AraVQA: Building a New Arabic Factoid Visual Question Answering Dataset from Wikipedia

Sultan Alrowili[1]*, Younes Samih[2]*, Abed Alhakim Freihat[3], Mathan Kumar Eswaran[1]

[1] IBM Research AI, Riyadh, Saudi Arabia
[2] IBM Research AI, Abu Dhabi, UAE
[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

{sultan.alrowili,younes.samih}@ibm.com

Abstract

The development of large-scale Visual Question Answering (VQA) datasets has traditionally relied on resource-intensive manual annotation. In addition, most of the existing Arabic VQA datasets focus on culturally-specific and dialect-aware domains. To address these limitations, we propose a new pipeline that leverages Wikipedia template tags to extract the relevant information for each image, which is subsequently utilized by the Large Language Model (LLM) to synthetically generate a new visual question answering dataset. Using this pipeline, we have constructed AraVQA, the most comprehensive Arabic Factoid Visual Question Answering dataset, containing more than 50,000 questions and covering over 20 varied primary subjects within Arabic general knowledge. Our detailed analysis shows that our dataset can serve as a post-training dataset to enhance the performance of existing Visual Language Models (VLMs) on Arabic VQA tasks. Furthermore, we present a novel benchmark, derived from our dataset and validated through manual annotation, that poses more challenges to Arabic VLMs than existing Arabic VQA datasets.

1 Introduction

Large Language Models (LLMs) designed for Arabic have recently achieved substantial progress, with systems such as ALLaM (Bari et al., 2024), Jais (Sengupta et al., 2023), and Command R7B Arabic (Alnumay et al., 2025) establishing new state-of-the-art results on knowledge-intensive benchmarks, including ArabicMMLU (Koto et al., 2024) and ArabicEXAMS (Hardalov et al., 2020). These advances illustrate the rapid maturation of Arabic Natural Language Processing (NLP) in text-only domains, particularly for tasks requiring factual reasoning and broad domain coverage. In parallel, Vision–Language Models (VLMs) have transformed multimodal research in English and other high-resource languages. Yet the progress of Arabic VLMs has been markedly slower, leaving multimodal reasoning in Arabic comparatively underexplored.

Existing benchmarks such as Henna (Alwajih et al., 2024), CAMEL-Bench (Ghaboura et al., 2025), Pearl (Alwajih et al., 2025), and JEEM (Kadaoui et al., 2025) have provided valuable benchmarks for evaluating Arabic VLMs. However, these resources predominantly emphasize culturally grounded, dialect-sensitive tasks or are not publicly accessible (e.g., CAMEL-Bench). Henna and Pearl focus on culturally salient domains such as cuisine, attire, and landmarks, while JEEM evaluates captioning and VQA across dialects. Although such datasets enrich the cultural and linguistic scope of evaluation, they do not systematically probe encyclopedic and factoid knowledge—a domain central to assessing whether VLMs can integrate visual input with general knowledge reasoning. Developing such resources presents significant challenges: large-scale image–question–answer annotation is prohibitively expensive, the intensive effort required to link images with relevant context, and the limitation of translation-based adaptation, which often omits Arabic-specific content and diminishes linguistic quality.

To bridge this gap, we introduce AraVQA, the first large-scale Arabic factoid visual question answering dataset. Our approach exploits the structural affordances of Arabic Wikipedia to automatically pair images with related captions and section-level textual context. These pairs are then processed by large-scale LLMs, guided by carefully designed prompts, to generate a multiple-choice Arabic visual question dataset. The resulting dataset is fully automatic and scalable, providing broad topical coverage that extends well beyond the cultural focus of existing benchmarks and explicitly targets factoid and world knowledge domains, while the underlying generation pipeline

*Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2026–2042

is resilient and applicable across Wikipedia languages.

Thus, the main contributions of this work are summarized as follows:

- **Automatic generation pipeline**. We introduce a fully automatic pipeline that leverages Arabic Wikipedia and large-scale LLMs to synthetically generate high-quality Arabic VQA data without costly manual annotation or reliance on translation. Our quality-control analysis shows that over 94% of the generated questions are valid, highlighting the robustness and reliability of our pipeline.

- **AraVQA dataset**. Using this pipeline, we construct AraVQA, the largest Arabic VQA dataset explicitly designed to support encyclopedic and factoid visual question answering. AraVQA spans a wide range of topics, including science, history, geography, and culture, establishing itself as the first large-scale resource of its kind in the Arabic language.

- **Human-verified benchmark**. From AraVQA, we derive a benchmark subset that is automatically generated from our pipeline and subsequently verified by six human annotators to ensure correctness and consistency.

- **Evaluation and community tools**. We conduct both quantitative and qualitative analyses showing that AraVQA exhibits greater topical breadth and difficulty than existing Arabic VQA datasets, and we demonstrate empirically that fine-tuning Arabic VLMs on AraVQA yields measurable performance improvements on existing Arabic VQA benchmarks. We release the resources used in this work via a public GitHub repository (`https://github.com/mbzuai-nlp/AraVQA`).

2 Related Work

Recent advances in Multimodal Large Language Models (MLLMs) have revealed the effectiveness of coupling frozen vision encoders with pretrained language models to achieve unified visual–textual reasoning. Architectures such as Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023b) project visual embeddings into a compact latent space via gated cross-attention or Q-Formers, while LLaVA (Liu et al., 2023a) employs a lightweight adapter to retain fine-grained perceptual signals. Collectively, these models demonstrate that large-scale multimodal alignment can support interactions between perception and reasoning. Yet, their success is contingent on massive English-centric data, a prerequisite absent for Arabic.

A complementary trajectory, visual instruction tuning, has made multimodal supervision scalable by automatically generating instruction–response pairs. MULTIINSTRUCT (Xu et al., 2023), MIMIC-IT, and LLaVA-Instruct (Li et al., 2023a; Liu et al., 2023b) have shown that synthetic data can unify captioning, VQA, and reasoning tasks under a shared schema. However, these pipelines remain exclusively English-based. Transferring such methods to Arabic introduces non-trivial linguistic and cultural challenges, as the language's complex morphology, diglossia, and culturally grounded semantics complicate automatic alignment.

Recent Arabic multimodal efforts—Henna (Peacock) (Alwajih et al., 2024), CAMEL-Bench (Ghaboura et al., 2025), Pearl (Alwajih et al., 2025), and JEEM (Kadaoui et al., 2025)—have advanced cultural and dialectal evaluation. Yet these datasets emphasize thematic domains (e.g., cuisine, landmarks) or dialect-specific reasoning. Thus, the domain of general knowledge remains overlooked in the existing Arabic Visual Question Answering (VQA) datasets. Moreover, the reliance on manual curation for these datasets poses challenges in terms of scalability, which can be observed from the number of images used to generate each dataset, as shown in Table 1.

