

Motivation

Why this matters?

- Western-centric image distribution
- Equivalent text and speech should yield the same decision
- Non-English speech remains under-evaluated

Benchmark at a glance

- 10,150 culturally grounded images from 18 MENA countries
- One supported and two plausible unsupported statements
- English and Arabic, as text and speech

Research Questions

- Does Contrastive Instability CI reveal failure beyond Q+/Q- accuracy?
- How does stability change from text to speech?
- Is the effect larger in Arabic than English?
- Do scale and acoustic robustness help?
- Can joint speech-text input restore stability?

M2CQA: Multilingual Multimodal Counterfactual Question Answering

Contrastive statements over MENA region

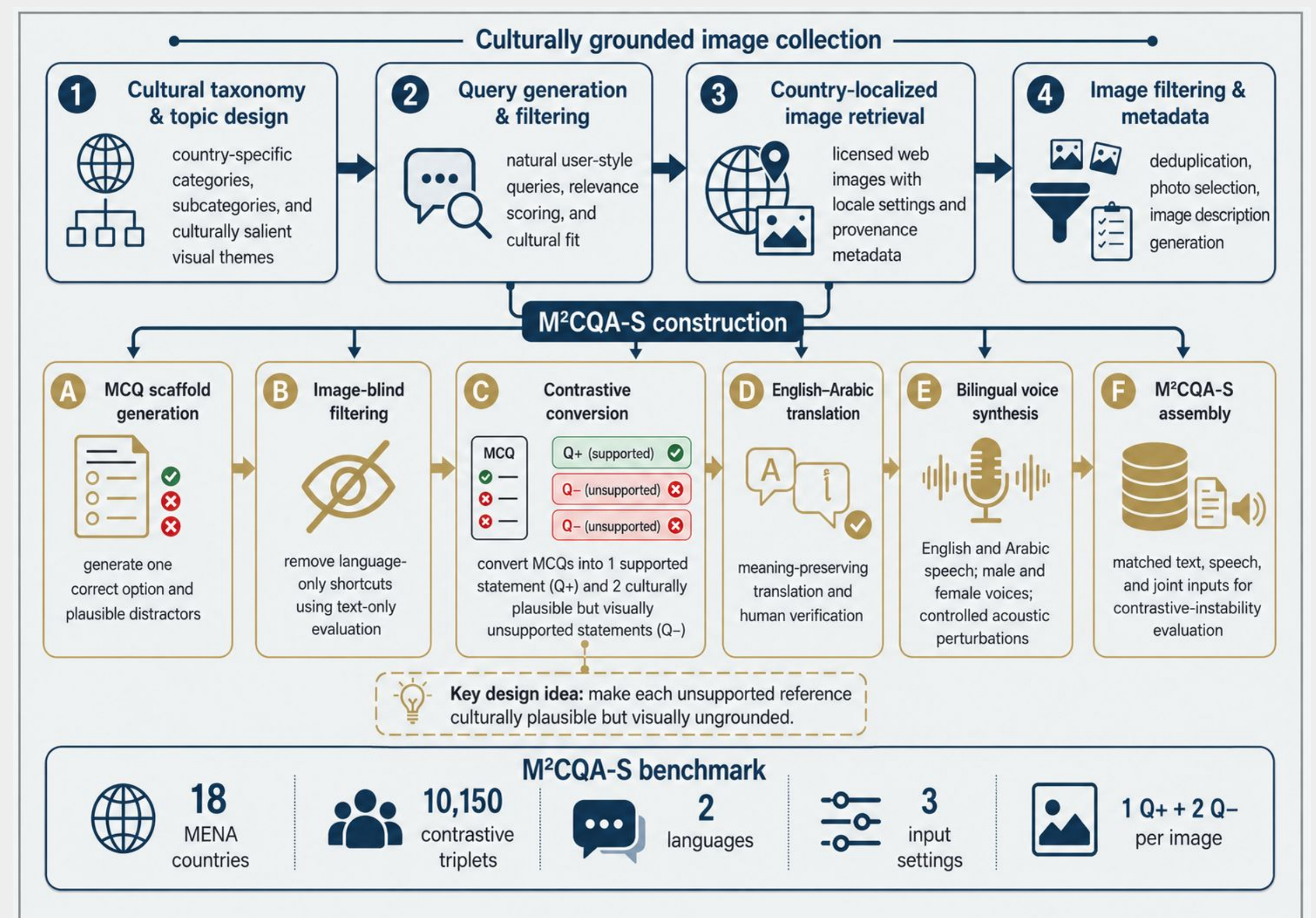
The images show:

Q+: national animal of Qatar
Q-: national animal of Morocco
Q-: national animal of Algeria

