

Motivation

What standard benchmarks miss

- Western-centric image distribution
- Generic object/relation semantics
- Limited dialectal evaluation

What MENA Settings expose

- Architecture, clothing, food, artifacts, market
- Country and regional visual cues
- MSA, Egyptian, Levantine prompts

Research Questions

- Do standard accuracy metrics **capture** hallucination?
- How does hallucination **vary across languages/dialects**?
- How **robust** are different VLM families?
- How do **prompting strategies** affect hallucination?
- Does **model scaling** reduce hallucination?

M2CQA: Multilingual Multimodal Counterfactual Question Answering



Contrastive statements over MENA region

The decorative lantern is representing:

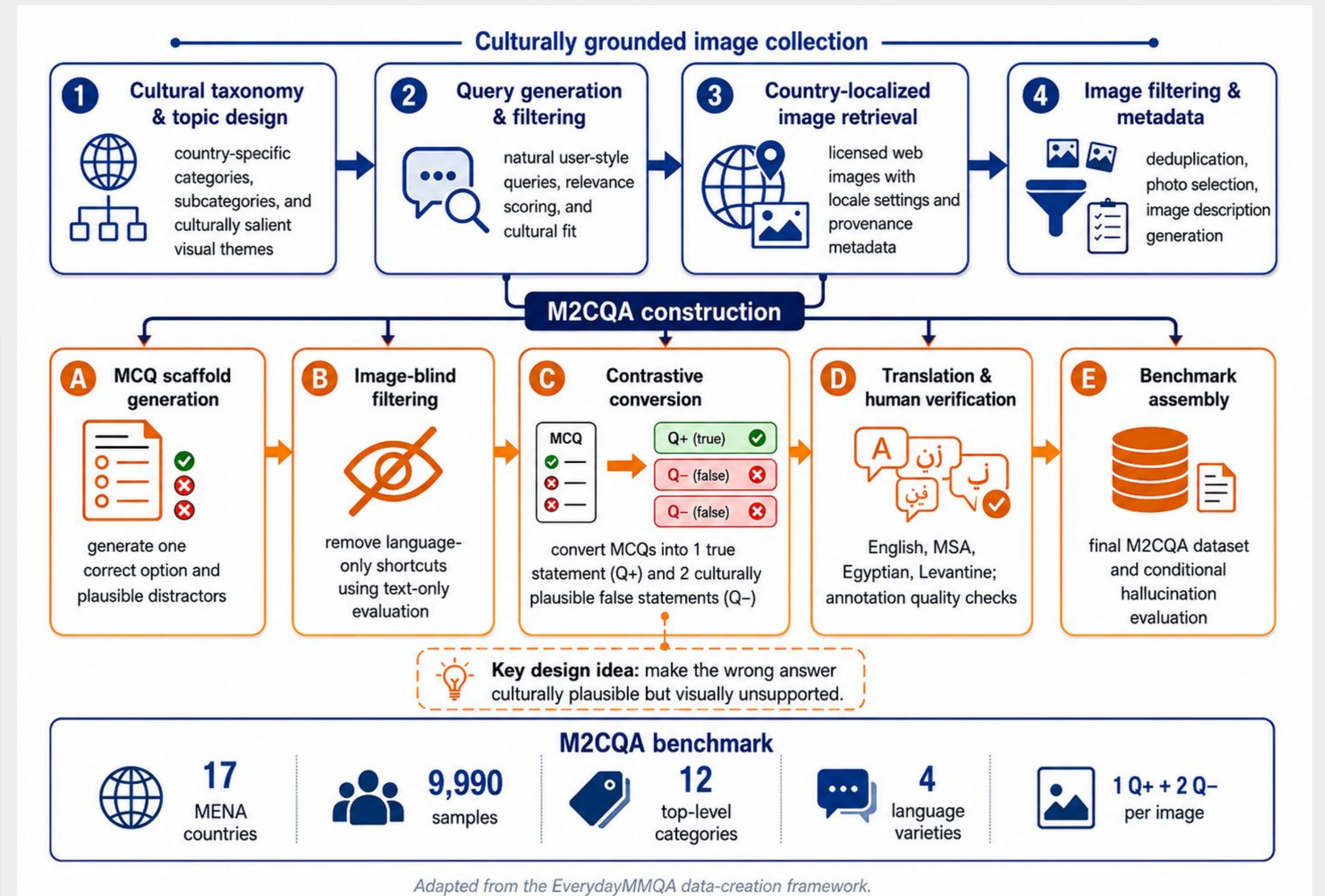
Q+: Ramadan Celebration

Q-: Christmas Eve

Q-: Diwali Celebration

Counterfactual Hallucination Rate CFHR = P (fails on a Q- | succeeded on Q+)

- Q+ Success shows the model can identify image-grounded interpretation
- Q- Failure after Q+ succeeds reveals over-acceptance of plausible false alternatives.
- CFHR to be interpreted jointly with Q+ accuracy

