ARB: A Comprehensive Arabic Multimodal Reasoning Benchmark

Sara Ghaboura[1*], Shubham Patle[1*], Ketan More[1], Wafa Alghallabi[1], Omkar Thawakar[1], Jorma Laaksonen[2], Hisham Cholakkal[1], Salman Khan[1,3], Rao Anwer[1]

[1]Mohamed bin Zayed University of AI, Abu Dhabi, UAE
[2]Aalto University, Espoo, Finland
[3]Australian National University, Canberra, Australia
{sara.ghaboura, shubham.patle, wafa.alghallabi, omkar.thawakar}@mbzuai.ac.ae

Abstract

As Large Multimodal Models (LMMs) become more capable, there is growing interest in evaluating their reasoning processes alongside their final outputs. However, most existing benchmarks remain focused on English, overlooking languages with rich linguistic and cultural depth such as Arabic. To address this gap, we introduce the Comprehensive Arabic Multimodal Reasoning Benchmark (ARB), the first benchmark designed to evaluate step-by-step reasoning in Arabic across both textual and visual modalities. ARB covers 11 diverse domains and over 40 subfields, including visual reasoning, optical character recognition, scientific analysis, and cultural interpretation. It comprises 2,219 multimodal samples paired with over 8K human-curated reasoning steps and corresponding actions, verified through a human-in-the-loop process. We evaluated 15 state-of-the-art open- and closed-source LMMs and found persistent challenges in coherence, faithfulness, and cultural grounding. ARB provides a structured framework for diagnosing multimodal reasoning in underrepresented languages, marking a critical step toward inclusive, transparent, and culturally aware AI systems. The benchmark [1], rubric, and evaluation suite[2] are publicly available.

Keywords: Arabic Multimodal Reasoning, Benchmarking Multimodal Models, Low-Resource Languages

1. Introduction

Arabic, spoken by more than 400 million people worldwide, embodies significant linguistic diversity and a profound cultural heritage. Despite its global importance, Arabic remains underrepresented in advanced AI systems, particularly those involving multimodal reasoning, simultaneous interpretation, and the logical processing of textual and visual information crucial for education, healthcare, and cultural preservation. This scarcity limits inclusive AI applications and hinders equitable technological advancement within Arabic-speaking communities.

Recent advances in reasoning-centric modeling, particularly chain-of-thought (CoT) prompting (Wei et al., 2022), have improved interpretability by encouraging models to articulate intermediate reasoning traces. These ideas have extended to multimodal contexts through models such as LLaVA-CoT (Xu et al., 2025), VisCoT (Shao et al., 2024), and LlamaV-o1 (Thawakar et al., 2025). However, most reasoning-oriented benchmarks remain English-only, overlooking the linguistic and cultural nuances essential for Arabic reasoning.

To address this gap, we introduce ARB, the Comprehensive Arabic Multimodal Reasoning Benchmark, the first dataset explicitly designed for step-by-step multimodal reasoning in Arabic. ARB contains 2,219 multimodal samples in 11 domains and over 40 fine-grained subdomains or tasks, spanning visual reasoning, document understanding, optical character recognition (OCR), cultural interpretation, biodiversity, medicine, and remote sensing. Each sample is paired with human-curated reasoning steps and actions, reviewed, verified, and validated by 5 native Arabic speakers volunteering as expert annotators through a human-in-the-loop process to ensure logical precision and cultural grounding.

We evaluate 15 open- and closed-source LMMs, including GPT-4V (OpenAI, 2024b,a, 2025b,c,a), Gemini (DeepMind, 2024; Comanici et al., 2025; DeepMind and AI, 2025), AIN (Heakl et al., 2025), Gemma 3 (Team-Gemma et al., 2025), Aya-vision (Cohere-Labs, 2025), InternVL3 (Chen et al., 2024b), LLaMA-4 Scout (Meta-AI, 2025), and Qwen (Qwen-Team, 2025; Team, 2025) variants, and identify persistent weaknesses in reasoning coherence, cultural awareness, and step consistency. ARB thus provides a unified framework for benchmarking multimodal reasoning in Arabic and lays the foundation for inclusive, culturally grounded AI systems.

In summary, (1) We introduce **ARB**, the first

* Equal contribution

[1]Dataset:https://huggingface.co/datasets/MBZUAI/ARB

[2]Page:https://mbzuai-oryx.github.io/ARB/.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9202–9216
11–16 May 2026.

Benchmark	MM*	MD*	R*	Open*	FA*/S*
Henna	✓	✗	✗	✗	FA*
CAMEL-Bench	✓	✓	✗	✓	FA*
AraDiCE	✗	✓	✓	✓	FA*
JEEM	✓	✓	✗	✓	FA*
ArabCulture	✗	✓	✓	✓	FA*
ARB (ours)	✓	✓	✓	✓	S

Table 1: Comparison of ARB with existing Arabic benchmarks. Note*: **MM**: Multimodal; **MD**: Multidomain; **R**: Reasoning support; **Open**: Open-source; **FA**: Final-Answer; **S**: Steps support.

large-scale benchmark for Arabic multimodal reasoning with human-curated step-by-step explanations across 11 domains; (2) We propose a structured Arabic evaluation framework combining similarity metrics, an LLM-as-Judge protocol, and inter-annotator agreement (IAA) analysis; and (3) We evaluate 15 state-of-the-art models, revealing critical gaps in Arabic reasoning, cultural grounding, and step coherence.

2. Related Work

Reasoning in Large Models. CoT prompting (Wei et al., 2022) enhanced large language models (LLMs) reasoning by detailing intermediate steps, inspiring extensions such as self-consistency (Wang et al., 2022), tree-of-thoughts (Yao et al., 2023), and reasoning-tuned instruction models (Vaillancourt and Thompson, 2024). Modern approaches, including OpenAI's o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025), reinforce logical fidelity through reinforcement learning and inference-time scaling.

Multimodal Reasoning. Integrating reasoning across vision and language has yielded models such as LLaVA-CoT (Xu et al., 2025) and LlamaV-o1 (Thawakar et al., 2025), which explicitly structure visual reasoning steps. Further studies (Chen et al., 2024a; Zhang et al., 2024) refine reasoning coherence through rationale distillation and preference optimization, highlighting the growing emphasis on interpretable multimodal thought.

Arabic Reasoning Resources. Existing Arabic datasets, summarized in Table 1, including Henna (Alwajih et al., 2024), CAMEL-Bench (Ghaboura et al., 2025a), AraDiCE (Mousi et al., 2024), JEEM (Kadaoui et al., 2025), and ArabCulture (Sadallah et al., 2025), remain limited to final-answer evaluation and do not include stepwise multimodal annotations. Although native Arabic models such as ALLaM-Thinking (Allam-Research, 2025) and Fanar (Fanar-Team et al., 2025) have advanced reasoning capabilities, their scope remains limited to textual tasks. Meanwhile, the multimodal AIN (Heakl et al., 2025) focuses on final-answer evaluation without stepwise reasoning.

ARB addresses this gap by providing the first resource for step-by-step multimodal reasoning in Arabic.

3. ARB: Step-by-Step Arabic Reasoning Benchmark

Figure 1 presents an overview of ARB data construction pipeline, outlining the main stages from data sourcing to reasoning-step generation and validation. The following subsections describe the stages in detail.

3.1. Data Collection

We adopt a domain-guided approach to curate data across a broad spectrum of categories relevant to Arabic multimodal reasoning, ensuring diversity in both content and modality. ARB spans 11 major domains, 1- Visual Reasoning (VR), 2- OCR and document understanding (OCR), 3- Charts, Diagrams, and Tables (CDT), 4- Mathematical and Logical Reasoning (M&L), 5- Social and Cultural Understanding (Soc.Cult.), 6- Complex Visual Perception (CVP), 7- Medical Image Analysis (Med), 8- Scientific Reasoning (Sci.R.), 9- Historical & Archaeological Interpretation (Hist.), 10- Remote Sensing Analysis (RS), and 11- Agri-Biodiversity Image Understanding (AgriB.). The selected domains, covering both textual and visual tasks in more than 40 subfields and tasks (Figure 2), are derived from existing benchmarks, human-authored questions, and synthetic content. These sources with their topics were chosen to capture various reasoning challenges and to promote linguistic, thematic, and cultural diversity in the dataset.