On the other hand, the emergence of AIN (Heakl et al., 2025) demonstrates the feasibility of Arabic multimodal modeling, achieving strong results on CAMEL-Bench. However, its improvements largely stem from architectural scale and bilingual pretraining rather than Arabic-native multimodal supervision, highlighting the limitations of current methods. While cultural Arabic datasets have made progress in closing the gap, there is still a need for a new dataset that encompasses general knowledge in the Arabic language and does not rely on manual annotation. Such a dataset would enable Arabic VLMs to leverage these resources in pre-training and fine-tuning phases, facilitating more comprehensive and scalable approaches to Arabic VQA.

Dataset	Images	Questions	Method	Data Split
Henna (Alwajih et al., 2024)	120	1,132	A + M	Test
CAMEL-Bench (Ghaboura et al., 2025)	-	29,036	A + M	Test
JEEM (Kadaoui et al., 2025)	2,196	10,890	A	Test
Pearl (Alwajih et al., 2025)	12,637	309k	A + M	Test
AraVQA (Ours)	46,958	50,757	A	Train / Test

Table 1: Details about the existing Arabic Visual Question Answering datasets compared to our dataset, AraVQA. **A**: Automatic, **M**: Manual Annotation.

3 Method

Our proposed method consists of three phases: (1) data collection, (2) pre-processing image metadata, and (3) synthetic visual question generation with LLMs. We will explain each of these three phases in Section 3.1, Section 3.2, and Section 3.3, respectively. In addition, to assess the quality of our pipeline in generating valid questions, we added a quality control phase, explained in detail in Section 3.4.

3.1 Data Collection

Our methodology begins with the acquisition and structural parsing of the Arabic Wikipedia dump, which retains the full set of MediaWiki template tags. These tags provide a rich source of metadata that we exploit to extract article sections, section titles, and the associated visual elements. In particular, the preserved markup enables a systematic segmentation of articles into coherent sections, each of which can be explicitly linked to its corresponding image and contextual snippet. From these same templates, we additionally extract auxiliary descriptors such as image captions, textual descriptions, and article-level titles.

A practical challenge in utilizing the raw Wikipedia dump is that template tags do not include complete image URLs. To resolve this, we compute the MD5 hash of each image filename and use it to reconstruct the canonical URL path for the corresponding image resource by following Wikimedia's storage convention, in which the first character and the first two characters of the MD5 hash determine the intermediate path components that lead to the full image URL.

A second limitation concerns the absence of explicit license metadata within the dump. To recover this information, we query the Wikidata Query Service[1] for each image, retrieving its copyright license property.[2]

As shown in Figure 1, the final dataset integrates all these components into a structured representation consisting of the following fields: article title, article identifier, section title, section text, image URL, image description, and image license. This organization facilitates downstream alignment between textual sections and visual content, forming the foundation for subsequent multimodal question–answer generation with LLMs.

3.2 Dataset Preprocessing

Variability in the dimensional properties of Wikipedia images presents a potential challenge for Vision–language models (VLMs), as high-resolution inputs substantially increase both fine-tuning and inference costs. To standardize image representations and ensure computational efficiency, all images are resized such that their longest dimension (height or width) does not exceed 224 pixels. This resolution threshold was determined empirically: preliminary experiments revealed that increasing image size to 768 or 1024 pixels yields only marginal performance improvements while imposing significant computational overhead. To maintain reproducibility and facilitate future research, the dataset retains the original image URLs, thereby enabling access to full-resolution versions when required.

To preserve unambiguous alignment between textual and visual modalities, we exclude all sections containing multiple images. We additionally remove sections whose textual content exceeds 1,024 tokens to prevent out-of-memory (OOM) failures during large-scale synthetic generation with large language models. Finally, sections lacking valid image URLs are discarded to ensure data completeness and structural consistency across the dataset.

[1] https://query.wikidata.org/

[2] https://www.wikidata.org/wiki/Property:P275

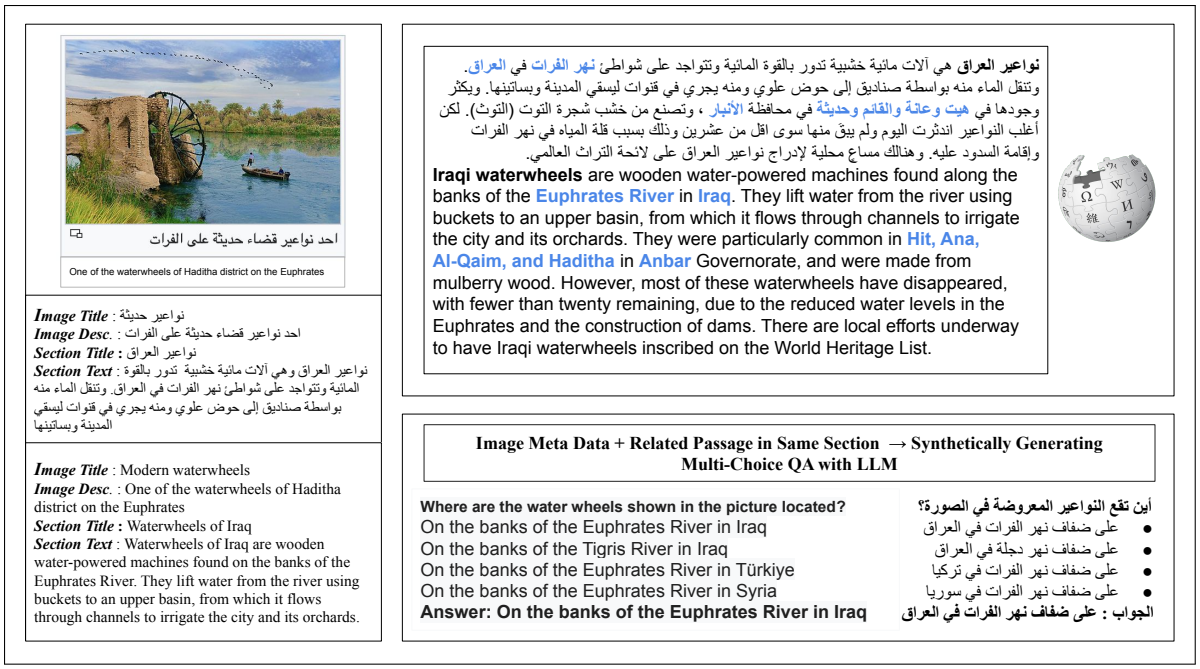


Figure 1: Illustration of our proposed method. Additional illustration of the Wikipedia template tags and the data collection phase are shown in Figure 4.

