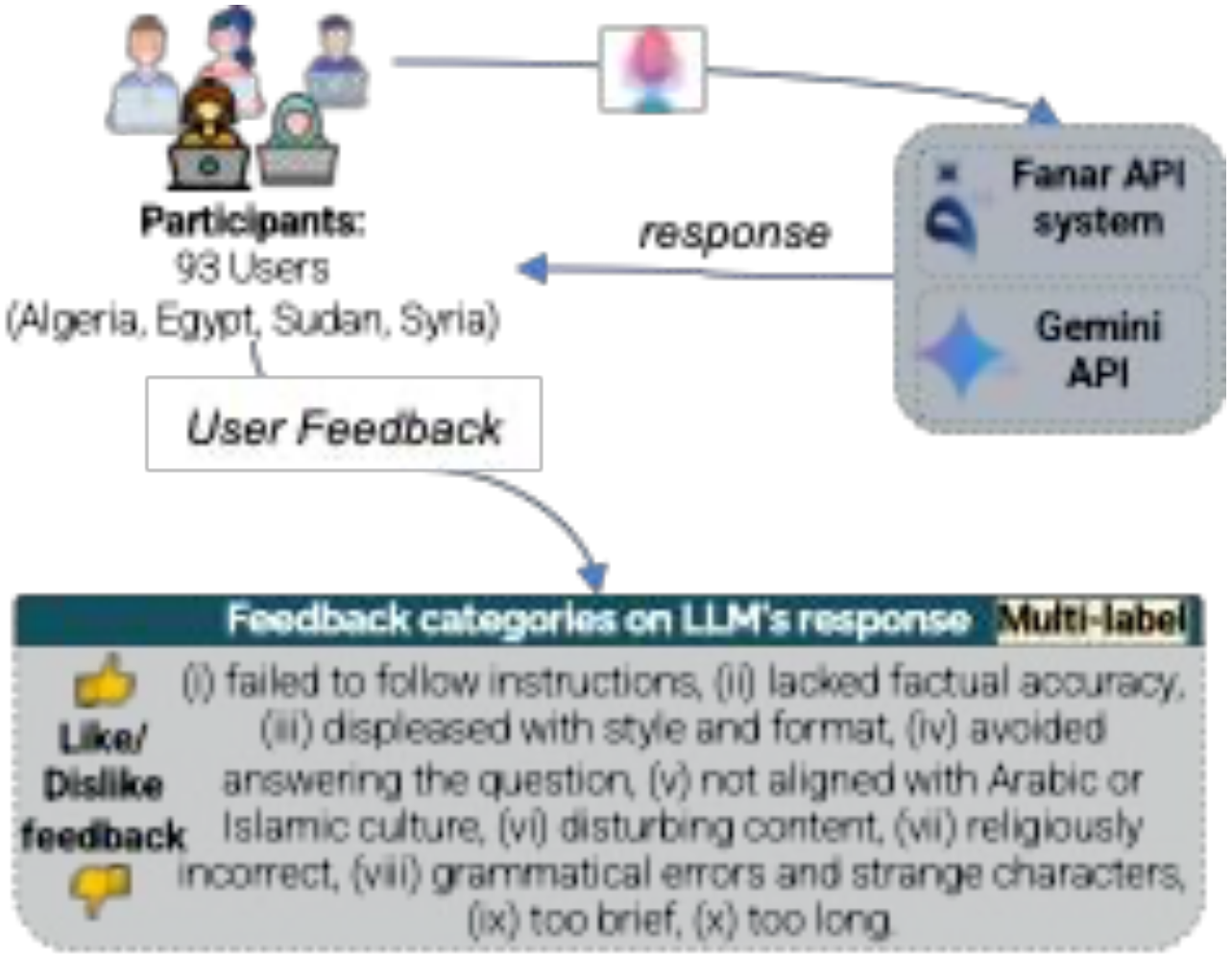
Motivation

The First in-the-wild Arabic spoken interaction log datasets with explicit fine-grained user feedbacks along with multi-layer of annotation: Dialect identification, transcription, MSA translation among others.

Can be used for:

- **Diagnostic Test set**
- **Preference understanding and modeling**
- **Benchmark general purpose text- and multimodal LLMs**

WASIL Step1: Spoken LLM Interaction Collection



9 days collection:
open discussion (any topic), follow-up questions, creative writing, scientific questions, safety-related queries, and Arabic culture.

