ARB: A Comprehensive Arabic Multimodal Reasoning Benchmark

Sara Ghaboura[1†] Ketan More[1†] Wafa Alghallabi[1] Omkar Thawakar1[1]
Jorma Laaksonen[3] Hisham Cholakkal[1] Salman Khan[1,2] Rao Muhammad Anwer[1,3]
[1]Mohamed bin Zayed University of AI, [2]Australian National University, [3]Aalto University
{sara.ghaboura, ketan.more}@mbzuai.ac.ae
https://mbzuai-oryx.github.io/ARB/

Abstract

As Large Multimodal Models (LMMs) become more capable, there is growing interest in evaluating their reasoning processes alongside their final outputs. However, most benchmarks remain focused on English, overlooking languages with rich linguistic and cultural contexts, 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 spans 11 diverse domains, including visual reasoning, document understanding, OCR, scientific analysis, and cultural interpretation. It comprises 1,356 multimodal samples paired with 5,119 human-curated reasoning steps and corresponding actions. We evaluated 12 state-of-the-art open- and closed-source LMMs and found persistent challenges in coherence, faithfulness, and cultural grounding. ARB offers a structured framework for diagnosing multimodal reasoning in underrepresented languages and marks a critical step toward inclusive, transparent, and culturally aware AI systems. We release the benchmark[1], rubric, and evaluation suit [2] to support future research and reproducibility.

1 Introduction

Arabic, spoken by more than 400 million people worldwide, embodies significant linguistic diversity and a profound cultural heritage. Despite its widespread usage, Arabic remains notably underrepresented in advanced AI systems, particularly those that involve multimodal reasoning, simultaneous interpretation, and logical processing of textual and visual data crucial for fields such as education, healthcare, and cultural preservation. This scarcity limits the deployment and inclusion of multimodal AI in Arabic-speaking communities.

Recent developments in LMMs reflect a growing emphasis on transparency and interpretability, achieved through explicit reasoning steps. Techniques such as chain-of-thought (CoT) prompting, initially introduced by Wei et al. (2022), encourage models to systematically articulate intermediate reasoning steps, significantly improving both performance and explainability. This paradigm has gained traction in English-based language models and has been effectively extended to multimodal settings in models such as LLaVA-CoT (Xu et al., 2025), VisCoT (Shao et al., 2024), and the recent LLamaV-o1 (Thawakar et al., 2025).

Current step-by-step reasoning benchmarks largely focus on English, overlooking the linguistic nuances and cultural contexts essential to Arabic. Recent work on cross-lingual reasoning (Yong et al., 2025) shows that English-trained models can generalize to other languages via test-time scaling; however, Arabic was not explicitly evaluated, and performance often falters in the presence of linguistic complexity and cultural commonsense. Existing Arabic multimodal data sets, such as CAMEL-Bench (Ghaboura et al., 2025a), Henna (Alwajih et al., 2024), and JEEM (Kadaoui et al., 2025), prioritize final answer accuracy with limited attention to intermediate reasoning. Meanwhile, benchmarks like AraDiCE (Mousi et al., 2024) and ArabCulture (Sadallah et al., 2025) remain confined to textual modalities. Together, these limitations signal the need for Arabic-specific multimodal reasoning benchmarks that reflect the linguistic and cultural demands of the target language.

To bridge this critical chasm, we introduce the Comprehensive Arabic Multimodal Reasoning Benchmark (ARB), the first explicitly designed benchmark for evaluating detailed step-by-step reasoning in Arabic multimodal contexts (Table 1). ARB comprises 1,356 multimodal samples in 11

[1]https://huggingface.co/datasets/MBZUAI/ARB
[2]https://github.com/mbzuai-oryx/ARB
[†]Equal contribution.

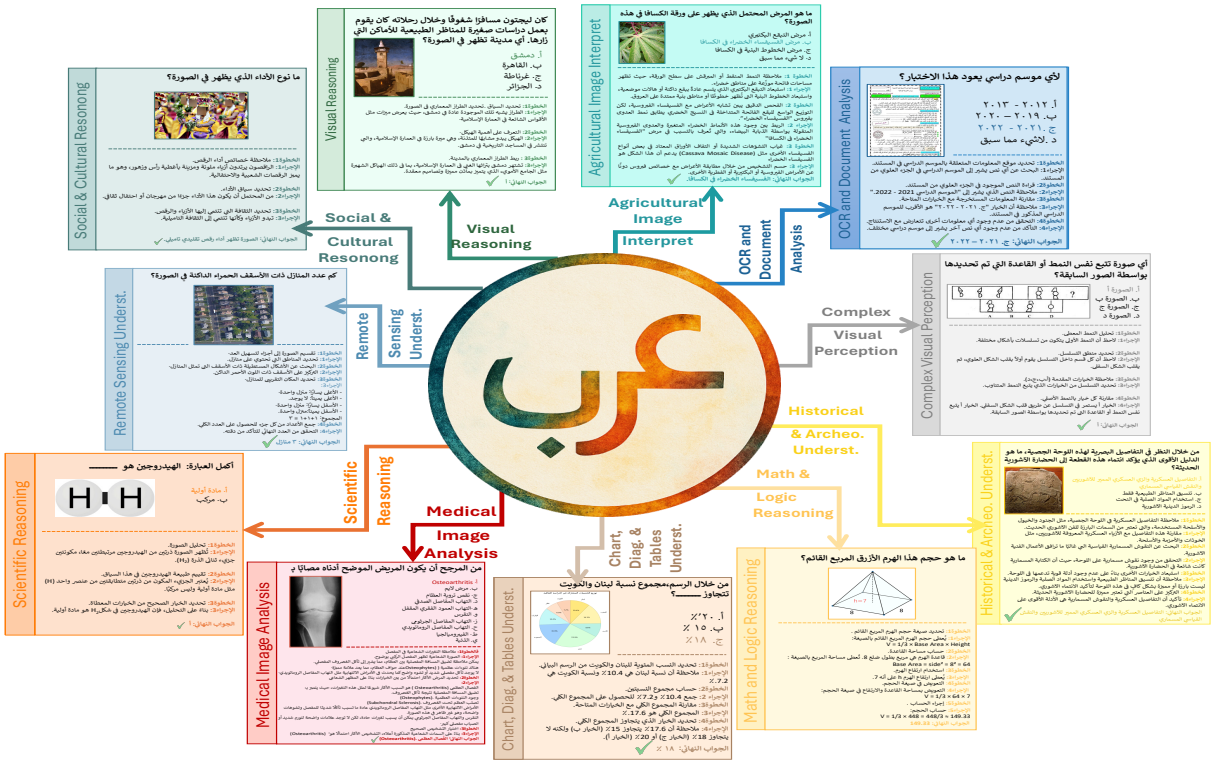


Figure 1: **ARB Dataset Diversity.** ARB comprises a wide array of multimodal reasoning samples, each combining a visual input with an Arabic question and detailed step-by-step reasoning with actions taken by step. The dataset spans 11 distinct domains, including visual reasoning, OCR and document understanding, chart and diagram interpretation, mathematical and logical inference, scientific and medical analysis, cultural and historical interpretation, remote sensing, agricultural image analysis, and complex visual perception—capturing the linguistic richness, cultural depth, and cross-domain complexity essential for evaluating reasoning in Arabic.

Benchmarks	Multi-modal?	Multi-domain?	Reasoning support?	Open-source?	Eval. Level
Henna	✓	✗	✗	✗	FA*
CAMEL-Bench	✓	✓	✗	✓	FA*
AraSTEM	✗	✗	✓	✓	FA*
AraDiCE	✗	✓	✓	✓	FA*
JEEM	✓	✓	✗	✓	FA*
PALM	✗	✓	✗	✗	FA*
ArabCulture	✗	✓	✓	✓	FA*
ARB (ours)	✓	✓	✓	✓	**FA* & Step***

Table 1: **Comparison of our ARB with existing Arabic LMM benchmarks and Reasoning Benchmarks.** Henna (Alwajih et al., 2024), CAMEL-Bench (Ghaboura et al., 2025a), AraSTEM (Mustapha et al., 2024), AraDiCE (Mousi et al., 2024), JEEM (Kadaoui et al., 2025), PALM (Alwajih et al., 2025), ArabCulture (Sadallah et al., 2025). FA*: Final Answer Evaluation. Step*: Step-level Evaluation.

domains, including visual reasoning, document understanding, optical character recognition (OCR), cultural interpretation, medical imaging, and remote sensing (Figure 1). Each sample includes meticulously curated annotations with more than 5.1k reasoning steps, each paired with a specific action, allowing nuanced assessment of coherence, fidelity, and cultural grounding beyond mere final-answer accuracy.

The construction of ARB involved systematic identification of critical reasoning domains and rigorous data sourcing, validated by domain experts. All annotations, reasoning chains, and actions were verified by native speakers through a human-in-the-loop process to ensure logical precision and cultural fidelity. We also performed a human evaluation to assess the correctness of the reasoning steps and to validate the reliability of using LLMs as automated judges.

Evaluations of 12 prominent open-source and closed-source LMMs - including GPT-4V (OpenAI, 2024b,a, 2025a,b), Gemini variants (Gemini Team, 2024; DeepMind, 2024), and open-source multilingual models such as Qwen2.5-VL (Qwen-Team, 2025), LlaMA variants (Meta-AI, 2024, 2025), Aya-Vision (Cohere-Labs, 2025), InternVL3 (Chen et al., 2024b), and Arabic-focused AIN (Heakl et al., 2025) - highlight significant deficiencies in Arabic reasoning coherence and cultural grounding despite robust English performance, underscoring the necessity of ARB.