3.2. Data Generation and Processing

The first part focuses on model setup and prompt design, followed by large-scale, domain-specific data generation to maintain consistent reasoning quality and linguistic alignment across domains.

3.2.1. Model and Prompt Configuration

We compared GPT-4o (OpenAI, 2024b) and GPT-4o-mini (OpenAI, 2024a), selected for their efficiency and strong multimodal reasoning capabilities (Heakl et al., 2025), under Arabic and English prompt settings. A pilot study of 50 samples across multiple domains, evaluated by the annotators and verified with LaBSE (Feng et al., 2020), showed that GPT-4o with Arabic prompts produced the most coherent, fluent, and culturally

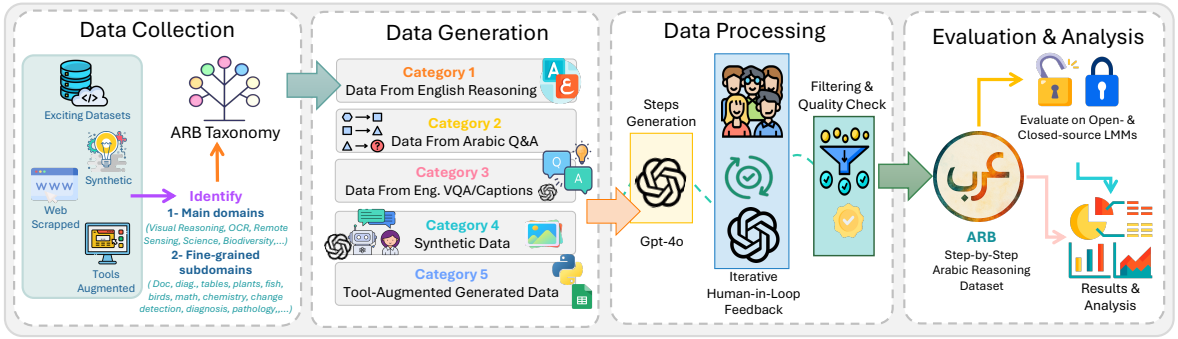


Figure 1: Overview of ARB dataset pipeline for evaluating Arabic multimodal reasoning. Data covers 11 domains derived from curated datasets, web sources, synthetic data, and tool-augmented content. Reasoning steps undergo human-in-the-loop validation and verifications to ensure logical consistency and cultural relevance.

Curriculum	Domain	Subfields / Tasks	
General	VR	Object localization Spatial reasoning Cause–effect recognition	Comparison Counting Visual analogy
	CVP	Scene understanding Cross-view inference	Fine-grained recognition Multiple-object interaction
	Soc.Cult.	Cultural artifacts Attire and traditions Event symbolism	Food identification Landmark recognition
	AgriB.	Plant disease detection Fruit & vegetable classification	Plant species identification Bird and fish breed recognition
	Hist.	Antiquities recognition Artwork classification	Artifact contextualization Historical object reasoning
Descriptive / Inference	RS	Land cover classification Geospatial scene parsing	Transportation detection Change detection (multi-image)
	OCR	Text localization Handwriting recognition	Document understanding Infographic interpretation
Scientific/ Medical	CDT	Chart interpretation Table reasoning	Visual data comparison Scientific diagram understanding
	Med.	Clinical diagnosis reasoning Anatomy & pathology ident.	Lab result interpretation Disease classification
	Sci.R.	Chemistry problem-solving Physics conceptual reasoning	Scientific method inference Social science interpretation
Computational	M&L	Geometry Arithmetic Trigonometry	Algebraic manipulation Numerical reasoning Symbolic logic

Figure 2: The ARB Domain Taxonomy defines four curricula and 49 subfields spanning diverse multimodal reasoning tasks and guiding step generation during data creation.

aligned reasoning traces. Consequently, this setting was adopted as the unified configuration for all reasoning-step generation and translation tasks.

Building on the insights of the pilot study, we grouped the ARB domains into four curriculum tracks (*General, Scientific/Medical, Computational, and Descriptive/Inferential*) following the subject-level reasoning approach of (Mustapha et al., 2024). Each curriculum was tailored to its respective domain (Figure 2). Curriculum-guided prompting improved reasoning accuracy by 12.8% in human evaluations over unguided prompting, confirming its role in producing domain-aligned reasoning chains. For readability, Figure 3 presents the English translation of the final Arabic prompt (see Appendix A.2) used across domains.

3.2.2. Data Generation

The dataset is structured into 5 primary categories, distinguished by the origin of the source data and the methodology employed for its creation and refinement.

Reasoning Steps Generation Prompt

You are a professional expert specialized in the field of **{Domain}**. Your task is to generate step-by-step logical analysis and reasoning for textual and visual questions, including the necessary action at each step to arrive at the correct answer. Your reasoning should be grounded in visual evidence from the image, the information provided in the question, and the available answer choices. Use the provided **{example}** as a structural template for formatting the reasoning steps and corresponding actions. Please follow the instructions below:

1. Read the question and available answer choices carefully.
2. Identify the core concepts, required skills, and domain-specific knowledge relevant to **{Domain}**.
3. Questions span multiple formats, and you must follow a curriculum-based approach specified by the **{Curriculum}** associated with each **{Domain}**.
4. The curricula are categorized into four types:
   - **First group - {Curriculum} = "Computational"**: Focuses on basic arithmetic operations, comparative reasoning, and mathematical logic.
   - **Second group - {Curriculum} = "Scientific/Medical"**: Involves scientific reasoning and domain-specific evidence-based analysis.
   - **Third group - {Curriculum} = "Descriptive/Inference"**: Emphasizes segmenting and analyzing the visual content to extract meaningful insights.
   - **Fourth group - {Curriculum} = "General"**: Relies on comparison, contrast, and what the question logically requires to reach the correct answer.

Please format your output according to the structure shown in **{example}**, and conclude with the phrase: "The correct answer is: _____".

Figure 3: English version of ARB prompt showing the translation of the original Arabic template used to guide reasoning-step generation. (For original Arabic prompt see 14 in Appendix A.2)

Table 2 summarizes the 5 categories along with their sources, languages, and creation methods. During translation, the model (GPT-4o) produced initial Arabic drafts that were iteratively reviewed and refined by the annotators to ensure linguistic fluency and cultural accuracy.

For multiple-choice questions (MCQ) and step-action chain creation, we applied domain-specific prompting strategies: few-shot CoT prompting for most reasoning and visual question-answering

Category	Data Source	Language	Type	Creation Method
Category 1: English Reasoning Benchmarks	VRC-Bench (Thawakar et al., 2025)	English	Reasoning/CoT	Translation + Human Refinement
	NaBirds-CoT (scottgeng00, 2025)	English	Reasoning/CoT	
Category 2: Arabic QA* Benchmarks	CAMEL-Bench (Ghaboura et al., 2025a)	Arabic	VQA	Plan-and-solve prompting framework(Wang et al., 2023)
	Exams-V (Das et al., 2024)	Arabic	VQA	+ Few-shot CoT prompting with HL*
Category 3: English Captioning and VQA* Benchmarks	TimeTravel (Ghaboura et al., 2025b)	English	Caption	Synthetic prompting framework (backward/forward strategy) with HL* (Shao et al., 2023) + Few-shot CoT prompting with HL*
	AgriCLIP (Nawaz et al., 2025)	English	Caption	
	ARCH (Gamper and Rajpoot, 2021)	English	Caption	
	Seg-Zero (Liu et al., 2025a,b)	English	Caption	
	PathVQA (He et al., 2020)	English	VQA	
	CDChat (Noman et al., 2024)	English	VQA	
Category 4: Synthetic Data	Pinterest (Pinterest, 2025)	Arabic	Caption	Few-shot CoT prompting with HL*
Category 5: Tool-augmented Generated Data	Human Created Themes & Topics	Arabic	Created MCQ	Visual part: **Chart:** Python(Python Software Foundation, 2024) & Matplotlib (Bisong and Bisong, 2019) **Tables:** Excel (Microsoft Corporation, 2024).
	AI2D(Kembhavi et al., 2016)	English	VQA	Visual diagrams were manually edited and relabeled in Arabic Few-shot CoT prompting with HL*

Table 2: Overview of ARB Data Categories. Each category lists its source dataset(s), language origin, task type, and data creation approach. Note HL*: Human-in-the-Loop.

