Invertible AI

PEARL: A Multimodal Culturally-Aware Arabic Instruction Dataset

Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmen, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El Aatar, Nour Qandos, Rahaf Alhamouri, Samar Ahmad, Mohammed Anwar Al-Ghrawi, Aminetou Yacoub, Ruwa AbuHweidi, Vatimetou Mohamed Lemin, Reem Abdel-Salam, Ahlam Bashiti, Aisha Alansari, Ahmed Ashraf, Nora Alturayeif, Alcides Alcoba Inciarte, Adel Ammar, Abdelrahim A. Elmadany, Mohamedou Cheikh Tourad, Ismail Berrada, Mustafa Jarrar, Shady Shehata, Muhammad Abdul-Mageed

The University of British Columbia

The Problem: Cultural Bias in AI

- **Western-Centric Bias**

  Encodes Western cultural perspectives

- **Shallow Understanding**

  Datasets lack cultural depth.

- **Cultural Uniformity**

  Overlook nuanced regional variations

- **Our Solution**

  Introduce PEARL dataset, benchmarks

Introducing PEARL

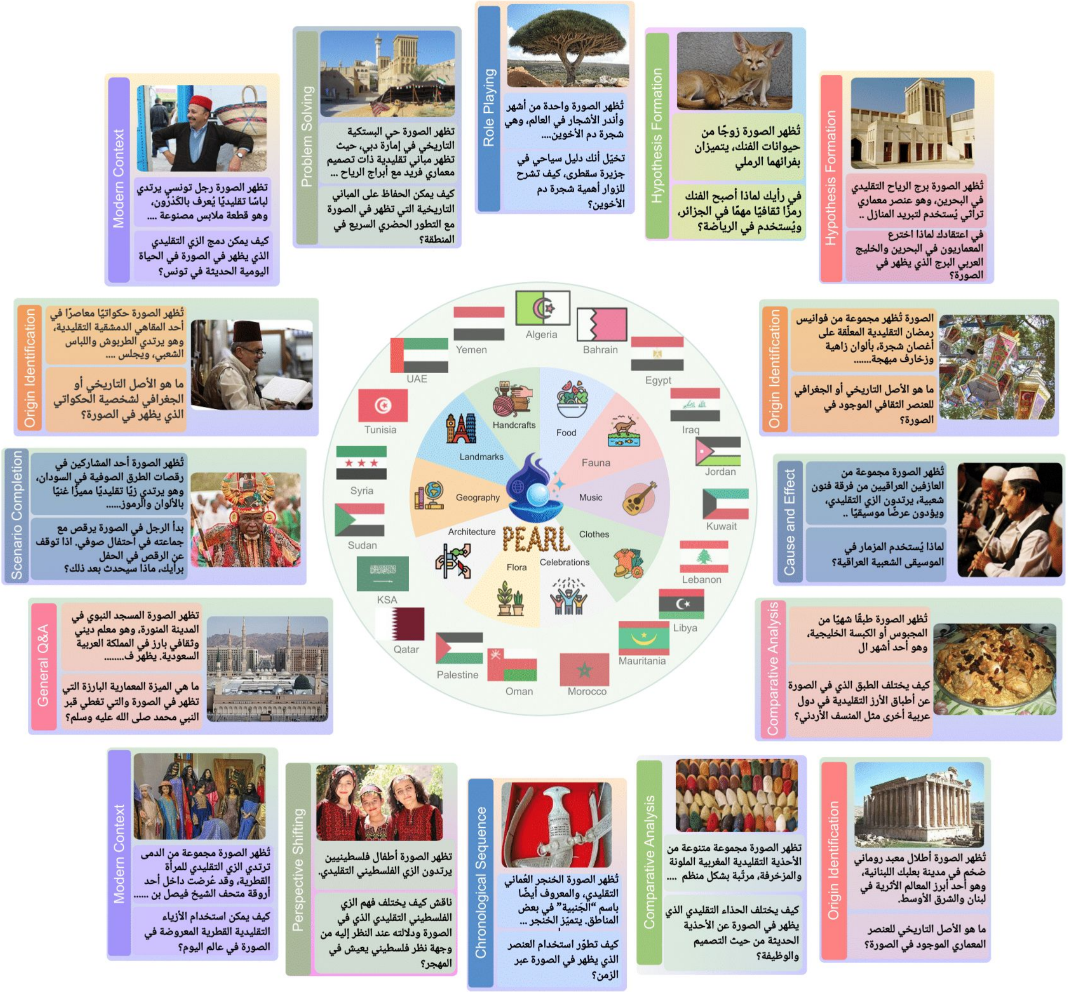
PEARL is a foundational resource designed to make LVLMs more culturally intelligent for Arabic.