In summary, (1) we introduce ARB, the first Arabic-centric benchmark designed to evaluate step-by-step multimodal reasoning across 11 cul-

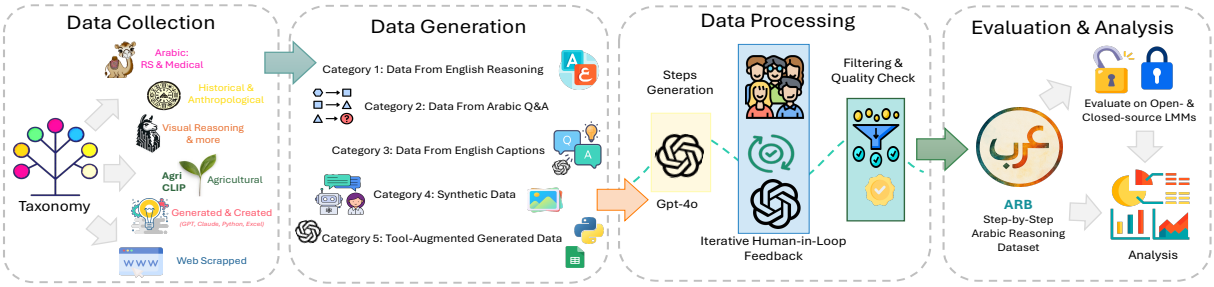


Figure 2: **The ARB Dataset Pipeline.** The figure illustrates the ARB pipeline for evaluating Arabic multimodal reasoning in LMMs. It begins with data collection across 11 domains—such as medical imaging, historical interpretation, visual reasoning, and agriculture—sourced from curated datasets (e.g., VRC-Bench, CAMEL-Bench), synthetic content, tool-augmented outputs, and web scraping. Data is generated across five categories: English reasoning chains, Arabic Q&A, English captions, synthetic samples, and tool-enhanced content. Reasoning steps are refined via human-in-the-loop feedback and filtered for logical consistency and cultural alignment. The benchmark supports fine-grained evaluation of open- and closed-source models on Arabic step-by-step reasoning.

turally and linguistically grounded domains; (2) we conduct extensive evaluations of 12 leading open- and closed-source LMMs, uncovering limitations in coherence, faithfulness, and reasoning quality in Arabic; (3) we integrate a human-in-the-loop pipeline with manual verification by native speakers and domain experts to ensure annotation accuracy; and (4) we perform human evaluations to validate reasoning correctness and assess the effectiveness of LLM-as-a-judge scoring.

2 Step-by-Step Arabic Reasoning Benchmark: ARB

Figure 2 presents an overview of the ARB data construction pipeline, which we describe in detail through the following subsections.

2.1 Data Collection

We adopt a domain-guided approach to curate data across a broad spectrum of categories relevant to Arabic multimodal reasoning. This ensures diversity in both content and modality, encompassing textual and visual tasks. The selected domains (Figure 1)—from visual perception to historical and anthropological interpretation—are sourced from existing benchmarks, human-authored questions, and synthetic content (Table 2). These sources were selected to capture diverse reasoning challenges and promote linguistic, cognitive, and cultural variety across the dataset.

2.2 Data Generation and Data Processing

We generated the dataset content in five main categories, each targeting a different source or creation method (Figure 3). For each category, we employed a strategically selected prompting technique and engaged human experts to iteratively review and refine the resulting reasoning steps.

Domains	English Bench	Arabic Bench	Human Created	Synthetic
Visual Reasoning	✓	–	–	–
OCR & Docs Anal.	–	–	✓	✓
CDT	✓	✓	✓	✓
Math &logic	✓	–	–	–
Social & Cult.	✓	–	–	–
Comp. Vis. Percept.	✓	–	–	–
Medica Img. Anal.	✓	✓	–	–
Scientific Reasoning	✓	–	–	–
Agricultural Interp.	✓	–	✓	✓
Remote Sensing Und.	–	✓	–	–
Histo. & Anthro.	✓	–	✓	✓

Table 2: **Source Types Across ARB Domains.** We show the sources for each of the 11 domains, indicating whether data originated from Arabic or English benchmarks, human-written questions, or synthetic content, highlighting the dataset's linguistic and cognitive diversity. **CDT:** Chart, Diagrams, & Table Understanding; **Social & Cult.:** Social & Cultural Reasoning; **Complex Vis. Percept.**: Complex Visual Perception; **Agricultural Interp.:** Agricultural Image Interpretation; **Histo. & Anthro.**: Historical & Anthropological Understanding.

Category 1: English Reasoning Benchmarks
We adapted the English step-by-step reasoning dataset VRC-Bench (Thawakar et al., 2025) by excluding domains with limited Arabic relevance (e.g., OCR, Charts, Diagrams & Tables). The remaining content was translated into Arabic using GPT-4o and reviewed by native speakers for step-level accuracy, coherence, and fluency. Particular attention was given to resolving translation challenges such as singular–plural and subject–verb agreement, sentence structure differences, and non-literal expressions. Figurative language and cultural references were carefully localized to preserve contextual relevance, meaning complexity, and naturalness in Arabic.

Category 2: Arabic QA Benchmarks
To further enrich the ARB collection, we incorporate two specialized domains, medical image anal-

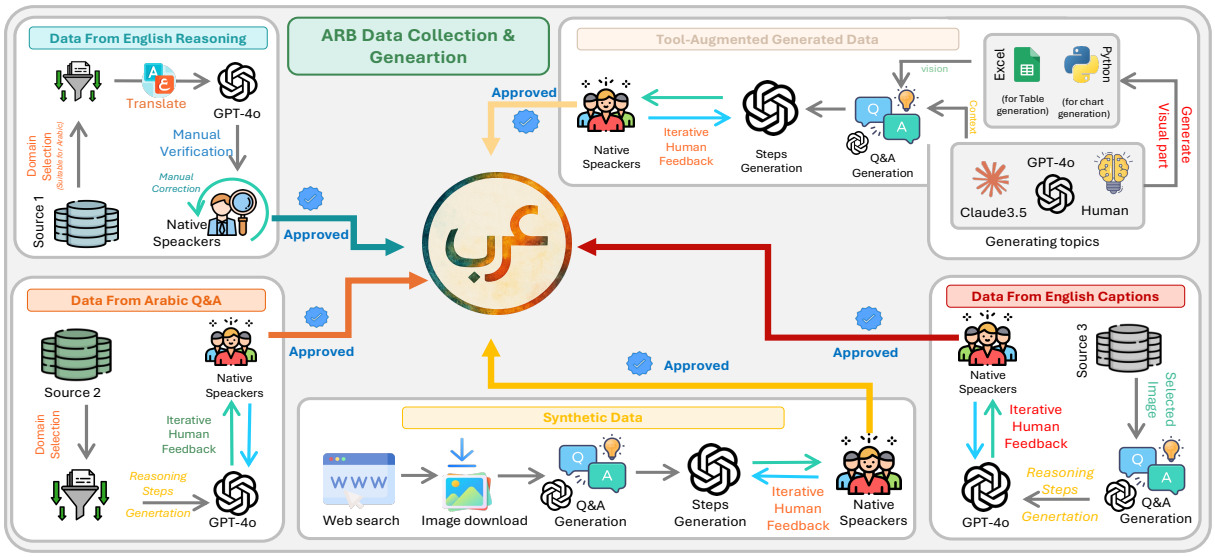


Figure 3: **Overview of the ARB Data Collection, Generation and Verification Framework.** The ARB benchmark is constructed from five primary data sources: (1) English reasoning benchmarks, (2) Arabic question–answer benchmarks, (3) English-captioned datasets, (4) Synthetic data, and (5) Tool-augmented data. All data undergoes iterative refinement through human-in-the-loop feedback and validation by native Arabic speakers to ensure cultural and linguistic fidelity.

ysis and remote sensing understanding, sourced from the CAMEL-Bench (Ghaboura et al., 2025a). For each QA pair, we generated detailed step-by-step reasoning traces to support interpretability and structured inference using GPT-4o. For the medical domain, we employed a **few-shot CoT prompting** strategy to produce coherent reasoning chains. However, this approach proved insufficient for the remote sensing domain, where questions often require spatial decomposition and complex visual inference. To address this, we adopted the **plan-and-solve prompting** framework (Wang et al., 2023), guiding the model to divide images into segments (e.g., quadrants or longitudinal zones) and apply a structured, divide-and-conquer reasoning approach. This significantly improved the fidelity and completeness of reasoning in the remote sensing domain.

Category 3: English Caption Benchmarks As an additional expansion of the ARB, we integrated two new domains—agricultural image interpretation and historical & archaeological understanding—using visual content and captions sourced from AgriCLIP (Nawaz et al., 2025) and TimeTravel (Ghaboura et al., 2025b), respectively. To generate Arabic reasoning questions with corresponding step-by-step answers, we adopted the **synthetic prompting** like framework inspired by (Shao et al., 2023) implemented using GPT-4o. This approach followed a backward–forward generation strategy; the model first synthesized a plausible reasoning chain (backward step), then generated a question that would logically yield that reasoning. In the forward step, the model refined the reasoning trace to ensure alignment and internal consistency. To ensure data quality and reasoning diversity, we applied a complexity-based selection criterion that prioritized samples involving multi-step inference or higher-order reasoning. This pipeline enabled scalable generation of rich, inference-oriented Arabic QA pairs without requiring exhaustive manual annotation.

Category 4: Synthetic Data For the OCR and Document Analysis domain, we curated a set of web-sourced images containing textual content from publicly available sources (Pinterest, 2025). Each image was processed using GPT-4o, which was prompted to generate Arabic QA pairs along with corresponding step-by-step reasoning. To guide the generation process, we employed a **few-shot CoT prompting** strategy, encouraging the model to produce inference-driven reasoning chains grounded in both visual and textual cues present in the images.

Category 5: Tool-augmented Generated Data In this category, we constructed the domain of Charts, Diagrams, and Tables by integrating external tools to create visual samples. For the charts subdomain, data was derived from both human-curated topics and synthetic scenarios

using GPT-4o under human guidance, with visualizations produced via Python and Matplotlib (Bisong and Bisong, 2019). The tables subdomain involved generating structured data using GPT-4o and Claude-3.5 (Anthropic, 20254), based on human-defined themes, and visualized in Excel to simulate realistic interpretation tasks. For diagrams, we adapted a subset of the AI2D dataset (Kembhavi et al., 2016), translating and extensively editing the content into Arabic. Human annotators refined the corresponding questions to prioritize reasoning over factual recall. Across all subdomains, GPT-4o was prompted using a **few-shot CoT** strategy to generate Arabic QA pairs with explicit step-by-step reasoning.

2.3 Data Filtering and Verification Process

To ensure the integrity and quality of ARB, we implemented a multi-stage filtering and verification pipeline (Figure 3). This process combined manual correction, semi-automated AI–human refinement, and native speaker validation, each tailored to the complexity and origin of the data.

Manual Review and Targeted Corrections:
In the initial review phase, human annotators—primarily native Arabic speakers—directly corrected minor issues such as typos, grammar errors, or subtle translation inconsistencies. This approach was especially effective for Category 1, where translated content from English required adjustments rather than full regeneration. To support this workflow, we developed a custom annotation interface for efficient review (see Figure 7a in Appendix C).

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. A second interface (see Figure 7b, Appendix C) allowed annotators to check, rate, flag, and finalize items efficiently.

Quality Filtering and Cultural Alignment:
Post-refinement, all question–answer–reasoning samples were evaluated against strict quality criteria: accuracy, coherence, reasoning completeness, and Arabic fluency. We applied both automated checks (e.g., verifying the answer aligns with the reasoning steps) and manual review. Over 200 samples were discarded at this stage due to cultural misalignment or insufficient reasoning depth. This filtering step ensured only high-quality, culturally appropriate, and challenging samples were retained.