Step2: Gold Transcript Creation

Transcribing dialectal Arabic from scratch is slow and expensive.

Post-editing

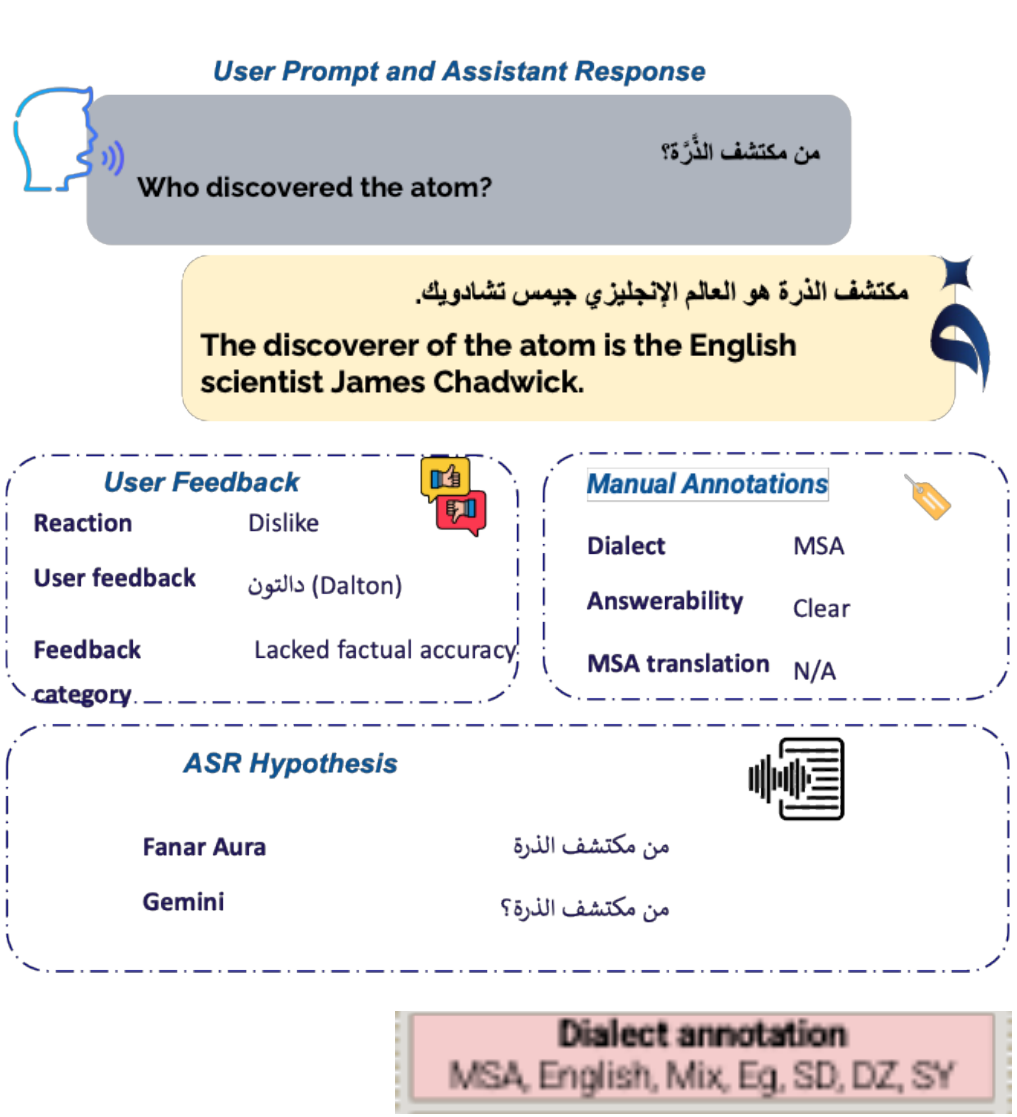


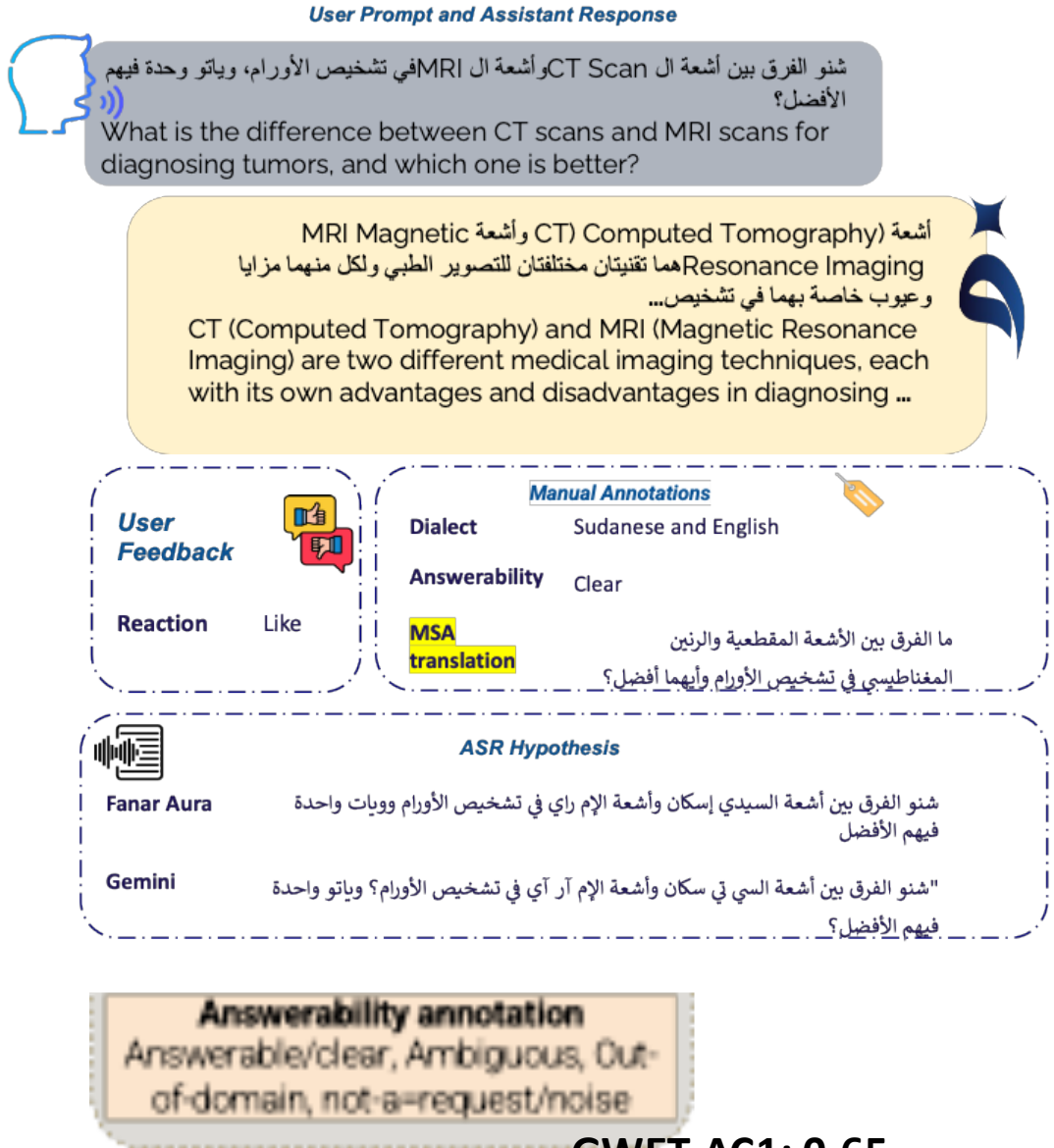
Dialect vs MSA
Scratch vs post-editing

Dialect-wise Differences (ASR transcript Vs. Post-edited transcript)

- **Easiest:** Egyptian, Syrian
- **Hard:** MSA
- **Hardest:** Mix, English

Step3: Dialect, Answerability and MSA Translation





Dialect annotation
MSA, English, Mix, Eg, SD, DZ, SY

Answerability annotation
Answerable/clear, Ambiguous, Out-of-domain, not-a-request/noise