- Massive Scale

  Over 309K multimodal examples

- Comprehensive Coverage
  - 19 Arab Countries
  - Spans 10 key cultural domains
- Authentic & Nuanced

  Built by 37 native annotators.



Modern Context

تظهر الصورة رجل تونسي يرتدي لباسًا تقليديًا يُعرف بالكَدْرُون، وهو قطعة ملابس مصنوعة ....

كيف يمكن دمج الزي التقليدي الذي يظهر في الصورة في الحياة اليومية الحديثة في تونس؟

Problem Solving

تظهر الصورة حي البستكية التاريخي في إمارة دبي، حيث تظهر مباني تقليدية ذات تصميم معماري فريد مع أبراج الرياح ...

كيف يمكن الحفاظ على المباني التاريخية التي تظهر في الصورة مع التطور الحضري السريع في المنطقة؟

Role Playing

تُظهر الصورة واحدة من أشهر وأندر الأشجار في العالم، وهي شجرة دم الأخوين....

تخيّل أنك دليل سياحي في جزيرة سقطرى، كيف تشرح للزوار أهمية شجرة دم الأخوين؟

Hypothesis Formation

تُظهر الصورة زوجًا من حيوانات الفنك، يتميزان بفرائهما الرملي

في رأيك لماذا أصبح الفنك رمزًا ثقافيًا مهمًا في الجزائر، ويُستخدم في الرياضة؟

Hypothesis Formation

تُظهر الصورة برج الرياح التقليدي في البحرين، وهو عنصر معماري تراثي يُستخدم لتبريد المنازل ..

في اعتقادك لماذا اخترع المعماريون في البحرين والخليج العربي البرج الذي يظهر في الصورة؟

Origin Identification

تُظهر الصورة حكواتيًا معاصرًا في أحد المقاهي الدمشقية التقليدية، وهو يرتدي الطربوش واللباس الشعبي، ويجلس ....

ما هو الأصل التاريخي أو الجغرافي لشخصية الحكواتي الذي يظهر في الصورة؟

Origin Identification

الصورة تُظهر مجموعة من فوانيس رمضان التقليدية المعلّقة على أغصان شجرة، بألوان زاهية وزخارف مبهجة.......

ما هو الأصل التاريخي أو الجغرافي للعنصر الثقافي الموجود في الصورة؟

Scenario Completion

تُظهر الصورة أحد المشاركين في رقصات الطرق الصوفية في السودان، وهو يرتدي زيًا تقليديًا مميزًا غنيًا بالألوان والرموز......

بدأ الرجل في الصورة يرقص مع جماعته في احتفال صوفي. اذا توقف عن الرقص في الحفل برأيك، ماذا سيحدث بعد ذلك؟

Cause and Effect

تُظهر الصورة مجموعة من العازفين العراقيين من فرقة فنون شعبية، يرتدون الزي التقليدي، ويؤدون عرضًا موسيقيًا ..

لماذا يُستخدم المزمار في الموسيقى الشعبية العراقية؟

General Q&A

تظهر الصورة المسجد النبوي في المدينة المنورة، وهو معلم ديني وثقافي بارز في المملكة العربية السعودية. يظهر ف........

ما هي الميزة المعمارية البارزة التي تظهر في الصورة والتي تغطي قبر النبي محمد صلى الله عليه وسلم؟

Comparative Analysis

تُظهر الصورة طبقًا شهيًا من المجبوس أو الكبسة الخليجية، وهو أحد أشهر ال

كيف يختلف الطبق الذي في الصورة عن أطباق الأرز التقليدية في دول عربية أخرى مثل المنسف الأردني؟

Modern Context

تُظهر الصورة مجموعة من الدمى ترتدي الزي التقليدي للمرأة القطرية، وقد عُرضت داخل أحد أروقة متحف الشيخ فيصل بن ......