English Original Text	Translated Text (Model Output)	Human Curated Text (ARB Dataset)	Issues/ Comments
The triangle ABC is isosceles with AB = BC = 16	المثلث ABC متساوي الأضلاع حيث AB = BC= 16...	المثلث ABC متساوي الساقين حيث AB = BC= 16...	Incorrect word choice — changed meaning entirely.
Subtract all big matte balls. Subtract all green rubber objects. How many objects are left?"	اطرح كل الكرات الكبيرة الغير لامعة. اطرح كل الأغراض المطاطية الخضراء. كم عدد الأغراض المتبقية؟	إطرح جميع الكرات الكبيرة غير اللامعة. إطرح جميع الأجسام المطاطية الخضراء. كم يبقى من الأجسام؟	Inaccurate term for "object" and missing "hamza" weakens the imperative form.
Extract values for females and males from the graph for 2009 to 2019.	استخراج القيم للنساء والرجال من الرسم البياني من 2009 إلى 2019.	استخراج القيم للإناث والذكور من الرسم البياني من 2009 إلى 2019.	"Women and men" misused — should be "females and males" (age-neutral).
the muse Henrietta Moraes sat ..	جلست العارضة هينريتا موراي..	جلست المُلهِمة هنرييتا موراي..	Contextually incorrect — "muse" mistranslated as "model."
Balancing pots and baskets on their heads symbolizes grace, balance...	يرمز موازنة الأواني والسلال على رؤوسهم إلى النعمة، التوازن...	ترمز موازنة الأواني والسلال على رؤوسهم إلى الرشاقة، التوازن...	Subject-verb mismatch and inaccurate rendering of "grace."
It is often accompanied by music, singing, and drumming.	غالبًا ما يرافقه الموسيقى، الغناء، والطبل	غالبًا ما يرافقه الموسيقى والغناء وقرع الطبول.	Misinterprets "drumming" — refers to the act, not the instrument.
This activity reinforces cultural identity..	هذه النشاطات تعزز الهوية الثقافية ...	هذا النشاط يعزز الهوية الثقافية ...	Singular/plural disagreement.

Figure 4: Examples of model translation errors. Human-curated cases showing translation and linguistic/cultural alignment issues found during verification.

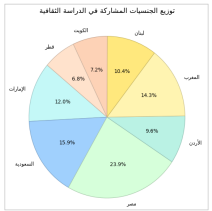
Input (Visual and Textual)	Model-Generated Reasoning Output	Post HL Verification Steps
السؤال: من خلال الرسم، مجموع نسبة لبنان والكويت تتجاوز ـ؟  **الخيارات:** أ. ٢٠٪ ب. ١٥٪ ج. ١٨٪	الخطوة 1: تحديد النسب المئوية للبنان والكويت من الرسم البياني. الإجراء 1: ملاحظة أن لبنان 4.15٪ والكويت هي 2.7٪. الخطوة 2: حساب مجموع النسبتين. الإجراء 2: جمع 4.15٪ و2.7٪. الخطوة 3: مقارنة الخيارات. الإجراء 3: المجموع الكلي هو 11.35٪. الخطوة 4: تحديد الخيار الذي يتجاوز المجموع الكلي. الإجراء 4: ملاحظة أن 11.35٪ لا يتجاوز 15٪ (الخيار ب) ولا يتجاوز 18٪ (الخيار ج) أو 20٪ (الخيار أ). الجواب النهائي:	الخطوة 1: تحديد النسب المئوية للبنان والكويت من الرسم البياني. الإجراء 1: ملاحظة أن نسبة لبنان هي 10.4٪ ونسبة الكويت هي 7.2٪. الخطوة 2: حساب مجموع النسبتين. الإجراء 2: جمع 10.4٪ و7.2٪ للحصول على المجموع الكلي. الخطوة 3: مقارنة المجموع الكلي مع الخيارات المتاحة. الإجراء 3: المجموع الكلي هو 17.6٪. الخطوة 4: تحديد الخيار الذي يتجاوز المجموع الكلي. الإجراء 4: ملاحظة أن 17.6٪ يتجاوز 15٪ (الخيار ب) ولكنه لا يتجاوز 18٪ (الخيار ج) أو 20٪ (الخيار أ). الجواب النهائي: ب

Figure 5: Examples of Post-Verification Reasoning Outputs. Model-generated Arabic reasoning traces refined through human-in-the-loop correction for logical and linguistic accuracy.

Figure 6: Example samples from ARB benchmark illustrating diverse domains, each showing the visual input, question, answer, and corresponding step-by-step reasoning traces.

(VQA) tasks; the plan-and-solve framework (Wang et al., 2023) for spatial and remote-sensing inference; and a synthetic backward–forward prompting framework (Shao et al., 2023) for generating Arabic question-answering (QA) pairs with coherent reasoning traces.

3.3. Data Filtering and Verification Process

To ensure ARB's integrity and quality, we implemented a multi-layered filtering and verification pipeline, combining manual review, AI–human refinement, and annotator validation adapted to data complexity and source.

Manual Review and Targeted Corrections: In the initial review phase, the annotators directly corrected minor issues such as typos, grammar errors, or subtle translation inconsistencies (see Figure 4). This approach was especially effective for Category 1, where the translated content from English required adjustments rather than full regeneration. To support this workflow, we developed a custom annotation interface using Excel VBA for efficient review (see Figure 11).

Iterative Human–AI Refinement: For all other categories, we adopted a semi-automated human-in-the-loop framework. GPT-4o generated step-by-step reasoning, which was then reviewed by native speakers and domain experts for logical consistency, linguistic clarity, and cultural alignment . When errors were found, such as unclear steps or reasoning gaps, the annotators provided targeted feedback, prompting partial regeneration or manual edits. This loop continued until each item met the desired quality standard (see Figure 5). A second VBA interface was provided to the annotators to check, rate, flag and finalize items efficiently (see Figure 12).

Quality Filtering and Cultural Alignment: All question–answer–reasoning samples were evaluated for accuracy, coherence, completeness, and Arabic fluency. Automated checks verified logical consistency between reasoning and answers, while annotators conducted manual review. About 8% of samples were discarded due to cultural misalignment or shallow reasoning, ensuring that only high-quality and contextually appropriate data were retained.

Final Approval and Integration: Validated samples underwent final formatting and consistency checks before inclusion. Approved entries were standardized and integrated into ARB benchmark, ensuring completeness, logical coherence, and readiness for multimodal reasoning evaluation (see Figure 6).

3.4. ARB Data Statistics

ARB benchmark comprises 2,219 multimodal samples distributed over 11 domains as showing in Figure 7. In total, the dataset contains about 8K reasoning steps with their corresponding actions, with an average of 3.78 and a median of 4 steps per sample, with the M&L domain exhibiting the greatest reasoning depth. Overall, 71.02% of samples are originally authored in Arabic, with 56.54% from purely Arabic datasets and 43.46% adapted from English captioning or VQA sources, while 28.98% correspond to translated CoT content. The dataset is broadly balanced across domains, with the largest portions from CDT (15%), M&L (14%), and Soc.Cult. (14%), followed by Sci.R. (12%) and OCR (10%). Domains with limited data availability, such as RS (4%), CVP (4%), and VR (5%), reflect the naturally lower availability of high-quality, Arabic-aligned visual resources in these fields. Most samples follow an MCQ format, though short-answer questions and multi-image inputs are also included to encourage comparative and cross-scene reasoning, further enhancing ARB's multimodal diversity.