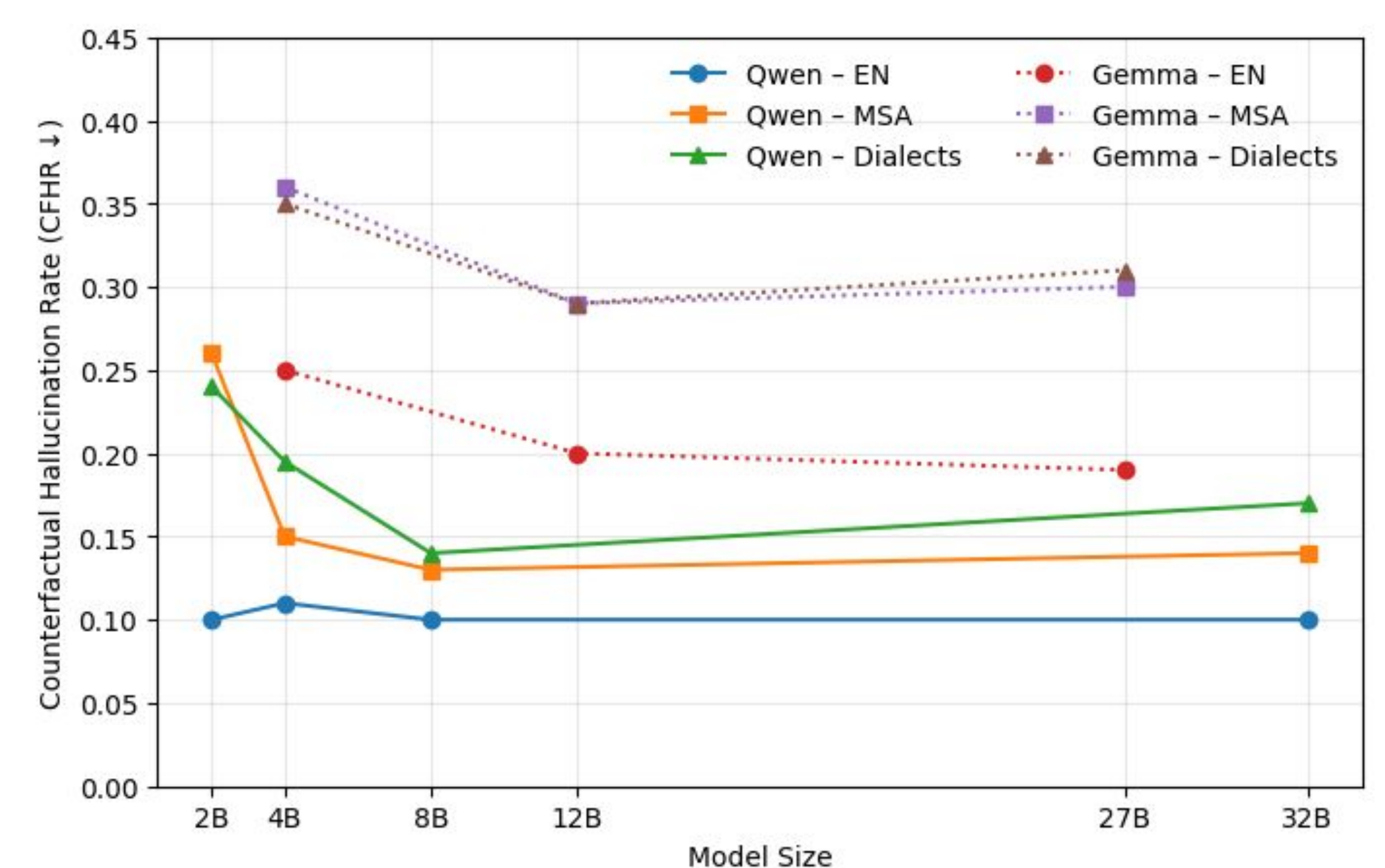


Findings

Lang	Model	A. True or False			CFHR ↓
		Q+	Q-	F1	
EN	Qwen3-VL-32B	0.93	0.94	0.93	0.10
	Gemma-3-27B	0.94	0.88	0.91	0.19
	FanarOryx-v1	0.94	0.93	0.93	0.12
	AIN	0.93	0.91	0.92	0.14
MSA	Qwen3-VL-32B	0.88	0.91	0.90	0.14
	Gemma-3-27B	0.93	0.81	0.86	0.30
	FanarOryx-v1	0.74	0.92	0.82	0.14
	AIN	0.95	0.60	0.74	0.55
Dialect	Qwen3-VL-32B	0.88	0.90	0.88	0.17
	Gemma-3-27B	0.93	0.79	0.86	0.32
	FanarOryx-v1	0.71	0.92	0.80	0.15
	AIN	0.97	0.38	0.56	0.79