كيف يمكن استخدام الأزياء التقليدية القطرية المعروضة في الصورة في عالم اليوم؟

Perspective Shifting

تظهر الصورة أطفال فلسطينيين يرتدون الزي الفلسطيني التقليدي.

ناقش كيف يختلف فهم الزي الفلسطيني التقليدي الذي في الصورة ودلالته عند النظر إليه من وجهة نظر فلسطيني يعيش في المهجر؟

Chronological Sequence

تُظهر الصورة الخنجر العُماني التقليدي، والمعروف أيضًا باسم "الجَنبية" في بعض المناطق. يتميّز الخنجر ...

كيف تطوّر استخدام العنصر الذي يظهر في الصورة عبر الزمن؟

Comparative Analysis

تظهر الصورة مجموعة متنوعة من الأحذية التقليدية المغربية الملونة والمزخرفة، مرتّبة بشكل منظم ....

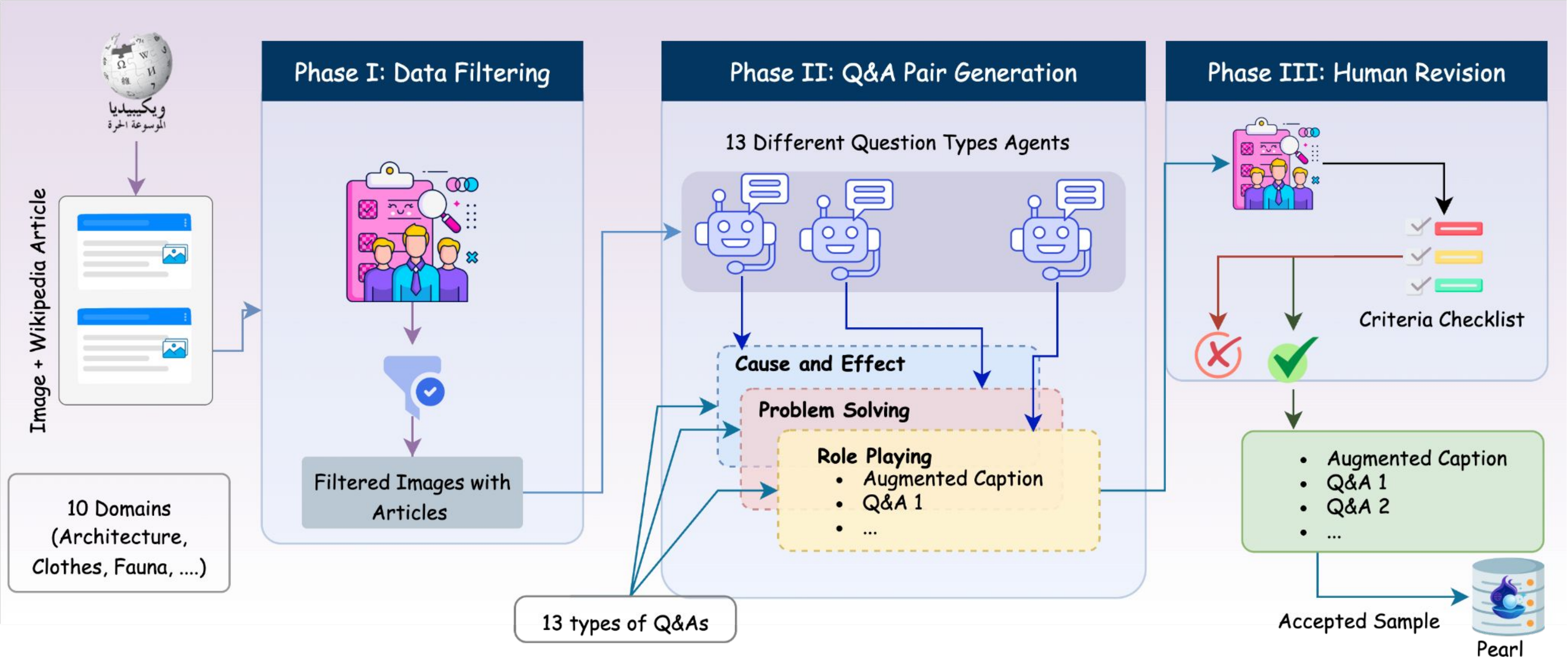
كيف يختلف الحذاء التقليدي الذي يظهر في الصورة عن الأحذية الحديثة من حيث التصميم والوظيفة؟

Origin Identification

تُظهر الصورة أطلال معبد روماني ضخم في مدينة بعلبك اللبنانية، وهو أحد أبرز المعالم الأثرية في لبنان والشرق الأوسط.