4. Evaluation Framework

We employ an integrated evaluation framework combining lexical–semantic similarity, an Arabic LLM-as-Judge for stepwise reasoning assess-

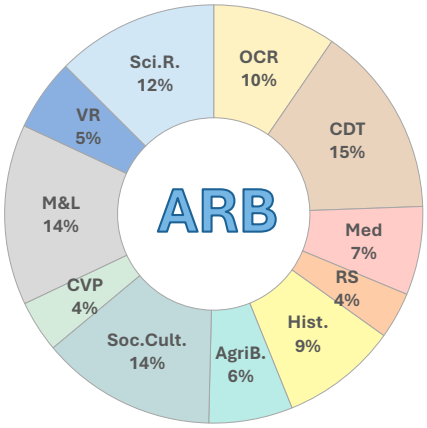


Figure 7: Domain distribution in ARB across 11 domains, showing balanced coverage in CDT, M&L, and Soc.Cult. domains, and fewer samples in resource-scarce domains such as RS, Agri.B, and CVP.

ment, and IAA analysis to validate consistency between humans, and human-model judgments.

Lexical and Semantic Similarity Metrics. We evaluated the alignment between generated reasoning steps and human-curated references using standard similarity metrics (Table 3). BLEU (Papineni et al., 2002) captured n-gram overlap, while ROUGE-1 and ROUGE-L (Lin, 2004) measured fluency and content recall. BERTScore (Zhang et al., 2019) and LaBSE (Feng et al., 2020) assessed token- and sentence-level semantic similarity, offering cross-lingual robustness for Arabic. Normalized Levenshtein Distance (NLD) (Yujian and Bo, 2007) quantified edit-based divergence between predicted and reference reasoning steps. Collectively, these metrics evaluate lexical and semantic similarity but not stepwise logical coherence.

Stepwise Evaluation Using LLM-as-Judge. To address the limitations of classical metrics, we adopted a structured LLM-as-Judge framework with a reference-based Arabic evaluation prompt (Figure 8; English version shown for readability), adapted from (Thawakar et al., 2025). GPT-4.1 served as the judge, assessing reasoning outputs across 10 dimensions—faithfulness (step and token), informativeness (step), repetition (step), redundancy, hallucination, semantic coverage (step), reasoning alignment, commonsense reasoning, and missing step—each rated from 1 to 10. The final quality of the reasoning was calculated as the mean percentage in all dimensions (Table 3 and 4).

IAA: Krippendorff's Alpha. To verify the reliability of human and model-based judgments, we conducted two IAA experiments on 16.18% of the dataset. In the first setting "Humans Only", 3 native Arabic annotators rated reasoning samples on a scale of 1–5 (see Figure 13); in the second setting "Humans + LLM", GPT-4.1 was introduced as a 4th annotator to assess human–model agreement. Krippendorff's-α (Krippendorff, 2018) was used to measure consistency between annotators.

Models Evaluation Prompt

```
You are a reasoning evaluator designed to assess the
alignment, coherence, and quality of reasoning steps
in text responses. Your task is to evaluate reasoning
steps between the *ground truth* and the *LLM
response* using the following metrics:

  1. Faithfulness-Step: Measure how well the reasoning
     steps align with the source sentences.

  2. Faithfulness-Token: Extend Faithfulness-Step by
     token-level alignment within reasoning steps.

  3. Informativeness-Step (Info-Step): Evaluate how
     well the reasoning steps extract relevant
     information from the source.

  4. Repetition-Token: Identify repeated or paraphrased
     reasoning steps within the hypothesis.

  5. Hallucination: Detect irrelevant reasoning steps
     not aligned with the source or reference chain.

  6. Redundancy: Identify redundant reasoning steps
     that are unnecessary for solving the problem.

  7. Semantic Coverage-Step: Evaluate how well the
     hypothesis captures essential elements from the
     source.

  8. Reasoning Alignment: Assess overall overlap and
     alignment between the hypothesis and reference
     chain.
  9. Commonsense: Detect missing commonsense reasoning
     required to solve the problem.

  10.Missing Step: Identify missing reasoning steps
     necessary to solve the problem.

Must give score between (1-10)

Output Format:
Provide your evaluation as follows (only give scores
not explanation.):
      - Metric Scores:
      - **Overall Score:
```

Figure 8: A translated version of the Arabic Evaluation Prompt for LLM-as- Judge. This prompt was used to evaluate reasoning steps across all models. (For original Arabic prompt see Figure 15 in Appendix A.3)

5. Results, Analysis, and Insights

Lexical, Semantic, and Reasoning Consistency. Table 3 provides a unified comparison of models across lexical, semantic, and reasoning metrics. Closed-source models show stronger overall alignment, with Gemini 2.5 Pro achieving the highest LaBSE (73.44%), ROUGE-1 (64.65%), and high LLM-Judge (9.01), reflecting superior fluency and reasoning fidelity. GPT-5 slightly exceeds it in coherence (9.11), confirming its stepwise robustness. Among open models, Gemma 3-27B attains the best BLEU (8.71) and lowest NLD (69.1), while Aya-Vision-8B yields the highest BERTScore (83.66) but lower logical cohesion. Overall, closed models maintain

	Model	BLEU↑	ROUGE-1↑	ROUGE-L↑	NLD$_{error}$ (%)↓	BERTScore↑	LaBSE↑	LLM-Judge↑
Closed-source	GPT-4o	6.95	40.69	28.21	69.3	74.00	68.65	8.14
	GPT-4o-mini	2.11	26.36	19.46	74.4	71.49	57.33	6.68
	GPT-4.1	7.15	47.70	31.07	70.1	73.30	69.83	8.61
	GPT-5	5.47	45.25	29.13	71.4	72.16	68.75	**9.11**
	o4-mini	3.82	35.86	24.27	70.9	71.28	64.59	8.55
	Gemini 2.0 Flash	5.50	35.35	25.31	71.9	72.91	63.64	7.77
	Gemini 2.5 Flash	5.11	63.79	42.93	73.0	72.68	72.33	8.96
	Gemini 2.5 Pro	5.04	**64.65**	42.58	72.2	73.08	**73.44**	9.01
Open-source	Gemma3-27B	**8.71**	52.07	34.80	**69.1**	73.89	71.23	8.24
	Qwen2.5-VL-7B	1.88	26.37	19.56	74.8	69.72	52.45	5.47
	Qwen3-VL-8B	4.04	33.84	23.48	71.8	72.45	63.20	7.25
	LLaMA-4 Scout	7.39	39.76	27.59	70.4	73.07	65.33	7.75
	Aya-Vision-8B	1.20	58.12	**55.78**	91.0	**83.66**	27.04	5.58
	InternVL3-8B	1.89	57.03	54.88	91.0	83.19	30.88	5.47
	AIN (MBZUAI)	1.32	22.50	17.37	75.3	66.49	45.69	4.84

Table 3: Comprehensive evaluation of reasoning quality across models using lexical, semantic, and LLM-as-Judge metrics. BLEU, ROUGE, and NLD assess surface-level similarity; BERTScore and LaBSE measure semantic alignment; and LLM-as-Judge reflects reasoning coherence and step-level fidelity. **Bold** indicates the best score per metric.