Q+: traditional musical performance
Q-: religious ceremony
Q-: preparing for a sports event

Contrastive Instability (CI) $CI = 1 - N(\text{consistent}) / N(\text{partial})$

- Full consistency: accepts Q+ and rejects both Q-.
- Condition on triplets with at least one correct answer.
- Isolates fragmented reasoning from complete failure



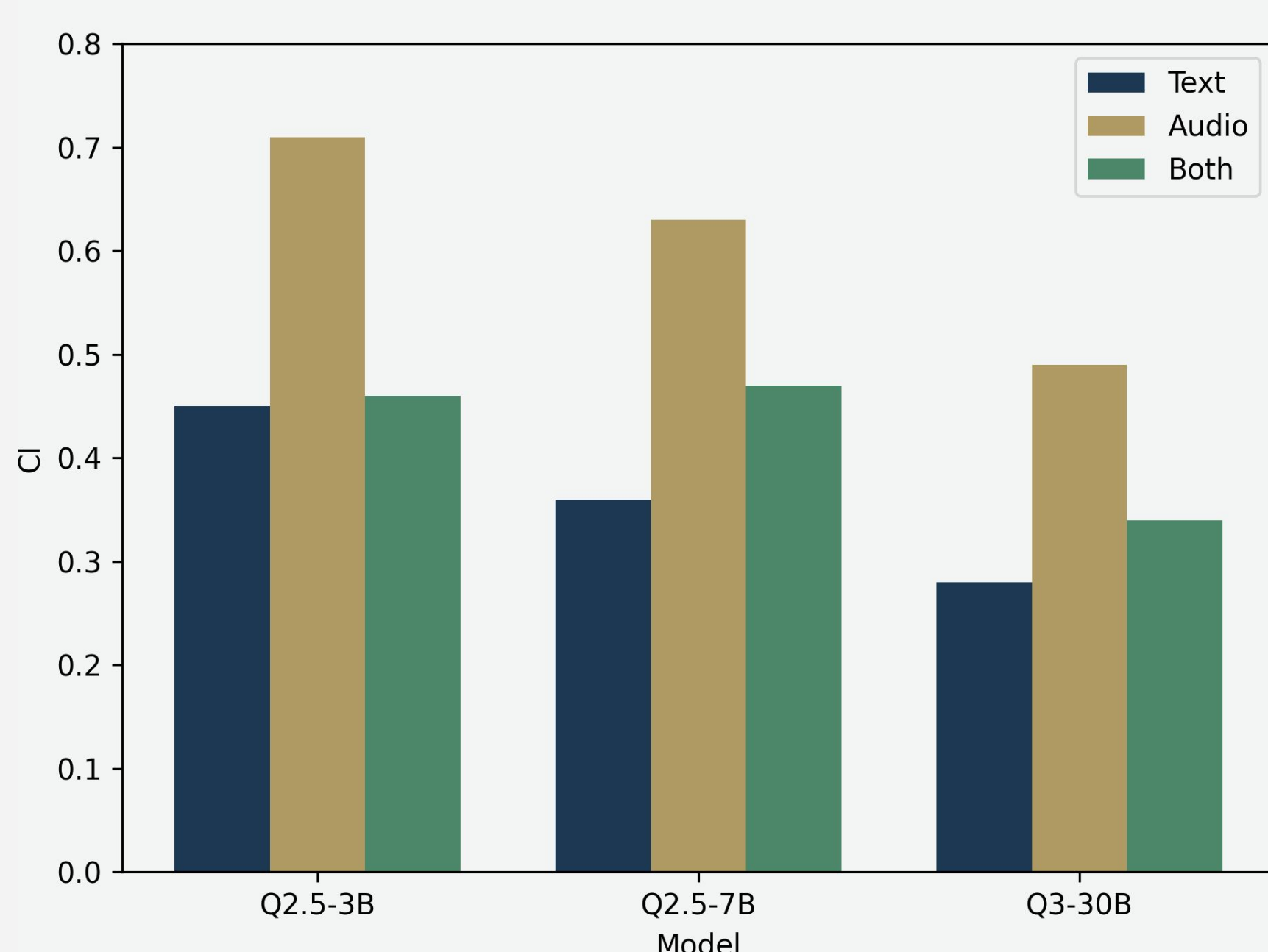
Findings

Model	Lg	Mode	Q+ ↑	Q- ↑	F1 ↑	CI ↓
Q2.5-3B	en	Text	0.93	0.91	0.92	0.20
	en	Speech	0.93	0.90	0.91	0.22
	ar	Text	0.66	0.92	0.77	0.43
	ar	Speech	0.35	0.94	0.51	0.71
Q2.5-7B	en	Text	0.94	0.92	0.93	0.17
	en	Speech	0.95	0.92	0.93	0.18
	ar	Text	0.90	0.82	0.86	0.35
	ar	Speech	0.88	0.62	0.73	0.63
Q3-30B	en	Text	0.93	0.91	0.92	0.20
	en	Speech	0.92	0.92	0.92	0.19