Final Approval and Integration:
Items that passed all prior checks were subjected to a final review to ensure proper formatting, logical coherence, and internal consistency. Upon approval by native Arabic reviewers, the data was standardized and formally integrated into the ARB benchmark. This final validation step ensured that each entry was complete, well-structured, and suitable for robust evaluation of Arabic multimodal reasoning. Further details on the filtering, verification procedures, and annotation interfaces are provided in the Appendix C.

2.4 ARB Data Statistics

The ARB benchmark consists of 1,356 multimodal samples distributed across 11 domains (Figure 19), with Math & Logic comprising the largest share, followed by Charts, Diagrams, & Tables. Each sample includes an image, an Arabic question, and a set of step–action pairs leading to a final answer. In total, ARB contains 5,119 reasoning steps, with no fixed limit imposed during generation to preserve flexibility based on task complexity. Most samples include 2–6 steps, with an average of 3.78 and a median of 4. The number of steps ranges from 1 to 16, with Math & Logic exhibiting the highest reasoning depth. Further statistics are presented in Appendix H Figure 20.

3 Evaluation Framework

3.1 Model and Prompt Selection

We selected GPT-4o and GPT-4o-mini as candidate models due to their demonstrated efficiency and effectiveness in multimodal tasks, referring to (Heakl et al., 2025). Recognizing the sensitivity of reasoning performance to prompt language, we evaluated both models using prompts in English and Arabic. A diverse set of 40 samples spanning multiple domains was assessed by three native Arabic speakers. To further support the evaluation of translated outputs, we employed LaBSE (Feng et al., 2020) to

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 4: **ARB Evaluation Prompt.** The standardized Arabic prompt used across all ARB domains to elicit structured, curriculum-based reasoning steps from evaluated models during inference. The English version is provided in Appendix E.

measure semantic similarity between English and Arabic responses.

Human evaluations consistently favored GPT-4o in both prompt settings. When incorporating LaBSE, GPT-4o with Arabic prompts achieved the highest similarity scores. However, across all settings, automated scores remained lower than human judgments, reflecting the models' difficulty in capturing acceptable variations in structure and order. To mitigate this, we adopted a few-shot prompting strategy, which improved similarity scores by 20–30%, while preserving GPT-4o with Arabic prompts as the best performer. Thus, we finalize GPT-4o with Arabic prompts for the generation of reasoning steps (Figure 4).

3.2 Evaluation Methodology and Metrics

Lexical and Semantic Similarity Metrics.

To assess similarity between predicted reasoning steps and human-curated references, we employed standard metrics (Table 4). BLEU (Papineni et al., 2002) showed weak n-gram alignment, while ROUGE variants (Lin, 2004) yielded mixed results with a sharp drop in ROUGE-2, indicating limited fluency. For semantic similarity, we used BERTScore (Zhang et al., 2019), which captures token-level alignment but lacks cross-lingual robustness, reducing its reliability for Arabic evaluation. To address this, we adopted LaBSE (Feng et al., 2020), a multilingual sentence-level model that provided more stable results, averaging 81.5%±2 for closed-weight models and 71.5%±5 for open-weight ones. Despite their utility, these metrics fall short in capturing logical structure, coherence, and factual grounding in multi-step reasoning.

Stepwise Evaluation Using LLM-as-Judge

To address the limitations of traditional evaluation metrics, we adopted a structured LLM-as-Judge framework, along with a reference-based protocol and Arabic prompt, adapted from (Thawakar et al., 2025) evaluation suite. Unlike reference-free metrics (Golovneva et al., 2022), this set-up enables a fine-grained, interpretable evaluation aligned with Arabic linguistic and contextual nuances. GPT-4o, used as LLM-as-Judge, is instructed to assess reasoning outputs across several dimensions, including faithfulness, informativeness, redundancy, hallucination, semantic coverage, and commonsense reasoning. Each attribute is rated on a scale from 1 to 10 (see Figure 15 and Figure 16), and the final score for reasoning steps is computed as the average across all dimensions (Table 3). The full evaluation prompt is provided in Appendix D.

Inter-Annotator Agreement: Krippendorff's Alpha. To ensure data quality and validate the efficiency of our LLM-as-Judge selection, we conducted an inter-annotator agreement analysis over 5% of the dataset. Three human annotators were provided with a user-friendly interface (Figure 8) to rate samples on a scale from 1 (lowest) to 5 (highest). Most samples received scores of 4 or higher, confirming the effectiveness of our earlier verification steps and reflecting strong agreement among annotators. We measured Krippendorff's Alpha (Krippendorff, 2018), achieving a score of 83.56% among human annotators. To further assess the reliability of GPT-4o as an LLM-as-Judge, we repeated the evaluation by including the model's judgments, resulting in an even higher Krippendorff's Alpha of 87.62%. These results demonstrate high consistency between human and LLM assessments, supporting the robustness of our evaluation framework.

Closed-source Models	GPT-4o	GPT-4o -mini	GPT-4.1	o4 -mini	Gemini 1.5 Pro	Gemini 2.0 Flash
Final Answer (%)	**60.22**	52.22	59.43	58.93	56.70	57.80
Reasoning Steps (%)	64.29	61.02	80.41	**80.75**	64.34	64.09
Open-source Model	Qwen2.5 VL-7B	Llama-3.2 11B-Vis-Inst.	AIN	Llama-4 Scout (17Bx16E)	Aya-vision-8B	InternVL3 -8B
Final Answer (%)	37.02	25.58	27.35	**48.52**	28.81	31.04
Reasoning Steps (%)	64.03	53.20	52.77	**77.70**	63.64	54.50

Table 3: **Stepwise Evaluation Using LLM-as-Judge.** Comparison of closed- and open-weight models based on final answer accuracy and aggregated quality scores of reasoning steps, using our LLM-as-Judge framework with Arabic prompts and evaluation metrics. The evaluation follows a reference-based, attribute-level protocol for assessing reasoning quality. The best model in each category (closed- and open-source) is shown in bold.

	Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	LaBSE
Closed-source	GPT-4o	6.21	63.61	42.71	58.70	76.33	82.82
	GPT-4o-mini	5.30	61.86	41.18	56.73	76.23	81.56
	GPT-4.1	6.35	**71.13**	48.83	65.33	77.32	**84.40**
	o4-mini	5.38	65.22	45.94	59.45	76.33	82.57
	Gemini 1.5 Pro	5.49	62.71	45.90	58.34	76.05	79.81
	Gemini 2.0 Flash	**8.27**	70.91	**54.81**	**65.95**	**78.56**	83.77
Open-source	Qwen2.5-VL-7B	3.21	48.51	31.19	45.97	73.03	73.67
	Llama-3.2-11B	1.75	22.83	11.20	19.63	66.89	65.41
	AIN	3.16	59.18	**43.54**	**55.41**	**73.26**	72.25
	Llama-4 Scout	**4.32**	47.74	27.52	41.07	73.06	**77.51**
	Aya-Vision-8B	3.39	**59.64**	38.98	53.80	72.54	76.84
	IntenVL3-8B	2.93	50.78	29.96	46.35	72.52	77.28

Table 4: **Lexical and Semantic Similarity Scores.** Evaluation of generated reasoning steps using classical metrics, including BLEU, ROUGE, BERTScore, and LaBSE. These metrics reflect surface-level lexical overlap and overall semantic similarity but fall short in capturing stepwise logical coherence. The best model in each category (closed- and open-source) is shown in bold.

المعيار	مستوى	التوصيف
التطابق	الخطوة	قياس مدى توافق و تطابق ودقة و موثوقية واتساق خطوات الاستدلال مع الجمل المصدرية.
التطابق	الرمز	توسيع مقياس التطابق - الخطوة (التوافق على مستوى الخطوات) عبر التحقق من التوافق والتطابق والدقة والموثوقية والاتساق على مستوى الرموز داخل خطوات الاستدلال.
الإثراء المعلوماتي	الخطوة	تقييم مدى قدرة خطوات الاستدلال على استخراج المعلومات ذات الصلة من المصدر.
تكرار	الرمز	تحديد الخطوات الاستدلالية المكررة غير اللازمة أو المعاد صياغتها داخل الفرضية.
الهلوسة	العام	اكتشاف خطوات استدلال غير المرتبطة أو غير المتوافقة مع المصدر أو سلسلة المرجع.
التكرار الزائد	العام	تحديد الخطوات الاستدلالية الزائدة التي لا تُضيف قيمة إلى عملية الحل.
التغطية الدلالية	الخطوة	تقييم مدى تغطية الفرضية للعناصر الأساسية في المصدر.
توافق الاستدلال	العام	قياس مدى التوافق والارتباط العام بين الفرضية وسلسلة المرجع.
المنطق العام	العام	الكشف عن غياب الاستدلال المبني على المنطق العام الضروري لحل المشكلة.
الخطوة المفقودة	الخطوة	تحديد ما إذا كانت هناك خطوات استدلال ضرورية مفقودة لحل المشكلة.

Figure 5: **Arabic Reasoning Evaluation Metrics.** We assess step-by-step reasoning using five core Arabic-specific dimensions: *Faithfulness (At-Tatābuq)*, *Informativeness (Al-Ithrā' Al-Ma'lūmātī)*, *Coherence (At-Tawāfuq)*, *Commonsense (Al-Mantiq Al-'Āmm)*, and *Reasoning Alignment (At-Tawāfuq Al-Istidlālī)*. Auxiliary checks cover hallucinations, redundancy, semantic gaps, and missing steps. Metrics are defined at the step and/or token level. The full evaluation rubric is provided in English in Appendix E.

4 Results and Analysis

Reasoning–Answer Performance Gap.

The ARB evaluation (Table 3) reveals a consistent gap between models' ability to generate coherent reasoning steps and their success in reaching correct final answers. For example, models like GPT-4.1 and o4-mini achieve reasoning coherence scores above 80%, while their final answer accuracy hovers around 58–60%. This pattern is even more pronounced in open models such as Qwen2.5-VL and Aya-vision, where reasoning steps are moderately strong (above 50–60%) but final answer correctness remains below 40%. These results demonstrate that well-structured reasoning does not guarantee correct conclusions—underscoring the need for

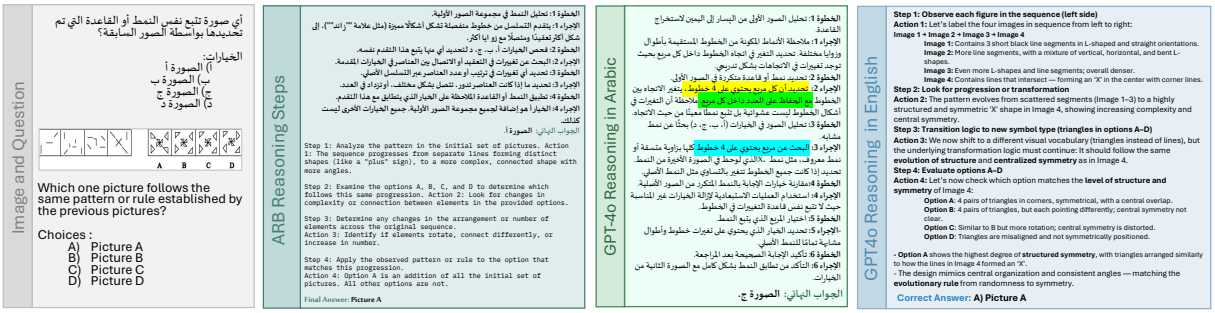


Figure 6: **Cross-Lingual Reasoning Comparison (Arabic vs. English).** This figure compares LMMs (GPT-4o) reasoning steps in Arabic and English for the same visual task. In the Arabic version, the model misinterprets structural constraints, yellow highlights incorrect assumptions about equal line counts 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.

step-level evaluation to accurately assess a model's reasoning capabilities.