Closed-source	**GPT-4o**	**GPT-4o -mini**	**GPT-4.1**	**GPT-5**	**o4 -mini**	**Gemini 2.0 Flash**	**Gemini 2.5 Flash**	**Gemini 2.5 Pro**
Final Answer	67.09	61.27	70.78	**74.88**	71.60	70.20	73.29	73.31
Reason. Steps	81.42	66.82	86.14	**91.14**	85.50	77.69	89.58	90.12
Open-source	**Gemma3 27B**	**Qwen2.5 VL-7B**	**Qwen3 VL-8B**	**LLaMA-4 Scout**	**Aya-Vision-8B**	**InternVL3 -8B**	**AIN MBZUAI**	
Final Answer	64.05	40.23	62.58	**66.19**	47.27	55.07	41.93	
Reason. Steps	**82.41**	54.70	72.53	77.53	55.81	54.71	48.39	

Table 4: Stepwise Evaluation Using LLM-as-Judge (%). Comparison of closed- and open-weight models on final-answer accuracy and aggregated reasoning quality. **Bold** indicates the best score.

	Model	VR	OCR	CDT	CVP	Soc.Cult.	Hist.	Med	M&L	Sci.R.	AgriB.	RS
Closed-source	GPT-4o	79.84	93.08	81.40	63.55	85.11	84.09	73.60	79.41	88.38	81.93	73.22
	GPT-4o-mini	70.71	73.27	68.73	55.41	69.63	69.01	63.87	63.83	72.69	69.01	62.73
	GPT-4.1	82.75	96.67	86.01	74.78	88.55	87.74	81.55	82.67	91.71	84.60	79.37
	GPT-5	**85.71**	**98.07**	**92.74**	**82.52**	**93.07**	**91.49**	**86.70**	**89.22**	**95.34**	**86.04**	**83.73**
	o4-mini	82.93	95.03	87.88	73.28	85.98	85.78	77.91	85.21	91.08	81.77	75.25
	Gemini 2.0 Flash	78.46	91.76	82.50	65.93	79.50	77.73	71.72	77.25	77.40	79.19	70.27
	Gemini 2.5 Flash	86.12	97.77	92.18	77.59	89.55	91.34	82.52	89.04	94.59	86.99	78.76
	Gemini 2.5 Pro	85.49	98.12	92.01	76.53	92.15	91.24	84.25	89.14	94.59	87.41	80.59
Open-source	Aya-Vision-8B	54.80	44.45	63.58	67.70	51.78	53.98	52.66	58.55	51.84	51.04	63.52
	InternVL3-8B	51.72	54.37	59.27	45.26	46.95	48.41	49.02	61.13	58.48	51.72	50.79
	LLaMA-4 Scout	78.46	87.30	80.00	65.57	73.83	82.63	70.00	79.43	85.56	76.93	73.94
	Qwen2.5-VL-7B	60.23	62.38	56.84	48.26	49.70	60.93	49.07	47.96	61.54	58.66	48.51
	Qwen3-VL-8B	76.73	84.50	74.19	58.90	71.88	73.98	65.30	72.40	76.55	71.23	73.73
	Gemma3-27B	**78.46**	**87.30**	**80.00**	**65.57**	**73.83**	**82.63**	**70.00**	**79.43**	**85.56**	**76.93**	**73.94**
	MBZUAI AIN	59.74	47.57	48.24	39.47	44.86	57.20	45.43	43.96	44.71	63.09	60.91

Table 5: LLM-as-Judge per-domain reasoning performance (%) across all domains, reflecting each model's step-level coherence and factual grounding. **Bold** indicates the best score per domain.

stronger cross-lingual and reasoning consistency, whereas open ones show localized precision but variable step alignment, highlighting the need for better adaptation to Arabic linguistic and cognitive patterns.

Reasoning–Answer Performance Gap. Table 4 reveals a consistent gap between reasoning coherence and final-answer accuracy. Among closed models, GPT-5 leads with 91.14% reasoning quality and 74.88% accuracy, followed by Gemini 2.5 Pro/Flash with balanced coherence and factual precision. Open models perform less uniformly: Gemma 3-27B yields the best reasoning (82.41%), and LLaMA-4 Scout ranks highest in answer accuracy (66.19%). Despite moderate reasoning strength, open models' weaker Arabic alignment and multimodal grounding reduce reliability. In general, closed models excel in structured reasoning, while open models show

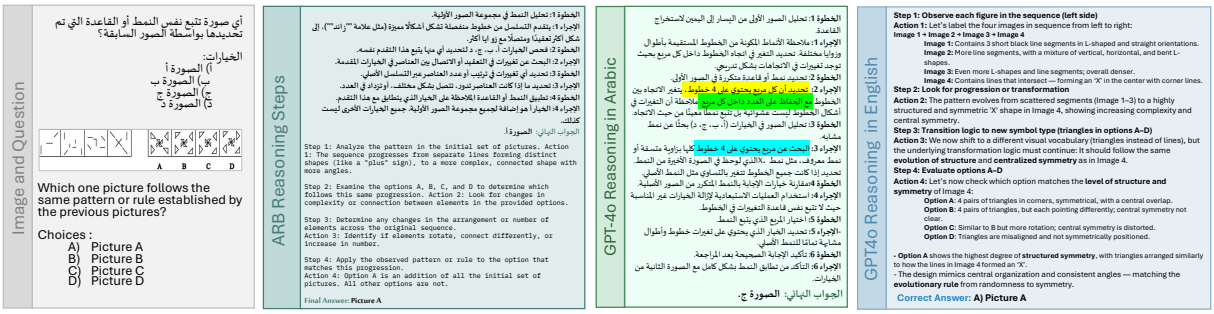


Figure 9: **Cross-Lingual Reasoning Comparison (Arabic vs. English).** This figure compares LMMs (GPT-4o) reasoning steps in Arabic and English for the same visual task. In the Arabic version, the model misinterprets structural constraints, yellow highlights incorrect assumptions about equal line counts across boxes, green emphasizes miscounted lines within the boxes, and cyan marks an irrelevant search for a box with exactly 4 lines. These reasoning flaws lead to the wrong answer (C). In contrast, the English reasoning is structured, accurate, and constraint-aware, correctly identifying the answer (A), highlighting the performance gap in Arabic.

promising but inconsistent progress, highlighting the need for Arabic-specific adaptation.

Qualitative Evaluation. Figures 16 and 17 present representative reasoning failures observed in open- and closed-source models, respectively. These examples highlight common issues such as incomplete or incorrect step transitions, shallow reasoning, and hallucinated content across diverse Arabic multimodal tasks. (see Appendix A.4)

Figure 9 further provides a cross-lingual comparison of GPT-4o reasoning generated in Arabic and English for the same visual task. The Arabic reference reasoning is presented alongside its English translation to support non-Arabic readers. While the English reasoning follows a structured, constraint-aware process that correctly identifies the answer, the Arabic reasoning exhibits several errors, including incorrect assumptions about structural constraints, miscounted line patterns within the boxes, and irrelevant reasoning steps. These issues ultimately lead to an incorrect final answer, highlighting disparities in reasoning reliability across languages and underscoring the importance of Arabic-centric evaluation benchmarks such as ARB.

IAA and Domain Consistency. The IAA results (Figure 10) confirm high agreement across domains, validating the reliability of human and LLM-augmented assessments. The strongest alignment appears in CDT (88.58%) and Sci.R. (83.79%), where tasks are visually grounded and less ambiguous, while RS (57.15%) and Soc.Cult. (65.70%) show lower consistency due to interpretive complexity. In some domains (e.g., Sci.R.), GPT-4.1 aligns closely with human scores, whereas others (Hist., AgriB, CVP) show stronger human agreement. The minimal gap between

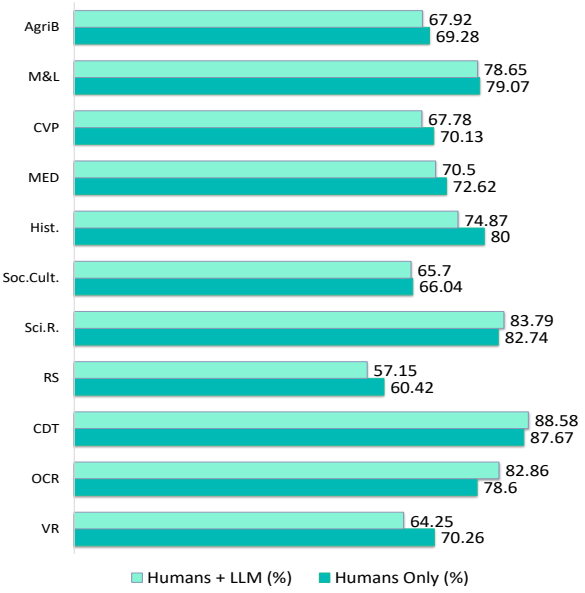


Figure 10: Krippendorff's α (%) for IAA across 11 domains. Results are shown for human-only and human–LLM evaluations.

human-only (76.45%) and human–LLM (76.00%) scores confirms the reliability of GPT-4.1 as a 4th annotator in Arabic reasoning evaluation (see Appendix A.1).