GWET AC1: 0.65

IN-THE-WILD

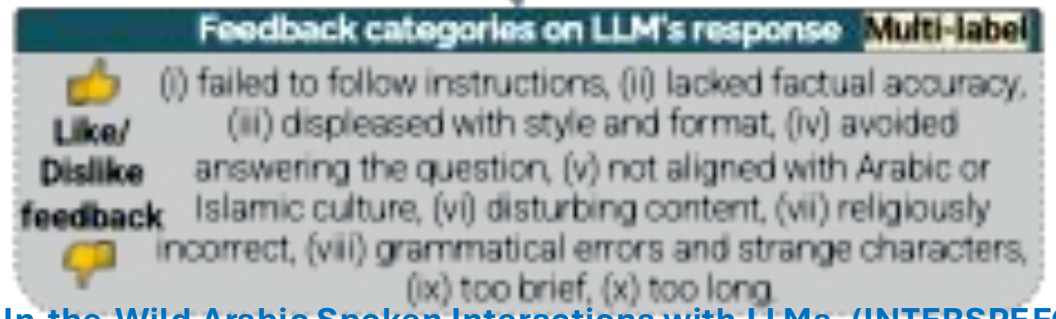
WASIL

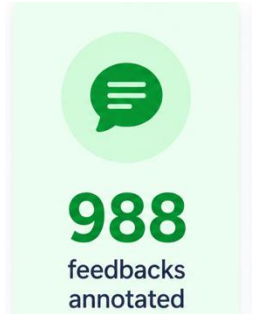
WASIL at a Glance

The first dataset of in-the-wild Arabic Spoken LLM Interaction

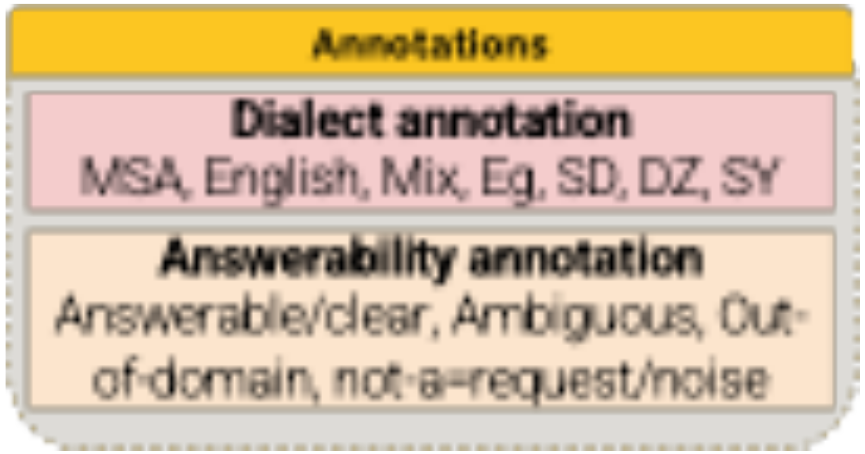


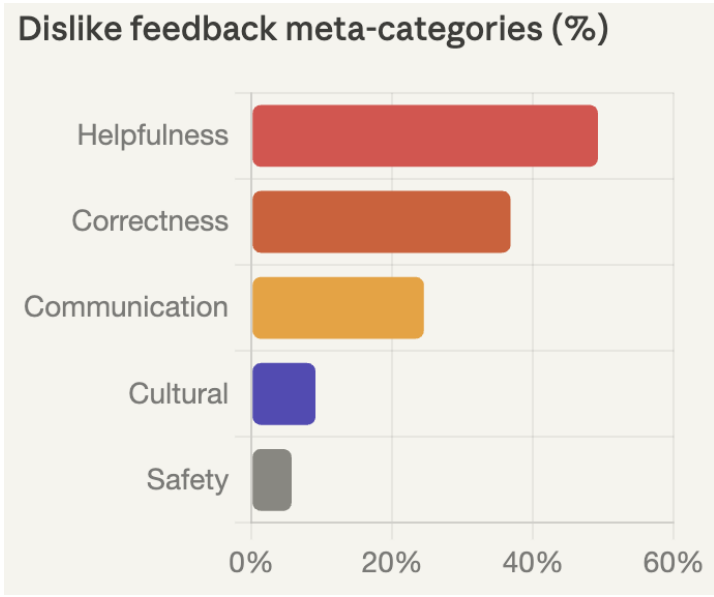












WASIL: In-the-Wild Arabic Spoken Interactions with LLMs. (INTERSPEECH 2026 Long)

IN-THE-WILD

WASIL

What Makes WASIL Different

Existing resources*	WASIL
Text chat logs	**Spoken conversations** Speech, ASR hypothesis and Gold Transcription
Preference judgments	**Feedback that explains why** Like/dislike with reasons, reference transcripts, and labels for whether a question is answerable
English-centric resources	**Arabic language** Modern Standard Arabic (MSA), 4 dialects, and Code-switching. MSA <-> DA translation
Task-driven evaluation	**Real-world Speech!** In-the-wild conversations with real users