Closed vs. Open-Source Model Performance.

Quantitative Evaluation. Closed-source models consistently outperform open-source ones in both reasoning and final answer accuracy. GPT-4.1 and o4-mini lead the closed category, with strong logical consistency and relatively high correctness. Among open models, LLaMA-4 Scout performs best, scoring 77.7% in reasoning steps and 48.5% in final answers—narrowing the gap with closed models but still trailing. Other open models such as LLaMA-3.2, AIN, Aya Vision, and InternVL3 demonstrate coherent reasoning but struggle with accurate conclusions, reflecting limitations in cross-lingual understanding and cultural grounding.

Qualitative Evaluation. To investigate reasoning gaps in Arabic, we conducted a qualitative comparison between model outputs and human-curated ARB references. Selected examples illustrate common pitfalls in both open- and closed-source models, including incomplete or incoherent step transitions, hallucinations, and shallow logical progression in Arabic responses (Figures 17 and 18).

We further examine the impact of language by comparing Arabic and English reasoning steps generated by the same model on identical visual inputs (Figure 6). This side-by-side analysis reveals notable inconsistencies in reasoning quality across languages, emphasizing the need for Arabic-specific benchmarks.

These findings underscore the importance of evaluating and improving Arabic multimodal reasoning, directly supporting ARB's core motivation.

Domain-Level Trends. Figures 13 and 14 (Appendix F) show a domain-level breakdown, illustrating the persistent reasoning-answer gap across task categories. Figures 15 and 16 offer fine-grained step-by-step scores, revealing domain-specific model behavior. These results underscore ARB's value in exposing nuanced reasoning patterns and highlighting the strengths and weaknesses of both closed- and open-source models across domains.

5 Conclusion

In this work, we presented ARB, the first benchmark designed to evaluate step-by-step multimodal reasoning in Arabic across 11 diverse domains. With 1.35K high-quality samples and over 5K human-curated reasoning steps, it was built through a hybrid pipeline combining prompting strategies, tool-assisted generation, and native-speaker validation. Our evaluation of 12 state-of-the-art open- and closed-weight models revealed persistent gaps in reasoning quality, coherence, and cultural alignment when operating in Arabic, despite their strong performance in English-centric settings. These findings underscore the need for step-level, culturally aware evaluation strategies tailored to underrepresented languages. Beyond benchmarking, the open-source ARB offers tools, protocols, and interfaces to support reproducibility and future research. It sets the foundation for training and evaluating Arabic-native LMMs and contributes toward building more inclusive, interpretable, and linguistically grounded AI systems.

6 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 still not fully capture the full linguistic, dialectal, or cultural variability present across the Arabic-speaking world. Additionally, reasoning evaluations rely on human judgment and model-specific prompts, which may introduce subjectivity or prompt-induced biases. The benchmark also focuses on Arabic exclusively, and does not offer multilingual alignment or cross-lingual transfer assessments, which could be valuable for 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.

References

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, et al. 2025. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. *arXiv preprint arXiv:2503.00151*.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Anthropic. 20254. Claude. AI assistant.

Ekaba Bisong and Ekaba Bisong. 2019. Matplotlib and seaborn. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, pages 151–165.

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.

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

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

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://arxiv.org/abs/2403.05530. Accessed: 2025-05-03.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad Shahbaz Khan, Salman H Khan, and Rao Muhammad Anwer. 2025a. Camel-bench: A comprehensive arabic lmm benchmark. *NAACL*.

Sara Ghaboura, Ketan More, Ritesh Thawkar, Wafa Alghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025b. Time travel: A comprehensive benchmark to evaluate lmms on historical and cultural artifacts. *arXiv preprint arXiv:2502.14865*.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

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*.

Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. 2024. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. *arXiv preprint arXiv:2408.07983*.

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*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step

length on large language models. *arXiv preprint arXiv:2401.04925*.

Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2025. Jeem: Vision-language understanding in four arabic dialects. *arXiv preprint arXiv:2503.21910*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

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

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Salima Lamsiyah, Kamyar Zeinalipour, Matthias Brust, Marco Maggini, Pascal Bouvry, Christoph Schommer, et al. 2025. Arabicsense: A benchmark for evaluating commonsense reasoning in arabic with large language models. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11.

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

Meta-AI. 2024. Llama 3.2 11b vision instruct. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct. Accessed: 2025-05-03.

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

Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. 2024. Reasoning beyond bias: A study on counterfactual prompting and chain of thought reasoning. *arXiv preprint arXiv:2408.08651*.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*.

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*.

Umair Nawaz, Awais Muhammad, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, Salman Khan, and Rao Anwer. 2025. Agriclip: Adapting clip for agriculture and livestock via domain-specialized cross-model alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9630–9639.

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

OpenAI. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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

OpenAI. 2025b. 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.

Leonardo Ranaldi and Andrè Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. *arXiv preprint arXiv:2405.00402*.

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

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in arab culture. *arXiv preprint arXiv:2502.12788*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

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.

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*.

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. *Preprint*, arXiv:2411.10440.

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.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

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*.

A Appendix

This appendix provides supplementary material supporting our contributions. It includes: (1) a brief overview of related work situating our approach within broader research on Arabic reasoning and multimodal data generation; (2) details of the filtering and verification pipeline, including interface designs used for human-in-the-loop validation and the inter-annotator agreement study; (3) additional details on the prompts used for model reasoning generation and evaluation; (4) English translations of the Arabic generation prompt and evaluation metrics; and (5) extended data statistics, such as domain and steps by domain distributions, token length distributions in questions and reasoning steps, as well as their ratios. These additions enhance transparency and offer deeper insight into the construction and quality control of the ARB benchmark.

B Related Work

Chain-of-Thought Reasoning in LLMs

CoT prompting was introduced by (Wei et al., 2022) to improve LLMs' logical reasoning, inspiring extensions like self-consistency (Wang et al., 2022), tree-of-thoughts (Yao et al., 2023), and instruction tuning for reasoning (Vaillancourt and Thompson, 2024; Ranaldi and Freitas, 2024). Recent work has also explored structural aspects of reasoning, including the impact of step length (Jin et al., 2024) and counterfactual prompting to reduce bias (Moore et al., 2024).

Building on these developments, state-of-the-art LLMs have adopted advanced post-training strategies to strengthen reasoning. Kumar et al. (Kumar et al., 2025) survey techniques such as fine-tuning, reinforcement learning, and test-time scaling. OpenAI's o1 model (Jaech et al., 2024) leverages reinforcement learning and inference-time scaling to improve reasoning fidelity. DeepSeek R1 (Guo et al., 2025) enhances CoT performance using reward models that prioritize logical soundness over natural phrasing.

Multimodal Reasoning in VLMs

Extending CoT reasoning to multimodal tasks has proven both challenging and rewarding. Models like LLaVA-CoT (Xu et al., 2025) explicitly incorporate structured visual reasoning steps into their outputs, enabling multi-stage perception and interpretation of images. Trained on a dataset of 100k CoT-annotated visual QA pairs, LLaVA-CoT achieves notable gains on reasoning benchmarks. Similarly, LlamaV-o1 (Thawakar et al., 2025) introduces a curriculum-based framework and benchmark for multi-step visual reasoning, demonstrating improvements in both accuracy and interpretability.

Recent studies have proposed methods to further enhance reasoning coherence and alignment. Chen et al. (Chen et al., 2024a) present metrics and a two-stage training strategy to improve consistency in vision-language reasoning. Zhang et al. (Zhang et al., 2024) enrich training data with rationales distilled from GPT-4o and apply Direct Preference Optimization (DPO) to guide models toward more faithful and coherent CoT outputs.

These developments reflect a growing consensus that multimodal models must reason systematically across modalities—not merely generate final answers—to ensure robustness and interoperability.

Arabic and Multilingual Reasoning Resources

Despite increasing multilingual training in LLMs, Arabic remains underrepresented in reasoning-focused benchmarks. Several datasets have emerged to address this gap. ArabicSense (Lamsiyah et al., 2025) evaluates commonsense reasoning in Arabic, while AraSTEM (Mustapha et al., 2024) offers over 11,000 science-focused multiple-choice questions in Arabic. ArabLegalEval (Hijazi et al., 2024) benchmarks Arabic legal reasoning using real-world legal documents and synthetic questions. ArabCulture (Sadallah et al., 2025) focuses on MSA commonsense reasoning across 13 Arab countries using culturally grounded, native-authored questions. AraDiCE (Mousi et al., 2024) evaluates dialectal and cultural reasoning across Arabic varieties using post-edited synthetic data.

These resources reveal substantial performance disparities between Arabic and English, particularly in reasoning-heavy tasks; however, they remain limited to the text modality and focus primarily on LLMs rather than LMMs.

Arabic-Native Reasoning Models

Recent efforts have introduced Arabic-native LLMs with enhanced reasoning capabilities. ALLaM-Thinking (Research, 2025) is a fine-tuned model specifically optimized for stepwise logic and arithmetic problem-solving, demonstrating improved chain-of-thought performance in math

tasks through Unsloth and Grouped Policy Optimization. Fanar (Team et al., 2025), a broader Arabic LLM, recently introduced the "Think Before Responding" feature, enabling intermediate reasoning traces during decoding and improving interpretability and alignment with structured reasoning. In contrast, models like AIN (Heakl et al., 2025) and Jais (Sengupta et al., 2023) offer general Arabic capabilities but lack fine-grained reasoning alignment.

ARB complements these resources by providing the first multimodal step-by-step reasoning benchmark in Arabic, creating a unified framework for evaluating reasoning transparency across vision-language tasks.

C Filtering and Verification Pipeline and 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 data and multiple annotators involved, the interface was designed to simplify inspection and accelerate review.