3.3 Synthetic Question Generation

We employ the GPT-OSS-120B model[3] to automatically generate multiple-choice questions from the dataset described in the preceding section. The generation process is executed using the vLLM inference framework (Kwon et al., 2023) and the Transformers library (Wolf et al., 2020) on a single NVIDIA H100 GPU, guided by the instruction prompt illustrated in Figure 2. To ensure structural consistency and adherence to the predefined template, we implement rigorous post-generation filtering. Outputs that deviate from the expected format—such as missing or incomplete answer options, or placeholder text (e.g., the model producing "[question]" instead of an actual interrogative form like "Who is the player shown in the picture?")—are automatically excluded during post-processing. We further enforce a balanced distribution of answer options by randomizing the order of the four candidate answers (A–D) and subsequently reassigning the corresponding correct label.

Sampling configuration plays a crucial role in controlling the factuality and precision of the generated content. Given our focus on factoid visual question answering, we adopt a low-temperature setting of 0.1 to reduce randomness, combined with a Top-P value of 0.9 to maintain lexical diversity without compromising determinism. The maximum sequence length is set to 2,048 tokens, with a generation limit of 1,024 new tokens per output. To expedite large-scale generation, we parallelize inference by dispatching concurrent requests to the vLLM server, achieving efficient throughput while preserving generation quality.

3.4 Quality Control with Human Annotation

We add a quality control phase, which aims to evaluate the quality of the visual question answering dataset generated from our proposed method using manual annotation. The annotation team comprises six native Arabic speakers representing diverse regional varieties, each holding at least a bachelor's degree. To ensure both linguistic coverage and inter-rater reliability, the test set is partitioned into six equal segments, with each segment independently annotated by two annotators. Thus, every example receives two judgments. The task follows a binary decision scheme (approve or reject), as the goal—unlike prior Arabic resources—is to evaluate the precision of our automatic generation pipeline rather than to collect human-written questions.

All annotators were provided with detailed guidelines and dataset documentation, as illustrated in Figure 5. The annotation was conducted through a custom web interface, as shown in Figure 6, that presents the question, corresponding image, and contextual information required for evaluation. An-

[3] https://huggingface.co/openai/gpt-oss-120b

notators were compensated at a rate of $15 per hour. Inter-Annotator Agreement (IAA) was established by retaining all examples jointly approved by both annotators for inclusion in the final test set. To quantify annotation reliability, we compute Cohen's κ (kappa) (Landis and Koch, 1977), yielding $\kappa = 0.92$ in the initial round and $\kappa = 0.96$ after adjudication—values that denote near-perfect agreement under the Landis and Koch interpretation.

4 Dataset Evaluation

Our evaluation integrates complementary quantitative and qualitative analyses. Section 4.1 reports a quantitative characterization of the final dataset, detailing its scale, distributional properties, and internal consistency. Section 4.2 conducts a qualitative assessment of AraVQA both as a benchmark and as a supervised fine-tuning corpus, comparing it to the Pearl dataset.[4] Multiple evaluation setups are employed to ensure a comprehensive and interpretable assessment of dataset quality and downstream utility.

4.1 Quantitative Analysis

Our proposed pipeline produced a total of 51,063 question–image pairs. Following standard practice in the literature, we allocated 10% of the data to the test split, yielding 5,097 examples. The subsequent quality control phase, described in Section 3.4, filtered out 306 invalid questions from the test set, resulting in a pipeline precision of 94%. After filtering, the final AraVQA dataset comprises 50,757 validated questions paired with 46,958 unique images. Of these, 45,969 belong to the training set, and 4,791 to the test set.

Table 2 presents the topic distribution across the combined training and test sets. The dataset spans over twenty distinct domains encompassing a wide spectrum of Arabic general knowledge. These include major academic and cultural fields such as History, Natural Sciences, Social Sciences, Geography, Medicine, Sports, Applied Sciences, and Religion. This diversity ensures broad topical coverage and supports the evaluation of multimodal reasoning across multiple domains of factual and contextual knowledge. Representative examples from the dataset are illustrated in Figure 3, demonstrating the range and quality of the generated image–question–answer triples.

Topic	Questions
Humanities & History	10,387
Natural Sciences & Technology	6,315
Social Sciences	4,826
Geography & Tourism	4,384
Medicine & Health Sciences	3,131
Sports & Physical Activities	2,940
Engineering & Applied Sciences	2,835
Religion & Theology	2,510
Business & Economics	2,102
Film, Television & Media	1,813
Architecture & Urban Design	1,596
Cultures & Folklore	1,527
Arts & Visual Culture	1,350
Literature & Poetry	1,331
Performing Arts	901
Education & Pedagogy	768
Mathematics & Computer Science	743
Law & Governance	599
Language & Linguistics	436
Philosophy & Ethics	263

Table 2: Topic distribution of the AraVQA dataset. We use the GPT-OSS-120B model to assign the topic for each question.

4.2 Qualitative Analysis

Our qualitative evaluation is conducted in two complementary setups designed to assess the utility and generalizability of the proposed AraVQA dataset.

In the first setup, we evaluate the zero-shot performance of a suite of state-of-the-art Vision–Language Models (VLMs) on the AraVQA test split and the multiple-choice question subset of the Pearl dataset. The evaluated models include Qwen2-VL-72B-Instruct (Wang et al., 2024), Llama-3.2-11B-Vision-Instruct (Meta AI, 2024), InternVL2-Llama3-76B (Chen et al., 2024c,b,a), Llama-4-Scout-17B-16E-Instruct (Meta, 2025), Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025), Granite-3.3-2B-Vision (Team et al., 2025), and Phi-4-Multimodal-Instruct (Microsoft et al., 2025). This configuration allows for a direct zero-shot comparison between Arabic and multilingual benchmarks, thereby assessing model performance in the absence of task-specific fine-tuning.

In the second setup, we fine-tune the Granite-3.3-2B-Vision and Phi-4-Multimodal-Instruct models

[4]Pearl is selected for comparison because it provides publicly available question–answer pairs derived from a substantially larger image base than alternative Arabic VQA datasets such as CAMEL-Bench.