ما هو الأصل التاريخي للعنصر المعماري الموجود في الصورة؟

How We Built PEARL: A 3-Phase Workflow



The PEARL Ecosystem: Datasets for All Needs

PEARL isn't just one dataset; it's a suite of resources available to the research community.

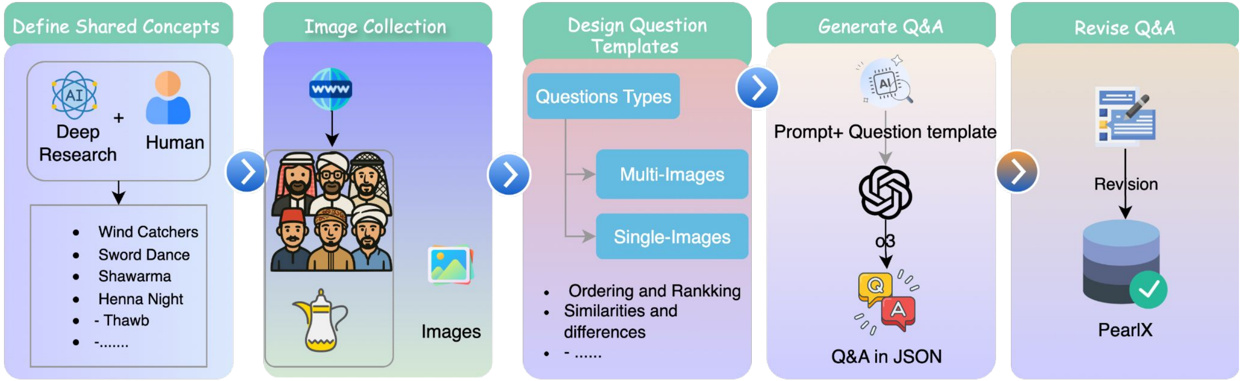
- **Core Assets:**
  - **12k** Images
  - **309k** Automated Q&A Pairs with captions.
  - **16k** Human-Revised Q&A Pairs
- **Evaluation Benchmarks:**
  - **PEARL**: The main benchmark with **6,301** samples.
  - **PEARL-LITE**: A smaller **893** samples.
  - **PEARL-X**: A specialized benchmark with **367** Q&As with **61** shared concepts