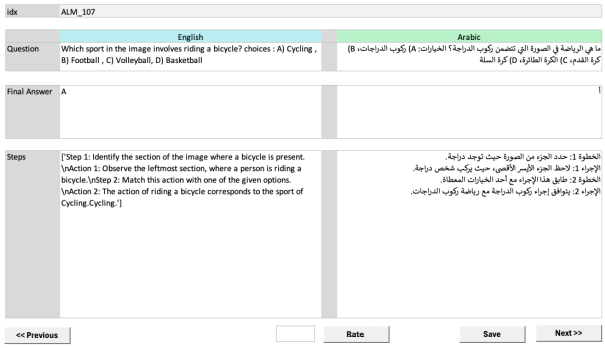
For translation tasks (see Figure 7a), the interface displays the original English text alongside the Arabic translation, allowing annotators to directly edit only the translated portion. For synthetic samples (see Figure 7b), the interface presents the image, Arabic question, step-by-step reasoning, predicted answer, and reference answer. Annotators assess the sample based on accuracy, clarity, cultural alignment, and faithful delivery of meaning, with an emphasis on conceptual correctness rather than word-for-word translation.

Each sample is rated on a 6-point scale, as shown below.

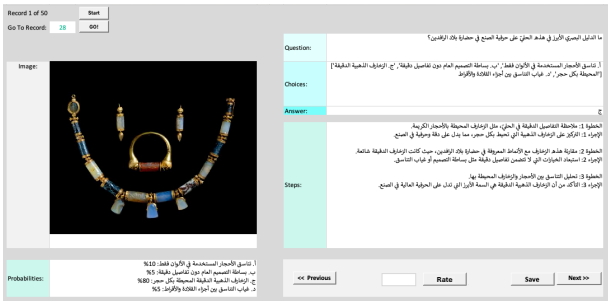
Rate	Description
0	Reject: Culturally inappropriate/ Irrelevant content
1	Reject: Requires full regeneration by the model
2	Poor: Major edits needed to fix reasoning or clarity
3	Fair: Moderate edits required
4	Good: Minor edits needed
5	Excellent: No edits needed; ready for inclusion

Table 5: **Filtering and Verification Rating Scale.** A standardized scoring scheme used by annotators to assess the quality of translations and reasoning steps. The scale guides decisions on whether a sample should be accepted, revised, or regenerated based on linguistic accuracy, reasoning clarity, and cultural appropriateness.

Each sample was independently reviewed by two annotators and then passed to a controller, with individual scores combined for a total of 10. If either annotator assigned a score of 0, the sample was immediately discarded due to cultural or contextual inappropriateness. Samples scoring 8–10 were approved without further review, while those scoring 2–4 were sent back for regeneration. Samples with intermediate scores (5–7) were escalated to a controller, who conducted a final review, resolved discrepancies, and made any necessary corrections. This multi-tiered evaluation process ensured both



(a) Example of ARB translation verification user interface.



(b) Example of ARB generated data verification user interface.

Figure 7: **Filtering and Verification User Interface.** The interface enables annotators to manually edit content when applicable and assign quality ratings to guide subsequent controller review and final approval.

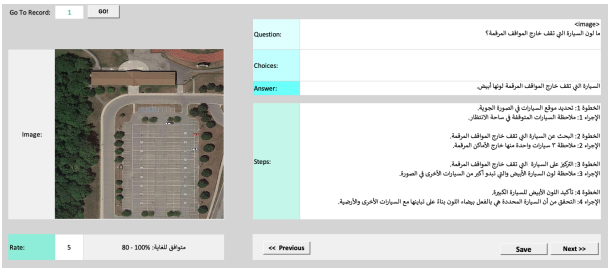


Figure 8: **Inter-Annotator Agreement Interface.** The interface allows annotators to evaluate each sample by assessing the compatibility of the model's step/action chain with the provided image, question, and choices (when applicable). Annotators assign a score by comparing the model's reasoning process to their own human reasoning approach for solving the question.

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 9: **Arabic Evaluation Prompt for LLM-as-Judge.** This prompt was used to evaluate reasoning steps across all models in Arabic. It guides models to assess reasoning quality using a set of structured criteria defined in the ARB framework.

the consistency and quality of the final dataset.

D 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 9). 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.

An English translation of the prompt is provided (Figure 10) to assist non-Arabic readers and enhance accessibility.

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 10: **English Translation of the Arabic Evaluation Prompt.** A translated version of the prompt used to evaluate reasoning steps in ARB (see Figure 9) to aid non-Arabic readers.

E English Translation of Generation Prompt and Evaluation Metrics

This section presents the English translations of two core components used in ARB: (1) the prompt for the generation of reasoning steps, originally designed in Arabic (see the Arabic version in Figure 4, the English translation in Figure 12); and (2) the evaluation metrics used to assess the quality of these reasoning steps (see original in Figure 5, the English translation in Figure 11). These metrics were also used in the evaluation prompt provided in Appendix D.

F Domain-Level Analysis of Reasoning and Final Answers

To gain deeper insight into model performance across various task categories, we present a domain-level analysis of ARB results for both closed- and open-source models. These visualizations illustrate how models perform in terms of both final answer accuracy and reasoning step quality across the 11 benchmark domains.

To support clarity and consistency across the fol-

Metric	Level	Definition / التوصيف	مستوى	المعيار
Faithfulness	Step	قياس مدى توافق و تطابق ودقة و موثوقية واتساق خطوات الاستدلال مع الجمل المصدرية. Measures the degree of alignment, consistency, accuracy, reliability, and coherence of the reasoning steps with the reference sentences.	الخطوة	التطابق
Faithfulness	Token	توسيع مقياس التطابق - الخطوة (التوافق على مستوى الخطوات) عبر التحقق من التوافق والتطابق والدقة والموثوقية والاتساق على مستوى الرموز داخل خطوات الاستدلال. Extends the Faithfulness-Step metric (alignment at the step level) by verifying alignment, accuracy, reliability, and coherence at the token level within the reasoning steps.	الرمز	التطابق
Informativeness	Step	تقييم مدى قدرة خطوات الاستدلال على استخراج المعلومات ذات الصلة من المصدر. Evaluates the extent to which the reasoning steps successfully extract relevant information from the source.	الخطوة	الإثراء المعلوماتي
Repetition	Token	تحديد الخطوات الاستدلالية المكررة غير اللازمة أو المعاد صياغتها داخل الفرضية. Identifies repeated or unnecessarily paraphrased reasoning steps in the reasoning chain.	الرمز	تكرار
Hallucination	Overall	اكتشاف خطوات استدلال غير المرتبطة أو غير المتوافقة مع المصدر أو سلسلة المرجع. Detects irrelevant or fabricated reasoning steps not aligned with the source.	العام	الهلوسة
Redundancy	Overall	تحديد الخطوات الاستدلالية الزائدة التي لا تُضيف قيمة إلى عملية الحل. Identifies redundant reasoning steps that do not add value.	العام	التكرار الزائد
Semantic Coverage	Step	تقييم مدى تغطية الفرضية للعناصر الأساسية في المصدر. Measures how well the response covers the essential semantic elements.	الخطوة	التغطية الدلالية
Reasoning Alignment	Overall	قياس مدى التوافق والارتباط العام بين الفرضية وسلسلة المرجع. Measures the overall alignment and consistency between the hypothesis and the reference reasoning chain.	العام	توافق الاستدلال
Commonsense	Overall	الكشف عن غياب الاستدلال المبني على المنطق العام الضروري لحل المشكلة. Detects the absence of essential commonsense reasoning needed to reach a solution.	العام	المنطق العام
Missing Step	Step	تحديد ما إذا كانت هناك خطوات استدلال ضرورية مفقودة لحل المشكلة. Identifies if any necessary reasoning steps are missing.	الخطوة	الخطوة المفقودة

Figure 11: **English Translation of ARB Evaluation Metrics.** An English version of the Arabic reasoning evaluation rubric used in ARB (see Figure 5), detailing the definitions of all step-level and overall reasoning quality metrics. These include measures for faithfulness, informativeness, repetition, hallucination, redundancy, semantic coverage, reasoning alignment, commonsense reasoning, and missing steps. This translation supports cross-lingual reproducibility and interpretability of the evaluation framework.

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 Category - {Curriculum} = "Computational"**: Focuses on basic arithmetic operations, comparative reasoning, and mathematical logic.
   - **Second Category - {Curriculum} = "Scientific/Medical"**: Involves scientific reasoning and domain-specific evidence-based analysis.
   - **Third Category - {Curriculum} = "Descriptive/Inference"**: Emphasizes segmenting and analyzing the visual content to extract meaningful insights.
   - **Fourth Category - {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 12: **English Version of the ARB Prompt.** This figure presents the English translation of the original Arabic prompt (see Figure 4) used to guide reasoning step generation across domains.

lowing visual analyses, we adopt the following standardized abbreviations for the 11 ARB domains:

Abb	Description
VR	Visual Reasoning;
OCR	OCR and Document Analysis;
CDT	Charts, Diagrams, and Tables;
M&L	Mathematical and Logical Reasoning;
Soc.Cult.	Social and Cultural Understanding;
CVP	Complex Visual Perception;
MED	Medical Image Analysis;
Sci.R	Scientific Reasoning;
Hist.	Historical & Archaeological Interpretation;
RS	Remote Sensing Analysis;
Agro	Agricultural Image Understanding.

The bar charts (Figures 13 and 14) provide an overview of the aggregated scores, while the heat maps (Figures 15 and 16) offer a more granular perspective on domain-level performance across individual evaluation metrics. Together, these figures reveal consistent discrepancies between reasoning coherence and final answer correctness, and highlight domain-specific strengths and weaknesses across model types.