Model	Size	Post-Training Setup	Pearl$_{MCQ}$	AraVQA$_{Test}$
Qwen2-VL-72B-Instruct	73.4B	Zero shot	0.7268	0.6369
Llama-3.2-11B-Vision-Instruct	10.7B	Zero shot	0.6505	0.5621
InternVL2-Llama3-76B	76.3B	Zero shot	**0.7405**	0.6729
Llama-4-Scout-17B-16E-Instruct	109B	Zero shot	0.7087	0.5874
Mistral-Small-3.2-24B-Instruct	24B	Zero shot	0.7282	0.6266
AIN Arabic MultiModel	8.29B	Zero Shot	0.6640	0.6057
Granite 3.3 2B Instruct	2.53B	Zero shot	0.4343	0.3648
Phi4 MultiModel Instruct	5.57B	Zero Shot	0.6235	0.5053
Granite 3.3 2B Instruct	2.53B	SFT on AraVQA$_{Train}$	0.5742	0.5746
Phi4 MultiModel Instruct	5.57B	SFT on AraVQA$_{Train}$	0.7009	**0.7384**

Table 3: The accuracy results of existing VLMs on the multiple-choice question–answering subset of the Pearl benchmark and AraVQA test set. The first section shows the results in the zero-shot setting, where no supervised fine-tuning (SFT) or few-shot examples were used. The second section shows the results after fine-tuning the Granite and Phi models on our AraVQA training set. All results in the table were generated by our team.

on the training portion of the AraVQA dataset. We evaluate the fine-tuned models on the multiple-choice question–answering subset of the Pearl benchmark as well as on our AraVQA test set. This experimental design enables us to quantify the contribution of AraVQA to improving the performance of existing VLMs and to examine their capability for cross-dataset generalization across Arabic visual question answering tasks.

4.2.1 Zero-Shot Evaluation

Our zero-shot results in Table 3 show that our dataset poses a greater challenge to Visual Language Models than the existing Pearl dataset. Most of the VLM models in the table show a $10\% - 12\%$ drop in performance when evaluated on our AraVQA test set compared to the Pearl dataset. This drop in accuracy suggests that our dataset poses a greater challenge to existing VLMs than the Pearl dataset. The results also indicate that the InternVL2-Llama3-76B model has the highest generalization ability among all other models, with only a 7% drop in performance. Overall, these results highlight the role that the AraVQA dataset can play in future research to detect out-of-distribution issues in the existing VQA datasets and evaluate the ability of VLMs to generalize across different domains.

4.2.2 Supervised SFT Evaluation

In contrast to the zero-shot evaluation, the Supervised Fine-Tuning (SFT) experiments aim to quantify the impact of training on the AraVQA dataset in enhancing VLM performance. As shown in the lower section of Table 3, fine-tuning on the AraVQA training split yields a substantial improvement of $21\% - 23\%$ in accuracy across both the Pearl and AraVQA test sets. This gain demonstrates that the dataset provides highly informative supervision signals that generalize beyond the training distribution.

Moreover, the performance gap between the two benchmarks narrows considerably following SFT, indicating that the degradation observed under the zero-shot condition stems largely from out-of-distribution effects. These findings highlight the capability of AraVQA to serve not only as a challenging evaluation benchmark but also as an effective training resource for improving multimodal reasoning in Arabic-centric VLMs.

4.2.3 Visual Grounding Analysis

Table 4 presents the accuracy of six vision-language models evaluated on two Arabic VQA benchmarks—**Pearl** and **AraVQA**—under two complementary conditions: one in which each question is paired with its correct image (*Related*), and another where the image is replaced by a random, unrelated one (*Random*). The resulting *Gap* column quantifies the performance drop between these conditions and thus serves as a direct measure of each model's dependence on visual grounding rather than linguistic priors. Formally, for each model m, we define the grounding gap as:

$$\Delta_m = A_m^{\text{related}} - A_m^{\text{random}}, \quad (1)$$

where A_m^{related} and A_m^{random} denote the model's accuracy under the related and random image conditions, respectively.

Model	Pearl Test			AraVQA Test		
	Related	Random	Δ_m	Related	Random	Δ_m
Llama-3.2-11B-Vision-Instruct	0.6505	0.5026	0.1479	0.5621	0.4051	0.1570
AIN Arabic MultiModel	0.6640	0.5325	0.1315	0.6057	0.4531	0.1526
Llama-4-Scout-17B-16E-Instruct	0.7087	0.6323	0.0764	0.5874	0.4999	0.0875
Qwen2-VL-72B-Instruct	0.7268	0.6016	0.1252	0.6369	0.4937	0.1432
Mistral-Small-3.2-24B-Instruct	0.7282	0.5555	0.1727	0.6266	0.4665	0.1601
InternVL2-Llama3-76B	0.7405	0.4993	0.2412	0.6729	0.4135	0.2594

Table 4: Accuracy results on Pearl MCQ and AraVQA test sets when models are evaluated using randomly assigned images instead of the relevant images.

Across both benchmarks, all models exhibit a consistent degradation when visual information is perturbed, validating that Arabic VQA tasks elicit genuine cross-modal reasoning rather than text-only inference. However, the extent of this degradation varies substantially across architectures, revealing systematic differences in the strength and robustness of visual–textual integration. **InternVL2-Llama3-76B** attains the highest overall accuracy yet displays the largest grounding gap ($\Delta \approx 0.25$–0.26), indicating a strong but brittle reliance on visual cues. In contrast, **Llama-4-Scout-17B-16E** manifests the smallest gap ($\Delta \approx 0.08$–0.09), suggesting that its relative stability stems from an overreliance on textual correlations. **Qwen2-VL-72B-Instruct**, **Mistral-Small-3.2-24B-Instruct**, and **AIN Arabic MultiModel** occupy an intermediate range ($\Delta \approx 0.13$–0.17), reflecting partial but uneven visual grounding.

Overall, results from **Pearl** and **AraVQA** converge on a coherent trend: current Arabic MLLMs demonstrate measurable visual sensitivity but limited multimodal robustness. The *random-image* paradigm therefore constitutes a rigorous and unified diagnostic for disentangling true visual reasoning from spurious textual priors, offering a principled framework for evaluating multimodal grounding in Arabic VQA.