A Deeper Challenge: PEARL-X

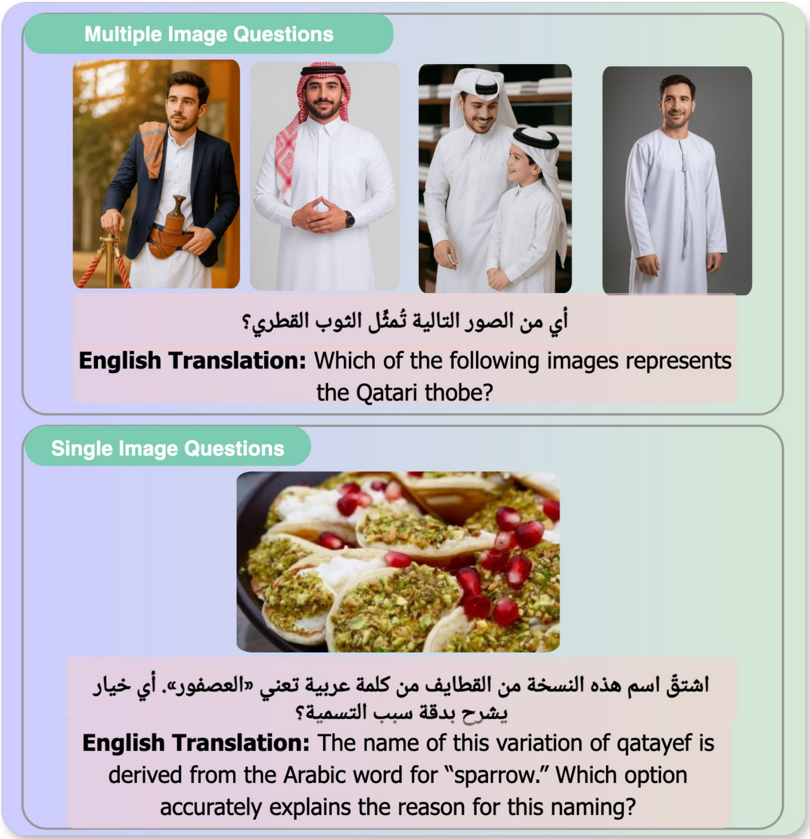
- ## The Goal of PEARL-X

  To test if models can recognize these nuanced differences.

- ## How it Works
  - It focuses on 61 shared cultural concepts.
  - It uses both single-image and multi-image questions.



PEARL-X (Examples)



Multiple Image Questions

أي من الصور التالية تُمثِّل الثوب القطري؟

English Translation: Which of the following images represents the Qatari thobe?

Single Image Questions

اشتقّ اسم هذه النسخة من القطايف من كلمة عربية تعني «العصفور». أي خيار يشرح بدقة سبب التسمية؟

English Translation: The name of this variation of qatayef is derived from the Arabic word for "sparrow." Which option accurately explains the reason for this naming?

The Experiment: Testing the Best AI Models

- **Models Tested**
  - **Open Models** and **Proprietary Models**
  - Different sizes and architectures
  - Reasoning and non reasoning models
- **Evaluation Protocol**
  - LVLM-as-Judge (InternVL3.5-38B)
  - ***Accuracy*** for Closed-form questions.
  - ***Correctness***, ***Coherence***, ***Fluency***, ***Detail***, and ***Overall*** for open-ended questions
  - ***Cultural Awareness Score*** (CAS) binary score for open-end questions
- **Validated by human evaluators and found a strong correlation (ICC=0.708)**

Key Finding 1: Reasoning Beats Scale

- **The Surprise Winner**

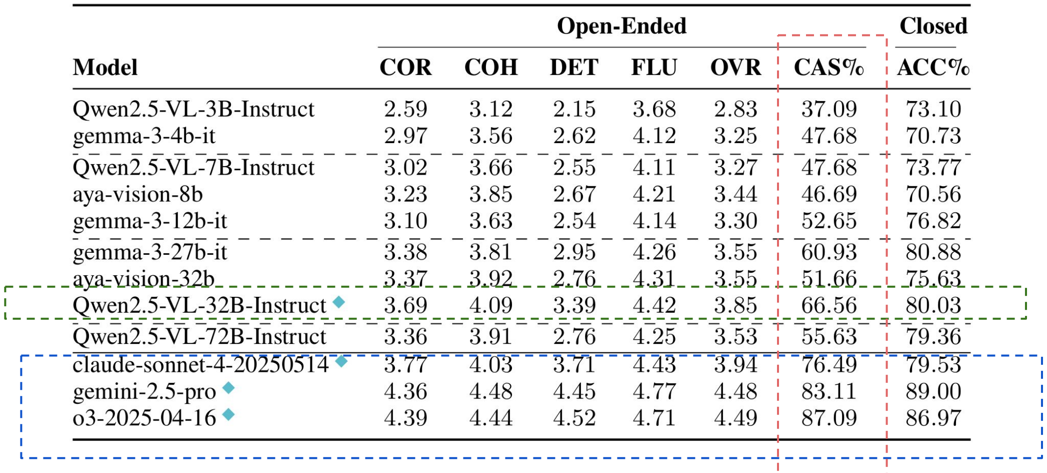
  Reasoning model outperformed larger one.

- **The Takeaway**

  Reasoning alignment is key.

- **Proprietary Models Lead**

  Proprietary models superior overall.