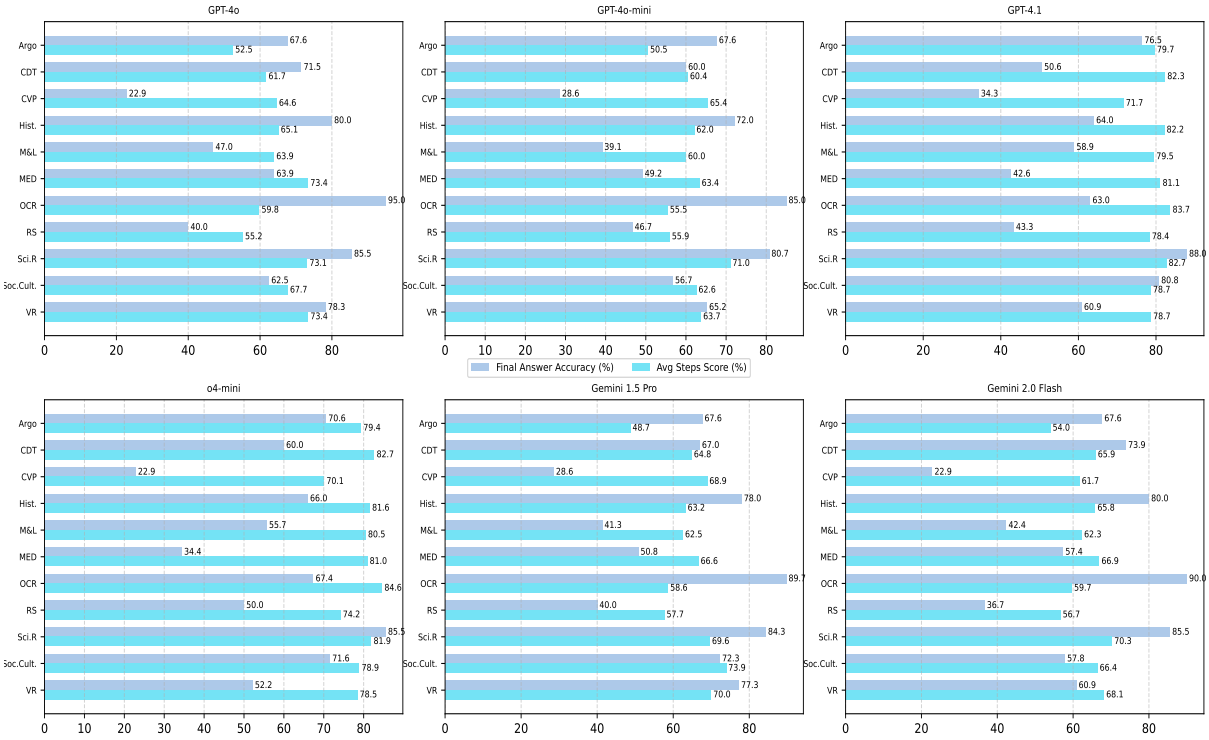


Figure 13: **Domain-Level Performance of Closed-Source Models.** Bar charts comparing final answer accuracy and average reasoning step quality across ARB domains for each closed-source model. GPT-4.1 and o4-mini show strong reasoning in domains like Sci.R, CDT, and Hist., while notable gaps appear in CVP and RS. All models consistently score higher on reasoning than final answers, underscoring the importance of step-level evaluation. The figure highlights both strengths and limits of closed models in Arabic multimodal reasoning.

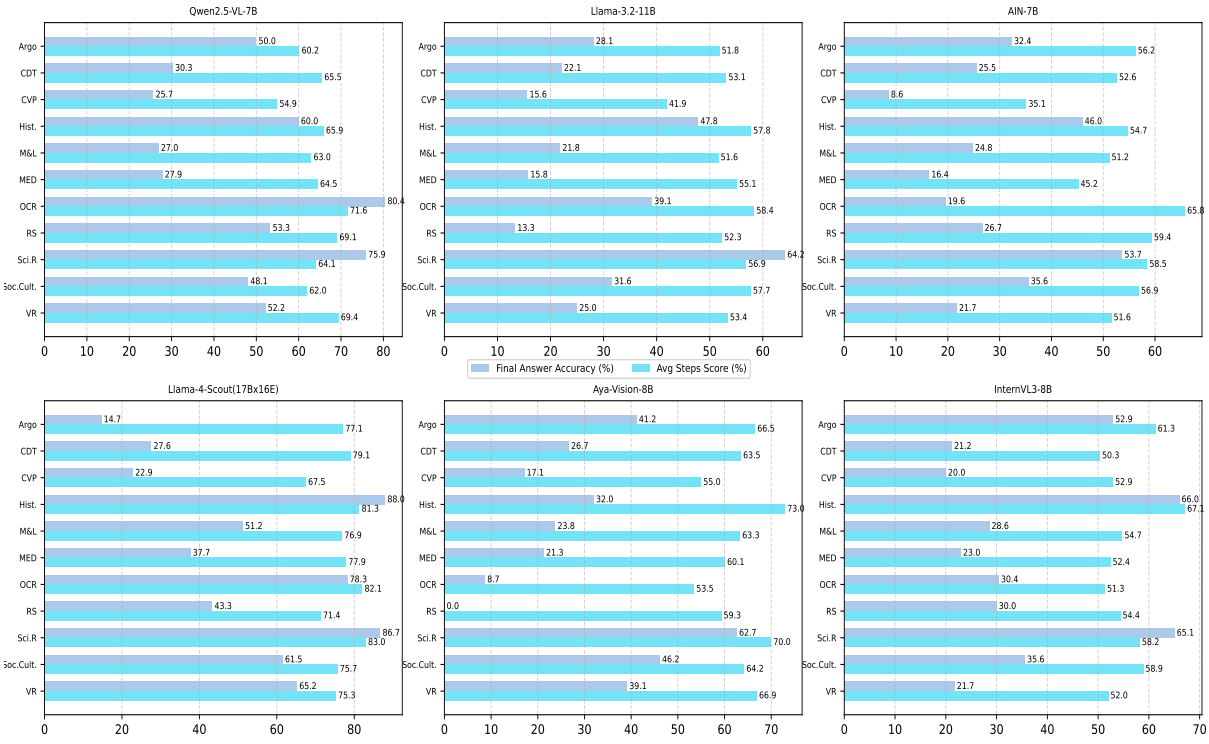


Figure 14: **Domain-Level Performance of Open-Source Models.** Comparison of final answer accuracy and reasoning step scores across ARB domains for six open-source models. LLaMA-4 and AIN perform well in Sci.R and OCR but struggle in RS and VR. Qwen2.5-VL and LLaMA-3.2 show large gaps between reasoning and answers, especially in culturally grounded domains (e.g., Hist., Soc.Cult.). The figure illustrates challenges open models face in Arabic cross-modal reasoning.

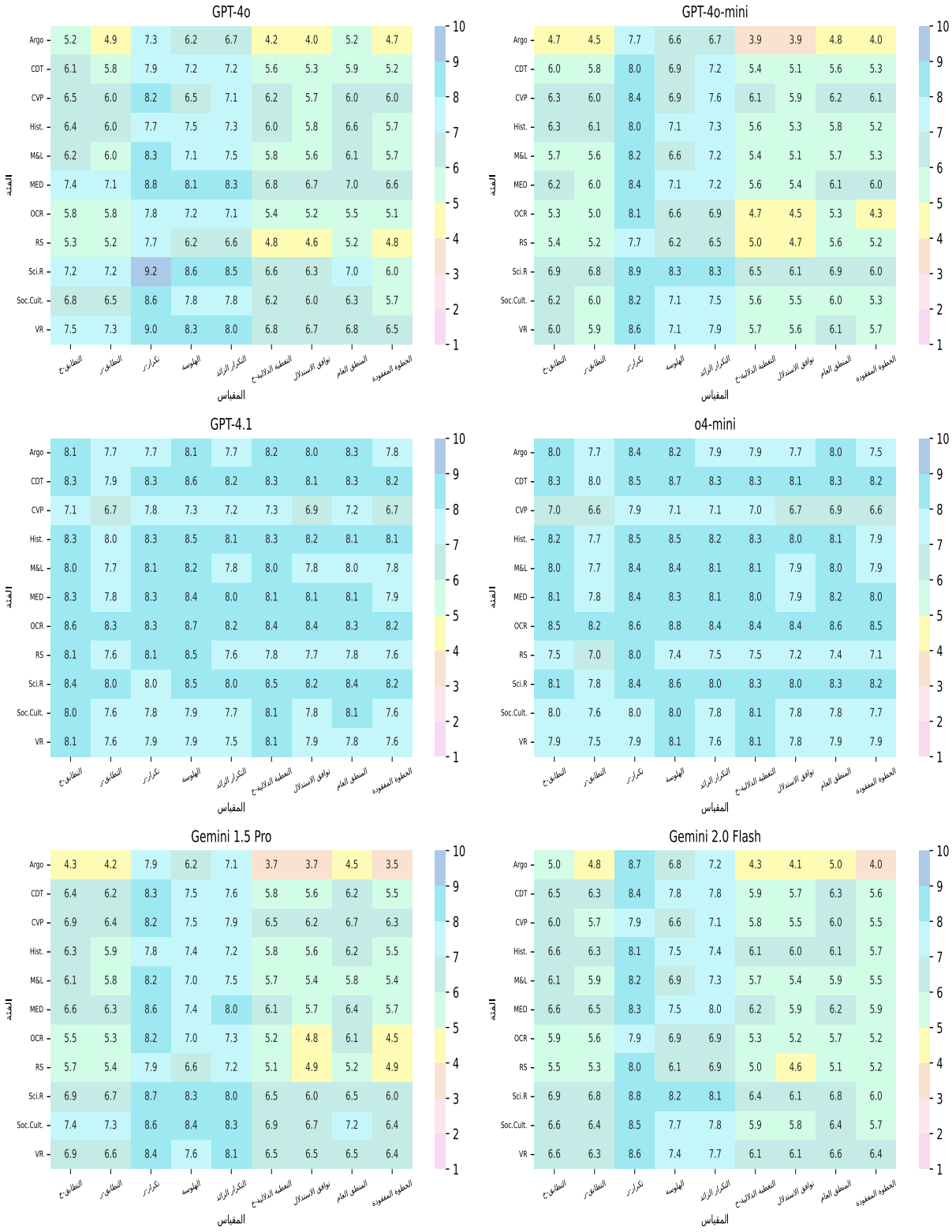


Figure 15: **Stepwise Attribute-Level Evaluation of Closed-Source Models.** Heatmaps illustrating the average scores (1–10 scale) across key reasoning attributes—faithfulness, coherence, informativeness, and other diagnostic criteria—within each ARB domain for six closed-source models, based on the LLM-as-Judge framework using Arabic prompts. Models such as GPT-4.1 and o4-mini consistently achieve high scores across most attributes and domains, particularly in Sci.R, CDT, and Hist., indicating strong reasoning reliability. In contrast, performance degrades in perceptual-heavy domains like CVP and RS, where scores drop across multiple attributes. The heatmaps also expose granular inconsistencies—e.g., faithfulness gaps in MED or informativeness variability in Agro—that would be obscured by aggregate metrics. These results emphasize the value of attribute-level evaluation in diagnosing model reasoning quality in Arabic multimodal tasks.