	ar	Text	0.92	0.86	0.89	0.28
	ar	Speech	0.92	0.73	0.81	0.47
Phi-4M	en	Text	0.57	0.94	0.71	0.50
	en	Speech	0.61	0.83	0.70	0.58
	ar	Text	0.45	0.77	0.57	0.75
	ar	Speech	0.04	0.96	0.08	0.97

Q denotes Qwen models. Higher Q+, Q-, and F1 are better; lower CI is better.

Joint Input Reduces Arabic CI

- Audio-only input is consistently worse
- Joint speech-text input substantially reduces CI
- The largest model comes closest to text-only stability
- Multi-modal redundancy helps, but does not close the gap



Cross-model Analysis

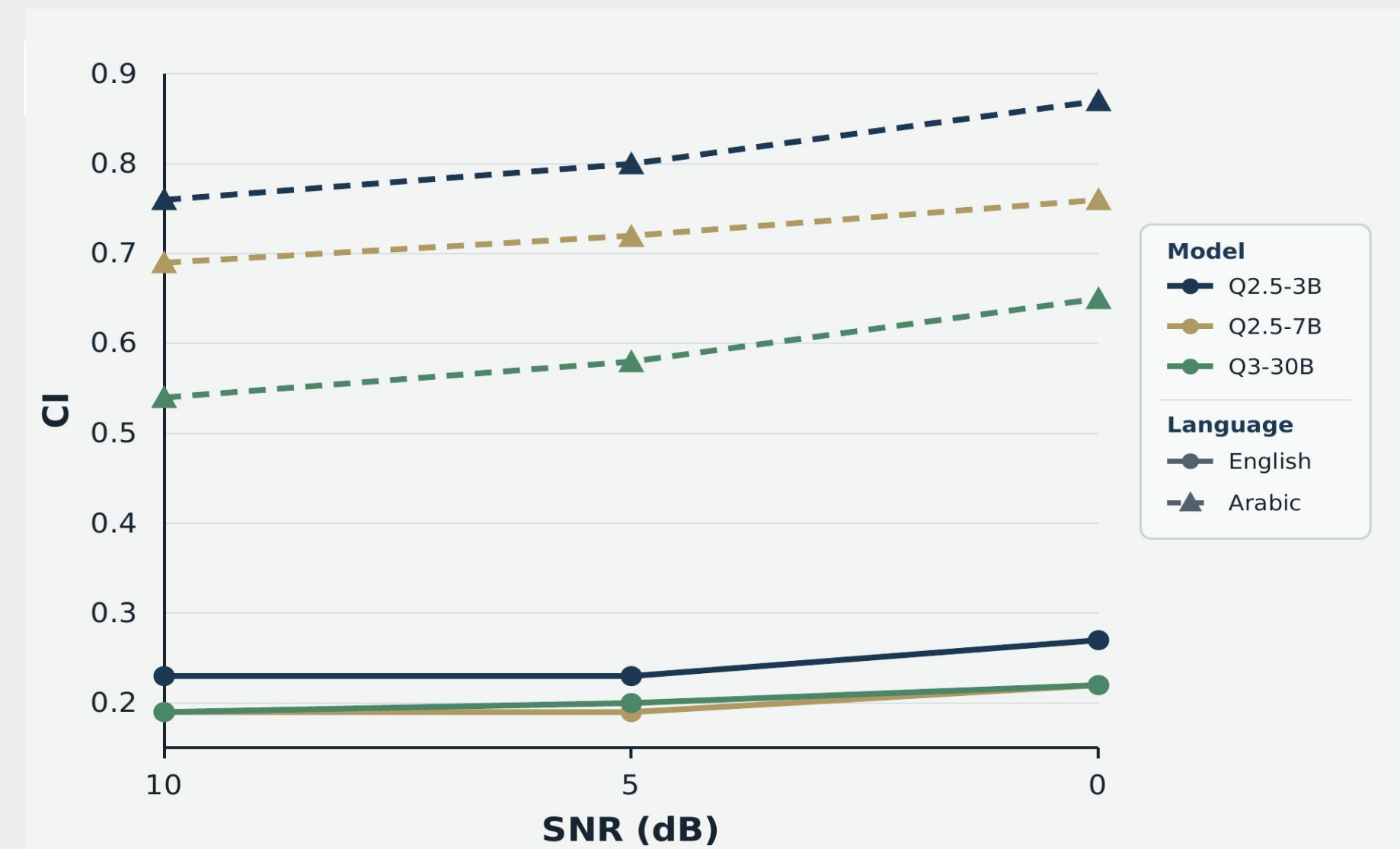
- English remains stable across modalities
- Arabic speech sharply increases CI for every model
- Q3-30B is most stable in Arabic
- Phi-4M collapses under Arabic speech (CI = 0.97)