Per-Domain Reasoning Analysis. Per-domain results in Table 5 reveal clear variations in reasoning performance across the 11 ARB domains. Closed-source models, particularly GPT-5 and Gemini-2.5 Pro, maintain consistently high accuracy and coherence across both structured and visually grounded tasks such as SciR. and CDT. In contrast, all models exhibit lower stability in culturally interpretive and specialized domains like Soc.Cult., RS, and AgriB., reflecting the nuanced reasoning and domain-specific vocabulary

required. These variations also reflect ARB's hybrid design, combining native Arabic and adapted multimodal samples, which challenges models to varying degrees across visual, textual, and cultural reasoning dimensions.

In summary, the size of the model, the origin of training, and the composition of the dataset jointly determine the fidelity of the reasoning. Closed models achieve higher coherence and factual grounding, while open models, though smaller and linguistically constrained, show steady progress toward culturally aware Arabic multimodal reasoning.

6. Conclusion

We introduced **ARB**, the first benchmark for step-by-step multimodal reasoning in Arabic across 11 domains, featuring 2.2K high-quality samples and over 8K human-curated reasoning steps. Built through a hybrid pipeline of prompting, tool-assisted generation, and native-speaker validation, ARB enables fine-grained evaluation of both open- and closed-weight models. Our analysis of 15 state-of-the-art LMMs revealed persistent gaps in coherence, reasoning quality, and cultural alignment when reasoning in Arabic, underscoring the need for step-level, culturally grounded evaluation for underrepresented languages. Beyond benchmarking, ARB provides open-source tools and protocols supporting reproducibility and future research, establishing a foundation for training Arabic-native LMMs and advancing inclusive, interpretable AI.

7. Limitations and Societal Impact

While ARB provides a valuable resource for evaluating Arabic multimodal reasoning, it has certain limitations. First, although it spans 11 diverse domains, the benchmark may not fully capture the linguistic, dialectal, and cultural diversity present across the Arabic-speaking world. Additionally, reasoning evaluations rely on human judgments and model-specific prompts, which may introduce a degree of subjectivity or prompt-induced bias. ARB also focuses primarily on static multimodal tasks and does not currently cover domains such as code explanation or video understanding. These areas were excluded mainly due to the scarcity of high-quality Arabic datasets suitable for structured reasoning evaluation, especially for modalities involving temporal information such as video. Finally, the benchmark is designed specifically for Arabic and does not include multilingual alignment or cross-lingual transfer settings, which could be valuable directions for future comparative studies.

From a societal perspective, ARB promotes more inclusive and culturally aware AI by centering Arabic, an underrepresented yet widely spoken language. Its focus on interpretable, step-by-step reasoning supports broader goals of AI transparency and accountability. Nonetheless, ethical considerations remain important, particularly to prevent the misuse or misinterpretation of culturally sensitive content in applications where AI decisions may have real-world consequences.

8. Bibliographical References

Mohammed Al-Maghrabi Allam-Research. 2025. Allam-thinking: Arabic large language model with enhanced reasoning capabilities. https://huggingface.co/almaghrabima/ALLaM-Thinking.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024a. Measuring and improving chain-of-thought reasoning in vision-language models. In *NAACL-HLT*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cohere-Labs. 2025. Aya vision 8b: A multilingual vision-language model. https://huggingface.co/CohereForAI/aya-vision-8b. Accessed: 2025-05-03.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google DeepMind. 2024. Gemini 2.0 flash thinking: Unlocking transparent reasoning in ai. https://deepmind.google/technologies/gemini/flash-thinking/. Accessed: 2025-05-03.

Google DeepMind and Google AI. 2025. Gemini 2.5 flash.

Fanar-Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri

Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. Ain: The arabic inclusive large multimodal model. *arXiv preprint arXiv:2502.00094*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. 2025a. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. 2025b. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*.

Meta-AI. 2025. Llama-4-scout-17b-16e-instruct. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct. Accessed: 2025-05-03.

Ahmad Mustapha, Hadi Al-Khansa, Hadi Al-Mubasher, Aya Mourad, Ranam Hamoud, Hasan El-Husseini, Marwah Al-Sakkaf, and Mariette Awad. 2024. Arastem: A native arabic multiple choice question benchmark for evaluating llms knowledge in stem subjects. *arXiv preprint arXiv:2501.00559*.

OpenAI. 2024a. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-05-03.

OpenAI. 2024b. Gpt-4o system card.

OpenAI. 2025a. GPT-5 model release. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-10-24.

OpenAI. 2025b. Introducing gpt-4.1 in the api. Accessed: 2025-05-03.

OpenAI. 2025c. Openai o3 and o4-mini system card. Accessed: 2025-05-03.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pinterest. 2025. Pinterest platform. https://www.pinterest.com/.

Qwen-Team. 2025. Qwen2.5-vl. Accessed: 2025-05-03.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: generating chain-of-thought demonstrations for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 30706–30775.

Qwen Team. 2025. Qwen3 technical report.

Team-Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.

Emily Vaillancourt and Christopher Thompson. 2024. Instruction tuning on large language models to improve reasoning performance. *Authorea Preprints*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Language Resource References

Alwajih, Fakhraddin and Nagoudi, El Moatez Billah and Bhatia, Gagan and Mohamed, Abdelrahman and Abdul-Mageed, Muhammad. 2024. *Peacock: A Family of Arabic Multimodal Large Language Models and Benchmarks*. arXiv.

Bisong, Ekaba and Bisong, Ekaba. 2019. *Matplotlib and seaborn*. Springer.

Das, Rocktim Jyoti and Hristov, Simeon Emilov and Li, Haonan and Dimitrov, Dimitar and Koychev, Ivan and Nakov, Preslav. 2024. *EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models*. Association for Computational Linguistics.

Jevgenij Gamper and Nasir Rajpoot. 2021. *Multiple Instance Captioning: Learning Representations from Histopathology Textbooks and Articles*. arXiv.

Ghaboura, Sara and Heakl, Ahmed and Thawakar, Omkar and Alharthi, Ali Husain Salem Abdulla and Riahi, Ines and Radman, Abduljalil and Laaksonen, Jorma and Khan, Fahad Shahbaz and Khan, Salman and Anwer, Rao Muhammad. 2025a. *CAMEL-Bench: A Comprehensive Arabic LMM Benchmark*. Association for Computational Linguistics.

Ghaboura, Sara and More, Ketan Pravin and Thawkar, Ritesh and Ghallabi, Wafa Al and Thawakar, Omkar and Khan, Fahad Shahbaz and Cholakkal, Hisham and Khan, Salman and Anwer, Rao Muhammad. 2025b. *Time Travel: A Comprehensive Benchmark to Evaluate LMMs on Historical and Cultural Artifacts*. Association for Computational Linguistics.

He, Xuehai and Zhang, Yichen and Mou, Luntian and Xing, Eric and Xie, Pengtao. 2020. *Pathvqa: 30000+ questions for medical visual question answering*. arXiv.

Kadaoui, Karima and Atwany, Hanin and Al-Ali, Hamdan and Mohamed, Abdelrahman and Mekky, Ali and Tilga, Sergei and Fedorova, Natalia and Artemova, Ekaterina and Aldarmaki, Hanan and Kementchedjhieva, Yova. 2025. *JEEM: Vision-Language Understanding in Four Arabic Dialects*. arXiv.