4.2.4 Error Analysis and Insights

To better understand the qualitative behavior of the best-performing model, we conducted a detailed analysis of 100 randomly sampled evaluation examples from the AraVQA test split. This scrutiny set was uniformly sampled to ensure balanced coverage of question types and difficulty levels. The analysis reveals systematic and interpretable patterns in the model's learned behavior following supervised fine-tuning (SFT). In the zero-shot setting, the base PHI model exhibits clear limitations in *visual grounding*, particularly for interrogatives such as أين (where), متى (when), and من (who), which require reasoning over spatial, temporal, or identity cues within the image. After SFT, these same examples are correctly answered, demonstrating that the fine-tuning data effectively taught multimodal alignment skills absent from the pre-trained model. Across the 100 examples, this improvement is both qualitative and consistent: SFT converted guess-like predictions into grounded, evidence-based reasoning. For example, in the question أين يقع ميناء القطيف على الخريطة؟ ("Where is Qatif Port located on the map?"), the zero-shot model incorrectly selected في البحر الأحمر ("in the Red Sea"), whereas the fine-tuned model correctly localized it as في الخليج العربي ("in the Arabian Gulf"). Similarly, for متى تم التقاط الصورة التي تُظهر فيها سفينة تايتانيك؟ ("When was the picture showing the Titanic taken?"), the base model produced an arbitrary date, but PHI-SFT accurately chose ٢ أبريل ١٩١٢ ("2 April 1912"), inferring the temporal context from historical visual cues. A comparable improvement appears in identity recognition: for من هو الملك الظاهر في الصورة؟ ("Who is the king shown in the picture?"), the zero-shot model predicted فريدريك الأول ("Frederick I"), while the fine-tuned version correctly answered بلدوين الثاني ("Baldwin II"). These examples typify a broader trend in which fine-tuning enables robust Arabic-specific grounding in *space*, *time*, and *identity*—three domains where cross-lingual transfer from English vision–language data remains weakest. Smaller yet meaningful gains also occur in "what" (ما) questions involving object, text, or concept recognition. For instance, in ما هو الجين المتسبب في مرض ويلسون كما هو

موضح في الصورة؟ (“What gene causes Wilson’s disease as shown in the picture?”), the zero-shot model predicted “*HFE*,” while the fine-tuned version correctly identified “*ATP7B*,” demonstrating that SFT enhanced the model’s ability to map biomedical images to factual entities.

Interestingly, we also observe a small subset of questions whose answers can be inferred directly from the textual content of the question itself, without requiring information from the accompanying image. For example, in من هو مؤلف كتاب الصناعتين؟ (“Who is the author of the book *Kitāb al-Sināatayn*?”), the question provides enough linguistic context for a knowledgeable model to respond “أبو هلال العسكري” (“Abu Hilal al-Askari”), yet the zero-shot PHI fails to answer correctly, selecting a random option instead. Only after fine-tuning does PHI-SFT provide the correct answer. This behavior indicates that AraVQA is not limited to assessing visual recognition but also diagnoses a model’s ability to leverage textual and world knowledge when the answer does not depend on the image. Although such cases are relatively few in the 100-item scrutiny set, their presence highlights the diagnostic value of AraVQA in distinguishing between visually grounded and textually sufficient reasoning. Residual errors remain primarily in causal or abstract reasoning, such as ما السبب في الظاهرة الظاهرة في الصورة؟ (“What is the cause of the phenomenon shown in the image?”), where success requires linking perceptual cues (e.g., glowing seawater) to underlying scientific explanations (e.g., bioluminescent dinoflagellates). These cases suggest that while SFT substantially improves *perceptual grounding*, it only partially enhances *higher-order conceptual reasoning*.

Overall, the 100-example analysis shows that fine-tuning reshaped the model’s internal behavior rather than merely boosting accuracy. PHI-SFT acquired the ability to answer Arabic multimodal questions through **grounded factual reasoning**, with the largest gains in spatial, temporal, and identity reasoning. These interpretable improvements validate the scientific value of our dataset: it delivers targeted supervision that systematically addresses weaknesses in multilingual vision–language models, enabling **genuine multimodal understanding in Arabic**. By bridging linguistic, cultural, and perceptual dimensions, our dataset provides a principled foundation for evaluating and enhancing grounded reasoning in low-resource multimodal contexts.

Question Word	Before	After	Improv.%
Which أي	6	10	66.67
Where أين	885	1295	46.33
What ما	755	994	31.66
When متى	536	894	66.79
Who is من هو	239	341	42.68
Total Questions	**2421**	**3534**	**45.97**

Table 5: Comparison of correct answers by question word on the AraVQA test set (4791) before and after fine-tuning Phi-4 Multimodal-Instruct on the AraVQA training set.

Question Word	Before	After	Improv.%
Which أي	54	59	9.26
Where أين	3	4	33.33
How many كم	2	4	100.00
How كيف	55	56	1.82
What ما	2375	2669	12.38
Who is من هو	10	9	-10.00
Total Questions	**2499**	**2801**	**12.08**

Table 6: Comparison of correct answers by question word on the Pearl test set (4009) before and after fine-tuning Phi-4 Multimodal-Instruct on the AraVQA training set.

4.2.5 Impact of Fine-Tuning by Question Word

Tables 5 and 6 present results before and after fine-tuning *Phi-4-Multimodal-Instruct* on the **AraVQA** training set. We evaluate performance on the **AraVQA$_{\text{test}}$** set and on **PEARL**, grouping results by question word.

On the **AraVQA$_{\text{test}}$** set, fine-tuning leads to substantial improvements across all question words, with an overall gain of **45.97%**. The largest improvements are observed for *“Which”* and *“When”* questions (both around **66%**), followed by *“Where”* (**46.33%**) and *“Who is”* (**42.68%**). These results indicate that training on **AraVQA** helps the model better align linguistic cues with visual content, particularly for questions involving location, time, and people.

On **PEARL**, fine-tuning also yields improvements, though the overall gains are smaller (**12.08%**). Notable improvements are observed

for *"How many"* questions (**100.00%**), *"Where"* questions (**33.33%**), and *"What is"* questions (**12.38%**). In contrast, *"How"* questions show minimal change (**1.82%**), while performance on *"Who is"* questions slightly decreases (**-10.00%**), suggesting limited cross-dataset transfer for certain question types.

5 Conclusion

This paper presented an end-to-end pipeline that leverages the Wikipedia template tags and the generative power of large language models (LLMs) to automatically construct visual question answering (VQA) datasets. Using this framework, we developed ARAVQA—the largest publicly available Arabic factoid VQA dataset—comprising over $50,000$ questions spanning a wide range of general-knowledge domains. The pipeline achieves high precision in question generation and produces visually grounded, semantically coherent items suitable for downstream vision–language modeling.

We further introduced a benchmark derived from ARAVQA and validated it through systematic human annotation. Fine-tuning compact Vision–language models (VLMs), including GRANITE-3.3-2B and PHI-4-MULTIMODAL, on 45,969 training examples from ARAVQA yields substantial performance gains—improving accuracy by up to 20 percentage points on both the PEARL benchmark and the held-out ARAVQA test set. Remarkably, these fine-tuned models surpass VLMs exceeding 70B parameters, demonstrating that fine-tuning small vision–language models with a high-quality dataset can achieve competitive multimodal reasoning even with limited model capability.