	Open-Ended						Closed
Model	**COR**	**COH**	**DET**	**FLU**	**OVR**	**CAS%**	**ACC%**
Qwen2.5-VL-3B-Instruct	2.59	3.12	2.15	3.68	2.83	37.09	73.10
gemma-3-4b-it	2.97	3.56	2.62	4.12	3.25	47.68	70.73
Qwen2.5-VL-7B-Instruct	3.02	3.66	2.55	4.11	3.27	47.68	73.77
aya-vision-8b	3.23	3.85	2.67	4.21	3.44	46.69	70.56
gemma-3-12b-it	3.10	3.63	2.54	4.14	3.30	52.65	76.82
gemma-3-27b-it	3.38	3.81	2.95	4.26	3.55	60.93	80.88
aya-vision-32b	3.37	3.92	2.76	4.31	3.55	51.66	75.63
Qwen2.5-VL-32B-Instruct ◆	3.69	4.09	3.39	4.42	3.85	66.56	80.03
Qwen2.5-VL-72B-Instruct	3.36	3.91	2.76	4.25	3.53	55.63	79.36
claude-sonnet-4-20250514 ◆	3.77	4.03	3.71	4.43	3.94	76.49	79.53
gemini-2.5-pro ◆	4.36	4.48	4.45	4.77	4.48	83.11	89.00
o3-2025-04-16 ◆	4.39	4.44	4.52	4.71	4.49	87.09	86.97

Results for the PEARL-LITE subset. The metrics are as follows: Correctness (COR), Coherence (COH), Detail (DET), Fluency (FLU), Overall Score (OVR), Cultural Awareness Score (CAS) for open-ended questions, and Accuracy (ACC) for closed-ended questions.

Key Finding 2: Nuance is Hard

The PEARL-X benchmark proved to be a difficult test for every model.

- **The Challenge**

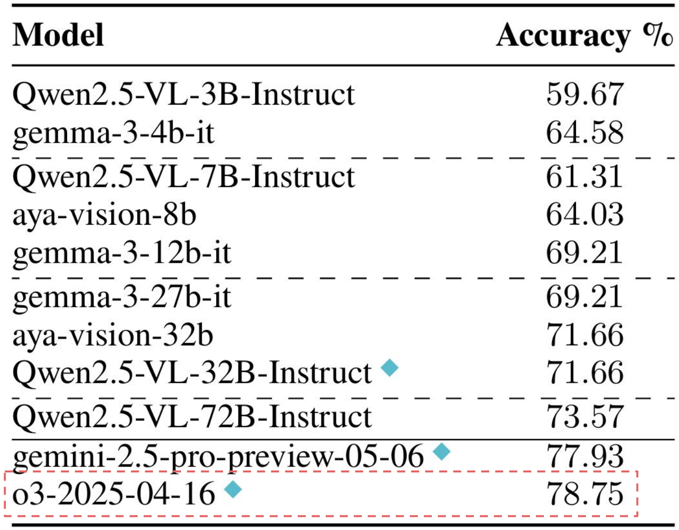
  Identifying subtle differences is tough.

- **Performance**

  Top models show difficulty.

- **Implication**

  Reasoning is critical for nuance.



Model	Accuracy %
Qwen2.5-VL-3B-Instruct	59.67
gemma-3-4b-it	64.58
Qwen2.5-VL-7B-Instruct	61.31
aya-vision-8b	64.03
gemma-3-12b-it	69.21
gemma-3-27b-it	69.21
aya-vision-32b	71.66
Qwen2.5-VL-32B-Instruct ◆	71.66
Qwen2.5-VL-72B-Instruct	73.57
gemini-2.5-pro-preview-05-06 ◆	77.93
o3-2025-04-16 ◆	78.75

Accuracy on the Pearl-X shared-concepts benchmark.

Conclusion

- **Our Contribution:** First large-scale Arabic benchmark.
- **Key Insight:** Reasoning-alignment is most effective.
- **New Frontiers:** With PEARL-X, we've provided a novel way to measure a model's understanding of nuanced, cross-country cultural variations.
- **Limitations:** Wikipedia bias, representation gaps.

Thank you!

arXiv

GitHub

Canada Research Chairs

NSERC
CRSNG

Digital Research
Alliance of Canada

Social Sciences and
Humanities Research
Council of Canada

Conseil de recherches
en sciences humaines
du Canada

Canada

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION
FONDATION CANADIENNE POUR L'INNOVATION