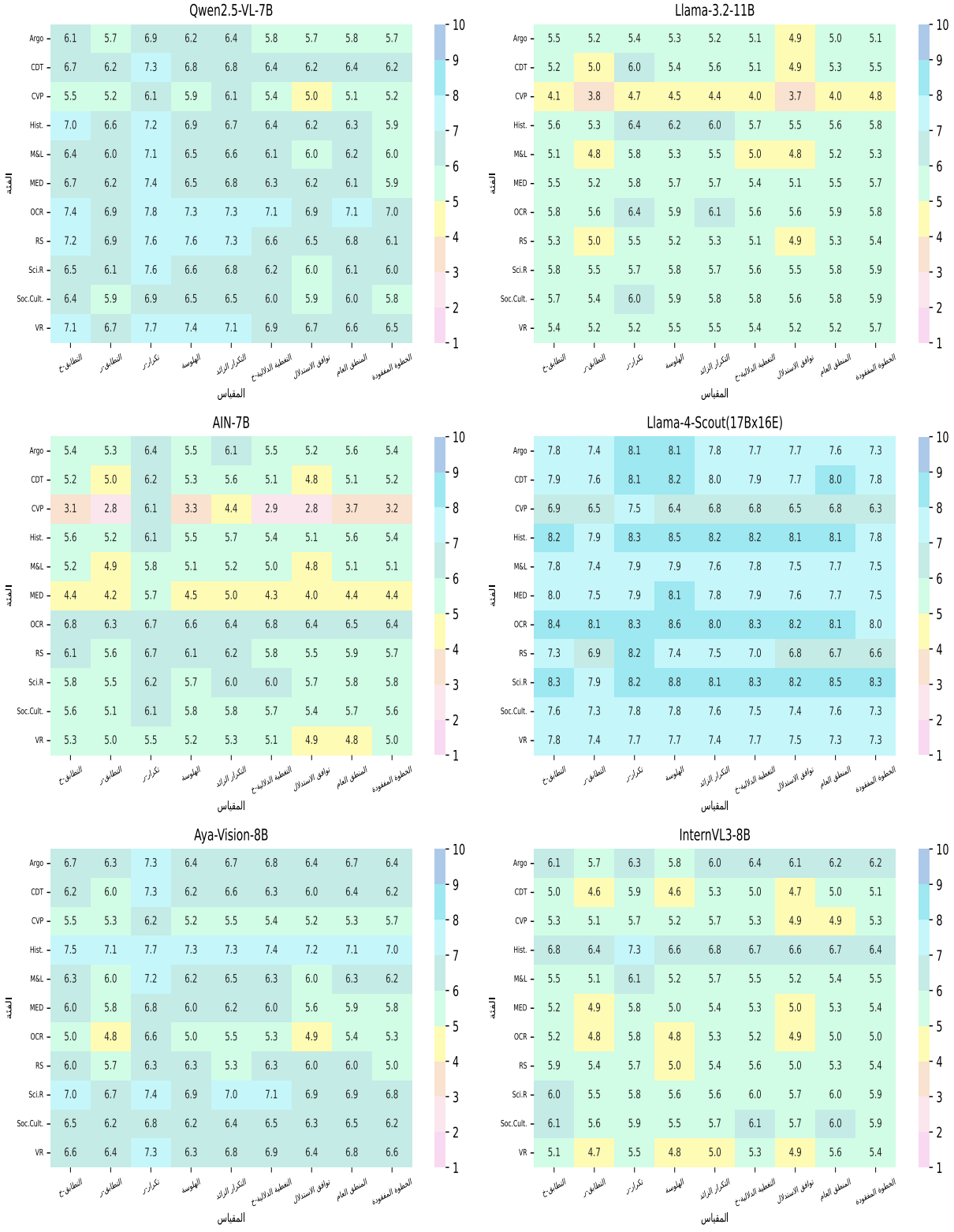
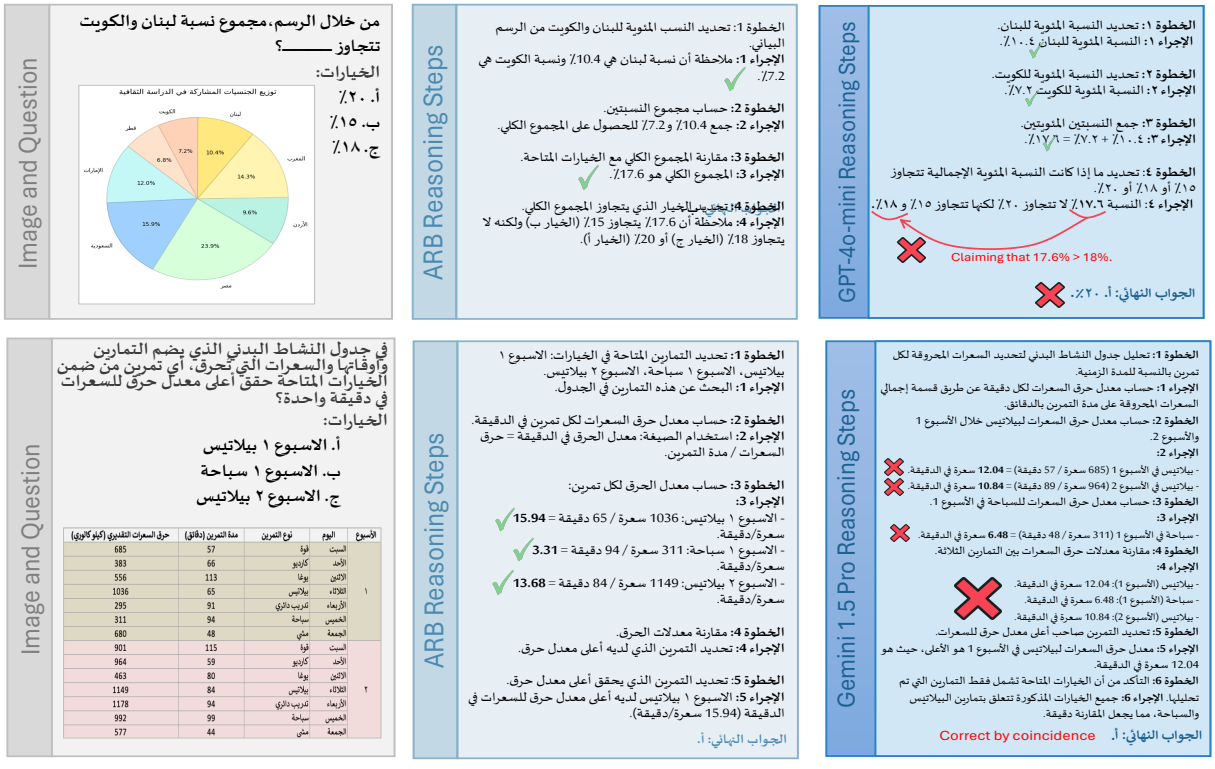


Figure 16: **Stepwise Attribute-Level Evaluation of Open-Source Models.** Heatmaps visualizing average attribute-level scores (1–10 scale) across ARB domains for six open-source models, based on the LLM-as-Judge framework using Arabic prompts. Each cell reflects the model's performance across core reasoning dimensions—faithfulness, coherence, informativeness, and error-related factors—per domain. Models such as LLaMA-4 and AIN demonstrate consistent stepwise quality across scientific and OCR tasks, while others like Qwen2.5-VL and LLaMA-3.2 struggle in culturally sensitive or perception-heavy domains (e.g., Hist., Soc.Cult., RS). These results offer fine-grained insight into open-model weaknesses and underscore the importance of domain- and attribute-aware evaluation in Arabic multimodal reasoning tasks.

Figure 17: **Qualitative Errors in Open-Source Models.** This figure showcases common reasoning flaws in open-source LMMs across diverse Arabic multimodal tasks. Errors include incomplete reasoning steps, inconsistent logic, and hallucinated interpretations not grounded in the input. These issues often result in incorrect answers or unreliable outputs, reflecting the challenges open models face in structured Arabic reasoning.



الأسبوع	اليوم	نوع التمرين	مدة التمرين (دقائق)	حرق السعرات التقديري (كيلو كالوري)
١	السبت	قوة	57	685
	الأحد	كارديو	66	383
	الاثنين	يوغا	113	556
	الثلاثاء	بيلاتيس	65	1036
	الأربعاء	تدريب دائري	91	295
	الخميس	سباحة	94	311
	الجمعة	مشي	48	680
٢	السبت	قوة	115	901
	الأحد	كارديو	59	964
	الاثنين	يوغا	80	463
	الثلاثاء	بيلاتيس	84	1149
	الأربعاء	تدريب دائري	94	1178
	الخميس	سباحة	99	992
	الجمعة	مشي	44	577

Figure 18: **Qualitative Errors in Closed-Source Models.** This figure highlights reasoning failures by closed-source LMMs across various Arabic multimodal tasks. Common issues include incorrect numerical comparisons, invalid assumptions, misinterpreted constraints, and logically inconsistent step sequences. These errors often lead to incorrect conclusions despite the appearance of structured reasoning, underscoring the limitations of current closed models when operating in Arabic.

G Qualitative Examples

As a further illustration of the quantitative trends discussed in section 4, we present qualitative examples of reasoning failures in both open- and closed-source models (Figures 17 and 18). These examples reveal persistent issues such as incomplete reasoning chains, hallucinated content, and misapplied constraints across a range of Arabic multimodal tasks. While some outputs appear structurally coherent, they often fail to adhere to task-specific logic or factual correctness. These qualitative insights reinforce the need for Arabic-centric benchmarks like ARB to diagnose and improve model behavior in complex reasoning scenarios.

H Data Statistics

H.1 Distribution of Reasoning Steps per Sample

To examine the structure of the ARB benchmark across domains, we report key statistical findings. Figure 20 illustrates the distribution of step counts in all ARB entries over their domains, revealing the frequency and variance of the step depth required for the completion of the task.

H.2 Token Count by Domain

Figure 21a shows the distribution of question token lengths across domains. Most questions are relatively concise, but domains such as Medical Reasoning (MED) and Historical and Archaeological Understanding (Hist.) exhibit higher variability and longer lengths. This reflects the inherent complexity and information density required in specialized domains. Similarly, Figure 21b presents the token length distribution of the reasoning steps. These are often longer in domains like Medical Reasoning, Math and Logic (M&L), and Historical and Archaeological Understanding, indicating the need for more elaborate multi-step reasoning in knowledge-intensive tasks.

H.3 Question-to-Reasoning Token Ratio

Figure 22 depicts the average ratio of question tokens to reasoning step tokens across domains. Generally, reasoning steps are significantly longer than the original questions, with ratios exceeding 30% in most cases. Notably, the Medical Reasoning (MED) and Agricultural Image Interpretation (Argo) domains show the highest ratios, suggesting that these tasks demand extensive inferential elaboration beyond the surface-level query.

H.4 Performance Correlation with Length

Preliminary analysis indicates that longer reasoning chains are modestly correlated with improved performance in complex domains such as Medical and Scientific Reasoning. However, excessive verbosity does not consistently yield higher accuracy, highlighting the importance of targeted, efficient reasoning over mere length.