In addition, our analysis reveals that ARAVQA constitutes a challenging and diagnostic benchmark for Arabic multimodal reasoning, exposing persistent gaps in visual grounding and cross-domain generalization. Collectively, the dataset, benchmark, and empirical findings establish a scalable foundation for advancing Arabic vision–language research and for developing culturally grounded, linguistically faithful visual understanding systems.

6 Future Work

Building on the success of our AraVQA pipeline, our next step is to extend the framework to additional low-resource languages represented in Wikipedia. This effort will assess the scalability and language-agnostic capacity of our approach, enabling the automatic construction of high-quality, culturally grounded VQA datasets across diverse linguistic settings. To achieve this, we plan to adapt the pipeline to handle complex morphology (e.g., Swahili), right-to-left orthographies (e.g., Urdu), and language-specific cultural semantics reflected in visual content, thereby supporting more inclusive data generation and annotation practices.

Beyond dataset creation, we aim to leverage these multilingual resources to advance model development and evaluation. We will explore cross-lingual transfer learning through multilingual vision–language pretraining to enable unified models capable of visual reasoning across languages and knowledge transfer from high-resource to low-resource contexts. In parallel, we plan to establish equitable VQA benchmarks for underrepresented languages such as Swahili, Urdu, and Amharic, promoting fair and culturally informed evaluation. Ultimately, this research seeks to bridge the linguistic divide in vision–language AI by fostering models that are both performant and culturally aware, contributing to a more inclusive and globally representative multimodal ecosystem.

Limitations

As our work focuses more on factoid visual question answering, we have less representation in the dataset for questions that address domains such as OCR (Optical Character Recognition) or logic reasoning. Our AraVQA dataset is aimed more toward serving as a complementary dataset to existing datasets that address these types of tasks (e.g., Pearl, CAMEL-Bench). In contrast to existing Arabic visual question answering datasets, AraVQA is aimed more toward assessing and improving the general knowledge inside Vision-Language Models (VLMs). However, in this work, we propose a pipeline method that can be generalized across other domains in future work.

In this work, we also choose to work with the open-source GPT-OSS-120B model during the synthetic question generation process rather than using other larger commercial LLMs (e.g., OpenAI GPT-4, Google Gemini). The decision is driven by our goal to illustrate the simplicity and affordability of our proposed pipeline that can be used with a single commercial GPU (e.g., RTX 5090, H100), which allows researchers interested in working with limited-resource domains to improve this

method further.

Furthermore, one of the limitations we encountered during the SFT evaluation presented in Table 3 is the out-of-memory issue when we attempted to fine-tune larger VLMs, as our work environments consist of a single H100 GPU. Despite this limitation, both models that we fine-tuned on our AraVQA dataset: Granite 3.3 2B Vision and Phi4 MultiModel 5B demonstrate a significant increase across all tasks. These two models outperform other existing VLMs (+70B parameters), including InternVL2-Llama3-76B, Qwen2-VL-72B-Instruct, and Llama-4-Scout-17B-16E-Instruct. Therefore, we can expect a substantial improvement across Arabic visual question tasks if we fine-tune larger VLM models (e.g., Llama-3.2-11B-Vision-Instruct).

Ethics Statement

The AraVQA dataset is synthetically generated from articles in Arabic Wikipedia. These articles are written and curated by human contributors and reviewed by Wikipedia editors. Wikipedia follows the Neutral Point of View (NPOV) policy, which is defined as "representing fairly, proportionately, and, as far as possible, without editorial bias, all significant views that have been published by reliable sources on a topic."

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. Command r7b arabic: A small, enterprise focused, multilingual, and culturally aware arabic llm. *Preprint*, arXiv:2503.14603.

Fakhraddin Alwajih, Samar Mohamed Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, and 26 others. 2025. Pearl: A multimodal culturally-aware arabic instruction dataset. *Preprint*, arXiv:2505.21979.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Radman, Jorma Laaksonen, Fahad Shahbaz Khan, Salman Khan, and Rao Muhammad Anwer. 2025. CAMEL-bench: A comprehensive Arabic LMM benchmark. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1970–1980, Albuquerque, New Mexico. Association for Computational Linguistics.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2502.00094.

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. *Preprint*, arXiv:2503.21910.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *Preprint*, arXiv:2306.05425.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Meta. 2025. Llama-4-scout-17b-16e-instruct. Machine learning model.

Meta AI. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*. Llama 3.2-11B-Vision-Instruct model, updated September 25, 2024.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Mistral AI. 2025. Mistral-Small-3.2-24B-Instruct-2506. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, Nimrod Shabtay, Pengyuan Li, Roei Herzig, Shafiq Abedin, Shaked Perek, Sivan Harary, Udi Barzelay, Adi Raz Goldfarb, Aude Oliva, and 44 others. 2025. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *Preprint*, arXiv:2502.09927.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *Preprint*, arXiv:2212.10773.

A Supplementary Materials

This appendix provides additional material that complements the main paper and supports reproducibility. The first subsection, §A.1, presents supplementary technical details, including the structure of the GPT-OSS prompt, representative AraVQA dataset examples, and an illustration of the data extraction process from Wikipedia. The second subsection, §A.2, describes the data annotation process, including the annotation interface and the guidelines followed by annotators to ensure consistency and quality.

A.1 GPT-OSS Prompt Specification and AraVQA Examples

This subsection provides supplementary details that support the methodology described in the main paper. Appendix Figure 2 illustrates the structure of the GPT-OSS prompt employed in our pipeline, showing how instructions, visual inputs, and expected outputs are organized within a unified prompt template. This figure offers implementation-level clarity that complements the high-level description provided in the main sections.

Appendix Figure 3 presents representative samples from the AraVQA dataset, highlighting the alignment between images, Arabic questions, and their corresponding answers. Finally, Appendix Figure 4 details the automated data extraction process from Wikipedia template tags. In addition, the figure illustrates how image URLs are constructed by computing the MD5 hash of the image filename and using the initial characters of the hash to generate the full URL. This visualization clarifies a critical step in the dataset construction pipeline.

GPT OSS 120B Prompt

ID : {id} (ignore this id in the prompt. This is just to retrieve original data)

Your task is to create a multi-choice vision question answering in Arabic Language.

I will give you a list of snippets taken from Arabic Wikipedia, and with each snippet, I will provide you with: A. The description of the image attached to this snippet. B. The article title in which this snippet has been written. C. The snippet title.