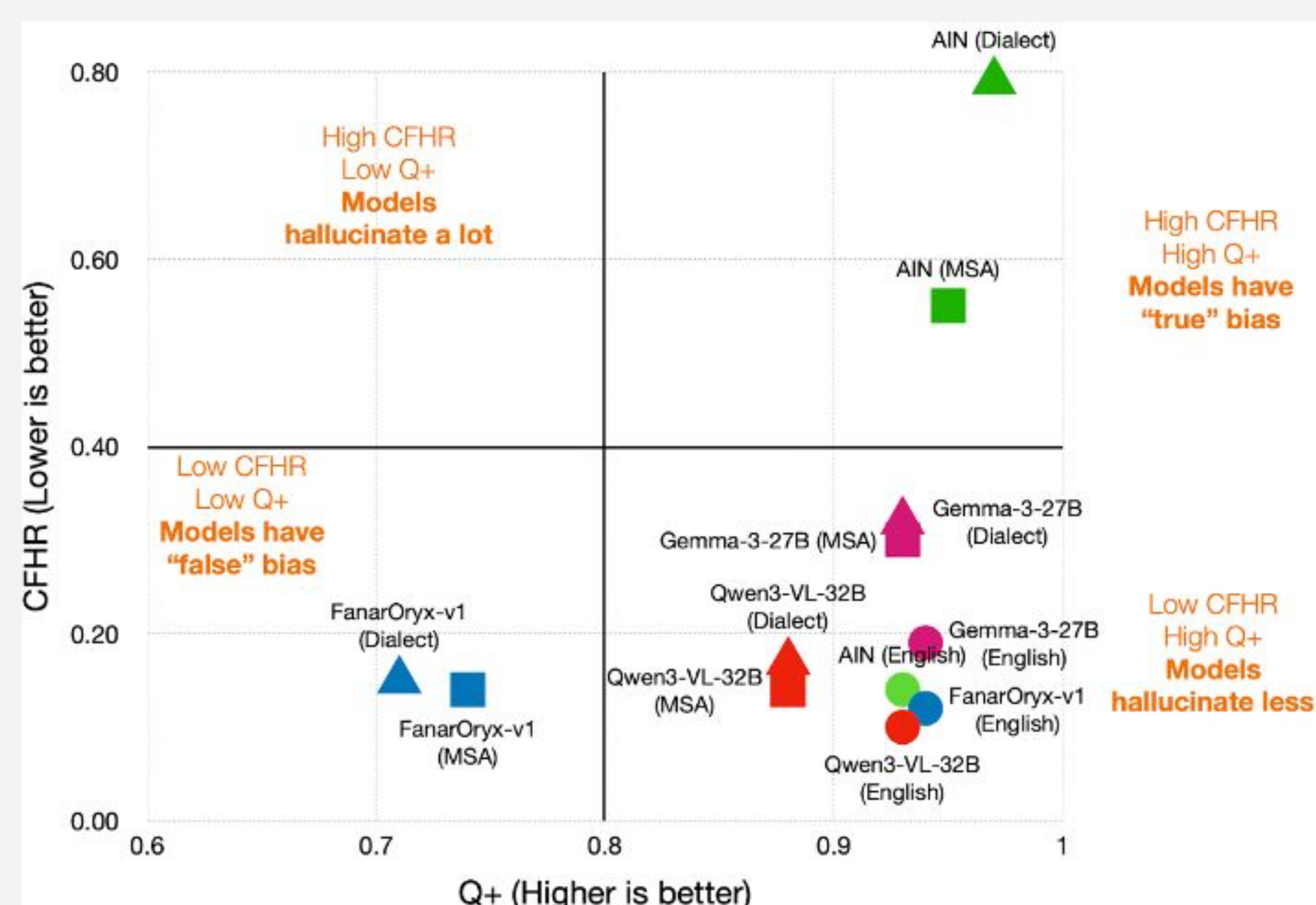
Cross-model Analysis

- Qwen3-VL** strongest counterfactual robustness across languages
- Gemma-3** has higher Arabic CFHR, scaling helps
- Fanar** exhibits conservative rejection: low CFHR, low Q+
- Ain** exhibits over-acceptance: high Q+, but also high CFHR