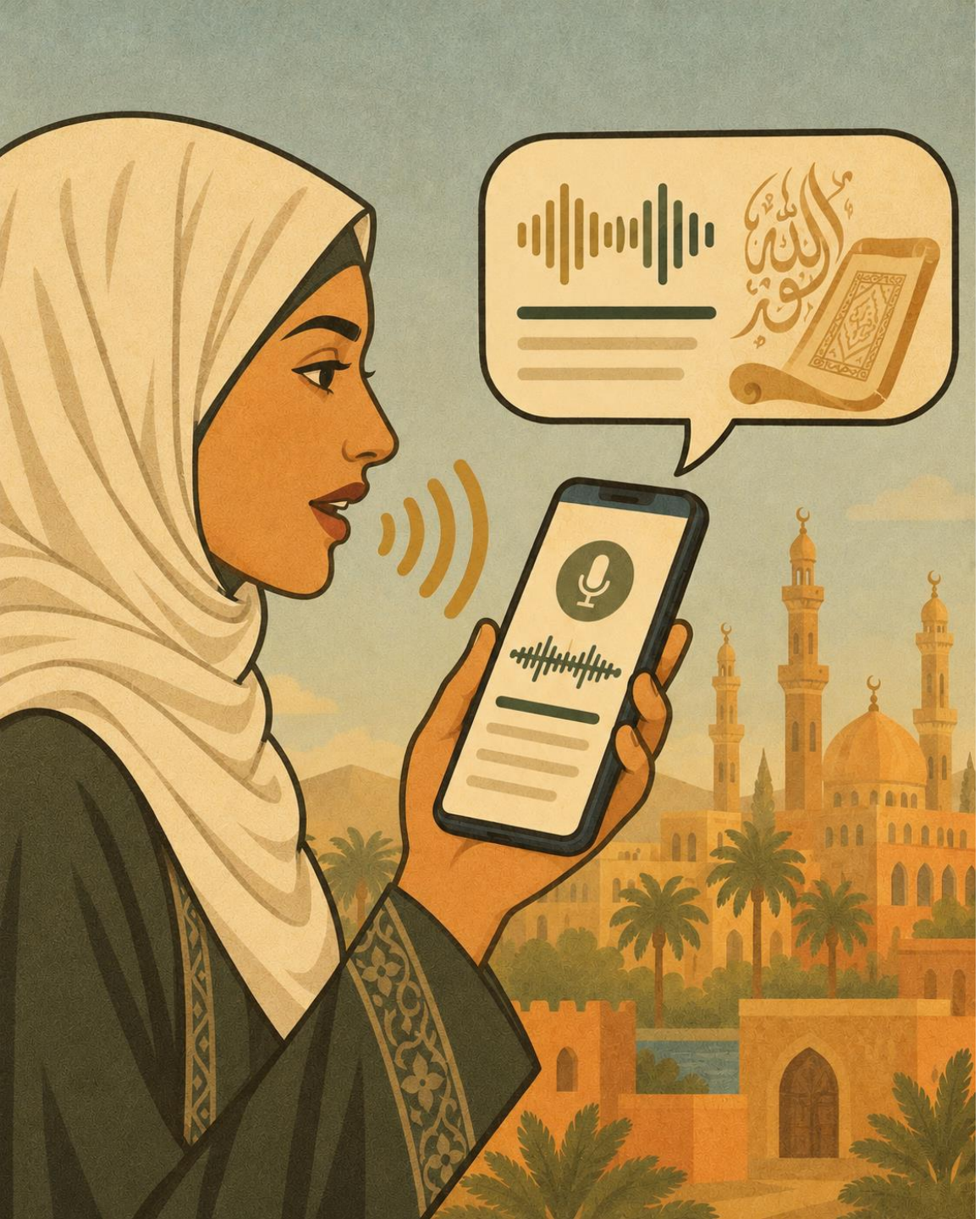


IN-THE-WILD

WASIL



https://huggingface.co/datasets/QCRI/WASIL



Faithful Recognition using Speech AI

Fanar Aura Arabic-English Speech Recognition Systems

Fanar-Aura-STT-1 (deprecated)