Read the whole lines and then form a multi-choice QA about the image following the guidelines below:

1. The question should start with one of the following question words (translated to Arabic): 'What', 'When', 'Where', 'Who'.
2. If the image description contains a date and this date is related to the image, then your question must start with "When" in the Arabic Language.
3. If the snippet has a place where the entity inside the image is located, then your question must start with 'Where' in the Arabic Language.
4. A few questions should ask to explain or describe the image in the Arabic Language. Do not make many of them.
5. The multi-choice question should be in open-domain trivia style, meaning we will not need the related paragraph to answer the question, but the image is needed.
6. Please do not explain the reasons behind creating each question. Your output should only contain the question, four choices, and the answer.
7. The answer should be only one of the choices, not two choices.
8. The answer should be exactly what is written in the text.
9. Generate only one question.
10. Do not always generate questions when the given lines do not have things that meet the given conditions. In this case, your output should be only "Not Applicable".
11. All questions should be in the Arabic Language.
12. You should not give any hint for the answer from the text because this is a visual question answering.
13. Multiple choices should be different. You cannot have the same choice.
14. All questions should be about factoid questions.
15. Do not give any hint from the snippet that could allow the reader to answer the question without having the image (e.g., Which area is London located in?).
16. The distractors (wrong choices) should not be easy to guess or close to each other (e.g., A. 19 Jan 2019 B. 20 Jan 2019 C. 21 Jan 2019).
17. Follow the following template when you create question-answer pairs (do not add **):

Question: [question]

A. [Choice 1]
B. [Choice 2]
C. [Choice 3]
D. [Choice 4]

Answer Letter: [answer letter]

Answer: [answer]

Only one question and answer. Do not create many.

Please find below the text snippet and the image description:

Article Title: {article_title} Text: {section_text} Image Name: {img_path} Image Description: {img_desc}

Figure 2: Structure of the GPT-OSS prompt used in our framework.

Where are the waterfalls shown in the picture located?
A. Between the borders of Argentina and Brazil
B. On the border between Canada and the United States
C. On the border between China and India
D. On the border between Australia and New Zealand
Answer: A. Between the borders of Argentina and Brazil

أين تقع الشلالات التي تظهر في الصورة؟
أ. بين حدود الأرجنتين والبرازيل .
ب. على الحدود بين كندا والولايات المتحدة الأمريكية
ج. على الحدود بين الصين والهند
د. على الحدود بين أستراليا ونيوزيلندا
الجواب : أ. بين حدود الأرجنتين والبرازيل .

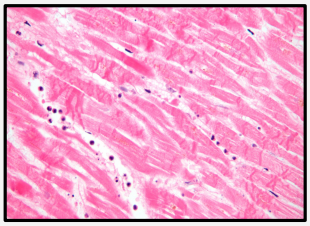
What are the names of the two teams shown in the picture ?
A. Argentina and North Korea
B. Germany and South Korea
C. Brazil and Honduras
D. Brazil and North Korea
Answer: D. Brazil and North Korea

ما هو اسم المنتخبين الظاهرين في الصورة؟
آ. الأرجنتين وكوريا الشمالية
ب. ألمانيا وكوريا الجنوبية
ج. البرازيل والهندوراس
د. البرازيل وكوريا الشمالية
الجواب : د. البرازيل وكوريا الشمالية

What plant is shown in the picture?
A. Basil
B. Thyme
C. Coriander
D. Mint
Answer: D. Mint

ما هو النبات المعروض في الصورة؟
أ. ريحان
ب. زعتر
ج. كزبرة
د. نعناع
الجواب : د. النعناع

What is the histopathological appearance seen in the microscopic image?
A. Myofibrosis
B. Contractile band necrosis
C. Muscle hypertrophy
D. Blood clot
Answer: B. Contractile band necrosis

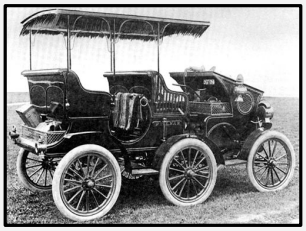


ما هو المظهر المرضي النسيجي الظاهر في الصورة المجهرية؟
أ. تليف عضلي
ب. نخر الشريط الانقباضي
ج. تضخم عضلي
د. تجلط دموي
الجواب : ب. . نخر الشريط الانقباضي

When was the astrolabe pictured made?
A. 1310
B. 1285
C. 1291
D. 1302
Answer: C. 1291

متى صُنع الأسطرلاب الموجود في الصورة؟
أ. 1310
ب. 1285
ج. 1291
د. 1302
الجواب : ج . 1291

How many wheels does the Pullman car shown in the picture have ?
A. Seven wheels
B. Six wheels
C. Five wheels
D. Four wheels
Answer: B. Six wheels



ما هو عدد العجلات في سيارة بولمان المعروضة في الصورة؟
أ. سبع عجلات
ب. ست عجلات
ج. خمس عجلات
د. أربع عجلات
الجواب : ب. ست عجلات

What is the name of the dish shown in the picture?
A. Kibbeh
B. Stuffed grape leaves
C. Maqluba
D. Mansaf
Answer: B. Stuffed grape leaves

ما هو اسم الطبق المعروض في الصورة؟
إ. كبة
ب. ورق العنب
ج. مقلوبة
د. منسف
الجواب : ب. ورق العنب

Figure 3: Representative examples from the AraVQA dataset.