Language x Modality

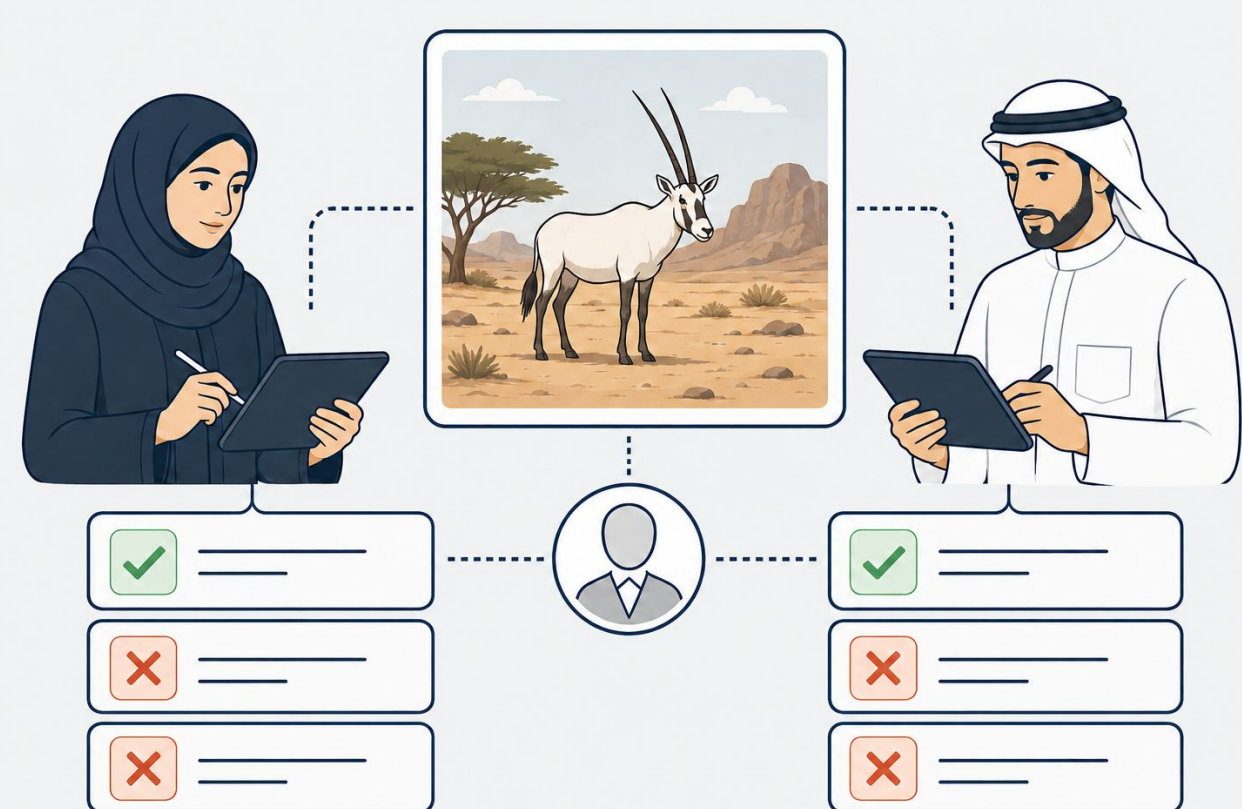
- Speech is not a neutral input channel
- Arabic is consistently less stable:** the difference is greater for speech
- Scaling helps, but does not solve the problem:** larger models reduce CI, yet the gap remains

What the SNR Stress Test Shows?

- Low SNR means more acoustic noise
- CI rises monotonically as SNR falls, especially in Arabic
- Q3-30B is most robust across SNR levels
- Noise amplifies an existing language gap



Inter Annotator Agreement



- Human verification using 15% stratified samples
- Need Image
 - AC1 0.996 | Agreement 0.996
- Q-Correctness
 - AC1 0.741 | Raw Agreement 0.775
 - Substantial agreement on correctness

References:

- EverydayMMQA: A Multilingual and Multimodal Framework for Cultural Grounded Spoken Visual QA. 2025.
- Once Correct, Still Wrong: Counterfactual Hallucination in Multilingual Visian Language Models, ACL 2026