Fanar-Aura-STT-LF-1 (deprecated)

Fanar-3.0-Aura-Transcribe

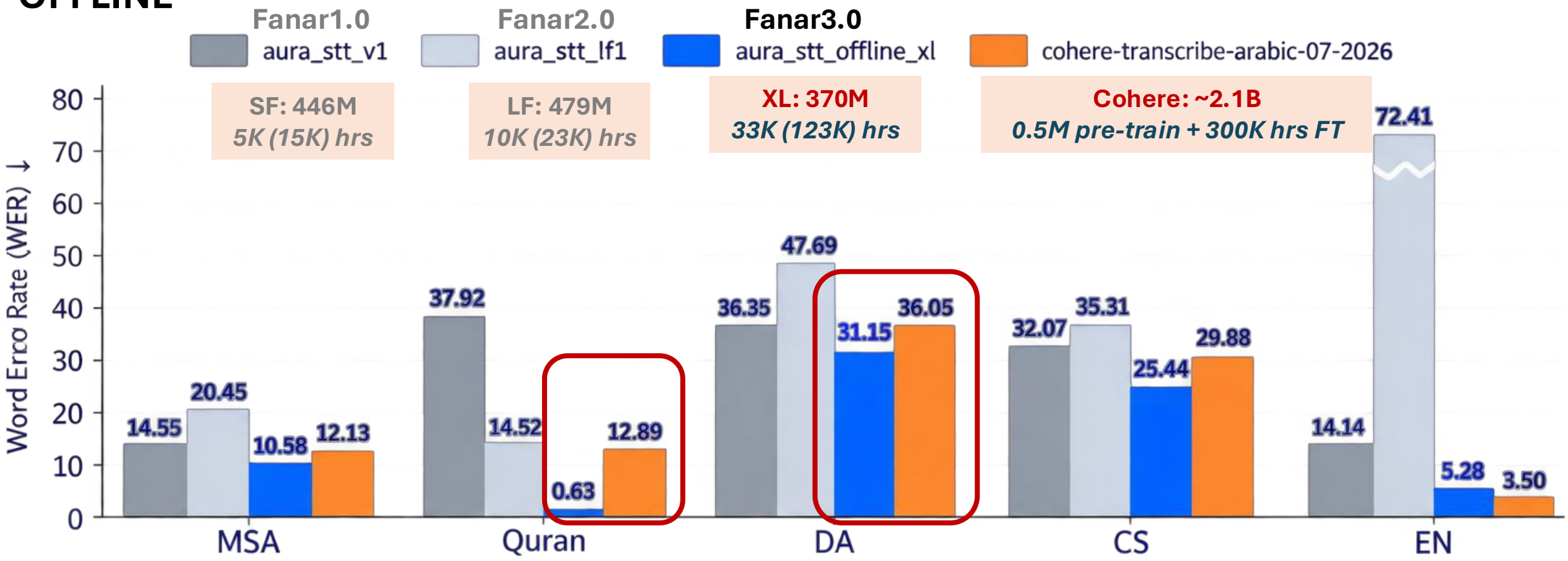
Fanar-3.0-Aura
Long-form ASR for
MSA, Dialect and Classical Arabic