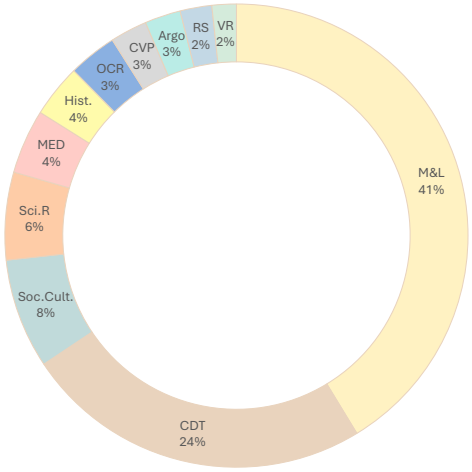


Figure 19: **Domain Distribution in ARB.** The figure shows the distribution of ARB samples across 11 domains. Math & Logic (41%) and Charts, Diagrams, & Tables (24%) dominate, reflecting the dataset's emphasis on structured reasoning. Other domains, including Social & Cultural, Scientific, and Medical, add thematic diversity.

H.5 Average Number of Steps and Domain Effects

On average, domains such as Medical, Scientific Reasoning, and Historical and Archaeological Understanding require a greater number of reasoning steps per question, compared to more straightforward domains like OCR or Remote Sensing (RS). This suggests that scientifically and historically grounded tasks inherently involve deeper multi-hop reasoning, presenting greater challenges for both human annotators and models.

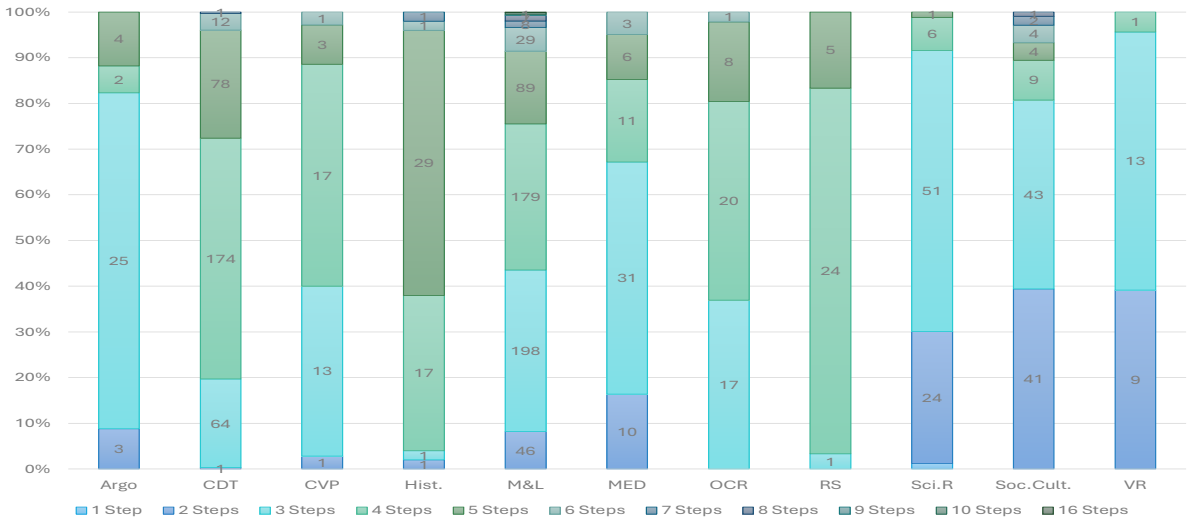
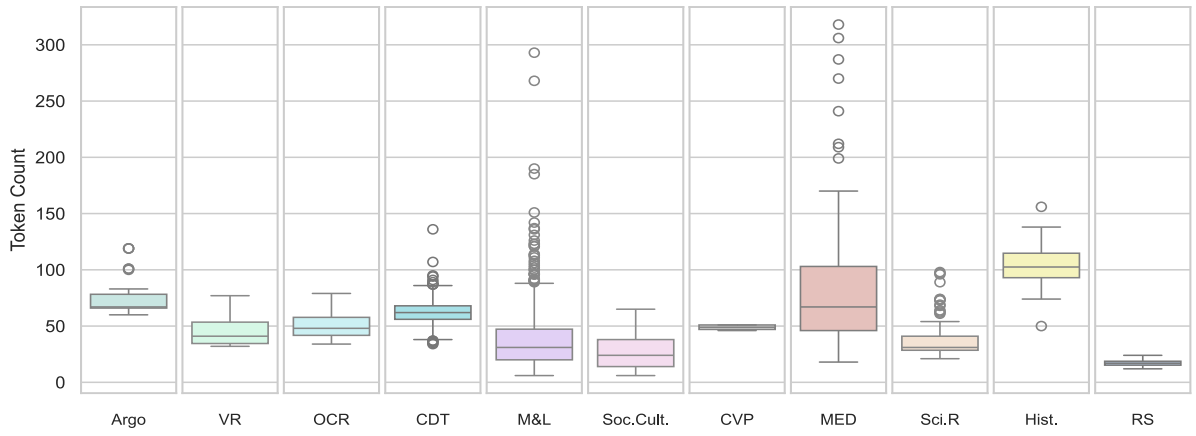
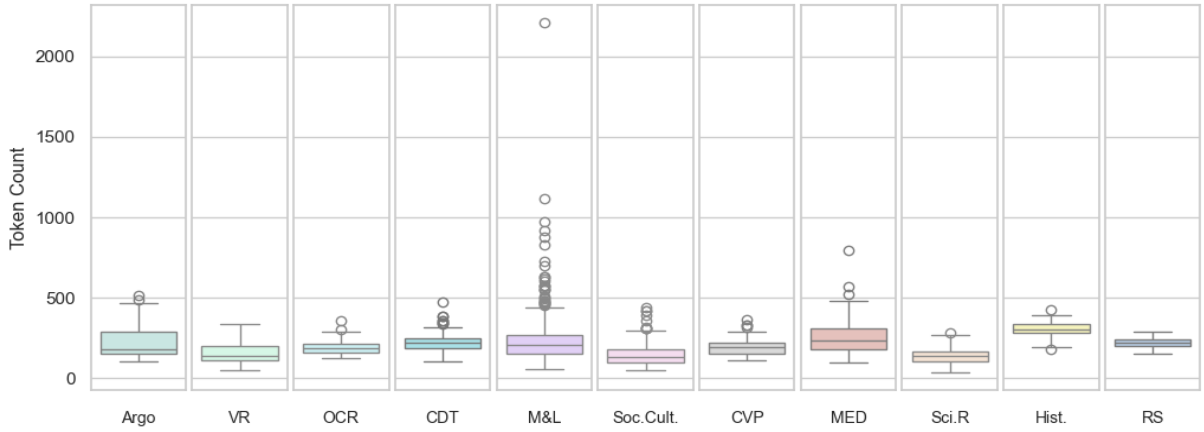


Figure 20: **Step Count Distribution by Domain.** This figure shows the frequency distribution of reasoning steps per sample across the 11 ARB domains. Most domains exhibit a concentration between 2 and 6 steps, with Math & Logic, History, and Remote Sensing containing a larger share of samples requiring extended reasoning chains.



(a) **Question Token Length Distribution by Domain.** The figure shows the distribution of token counts for questions across different domains in ARB. Domains such as Medical Reasoning (MED) and Historical and Archeological Understanding (Hist.) exhibit higher variability and longer questions, reflecting their inherent complexity.



(b) **Reasoning Steps Token Length Distribution by Domain.** The figure presents the distribution of token counts for the generated reasoning steps across domains. Reasoning steps tend to be longer in complex domains such as Medical, Math & Logic, and Historical & Archaeological Understanding (Hist.), highlighting the need for extended multi-hop reasoning.

Figure 21: Question token analysis in ARB: (a) token length by domain, and (b) [describe the second figure].

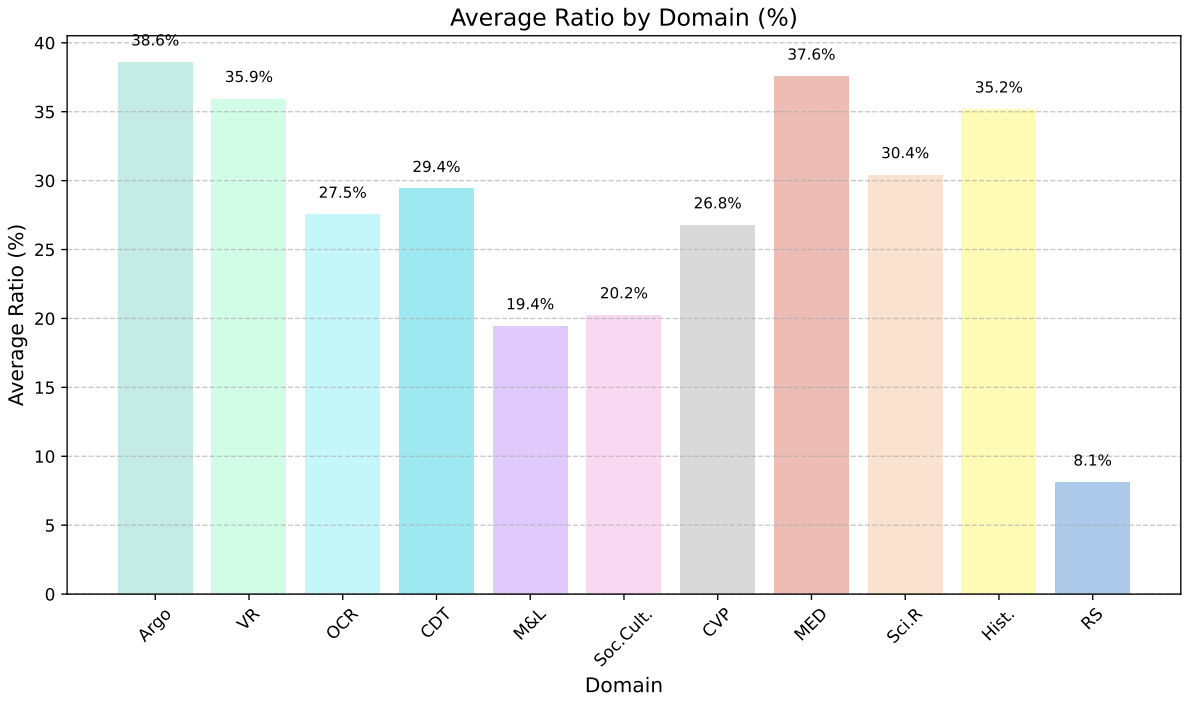


Figure 22: **Question-to-Reasoning Token Ratio by Domain.** The figure illustrates the average ratio between question token lengths and reasoning step token lengths across domains. Higher ratios in domains like Argo and MED indicate that these tasks require significantly more elaborate reasoning chains compared to the original question length.