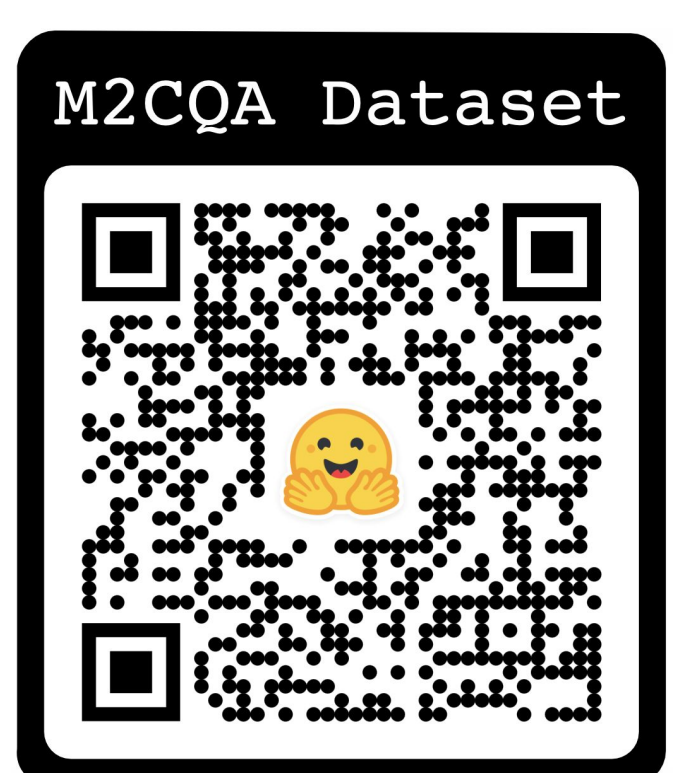
Joint Interpretation: Q+ & CFHR

- High Q+, Low CFHR**: Grounded and robust
- Low Q+, Low CFHR**: Conservative rejection; answer is not necessarily grounded
- High Q+, High CFHR**: Hidden over-acceptance bias towards *True*