Fanar-3.0-Aura
Streaming ASR for
MSA, Dialect and Classical Arabic

API: https://api.fanar.qa
Chat Platform: https://chat.fanar.qa/

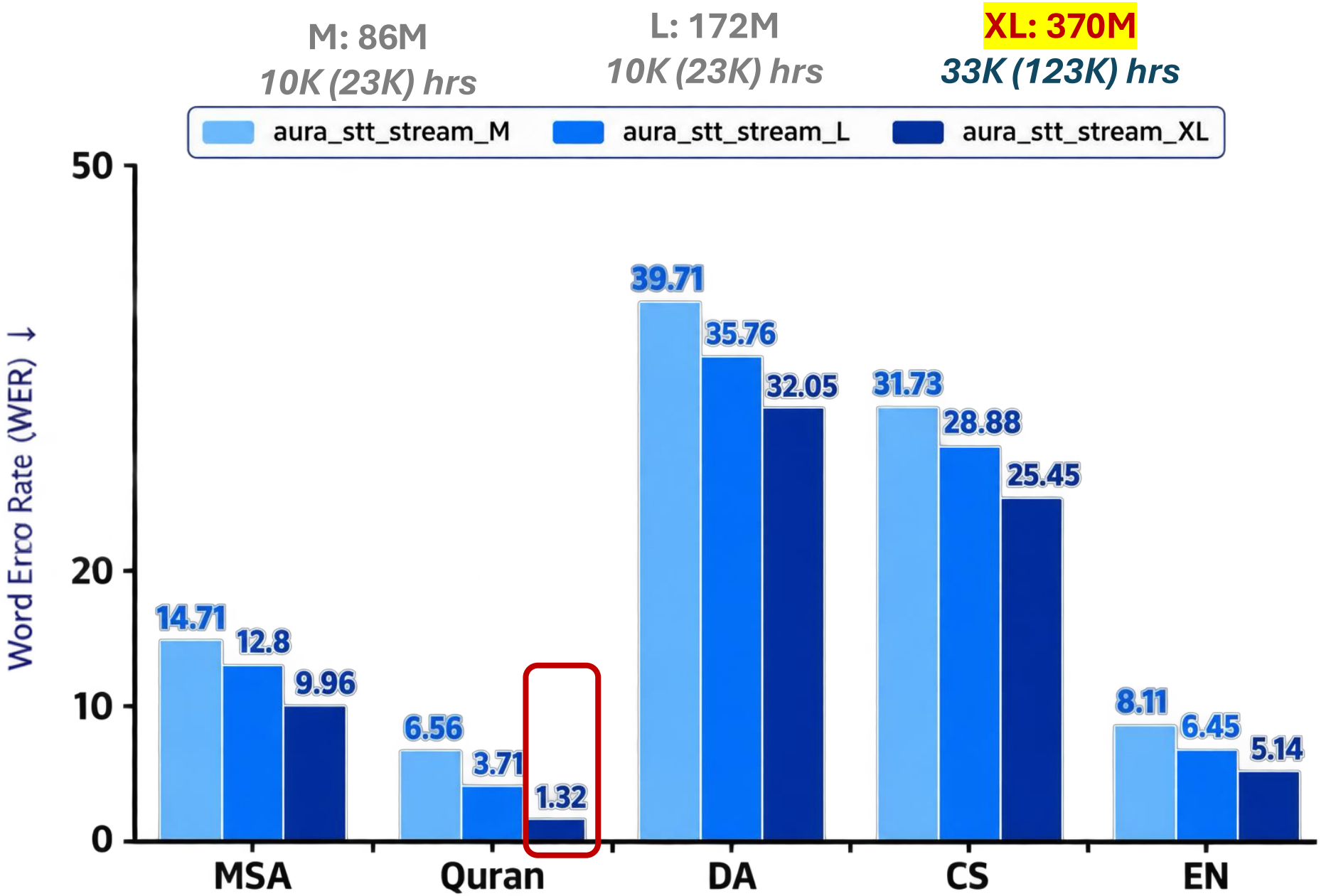
https://huggingface.co/spaces/QCRI/AURA-Arabic-English-ASR-leaderboard

OFFLINE



Set	Offline XL vs v1	Offline XL vs lf1	vs Cohere
MSA	↓ 27.3%	↓ 48.3%	1.55 better
Quran	↓ 98.3%	↓ 95.7%	12.26 better
DA	↓ 14.3%	↓ 34.7%	4.90 better
CS	↓ 20.7%	↓ 28.0%	4.44 better
EN	↓ 62.7%	↓ 92.7%	1.78 worse

Streaming AR-EN ASR 2026



ASR Data in the Aura-STT-Stream-XL model (~370M)

Category	Value
Total cuts	**53,634,239**
Raw speech hours	**33,434.27 h**
Estimated after augmentation	**~123,657 h**
Approx augmentation factor	**~3.7x vs updated raw**

Tag	Overall hours	Share of overall
<en>	**10,012.54**	**29.9%**
<ar> MSA	**3,784.22**	**11.3%**
<da>	**6,808.05**	**20.4%**
<qa>	**2,002.00**	**6.0%**
<mix>	**7,718.49**	**23.1%**
No tag	**3,108.97**	**9.3%**
Total	**33,434.27**	**100.0%**

	Unique tokens	Total Token
Total	1,958,972	794,150,581
Arabic	1,302,414	338,610,178
English	648,155	454,358,563
Numbers	4,000	932,217
Tags	6	55,909,745
Mixed	4,403	100,809

Aura-Stream-xl-1