Kembhavi, Aniruddha and Salvato, Mike and Kolve, Eric and Seo, Minjoon and Hajishirzi, Hannaneh and Farhadi, Ali. 2016. *A diagram is worth a dozen images*. Springer.

Microsoft Corporation. 2024. *Microsoft Excel*. Microsoft Corporation. Version 16.0.

Mousi, Basel and Durrani, Nadir and Ahmad, Fatema and Hasan, Md Arid and Hasanain, Maram and Kabbani, Tameem and Dalvi, Fahim and Chowdhury, Shammur Absar and Alam,

Firoj. 2024. *Aradice: Benchmarks for dialectal and cultural capabilities in llms*. arXiv.

Nawaz, Umair and Muhammad, Awais and Gani, Hanan and Naseer, Muzammal and Khan, Fahad Shahbaz and Khan, Salman and Anwer, Rao. 2025. *AgriCLIP: Adapting CLIP for Agriculture and Livestock via Domain-Specialized Cross-Model Alignment*. Association for Computational Linguistics.

Noman, Mubashir and Ahsan, Noor and Naseer, Muzammal and Cholakkal, Hisham and Anwer, Rao Muhammad and Khan, Salman H and Khan, Fahad Shahbaz. 2024. *CDChat: A Large Multimodal Model for Remote Sensing Change Description*.

Python Software Foundation. 2024. *Python Language Reference, version 3.12*.

Sadallah, Abdelrahman and Tonga, Junior Cedric and Almubarak, Khalid and Almheiri, Saeed and Atif, Farah and Qwaider, Chatrine and Kadaoui, Karima and Shatnawi, Sara and Alesh, Yaser and Koto, Fajri. 2025. *Commonsense Reasoning in Arab Culture*. Association for Computational Linguistics.

scottgeng00. 2025. *NaBirds*. [Accessed 10-10-2025].

A. Appendix

This appendix provides supplementary material supporting our contributions. It includes: (1) details of the filtering and verification pipeline and annotation interfaces used for human-in-the-loop validation and inter-annotator agreement; (2) prompts used for reasoning generation and evaluation; (3) the original Arabic versions of the generation prompts and evaluation metrics; and (4) additional qualitative examples from open- and closed-source models. These materials provide greater transparency into the construction and quality control of the ARB benchmark.

A.1. Manual Verification Pipeline and Annotation Interface

To ensure quality and consistency across all samples, we developed a streamlined and user-friendly annotation interface to support manual verification and scoring. Given the scale of the dataset and the involvement of multiple annotators, the interface was designed to simplify inspection and accelerate the review process.

For translation tasks (Figure 11), the interface displays the original English text alongside the corresponding Arabic translation, allowing annotators to edit only the translated portion when necessary. For synthetic samples (Figure 12), the interface presents the image, the Arabic question, the step-by-step reasoning, the predicted answer, and the reference answer. Annotators evaluate each sample based on accuracy, clarity, cultural alignment, and faithful delivery of meaning, with emphasis placed on conceptual correctness rather than word-for-word translation.

Each sample is rated on a six-point scale, summarized in Table 6. The scoring scheme guides annotators in determining whether a sample should be accepted, revised, or regenerated.

Rate	Description
0	Reject: Culturally inappropriate or irrelevant content
1	Reject: Requires full regeneration by the model
2	Poor: Major edits needed to fix reasoning or clarity
3	Fair: Moderate edits required
4	Good: Minor edits needed
5	Excellent: No edits needed; ready for inclusion

Table 6: **Filtering and Verification Rating Scale.** A standardized scoring scheme used by annotators to assess the quality of translations and reasoning steps.

Each sample was independently reviewed by four annotators, and their scores were averaged to obtain a final rating on a scale of 0–5. Samples with an average score equal to 0 were immediately discarded due to cultural or contextual inappropriateness. Samples with an average score of 5 were approved without further review, while those with an average score of 1 were returned for regeneration. Samples with intermediate scores (greater than 1 and less than 5) were escalated to a fifth annotator for final assessment, discrepancy resolution, and necessary corrections.

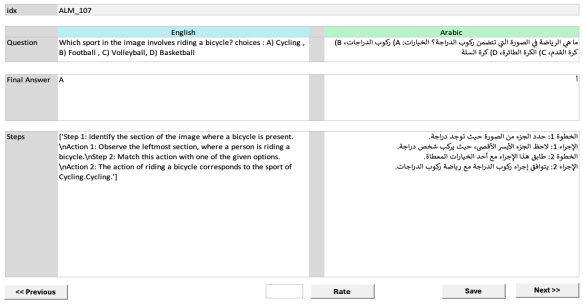


Figure 11: Example of ARB translation verification user interface.

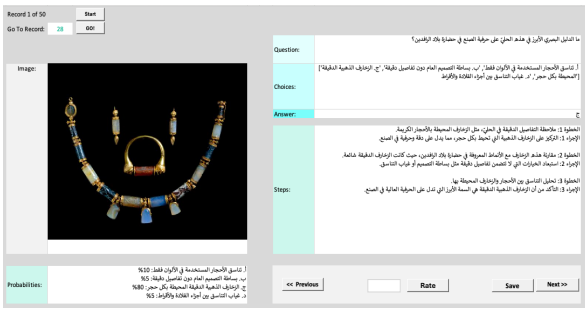


Figure 12: Example of ARB generated data verification user interface.

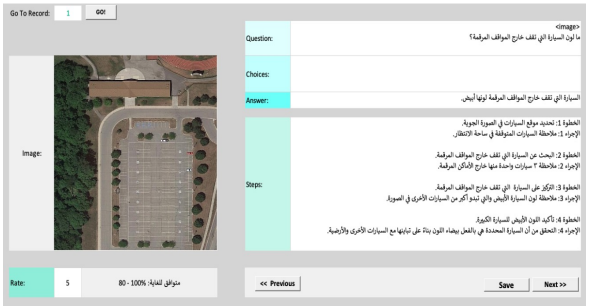


Figure 13: IAA interface used for human validation of reasoning steps.

This multi-tiered evaluation process ensured both the consistency and overall quality of the final dataset.

Figure 13 illustrates the interface used for the IAA study. Annotators evaluated whether the model's reasoning chain was consistent with the image, question, and available answer choices, assigning scores (from 0 to 5) based on the alignment between the model's reasoning steps and their own reasoning process.

A.2. Original Arabic Generation Prompt

This section presents the original Arabic generation prompt used to produce reasoning steps in ARB, shown in Figure 14. An English translation of this prompt is provided in the main paper (Figure 3) for accessibility to non-Arabic readers.

Reasoning Steps Generation Prompt

أنت خبير محترف متخصص في {Domain} مهمتك توليد خطوات التحليل المنطقي وخطوات الاستدلال للبيانات وللأسئلة النصية والبصرية مع الإجراء اللازم لكل خطوة للوصول إلى الجواب الصحيح استناداً إلى القرائن البصرية في الصورة والمعلومات في السؤال والاختيارات المتوفرة، مع الاسترشاد بالمثال{example} التالي كنمط للهيكل المستخدم في توليد خطوات التحليل والإجراءات التابعة لها. استخدم التعليمات التالية:

1. اقرأ بتمعن السؤال والخيارات المتوفرة – إن وجدت.
2. حدد المفاهيم الأساسية للموضوع {Domain} والمهارات والمعرفة المطلوبة.
3. الأسئلة متنوعة وعليك اتباع منهج {Curriculum} محدد لكل موضوع {Domain}.
4. تقع المناهج {Curriculum} ضمن أربع الفئات:

- **الفئة الأولى - {Curriculum} = "حسابي"**: يجب عليك استخدام العمليات الحسابية الأساسية والعمليات الحسابية النسبية والمنطق الرياضي
- **الفئة الثانية - {Curriculum} = "علمي/طبي"**: عليك استخدام المنطق والقواعد العلمية لكل مجال تخصصي.
- **الفئة الثالثة - {Curriculum} = "نصي/جزئي"** من الصورة: عليك التركيز على تجزيء الصورة والتفحص.
- **الفئة الرابعة - {Curriculum} = "عامي"**: عليك استخدام المقارنة والمفارقة وما يفرضه السؤال للوصول إلى الإجابة الصحيحة.