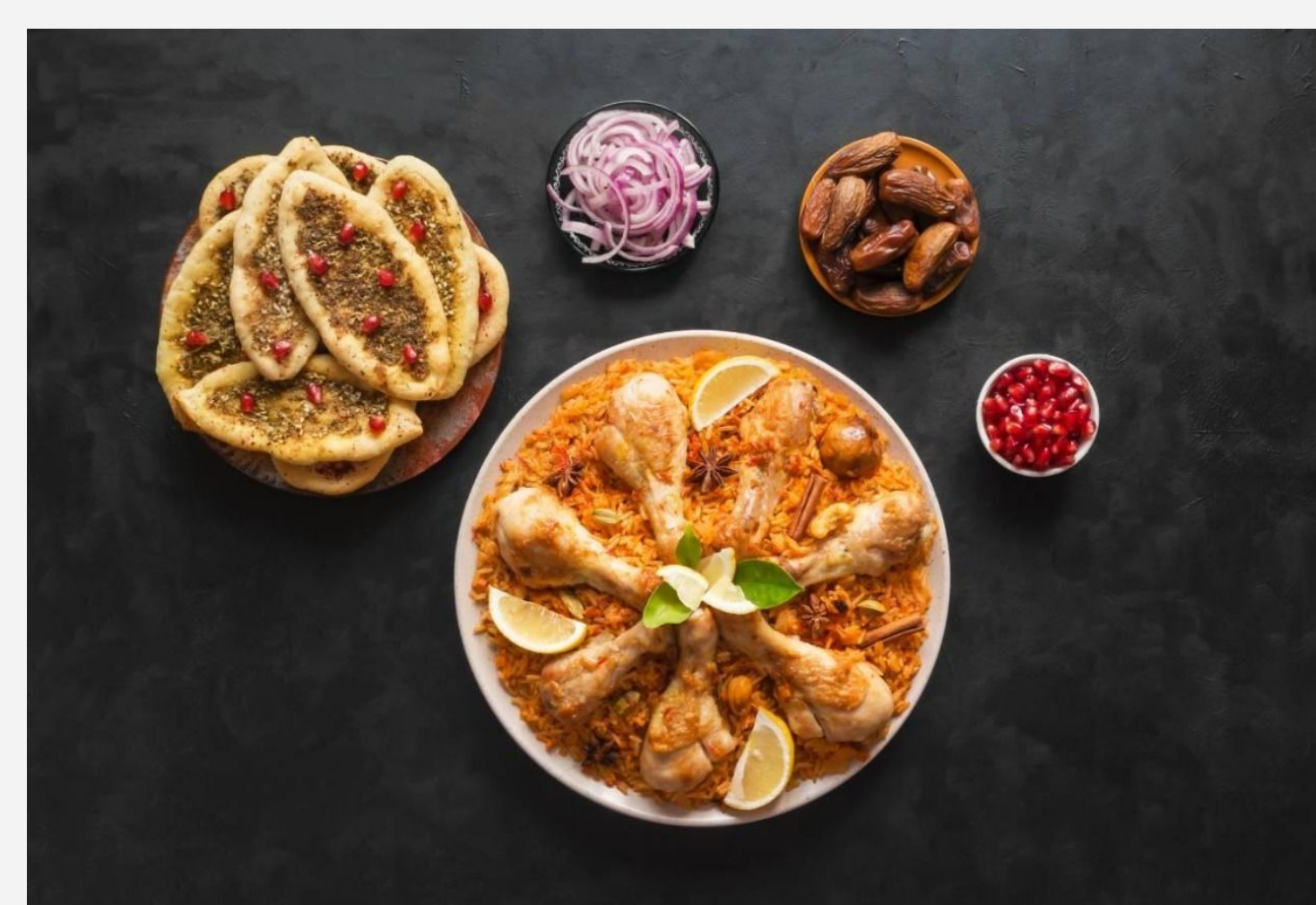


Cross-lingual Analysis

- Models are most **robust in English**
- Arabic** introduces more **counterfactual susceptibility**
- Egyptian** and **Levantine** **amplify** the gap



Prompt Analysis



Question: Star anise is commonly used in Middle Eastern cuisine and is visible in the chicken biryani shown in the image. True or False?

- A:** Answer only (True/False)
- B:** Answer then Reason
- C:** Reason then Reason

Observations:

- Answer then reason (B) generally reduces hallucination over A
- Reasoning first increases hallucination over B
- Pre-answer reasoning can override image-grounded evidence by introducing culturally plausible but visually unsupported constraints