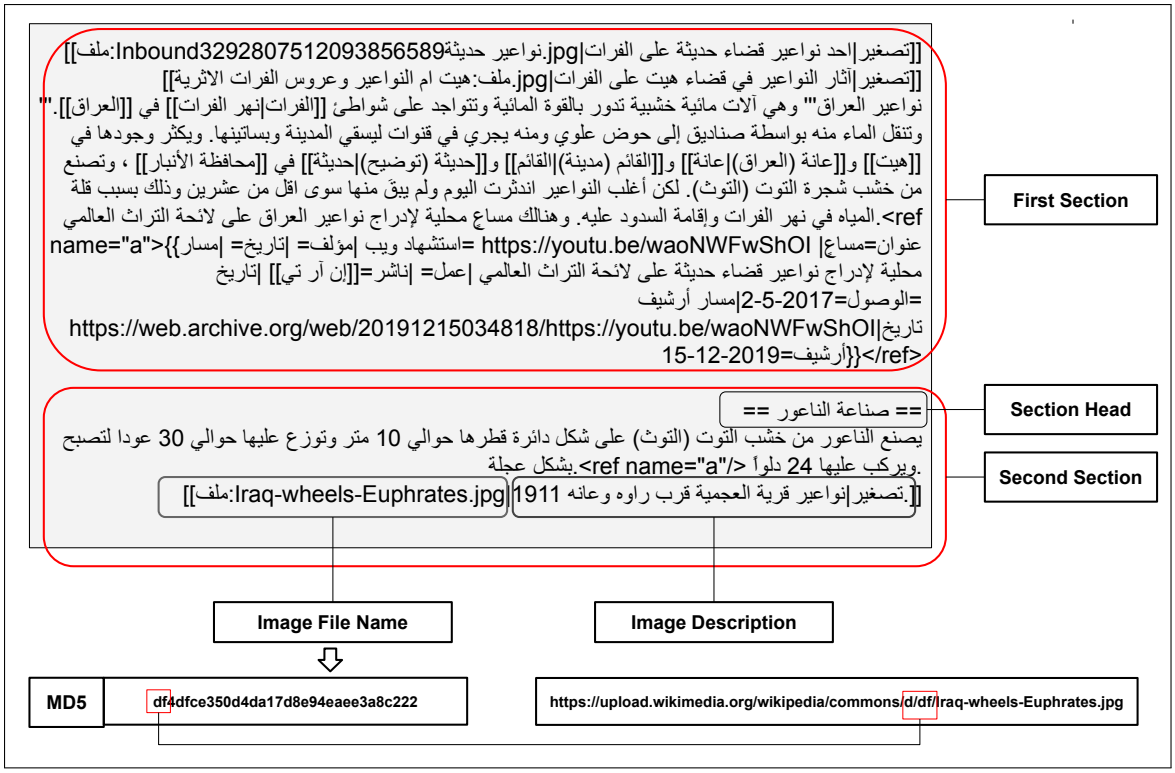


Figure 4: Overview of the data extraction pipeline from Wikipedia template tags, including image URL generation using MD5 hashing.

A.2 Human Annotation Interface and Guidelines

This subsection describes the annotation setup used to construct and validate the dataset. Appendix Figure 6 shows the user interface provided to annotators, including the visual layout and components used to review images and submit annotations. The interface was designed to facilitate efficient and consistent annotation.

Appendix Figure 5 presents the annotation guidelines distributed to annotators prior to the labeling task. These guidelines define the annotation criteria, quality requirements, and consistency rules followed throughout the annotation process, ensuring reliability and uniformity across annotations.

Data Description: This dataset was collected from Wikipedia, where we pair each image with its context taken from the passage related to each image. Each example has a question, four choices, an answer, a related image, and a description (caption) for the related image.

Annotator task: The main task for the annotator is just to verify that each question is factually correct. Thus, it's simply a binary task for an annotator, and this task has two possible outcomes (correct, not correct).
The annotators will be able to verify each question by comparing the actual image with the image description first. If both the image and image description are enough to verify the answer, then the final decision will be "correct". If these two inputs are not enough, the annotators will read the "section" text on the bottom right for further verification. Then the final decision will be made.

Things to pay attention to while doing the annotation :
A. If the answer exists in the question, then the example should be flagged as incorrect.
B. In some questions, the information given inside the question is enough to answer the question without having the image. Do not flag this question as incorrect in this case. If everything is ok with this question, other than this issue, then it should be flagged as correct.
C. If the answer can be more than one choice, then flag it as incorrect

Figure 5: Annotation guidelines provided to annotators.

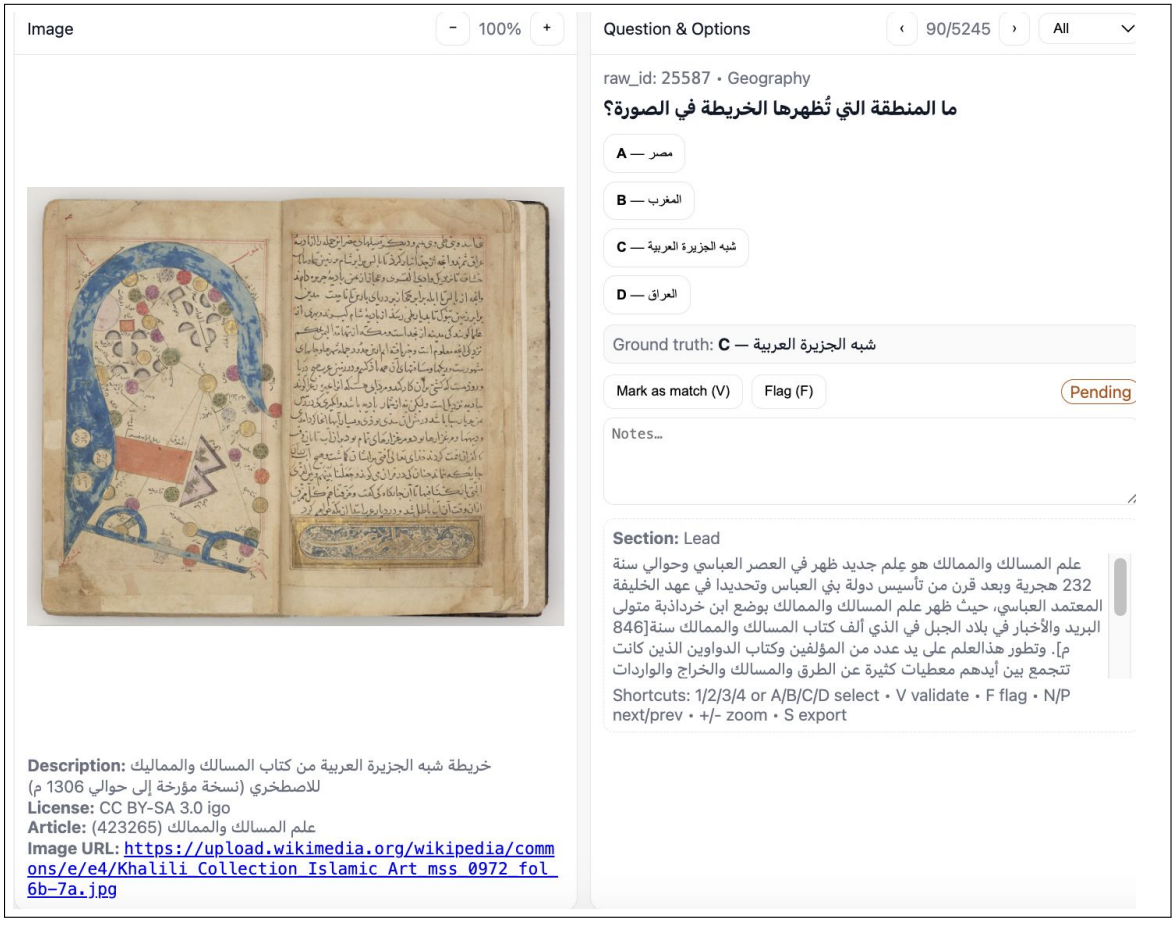


Figure 6: Annotation interface used during the data labeling process.