يرجى إخراج الملف وفقًا للصيغة المحددة{example}؛ وحدد الجواب النهائي من خلال "الجواب هو:_____".

Figure 14: **Original Arabic Generation Prompt.** The original Arabic version of the prompt used to generate reasoning steps in ARB (see Figure 3) to aid non-Arabic readers.

A.3. Models' Evaluation Prompts

This section presents the evaluation prompts used to assess the step-by-step reasoning quality of LMMs in our study. The prompt was adapted from the LLamaV-o1 evaluation protocol (Thawakar et al., 2025) and tailored to the Arabic multimodal reasoning context of ARB (Figure 15). To ensure consistency between the generation and evaluation phases, all assessments were performed using Arabic prompts exclusively in open-source and closed-source models. This design choice maintained linguistic alignment with model outputs and minimized potential cross-lingual biases during judgment.

Evaluation Prompt

أنت مُقيِّم للاستدلال مصمم لتقييم مدى التوافق والتماسك وجودة خطوات الاستدلال في الاستجابات النصية. مهمتك هي تقييم خطوات الاستدلال بين الجواب المرجعي (الحقيقي) للسؤال و استجابة النموذج اللغوي للسوال (أي الجواب الصادر عن النموذج) باستخدام المقاييس التالية:

1.التطابق - الخطوة (Faithfulness-Step): قياس مدى توافق و تطابق ودقة و موثوقية واتساق خطوات الاستدلال مع الجمل المصدرية.

1.التطابق - الرمز (Faithfulness-Token):توسيع مقياس التطابق - الخطوة (التوافق على مستوى الخطوات) عبر التحقق من التوافق و التطابق والدقة و الموثوقية والاتساق على مستوى الرموز داخل خطوات الاستدلال.

2.الإثراء المعلوماتي - الخطوة (Informativeness-Step - Info-Step):تقييم مدى قدرة خطوات الاستدلال على استخراج المعلومات ذات الصلة من المصدر.

3.تكرار - الرمز (Repetition-Token): تحديد الخطوات الاستدلالية المكررة أو المعاد صياغتها داخل الفرضية.

4.الهلوسة (Hallucination): اكتشاف خطوات استدلال غير المرتبطة أو غير المتوافقة مع المصدر أو سلسلة المرجع.

5.التكرار الزائد (Redundancy):تحديد الخطوات الاستدلالية الزائدة وغير الضرورية لحل المشكلة.

6.التغطية الدلالية - الخطوة (Semantic Coverage-Step): تقييم مدى تغطية الفرضية للعناصر الأساسية في المصدر.

7.توافق الاستدلال (Reasoning Alignment): قياس مدى التوافق والارتباط العام بين الفرضية وسلسلة المرجع.

8.المنطق العام (Commonsense):الكشف عن غياب المنطق العام المطلوب لحل المشكلة.

9.الخطوة المفقودة (Missing Step): تحديد خطوات الاستدلال الناقصة والضرورية لحل المشكلة.

يجب أن تعطي درجة بين (1-10)

يرجى إخراج الملف وفقًا للصيغة المحددة:
قدم تقييمك كما يلي (قم بتقديم الدرجات فقط بدون تفسير):
- **درجات المقاييس:**
- **الدرجة الإجمالية:**

Figure 15: **Arabic Evaluation Prompt for LLM-as-Judge.** This prompt was used to evaluate reasoning steps across all models in Arabic. Refer to Figure 8 for English translation.

A.4. Qualitative Examples

Figures 16 and 17 present qualitative examples of reasoning failures in both open- and closed-source models, complementing the quantitative trends discussed in Section 5.

Open-source models frequently produce incomplete reasoning chains, inconsistent logical steps, or hallucinated interpretations that are not grounded in the visual or textual inputs. Closed-source models, while often generating more structured reasoning, still exhibit errors such as incorrect numerical comparisons, invalid assumptions, and misinterpreted constraints. These examples show that seemingly coherent step-by-step reasoning does not necessarily ensure logical correctness, highlighting persistent challenges in achieving reliable multimodal reasoning in Arabic and underscoring the importance of benchmarks like ARB for diagnosing such failures.

Figure 16: **Qualitative reasoning errors in open-source models across Arabic multimodal tasks.** Examples illustrate common failures such as incomplete reasoning chains, inconsistent logic, and hallucinated interpretations that are not grounded in the input.

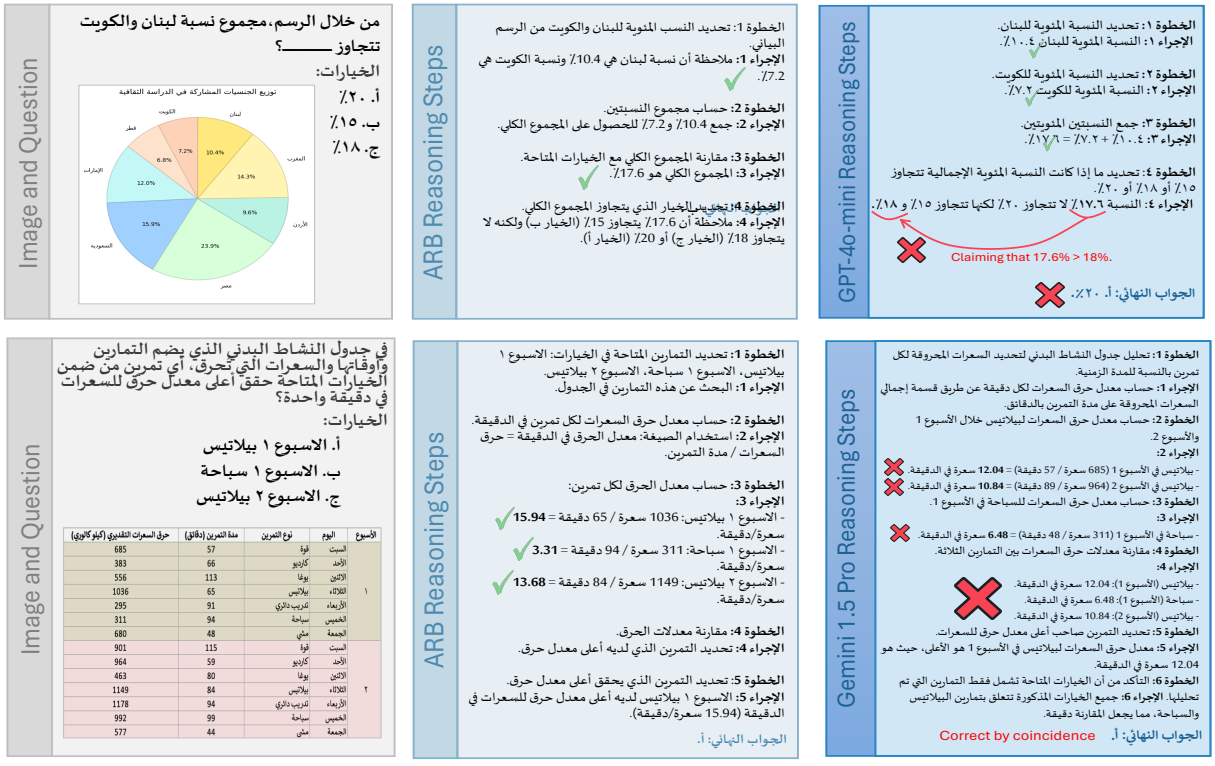


Figure 17: **Qualitative reasoning errors in closed-source models across Arabic multimodal tasks.** Examples highlight issues including incorrect numerical comparisons, invalid assumptions, and logically inconsistent reasoning steps leading to incorrect conclusions.